Introduction to Real Analysis

Liviu I. Nicolaescu

University of Notre Dame

Last revised 2025-05-02, 08:30:56.

Notes for the Honors Calculus class at the University of Notre Dame.
Started August 21, 2013. Completed on April 30, 2014.
Last modified on 2025-05-02,08:30:56.

Introduction

What follows started as notes for the Freshman Honors Calculus course at the University of Notre Dame. The word “calculus” is a misnomer since this course was intended to be an introduction to real analysis or, if you like, “calculus with proofs”.

For most students this class is the first encounter with mathematical rigor and it can be a bit disconcerting. In my view the best way to overcome this is to confront rigor head on and adopt it as standard operating procedure early on. This makes for a bumpy early going, but with a rewarding payoff.

A proof is an argument that uses the basic rules of Aristotelian logic and relies on facts everyone agrees to be true. The course is based on these basic rules of the mathematical discourse. It starts from a meagre collection of obvious facts (postulates) and ends up constructing the main contours of the impressive edifice called real analysis.

No prior knowledge of calculus is assumed, but being comfortable performing algebraic manipulations is something that will make this journey more rewarding.

In writing these notes I have benefitted immensely from the students who took the Honors Calc Course during the academic years 2013-2016. Their questions and reactions in class, and their expert editing have improved the original product. I asked a lot of them and I got a lot in return. I want to thank them for their hard work, curiosity and enthusiasm which made my job so much more enjoyable.

The first 9 chapters correspond to subjects covered in the Freshman course. I rarely was able to complete the brief Chapter 10 on complex numbers. Chapters 11 and above deal with several variables calculus topics, corresponding to the sophomore Honors Calculus offered at the University of Notre Dame.

This is probably not the final form of the notes, but close to final. I will probably adjust them here and there, taking into account the feedback from future students.

Notre Dame, 2025-05-02.

The Greek Alphabet

A	α	Alpha	N	ν	Nu
B	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	O	o	Omicron
Δ	δ	Delta	Π	π	Pi
E	ε	Epsilon	P	ρ	Rho
Z	ζ	Zeta	Σ	σ	Sigma
H	η	Eta	T	τ	Tau
Θ	θ	Theta	Υ	υ	Upsilon
I	ι	Iota	Φ	φ	Phi
K	κ	Kappa	X	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	Mu	Ω	ω	Omega

Contents

Chapter 1

The basics of mathematical reasoning

1.1. Statements and predicates

Mathematics deals in *statements*. These are sentences that have a definite truth value. What does this mean? The classical text [24] does a marvelous job explaining this point of view. I will not even attempt a rigorous or exhaustive explanation. Instead, I will try to suggest it to you through examples.

Example 1.1.1. (a) Consider the following sentence: "*if you walk in the rain without an umbrella, you will get wet*". This is a true sentence and we say that its truth value is $TRUE$ or T. This is an example of a *statement*.

(b) Consider the sentence: "*the number x is bigger than the number y*" or, in mathematical notation, $x > y$. This sentence could be $TRUE$ or $FALSE$ (F), depending on the choice of x and y. This is not a statement because it does not have a definitive truth value. It is a type of sentence called *predicate* that is encountered often in mathematics.

A *predicate* or *formula* is a sentence that depends on some parameters (or variables). In the above example the parameters were x and y. For some choices of parameters (or variables) it becomes a $TRUE$ statement, while for other values it could be $FALSE$.

When expressed in everyday language, statements and predicates must contain a verb.

Often a predicate comes in the guise of a *property*. For example the property "*the integer n is even*" stands for the predicate "*the integer n is twice an integer m*".

(c) Consider the following sentence: " *This sentence is false.*" Is this sentence true? Clearly it cannot be true because if it were, then we would conclude that the sentence is false. Thus the sentence is false so the opposite must be true, i.e., the sentence is true. Something is obviously amiss. This type of sentence is *not* a statement because it does not have a truth value, and it is also *not* a predicate. It is a *paradox*. Paradoxes are to be avoided in mathematics. □

✍ **Notation.** It is time to explain the usage of the notation :=. For example an expression such as

$$x := \text{bla-bla-bla}$$

reads "*x is defined to be bla-bla-bla*", or "*x is short-hand for bla-bla-bla*".

The manipulations of statements and predicates are governed by the rules of *Aristotelian logic*. This and the following section will provide you with a very sparse introduction to logic. For more details and examples I refer to [**37**].

All the predicates used in mathematics are obtained from simpler ones called *atomic predicates* using the following *logical operators*.

- *NEGATION* $\neg$ (read as *not*).
- *CONJUNCTION* $\wedge$ (read as *and*).
- *DISJUNCTION* $\vee$ (read as *or*).
- *IMPLICATION* $\Rightarrow$ (read as *implies*).

To describe the effect of these operations we need to look Table 1.1 describing the *truth tables* of these operations.

p	T	F
$\neg p$	F	T

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

Table 1.1. The truth tables of $\neg, \wedge, \vee, \Rightarrow$

Here is how one reads Table 1.1. When p is true (T), then $\neg p$ must be false (F), and when p is false, then $\neg p$ is true. To put it in simpler terms

$$\neg T = F, \;\; \neg F = T.$$

The truth table for $\wedge$ can be presented in the simplified form

$$T \wedge T = T, \;\; T \wedge F = F \wedge T = F \wedge F = F.$$

Observe that the predicate $p \vee q$ is true when at least one of the predicates p and q is true. *It is NOT an exclusive OR.* Another way of saying this is

$$T \vee T = T \vee F = F \vee T = T, \;\; F \vee F = F.$$

The equivalence $\Leftrightarrow$ is the operation

$$p \Leftrightarrow q := (p \Rightarrow q) \wedge (q \Rightarrow p).$$

Its truth table is described in Table 1.2

p	q	$p \Leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

Table 1.2. The truth table of $\Leftrightarrow$

Remark 1.1.2. (a) In everyday language, when we say that p *implies* q we mean that the statement $p \Rightarrow q$ is true. This signifies that either both p and q are true, or that p is false. Often we express this in the conditional form *if* p*, then* q.

If the implication $p \Rightarrow q$ is true, then we say that q *is a necessary condition for* p and that p *is a sufficient condition for* q. In everyday language the implications are the *if* bla-bla, *then* bla-bla statements.

The truth table for $\Rightarrow$ hides certain subtleties best illustrated by the following example. Consider the statement

$$s := \text{if an elephant can fly, then it can also drive a car.}$$

This statement is composed of two simpler statements

$$p := \text{an elephant can fly},\quad q := \text{an elephant can drive a car}.$$

We note that the statement s coincides with the implication $p \Rightarrow q$. Obviously, both statements p and q are false, but according to the truth table for $\Rightarrow$, the implication $p \Rightarrow q$ is true, and thus s is true as well. This conclusion is rather unsettling. It may be easier to accept it if we rephrase s as follows:

> *if you can show me a flying elephant, then I can show you that it can also drive a car.*

(b) In everyday language when we say that p *is equivalent to* q we mean that the statement $p \Leftrightarrow q$ is true. This signifies that either both p and q are true, or both are false. If p is equivalent to q, we say that q *is a necessary and sufficient condition for* p and that p *is a necessary and sufficient condition for* q.

We often express this in one of the following forms: p *if and only if* q. The mathematicians' abbreviation for the oft encountered construct "*if and only if*" is *iff*. □

Example 1.1.3. Consider the following predicate.

$s :=$ *if you do not clean your room, then you will not go to the movies.*

Figure 1.1. *The elusive flying elephant.*

This is composed of two simpler predicates

- $p :=$ you do not clean the room.
- $q :=$ you do not go to the movies.

Observe that s is the compound predicate $p \Rightarrow q$. For s to be true, one of the following two mutually exclusive situations must happen:

- either you do not clean your room AND you do not go to the movies A
- or you clean the room.

Note that there is no implied guarantee that if you clean your room, then you go to the movies. □

Example 1.1.4. Consider the following *true* statement: mathematicians like to be precise.

First, let us phrase this in a less ambiguous way. The above statement can be equivalently rephrased as: if you are a mathematician, then you are precise. To put it in symbolic terms

$$\underbrace{\text{you are a mathematician}}_{p} \Rightarrow \underbrace{\text{you are precise}}_{q}.$$

Thus, to be a mathematician it is necessary to be precise and to be precise it suffices to be a mathematician. However, to be precise it is not necessary to be a mathematician. □

A *tautology* is a compound predicate which is true no matter what the truth values of its atoms are.

Example 1.1.5. The predicate $p \vee \neg p$ is a tautology. In other words, in mathematics, a statement is either true, or false. There is no in-between. □

Two compound predicates P and Q are called *equivalent*, and we indicate this with the notation $P \longleftrightarrow Q$, if they have identical truth tables. In other words, P and Q are equivalent if the compound predicate $P \Longleftrightarrow Q$ is a tautology.

Example 1.1.6. Let us observe that the compound predicate $p \Rightarrow q$ is equivalent to the compound statement $(\neg p) \vee q$, i.e.

$$p \Rightarrow q \longleftrightarrow (\neg p) \vee q. \tag{1.1.1}$$

Indeed if p is false then $p \Rightarrow q$ and $\neg p$ are true, no matter what q. In particular $(\neg p) \vee q$ is also true, no matter what q. If p is true, then $\neg p$ is false, and we deduce that $p \Rightarrow q$ and $(\neg p) \vee q$ are either simultaneously true, or simultaneously false. □

1.2. Quantifiers

Example 1.2.1. Consider the following property of a person x

$$x \textit{ is at least } 6ft \textit{ tall.}$$

This does not have a definite truth value because the truth value depends on the person x. However the claims

$$C_1 := \textit{there exists a} \text{ person } x \text{ that is at least 6ft tall,}$$

and

$$C_2 := \textit{any} \text{ person } x \text{ is at least 6ft tall}$$

have definite truth values. The claim C_1 is true, while the claim C_2 is false. □

Example 1.2.2. Consider the following property involving the numbers x, y

$$x > y.$$

This does not have a definite truth value. However, we can modify it to obtain statements that have definite truth values. Here are several possible modifications. (Below and in the sequel the abbreviation s.t. stands for *such that*)

$$S_1 := \underline{\text{For any}}\ x,\ \underline{\text{for any}}\ y,\ x > y.$$

$$S_2 := \underline{\text{For any}}\ x,\ \underline{\text{there exists}}\ y \text{ s.t. } x > y.$$

$$S_3 := \underline{\text{There exists}}\ y \text{ s.t. } \underline{\text{for any}}\ x, \quad x > y.$$

Observe that the statements S_1 and S_3 are false, while S_2 is a true statement. Notice a *very important fact*. The statement S_3 is obtained from S_2 by a seemingly innocuous transformation: we changed the order of some words. However, in doing so, we have dramatically altered the meaning of the statement. *Let this be a warning!* □

The expressions *for any*, *for all*, *there exists*, *for some* appear very frequently in mathematical communications and for this reason they were given a name, and special abbreviations. These expressions are called *quantifiers* and they are abbreviated as follows.

$$\forall := \text{for any, for all,}$$

$$\exists := \text{there exists, there exist, for some.}$$

The symbol $\forall$ is also called *the universal quantifier*, while the symbol $\exists$ is called *the existential quantifier*. There is another quantifier encountered quite frequently namely

$$\exists! := \text{there exists a unique.}$$

The above examples illustrate the roles of the quantifiers: they are used to convert predicates, which have no definite truth value, to statements which have definite truth value. To achieve this, we need to attach a quantifier to each variable in the predicate. In Example 1.2.2 we used a quantifier for the variable x and a quantifier for the variable y. We cannot overemphasize the following fact.

☞ ***The meaning of a statement is sensitive to the order of the quantifiers in that statement!***

Example 1.2.3. Let us put to work the above simple principles in a concrete situation. Consider the statement:

$S :=$ *there is a person in this class such that, if he or she gets an A in the final, then everyone will get an A in the final.*

Is this a true statement or a false statement? There are two ways to decide this. The fastest way is to think of the persons who get the lowest grade in the final. If those persons get A's, then, obviously, everybody else will get A's.

We can use a more formal way of deciding the truth value of the above statement. Consider the predicate $P(x) :=$*the person x gets an A in the final.* The quantified form of S is then

$$\exists x : \;\; \bigl(P(x) \Rightarrow \forall y P(y)\bigr).$$

As we know, an implication $p \Rightarrow q$ is equivalent to the disjunction $\neg p \vee q$; see (1.1.1). We can rewrite the above statement as

$$\exists x : \;\; \bigl(\neg P(x) \vee \forall y P(y)\bigr).$$

In everyday language the above statement says that either there is a person who did not get an A or everybody gets an A. This is a *Duh!* statement or, as mathematicians like to call it, a *tautology*. □

Let us discuss how to concretely describe the negation of a statement involving quantifiers. Take for example the statements S_1, S_2, S_3 in Example 1.2.2. Their opposites are

$$\neg S_1 := \;\underline{\text{There exists}}\; x,\; \underline{\text{there exists}}\; y \text{ s.t. } x \leqslant y,$$
$$\neg S_2 := \;\underline{\text{There exists}}\; x \text{ s.t. } \underline{\text{for any}}\; y\text{: } x \leqslant y$$

,

$$\neg S_3 := \;\underline{\text{For any }}\; y,\; \underline{\text{there exists}}\; x \text{ s.t. } x \leqslant y.$$

Observe that all the opposites were obtained by using the following simple operations.

- Globally replacing the existential quantifier $\exists$ with its opposite $\forall$.

- Globally replace the universal quantifier $\forall$ with its opposite, $\exists$.
- Replace the predicate $x > y$ with its opposite, $x \leqslant y$.

When dealing with more complex statements it is very useful to remember the above rules. We summarize them below.

☞ ***The opposite of a statement that contains quantifiers is obtained by replacing each quantifier with its opposite, and each predicate with its opposite.***

Example 1.2.4. Consider the following portion of a famous Abraham Lincoln quote: *you can fool all of the people some of the time.* There are two conceivable ways of phrasing this rigorously.

1. *For any person y there exists a moment of time t when y can be fooled by you at time t.*

2. *There exists a moment of time t such that any person y there can be fooled by you at time t.*

We can now easily transform these into *quantified statements*.

1. $S_1 :=$ $\forall$ *person* y, $\exists$ *moment* t, *s.t.*, y *can be fooled by you at time* t.

2. $S_2 :=$ $\exists$ *moment* t, *s.t.*, $\forall$ *person* y, y *can be fooled by you at time* t.

Note that the two statements carry different meanings. Which do you think was meant by Lincoln? Observe also

$\neg S_1 :=$ $\exists$ *person* y *s.t.* $\forall$ *moment* t*:* y *cannot be fooled by you at time* t.

In plain English this reads: *some people cannot be fooled at any time.* □

1.3. Sets

Now that we have learned a bit about the language of mathematics, let us mention a few fundamental concepts that appear in all the mathematical discourses. The most important concept is that of *set*.

Any attempt at a rigorous definition of the concept of set unavoidably leads to treacherous logical and philosophical marshes.[1] A more productive approach is not to define what a set is, but agree on a list of "uncontroversial" properties (or *axioms*) our intuition tells us the sets ought to satisfy.[2] Once these axioms are adopted, then the entire edifice of mathematics should be built on them. I refer to [25] for a detailed description of this point of view.

The axiomatic approach mentioned above is very labor intensive, and would send us far astray. Our goal for now is a bit more modest. We will adopt a more elementary

[1] For more details on the possible traps; see Wikipedia's article on set theory.

[2] See the above footnote.

(or naive) approach relying on the intuition of a set X as a collection of objects, usually referred to as the *elements of* X. In mathematics, a set is described by the "list" of its elements enclosed by brackets. In this list, no two objects are identical. For example, the set

$$\{winter, spring, summer, fall\}$$

is the set of seasons in a temperate region such as Indiana. However, the list

$$\{winter, winter, spring\}$$

is not a set.

We will use the notation $x \in X$ (or $X \ni x$) to indicate that the *object x belongs to* the *set X*, i.e., the object x is an element of X. The notation $x \notin X$ indicates that x is not an element of X. Two sets A and B are considered identical if they consist of the same elements, i.e., the following (quantified) statement is true

$$\forall x \,\bigl(\, x \in A \Longleftrightarrow x \in B\,\bigr).$$

In words, an object belongs to A iff it also belongs to B. For example, we have the equality of sets

$$\{winter, spring, summer, fall\} = \{spring, summer, fall, winter\}.$$

There exists a distinguished set, called the *empty set* and denoted by $\varnothing$. Intuitively, $\varnothing$ is the set with no elements.

Remark 1.3.1. The nature of the elements of a set is not important in set theory. In fact, the elements of a set can have varied natures. For example, we have the set

$$\bigl\{\, 1,\ \varnothing,\ apple\,\bigr\}$$

which consists of three elements: of the number 1, the empty set, and the word *apple*. Another more subtle example is the set $\{\varnothing\}$ which consists of the single element, the empty set $\varnothing$. Let us observe that $\varnothing \neq \{\varnothing\}$. □

We say that a set A is a *subset* of B, and we write this $A \subset B$, if any element of A is also an element of B. In other words, $A \subset B$ signifies that the following statement is true:

$$\forall x \,\bigl(\, x \in A \Rightarrow x \in B\,\bigr).$$

A *proper subset* of B is a subset $A \subset B$ such that $A \neq B$. We will use the notation $A \subsetneq B$ to indicate that A is a proper subset of B.

The *union* of two sets A, B is a new set denoted by $A \cup B$. More precisely,

$$x \in A \cup B \iff (x \in A) \vee (x \in B).$$

The *intersection* of two sets A, B is a new set denoted by $A \cap B$. More precisely,

$$x \in A \cap B \iff (x \in A) \wedge (x \in B).$$

The sets A and B are said to be *disjoint* if $A \cap B = \varnothing$.

More generally, if $(A_i)_{i\in I}$ is a collection of sets, then we can define their union

$$\bigcup_{i\in I} A_i := \bigl\{\, x;\ \ \exists i \in I :\ \ x \in A_i \,\bigr\},$$

and their intersection

$$\bigcap_{i\in I} A_i := \bigl\{\, x;\ \ \forall i \in I;\ \ x \in A_i \,\bigr\}.$$

The *difference* between a set A and a set B is a new set $A\backslash B$ defined by

$$x \in A\backslash B \iff (x \in A) \wedge (x \notin B).$$

If A is a subset of X, then we will use the alternative notation C_XA when referring to the difference $X\backslash A$. The set C_XA is called the *complement of A in X*. Observe that

$$C_X\bigl(C_XA\bigr) = A.$$

It is sometimes convenient to visualize sets using *Venn diagrams*. A Venn diagram identifies a set with a region in the plane.

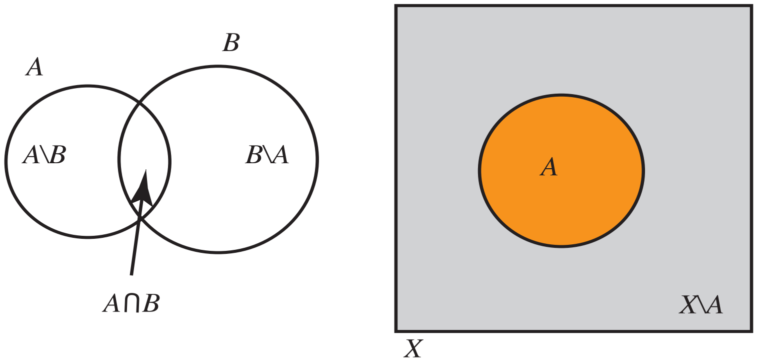

Figure 1.2. *Venn diagrams.*

Proposition 1.3.2 (De Morgan Laws)**.** *If A, B are subsets of a set X then*

$$C_X(A\cup B) = (C_XA)\cap(C_XB),\ \ C_X(A\cap B) = (C_XA)\cup(C_XB). \qquad \square$$

Given two sets A and B we can form a new set $A\times B$ which consists of all ordered pairs of objects (a,b) where $a\in A$ and $b\in B$. The set $A\times B$ is called the *Cartesian product* of A and B.

Remark 1.3.3. As a curiosity, and to give you a sense of the intricacies of the axiomatic set theory, let us point out that above the concept of *ordered pair*, while intuitively clear, it is not rigorous. One rigorous definition of an ordered pair is due to *Norbert Wiener* who defined the ordered pair (a,b) to be the set consisting of two elements that are themselves sets: one element is the set $\{a,\varnothing\}$ and the other element is the set $\{\,\{b\}\,\}$, i.e.,

$$(a,b) := \bigl\{\,\{a,\varnothing\},\{\,\{b\}\,\}\,\bigr\}. \qquad \square$$

Most of the time sets are defined by properties. For example, the interval $[0,1]$ consists of the real numbers x satisfying the property

$$P(x) := (x \geqslant 0) \wedge (x \leqslant 1).$$

As we discussed in the previous section, a synonym for the term *property* is the term *predicate*. Proving that an object satisfying a property P also satisfies a property Q is tantamount to showing that the set of objects satisfying property P is contained in the set of objects satisfying property Q.

Remark 1.3.4. To prove that two sets A and B are equal one has to prove two inclusions: $A \subset B$ and $B \subset A$. In other words one has to prove two facts:

- If x is in A, then x is also in B.
- If x is in B, then x is also in A.

□

1.4. Functions

Suppose that we are given two sets X, Y. Intuitively, a *function* f from X to Y is a "device" that feeds on elements of X. Once we feed this machine an element $x \in X$ it spits out an element of Y denoted by $f(x)$. The elements of X are called *inputs*, and those of Y, *outputs*. In Figure 1.3 we have depicted such a device. Each arrow starts at some input and its head indicates the resulting output when we feed that input to the function f.

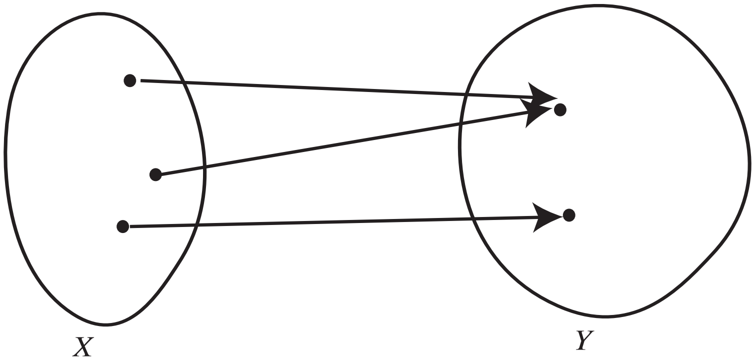

Figure 1.3. *A Venn diagram depiction of a function from X to Y.*

The above definition may not sound too academic, but at least it gives an idea of what a function is supposed to do. Mathematically, a function is described by listing its effect on each and every one of the inputs $x \in X$. The result is a list G which consists of pairs $(x,y) \in X \times Y$, where the appearance of a pair (x,y) in the list indicates the fact that when the device is fed the input x, the output will be y. Note that the list G is a subset of $X \times Y$ and has two properties.

- For any $x \in X$ there exists $y \in Y$ such that $(x, y) \in G$. Symbolically
$$\forall x \in X \;\; \exists y \in Y : \;\; (x, y) \in G. \tag{F_1}$$
- For any $x \in X$ and any $y_1, y_2 \in Y$, if $(x, y_1), (x, y_2) \in G$, then $y_1 = y_2$. Symbolically
$$\forall x \in X, \;\; \forall y_1, y_2 \in Y, \Big((x, y_1) \in G \wedge (x, y_2) \in G \Big) \Rightarrow (y_1 = y_2). \tag{F_2}$$

Property F_1 states that to any input there corresponds at least one output, while property F_2 states that each input has at most one output.

Definition 1.4.1. A function from X to Y is a subset of $X \times Y$ satisfying the conditions (F_1) and (F_2) above. □

We can use any symbol to name functions. The notation $f : X \to Y$ indicates that f is a function from X to Y. Often we will use the alternate notation $X \xrightarrow{f} Y$ to indicate that f is a function from X to Y. In mathematics there are many synonyms for the term function. They are also called *maps*, *mappings*, *operators*, or *transformations*.

Given a function $f : X \to Y$ we will refer to the set of inputs X as the *domain* of the function. The set Y is called the *codomain* of f. The result of feeding f the input $x \in X$ is denoted by $f(x)$. By definition $f(x) \in Y$. We say that x *is mapped to* $f(x)$ by f. The set
$$G_f := \big\{ (x, f(x)); \;\; x \in X \big\} \subset X \times Y$$
lists the effect of f on each possible input $x \in X$, and it is usually referred to as the *graph* of f.

The *range* or *image* of a function $f : X \to Y$ is the set of all outputs of f. More precisely, it is the subset $f(X)$ of F defined by
$$f(X) := \big\{ y \in Y; \;\; \exists x \in X : \;\; y = f(x) \big\}.$$
The range of f is also denoted by $\boldsymbol{R}(f)$. More generally, for any subset $A \subset X$ we define
$$f(A) = \big\{ y \in Y; \;\; \exists a \in A; \;\; f(a) = y \big\} \subset Y. \tag{1.4.1}$$
The set $f(A)$ is called the *image* of A via f.

For a subset $S \subset Y$, we define the *preimage* of S via f to be the set of all inputs that are mapped by f to an element in S. More precisely the preimage of S is the set
$$f^{-1}(S) := \big\{ x \in X; \;\; f(x) \in S \big\} \subset X. \tag{1.4.2}$$
When S consists of a single point, $S = \{y_0\}$ we use the simpler notation $f^{-1}(y_0)$ to denote the preimage of $\{y_0\}$ via f. The set $f^{-1}(y_0)$ is a subset of X called the *fiber of f over y_0*.

A function $f : X \to Y$ is called *surjective*, or *onto*, if $f(X) = Y$. Using the visual description of a function given in Figure 1.3 we see that a function is onto if any element in Y is hit by an arrow originating at some element $x \in X$. Symbolically

$$f : X \to Y \text{ is surjective} \Longleftrightarrow \forall y \in Y,\ \ \exists x \in X : \ \ y = f(x).$$

A function $f : X \to Y$ is called *injective*, or *one-to-one*, if different inputs have different outputs under f. More precisely

$$f : X \to Y \text{ is injective} \Longleftrightarrow \forall x_1, x_2 \in X : \ \ x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$$

$$\Longleftrightarrow \forall x_1, x_2 \in X : \ \ f(x_1) = f(x_2) \Rightarrow x_1 = x_2.$$

A function $f : X \to Y$ is called *bijective* if it is both injective and surjective. We see that

$$f : X \to Y \text{ is bijective} \Longleftrightarrow \forall y \in Y \ \ \exists!\, x \in X : \ \ y = f(x).$$

Example 1.4.2. (a) For any set X we denote by $\mathbb{1}_X$ or by e_X the function $X \to X$ which maps any $x \in X$ to itself. The function $\mathbb{1}_X$ is called the *identity map*. The identity map is clearly injective.

(b) Suppose that X, Y are two sets. We denote π_X the mapping $X \times Y \to X$ which sends a pair (x, y) to x. We say that π_X is the *natural projection* of $X \times Y$ onto X.

(c) Given a function $f : X \to Y$ and a subset $A \subset X$ we can construct a new function $f|_A : A \to Y$ called the *restriction* of f to A and defined in the obvious way

$$f|_A(a) = f(a), \ \ \forall a \in A.$$

(d) If X is a set and $A \subset X$, then we denote by i_A the function $A \to X$ defined as the restriction of $\mathbb{1}_X$ to A. More precisely

$$i_A(a) = a, \ \ \forall a \in A.$$

The function i_A is called the *natural inclusion map* associated to the subset $A \subset X$. □

Given two functions

$$X \xrightarrow{f} Y, \ \ Y \xrightarrow{g} Z$$

we can form their *composition* which is a function $g \circ f : X \to Z$ defined by

$$g \circ f(x) := g\big(\, f(x)\,\big).$$

Intuitively, the action of $g \circ f$ on an input x can be described by the diagram

$$x \overset{f}{\mapsto} f(x) \overset{g}{\mapsto} g\big(\, f(x)\,\big).$$

In words, this means the following: take an input $x \in X$ and drop it in the device $f : X \to Y$; out comes $f(x)$, which is an element of Y. Use the output $f(x)$ as an input for the device $g : Y \to Z$. This yields the output $g\big(\, f(x)\,\big)$.

Proposition 1.4.3. *Let $f : X \to Y$ be a function. The following statements are equivalent.*

(i) *The function f is bijective.*

(ii) *There exists a function $g : Y \to X$ such that*

$$f \circ g = \mathbb{1}_Y, \quad g \circ f = \mathbb{1}_X. \tag{1.4.3}$$

(iii) *There exists a* ***unique*** *function $g : Y \to X$ satisfying (1.4.3).*

Proof. (i) ⇒ (ii) Assume (i), so that f is bijective. Hence, for any $y \in Y$ there exists a unique $x \in X$ such that $f(x) = y$. This unique x depends on y and we will denote it by $g(y)$; see Figure 1.4.

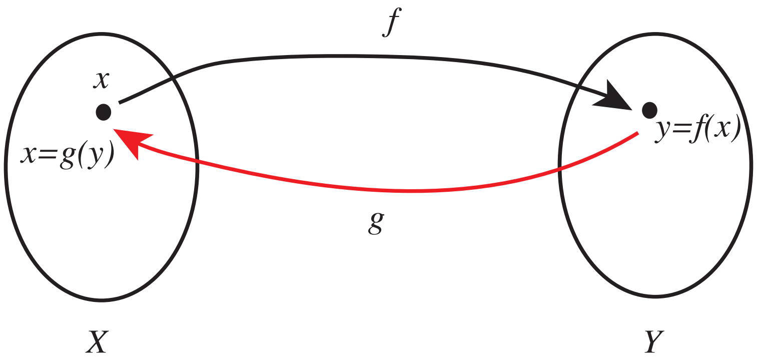

Figure 1.4. *Constructing the inverse of a bijective function $X \to Y$.*

The correspondence $y \mapsto g(y)$ defines a function $g : Y \to X$. By construction, if $x = g(y)$, then

$$y = f(x) = f\big(\,g(y)\,\big) = f \circ g(y) \;\; \forall y \in Y$$

so that $f \circ g = \mathbb{1}_Y$. Also, if $y = f(x)$, then

$$x = g(y) = g\big(\,f(x)\,\big) = g \circ f(x), \;\; \forall x \in X.$$

Hence $g \circ f = \mathbb{1}_X$. This proves the implication (i) ⇒ (ii)

(ii) ⇒ (iii) Assume (i). We need to show that if $g_1, g_2 : Y \to X$ are two functions satisfying (1.4.3), then $g_1 = g_2$, i.e., $g_1(y) = g_2(y)$, $\forall y \in Y$.

Let $y \in Y$. Set $x_1 = g_1(y)$. Then

$$f(x_1) = f\big(\,g_1(y)\,\big) = f \circ g_1(y) \overset{(1.4.3)}{=} y.$$

On the other hand,

$$g_2(y) = g_2\big(\,f(x_1)\,\big) = g_2 \circ f(x_1) \overset{(1.4.3)}{=} x_1 = g_1(y).$$

This proves the implication (ii) ⇒ (iii).

(iii) ⇒ (i). We assume that there exists a function $g : Y \to X$ satisfying (1.4.3) and we will show that f is bijective. We first prove that f is injective, i.e.,

$$\forall x_1, x_2 \in X : f(x_1) = f(x_2) \Rightarrow x_1 = x_2.$$

Indeed, if $f(x_1) = f(x_2)$, then

$$x_1 \overset{(1.4.3)}{=} g \circ f(x_1) = g\big(f(x_1)\big) = g\big(f(x_2)\big) = g \circ f(x_2) \overset{(1.4.3)}{=} x_2.$$

To prove surjectivity we need to show that for any $y \in Y$, there exists $x \in X$ such that $f(x) = y$. Let $y \in Y$. Set $x = g(y)$. Then

$$y \overset{(1.4.3)}{=} f \circ g(y) = f\big(g(y)\big) = f(x).$$

This proves the surjectivity of f and completes the proof of Proposition 1.4.3. □

Definition 1.4.4. Let $f : X \to Y$ be a bijective function. The *inverse* of f is the unique function $g : Y \to X$ satisfying (1.4.3). The inverse of a bijective function f is denoted by f^{-1}. □

1.5. Exercises

Exercise 1.1. Show that

$$\neg(p \vee q) \longleftrightarrow (\neg p \wedge \neg q), \quad \neg(p \wedge q) \longleftrightarrow \neg p \vee \neg q. \qquad \square$$

Exercise 1.2. (a) Show that

$$(p \Rightarrow q) \longleftrightarrow (\neg q \Rightarrow \neg p), \quad \neg(p \Rightarrow q) \longleftrightarrow (p \wedge \neg q).$$

(b) Consider the predicates

$$p := \text{the elephant } x \text{ can fly}, \quad q := \text{the elephant } x \text{ can drive}.$$

Let us stipulate that p is false. Show that the predicate $p \Rightarrow q$ is true by showing that its negation $\neg(p \Rightarrow q)$ is false. □

Exercise 1.3. Consider the exclusive-OR operation $\vee^*$ with truth table

p	q	$p \vee^* q$
T	T	F
T	F	T
F	T	T
F	F	F

Table 1.3. The truth table of "$\vee^*$"

Show that

$$(p \vee^* q) \longleftrightarrow (p \wedge \neg q) \vee (\neg p \wedge q) \longleftrightarrow (p \Longleftrightarrow \neg q) \longleftrightarrow (p \Rightarrow \neg q) \wedge (\neg p \Rightarrow q). \qquad \square$$

Exercise 1.4 (Modus ponens)**.** Show that the compound predicate

$$\big((p \Rightarrow q) \wedge p \big) \Rightarrow q$$

is a tautology. □

Exercise 1.5 (Modus tollens)**.** Show that the compound predicate

$$\big((p \Rightarrow q) \wedge \neg q \big) \Rightarrow \neg p$$

is a tautology. □

Exercise 1.6. Translate each of the following propositions into a *quantified statement* in standard form, write its symbolic negation, and then state its negation in words. (Use Example 1.2.4 as guide.)

(i) You can fool some of the people all of the time.

(ii) Everybody loves somebody sometime.

(iii) You cannot teach an old dog new tricks.

(iv) When it rains, it pours.

□

Exercise 1.7. Consider the following predicates.

$$P := \text{I will attend your party.}$$

$$Q := \text{I go to a movie.}$$

Rephrase the predicate

$$\text{I will attend your party unless I go to a movie}$$

using the predicates P, Q and the logical operators $\neg, \vee, \wedge, \Rightarrow$. □

Exercise 1.8. Give an example of three sets A, B, C satisfying the following properties

$$A \cap B \neq \varnothing,\;\; B \cap C \neq \varnothing,\;\; C \cap A \neq \varnothing,\;\;\; A \cap B \cap C = \varnothing.$$ □

Exercise 1.9. Suppose that A, B, C are three arbitrary sets. Show that

$$A \cap \big(B \cup C \big) = \big(A \cap B \big) \cup \big(A \cap C \big),$$
$$A \cup \big(B \cap C \big) = \big(A \cup B \big) \cap \big(A \cup C \big),$$

and

$$A \setminus \big(B \cup C \big) = \big(A \setminus B \big) \cap \big(A \setminus C \big).$$

(In the above equalities it should be understood that the operations enclosed by parentheses are to be performed first.)

Hint. Use Remark 1.3.4. □

Exercise 1.10. Suppose that $f : X \to Y$ is a function and $A, B \subset Y$ are subsets of the codomain. Prove that

$$f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B),\;\; f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B).$$

Hint. Take into account (1.4.2) and Remark 1.3.4. □

Exercise 1.11. Let $f : X \to Y$ be a map between the sets X, Y. Prove that f is one-to-one *if and only if* for any subsets $A, B \subset X$ we have

$$f(A \cap B) = f(A) \cap f(B).$$ □

Exercise 1.12. Suppose A, B are sets and $f : A \to B$ is a map.[3] Define the maps

$$\varphi : A \to A \times B,\;\; \rho : A \times B \to B$$

by setting

$$\varphi(a) := \big(a, f(a) \big),\;\; \forall a \in A,\;\; \rho(a, b) := b,\;\; \forall (a, b) \in A \times B.$$

Prove that the following hold.

(i) The map φ is injective.

[3] Recall that a map is a function.

(ii) The map ρ is surjective.

(iii) $f = \rho \circ \varphi$.

□

Exercise 1.13. Suppose that $f : X \to Y$ and $g : Y \to Z$ are two bijective maps. Prove that the composition $g \circ f$ is also bijective and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

□

Exercise 1.14. Suppose that $f : X \to Y$ is a function. Prove that the following statements are equivalent.

(i) The function f is injective.

(ii) There exists a function $g : Y \to X$ such that $g \circ f = \mathbb{1}_X$.

Exercise 1.15. Suppose that $f : X \to Y$ is a function. Prove that the following statements are equivalent.

(i) The function f is surjective.

(ii) There exists a function $g : Y \to X$ such that $f \circ g = \mathbb{1}_Y$.

□

1.6. Exercises for extra credit

Exercise* 1.1. Two old ladies left from A to B and from B to A at dawn heading towards one another along the same road. They met at noon, but did not stop, each carried on walking with the same speed as before they met. The first lady arrives at B at 4 pm, and the second lady arrives at A at 9 pm. What time was the dawn that day? □

Exercise* 1.2. A farmer must take a wolf, a goat and a cabbage across a river in a boat. However the boat is so small that he is able to take only one of the three on board with him. How should he transport all three across the river? (The wolf cannot be left alone with the goat, and the goat cannot be left alone with the cabbage.) □

Chapter 2

The Real Number System

Any attempt to define the concept of number is fraught with perils of a logical kind: we will eventually end up chasing our tails. Instead of trying to explain *what numbers are* it is more productive to explain *what numbers do*, and *how they interact with each other*.

In this section we gather in a coherent way some of the basic properties our intuition tells us that real numbers[1] ought to satisfy. We will formulate them precisely and we will declare, by fiat, that *these are true statements*. We will refer to these as the *axioms of the real number system*. (Things are a bit more subtle, but that's the gist of our approach.) All the other properties of the real numbers follow from these axioms. Such deductible properties are known in mathematics as *Proposition*s or *Theorem*s. The term *Theorem* is used sparingly and it is reserved to the more remarkable properties.

The process of deducing new properties from the already established ones is called a mathematical *proof*. Intuitively, a proof is a complete, precise and coherent explanation of a fact. In this course we will prove all of the calculus facts you are familiar with, and much more.

The first thing that we observe is that the real numbers, whatever their nature, form a set. We will encounter this set so often in our mathematical discourse that it deserves a short name and symbol. We will denote the set of real numbers by $\mathbb{R}$. More importantly this set of mysterious objects called numbers satisfies certain properties that we use every day. We take them for granted, and do not bother to prove them. These are the axioms of the real numbers and they are of three types.

- Algebraic axioms.

[1] You may know them as *decimal numbers* or *decimals*.

- Order axioms.
- The completeness axiom.

In this chapter we discuss these axioms in some details and then we show some of their immediate consequences.

Remark 2.0.1. There is one rather delicate issue that we do not address in these notes. We introduce a set of objects whose nature we do not explain and then we take for granted that they satisfy certain properties.

Naturally, one should ask if such things exist, because, for all we know, we might be investigating the set of flying elephants. This is a rather subtle question, and answering it would force us to dig deep at the foundations of mathematics. Historically, this question was settled relatively recently during the twentieth century but, mercifully, science progressed for two millennia before people thought of formulating and addressing this issue. To cut to the chase, no, we are not investigating flying elephants. □

2.1. The algebraic axioms of the real numbers

Another thing we know from experience is that we can operate with numbers. More precisely we can add, subtract, multiply and divide real numbers. Of these four operations, the addition and the multiplication are the fundamental ones. These are special instances of a more general mathematical concept, that of *binary operation*.

A binary operation on a set S is, by definition, a function $S\times S\to S$. Loosely, a binary operation is a gizmo that feeds on *ordered* pairs of elements of S, processes such a pair in some fashion, and produces a single element of S. We list the first axioms describing the set of real numbers.

Axiom 1. The set $\mathbb{R}$ of real numbers $\mathbb{R}$ is equipped with two binary operations,

- *addition*
$$+:\mathbb{R}\times\mathbb{R}\to\mathbb{R},\;\;(x,y)\mapsto x+y,$$
- and *multiplication*
$$\cdot:\mathbb{R}\times\mathbb{R}\to\mathbb{R},\;\;(x,y)\mapsto x\cdot y.$$

□

The operation of multiplication is sometimes denoted by the symbol ×.

Axiom 2. The addition is *associative*, i.e.,

$$\forall x,y,z\in\mathbb{R};\;\;(x+y)+z=x+(y+z).$$ □

The usage of parentheses $(\,-\,)$ indicates that we first perform the operation enclosed by them.

Axiom 3. The addition is *commutative*, i.e.,

$$\forall x,y\in\mathbb{R}:\;\;x+y=y+x.$$ □

Axiom 4. An *additive identity element* exists. This means that there exists at least one real number u such that

$$x+u=u+x=x,\quad \forall x\in\mathbb{R}. \tag{2.1.1}$$

□

Before we proceed to our next axiom, let us observe that there exists precisely one additive identity element.

Proposition 2.1.1. *If $u_0, \hat{u}_0\in\mathbb{R}$ are additive identity elements, then $u_0=\hat{u}_0$.*

Proof. Since u_0 is an identity element, if we choose $x=\hat{u}_0$ in (2.1.1) we deduce that

$$\hat{u}_0+u_0=u_0+\hat{u}_0=\hat{u}_0.$$

On the other hand, $\hat{u}_0$ is also an identity element and if we let $x=u_0$ in (2.1.1) we conclude that

$$u_0+\hat{u}_0=\hat{u}_0+u_0=u_0.$$

Thus $u_0=\hat{u}_0$. □

Definition 2.1.2. The unique additive identity element of $\mathbb{R}$ is denoted by 0. □

Axiom 5. *Additive inverses exist.* More precisely, this means that for any $x\in\mathbb{R}$ there exists at least one real number $y\in\mathbb{R}$ such that

$$x+y=y+x=0.$$ □

We have the following result whose proof is left to you as an exercise.

Proposition 2.1.3. *Additive inverses are unique. This means that if x, y, y' are real numbers such that*

$$x+y=y+x=0=x+y'=y'+x,$$

then $y=y'$. □

Definition 2.1.4. The unique additive inverse of a real number x is denoted by $-x$. Thus

$$x+(-x)=(-x)+x=0,\quad \forall x\in\mathbb{R}.$$ □

Axiom 6. The multiplication is *associative*, i.e.,

$$\forall x,y,z\in\mathbb{R};\quad (x\cdot y)\cdot z=x\cdot(y\cdot z).$$ □

Axiom 7. The multiplication is *commutative*, i.e.,

$$\forall x,y\in\mathbb{R}:\quad x\cdot y=y\cdot x.$$ □

Axiom 8. A *multiplicative identity element* exists. This means that there exists at least one <u>nonzero</u> real number u such that

$$x\cdot u=u\cdot x=x,\quad \forall x\in\mathbb{R}. \tag{2.1.2}$$

□

Arguing as in the proof of Proposition 2.1.1 we deduce that there exists precisely one multiplicative identity element. We denote it by 1. We define

$$2 := 1 + 1,\quad x^2 := x \cdot x,\quad \forall x \in \mathbb{R}. \tag{✎}$$

Axiom 9. *Multiplicative inverses exist.* More precisely, this means that for any $x \in \mathbb{R}$, $x \neq 0$, there exists at least one real number $y \in \mathbb{R}$ such that

$$x \cdot y = y \cdot x = 1.$$ □

Proposition 2.1.3 has a multiplicative counterpart that states that multiplicative inverses are unique. The multiplicative inverse of the *nonzero* real number x is denoted by x^{-1}, or $1/x$, or $\frac{1}{x}$. Also, we will frequently use the notation

$$\frac{x}{y} := x \cdot y^{-1},\quad y \neq 0.$$

☞ ***The real number zero does not have an inverse. For this reason division by zero is an illegal and very dangerous operation. NEVER DIVIDE BY ZERO!***

Axiom 10. *Distributivity.*

$$\forall x, y, z \in \mathbb{R}:\quad x \cdot (y + z) = x \cdot y + x \cdot z.$$ □

✍ *To save energy and time we agree to replace the notation $x \cdot y$ with the simpler one, xy, whenever no confusion is possible.*

Definition 2.1.5. A set satisfying Axioms 1 through 10 is called a *field.* □

The above axioms have a number of "obvious" consequences.

Proposition 2.1.6. (i) *$\forall x \in \mathbb{R}$, $x \cdot 0 = 0$.*

(ii) *$\forall x, y \in \mathbb{R}$, $(xy = 0) \Rightarrow (x = 0) \vee (y = 0)$.*

(iii) *$\forall x \in \mathbb{R}$, $-x = (-1) \cdot x$.*

(iv) *$\forall x \in \mathbb{R}$, $(-1) \cdot (-x) = x$.*

(v) *$\forall x, y \in \mathbb{R}$, $(-x) \cdot (-y) = xy$.*

Proof. We will prove only part (i). The rest are left as exercises. Since 0 is the additive identity element we have $0 + 0 = 0$ and

$$x \cdot 0 = x \cdot (0 + 0) = x \cdot 0 + x \cdot 0.$$

If we add $-(x \cdot 0)$ to both sides of the equality $x \cdot 0 = x \cdot 0 + x \cdot 0$ we deduce $0 = x \cdot 0$. □

2.2. The order axiom of the real numbers

Experience tells us that we can compare two real numbers, i.e., given two real numbers we can decide which is smaller than the other. In particular, we can decide whether a number is positive or not. In more technical terms we say that we can *order* the set of real numbers. The next axiom formalizes this intuition.

Axiom 11. There exists a subset $\boldsymbol{P} \subset \mathbb{R}$ called the *subset of positive real numbers* satisfying the following two conditions.

(i) If x and y are in $\boldsymbol{P}$, then so are their sum and product, $x+y \in \boldsymbol{P}$ and $xy \in \boldsymbol{P}$.

(ii) If $x \in \mathbb{R}$, then *exactly one* of the following statements is true:

$$x \in \boldsymbol{P}, \text{ or } x = 0, \text{ or } -x \in \boldsymbol{P}.$$

□

Definition 2.2.1. Let $x, y \in \mathbb{R}$.

(i) We say that x is *negative* if $-x \in \boldsymbol{P}$.

(ii) We say that x is *greater than* y, and we write this $x > y$, if $x - y$ is positive. We say that x is *less than* y, written $x < y$, if y is greater than x.

(iii) We say that x is *greater than or equal to* y, and we write this $x \geqslant y$, if $x > y$ or $x = y$. We say that x is *less than or equal to* y, and we write this $x \leqslant y$, if $y \geqslant x$.

(iv) A real number x is called *nonnegative* if $x \geqslant 0$.

□

Observe that $x > 0$ signifies that $x \in \boldsymbol{P}$.

Proposition 2.2.2. (i) *$1 > 0$, i.e. $1 \in \boldsymbol{P}$.*

(ii) *If $x > y$ and $y > z$, then $x > z$, $x, y, z \in \mathbb{R}$.*

(iii) *If $x > y$, then for any $z \in \mathbb{R}$, $x + z > y + z$.*

(iv) *If $x > y$ and $z > 0$, then $xz > yz$.*

(v) *If $x > y$ and $z < 0$, then $xz < yz$.*

Proof. We will prove only (i) and (ii). The proofs of the other statements are left to you as exercises. To prove (i) we argue by contradiction. Thus we assume that $1 \notin \boldsymbol{P}$. By Axiom 8, $1 \neq 0$, so Axiom 11 implies that $-1 \in \boldsymbol{P}$ and $(-1) \cdot (-1) \in \boldsymbol{P}$. Using Proposition 2.1.6(v) we deduce that

$$1 = (-1) \cdot (-1) \in \boldsymbol{P}.$$

We have reached a contradiction which proves (i).

To prove (ii) observe that

$$x > y \Rightarrow x - y \in \boldsymbol{P}, \quad y > z \Rightarrow y - z \in \boldsymbol{P}$$

so that

$$x - z = (x - y) + (y - z) \in \boldsymbol{P} \Rightarrow x > z.$$

□

Definition 2.2.3 (Intervals)**.** Let $a, b \in \mathbb{R}$. We define the following sets.

(i) $(a,b) =]a,b[:= \{ x \in \mathbb{R};\ \ a < x < b \}$.

(ii) $(a,b] =]a,b] := \{ x \in \mathbb{R};\ \ a < x \leqslant b \}$.

(iii) $[a,b) = [a,b[:= \{ x \in \mathbb{R};\ \ a \leqslant x < b \}$.

(iv) $[a,b] := \{ x \in \mathbb{R};\ \ a \leqslant x \leqslant b \}$.

(v) $[a,\infty) = [a,\infty[:= \{ x \in \mathbb{R};\ \ a \leqslant x \}$.

(vi) $(a,\infty) =]a,\infty[:= \{ x \in \mathbb{R};\ \ a < x \}$.

(vii) $(-\infty,a) =]-\infty,a[:= \{ x \in \mathbb{R};\ \ x < a \}$.

(viii) $(-\infty,a] =]-\infty,a] := \{ x \in \mathbb{R};\ \ x \leqslant a \}$.

A subset I of $\mathbb{R}$ is called an *interval* if $I = \mathbb{R}$ or if it is of one of the types (i)-(viii). . The intervals of the form $[a,b]$, $[a,\infty)$, or $(-\infty,a]$ are called *closed*, while the intervals of the form (a,b), (a,∞), or $(-\infty,a)$ are called *open*. □

I would like to emphasize that in the above definition we made no claim that any or some of the intervals are nonempty. This is indeed the case, but this fact requires a proof.

Definition 2.2.4. For any $x \in \mathbb{R}$ we define the *absolute value* of x to be the quantity

$$|x| := \begin{cases} x & \text{if } x \geqslant 0, \\ -x & \text{if } x < 0. \end{cases}$$

□

Proposition 2.2.5. (i) *Let $\varepsilon > 0$. Then $|x| < \varepsilon$ if and only if $-\varepsilon < x < \varepsilon$, i.e.,*

$$(-\varepsilon,\varepsilon) = \{ x \in \mathbb{R};\ \ |x| < \varepsilon \}.$$

(ii) $x \leqslant |x|$, $\forall x \in \mathbb{R}$.

(iii) $|xy| = |x| \cdot |y|$, $\forall x,y \in \mathbb{R}$. *In particular,* $|-x| = |x|$

(iv) $|x+y| \leqslant |x| + |y|$, $\forall x,y \in \mathbb{R}$.

Proof. We prove only (i) leaving the other parts as an exercise. We have to prove two things,

$$|x| < \varepsilon \Rightarrow -\varepsilon < x < \varepsilon, \tag{2.2.1}$$

and

$$-\varepsilon < x < \varepsilon \Rightarrow |x| < \varepsilon. \tag{2.2.2}$$

To prove (2.2.1) let us assume that $|x| < \varepsilon$. We distinguish two cases. If $x \geqslant 0$, then $|x| = x$ and we conclude that $-\varepsilon < 0 \leqslant x < \varepsilon$. If $x < 0$, then $|x| = -x$ and thus $0 < -x = |x| < \varepsilon$. This implies $-\varepsilon < -(-x) = x < 0 < \varepsilon$.

Conversely, let us assume that $-\varepsilon < x < \varepsilon$. Multiplying this inequality by -1 we deduce that $-\varepsilon < -x < \varepsilon$. If $0 \leqslant x$, then $|x| = x < \varepsilon$. If $x < 0$ then $|x| = -x < \varepsilon$. □

Definition 2.2.6. The distance between two real numbers x, y is the nonnegative number $\operatorname{dist}(x, y)$ defined by

$$\operatorname{dist}(x, y) := |x - y|. \qquad \square$$

Very often in calculus we need to solve *inequalities*. The following examples describe some simple ways of doing this.

Example 2.2.7. (a) Suppose that we want to find all the real numbers x such that

$$(x-1)(x-2) > 0.$$

To solve this inequality we rely on the following simple principle: the product of two real numbers is positive if and only if both numbers are positive or both numbers are negative; see Exercise 2.8. In this case the answer is simple: the numbers $(x-1)$ and $(x-2)$ are both positive iff $x > 2$ and they are both negative iff $x < 1$. Hence

$$(x-1)(x-2) > 0 \iff x \in (-\infty, 1) \cup (2, \infty).$$

(b) Consider the more complicated problem: find all the real numbers x such that

$$(x-1)(x-2)(x-3) > 0.$$

The answer to this question is also decided by the multiplicative rule of signs, but it is convenient to organize or work in a table. In each of row we read the sign of the quantity listed at the beginning of the row. The signs in the bottom row are obtained by multiplying the signs in the column above them. We read

x	$-\infty$		1		2		3		∞
$(x-1)$	$-\infty$	$- - - -$	0	$+ + +$	$+$	$+ + +$	$+$	$+ + +$	∞
$(x-2)$	$-\infty$	$- - - -$	$-$	$- - -$	0	$+ + +$	$+$	$+ + +$	∞
$(x-3)$	$-\infty$	$- - - -$	$-$	$- - -$	$-$	$- - -$	0	$+ + +$	∞
$(x-1)(x-2)(x-3)$	$-\infty$	$- - - -$	0	$+ + +$	0	$- - -$	0	$+ + +$	∞

$$(x-1)(x-2)(x-3) > 0 \iff x \in (1, 2) \cup (3, \infty).$$

(c) Consider the related problem: find all the real numbers x such that

$$\frac{(x-1)}{(x-2)(x-3)} \geqslant 0.$$

Before we proceed we need to eliminate the numbers $x = 2$ and $x = 3$ from our considerations because the denominator of the above fraction vanishes for these values of x and the ***division by* 0 *is an illegal operation***. We obtain a similar table

x	$-\infty$		1		2		3		∞
$(x-1)$	$-\infty$	$----$	0	$+++$	$+$	$+++$	$+$	$+++$	∞
$(x-2)$	$-\infty$	$----$	$-$	$---$	0	$+++$	$+$	$+++$	∞
$(x-3)$	$-\infty$	$----$	$-$	$---$	$-$	$---$	0	$+++$	∞
$\frac{(x-1)}{(x-2)(x-3)}$	$-\infty$	$----$	0	$+++$	$!$	$---$	$!$	$+++$	∞

The exclamation signs at the bottom row are warning us that for the corresponding values of x the fraction has no meaning. We read

$$\frac{(x-1)}{(x-2)(x-3)} \geqslant 0 \iff x \in [1,2) \cup (3,\infty). \qquad \square$$

Example 2.2.8. We want to discuss a question involving inequalities frequently encountered in real analysis. Consider the statement

$$P(M): \ \forall x \in \mathbb{R}, \ \ x > M \Rightarrow \left|\frac{x^2}{x^2+x-2} - 1\right| < \frac{1}{10}.$$

We want to show that there exists at least one positive number M such that $P(M)$ is true, i.e., we want to prove that the statement

$$\exists M > 0 \text{ such that, } \ \forall x \in \mathbb{R}, \ \ x > M \Rightarrow \left|\frac{x^2}{x^2+x-2} - 1\right| < \frac{1}{10}.$$

Let us observe that if $P(M)$ is true and $M' \geqslant M$, then $P(M')$ is also true. Thus, once we find one M such that $P(M)$ is true, then $P(M')$ is true for all $M' \in [M,\infty)$.

We are content with finding only one M such that $P(M)$ is true and the above observation shows that in our search we can assume that M is very large. This is a bit vague, so let us see how this works in our special case.

First, we need to make sure that our algebraic expression is well defined so we need to require that the denominator $x^2+x-2 = (x-1)(x+2)$ is not zero. Thus we need to assume that $x \neq 1, -2$. In particular, we will restrict our search for M to numbers larger than 1. We have

$$\left|\frac{x^2}{x^2+x-2} - 1\right| = \left|\frac{x^2-(x^2+x-2)}{x^2+x-2}\right| = \left|\frac{-x+2}{x^2+x-2}\right| = \left|\frac{x-2}{x^2+x-2}\right|.$$

Since we are investigating the properties of the last expression for $x > M > 1$ we deduce that for $x > 2$ both quantities $x-2$ and $(x-1)(x+2)$ are positive and thus

$$\left|\frac{x-2}{x^2+x-2}\right| = \frac{x-2}{x^2+x-2}.$$

We want this fraction to be small, smaller than $\frac{1}{10}$. Note that for $x > 2$ we have

$$\frac{x-2}{x^2+x-2} \leqslant \frac{x-1}{x^2+x-2} = \frac{x-1}{(x-1)(x+2)} = \frac{1}{x+2},$$

and

$$x > 2 \ \wedge \ \frac{1}{x+2} < \frac{1}{10} \Longleftrightarrow x+2 > 10 \Longleftrightarrow x > 8\ .$$

We deduce that if $x > 8$, then

$$\frac{1}{10} > \frac{1}{x+2} > \left| \frac{x^2}{x^2+x-2} - 1 \right|.$$

Hence $P(8)$ is true. □

2.3. The completeness axiom

Definition 2.3.1. Let $X \subset \mathbb{R}$ be a nonempty set of real numbers.

(i) A real number M is called an *upper bound* for X if

$$\forall x \in X: \quad x \leqslant M. \tag{2.3.1}$$

(ii) The set X is said to be *bounded above* if it admits an upper bound.

(iii) A real number m is called a *lower bound* for X if

$$\forall x \in X: \quad x \geqslant m. \tag{2.3.2}$$

(iv) The set X is said to be *bounded below* if it admits a lower bound.

(v) The set X is said to be *bounded* if it is bounded both above and below.

□

Example 2.3.2. (a) The interval $(-\infty, 0)$ is bounded above, but not below. The interval $(0, \infty)$ is bounded below, but not above, while the interval $(0, 1)$ is bounded. □

(b) Consider the set R consisting of positive real numbers x such that $x^2 < 2$. This set is not empty because $1^2 = 1 < 2$ so that $1 \in R$. Let us show that this set is bounded above. More precisely, we will prove that

$$x^2 < 2 \Rightarrow x \leqslant 2.$$

We argue by contradiction. Suppose that $x \in R$ yet $x > 2$. Then

$$x^2 - 2^2 = (x-2)(x+2) > 0.$$

Hence $x^2 > 2^2 > 2$ which shows that $x \notin R$. This contradiction proves that 2 is an upper bound for R. □

Definition 2.3.3. Let $X \subset \mathbb{R}$ be a nonempty set of real numbers.

(i) A *least upper bound* for X is an upper bound M with the following additional property: if M' is another upper bound of X, then $M \leqslant M'$.

(ii) A *greatest lower bound* for X is a lower bound m with the following additional property: if m' is another lower bound of X, then $m \geqslant m'$.

□

Thus, M is a least upper bound for X if

- $\forall x \in X$, $x \leqslant M$, and
- if $M' \in \mathbb{R}$ is such that $\forall x \in X$, $x \leqslant M'$, then $M \leqslant M'$.

Proposition 2.3.4. *Any nonempty set $X \subset \mathbb{R}$ admits at most one least upper bound, and at most one greatest lower bound.*

Proof. We prove only the statement concerning upper bounds. Suppose that M_1, M_2 are two least upper bounds. Since M_1 is a least upper bound, and M_2 is an upper bound we have $M_1 \leqslant M_2$. Similarly, since M_2 is a least upper bound we deduce $M_2 \leqslant M_1$. Hence $M_1 \leqslant M_2$ and $M_2 \leqslant M_1$ so that $M_1 = M_2$. □

Definition 2.3.5. Let $X \subset \mathbb{R}$ be a nonempty set of real numbers.

(i) The least upper bound of X, when it exists, is called the *supremum* of X and it is denoted by $\sup X$.

(ii) The greatest lower bound of X, when it exists, is called the *infimum* of X and it is denoted by $\inf X$.

□

Example 2.3.6. Suppose that $X = [0, 1)$. Then $\sup X = 1$ and $\inf X = 0$. Note that $\sup X$ is not an element of X. □

Proposition 2.3.7. *Let $X \subset \mathbb{R}$ be a nonempty set of real numbers and $M \in \mathbb{R}$. The following statements are equivalent.*

(i) $M = \sup X$.

(ii) *The number M is an upper bound for X and for any $\varepsilon > 0$ there exists $x \in X$ such that $x > M - \varepsilon$.*

Proof. (i) $\Rightarrow$ (ii) Assume that M is the least upper bound of X. Then clearly M is an upper bound and we have to show that for any $\varepsilon > 0$ we can find a number $x \in X$ such that $x > M - \varepsilon$.

Because M is the least upper bound and $M - \varepsilon < M$, we deduce that $M - \varepsilon$ is *not* an upper bound for X. In other words, the opposite of (2.3.1) must be true, i.e., there must exist $x \in X$ such that x is not less or equal to $M - \varepsilon$.

(ii) $\Rightarrow$ (i) We have to show that if M' is another upper bound then $M \leqslant M'$. We argue by contradiction. Suppose that $M' < M$. Then $M' = M - \varepsilon$ for some positive number ε. The assumption (ii) implies that $x > M - \varepsilon$ for some number $x \in X$ so that $M' = M - \varepsilon$ is not an upper bound. We reached a contradiction which completes the proof. □

The Completeness Axiom. *Any nonempty set of real numbers that is bounded above* ***admits*** *a least upper bound.* □

From the completeness axiom we deduce the following result whose proof is left to you as Exercise 2.22.

Proposition 2.3.8. *If the nonempty set $X \subset \mathbb{R}$ is bounded below, then it admits a greatest lower bound.* □

Definition 2.3.9. Let $X \subset \mathbb{R}$ be a nonempty subset.

(i) We say that X admits a *maximal element* if X is bounded above and $\sup X \in X$. In this case we say that $\sup X$ is the maximum of X and it is denoted by $\max X$.

(ii) We say that X admits a *minimal element* if X is bounded below and $\inf X \in X$. In this case we say that $\inf X$ is called the *minimum* of X and it is denoted by $\min X$.

□

Note that the interval $I = [0, 1)$ has no maximal element, but it has a minimal element

$$\min I = 0.$$

2.4. Visualizing the real numbers

The approach we have adopted in introducing the real numbers differs from the historical course of things. For centuries scientists did not bother to ask what are the real numbers, often relying on intuition to prove things. This lead to various contradictory conclusions which prompted mathematicians to think more carefully about the concept of number and to treat the intuition more carefully.

This does not mean that the intuition stopped playing an important part in the modern mathematical thinking. On the contrary, intuition is still the first guide, but it always needs to be checked and backed by rigorous arguments.

For example, you learned to visualize the numbers as points on a line called the *real line*. We will not even attempt to explain what a line is. Instead we will rely on our physical intuition of this geometric concept. The real line is more than just a line, it is a line enriched with several attributes.

- It has a distinguished point called the *origin* which should be thought of as the real number 0.
- It is equipped with an *orientation*, i.e., a direction of running along the line visually indicated by an arrowhead at one end of the line; see Figure 2.1. Equivalently, the origin splits the line into two sides, and choosing an orientation is equivalent to declaring one side to be the *positive side* and the other side to be

the *negative side*. Traditionally the above arrowhead points towards the positive side; see Figure 2.1

- There is a way of measuring the distance between two points on the line.

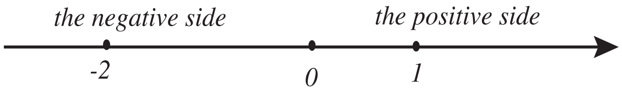

Figure 2.1. *The real line.*

For example, the number -2 can be visualized as the point on the negative side situated at distance 2 from the origin; see Figure 2.1.

Now that we have identified the set $\mathbb{R}$ of real numbers with the set of points on a line, we can visualize the Cartesian product $\mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$ with the set of points in a plane, called the *Cartesian plane*; see the top of Figure 2.2.

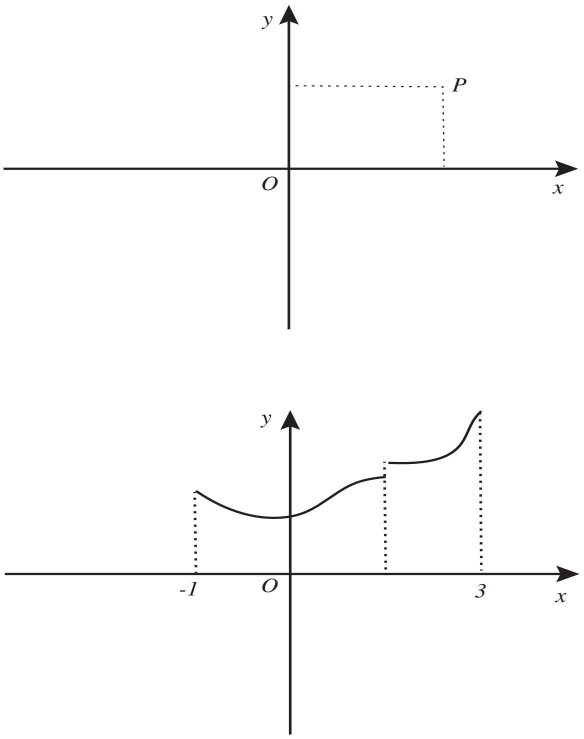

Figure 2.2. *The real line.*

Just like the real line, the Cartesian plane is more than a plane: it is a plane enriched by several attributes.

- It contains a distinguished point, called the origin and denoted by O.

- It contains two distinguished perpendicular lines intersecting at O. These lines are called the *axes* of the Cartesian plane. One of the axes is declared to be horizontal and the other is declared to be vertical.
- Each of these two axes is a real line, i.e., it has the additional attributes of a real line: each has a distinguished point, O, each has an orientation, and each is equipped with a way of measuring distances along that respective line. The horizontal axis is also known as the x-axis, while the vertical one is also known as the y-axis.

The position of a point P in that plane is determined by a pair of real numbers called the *Cartesian coordinates* of that point. These two numbers are obtained by intersecting the two axes with the lines through P which are perpendicular to the axes.

An interval of the real line can be visualized as a segment on the real line, possibly with one or both endpoints removed. If I is an interval of the real line and $f : I \to \mathbb{R}$ is a function, then its graph looks typically like a curve in the Cartesian plane. For example, the bottom of Figure 2.2 depicts the graph of a function $f : [-1, 3] \to \mathbb{R}$.

2.5. Exercises

Exercise 2.1. (a) Prove Proposition 2.1.3.

(b) State and prove the multiplicative counterpart of Proposition 2.1.1. □

Exercise 2.2. Prove parts (ii)-(v) of Proposition 2.1.6. □

Exercise 2.3. (a) Prove that

$$(x+y)+(z+t)=\big((x+y)+z\big)+t,\ \ \forall x,y,z,t\in\mathbb{R}.$$

(b) Prove that for any $x,y,z,t,\in\mathbb{R}$ the sum $x+y+z+t$ is independent of the manner in which parentheses are inserted. □

Exercise 2.4. Prove parts (iii)-(v) of Proposition 2.2.2. □

Exercise 2.5. Show that for any real numbers x,y,z such that $y,z\neq 0$, we have

$$\frac{xz}{yz}=\frac{x}{y}.$$ □

Exercise 2.6. (a) Show that for any real numbers x,y,z,t such that $y,t\neq 0$ we have the equality

$$\frac{x}{y}+\frac{z}{t}=\frac{xt+yz}{yt}.$$

(b) Prove that for any real numbers x,y we have

$$x^2-y^2=(x-y)(x+y).$$

(c) Prove that the function $f:(0,\infty)\to\mathbb{R}$, $f(x)=x^2$, is injective but not surjective. □

Exercise 2.7. Prove that if $x\leqslant y$ and $y\leqslant x$, then $x=y$. □

Exercise 2.8. (a) Prove that if $xy>0$, then either $x>0$ and $y>0$, or $x<0$ and $y<0$. □

(b) Prove that if $x>0$, then $1/x>0$.

(c) Let $x>0$. Show that $x>1$ if and only if $1/x<1$.

(d) Prove that if $y>x\geqslant 1$, then

$$x+\frac{1}{x}<y+\frac{1}{y}.$$ □

Exercise 2.9. (a) Prove that $x^2>0$ for any $x\in\mathbb{R}$, $x\neq 0$.

(b) Consider the functions

$$f,g:\mathbb{R}\to\mathbb{R},\ \ f(x)=x^2+1,\ \ g(x)=2x+1.$$

Decide if any of these two functions is injective or surjective.

(c) With f and g as above, describe the functions $f \circ g$ and $g \circ f$. □

Exercise 2.10. Using the technique described in Example 2.2.7 find all the real numbers x such that

$$\frac{x^2}{(x-1)(x+2)} \leqslant 1. \qquad \square$$

Exercise 2.11. (a) Find a positive number M with the following property:

$$\forall x : x > M \Rightarrow \frac{x^2}{x+1} > 10^5.$$

(b) Find a positive number M with the following property:

$$\forall x : \quad x > M \Rightarrow \frac{x^2}{x-1} > 10^6.$$

(c) Find a real number M with the following property:

$$\forall x : \quad x > M \Rightarrow \left|\frac{x^2}{(x-1)(x-2)} - 1\right| < \frac{1}{100}. \qquad \square$$

Exercise 2.12. Let $a < b$. Show that

$$a < \frac{1}{2}(a+b) < b,$$

where 2 is the real number $2 := 1 + 1$. Conclude that the interval (a, b) is nonempty. □

Exercise 2.13. Prove that $x^2 + y^2 \geqslant 2xy$, for any $x, y \in \mathbb{R}$. Use this inequality to prove that

$$x^2 + y^2 + z^2 \geqslant xy + yz + zx, \quad \forall x, y, z \in \mathbb{R}. \qquad \square$$

Exercise 2.14. Prove that if $0 \leqslant x \leqslant \varepsilon$, $\forall \varepsilon > 0$, then $x = 0$. (The Greek letter ε (read *epsilon*) is ubiquitous in analysis and it is almost exclusively used to denote quantities that are extremely small.) □

Exercise 2.15. (a) Consider the function $f : [0, 2] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & x \in [0, 1], \\ 1, & x \in (1, 2]. \end{cases}$$

Decide which of the following statements is true.

(i) $\exists L > 0$ such that $\forall x_1, x_2 \in [0, 2]$ we have $|f(x_1) - f(x_2)| \leqslant L|x_1 - x_2|$.

(ii) $\forall x_1, x_2 \in [0, 2]$, $\exists L > 0$ such that $|f(x_1) - f(x_2)| \leqslant L|x_1 - x_2|$.

(b) Same question, when we change the definition of f to $f(x) = x^2$, for all $x \in [0, 2]$. □

Exercise 2.16. Show that for any $\delta > 0$ and any $a \in \mathbb{R}$ we have
$$(a-\delta, a+\delta) = \big\{ x \in \mathbb{R};\ \ |x-a| < \delta \big\}. \qquad \square$$

Exercise 2.17. Prove the statements (ii)-(iv) of Proposition 2.2.5. □

Exercise 2.18. Prove that for any real numbers a, b, c we have
$$\operatorname{dist}(a,c) \leqslant \operatorname{dist}(a,b) + \operatorname{dist}(b,c). \qquad \square$$

Exercise 2.19. Prove that a set $X \subset \mathbb{R}$ is bounded if and only if there exists $C > 0$ such that $|x| \leqslant C$, $\forall x \in X$. □

Exercise 2.20. Fix two real numbers a, b such that $a < b$. Prove that for any $x, y \in [a,b]$ we have
$$|x-y| \leqslant b-a. \qquad \square$$

Exercise 2.21. State and prove the version of Proposition 2.3.7 involving the infimum of a bounded below set $X \subset \mathbb{R}$. □

Exercise 2.22. Let $X \subset \mathbb{R}$ be a nonempty set of real numbers. For $c \in \mathbb{R}$ define
$$cX := \big\{ cx;\ \ x \in X \big\} \subset \mathbb{R}.$$

(i) Show that if $c > 0$ and X is bounded above, then cX is bounded above and
$$\sup cX = c \sup X.$$

(ii) Show that if $c < 0$ and X is bounded above, then cX is bounded below and
$$\inf cX = c \sup X.$$

□

Exercise 2.23. (a) Let
$$A := \left\{ \frac{a}{a+1};\ \ a > 0 \right\}.$$
Compute $\inf A$ and $\sup A$.

(b) Let
$$B := \left\{ \frac{b}{b+1};\ \ b \in \mathbb{R}\backslash\{-1\} \right\}.$$
Prove that the set B is not bounded below or above. □

Chapter 3

Special classes of real numbers

3.1. The natural numbers and the induction principle

The numbers of the form

$$1, \ \ 1+1, \ \ (1+1)+1$$

and so forth are denoted respectively by $1, 2, 3, \ldots$ and are called *natural numbers.* The term *and so forth* is rather ambiguous and its rigorous justification is provided by the *principle of mathematical induction.*

Definition 3.1.1. A set $X \subset \mathbb{R}$ is called *inductive* if

$$\forall x : \ \ (x \in X \Rightarrow x+1 \in X).$$

□

Example 3.1.2. The set $\mathbb{R}$ is inductive and so is any interval (a, ∞) If $(X_a)_{a \in A}$ is a collection of inductive sets, then so is their intersection

$$\bigcap_{a \in A} X_a.$$ □

Definition 3.1.3. The set of *natural numbers* is the smallest inductive set containing 1, i.e., the intersection of all inductive sets that contain 1. The set of natural numbers is denoted by $\mathbb{N}$. □

To unravel the above definition, the set $\mathbb{N}$ is the subset of $\mathbb{R}$ uniquely characterized by the following requirements.

- The set $\mathbb{N}$ is inductive and $1 \in \mathbb{N}$.
- If $S \subset \mathbb{R}$ is an inductive set that contains 1, then $\mathbb{N} \subset S$.

The set $\mathbb{N}$ consists of the numbers

$$1,\;\; 2 := 1+1,\;\; 3 := 2+1,\;\; 4 := 3+1, \dots .$$

Note that $0 \notin \mathbb{N}$. Indeed, the interval $[1,\infty)$ is an inductive set, containing 1 and thus must contain $\mathbb{N}$. On the other hand, this interval does not contain 0. The above argument proves that $\mathbb{N} \subset [1,\infty)$, i.e.,

$$n \geqslant 1,\;\; \forall n \in \mathbb{N}. \tag{3.1.1}$$

We set

$$\boxed{\mathbb{N}_0 := \{0\} \cup \mathbb{N} = \big\{\, 0, 1, 2, , 3, \dots, \big\}.}$$

☆ **The Principle of Mathematical Induction.** *If E is an inductive subset of the set of natural numbers such that $1 \in E$, then $E = \mathbb{N}$.*

‘

In applications the set E consists of the natural numbers n satisfying a property $P(n)$. To prove that any natural number n satisfies the property $P(n)$ it suffices to prove two things.

- Prove $P(1)$. This is called the *initial step.*
- Prove that if $P(n)$ is true, then so is $P(n+1)$. This is called the *inductive step.*

Sometimes we need an alternate version of the induction principle.

☆ **The Principle of Mathematical Induction: alternate version.** Suppose that for any natural number n we are given a statement $P(n)$ and we know the following.

- The statement $P(1)$ is true.
- For any $n \in \mathbb{N}$, if $P(k)$ is true for any $k < n$, then $P(n)$ is true as well.

Then $P(n)$ is true for any $n \in \mathbb{N}$. □

‘

We will spend the rest of this section presenting various instances of the induction principle at work.

Proposition 3.1.4. *The sum and the product of two natural numbers are also natural numbers.*

Proof. [1] Fix a natural number m. For each $n \in \mathbb{N}$ consider the statement

$$P(n) := \;\; m+n \text{ is a natural number.}$$

[1] The proof can be omitted.

We have to prove that $P(n)$ is true for any $n \in \mathbb{N}$. We will achieve this using the principle of induction. We first need to check that $P(1)$ is true, i.e., that $m+1$ is a natural number. This follows from the fact that $m \in \mathbb{N}$ and $\mathbb{N}$ is an inductive set.

To complete the inductive step assume that $P(n)$ is true, i.e., $m+n \in \mathbb{N}$. Thus $(m+n)+1 \in \mathbb{N}$ and

$$m+(n+1) = (m+n)+1 \in \mathbb{N}.$$

This shows that $P(n+1)$ is also true. □

Lemma 3.1.5. *$\forall n \in \mathbb{N}$, $(n \neq 1) \Rightarrow (n-1) \in \mathbb{N}$.*

Proof. For $n \in \mathbb{N}$ consider the statement

$$P(n) := n \neq 1 \Rightarrow (n-1) \in \mathbb{N}.$$

We want to prove that this statement is true for any $n \in \mathbb{N}$. The initial step is obvious since for $n = 1$ the statement $n \neq 1$ is false and thus the implication is true.

For the inductive step assume that the statement $P(n)$ is true and we prove that $P(n+1)$ is also true. Observe that $n+1 \neq 1$ because $n \in \mathbb{N}$ and thus $n \neq 0$. Clearly $(n+1)-1 = n \in \mathbb{N}$. □

Lemma 3.1.6. *The set*

$$I_1 = \{ x \in \mathbb{N};\;\; x > 1 \}$$

admits a minimal element and $\min I_1 = 2$.

Proof. Consider the set

$$E := \{ x \in \mathbb{N};\;\; x = 1 \vee x \geqslant 2 \} \subset \mathbb{N}.$$

We will prove by induction that

$$E = \mathbb{N}. \tag{3.1.2}$$

Thus we need to show that $1 \in E$ and $x \in E \Rightarrow x+1 \in E$. Clearly $1 \in E$.

If $x \in E$, then

- either $x = 1$ so that $x+1 = 2 \geqslant 2$ so that $x+1 \in E$,
- or $x \geqslant 2$ which implies $x+1 \geqslant 2$ and thus $x+1 \in E$.

The equality $E = \mathbb{N}$ implies that a natural number n is either equal to 1, or it is $\geqslant 2$. Thus

$$x \in \mathbb{N} \wedge x > 1 \Rightarrow x \geqslant 2.$$

This shows that

$$x \geqslant 2,\;\; \forall x \in I_1.$$

Clearly $2 \in I_1$ so that $2 = \min I$. □

Corollary 3.1.7. *For any $n \geqslant 1$ the set*

$$H_n = \{ x \in \mathbb{N};\;\; x > n \}$$

admits a minimal element and

$$\min H_n = n+1.$$

Proof. We will prove that for any $n \in \mathbb{N}$ the statement

$$P(n): \quad \min H_n = n+1$$

is true. Lemma 3.1.6 shows that $P(1)$ is true.

Let us show that $P(n) \Rightarrow P(n+1)$. Since $n+2 \in H_{n+1}$ it suffices to show that $x \geqslant n+2$, $\forall x \in H_{n+1}$. Let $x \in H_{n+1}$. Lemma 3.1.5 implies that $x-1 \in \mathbb{N}$ and $x-1 > n$ so that $x-1 \in H_n$. Since $P(n)$ is true, we deduce $x-1 \geqslant n+1$, i.e., $x \geqslant n+2$.

□

Corollary 3.1.8. *Suppose that n is a natural number. Any natural number x such that $x > n$ satisfies $x \geqslant n+1$.* □

Corollary 3.1.9. *For any natural number n, the open interval $(n, n+1)$ contains no natural number.*

Proof. From Corollary 3.1.8 we deduce that if x is a natural number such that $x > n$, then $x \geqslant n+1$. Thus there cannot exist any natural number x such that $n < x < n+1$.□

The above results imply the following important theorem.

Theorem 3.1.10 (Well Ordering Principle)**.** *Any set of natural numbers $S \subset \mathbb{N}$ has a minimal element.* □

For a proof of this theorem we refer to [**44**, §2.2.1].

Definition 3.1.11. For any $n \in \mathbb{N}$ we denote by $\mathbb{I}_n$ the set

$$\mathbb{I}_n := \big\{ x \in \mathbb{N};\ \ 1 \leqslant x \leqslant n \big\} = [1, n] \cap \mathbb{N}.$$ □

Definition 3.1.12. We say two sets X and Y are said to have the *same cardinality*, and we write this $X \sim Y$, if and only if there exists a bijection $f: X \to Y$. A set X is called *finite* if there exists a natural number n such that $X \sim \mathbb{I}_n$. □

Let us observe that if X, Y, Z are three sets such that $X \sim Y$ and $Y \sim Z$, then $X \sim Z$; see Exercise 3.1. This implies that any set X equivalent to a finite set Y is also finite. Indeed, if $X \sim Y$ and $Y \sim \mathbb{I}_n$, then $X \sim \mathbb{I}_n$.

At this point we want to invoke (without proof) the following result.

Proposition 3.1.13. *For any $m, n \in \mathbb{N}$ we have*

$$\mathbb{I}_n \sim \mathbb{I}_m \Longleftrightarrow m = n.$$ □

The above result implies that if X is a finite set, then there exists a ***unique*** natural number n such that $X \sim \mathbb{I}_n$. This unique natural number is called the *cardinality* of X and it is denoted by $|X|$ or $\#X$. You should think of the cardinality of a finite set as the number of elements in that set.

An *infinite* set is a set that is not finite. We have the following *highly nontrivial* result. Its proof is too complex to present here.

Theorem 3.1.14. *A set X is infinite if and only if it is equivalent to one of its* proper[2] *subsets.* □

Theorem 3.1.15. *The set of natural numbers $\mathbb{N}$ is infinite.*

Proof. Consider the proper subset

$$H := \big\{ n \in \mathbb{N};\ \ n > 1 \big\} \subset \mathbb{N}.$$

Lemma 3.1.5 implies that if $n \in H$, then $(n-1) \in \mathbb{N}$. Consider the map

$$f : H \to \mathbb{N},\ \ f(n) = n - 1.$$

Observe that this map is injective. Indeed, if $f(n_1) = f(n_2)$, then $n_1 - 1 = n_2 - 1$ so that $n_1 = n_2$. This map is also surjective. Indeed, if $m \in \mathbb{N}$. Then, according to (3.1.1) the natural number $n := m + 1$ is greater than 1 so it belongs to H. Clearly $f(n) = (m+1) - 1 = m$ which proves that f is also surjective. □

Definition 3.1.16. A set X is called *countable* if it is equivalent with the set of natural numbers. □

Example 3.1.17. The set $\mathbb{N} \times \mathbb{N}$ is countable. To see this arrange the elements of $\mathbb{N} \times \mathbb{N}$ in a sequence as follows:

$$(1,1),\ \underbrace{(2,1),(2,2)}_{S_2},\ \underbrace{(3,1),(3,2),(3,3)}_{S_3},\ \underbrace{(4,1),(4,2),(4,3),(4,4)}_{S_4},$$

Now denote by $\phi(m,n)$ the location of the pair (m,n) in the above string. For example, $\phi(1,1) = 1$ since $(1,1)$ is the first term in the above sequence. Note that

$$\phi(4,2) = 1 + 2 + 3 + 2 = 8,$$

i.e., $(4,2)$ occupies the 8-th position in the above string. More precisely ϕ is the function

$$\phi : \mathbb{N} \times \mathbb{N} \to \mathbb{N},\ \ \phi(m,n) = \#S_1 + \cdots \#S_{m-1} + n.$$

It should be clear that ϕ is bijective proving that $\mathbb{N} \times \mathbb{N}$ has the same cardinality as $\mathbb{N}$. □

[2] We recall that a subset $S \subset X$ is called proper if $S \neq X$.

3.2. Applications of the induction principle

In this section we discuss some traditional applications of the induction principle. This serves two purposes: first, it familiarizes you with the usage of this principle, and second, some of the results we will discuss here will be needed later on in this class.

First let us introduce some notations. If n is a natural number, $n > 1$, and we are given n real numbers $a_1, \dots, a_n$, then define inductively

$$a_1 + \cdots + a_n := (a_1 + \cdots + a_{n-1}) + a_n,$$

$$a_1 \cdots a_n = (a_1 \cdots a_{n-1})a_n.$$

We will use the following notations for the sum and products of a string of real numbers. Thus

$$\sum_{k=1}^{n} a_k := a_1 + \cdots + a_n, \quad \prod_{k=1}^{n} a_k := a_1 \cdots a_n.$$

Similarly, given real numbers $a_0, a_1, \dots, a_n$ we define

$$\sum_{k=0}^{n} a_k = a_0 + a_1 + \cdots + a_n, \quad \prod_{k=0}^{n} a_k := a_0 \cdots a_n.$$

For any natural number n and any real number x we define inductively

$$x^n := \begin{cases} x & \text{if } n = 1 \\ (x^{n-1}) \cdot x & \text{if } n > 1. \end{cases}$$

Intuitively

$$x^n = \underbrace{x \cdot x \cdots x}_{n}.$$

If x is a *nonzero* real number we set

$$x^0 := 1.$$

Let us observe that for any natural numbers m, n and any real number x we have the equality

$$x^{m+n} = (x^m) \cdot (x^n).$$

Exercise 3.2 asks you to prove this fact.

Example 3.2.1. Let us prove that

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \forall n \in \mathbb{N}. \tag{3.2.1}$$

The expanded form of the last equality is

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}, \quad \forall n \in \mathbb{N}.$$

Let us denote by S_n the sum $1+2+\cdots+n$. We argue by induction. The initial case $n=1$ is trivial since

$$\frac{1\cdot(1+1)}{2}=1=S_1.$$

For the inductive case we assume that

$$S_n=\frac{n(n+1)}{2},$$

and we have to prove that

$$S_{n+1}=\frac{(n+1)(\,(n+1)+1)}{2}=\frac{(n+1)(n+2)}{2}.$$

Indeed we have

$$S_{n+1}=S_n+(n+1)=\frac{n(n+1)}{2}+n+1=\frac{n(n+1)}{2}+\frac{2(n+1)}{2}=\frac{n(n+1)+2(n+1)}{2}$$

(factor out $(n+1)$)

$$=\frac{(n+1)(n+2)}{2}.$$

□

Example 3.2.2 (Bernoulli's inequality)**.** We want to prove a simple but very versatile inequality that goes by the name of *Bernoulli's inequality.* It states that

$$\forall x\geqslant -1,\;\; \forall n\in\mathbb{N}:\;\; (1+x)^n\geqslant 1+nx. \tag{3.2.2}$$

We argue by induction. Clearly, the inequality is obviously true when $n=1$ and the initial case is true. For the inductive case, we assume that

$$(1+x)^n\geqslant 1+nx,\;\; \forall x\geqslant -1 \tag{3.2.3}$$

and we have to prove that

$$(1+x)^{n+1}\geqslant 1+(n+1)x,\;\; \forall x\geqslant -1.$$

Since $x\geqslant -1$ we deduce $1+x\geqslant 0$. Multiplying both sides of (3.2.3) with the nonnegative number $1+x$ we deduce

$$(1+x)^{n+1}\geqslant (1+x)(1+nx)=1+nx+x+nx^2\geqslant 1+nx+x=1+(n+1)x.$$

□

Example 3.2.3 (***Newton's Binomial Formula***)**.** Before we state this very important formula we need to introduce several notations widely used in mathematics. For $n\in\mathbb{N}\cup\{0\}$ we define $n!$ (read n factorial) as follows

$$0!:=1,\;\; 1!:=1,\;\; 2!=1\cdot 2,\;\; 3!=1\cdot 2\cdot 3,\cdots n!=1\cdot 2\cdots n.$$

Given $k,n\in\mathbb{N}\cup\{0\}$, $k\leqslant n$ we define the *binomial coefficient* $\binom{n}{k}$ (read n *choose* k)

$$\binom{n}{k}:=\frac{n!}{k!\cdot(n-k)!}.$$

We record below the values of these binomial coefficients for small values of n

$$\binom{0}{0}=1,\;\; \binom{1}{0}=\binom{1}{1}=1,$$

$$\binom{2}{0}=\frac{2!}{(0!)(2!)}=1,\ \ \binom{2}{1}=\frac{2!}{(1!)(1!)}=2,\ \ \binom{2}{2}=\frac{2!}{(2!)(0!)}=1,$$
$$\binom{3}{0}=\frac{3!}{(0!)(3!)}=\binom{3}{3}=1,\ \ \binom{3}{1}=\frac{(3!)}{(1!)(2!)}=\frac{3!}{(2!)(1!)}=\binom{3}{2}=3.$$

Here is a more involved example

$$\binom{7}{3}=\frac{7!}{(3!)(4!)}=\frac{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{(3!)1\cdot 2\cdot 3\cdot 4}=\frac{7\cdot 6\cdot 5}{3!}=\frac{7\cdot 6\cdot 5}{6}=35.$$

The binomial coefficients can be conveniently arranged in the so called *Pascal triangle*

$$\begin{array}{cccccccccc}
\binom{0}{k}: & & & & & 1 & & & & \\
\binom{1}{k}: & & & & 1 & & 1 & & & \\
\binom{2}{k}: & & & 1 & & 2 & & 1 & & \\
\binom{3}{k}: & & 1 & & 3 & & 3 & & 1 & \\
\binom{4}{k}: & 1 & & 4 & & 6 & & 4 & & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

Observe that each entry in the Pascal triangle is the sum of the closest neighbors above it.

The binomial coefficients play an important role in mathematics. One reason behind their usefulness is *Newton's binomial formula* which states that, for any natural number n, and any real numbers x, y, we have the equality below

$$\begin{aligned}(x+y)^n&=\binom{n}{0}x^n+\binom{n}{1}x^{n-1}y+\binom{n}{2}x^{n-2}y^2+\cdots+\binom{n}{n-1}xy^{n-1}+\binom{n}{n}y^n\\ &=\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k.\end{aligned}\tag{3.2.4}$$

We will prove this equality by induction on n. For $n=1$ we have

$$(x+y)^1=x+y=\binom{1}{0}x+\binom{1}{1}y,$$

which shows that the case $n=1$ of (3.2.4) is true.

As for the inductive steps, we assume that (3.2.4) is true for n and we prove that it is true for $n+1$. We have

$$(x+y)^{n+1}=(x+y)(x+y)^n$$

(use the inductive assumption)

$$=(x+y)\left(\binom{n}{0}x^n+\binom{n}{1}x^{n-1}y+\binom{n}{2}x^{n-2}y^2+\cdots+\binom{n}{n-1}xy^{n-1}+\binom{n}{n}y^n\right)$$

$$\begin{aligned}
&= x\left(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n\right)\\
&\quad +y\left(\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n\right)
\end{aligned}$$

$$\begin{aligned}
&= \binom{n}{0}x^{n+1} + \binom{n}{1}\boxed{x^ny} + \binom{n}{2}\boxed{x^{n-1}y^2} + \cdots + \binom{n}{n-1}x^2y^{n-1} + \binom{n}{n}\boxed{xy^n}\\
&\qquad + \binom{n}{0}\boxed{x^ny} + \binom{n}{1}\boxed{x^{n-1}y^2} + \binom{n}{2}x^{n-2}y^3 + \cdots + \binom{n}{n-1}\boxed{xy^n} + \binom{n}{n}y^{n+1}
\end{aligned}$$

$$\begin{aligned}
&= \binom{n}{0}x^{n+1} + \left(\binom{n}{1} + \binom{n}{0}\right)\boxed{x^ny} + \left(\binom{n}{2} + \binom{n}{1}\right)\boxed{x^{n-1}y^2} + \cdots\\
&\quad + \left(\binom{n}{k} + \binom{n}{k-1}\right)x^{n+1-k}y^k + \cdots + \left(\binom{n}{n} + \binom{n}{n-1}\right)\boxed{xy^n} + \binom{n}{n}y^{n+1}.
\end{aligned}$$

Clearly

$$\binom{n}{0} = 1 = \binom{n+1}{0}, \quad \binom{n}{n} = 1 = \binom{n+1}{n+1}.$$

We want to show $1 \leqslant k \leqslant n$ we have the *Pascal's formula*

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}. \tag{3.2.5}$$

Indeed, we have

$$\begin{aligned}
\binom{n}{k} + \binom{n}{k-1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!}\\
&= \frac{n!}{k(k-1)!(n-k)!} + \frac{n!}{(k-1)!(n-k)!(n-k+1)}\\
&= \frac{n!}{(k-1)!(n-k)!}\left(\frac{1}{k} + \frac{1}{n-k+1}\right)\\
&= \frac{n!}{(k-1)!(n-k)!}\cdot\left(\frac{(n-k+1)}{k(n-k+1)} + \frac{k}{k(n-k+1)}\right)\\
&= \frac{n!}{(k-1)!(n-k)!}\cdot\frac{n+1}{k(n-k+1)}\\
&= \frac{(n+1)n!}{(\,k(k-1)!\,)\cdot(\,(n-k+1)(n-k)!\,)} = \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}.
\end{aligned}$$

This completes the inductive step. □

3.3. Archimedes' Principle

We begin with a simple but fundamental observation.

Proposition 3.3.1. *Suppose that the nonempty subset $E \subset \mathbb{N}$ is bounded above. Then E has a maximal element n_0, i.e., $n_0 \in E$ and $n \leqslant n_0$, $\forall n \in E$.*

Proof. From the completeness axiom we deduce that E has a least upper bound

$$M := \sup E \in \mathbb{R}.$$

We want to prove that $M \in E$. We argue by contradiction. Suppose that $M \notin E$. In particular, this means that any number in E is strictly smaller than M.

From the definition of the *least* upper bound we deduce that there must exist $n_0 \in E$ such that

$$M - 1 < n_0 \leqslant M.$$

On the other hand, any natural number n greater than n_0 must be greater than or equal to $n_0 + 1$, $n \geqslant n_0 + 1$. Observing that $n_0 + 1 > M$, we deduce that any natural number $> n_0$ is also $> M$. Since $M \notin E$, then $n_0 < M$, and the above discussion show that the interval (n_0, M) contains no natural numbers, thus no elements of E. Hence, any real number in (n_0, M) will be an upper bound for E, contradicting that M is the *least* upper bound. □

Theorem 3.3.2 (Archimedes' Principle)**.** *Let ε be a positive real number. Then for any $x > 0$ there exists $n \in \mathbb{N}$ such that $n\varepsilon > x$.* [3]

Proof. Consider the set

$$E := \{\, n \in \mathbb{N};\ \ n\varepsilon \leqslant x \,\}.$$

If $E = \varnothing$, then this means that $n\varepsilon > x$ for any $n \in \mathbb{N}$ and the conclusion of the theorem is guaranteed. Suppose that $E \neq \varnothing$. Observe that

$$n \leqslant \frac{x}{\varepsilon},\ \ \forall n \in E.$$

Hence, the set E is bounded above, and the previous proposition shows that it has a maximal element n_0. Then $n_0 + 1 \notin E$, so that $(n_0 + 1)\varepsilon > x$. □

Definition 3.3.3. The set of *integers* is the subset $\mathbb{Z} \subset \mathbb{R}$ consisting of the natural numbers, the negatives of natural numbers and 0. □

The proof of the following results are left to you as an exercise.

Proposition 3.3.4. *If $m, n \in \mathbb{Z}$, then $m + n, mn \in \mathbb{Z}$.* □

[3] A popular formulation of Archimedes' principle reads: one can fill an ocean with grains of sand.

Proposition 3.3.5. *For any real number x the interval $(x-1, x]$ contains exactly one integer.* □

Corollary 3.3.6. *For any real number x there exists a unique integer n such that*

$$n \leqslant x < n+1.$$

*This integer is called the **integer part** of x and it is denoted by $\lfloor x \rfloor$.*

Proof. Observe that the inequalities $n \leqslant x < n+1$ are equivalent to the inequalities

$$x-1 < n \leqslant x.$$

By Proposition 3.3.5, the interval $(x-1, x]$ contains exactly one integer. This proves the existence and uniqueness of the integer with the postulated properties. □

Observe for example that

$$\left\lfloor \frac{1}{2} \right\rfloor = 0, \quad \left\lfloor -\frac{1}{2} \right\rfloor = -1,$$

Theorem 3.3.7 (Division with remainder)**.** *Let $m, n \in \mathbb{Z}$, $n > 0$. There exists a unique pair of integers $(q, r) \in \mathbb{Z} \times \mathbb{Z}$ satisfying the following properties.*

(i) $m = qn + r$.

(ii) $0 \leqslant r < n$.

The number r is called the remainder of the division of m by n.

Proof. *Uniqueness.* Suppose that there exist two pairs of integers (q_1, r_1) and (q_2, r_2) satisfying (i) and (ii). Then

$$nq_1 + r_1 = m = nq_2 + r_2,$$

so that,

$$nq_1 - nq_2 = r_2 - r_1 \Rightarrow n(q_1 - q_2) = r_2 - r_1 \Rightarrow n \cdot |q_1 - q_2| = |r_2 - r_1|.$$

The natural numbers r_1, r_2 satisfy $0 \leqslant r_1, r_2 < |n|$ so that $r_1, r_2 \in [0, n-1]$. Using Exercise 2.20 we deduce $|r_2 - r_1| \leqslant n-1$. Hence $n \cdot |q_1 - q_2| \leqslant n-1$ which implies

$$|q_1 - q_2| \leqslant \frac{n-1}{n} < 1.$$

The quantity $|q_1 - q_2|$ is a nonnegative integer < 1 so that it must equal 0. This implies $q_1 = 2$ and

$$r_2 - r_1 = n(q_1 - q_2) = 0.$$

This proves the uniqueness.

Existence. Let

$$q := \left\lfloor \frac{m}{n} \right\rfloor \in \mathbb{Z}.$$

Then

$$q \leqslant \frac{m}{n} < q+1 \Rightarrow nq \leqslant m < n(q+1) = nq + n \Rightarrow 0 \leqslant m - nq < n.$$

We set $r := m - nq$ and we observe that the pair (q, r) satisfies all the required properties. □

Definition 3.3.8. (a) Let $m, n \in \mathbb{Z}$, $m \neq 0$. We say that m *divides* n, and we write this $m|n$ if there exists an integer k such that $n = km$. When m divides n we also say that m is a *divisor* of n, or that n *is a multiple of* m, or that n *is divisible by* m.

(b) A *prime number* is a natural number $p > 1$ whose only divisors are ± 1 and $\pm p$. □

Observe that if d is a divisor of m, then $-d$ is also a divisor of m. An *even integer* is an integer divisible by 2. An *odd integer* is an integer not divisible by 2.

Given two integers m, n consider the set of common positive divisors of m and n, i.e., the set

$$D_{m,n} := \{\, d \in \mathbb{N};\;\; d|m \wedge d|n \,\}.$$

This set is not empty because 1 is a common positive divisor. This is bounded above because any divisor of m is $\leqslant |m|$. Thus the set $D_{m,n}$ has a maximal element called the *greatest common divisor* of m and n and denoted by $\gcd(m, n)$. Two integers are called *coprime* if $\gcd(m, n) = 1$, i.e., 1 is their only positive common divisor.

The next result describes on the most important property of the set $\mathbb{Z}$ of integers. We will not include its rather elaborate and tricky proof. The curious reader can find the proof in any of the books [8, 11, 34].

Theorem 3.3.9 (Fundamental Theorem of Arithmetic)**.** *(a) If p is a prime number that divides a product of integers mn, then $p|m$ or $p|n$.*

(b) Any natural number n can be written in a unique fashion as a product

$$n = p_1^{\alpha_1} \cdots p_k^{\alpha_k},$$

where $p_1 < p_2 < \cdots < p_k$ are prime numbers and $\alpha_1, \dots, \alpha_k$ are natural numbers. □

3.4. Rational and irrational numbers

We want to isolate another important subclasses of real numbers.

Definition 3.4.1. The set of *rationals* (or *rational numbers*) is the subset $\mathbb{Q} \subset \mathbb{R}$ consisting of real numbers of the form m/n where $m, n \in \mathbb{Z}$, $n \neq 0$. □

If q is a rational number, then it can be written as a fraction of the form $q = \frac{m}{n}$, $n \neq 0$. We denote by d the $\gcd(m, n)$. Thus there exist integers m_1 and n_1 such that

$$m = dm_1, \;\; n = dn_1.$$

Clearly the numbers m_1, n_1 are coprime, and we have

$$q = \frac{dm_1}{dn_1} = \frac{m_1}{n_1}.$$

We have thus proved the following result.

Proposition 3.4.2. *Every rational number is the ratio of two coprime integers.* □

The proof of the following result is left to you as an exercise.

Proposition 3.4.3. *If $q, r \in \mathbb{Q}$, then $q + r, qr \in \mathbb{Q}$.* □

We have a sequence of inclusions

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

Clearly $\mathbb{N} \neq \mathbb{Z}$ because $-1 \in \mathbb{Z}$, but $-1 \notin \mathbb{N}$. Note however that, although $\mathbb{Z}$ contains $\mathbb{N}$, the set of integers $\mathbb{Z}$ is countable, i.e., it has the same cardinality as $\mathbb{N}$.

Next observe that $\mathbb{Z} \neq \mathbb{Q}$. Indeed, the rational number $1/2$ is not an integer, because it is positive and smaller than any natural number.

Similarly, although $\mathbb{Q}$ strictly contains $\mathbb{Z}$, these two sets have the same cardinality: they are both countable. However, the following **very important result** shows that, loosely speaking, there are "many more rational numbers".

Proposition 3.4.4 (Density of rationals)**.** *Any open interval $(a, b) \subset \mathbb{R}$, no matter how small, contains at least one rational number.*

Proof. From Archimedes' principle we deduce that there exists at least one natural number n such that $n > \frac{1}{b-a}$. Observe that $(b - a)$ is the length of the interval (a, b). This inequality is obviously equivalent to the inequality

$$\frac{1}{n} < b - a \Longleftrightarrow n(b - a) > 1$$

(This last equality codifies a rather intuitive fact: one can divide a stick of length one into many equal parts so that the subparts are as small as we please.)

We will show that we can find an integer m such that $\frac{m}{n} \in (a, b)$. Observe that

$$a < \frac{m}{n} < b \Longleftrightarrow na < m < nb \Longleftrightarrow m \in (na, nb).$$

This shows that the interval (a, b) contains a rational number if the interval (na, nb) contains an integer.

Since $n(b - a) > 1$, we deduce $nb > na + 1$. In particular, this shows that the interval $(na, na + 1]$ is contained in the interval (na, nb). From Proposition 3.3.5 we deduce that the interval $(na, na + 1]$ contains an integer m. □

This abundance of rational numbers lead people to believe for quite a long while that all real numbers must be rational. Then the ancient Greeks showed that there must exist real numbers that cannot be rational. These numbers were called *irrational*. In the remainder of this section we will describe how one can produce a large supply of irrational numbers. We start with a baby case.

Proposition 3.4.5. *There exists a unique positive number r such that $r^2 = 2$. This number is called the square root of 2 and it is denoted by $\sqrt{2}$*

Proof. We begin by observing the following *useful fact*:

$$\forall x, y > 0: \quad x < y \Longleftrightarrow x^2 < y^2. \tag{3.4.1}$$

Indeed

$$y^2 - x^2 > 0 \Longleftrightarrow (y-x)(y+x) > 0 \Longleftrightarrow y > x.$$

This useful fact takes care of the uniqueness because, if r_1, r_2 are two *positive* real numbers such that $r_1^2 = r_2^2 = 2$, then $r_1 = r_2 \Longleftrightarrow r_1^2 = r_2^2$.

To establish the existence of a positive r such that $r^2 = 2$ consider as in Example 2.3.2(b) the set

$$R = \bigl\{ x > 0; \;\; x^2 < 2 \bigr\}.$$

We have seen that this set is bounded above and thus it admits a least upper bound

$$r := \sup R.$$

We want to prove that $r^2 = 2$. We argue by contradiction and we assume that $r^2 \neq 2$. Thus, either $r^2 < 2$ or $r^2 > 2$.

Case 1. $r^2 < 2$. We will show that there exists ε_0 such that $(r + \varepsilon_0)^2 < 2$. This would imply that $r + \varepsilon_0 \in R$ and would contradict the fact that r is an upper bound for R because r would be smaller than the element $r + \varepsilon_0$ of R.

Set $\delta := 2 - r^2$. For any $\varepsilon \in (0,1)$ we have

$$(r+\varepsilon)^2 - r^2 = \bigl((r+\varepsilon) - r \bigr)\bigl((r+\varepsilon) + r \bigr) = \varepsilon(2r + \varepsilon) < \varepsilon(2r+1).$$

Now choose a number $\varepsilon_0 \in (0,1)$ such that

$$\varepsilon_0 < \frac{\delta}{2r+1}.$$

Then

$$(r+\varepsilon_0)^2 - r^2 < \varepsilon_0(2r+1) < \delta$$
$$\Rightarrow (r+\varepsilon_0)^2 < r^2 + \delta = r^2 + 2 - r^2 = 2 \Rightarrow r + \varepsilon_0 \in R.$$

Case 2. $r^2 > 2$. We will prove that under this assumption

$$\exists \varepsilon_0 \in (0,1) \;\text{ such that }\; r - \varepsilon_0 > 0 \;\text{ and }\; (r - \varepsilon_0)^2 > 2. \tag{3.4.2}$$

Let us observe that (3.4.2) leads to a contradiction. Indeed, observe that $(r - \varepsilon_0)$ is an upper bound for R. Indeed, if $x \in R$, then

$$x^2 < 2 < (r - \varepsilon_0)^2 \overset{(3.4.1)}{\Rightarrow} x < r_0 - \varepsilon.$$

Thus, $r - \varepsilon_0$ is an upper bound of R and this upper bound is obviously strictly smaller than r, the *least* upper bound of R. This is a contradiction which shows that the situation $r^2 > 2$ is also not possible. Let us now prove (3.4.2).

Denote by δ the difference $\delta = r^2 - 2 > 0$. For any $\varepsilon \in (0, r)$ we have

$$r^2 - (r-\varepsilon)^2 = \Big(r - (r-\varepsilon)\Big)\Big(r + (r-\varepsilon)\Big) = \varepsilon(2r-\varepsilon) < 2r\varepsilon.$$

We have thus shown that for any $\varepsilon \in (0, r)$ we have $(r-\varepsilon) > 0$ and

$$r^2 - (r-\varepsilon)^2 \leqslant 2r\varepsilon \Longleftrightarrow (r-\varepsilon)^2 \geqslant r^2 - 2r\varepsilon.$$

Now choose $\varepsilon_0 \in (0, r)$ small enough so that $\varepsilon_0 < \frac{\delta}{2r}$. Hence $2r\varepsilon_0 < \delta$ so that $-2r\varepsilon_0 > -\delta$ and

$$(r-\varepsilon_0)^2 > r^2 - 2r\varepsilon_0 > r^2 - \delta = r^2 - (r^2 - 2) = 2.$$

We deduce again that the situation $r^2 > 2$ is not possible so that $r^2 = 2$.

□

The result we have just proved can be considerably generalized.

Theorem 3.4.6. *Fix a natural number $n \geqslant 2$. Then for any positive real number a* there exists a unique, positive *real number r such that $r^n = a$.*

Proof. *Existence.* Consider the set

$$S := \big\{ s \in \mathbb{R};\ \ s \geqslant 0 \wedge s^n \leqslant a \big\}.$$

Observe that this is a nonempty set since $0 \in S$. We want to prove that S is also bounded. To achieve this we need a few auxiliary results.

Lemma 3.4.7 (**A very handy identity**). *For any real numbers x, y and any natural number n we have the equality*

$$x^n - y^n = \big(\,x - y\,\big) \cdot \big(\,x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}\,\big) \tag{3.4.3}$$

Proof. We have

$$\begin{aligned} &\big(\,x-y\,\big)\cdot\big(\,x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}\,\big)\\ &= x\big(\,x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}\,\big) - y\big(\,x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}\,\big)\\ &= x^n + x^{n-1}y + x^{n-2}y^2 + \cdots + x^2y^{n-2} + xy^{n-1}\\ &\quad - x^{n-1}y - x^{n-2}y^2 - \cdots - x^2y^{n-2} - xy^{n-1} - y^n\\ &= x^n - y^n. \end{aligned}$$

□

Here is an immediate useful consequence of this identity.

$$\forall n \in \mathbb{N},\ \ \forall x, y > 0:\ \ x < y \Longleftrightarrow x^n < y^n. \tag{3.4.4}$$

Indeed

$$y^n - x^n > 0 \Longleftrightarrow (y-x)\big(\,y^{n-1} + y^{n-2}x + \cdots + x^{n-1}\,\big) > 0 \Longleftrightarrow y - x > 0.$$

Lemma 3.4.8. *Any positive real number x such that $x^n \geqslant a$ is an upper bound for S. In particular, any natural number $k > a$ is an upper bound for S so that S is a bounded set.*

Proof. Let x be a positive real number such that $x^n \geqslant a$. We want to prove that $x \geqslant s$ for any $s \in S$. Indeed

$$s \in S \Rightarrow s^n \leqslant a \leqslant x^n \overset{(3.4.4)}{\Rightarrow} s \leqslant x.$$

This proves the first part of the lemma.

Suppose now that k is a natural number such that $k > a$. Observe first that

$$k^n > k^{n-1} > \cdots > k > a.$$

From the first part of the lemma we deduce that k is an upper bound for S. □

The nonempty set S is bounded above. The Completeness Axiom implies that it admits a least upper bound

$$r := \sup S.$$

We will show that $r^n = a$. We argue by contradiction and we assume that $r^n \neq a$. Thus, either $r^n < a$, or $r^n > a$.

Case 1. $r^n < a$. We will show that we can find $\varepsilon_0 \in (0,1)$ such that $(r+\varepsilon_0)^n < a$. This would imply that $r+\varepsilon_0 \in S$ and it would contradict the fact that r is an upper bound for S because r is less than the number $r+\varepsilon_0 \in S$.

Denote by δ the difference $\delta := a - r^n > 0$. For any $\varepsilon \in (0,1)$ we have

$$(r+\varepsilon)^n - r^n = \Big((r+\varepsilon) - r \Big)\Big((r+\varepsilon)^{n-1} + (r+\varepsilon)^{n-2}r^2 + \cdots + r^{n-1} \Big)$$

$(r+\varepsilon < r+1)$

$$\leqslant \varepsilon \underbrace{\Big((r+1)^{n-1} + (r+1)^{n-2}r + \cdots + r^{n-1} \Big)}_{=:q}$$

We have thus proved that

$$(r+\varepsilon)^n \leqslant r^n + \varepsilon q, \;\; \forall \varepsilon \in (0,1).$$

Choose $\varepsilon_0 \in (0,1)$ small enough so that

$$\varepsilon_0 < \frac{\delta}{q} \Longleftrightarrow \varepsilon_0 q < \delta.$$

Then

$$(r+\varepsilon_0)^n \leqslant r^n + \varepsilon_0 q < r^n + \delta = a \Rightarrow r + \varepsilon_0 \in S.$$

This contradicts the fact that r is an upper bound for S and shows that the inequality $r^n < a$ is impossible.

Case 2. $r^n > a$. We will prove that under this assumption

$$\exists \varepsilon_0 \in (0,1) \;\text{ such that }\; r - \varepsilon_0 > 0 \;\text{ and }\; (r-\varepsilon_0)^n > a. \tag{3.4.5}$$

Let us observe that (3.4.5) leads to a contradiction. Indeed, Lemma 3.4.8 implies that $r - \varepsilon_0$ is an upper bound of S and this upper bound is obviously strictly smaller than r, the *least* upper bound of S. This is a contradiction which shows that the situation $b^n > a$ is also not possible. Let us now prove (3.4.5).

Denote by δ the difference $\delta = r^n - a > 0$. For any $\varepsilon \in (0, r)$ we have

$$r^n - (r-\varepsilon)^n = \Big(r - (r-\varepsilon)\Big)\Big(r^{n-1} + r^{n-2}(r-\varepsilon) + \cdots + \cdots + (r-\varepsilon)^{n-1}\Big)$$

$((r-\varepsilon) < b)$

$$\leqslant \varepsilon\, \underbrace{\left(r^{n-1} + r^{n-2}r + \cdots + r^{n-1}\right)}_{=:q}.$$

We have thus shown that for any $\varepsilon \in (0, r)$ we have $(r-\varepsilon) > 0$ and

$$r^n - (r-\varepsilon)^n \leqslant \varepsilon q \Longleftrightarrow (r-\varepsilon)^n \geqslant r^n - \varepsilon q.$$

Now choose $\varepsilon_0 \in (0, c)$ small enough so that $\varepsilon_0 < \frac{\delta}{q}$. Hence $\varepsilon_0 q < \delta$ so that $-\varepsilon_0 q > -\delta$ and

$$(r-\varepsilon_0)^n > r^n - \varepsilon_0 q > r^n - \delta = r^n - (r^n - a) = a.$$

We deduce again that the situation $r^n > a$ is not possible so that $r^n = a$. This completes the existence part of the proof.

Uniqueness. Suppose that r_1, r_2 are two positive numbers such that $r_1^n = r_2^n = a$. Using (3.4.4) we deduce that $r_1 = r_2$. This completes the proof of Theorem 3.4.6. □

The above result leads to the following important concept.

Definition 3.4.9. Let a be a positive real number and $n \in \mathbb{N}$. The *n-th root* of a, denoted by $a^{\frac{1}{n}}$ or $\sqrt[n]{a}$ is the ***unique positive real number*** r such that $r^n = a$. □

Theorem 3.4.10. *The positive number $\sqrt{2}$ is not rational.*

Proof. We argue by contradiction and we assume that $\sqrt{2}$ is rational. It can therefore be represented as a fraction,

$$\sqrt{2} = \frac{m}{n}, \quad m, n \in \mathbb{N}, \quad \gcd(m, n) = 1.$$

Thus $2 = \frac{m^2}{n^2}$ and we deduce

$$2n^2 = m^2. \tag{3.4.6}$$

Since 2 is a prime number and $2|m^2$ we deduce that $2|m$, i.e., $m = 2m_1$ for some natural number m_1. Using this last equality in (3.4.6) we deduce

$$2n^2 = (2m_1)^2 = 4m_1^2 \Rightarrow n^2 = 2m_1^2.$$

Thus $2|n^2$, and arguing as above we deduce that $2|n$. Hence 2 is a common divisor of both m and n. This contradicts the starting assumption that $\gcd(m, n) = 1$ and proves that $\sqrt{2}$ cannot be rational. □

Now that we know that there exist irrational numbers, we can ask, how many there are. It turns out that most real numbers are irrational, but we will not prove this fact now.

3.5. Exercises

Exercise 3.1. (a) Suppose that X, Y are two sets such that $X \sim Y$. Prove that $Y \sim X$.
(b) Prove that if X, Y, Z are sets such that $X \sim Y$ and $Y \sim Z$, then $X \sim Z$. □

Exercise 3.2. Prove by induction that for any natural numbers m, n and any real number x we have the equality

$$x^{m+n} = (x^m) \cdot (x^n). \qquad \Box$$

Exercise 3.3. (a) Prove that for any natural number n and any real numbers

$$a_1, a_2, \dots, a_n, b_1, \dots, b_n, c$$

we have the equalities

$$\sum_{k=1}^{n} (a_k + b_k) = \sum_{i=1}^{n} a_i + \sum_{j=1}^{n} b_j, \;\; \sum_{k=1}^{n} (ca_k) = c\left(\sum_{k=1}^{n} a_k\right). \qquad \Box$$

(b) Using (a) and (3.2.1) prove that for any natural number n and any real numbers a, r we have the equality

$$\sum_{k=0}^{n} (a + kr) = a + (a + r) + (a + 2r) \cdots + (a + nr) = (n + 1)a + \frac{rn(n + 1)}{2}.$$

(c) Use (b) to compute

$$3 + 7 + 11 + 15 + 19 + \cdots + 999,999.$$

Express the above using the symbol $\sum$.
(d) Prove that for any natural number n we have the equality

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2. \qquad \Box$$

Exercise 3.4. Prove that for any natural number n we have the equalities

$$\sum_{k-1}^{n} k^2 = \frac{1}{6}n(n + 1)(2n + 1),$$

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

□

Exercise 3.5. Prove that for any natural number n and any positive real numbers x, y such that $x < 1 < y$ we have

$$x^n \leqslant x, \;\; y \leqslant y^n. \qquad \Box$$

Exercise 3.6. Prove that if $0 < a < b$, and $n \geqslant 2$, then

$$\sqrt[n]{a} < \sqrt[n]{b},\ \ a < \sqrt{ab} < \frac{a+b}{2} < b. \qquad \square$$

Exercise 3.7. Find a natural number N_0 with the following property: for any $n > N_0$ we have

$$0 < \frac{n}{n^2+1} < \frac{1}{10^6} = \frac{1}{1,000,000}. \qquad \square$$

Exercise 3.8. Prove that for any natural number n and any real number $x \neq 1$ we have the equality.

$$\frac{1-x^n}{1-x} = 1 + x + x^2 + \cdots + x^{n-1}. \qquad \square$$

Exercise 3.9. (a) Compute

$$\binom{11}{2}, \binom{11}{3}, \binom{11}{8}, \binom{15}{4}, \binom{15}{11}.$$

(b) Show that for any $n, k \in \mathbb{N} \cup \{0\}$, $k \leqslant n$ we have

$$\binom{n}{k} = \binom{n}{n-k}.$$

(c) Use Newton's binomial formula to show that for any natural number n we have the equalities

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n,$$

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \cdots + (-1)^n\binom{n}{n} = 0.$$

Deduce that

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1}. \qquad \square$$

Exercise 3.10. Show that for any real number x, the interval $(x-1, x]$ contains exactly one integer.

Hint: For uniqueness use the Corollaries 3.1.8 and 3.1.9. To prove existence consider separately the cases

- $x \in \mathbb{Z}$.
- $(x \in \mathbb{R}\backslash\mathbb{Z}) \wedge (x > 0)$.
- $(x \in \mathbb{R}\backslash\mathbb{Z}) \wedge (x < 0)$.

□

Exercise 3.11. Let a, b, c be real numbers, $a \neq 0$.

(a) Show that

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \left(\frac{b^2 - 4ac}{4a}\right), \quad \forall x \in \mathbb{R}.$$

(b) Prove that the following statements are equivalent.

(i) There exist $r_1, r_2 \in \mathbb{R}$ such that

$$ax^2 + bx + c = a(x - r_1)(x - r_2).$$

(ii) There exists $r \in \mathbb{R}$ such that $ar^2 + br + c = 0$.

(iii) $b^2 - 4ac \geqslant 0$.

□

Exercise 3.12. Find the ranges of the functions

$$f : (-\infty, 5) \to \mathbb{R}, \quad f(x) = \frac{x+1}{x-5},$$

and

$$g : \mathbb{R} \to \mathbb{R}, \quad g(x) = \frac{x}{x^2+1}.$$

□

Exercise 3.13. (a) Show that the equation $x^2 - x - 1 = 0$ has two solutions $r_1, r_2 \in \mathbb{R}$ and then prove that r_1, r_2 satisfy the equalities

$$r_1 + r_2 = 1, \quad r_1 r_2 = -1.$$

(b) For any nonnegative integer n we set

$$F_n = \frac{r_1^{n+1} - r_2^{n+1}}{r_1 - r_2},$$

where r_1, r_2 are as in (a). Compute F_0, F_1, F_2.

(c) Prove by induction that for any nonnegative integer n we have

$$r_1^{n+2} = r_1^{n+1} + r_1^n, \quad r_2^{n+2} = r_2^{n+1} + r_2^n,$$

and

$$F_{n+2} = F_{n+1} + F_n.$$

(d) Use the above equality to compute $F_3, \dots, F_9$. □

Exercise 3.14. Prove Propositions 3.3.4 and 3.4.3 . □

Exercise 3.15. (a) Verify that for any $a, b > 0$ and any $m, n \in \mathbb{N}$ we have the equalities

$$(ab)^{\frac{1}{n}} = a^{\frac{1}{n}} \cdot b^{\frac{1}{n}}, \quad \left(a^{\frac{1}{n}}\right)^{\frac{1}{m}} = a^{\frac{1}{mn}},$$

$$(a^m)^{\frac{1}{n}} = \left(a^{\frac{1}{n}}\right)^m =: a^{\frac{m}{n}},$$

$$a^{\frac{km}{kn}}=a^{\frac{m}{n}},\ \ \forall k\in\mathbb{N}.$$

$$(a^{\frac{m}{n}})^{-1}=\big(a^{-1}\big)^{\frac{m}{n}}=:a^{-\frac{m}{n}}.$$

$$a^{-\frac{km}{kn}}=a^{-\frac{m}{n}},\ \ \forall k\in\mathbb{N}.$$

☞ *Recall that an expression of the form "bla-bla-bla =: x" signifies that the quantity x is defined to be whatever bla-bla-bla means. In particular the notation*

$$\big(a^{\frac{1}{n}}\big)^m=:a^{\frac{m}{n}}$$

indicates that the quantity $a^{\frac{m}{n}}$ is defined to be the m-th power of the n-th root of a.

(b) Prove that if $a>0$, then for any $m,m'\in\mathbb{Z}$ and $n,n'\in\mathbb{N}$ such that

$$\frac{m}{n}=\frac{m'}{n'},$$

then

$$a^{\frac{m}{n}}=a^{\frac{m'}{n'}}$$

Any rational number r admits a nonunique representation as a fraction

$$r=\frac{m}{n},\ \ m\in\mathbb{Z},\ \ n\in\mathbb{N}.$$

Part (b) allows us to give a well defined meaning to a^r, >0, $r\in\mathbb{Q}$.

(c) Show that for all $r_1,r_2\in\mathbb{Q}$ and any $a>0$ we have

$$a^{r_1}\cdot a^{r_2}=a^{r_1+r_2}.$$

(d) Suppose that $a>b>0$. Prove that for any rational number $r>0$ we have

$$a^r>b^r.$$

(e) Suppose that $a>1$. Prove that for any rational numbers r_1,r_2 such that $r_1<r_2$ we have

$$a^{r_1}<a^{r_2}.$$

(f) Suppose that $a\in(0,1)$. Prove that for any rational numbers r_1,r_2 such that $r_1<r_2$ we have

$$a^{r_1}>a^{r_2}.$$

□

3.6. Exercises for extra credit

Exercise* 3.1. There are 5 heads and 14 legs in a family. How many people and how many dogs are in the family? . □

Exercise* 3.2. You have two vessels of volumes 5 liters and 3 litters respectively. Measure one liter, producing it in one of the vessels. □

Exercise* 3.3. Each number from from 1 to 10^{10} is written out in formal English (e.g., "two hundred eleven", "one thousand forty-two") and then listed in alphabetical order (as in a dictionary, where spaces and hyphens are ignored). What is the first odd number in the list? □

Exercise* 3.4. Consider the map $f : \mathbb{N} \to \mathbb{Z}$ defined by

$$f(n) = (-1)^{n+1} \left\lfloor \frac{n}{2} \right\rfloor .$$

(i) Compute $f(1), f(2), f(3), f(4), f(5), f(6), f(7)$.

(ii) Given a natural number k, compute $f(2k)$ and $f(2k-1)$.

(iii) Prove that f is a bijection.

□

Exercise* 3.5. (a) Let p be a prime number and n a natural number > 1. Prove that $\sqrt[n]{p}$ is irrational.

(b) Let m, n be natural numbers and p, q prime numbers. Prove that

$$p^{1/m} = q^{1/n} \iff (p = q) \ \wedge \ (m = n).$$ □

Exercise* 3.6. Start with the natural numbers $1, 2, \dots, 999$ and change it as follows: select any two numbers, and then replace them by a single number, their difference. After 998 such changes you are left with a single number. Show that this number must be even. □

Exercise* 3.7. Let $S \subset [0, 1]$ be a set satisfying the following two properties.

(i) $0, 1 \in S$.

(ii) For any $n \in \mathbb{N}$ and any pairwise distinct numbers $s_1, \dots, s_n \in S$ we have

$$\frac{s_1 + \cdots + s_n}{n} \in S.$$

Show that $S = \mathbb{Q} \cap [0, 1]$. □

Exercise* 3.8. Given 25 positive real numbers, prove that you can choose two of them x, y so none of the remaining numbers is equal to the sum $x + y$ or the differences $x - y$, $y - x$. □

Exercise* 3.9. At a stockholders' meeting, the board presents the month-by-month profit (or losses) since the last meeting. "Note" says the CEO, "that we made a profit over every consecutive eight-month period."

"Maybe so", a shareholder complains, "but I also see that we *lost* over every consecutive *five*-month period!"

What is the maximum number of months that could have passed since the last meeting? □

Exercise* 3.10 (Erdös-Szekeres)**.** Suppose we are given an injection $f : \{1, \ldots, 10001\} \to \mathbb{R}$. Prove that there exists a subset $I \subset \{1, \ldots, 10001\}$ of cardinality 101 such that, either

$$f(i_1) < f(i_2), \ \ \forall i_1, i_2 \in I, \ \ i_1 < i_2,$$

or

$$f(i_1) > f(i_2), \ \ \forall i_1, i_2 \in I, \ \ i_1 < i_2. \qquad \square$$

Exercise* 3.11 (Chebyshev)**.** Suppose that $p_1, \ldots, p_n$ are positive numbers such that

$$p_1 + \cdots + p_n = 1.$$

Prove that if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are real numbers such that

$$x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \ \text{ and } \ y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n,$$

then

$$\sum_{k=1}^{n} x_k y_k p_k \geqslant \left(\sum_{i=1}^{n} x_i p_i \right) \left(\sum_{j=1}^{n} y_j p_j \right). \qquad \square$$

Exercise* 3.12. Let $k \in \mathbb{N}$. We are given k pairwise disjoint intervals $I_1, \ldots, I_k \subset [0, 1]$. Denote by S their union. We know that for any $d \in [0, 1]$ there exist two points $p, q \in S$ such that $\operatorname{dist}(p, q) = d$. Prove that

$$\text{length}\,(I_1) + \cdots + \text{length}\,(I_k) \geqslant \frac{1}{k}. \qquad \square$$

Chapter 4

Limits of sequences

The concept of limit is the central concept of this course. This chapter deals with the simplest incarnation of this concept namely, the notion of limit of a sequence of real numbers.

4.1. Sequences

Formally, a *sequence* of real numbers is a function $x : \mathbb{N} \to \mathbb{R}$. We typically describe a sequence $x : \mathbb{N} \to \mathbb{R}$ as a list $(x_n)_{n \in \mathbb{N}}$ consisting of one real number for each natural number n,

$$x_1, x_2, x_3, \dots, x_n, \dots .$$

Often we will allow lists that start at time 0, $(x_n)_{n \geqslant 0}$,

$$x_0, x_1, x_2, \dots .$$

If we use our intuition of a real number as corresponding to a point on a line, we can think of a sequence $(x_n)_{n \geqslant 1}$ as describing the motion of an object along the line, where x_n describes the position of that object at time n.

Example 4.1.1. (a) The natural numbers form a sequence $(n)_{n \in \mathbb{N}}$,

$$1, 2, 3, \dots .$$

(b) The *arithmetic progression* with initial term $a \in \mathbb{R}$ and ratio $r \in \mathbb{R}$ is the sequence

$$a, a + r, a + 2r, a + 3r, \dots .$$

For example, the sequence

$$3, 7, 11, 15, 19, \cdots$$

is an arithmetic progression with initial term 3 and ratio 4. The constant sequence

$$a, a, a, \dots ,$$

is an arithmetic progression with initial term a and ratio 0.

(c) The *geometric progression* with initial term $a \in \mathbb{R}$ and ratio $r \in \mathbb{R}$ is the sequence

$$a, ar, ar^2, ar^3, \dots .$$

For example, the sequence

$$1, -1, 1, -1,$$

is the geometric progression with initial term 1 and ratio -1.

(d) The *Fibonacci sequence* is the sequence $F_0, F_1, F_2, \dots$ given by the initial condition

$$F_0 = F_1 = 1,$$

and the recurrence relation

$$F_{n+2} = F_{n+1} + F_n, \quad \forall n \geqslant 0.$$

For example

$$F_2 = 1 + 1 = 2, \quad F_3 = 2 + 1 = 3, \quad F_4 = 3 + 2 = 5, \quad F_5 = 5 + 3 = 8, \dots .$$

In Exercise 3.13 we gave an alternate description to the Fibonacci sequence. □

Definition 4.1.2. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real numbers.

(i) The sequence $(x_n)_{n\in\mathbb{N}}$ is called *increasing* if

$$x_n < x_{n+1}, \quad \forall n \in \mathbb{N}.$$

(ii) The sequence $(x_n)_{n\in\mathbb{N}}$ is called *decreasing* if

$$x_n > x_{n+1}, \quad \forall n \in \mathbb{N}.$$

(iii) The sequence $(x_n)_{n\in\mathbb{N}}$ is called *nonincreasing* if

$$x_n \geqslant x_{n+1}, \quad \forall n \in \mathbb{N}.$$

(iv) The sequence $(x_n)_{n\in\mathbb{N}}$ is called *nondecreasing* if

$$x_n \leqslant x_{n+1}, \quad \forall n \in \mathbb{N}.$$

(v) A sequence $(x_n)_{n\in\mathbb{N}}$ is called *monotone* if it is either nondecreasing, or nonincreasing. It is called *strictly monotone* if it is either increasing, or decreasing.

(vi) The sequence $(x_n)_{n\in\mathbb{N}}$ is called *bounded* if there exist real numbers m, M such that

$$m \leqslant x_n \leqslant M, \quad \forall n \in \mathbb{N}.$$ □

Note that an arithmetic progression is increasing if and only if its ratio is positive, while a geometric progression with positive initial term and positive ratio is monotone: it is increasing if the ratio is > 1, decreasing if the ratio < 1 and constant if the ratio is $= 1$. A geometric progression is bounded if and only if its ratio r satisfies $|r| \leqslant 1$.

A *subsequence* of a sequence $x : \mathbb{N} \to \mathbb{R}$ is a restriction of x to an infinite subset $S \subset \mathbb{N}$. An infinite subset $S \subset \mathbb{N}$ can itself be viewed as an *increasing* sequence of natural numbers

$$n_1 < n_2 < n_3 < \dots,$$

where

$$n_1 := \min S,\ \ n_2 := \min S \backslash \{n_1\}, \dots, n_{k+1} := \min S \backslash \{n_1, \dots, n_k\}, \dots .$$

Thus a subsequence of a sequence $(x_n)_{n \in \mathbb{N}}$ can be described as a sequence $(x_{n_k})_{k \in \mathbb{N}}$, where $(n_k)_{k \in \mathbb{N}}$ is an increasing sequence of natural numbers.

4.2. Convergent sequences

Definition 4.2.1. We say that the sequence of real numbers (x_n) *converges to the number* $x \in \mathbb{R}$ if

$$\forall \varepsilon > 0 :\ \ \exists N = N(\varepsilon) \in \mathbb{N} \ \text{ such that } \ \forall n > N(\varepsilon) \ \text{ we have } \ |x_n - x| < \varepsilon. \tag{4.2.1}$$

A sequence (x_n) is called *convergent* if it converges to some number x. More precisely, this means

$$\exists x \in \mathbb{R},\ \ \forall \varepsilon > 0 :\ \ \exists N = N(\varepsilon) \in \mathbb{N} \ \text{ such that } \ \forall n > N(\varepsilon) \ \text{ we have } \ |x_n - x| < \varepsilon. \tag{4.2.2}$$

The number x is called *a limit* of the sequence (a_n). A sequence is called *divergent* if it is not convergent. □

Observe that condition (4.2.1) can be rephrased as follows

$$\forall \varepsilon > 0 :\ \ \exists N = N(\varepsilon) \in \mathbb{N} \ \text{ such that } \ \forall n > N(\varepsilon) \ \text{ we have } \ \operatorname{dist}(x_n, x) < \varepsilon. \tag{4.2.3}$$

Before we proceed further, let us observe the following simple fact.

Proposition 4.2.2. *Given a sequence (x_n) there exists at most one real number x satisfying the convergence property (4.2.1).*

Proof. Suppose that x, x' are two real numbers satisfying (4.2.1). Thus,

$$\forall \varepsilon > 0 :\ \ \exists N = N(\varepsilon) \in \mathbb{N} \ \text{ such that } \ \forall n > N \ \text{ we have } \ |x_n - x| < \varepsilon,$$

and

$$\forall \varepsilon > 0 :\ \ \exists N' = N'(\varepsilon) \in \mathbb{N} \ \text{ such that } \ \forall n > N', \ \text{ we have } \ |x_n - x'| < \varepsilon.$$

Thus, if $n > N_0(\varepsilon) := \max(\, N(\varepsilon), N'(\varepsilon)\,)$ then

$$|x_n - x|,\ |x_n - x'| < \varepsilon.$$

We observe that if $n > N_0(\varepsilon)$, then

$$|x - x'| = |(x - x_n) + (x_n - x')| \leqslant |x - x_n| + |x_n - x'| < 2\varepsilon.$$

In other words

$$\forall \varepsilon > 0 :\ \ |x - x'| < 2\varepsilon,\ \ \forall n > N_0(\varepsilon).$$

In the above statement the variable n really plays no role: if $|x - x'| < 2\varepsilon$ for some n, then clearly $|x - x'| < 2\varepsilon$ for any n. We conclude that

$$\forall \varepsilon > 0 : \quad |x - x'| < 2\varepsilon.$$

In other words, the distance $\operatorname{dist}(x, x') = |x - x'|$ between x and x' is smaller than any positive real number, so that this distance must be zero (Exercise 2.14) and hence $x = x'$. □

Definition 4.2.3. Given a convergent sequence (x_n), the unique real number x satisfying the convergence condition (4.2.1) is called *the limit* of the sequence (x_n) and we will indicate this using the notations

$$x = \lim_{n\to\infty} x_n \quad \text{or} \quad x = \lim_n x_n.$$

We will also say that (x_n) *tends (or converges) to x as n goes to ∞.* □

Observe that

$$\lim_{n\to\infty} x_n = x \Longleftrightarrow \lim_{n\to\infty} |x_n - x| = 0. \tag{4.2.4}$$

The next example shows that convergent sequences do exist.

Example 4.2.4. (a) If (x_n) is the constant sequence, $x_n = x$, for all n, then (x_n) is convergent and its limit is x.

(b) We want to show that

$$\boxed{\lim_{n\to\infty} \frac{C}{n} = 0, \quad \forall C > 0}. \tag{4.2.5}$$

Let $\varepsilon > 0$ and set $N(\varepsilon) := \lfloor \frac{C}{\varepsilon} \rfloor + 1 \in \mathbb{N}$. We deduce

$$N(\varepsilon) > \frac{C}{\varepsilon}, \quad \text{i.e.,} \quad \frac{N(\varepsilon)}{C} > \frac{1}{\varepsilon}.$$

For any $n > N(\varepsilon)$ we have

$$\frac{n}{C} > \frac{N(\varepsilon)}{C} > \frac{1}{\varepsilon} \Rightarrow \frac{C}{n} < \varepsilon.$$

Hence for any $n > N(\varepsilon)$ we have

$$|x_n| = \frac{C}{n} < \varepsilon.$$

□

Definition 4.2.5. (a) A *neighborhood* of a real number x is defined to be an *open* interval (α, β) that contains x, i.e., $x \in (\alpha, \beta)$.

(b) A *neighborhood of* ∞ is an interval of the form (M, ∞), while a *neighborhood of* $-\infty$ is an interval of the form $(-\infty, M)$. □

We have the following equivalent description of convergence. Its proof is left to you as an exercise.

Proposition 4.2.6. *Let (x_n) be a sequence of real numbers. Prove that the following statements are equivalent.*

(i) *The sequence (x_n) converges to $x \in \mathbb{R}$ as $n \to \infty$.*

(ii) *For any neighborhood U of x there exists a natural number N such that*

$$\forall n\big(\, n > N \Rightarrow x_n \in U\,\big). \qquad \square$$

The proof of the following result is left to you as an exercise.

Proposition 4.2.7. *Suppose that $(x_n)_{n \in \mathbb{N}}$ is a convergent sequence and $x = \lim_{n\to\infty} x_n$.*

(i) *If $(x_{n_k})_{k \geqslant 1}$ is a subsequence of (x_n), then*

$$\lim_{k\to\infty} x_{n_k} = x.$$

(ii) *Suppose that $(x'_n)_{n\in\mathbb{N}}$ is another sequence with the following property*

$$\exists N_0 \in \mathbb{N}: \;\; \forall n > N_0 \;\; x'_n = x_n.$$

Then

$$\lim_{n\to\infty} x'_n = x. \qquad \square$$

Part (ii) of the above proposition shows that the convergence or divergence of a sequence is not affected if we modify only finitely many of its terms. The next result is very intuitive.

Proposition 4.2.8 (Squeezing Principle)**.** *Let (a_n) (x_n), (y_n) be sequences such that*

$$\exists N_0 \in \mathbb{N} : \forall n > N_0, \;\; x_n \leqslant a_n \leqslant y_n.$$

If

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = a,$$

then

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

Proof. We have

$$\operatorname{dist}(a_n, a) \leqslant \operatorname{dist}(a_n, x_n) + \operatorname{dist}(x_n, a).$$

Since a_n lies in the interval $[x_n, y_n]$ for $n > N_0$ we deduce that

$$\operatorname{dist}(a_n, x_n) \leqslant \operatorname{dist}(y_n, x_n), \;\; \forall n > N_0,$$

so that

$$\operatorname{dist}(a_n, a) \leqslant \operatorname{dist}(y_n, x_n) + \operatorname{dist}(x_n, a), \;\; \forall n > N_0.$$

Now observe that

$$\operatorname{dist}(y_n, x_n) \leqslant \operatorname{dist}(y_n, a) + \operatorname{dist}(a, x_n).$$

Hence,

$$\begin{aligned}\operatorname{dist}(a_n,a) &\leqslant \operatorname{dist}(y_n,a)+\operatorname{dist}(a,x_n)+\operatorname{dist}(x_n,a)\\ &=\operatorname{dist}(y_n,a)+2\operatorname{dist}(x_n,a),\ \ \forall n>N_0.\end{aligned}\tag{4.2.6}$$

Let $\varepsilon>0$. Since $x_n\to a$ there exists $N_x(\varepsilon)\in\mathbb{N}$ such that

$$\forall n>N_x(\varepsilon):\ \ \operatorname{dist}(x_n,a)<\frac{\varepsilon}{3}.$$

Since $y_n\to a$ there exists $N_y(\varepsilon)\in\mathbb{N}$ such that

$$\forall n>N_y(\varepsilon):\ \ \operatorname{dist}(y_n,a)<\frac{\varepsilon}{3}.$$

Set $N(\varepsilon):=\max\{N_0,N_x(\varepsilon),N_y(\varepsilon)\}$. For $n>N(\varepsilon)$ we have

$$\operatorname{dist}(x_n,a)<\frac{\varepsilon}{3},\ \ \operatorname{dist}(y_n,a)<\frac{\varepsilon}{3}$$

and thus

$$\operatorname{dist}(y_n,a)+2\operatorname{dist}(x_n,a)<\varepsilon.$$

Using this in (4.2.6) we conclude that

$$\forall n>N(\varepsilon)\ \ \operatorname{dist}(a_n,a)<\varepsilon.$$

This proves that $a_n\to a$ as $n\to\infty$. □

Corollary 4.2.9. *Suppose that $a\in\mathbb{R}$ and (a_n), (x_n) are sequences of real numbers such that*

$$|a_n-a|\leqslant x_n\ \ \forall n,\ \ \lim_{n\to\infty}x_n=0.$$

Then

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

Proof. We have squeezed the sequence $|a_n-a|$ between the sequences (x_n) and the constant sequence 0, both converging to 0. Hence $|a_n-a|\to 0$ and, in view of (4.2.4), we deduce that also $a_n\to a$. □

Example 4.2.10. We want to show that

$$\boxed{\forall M>0,\ \ \forall r\in(-1,1)\lim_{n\to\infty}Mr^n=0}.\tag{4.2.7}$$

Clearly, it suffices to show that $M|r|^n\to 0$. This is clearly the case if $r=0$. Assume $r\neq 0$. Set

$$R:=\frac{1}{|r|}.$$

Then $R>1$ so that $R=1+\delta$, $\delta>0$. Bernoulli's inequality (3.2.2) implies that $\forall n\in\mathbb{N}$ we have $R^n\geqslant 1+n\delta$ so that

$$M|r|^n=\frac{M}{R^n}\leqslant\frac{M}{1+n\delta}\leqslant\frac{M}{n\delta}=\frac{C}{n},\ \ C:=\frac{M}{\delta}.$$

From Example 4.2.4 (b) we deduce that

$$\lim_n \frac{C}{n} = 0.$$

The desired conclusion now follows from the Squeezing Principle. □

Example 4.2.11. We want to prove that

$$\boxed{\lim_n \frac{r^n}{n!} = 0, \ \ \forall r \in \mathbb{R}}. \tag{4.2.8}$$

We will rely again on the Squeezing Principle. Fix $N_0 \in \mathbb{N}$ such that $N_0 > 2|r|$. Then for any $n > N_0$ we have

$$\left|\frac{r^n}{n!}\right| = \frac{|r|^n}{n!} = \frac{|r|^{N_0} r^{n-N_0}}{1 \cdot 2 \cdots N_0 \cdot (N_0+1)(N_0+2)\cdots n}$$
$$= \underbrace{\frac{|r|^{N_0}}{N_0!}}_{=:C_0} \cdot \underbrace{\frac{|r|}{N_0+1} \cdot \frac{|r|}{N_0+2} \cdots \frac{|r|}{n}}_{(n-N_0)\text{ terms}}.$$

Now observe that

$$\frac{|r|}{N_0+1}, \frac{|r|}{N_0+2}, \dots, \frac{|r|}{n} < \frac{|r|}{N_0} < \frac{1}{2},$$

and we deduce

$$\left|\frac{r^n}{n!}\right| < C_0 \left(\frac{1}{2}\right)^{n-N_0} = C_0 \left(\frac{1}{2}\right)^{-N_0} \left(\frac{1}{2}\right)^n = 2^{N_0} C_0 2^{-n}.$$

If we denote by M the constant $2^{N_0}C_0$ and we set $x_n := M2^{-n}$, $n \in \mathbb{N}$, we deduce that

$$\forall n > N_0 : \ \left|\frac{r^n}{n!}\right| < x_n.$$

Example 4.2.10 shows that $x_n \to 0$ and the conclusion (4.2.8) now follows from the Squeezing Principle. □

Proposition 4.2.12. *Any convergent sequence of real numbers is bounded.*

Proof. Suppose that $(a_n)_{n \geqslant 1}$ is a convergent sequence

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

There exists $N \in \mathbb{N}$ such that, for any $n > N$ we have

$$|a_n - a| < 1.$$

Thus, for any $n > N$ we have $a_n \in (a-1, a+1)$. Now set

$$m := \min\{a_1, a_2, \dots, a_N, a-1\}, \ \ M := \max\{a_1, a_2, \dots, a_N, a+1\}.$$

Then for any $n \geqslant 1$ we have

$$m \leqslant a_n \leqslant M,$$

i.e., the sequence (a_n) is bounded.

□

4.3. The arithmetic of limits

This section describes a few simple yet basic techniques that reduce the study of the convergence of a sequence to a similar study of potentially simpler sequences. Thus, we will prove that the sum of two convergent sequences is a convergent sequence etc.

Proposition 4.3.1 (Passage to the limit). *Suppose that $(a_n)_{n\geqslant 1}$ and $(b_n)_{n\geqslant 1}$ are two convergent sequences,*

$$a := \lim_{n\to\infty} a_n, \;\; b = \lim_{n\to\infty} b_n.$$

The following hold.

(i) *The sequence $(a_n + b_n)_{n\geqslant 1}$ is convergent and*

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n = a + b.$$

(ii) *If $\lambda \in \mathbb{R}$ then*

$$\lim_{n\to\infty}(\lambda a_n) = \lambda \lim_{n\to\infty} a_n = \lambda a.$$

(iii)

$$\lim_{n\to\infty}(a_n \cdot b_n) = \big(\lim_{n\to\infty} a_n\big) \cdot \big(\lim_{n\to\infty} b_n\big) = ab.$$

(iv) *Suppose that $b \neq 0$. Then there exists $N_0 > 0$ such that $b_n \neq 0$, $\forall N > N_0$ and*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b}.$$

(v) *Suppose that m, M are real numbers such that $m \leqslant a_n \leqslant M$, $\forall n$. Then*

$$m \leqslant \lim_{n\to\infty} a_n = a \leqslant M.$$

Proof. (i) Because (a_n) and (b_n) are convergent, for any $\varepsilon > 0$ there exist $N_a(\varepsilon), N_b(\varepsilon) \in \mathbb{N}$ such that

$$|a_n - a| < \frac{\varepsilon}{2}, \;\; \forall n > N_a(\varepsilon), \tag{4.3.1a}$$

$$|b_n - b| < \frac{\varepsilon}{2}, \;\; \forall n > N_b(\varepsilon). \tag{4.3.1b}$$

Let

$$N(\varepsilon) := \max\big\{N_a(\varepsilon),\; N_b(\varepsilon)\big\}.$$

Then for any $n > N(\varepsilon)$ we have $n > N_a(\varepsilon)$ and $n > N_b(\varepsilon)$ and

$$\big|\,(a_n + b_n) - (a + b)\,\big| = \big|\,(a_n - a) + (b_n - b)\,\big| \leqslant |a_n - a| + |b_n - b| \overset{(4.3.1a),(4.3.1b)}{<} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This proves that $\lim_{n\to\infty}(a_n + b_n) = a + b$.

(ii) If $\lambda = 0$, then the sequence (λa_n) is the constant sequence $0, 0, 0, \dots$ and the conclusion is obvious. Assume that $\lambda \neq 0$. The sequence (a_n) is convergent so for any $\varepsilon > 0$ there exists $N = N(\varepsilon) \in \mathbb{N}$ such that

$$|a_n - a| < \frac{\varepsilon}{|\lambda|}, \ \ \forall n > N(\varepsilon).$$

Hence for any $n > N(\varepsilon)$ we have

$$|\lambda a_n - \lambda a| = |\lambda| \cdot |a_n - a| < |\lambda| \cdot \frac{\varepsilon}{|\lambda|} = \varepsilon.$$

(iii) The sequences (a_n), (b_n) are convergent and thus, according to Proposition 4.2.12 they are bounded so that

$$\exists M > 0: \ \ |a_n|, |b_n| \leqslant M, \ \ \forall n.$$

We have

$$\begin{aligned} |a_nb_n - ab| &= |(a_nb_n - ab_n) + (ab_n - ab)| \leqslant |a_nb_n - ab_n| + |ab_n - ab| \\ &= |b_n| \cdot |a_n - a| + |a| \cdot |b_n - b| \leqslant M|a_n - a| + |a| \cdot |b_n - b|. \end{aligned}$$

Part (ii) coupled with the convergence of (a_n) and (b_n) show that

$$\lim_{n\to\infty} M|a_n - a| = \lim_{n\to\infty} |a| \cdot |b_n - b| = 0.$$

Using (i) we deduce

$$\lim_{n\to\infty} \big(M|a_n - a| + |a| \cdot |b_n - b| \big) = 0.$$

The squeezing principle shows that $|a_nb_n - ab| \to 0$.

(iv) Let us first show that if $b \neq 0$, then $b_n \neq 0$ for n sufficiently large. Since $b_n \to b$ there exists $N_0 \in \mathbb{N}$ such that

$$\forall n > N_0 \ \ |b_n - b| < \frac{|b|}{2}.$$

Thus, for any $n > N_0$, we have

$$\operatorname{dist}(b_n, b) = |b_n - b| < \frac{1}{2}|b| = \frac{1}{2}\operatorname{dist}(b, 0).$$

This shows that for $n > N_0$ we cannot have $b_n = 0$. In fact

$$|b_n| > \frac{|b|}{2}, \ \ \forall n > N_0. \tag{4.3.2}$$

Thus, the ratio $\frac{b_n}{b_n}$ is well defined at least for $n > N_0$. We have

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = \frac{|b_n - b|}{|b_n| \cdot |b|}.$$

The inequality (4.3.2) implies

$$\frac{1}{|b_n|} < \frac{2}{|b|}, \ \ \forall n > N_0.$$

Hence, for $n > N_0$ we have

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| < \frac{2}{|b|^2}|b_n - b| \to 0.$$

This implies

$$\lim_{n\to\infty} \frac{1}{b_n} = \frac{1}{b}.$$

Thus

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} a_n \cdot \lim_{n\to\infty} \frac{1}{b_n} = \frac{a}{b}.$$

(v) We argue by contradiction. Suppose that $a > M$ or $a < m$. We discuss what happens if $a > M$, the other situation being entirely similar. Then $\delta = a - M = \operatorname{dist}(a, M) > 0$. Since $a_n \to a$, there exists $N \in \mathbb{N}$ such that if $n > N$, then

$$\operatorname{dist}(a_n, a) = |a_n - a| < \frac{\delta}{2}.$$

Thus, for $n > N_0$ we have

$$a - \frac{\delta}{2} < a_n < a + \frac{\delta}{2}.$$

Clearly $M = a - \delta < a - \frac{\delta}{2}$ and thus, a fortiori, $a_n > M$ for $n > N_0$. Contradiction! □

Corollary 4.3.2. *Suppose that (a_n) and (b_n) are convergent sequences such that $a_n \geqslant b_n$, $\forall n$. Then*

$$\lim_{n\to\infty} a_n \geqslant \lim_{n\to\infty} b_n.$$

Proof. Let $c_n = a_n - b_n$. Then $c_n \geqslant 0$ $\forall n$ and thus

$$\lim_{n\to\infty} a_n - \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n \geqslant 0.$$

□

Let us see how the above simple principles work in practice.

Example 4.3.3. We already know that

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

We deduce that for any $k \in \mathbb{N}$ we have

$$\lim_{n\to\infty} \frac{1}{n^k} = 0.$$

Consider the sequence

$$a_n := \frac{5n^2 + 3n + 2}{3n^2 - 2n + 1}$$

We have

$$a_n = \frac{n^2(5 + \frac{3}{n} + \frac{2}{n^2})}{n^2(3 - \frac{2}{n} + \frac{1}{n^2})} = \frac{(5 + \frac{3}{n} + \frac{2}{n^2})}{(3 - \frac{2}{n} + \frac{1}{n^2})}.$$

Now observe that as $n \to \infty$

$$5 + \frac{3}{n} + \frac{2}{n^2} \to 5, \;\; 3 - \frac{2}{n} + \frac{1}{n^2} \to 3,$$

so that

$$\lim_{n\to\infty} a_n = \frac{5}{3}.$$

More generally, given $k \in \mathbb{N}$ and real numbers $a_0, b_0, \dots, a_k, b_k$ such that $b_k \neq 0$ then

$$\boxed{\lim_{n\to\infty} \frac{a_k n^k + \cdots + a_1 n + a_0}{b_k n^k + \cdots + b_1 n + b_0} = \frac{a_k}{b_k}}. \tag{4.3.3}$$

The proof is left to you as an exercise. □

Example 4.3.4. We want to show that

$$\boxed{\forall r > 1 \;\; \lim_n \frac{n}{r^n} = 0}. \tag{4.3.4}$$

We plan to use the Squeezing Principle and construct a sequence $(x_n)_{n\geqslant 1}$ of positive numbers such that

$$\frac{n}{r^n} \leqslant x_n \;\; \forall n \geqslant 2,$$

and

$$\lim_n x_n = 0.$$

Observe that since $r > 1$, we have $r - 1 > 0$. Set $a := r - 1$ so that $r = 1 + a$. Then, using Newton's binomial formula we deduce that if $n \geqslant 2$ then

$$r^n = (1+a)^n = 1 + \binom{n}{1}a + \binom{n}{2}a^2 + \cdots \geqslant 1 + \binom{n}{1}a + \binom{n}{2}a^2$$

$$= 1 + na + \frac{n(n-1)}{2}a^2 = 1 + na + \frac{a^2}{2}(n^2 - n).$$

Hence for $n \geqslant 2$ we have

$$\frac{1}{r^n} \leqslant \frac{1}{\frac{1}{2}(n^2 - n)a^2 + na + 1}$$

so that

$$\frac{n}{r^n} \leqslant \frac{n}{\frac{a^2}{2}(n^2 - n) + na + 1} =: x_n.$$

Now observe that

$$x_n = \frac{n}{n^2\big(\frac{a^2}{2}(1 - \frac{1}{n}) + \frac{a}{n} + \frac{1}{n^2}\big)} = \frac{\frac{1}{n}}{\frac{a^2}{2}(1 - \frac{1}{n}) + \frac{a}{n} + \frac{1}{n^2}} \overset{n\to\infty}{\longrightarrow} 0.$$ □

Example 4.3.5. We want to show that

$$\boxed{\lim_n \sqrt[n]{n} = 1}. \tag{4.3.5}$$

Let $\varepsilon > 0$. The number $r_\varepsilon = 1 + \varepsilon$ is > 1. Since $\frac{n}{r_\varepsilon^n} \to 0$ we deduce that there exists $N = N(\varepsilon) \in \mathbb{N}$ such that

$$\frac{n}{r_\varepsilon^n} < 1, \;\; \forall n > N(\varepsilon).$$

This translates into the inequality

$$n < r_\varepsilon^n = (1+\varepsilon)^n, \;\; \forall n > N(\varepsilon).$$

In particular

$$1 \leqslant \sqrt[n]{n} < \sqrt[n]{(1+\varepsilon)^n} = 1 + \varepsilon.$$

We have thus proved that for any $\varepsilon > 0$ we can find $N = N(\varepsilon) \in \mathbb{N}$ so that, as soon as $n > N(\varepsilon)$ we have

$$1 \leqslant \sqrt[n]{n} < 1 + \varepsilon.$$

Clearly this proves the equality (4.3.5). □

Definition 4.3.6 (Infinite limits)**.** Let $(a_n)_{n \in \mathbb{N}}$ be a sequence of real numbers.

(i) We say that a_n tends to ∞ as $n \to \infty$, and we write this

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

if

$$\forall C > 0 \;\; \exists N = N(C) \in \mathbb{N}: \;\; \forall n (n > N \Rightarrow a_n > C).$$

(ii) We say that a_n tends to $-\infty$ as $n \to \infty$, and we write this

$$\lim_{n\to\infty} a_n = -\infty$$

if

$$\forall C > 0 \;\; \exists N = N(C) \in \mathbb{N}: \;\; \forall n (n > N \Rightarrow a_n < -C).$$

□

Proposition 4.3.1 continues to hold if one or both of limits a, b are $\pm\infty$ provided we use the following conventions

$$\boxed{\infty + \infty = \infty \cdot \infty = \infty, \;\; \frac{C}{\infty} = 0, \;\; \forall C \in \mathbb{R}},$$

$$\boxed{C \cdot \infty = \begin{cases} \infty, & C > 0 \\ -\infty, & C < 0 \\ \text{undefined}, & C = 0, \end{cases}}$$

$$\infty - \infty = \text{undefined}, \;\; 0 \cdot \infty = \text{undefined}, \;\; \frac{\infty}{\infty} = \text{undefined}.$$

Example 4.3.7. (a) If we let $a_n = n$ and $b_n = \frac{1}{n}$, then Archimedes' Principle shows that $a_n \to \infty$ and $b_n \to 0$. We observe that $a_n b_n = 1 \to 1$. In this case $\infty \cdot 0 = 1$. On the other hand, if we let

$$a_n = n, \;\; b_n = \frac{1}{2^n}$$

then $a_n \to \infty$, $b_n \to 0$ and (4.3.4) shows that $a_n b_n \to 0$. In this case $\infty \cdot 0 = 0$.

(b) Consider the sequences $a_n = n$, $b_n = 2n$, $c_n = 3n$, $\forall n \in \mathbb{N}$. Observe that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \infty.$$

However

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\infty}{\infty} = \frac{1}{2}, \;\; \lim_{n\to\infty} \frac{a_n}{c_n} = \frac{\infty}{\infty} = \frac{1}{3}.$$ □

☛ ***Important Warning!*** When investigating limits of sequences you should keep in mind that the following arithmetic operations are treacherous and should be dealt with using *extreme care.*

$$\frac{\text{anything}}{0}, \;\; 0 \cdot \infty, \;\; \infty - \infty, \;\; \frac{\infty}{\infty}.$$

□

4.4. Convergence of monotone sequences

The definition of convergence has one drawback: to verify that a sequence is convergent using the definition we need to a priori know its limit. In most cases this is a nearly impossible job. In this section and the next we will discuss techniques for proving the convergence of a sequence without knowing the precise value of its limit.

> **Theorem 4.4.1** (Weierstrass). [a]*Any bounded and monotone sequence is convergent.*
>
> [a] Karl Weierstrass (1815-1897) was a German mathematician often cited as the "father of modern analysis"; see Wikipedia.

Proof. Suppose that (a_n) is a bounded and monotone sequence, i.e., it is either non-decreasing, or non-increasing. We investigate only the case when (a_n) is nondecreasing, i.e.,

$$a_1 \leqslant a_2 \leqslant a_3 \leqslant \cdots .$$

The situation when (a_n) is nonincreasing is completely similar.

The set of real numbers

$$A := \{\, a_n; \;\; n \geqslant 1 \,\}$$

is bounded because the sequence (a_n) is bounded. The Completeness Axiom implies it has a least upper bound

$$a := \sup A.$$

We will prove that

$$\lim_{n\to\infty} a_n = a. \tag{4.4.1}$$

Since a is an upper bound for the sequence we have

$$a_n \leqslant a, \ \ \forall n. \tag{4.4.2}$$

Since a is the *least* upper bound of A we deduce that for any $\varepsilon > 0$ the number $a-\varepsilon$ cannot be an upper bound of A. Hence, for any $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$ such that

$$a-\varepsilon < a_{N(\varepsilon)}.$$

Since (a_n) is nondecreasing we deduce that

$$a-\varepsilon < a_{N(\varepsilon)} \leqslant a_n, \ \ \forall n > N(\varepsilon) \tag{4.4.3}$$

Putting together (4.4.2) and (4.4.3) we deduce that

$$\forall \varepsilon > 0 \ \ \exists N = N(\varepsilon) \in \mathbb{N}: \ \ \forall n \ (n > N(\varepsilon) \Rightarrow a - \varepsilon < a_n \leqslant a).$$

This implies the claimed convergence (4.4.1) because $a-\varepsilon < a_n \leqslant a \Rightarrow |a_n - a| < \varepsilon$. □

We will spend the rest of this section presenting applications of the above *very important* theorem.

Example 4.4.2 (L. Euler)**.** Consider the sequence of positive numbers

$$x_n = \left(1+\frac{1}{n}\right)^n, \ \ n \in \mathbb{N}.$$

We will prove that this sequence is convergent. Its limit is called the *Euler*[1] *number* e.

We plan to use Weierstrass' theorem applied to a new sequence of positive numbers

$$y_n = \left(1+\frac{1}{n}\right)^{n+1}, \ \ n \in \mathbb{N}.$$

Note that

$$y_n = \left(\frac{n+1}{n}\right)^{n+1}$$

and for $n \geqslant 2$ we have

$$\frac{y_{n-1}}{y_n} = \frac{\left(\frac{n}{n-1}\right)^n}{\left(\frac{n+1}{n}\right)^{n+1}} = \left(\frac{n}{n-1}\right)^n \cdot \left(\frac{n}{n+1}\right)^{n+1}$$

$$= \frac{n^{2n+1}}{(n-1)^n(n+1)^n\cdot(n+1)} = \frac{n^{2n}}{(n^2-1)^n}\cdot\frac{n}{n+1}$$

[1] Leonhard Euler (1707-1783) was a Swiss mathematician, physicist, astronomer, logician and engineer who made important and influential discoveries in many branches of mathematics, He is also widely considered to be the most prolific mathematician of all time; see Wikipedia.

$$= \left(\frac{n^2}{n^2-1}\right)^n \cdot \frac{n}{n+1} = \underbrace{\left(1+\frac{1}{n^2-1}\right)^n}_{=:q_n} \cdot \frac{n}{n+1}.$$

Bernoulli's inequality implies that

$$q_n := \left(1+\frac{1}{n^2-1}\right)^n \geqslant 1+\frac{n}{n^2-1} > 1+\frac{n}{n^2} = 1+\frac{1}{n} = \frac{n+1}{n}.$$

Hence

$$\frac{y_{n-1}}{y_n} = q_n \cdot \frac{n}{n+1} > \frac{n+1}{n} \cdot \frac{n}{n+1} = 1.$$

Hence $y_{n-1} > y_n$ $\forall n \geqslant 2$, i.e., the sequence (y_n) is decreasing. Since it is bounded below by 1 we deduce that the sequence (y_n) is convergent.

Now observe that $y_n = x_n \cdot \left(1+\frac{1}{n}\right) = x_n \cdot \frac{n+1}{n}$ so that

$$x_n = y_n \cdot \frac{n}{n+1}.$$

Since

$$\lim_n \frac{n}{n+1} = 1$$

we deduce that (x_n) is convergent and has the same limit as the sequence (y_n). □

Definition 4.4.3. The *Euler number*, denoted e is defined to be

$$\boxed{e := \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n}.$$ □

The arguments in Example 4.4.2 show that

$$4 = y_1 \geqslant e \geqslant 2.$$

Using more sophisticated methods one can show that

$$e = 2.71828182845905\dots.$$

Example 4.4.4 (Babylonians and I. Newton)**.** Consider the sequence $(x_n)_{n\in\mathbb{N}}$ defined recursively by the requirements

$$x_1 = 1, \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right), \quad \forall n \in \mathbb{N}.$$

Thus

$$x_2 = \frac{1}{2}\left(1+\frac{2}{1}\right) = \frac{3}{2},$$

$$x_3 = \frac{1}{2}\left(\frac{3}{2}+\frac{2}{\frac{3}{2}}\right) = \frac{1}{2}\left(\frac{3}{2}+\frac{4}{3}\right) = \frac{17}{12} \text{ etc.}$$

We want to prove that this sequence converges to $\sqrt{2}$. We proceed gradually.

Lemma 4.4.5.

$$x_n \geqslant \sqrt{2}, \quad \forall n \geqslant 2. \tag{4.4.4}$$

Proof. Multiplying with $2x_n$ both sides of the equality

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$$

we deduce $2x_nx_{n+1} = x_n^2 + 2$, or equivalently

$$x_n^2 - 2x_{n+1}x_n + 2 = 0. \tag{4.4.5}$$

This shows that the quadratic equation

$$t^2 - 2x_{n+1}t + 2 = 0$$

has at least one real solution, $t = x_{n+1}$ so that (see Exercise 3.11)

$$\Delta = 4x_{n+1}^2 - 8 \geqslant 0,$$

i.e., $x_{n+1}^2 \geqslant 2$, or $x_{n+1} \geqslant \sqrt{2}$, $\forall n \in \mathbb{N}$. □

Lemma 4.4.6. *For any $n \geqslant 2$ we have*

$$x_{n+1} \leqslant x_n.$$

Proof. Let $n \geqslant 2$. We have

$$x_n - x_{n+1} = x_n - \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) = \frac{1}{2}\frac{x_n^2 - 2}{x_n} \overset{(4.4.4)}{\geqslant} 0.$$

□

Thus the sequence $(x_n)_{n\geqslant 2}$ is decreasing and bounded below and thus it is convergent. Denote by $\bar{x}$ the limit. The inequality (4.4.4) implies that $\bar{x} \geqslant \sqrt{2}$. Letting $n \to \infty$ in (4.4.5) we deduce

$$\bar{x}^2 - 2\bar{x}^2 + 2 = 0 \Rightarrow 2 = \bar{x}^2 \Rightarrow \bar{x} = \sqrt{2}.$$

For example

$$x_2 = 1.5, \quad x_3 = 1.4166..., \quad x_4 := 1.4142..., \quad x_5 := 1.4142....$$

Note that

$$(1.4142)^2 = 1.99996164.$$

□

Theorem 4.4.7 (Nested Intervals Theorem)**.** *Consider a* nested sequence *of* ***closed*** *intervals $[a_n, b_n]$, $n \in \mathbb{N}$, i.e.,*

$$[a_1, b_1] \supset [a_2, b_2] \supset [a_3, b_3] \supset \cdots.$$

Then there exists $x \in \mathbb{R}$ that belongs to all the intervals, i.e.,

$$\bigcap_{n \in \mathbb{N}} [a_n, b_n] \neq \varnothing.$$

Proof. The nesting condition implies that for any $n \in \mathbb{N}$ we have

$$a_n \leqslant a_{n+1} \leqslant b_{n+1} \leqslant b_n.$$

This shows that the sequence (a_n) is nondecreasing and bounded while the sequence (b_n) is non-increasing. Therefore, these sequences are convergent and we set

$$a := \lim_n a_n, \quad b := \lim_n b_n$$

the condition $a_n \leqslant b_n$, $\forall n$ implies that

$$a_n \leqslant a \leqslant b \leqslant b_n, \quad \forall n.$$

Hence $[a, b] \subset [a_n, b_n]$, $\forall n$. □

Theorem 4.4.8 (Bolzano-Weierstrass)**.** *Any bounded sequence has a convergent subsequence.*

Proof. Let (x_n) be a bounded sequence of real numbers. Thus, there exist real numbers a_1, b_1 such that $x_n \in [a_1, b_1]$, for all n. We set

$$n_1 := 1.$$

Divide the interval $[a_1, b_1]$ into two intervals of equal length. At least one of these intervals will contain infinitely many terms of the sequence (x_n). Pick such an interval and denote it by $[a_2, b_2]$. Thus

$$[a_1, b_1] \supset [a_2, b_2], \quad b_2 - a_2 = \frac{1}{2}(b_1 - a_1).$$

Choose $n_2 > 1$ such that $x_{n_2} \in [a_2, b_2]$. We now proceed inductively.

Suppose that we have produced the intervals

$$[a_1, b_1] \supset [a_2, b_2] \supset \cdots \supset [a_k, b_k]$$

and the natural numbers $n_1 < n_2 < \cdots < n_k$ such that

$$b_2 - a_2 = \frac{1}{2}(b_1 - a_1), \quad b_3 - a_3 = \frac{1}{2}(b_2 - a_2), \quad b_k - a_k = \frac{1}{2}(b_{k-1} - a_{k-1}),$$

$$x_{n_1} \in [a_1, b_1], \quad x_{n_2} \in [a_2, b_2], \cdots x_{n_k} \in [a_k, b_k],$$

and the interval $[a_k, b_k]$ contains infinitely many terms of the sequence (x_n). We then divide $[a_k, b_k]$ into two intervals of equal lengths. One of them will contain infinitely many terms of (x_n). Denote that interval by $[a_{k+1}, b_{k+1}]$. We can then find a natural number $n_{k+1} > n_k$ such that $x_{n_{k+1}} \in [a_{k+1}, b_{k+1}]$. By construction

$$b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k) = \cdots = \frac{1}{2^k}(a_1 - b_1).$$

We have thus produced sequences (a_k), (b_k) (x_{n_k}) with the properties

$$a_1 \leqslant a_2 \leqslant \cdots \leqslant a_k \leqslant x_{n_k} \leqslant b_k \leqslant \cdots \leqslant b_2 \leqslant b_1, \tag{4.4.6a}$$

$$b_k - a_k = \frac{1}{2^{k-1}}(b_1 - a_1). \tag{4.4.6b}$$

The inequalities (4.4.6a) show that the sequences (a_k) and (b_k) are monotone and bounded, and thus have limits which we denote by a and b respectively. By letting $k \to \infty$ in (4.4.6b) we deduce that $a = b$.

The subsequence (x_{n_k}) is squeezed between two sequences converging to the same limit so the squeezing theorem implies that it is convergent. □

Definition 4.4.9. A *limit point* of a sequence of real numbers (x_n) is a real number which is the limit of some subsequence of the original sequence (x_n). □

Example 4.4.10. Consider the sequence

$$x_n = (-1)^n + \frac{1}{n}, \quad n \in \mathbb{N}.$$

Thus

$$x_{2n} = 1 + \frac{1}{2n}, \quad x_{2n+1} = -1 + \frac{1}{2n+1}.$$

Then the numbers 1 and -1 are limit points of this sequence because

$$\lim_{n\to\infty} x_{2n} = \lim_{n\to\infty}\left(1 + \frac{1}{2n}\right) = 1,$$

$$\lim_{n\to\infty} x_{2n+1} = \lim_{n\to\infty}\left(-1 + \frac{1}{2n+1}\right) = -1.$$ □

4.5. Fundamental sequences and Cauchy's characterization of convergence

We know that any convergent sequence is bounded. In other words, so boundedness is a necessary condition for a sequence to be convergent. However, it is not also a sufficient condition. For example, the sequence

$$1, -1, 1, -1, \dots$$

is bounded, but it is not convergent.

Weierstrass's theorem on bounded monotone sequences shows that monotonicity is a sufficient condition for a bounded sequence to be convergent. However, monotonicity is not a necessary condition for convergence. Indeed, the sequence

$$x_n = \frac{(-1)^n}{n}, \quad n \in \mathbb{N}$$

converges to zero, yet it is not monotone because the even order terms are positive while the odd order terms are negative. In this subsection we will present a fundamental necessary and sufficient condition for a sequence to be convergent that makes no reference to the precise value of the limit. We begin by defining a very important concept.

Definition 4.5.1. A sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ is called *Cauchy*[2] or *fundamental* if the following holds:

$$\forall\varepsilon>0\ \ \exists N=N(\varepsilon)\in\mathbb{N}\ \text{ such that }\ \forall m,n>N(\varepsilon):\ \ |a_m-a_n|<\varepsilon. \tag{4.5.1}$$

□

Theorem 4.5.2 (Cauchy)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. Then the following statements are equivalent.*

(i) *The sequence (a_n) is convergent.*

(ii) *The sequence (a_n) is Cauchy.*

Proof. (i) ⇒ (ii). We know that there exists $a\in\mathbb{R}$ such that

$$\forall\varepsilon>0\ \ \exists N=N(\varepsilon)\in\mathbb{N}:\ \ \forall n>N(\varepsilon)\ \ |a_n-a|<\varepsilon.$$

Observe that for any $m,n>N(\varepsilon/2)$ we have

$$|a_m-a_n|\leqslant|a_m-a|+|a-a_n|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

This proves that (a_n) is fundamental.

(ii) ⇒ (i) This is the "meatier" part of the theorem. We will reach the conclusion in three conceptually distinct steps.

1. Using the fact that the sequence (a_n) is fundamental we will prove that it is bounded.
2. Since (a_n) is bounded, the Bolzano-Weierstrass theorem implies that it has a subsequence that converges to a real number a.
3. Using the fact that the sequence (a_n) is fundamental we will prove that it converges to the real number a found above.

Here are the details. Since (a_n) is fundamental, there exists $n_1>0$ such that, for any $m,n\geqslant n_1$ we have $|a_m-a_n|<1$. Hence if we let $m=n_1$ we deduce that for any $n\geqslant n_1$ we have

$$|a_{n_1}-a_n|<1\Rightarrow a_{n_1}-1<a_n<a_{n_1}+1,\ \ \forall n\geqslant n_1.$$

Now let

$$m:=\min\{a_1,a_2,\dots,a_{n_1-1},a_{n_1}-1\},\ \ M:=\max\{a_1,a_2,\dots,a_{n_1-1},a_{n_1}+1\}.$$

Clearly

$$m\leqslant a_n\leqslant M,\ \ \forall n\in\mathbb{N}$$

[2] Named after August-Louis Cauchy (1789-1857), French mathematician, reputed as a pioneer of analysis. He was one of the first to state and prove theorems of calculus rigorously, rejecting the heuristic principle of the generality of algebra of earlier authors; see Wikipedia.

so that the sequence (a_n) is bounded.

Invoking the Bolzano-Weierstrass theorem we deduce that there exists a subsequence $(a_{n_k})_{k\geqslant 1}$ and a real number a such that

$$\lim_{k\to\infty} a_{n_k} = a.$$

Let $\varepsilon > 0$. Since $a_{n_k} \to a$ as $k \to \infty$ we deduce that

$$\exists K = K(\varepsilon) \in \mathbb{N} \;\text{ such that }\; \forall k > K(\varepsilon) : \;\; |a_{n_k} - a| < \frac{\varepsilon}{2}.$$

On the other hand, the sequence $(a_n)_{n\in\mathbb{N}}$ is fundamental so that

$$\exists N' = N'(\varepsilon) \in \mathbb{N} \;\text{ such that }\; \forall m, n > N'(\varepsilon) : \;\; |a_m - a_n| < \frac{\varepsilon}{2}.$$

Now choose a natural number $k_0(\varepsilon) > K(\varepsilon)$ such that $n_{k_0(\varepsilon)} > N'(\varepsilon)$. Define

$$N(\varepsilon) = n_{k_0(\varepsilon)}.$$

If $n > N(\varepsilon)$ then $n, n_{k_0} > N'(\varepsilon)$ and thus

$$|a_n - a_{n_{k_0}}| < \frac{\varepsilon}{2}.$$

On the other hand, since $k_0(\varepsilon) > K(\varepsilon)$ we deduce that

$$|a_{n_{k_0}} - a| < \frac{\varepsilon}{2}.$$

Hence, for any $n > N(\varepsilon)$ we have

$$|a_n - a| \leqslant |a_n - a_{n_{k_0}}| + |a_{n_{k_0}} - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} < \varepsilon.$$

Since ε was arbitrary we conclude that (a_n) converges to a. □

4.6. Series

Often one has to deal with sums of infinitely many terms. Such a sum is called a *series*. Here is the precise definition.

Definition 4.6.1. The *series* associated to a sequence $(a_n)_{n\geqslant 0}$ of real numbers is the ***new*** sequence $(s_n)_{n\geqslant 0}$ defined by the *partial sums*

$$s_0 = a_0, \;\; s_1 = a_0 + a_1, \;\; s_2 = a_0 + a_1 + a_2, \dots, s_n = a_0 + a_1 + \cdots + a_n = \sum_{i=0}^{n} a_i, \; \dots .$$

The series associated to the sequence $(a_n)_{n\geqslant 0}$ is denoted by the symbol

$$\sum_{n=0}^{\infty} a_n \quad\text{or}\quad \sum_{n\geqslant 0} a_n.$$

The series is called *convergent* if the sequence of partial sums $(s_n)_{n\geqslant 0}$ is convergent. The limit $\lim_{n\to\infty} s_n$ is called *the sum* series. We will use the notation

$$\sum_{n\geqslant 0} a_n = S$$

to indicate that the series is convergent and its sum is the real number S. □

Example 4.6.2 (Geometric series. Part 1)**.** Let $r \in (-1, 1)$. The *geometric series*

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots$$

is convergent and we have the following *very useful equality*

$$\boxed{\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}}. \tag{4.6.1}$$

Indeed, the n-th partial sum of this series is

$$s_n = 1 + r + \cdots + r^n = \frac{1 - r^{n+1}}{1-r}.$$

Example 4.2.10 shows that when $|r| < 1$ we have $\lim_n r^{n+1} = 0$ so that

$$\sum_{n=0}^{\infty} r^n = \lim_n s_n = \frac{1}{1-r}.$$

Observe that if we set $r = \frac{1}{2}$ in (4.6.1) we deduce

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 2.$$

□

The proof of the following result is left to you as an exercise.

Proposition 4.6.3. *Consider two series*

$$\sum_{n\geqslant 0} a_n \;\; \textit{and} \;\; \sum_{n\geqslant 0} a_n'$$

such that there exists $N_0 > 0$ with the property

$$a_n = a_n' \;\; \forall n > N_0.$$

Then

$$\sum_{n\geqslant 0} a_n \;\; \textit{is convergent} \Longleftrightarrow \sum_{n\geqslant 0} a_n' \;\; \textit{is convergent.}$$

□

Proposition 4.6.4. *If the series $\sum_{n=0}^{\infty} a_n$ is convergent, then*

$$\lim_{n\to\infty} a_n = 0.$$

Proof. Observe that for $n \geqslant 1$

$$s_n = a_0 + a_1 + \cdots + a_{n-1} + a_n = s_{n-1} + a_n.$$

Hence

$$a_n = s_n - s_{n-1}.$$

The sequences $(s_n)_{n \geqslant 1}$ and $(s_{n-1})_{n \geqslant 1}$ converge to the same finite limit so that

$$\lim_n a_n = \lim_n s_n - \lim_n s_{n-1} = 0.$$

□

Example 4.6.5 (Geometric series. Part 2)**.** Let $|r| \geqslant 1$. Then the geometric series

$$1 + r + r^2 + \cdots + r^n + \cdots = \sum_{n=0}^{\infty} r^n$$

is divergent. Indeed, if it were convergent, then the above proposition would imply that $r^n \to 0$ as $n \to \infty$. This is not the case when $|r| \geqslant 1$. □

Proposition 4.6.6. *A series of positive numbers*

$$\sum_{n \geqslant 0} a_n, \quad a_n > 0 \ \forall n$$

is convergent if and only if the sequence of partial sums

$$s_n = a_0 + \cdots + a_n$$

is bounded.

Proof. Observe that the sequence of partial sums is increasing since

$$s_{n+1} - s_n = a_{n+1} > 0, \quad \forall n.$$

If the sequence (s_n) is also bounded, then Weierstrass' Theorem on monotone sequences implies that it must be convergent.

Conversely, if the sequence (s_n) is convergent, then Proposition 4.2.12 shows that it must also be bounded. □

Example 4.6.7. (a) The *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots .$$

is *divergent*. Here is why.

This is a series with positive terms. Observe that

$$s_1 = 1 \geqslant 1, \quad s_2 = 1 + \frac{1}{2} \geqslant 1 + \frac{1}{2},$$

$$s_{2^2} = s_4 = s_2 + \frac{1}{3} + \frac{1}{4} > s_2 + \frac{1}{4} + \frac{1}{4} = s_2 + \frac{1}{2} = 1 + \frac{2}{2}$$

$$\begin{aligned}s_{2^3}=s_8=s_4+\underbrace{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}_{\text{4 terms}}&>1+\frac{2}{2}+\underbrace{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}_{\text{4 terms}}\\&>1+\frac{2}{2}+\underbrace{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}_{\text{4 terms}}=1+\frac{3}{2}.\end{aligned}$$

Thus

$$s_{2^3}>1+\frac{3}{2}.$$

We want to prove that

$$s_{2^n}>1+\frac{n}{2},\ \ \forall n\geqslant 2. \tag{4.6.2}$$

We have shown this for $n=2$ and $n=3$. The general case follows inductively. Observe that $2^{n+1}=2\cdot 2^n=2^n+2^n$ and thus

$$\begin{aligned}s_{2^{n+1}}&=s_{2^n}+\underbrace{\frac{1}{2^n+1}+\cdots+\frac{1}{2^{n+1}}}_{2^n\text{-terms}}\\&>s_{2^n}+\underbrace{\frac{1}{2^{n+1}}+\cdots+\frac{1}{2^{n+1}}}_{2^n\text{-terms}}=s_{2^n}+\frac{2^n}{2^{n+1}}=s_{2^n}+\frac{1}{2}\end{aligned}$$

(use the inductive assumption)

$$>1+\frac{n}{2}+\frac{1}{2}.$$

This proves that (4.6.2) which shows that the sequence s_{2^n} is not bounded. Invoking Proposition 4.6.6 we conclude that the harmonic series is not convergent.

(b) Let $r>1$ be a rational number and consider the series

$$\sum_{n=1}^{\infty}\frac{1}{n^r}.$$

We want to show that this series is *convergent*.

We have

$$s_2=1+\frac{1}{2^r},$$

$$s_4=s_2+\frac{1}{3^r}+\frac{1}{4^r}<s_2+\frac{1}{2^r}+\frac{1}{2^r}<s_2+\frac{2}{2^r}=\frac{1}{2^r}+1+\frac{1}{2^{(r-1)}},$$

$$s_{2^3}=s_8=s_4+\frac{1}{5^r}+\frac{1}{6^r}+\frac{1}{7^r}+\frac{1}{8^r}<s_4+\frac{4}{4^r}=\frac{1}{2^r}+1+\frac{1}{2^{(r-1)}}+\frac{1}{2^{2(r-1)}}.$$

We claim that for any $n\geqslant 1$ we have

$$s_{2^{n+1}}<\frac{1}{2^r}+1+\frac{1}{2^{(r-1)}}+\frac{1}{2^{2(r-1)}}+\cdots+\frac{1}{2^{n(r-1)}}. \tag{4.6.3}$$

We argue inductively. The result is clearly true for $n = 1, 2$. We assume it is true for n and we prove it is true for $n+1$. We have

$$s_{2^{n+1}} = s_{2^n} + \underbrace{\frac{1}{(2^n+1)^r} + \frac{1}{(2^n+2)^r} + \cdots + \frac{1}{(2^{n+1})^r}}_{2^n \text{ terms}}$$

$$< s_{2^n} + \underbrace{\frac{1}{(2^n)^r} + \frac{1}{(2^n)^r} + \cdots + \frac{1}{(2^n)^r}}_{2^n \text{ terms}} = s_{2^n} + \frac{1}{2^{n(r-1)}}$$

(use the induction assumption)

$$< \frac{1}{2^r} + 1 + \frac{1}{2^{(r-1)}} + \frac{1}{2^{2(r-1)}} + \cdots + \frac{1}{2^{(n-1)(r-1)}} + \frac{1}{2^{n(r-1)}}.$$

If we set

$$q := \frac{1}{2^{r-1}} = \left(\frac{1}{2}\right)^{r-1},$$

then we observe that the condition $r > 1$ implies $q \in (0, 1)$ and we can rewrite (4.6.3) as

$$s_{2^{n+1}} < \frac{1}{2^r} + 1 + q + \cdots + q^n < \frac{1}{2^r} + \frac{1}{1-q}, \quad \forall n \in \mathbb{N}.$$

This implies that the sequence (s_{2^n}) is bounded and thus the series

$$\sum_{n=1}^{\infty} \frac{1}{n^r}$$

is convergent for any $r > 1$. Its sum is denoted by $\zeta(r)$ and it is called *Riemann zeta function* For most r's, the actual value $\zeta(r)$ is not known. However, L. Euler has computed the values $\zeta(r)$ when r is an even natural number. For example

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

All the known proofs of the above equality are very ingenious. □

Theorem 4.6.8 (Comparison principle)**.** *Suppose that*

$$\sum_{n\geqslant 0} a_n \quad \textit{and} \quad \sum_{n\geqslant 0} b_n$$

are two series of positive real numbers such that

$$\exists N_0 \in \mathbb{N} \ \textit{ such that } \ \forall n > N_0 : \ \ a_n < b_n.$$

Then the following hold.

(a) $\sum_{n\geqslant 0} a_n$ *divergent* $\Rightarrow \sum_{n\geqslant 0} b_n$ *divergent.*

(b) $\sum_{n\geqslant 0} b_n$ *convergent* $\Rightarrow \sum_{n\geqslant 0} a_n$ *convergent.*

Proof. We set

$$s_n(a)=\sum_{k=0}^n a_n,\ \ s_n(b)=\sum_{k=1}^n b_n.$$

In view of Proposition 4.6.3 the convergence or divergence of a series is not affected if we modify finitely many of its terms. Thus, we may assume that

$$a_n\leqslant b_n,\ \ \forall n\geqslant 0.$$

In particular, we have

$$s_n(a)\leqslant s_n(b),\ \ \forall n\geqslant 0. \tag{4.6.4}$$

Note that since the terms a_n are *positive*

$$\sum_{n\geqslant 0} a_n \ \text{ divergent} \Rightarrow s_n(a)\to\infty\Rightarrow s_n(b)\to\infty\Rightarrow\sum_{n\geqslant 0} b_n\ \text{ divergent}$$

and

$$\sum_{n\geqslant 0} b_n\ \text{ convergent}\Rightarrow s_n(b)\ \text{ bounded}\Rightarrow s_n(a)\ \text{ bounded}\Rightarrow\sum_{n\geqslant 0} a_n\ \text{ convergent.}$$

□

The above result has an immediate and very useful consequence whose proof is left to you as an exercise.

Corollary 4.6.9. *Suppose that*

$$\sum_{n\geqslant 0} a_n\ \textit{ and }\ \sum_{n\geqslant 0} b_n$$

are two series with positive terms.

(a) If the sequence $(\frac{a_n}{b_n})_{n\geqslant 0}$ is convergent and the series $\sum_{n\geqslant 0} b_n$ is convergent, then the series $\sum_{n\geqslant 0} a_n$ is also convergent.

(b) If the sequence $(\frac{a_n}{b_n})_{n\geqslant 0}$ has a limit r which is either positive, $r>0$, or $r=\infty$ and the series $\sum_{n\geqslant 0} b_n$ is divergent, then the series $\sum_{n\geqslant 0} a_n$ is also divergent. □

Example 4.6.10 (L. Euler)**.** Consider the series

$$\sum_{n=0}^\infty\frac{1}{n!}=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots. \tag{4.6.5}$$

Observe that if $n\geqslant 2$, then

$$\frac{1}{n!}=\frac{1}{2}\cdot\frac{1}{3}\cdots\frac{1}{n}\leqslant\underbrace{\frac{1}{2}\cdots\frac{1}{2}}_{(n-1)-\text{times}}=\frac{1}{2^{n-1}}=\frac{2}{2^n}.$$

Since the series

$$\sum_{n\geqslant 0}\frac{2}{2^n}$$

is convergent we deduce from the Comparison Principle that the series (4.6.5) is also convergent. Its sum is the Euler number

$$\sum_{n=0}^{\infty}\frac{1}{n!}=e=\lim_{n}\left(1+\frac{1}{n}\right)^{n}. \tag{4.6.6}$$

This is a nontrivial result. We will describe a more conceptual proof in Corollary 8.1.8. However, that proof relies on the full strength of differential calculus.

Here is an elementary proof. We set

$$e_n:=\left(1+\frac{1}{n}\right)^{n},\ \ s_n=1+\frac{1}{1!}+\frac{1}{2!}+\cdots+\frac{1}{n!},\ \ \forall n\in\mathbb{N},$$

We will prove two things.

$$e_n<s_n,\ \ \forall n\geqslant 1 \tag{4.6.7a}$$

$$s_k\leqslant e,\ \ \forall k\geqslant 1. \tag{4.6.7b}$$

Assuming the validity of the above inequalities, we observe that by letting $n\to\infty$ in (4.6.7a) we deduce that

$$e\leqslant\lim_{n}s_n.$$

On the other hand, if we let $k\to\infty$ in (4.6.7b), then we conclude that

$$\lim_{k}s_k\leqslant e.$$

Hence (4.6.7a, 4.6.7b) imply that

$$e=\lim_{n}s_n=\sum_{n=0}^{\infty}\frac{1}{n!}.$$

Proof of (4.6.7a). Using Newton's binomial formula we deduce

$$\begin{aligned}e_n&=\left(1+\frac{1}{n}\right)^{n}=1+\binom{n}{1}\frac{1}{n}+\binom{n}{2}\frac{1}{n^2}+\cdots+\binom{n}{n}\frac{1}{n^n}\\&=1+\frac{n}{1!}\frac{1}{n}+\frac{n(n-1)}{2!}\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n^3}+\cdots+\frac{n(n-1)\cdots 1}{n!}\frac{1}{n^n}\\&=1+\frac{n}{n}\frac{1}{1!}+\underbrace{\frac{n(n-1)}{n^2}}_{<1}\cdot\frac{1}{2!}+\underbrace{\frac{n(n-1)(n-2)}{n^3}}_{<1}\cdot\frac{1}{3!}+\cdots+\underbrace{\frac{n(n-1)\cdots 1}{n^n}}_{<1}\cdot\frac{1}{n!}\\&<1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}=s_n.\end{aligned}$$

Proof of (4.6.7b). Fix $k\in\mathbb{N}$. Then from the same formula above we deduce that if $k\leqslant n$, then

$$e_n=1+\frac{n}{1!}\frac{1}{n}+\frac{n(n-1)}{2!}\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n^3}+\cdots+\frac{n(n-1)\cdots 1}{n!}\frac{1}{n^n}$$

(neglect the terms containing the powers $\frac{1}{n^j}$, $j>k$)

$$\begin{aligned}&>1+\frac{n}{1!}\frac{1}{n}+\frac{n(n-1)}{2!}\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n^3}+\cdots+\frac{n(n-1)\cdots(n-k+1)}{k!}\frac{1}{n^k}\\&=1+\frac{1}{1!}+\frac{n-1}{n}\frac{1}{2!}+\frac{n-1}{n}\frac{n-2}{n}\cdot\frac{1}{3!}+\cdots+\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdot\frac{1}{k!}\\&=1+\frac{1}{1!}+\left(1-\frac{1}{n}\right)\frac{1}{2!}+\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\frac{1}{3!}+\cdots+\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{1}{k!}.\end{aligned}$$

If we let $n\to\infty$, while keeping k fixed we deduce

$$\begin{aligned}e&=\lim_{n\to\infty}e_n\\&\geqslant 1+\frac{1}{1!}+\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\frac{1}{2!}+\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\frac{1}{3!}+\cdots\end{aligned}$$

$$\cdots + \lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{1}{k!} = s_k.$$

Let us now estimate the error

$$\begin{aligned}\varepsilon_n = e - s_n &= \left(1+\frac{1}{1!}+\frac{1}{2!}+\cdots\right) - \left(1+\frac{1}{1!}+\frac{1}{2!}+\cdots+\frac{1}{n!}\right)\\ &= \frac{1}{(n+1)!}+\frac{1}{(n+2)!}+\cdots.\end{aligned}$$

Clearly $\varepsilon_n > 0$ and

$$\begin{aligned}\varepsilon_n &= \frac{1}{(n+1)!}\left(1+\frac{1}{n+2}+\frac{1}{(n+2)(n+3)}+\cdots\right)\\ &< \frac{1}{(n+1)!}\left(1+\frac{1}{n+2}+\frac{1}{(n+2)^2}+\frac{1}{(n+2)^3}+\cdots\right)\\ &= \frac{1}{(n+1)!}\cdot\frac{1}{1-\frac{1}{n+2}} = \frac{1}{(n+1)!}\cdot\frac{n+2}{n+1}.\end{aligned}$$

For example, if we let $n = 6$, then we deduce that

$$0 < \varepsilon_6 < \frac{8}{7\cdot 6!} = \frac{8}{7\cdot 720} \approx 0.0002\ldots.$$

This shows that s_5 computes e with a 2-decimal precision. We have

$$s_5 = 1+1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\frac{1}{120} = 2.71\ldots,$$

so that

$$e = 2.71\ldots$$

□

Given a series $\sum_{n=0}^{\infty} a_n$ and natural numbers $m < n$ we have

$$s_n - s_m = (a_{m+1} + a_{m+1} + \cdots + a_n) = \sum_{k=m+1}^{n} a_k.$$

Cauchy's Theorem 4.5.2 implies the following useful result.

Theorem 4.6.11 (Cauchy). *Let $\sum_{n=0}^{\infty} a_n$ be a series of real numbers. Then the following statements are equivalent.*

(i) *The series $\sum_{n=0}^{\infty} a_n$ is convergent.*

(ii)

$$\forall\varepsilon > 0\ \ \exists N = N(\varepsilon)\in\mathbb{N}\ \ \textit{such that}\ \ \forall n > m > N(\varepsilon)\ \ |a_{m+1}+\cdots+a_n| < \varepsilon.$$

□

Definition 4.6.12. The series of real numbers

$$\sum_{n\geqslant 0} a_n$$

is called *absolutely convergent* if the series of absolute values

$$\sum_{n\geqslant 0} |a_n|$$

is convergent. □

Theorem 4.6.13 (Absolute Convergence Theorem)**.** *If the series*

$$\sum_{n\geqslant 0} a_n$$

is absolutely convergent, then it is also convergent.

Proof. Since

$$\sum_{n\geqslant 0} |a_n|$$

is convergent, then Theorem 4.6.11 implies that

$$\forall \varepsilon > 0 \ \exists N = N(\varepsilon) \in \mathbb{N}: \ \ \forall n > m > N(\varepsilon): \ \ |a_{m+1}| + \cdots + |a_n| < \varepsilon.$$

On the other hand, we observe that

$$|a_{m+1} + \cdots + a_n| \leqslant |a_{m+1}| + \cdots + |a_n|$$

so that

$$\forall \varepsilon > 0 \ \exists N = N(\varepsilon) \in \mathbb{N}: \ \ \forall n > m > N(\varepsilon): \ \ |a_{m+1} + \cdots + a_n| < \varepsilon.$$

Invoking Theorem 4.6.11 again we deduce that the series $\sum_{n\geqslant 0} a_n$ is convergent as well. □

The Comparison Principle has the following immediate consequence.

Corollary 4.6.14 (Weierstrass M-test)**.** *Consider two series*

$$\sum_{n\geqslant 0} a_n, \ \ \sum_{n\geqslant 0} b_n$$

such that $b_n > 0$ for any n and there exists $N_0 \in \mathbb{N}$ such that

$$|a_n| < b_n, \ \ \forall n > N_0.$$

If the series $\sum_{n\geqslant 0} b_n$ is convergent, then the series $\sum_{n\geqslant 0} a_n$ converges absolutely. □

The Weierstrass M-test leads to a simple but very useful convergence test, called *the d'Alembert test* or the *ratio test*.

Corollary 4.6.15 (Ratio Test)**.** *Let*

$$\sum_{n\geqslant 0} a_n$$

be a series such that $a_n \neq 0$ $\forall n$ and the limit

$$L = \lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} \geqslant 0$$

exists, but it could also be infinite. Then the following hold.

(i) *If $L < 1$, then the series $\sum_{n\geqslant 0} a_n$ is absolutely convergent.*

(ii) *If $L > 1$ then the series $\sum_{n\geqslant 0} a_n$ is not convergent.*

Proof. (i) We know that $L < 1$. Choose r such that $L < r < 1$. Since

$$\frac{|a_{n+1}|}{|a_n|} \to L$$

there exists $N_0 \in \mathbb{N}$ such that

$$\frac{|a_{n+1}|}{|a_n|} \leqslant r,\ \ \forall n > N_0 \iff |a_{n+1}| \leqslant |a_n| r,\ \ \forall n > N_0.$$

We deduce that

$$|a_{N_0+1}| \leqslant |a_{N_0}| r,\ \ |a_{N_0+2}| \leqslant |a_{N_0+1}| r \leqslant |a_{N_0}| r^2,$$

and, inductively

$$|a_{N_0+k}| \leqslant r^k |a_{N_0}|,\ \ \forall k \in \mathbb{N}.$$

If we set $n = N_0 + k$ so that $k = n - N_0$, then we conclude from above that for any, $n > N_0$ we have

$$|a_n| \leqslant |a_{N_0}|\, r^{n-N_0} = \underbrace{\frac{|a_{N_0}|}{r^{N_0}}}_{=:C}\, r^n.$$

In other words

$$|a_n| \leqslant C r^n,\ \ \forall n \geqslant N_0.$$

The geometric series $\sum_{n\geqslant 0} b_n$, $b_n = Cr^n$, is convergent for $r \in (0,1)$ and we deduce from Weierstrass' Test that the series $\sum_{n\geqslant 0} |a_n|$ is also convergent.

(ii) We argue by contradiction and assume that the series $\sum_{n\geqslant 0} |a_n|$ is convergent. Since $L > 1$ we deduce that there exists a $N_0 \in \mathbb{N}$ such that

$$\frac{|a_{n+1}|}{|a_n|} > 1,\ \ \forall n > N_0 \iff |a_{n+1}| > |a_n|,\ \ \forall n > N_0.$$

Since the series $\sum_{n\geqslant 0}$ we deduce that $\lim_n a_n = 0$. On the other hand, $|a_n| > |a_{N_0}|$ for $n > N_0$ so that

$$0 = \lim_n |a_n| \geqslant |a_{N_0}| > 0.$$

This contradiction shows that the series $\sum_{n\geqslant 0} |a_n|$ cannot be convergent. □

Example 4.6.16. (a) Consider the series

$$\sum_{n\geqslant 1} (-1)^n \frac{n^2}{2^n}.$$

Then

$$\frac{|a_{n+1}|}{|a_n|} = \frac{\frac{(n+1)^2}{2^{n+1}}}{\frac{n^2}{2^n}} = \frac{1}{2}\left(\frac{n+1}{n}\right)^2 \to \frac{1}{2} \to \frac{1}{2} \text{ as } n \to \infty.$$

The Ratio Test implies that the series is absolutely convergent.

(b) Consider the series

$$\sum_{n\geqslant 1} \frac{1}{\sqrt{n(n+1)}}.$$

We observe that

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

Hence

$$\lim_{n\to\infty} \frac{\frac{1}{\sqrt{n(n+1)}}}{\frac{1}{n}} = 1$$

so that there exists $N_0 > 0$ such that

$$\frac{\frac{1}{\sqrt{n(n+1)}}}{\frac{1}{n}} > \frac{1}{2} \;\; \forall n > N_0,$$

i.e.,

$$\frac{1}{\sqrt{n(n+1)}} > \frac{1}{2n}, \quad \forall n > N_0.$$

In Example 4.6.7(a) we have shown that the series $\sum_{n\geqslant 1}\frac{1}{2n}$ is divergent. Invoking the comparison principle we deduce that the series $\sum_{n\geqslant 1}\frac{1}{\sqrt{n(n+1)}}$ is also divergent. □

Definition 4.6.17. A series is called *conditionally convergent* if it is convergent, but not absolutely convergent. □

Example 4.6.18. Consider the series

$$\sum_{n\geqslant 0}\frac{(-1)^n}{n+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots .$$

Example 4.6.7(a) shows that this series is not absolutely convergent. However, it is a convergent series. To see this observe first that

$$s_0 = 1, \;\; s_2 = s_0 - \frac{1}{2} + \frac{1}{3} = s_0 - \left(\frac{1}{2} - \frac{1}{3}\right) < s_0,$$

$$s_{2n+2} = s_{2n} - \frac{1}{(2n+2)} + \frac{1}{2n+3} = s_{2n} - \left(\frac{1}{2n+2} - \frac{1}{2n+3}\right) < s_{2n}.$$

Thus the subsequence $s_0, s_2, s_4, \dots,$ is decreasing.

Next observe that

$$s_1 = 1 - \frac{1}{2} > 0, \;\; s_3 = s_1 + \frac{1}{3} - \frac{1}{4} > s_1,$$

$$s_{2n+3} = s_{2n+1} + \frac{1}{2n+3} - \frac{1}{2n+4} > s_{2n+1}.$$

Thus, the subsequence $s_1, s_3, s_5, \dots,$ is increasing. Now observe that

$$s_{2n+2} - s_{2n+1} = \frac{1}{2n+3} > 0.$$

Hence

$$s_0 > s_{2n+2} > s_{2n+1} \geqslant s_1.$$

This proves that the increasing subsequence (s_{2n+1}) is also bounded above and the decreasing sequence (s_{2n+2}) is bounded below. Hence these two subsequences are convergent and since

$$\lim_n (s_{2n+2} - s_{2n+1}) = \lim_n \frac{1}{2n+3} = 0$$

we deduce that they converge to the same real number. This implies that the full sequence $(s_n)_{n\geqslant 0}$ converges to the same number; see Exercise 4.23.

The sum of this alternating series is $\ln 2$, but the proof of this fact is more involved an requires the full strength of the calculus techniques; see Example 9.6.10. □

4.7. Power series

Definition 4.7.1. A *power series* in the variable x and real coefficients $a_0, a_1, a_2, \dots$ is a series of the form

$$s(x) = a_0 + a_1 x + a_2 x^2 + \cdots .$$

The *domain of convergence* of the power series is the set of real numbers x such that the corresponding series $s(x)$ is convergent. □

Example 4.7.2. (a) The geometric series

$$1 + x + x^2 + \cdots$$

is a power series. It converges for $|x| < 1$ and diverges for $|x| \geqslant 1$.

(b) Consider the power series

$$\sum_{n\geqslant 1} \frac{x^n}{n} = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots .$$

Note that

$$\left| \frac{\frac{x^{n+1}}{n+1}}{\frac{x^n}{n}} \right| = |x| \frac{n}{n+1} \to |x| \;\text{ as } n \to \infty.$$

The Ratio Test shows that this series converges absolutely for $|x| < 1$ and diverges for $|x| > 1$.

When $x = 1$ the series becomes the harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

which is divergent. When $x = -1$ the series becomes the alternating series

$$-1 + \frac{1}{2} - \frac{1}{3} + \cdots = -\sum_{n\geqslant 1} \frac{(-1)^n}{n}.$$

As explained in Example 4.6.18, this series is convergent.

(c) Consider the power series

$$\sum_{n\geqslant 0}\frac{x^n}{n!}=1+\frac{x}{1!}+\frac{x^2}{2!}+\cdots .$$

Note that

$$\left|\frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}}\right|=\frac{|x|}{n+1}\to 0 \;\text{ as } n\to\infty.$$

The Ratio Test implies that this series converges absolutely for any $x\in\mathbb{R}$. □

Proposition 4.7.3. *Consider a power series in the variable x with real coefficients*

$$s(x)=a_0+a_1x+a_2x^2+\cdots .$$

Suppose that the nonzero real number x_0 is in the domain of convergence of the series. Then for any real number x such that $|x|<|x_0|$ the series $s(x)$ is absolutely *convergent.*

Proof. Since the series

$$a_0+a_1x_0+a_2x_0^2+\cdots$$

is convergent, the sequence $(a_nx_0^n)$ converges to zero. In particular, this sequence is bounded and thus there exists a positive constant C such that

$$|a_nx_0^n|<C,\;\;\forall n=0,1,2,\dots .$$

We set

$$r:=\frac{|x|}{|x_0|}$$

and we observe that $0\leqslant r<1$. Next we notice that

$$|a_nx^n|=|a_nx_0^n|\frac{|x|^n}{|x_0|^n}=|a_nx_0^n|r^n<Cr^n,\;\;\forall n.$$

Since $0\leqslant r<1$ we deduce that the positive geometric series

$$C+Cr+Cr^2+\cdots$$

is convergent. The comparison principle then implies that the series

$$|a_0|+|a_1x|+|a_2x^2|+\cdots$$

is also convergent. □

The above result has a very important consequence whose proof is left to you as an exercise.

Corollary 4.7.4. *Consider a power series in the variable x and real coefficients*

$$s(x)=a_0+a_1x+a_2x^2+\cdots .$$

We denote by D the domain of convergence of the series. We set

$$R := \begin{cases} \sup D, & \textit{if } D \textit{ is bounded above,} \\ \infty, & \textit{if } D \textit{ is not bounded above.} \end{cases} \tag{4.7.1}$$

Then the following hold.

(i) $R \geqslant 0$.

(ii) *If x is a real number such that $|x| < R$, then the series $s(x)$ is absolutely convergent.*

(iii) *If x is a real number such that $|x| > R$, then the series $s(x)$ is divergent.*

□

Definition 4.7.5. The quantity R defined in (4.7.1) is called the *radius of convergence* of the power series $s(x)$. □

Example 4.7.6. The power series in Example 4.7.2(a),(b) have radii of convergence 1, while the power series in Example 4.7.2(c) has radius of convergence ∞. □

4.8. Some fundamental sequences and series

$$\lim_{n\to\infty} \frac{C}{n} = 0, \;\; \forall C > 0.$$

$$\lim_{n\to\infty} Cn = \infty, \;\; \forall C > 0.$$

$$\lim_{n\to\infty} r^n = \lim_{n\to\infty} \frac{1}{a^n} = 0, \;\; \forall r \in (0,1), \;\; \forall a > 1.$$

$$\lim_{n\to\infty} \frac{n}{r^n} = 0, \;\; \forall r > 1.$$

$$\lim_{n\to\infty} r^{\frac{1}{n}} = 1, \;\; \forall r > 0.$$

$$\lim_{n\to\infty} n^{\frac{1}{n}} = 1.$$

$$\lim_{n\to\infty} \frac{r^n}{n!} = 0, \;\; \forall r \in \mathbb{R}.$$

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e.$$

$$1 + r + r^2 + \cdots + r^n + \cdots = \frac{1}{1-r}, \;\; \forall |r| < 1.$$

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots = e.$$

$$\sum_{n\geqslant 1} \frac{1}{n^s} = \begin{cases} \text{convergent}, & s > 1 \\ \text{divergent}, & s \leqslant 1. \end{cases}$$

4.9. Exercises

Exercise 4.1. Prove, *using the definition*, the following equalities.

$$\lim_{n\to\infty} \frac{n}{n^2+1} = 0, \tag{a}$$

$$\lim_{n\to\infty} \frac{3n+1}{2n+5} = \frac{3}{2}, \tag{b}$$

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0. \tag{c}$$

Exercise 4.2. Prove Proposition 4.2.6. □

Exercise 4.3. Prove Proposition 4.2.7. □

Exercise 4.4. Let $(x_n)_{n\geqslant 0}$ be a sequence of real numbers and $x \in \mathbb{R}$. Consider the following statements.

(i) $\forall \varepsilon > 0$, $\exists N \in \mathbb{N}$ such that, $n > N \Rightarrow |x_n - x| < \varepsilon$.

(ii) $\exists N \in \mathbb{N}$ such that, $\forall \varepsilon > 0$, $n > N \Rightarrow |x_n - x| < \varepsilon$.

Prove that (ii) $\Rightarrow$ (i) and construct an example of sequence $(x_n)_{n\geqslant 1}$ and real number x satisfying (i) but not (ii). □

Exercise 4.5. (a) Prove that for any real numbers a, b we have

$$\big|\, |a| - |b| \,\big| \leqslant |a - b|.$$

(b) Let $(x_n)_{n\geqslant 0}$ be a sequence of real numbers that converges to $x \in \mathbb{R}$. Prove that

$$\lim_{n\to\infty} |x_n| = |x|.$$ □

Exercise 4.6. Compute

$$\lim_{n\to\infty} \left(\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} \right).$$

Hint. Observe that

$$\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} = \frac{1}{2}\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{n-1}} \right).$$

At this point you might want to use Exercise 3.8. □

Exercise 4.7. Compute

$$\lim_{n\to\infty} \frac{2^1 + 2^3 + 2^5 + \cdots + 2^{2n+1}}{2^{2n+3}}.$$

Hint. Use Exercise 3.8. □

Exercise 4.8. Let $X \subset \mathbb{R}$ be a bounded above set of real numbers. Denote by x^* the supremum of X. (The existence of the least upper bound of X is guaranteed by the Completeness Axiom.) Prove that there exists a sequence of real numbers $(x_n)_{n\in\mathbb{N}}$ satisfying the following properties.

(i) $x_n \in X$, $\forall n \in \mathbb{N}$.

(ii) $\lim_{n\to\infty} x_n = x^*$.

Hint. Use Proposition 2.3.7 and Corollary 4.2.9. □

Exercise 4.9. Prove the equality (4.3.3). □

Exercise 4.10. Let $0 < a < b$. Compute

$$\lim_{n\to\infty} \frac{a^{n+1} + b^{n+1}}{a^n + b^n}.$$ □

Exercise 4.11. (a) Let (a_n) be a sequence of positive real numbers such that $\lim_n a_n = 1$. Prove that

$$\lim_n \sqrt{a_n} = 1.$$

(b) Compute

$$\lim_{n\to\infty} \sqrt{n}\big(\sqrt{n+1} - \sqrt{n}\,\big).$$

Hint. Prove first that

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}, \quad \forall x > 0.$$

□

Exercise 4.12. Prove that if $a > 0$, then

$$\lim_{n\to\infty} a^{\frac{1}{n}} = 1.$$

Hint. Consider first the case $a > 1$. Write $a^{\frac{1}{n}} = 1 + \varepsilon_n$ and then use Bernoulli's inequality. Show that the case $a < 1$ follows from the case $a > 1$. □

Exercise 4.13. Prove that for any real number x there exists an *increasing* sequence of *rational* numbers that converges to x and also a *decreasing* sequence of *rational* numbers that converges to x.

Hint. Use Proposition 3.4.4. □

Exercise 4.14. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers that converges to a *positive* number a. Prove that

$$\exists c > 0 \ \text{ such that } \ \forall n \in \mathbb{N} \ \ a_n > c.$$

Hint. Argue by contradiction. □

Exercise 4.15. Let $k \in \mathbb{N}$ and suppose that $(a_n)_{n\in\mathbb{N}}$ is a sequence of positive numbers that converges to a *positive* number a.

(a) Using Exercise 4.14 prove that there exists $r > 0$ such that $a_n > r$, $\forall n$, so that $a_n^{\frac{1}{k}} > r^{\frac{1}{k}}$, $\forall n$.

(b) Prove that there exists a constant $M > 0$ such that

$$\left|a_n^{\frac{1}{k}} - a^{\frac{1}{k}}\right| \leqslant M|a_n - a|, \quad \forall n \in \mathbb{N}.$$

Hint. Set $b_n := a_n^{\frac{1}{k}}$, $b := a^{\frac{1}{k}}$ and use the equality (3.4.3) to deduce.

$$a_n - a = b_n^k - b^k = (b_n - b)(b_n^{k-1} + b_n^{k-2}b + \cdots + b_n b^{k-2} + b^{k-1})$$

which implies

$$|b_n - b| = \frac{|a_n - a|}{b_n^{k-1} + b_n^{k-2}b + \cdots + b_n b^{k-2} + b^{k-1}}.$$

Now use part (a).

(c) Show that

$$\lim_n a_n^{\frac{1}{k}} = a^{\frac{1}{k}}.$$

(d) Show that if $r \in \mathbb{Q}$, then

$$\lim_n a_n^r = a^r. \qquad \square$$

Exercise 4.16. Let $r > 1$ and $k \in \mathbb{N}$. Prove that

$$\lim_{n\to\infty} r^n = \infty.$$

and

$$\lim_{n\to\infty} \frac{n^k}{r^n} = 0.$$

Hint. Let $a = r^{\frac{1}{k}}$. Then

$$\frac{n^k}{r^n} = \left(\frac{n}{a^n}\right)^k.$$

□

Exercise 4.17. Compute

$$\lim_{n\to\infty} \left(1 + \frac{1}{2n}\right)^n. \qquad \square$$

Exercise 4.18. (a) Using Example 4.4.2 as inspiration prove that the sequence

$$x_n = \left(1 + \frac{1}{n}\right)^n$$

is increasing.

(b) Prove that the Euler number e satisfies the inequalities

$$\left(1 + \frac{1}{n}\right)^n < e < \left(1 + \frac{1}{n}\right)^{n+1}, \quad \forall n \in \mathbb{N}.$$

Deduce from the above inequalities that $2 < e < 3$. □

Exercise 4.19. Consider the sequence (x_n) defined by the recurrence

$$x_1 = \sqrt{2}, \;\; x_{n+1} = \sqrt{2 + x_n}, \;\; \forall n \in \mathbb{N}.$$

Thus

$$x_2 = \sqrt{2 + \sqrt{2}}, \;\; x_3 = \sqrt{2 + \sqrt{2 + \sqrt{2}}}, \;\; x_4 = \sqrt{2 + \sqrt{2 + \sqrt{2 + \sqrt{2}}}}, \dots .$$

(a) Prove by induction that the sequence (x_n) is increasing.

(b) Prove by induction that $x_n < \sqrt{2} + 1$, $\forall n \in \mathbb{N}$.

(c) Find $\lim_{n\to\infty} x_n$.

Hint. Consider the function $f : (0, \infty) \to (0, \infty)$, $f(x) = \sqrt{2 + x}$ and prove that

$$0 < x < y \Rightarrow f(x) < f(y) \;\text{ and }\; x > 0 \wedge x = f(x) \Longleftrightarrow x = 2.$$

□

Exercise 4.20. Fix $a > 0$, $a \neq 1$ and define $f : (0, \infty) \to (0, \infty)$ by

$$f(x) = \frac{1}{2}\left(x + \frac{a}{x}\right) = \frac{x^2 + a}{2x}.$$

Consider the sequence of positive real numbers $(x_n)_{n \geqslant 1}$ defined by the recurrence

$$x_1 = 1, \;\; x_{n+1} = f(x_n), \; \forall n \in \mathbb{N}.$$

Use the strategy employed in Example 4.4.4 to show that

$$\lim_{n\to\infty} x_n = \sqrt{a}.$$

□

Exercise 4.21 (Gauss)**.** Let a_0, b_0 be two real numbers such that

$$0 < a_0 < b_0.$$

Define inductively

$$a_1 := \sqrt{a_0 b_0}, \;\; b_1 = \frac{a_0 + b_0}{2},$$

$$a_{n+1} = \sqrt{a_n b_n}, \;\; b_{n+1} = \frac{a_n + b_n}{2}.$$

(a) Prove by induction that

$$a_1 \leqslant a_2 \leqslant \cdots \leqslant a_n \leqslant b_n \leqslant \cdots \leqslant b_2 \leqslant b_1.$$

(b) Prove that the sequences (a_n) and (b_n) are convergent and

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

Hint: For part (a) use Exercise 3.6. For part (b) use Weierstrass' Theorem on the convergence of bounded monotone sequences, Theorem 4.4.1.

□

Exercise 4.22. Establish the convergence or divergence of the sequence

$$a_n = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}, \quad n \in \mathbb{N}. \qquad \square$$

Exercise 4.23. Let (a_n) be a sequence of real numbers. For each $n \in \mathbb{N}$ we set

$$b_n := a_{2n-1}, \quad c_n := a_{2n}.$$

Prove that the following statements are equivalent.

(i) The sequence (a_n) is convergent and its limit is $a \in \mathbb{R}$.

(ii) The subsequences $(b_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ converge to the same limit a.

□

Exercise 4.24. Suppose $(a_n)_{n\in\mathbb{N}}$ is a *contractive* sequence of real numbers, i.e., there exists $r \in (0,1)$ such that

$$|a_n - a_{n+1}| < r|a_n - a_{n-1}|, \quad \forall n \in \mathbb{N}, n \geqslant 2.$$

Prove that the sequence $(a_n)_{n\in\mathbb{N}}$ is convergent.

Hint. Set $x_1 := a_1, \ x_2 := a_2 - a_1, \ x_3 := a_3 - a_2, \dots$. Observe that $x_1 + x_2 + \cdots + x_n = a_n$, $\forall n \in \mathbb{N}$ so that the sequence $(a_n)_{n\in\mathbb{N}}$ is the sequence of partial sums of the series

$$x_1 + x_2 + x_3 + \cdots$$

Use the Comparison Principle to show that this series is absolutely convergent. □

Exercise 4.25. Consider the sequence of positive real numbers $(x_n)_{n\geqslant 1}$ defined by the recurrence

$$x_1 = 1, \quad x_{n+1} = 1 + \frac{1}{x_n}, \quad \forall n \in \mathbb{N}.$$

Thus

$$x_2 = 1+1, \quad x_3 = 1 + \frac{1}{1+1} = \frac{3}{2}, \quad x_4 = 1 + \frac{1}{1+\frac{1}{1+1}} = 1 + \frac{2}{3} = \frac{5}{3},$$

$$x_5 = 1 + \cfrac{1}{1 + \cfrac{1}{1+\frac{1}{1+1}}}, \quad x_6 = 1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\frac{1}{1+1}}}} \cdots$$

(a) Prove that

$$x_1 < x_3 < \cdots < x_{2n+1} < x_{2n+2} < x_{2n} < \cdots < x_2, \quad \forall n \geqslant 1.$$

(b) Prove that for $n \geqslant 3$ we have

$$|x_{n+1} - x_n| = \frac{|x_n - x_{n-1}|}{x_n x_{n-1}} \leqslant \frac{4}{9}|x_n - x_{n-1}|.$$

(c) Conclude that the sequence (x_n) is convergent and find its limit. **Hint.** Use Exercise 4.24. □

Exercise 4.26. If $a_1 < a_2$

$$a_{n+2} = \frac{1}{2}\big(\,a_{n+1} + a_n\,\big), \quad \forall n \in \mathbb{N}$$

show that the sequence $(a_n)_{n\in\mathbb{N}}$ is convergent.

Hint. Use Exercise 4.24. □

Exercise 4.27. Consider a sequence of positive numbers $(x_n)_{n\geqslant 1}$ satisfying the recurrence relation

$$x_{n+1} = \frac{1}{2+x_n}, \quad \forall n \in \mathbb{N}.$$

Show that $(x_n)_{n\in\mathbb{N}}$ is a contractive sequence (Exercise 4.24) and then compute its limit. □

Exercise 4.28. Find all the limit points (see Definition 4.4.9) of the sequence

$$a_n = (-1)^n \frac{n-1}{n}.$$ □

Exercise 4.29. Let $(a_n)_{n\in\mathbb{N}}$ be a bounded sequence of real numbers, i.e.,

$$\exists C \in \mathbb{R} : \;\; |a_n| \leqslant C, \;\; \forall n.$$

For any $k \in \mathbb{N}$ we set

$$b_k := \sup\{a_n;\;\; n \geqslant k\,\}.$$

(a) Show that the sequence $(b_k)_{k\in\mathbb{N}}$ is *nonincreasing* and conclude that it is convergent. Denote by b its limit.

(b) Show that b is a limit point of the sequence $(a_n)_{n\in\mathbb{N}}$, i.e., there exists a subsequence $(a_{n_k})_{k\geqslant 1}$ of $(a_n)_{n\geqslant 1}$ such that

$$\lim_{k\to\infty} a_{n_k} = b.$$

(c) Show that if α is a limit point of the sequence (a_n), then $\alpha \leqslant b$.

The number b is called the *superior limit* of the sequence (a_n) and it is denoted by $\limsup_n a_n$. The above exercise shows that the superior limit is the largest limit point of a bounded sequence. □

Exercise 4.30. Prove Proposition 4.6.3. □

Exercise 4.31. Prove that if $\sum_{n\geqslant 0} a_n$ and $\sum_{n\geqslant 0} b_n$ are convergent series of real numbers and $\alpha, \beta \in \mathbb{R}$, then the series $\sum_{n\geqslant 0}(\alpha a_n + \beta b_n)$ is convergent and

$$\sum_{n\geqslant 0}(\alpha a_n + \beta b_n) = \alpha \sum_{n\geqslant 0} a_n + \beta \sum_{n\geqslant 0} b_n.$$ □

Exercise 4.32. Can you give an example of convergent series $\sum_{n\geqslant 0} a_n$ and a divergent series $\sum_{n\geqslant 0} b_n$ such that $\sum_{n\geqslant 0}(a_n + b_n)$ is convergent? Explain. □

Exercise 4.33. Prove Corollary 4.6.9. □

Exercise 4.34. Consider the sequence

$$a_n = \frac{n^3+2n^2+2n+4}{n^5+n^4+7n^2+1}, \quad n \geqslant 0.$$

Prove that the series

$$\sum_{n\geqslant 0} a_n$$

is absolutely convergent.

Hint. Example 4.6.7(b) and Corollary 4.6.9. □

Exercise 4.35 (Leibniz)**.** Suppose that (a_n) is a decreasing sequence of positive real numbers such that

$$\lim_{n\to\infty} a_n = 0.$$

Prove that the series

$$\sum_{n\geqslant 0} (-1)^n a_n$$

is convergent.

Hint. Imitate the strategy in Example 4.6.18. □

Exercise 4.36 (Cauchy)**.** Suppose that $(a_n)_{n\geqslant 0}$ is a decreasing sequence of positive numbers that converges to 0. Prove that the series

$$\sum_{n\geqslant 0} a_n$$

converges if and only if the series

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + 8a_8 + 16a_{16} + \cdots$$

converges.

Hint. Imitate the strategy employed in Example 4.6.7. □

Exercise 4.37. Prove that for any $x \in [0,1)$ there exists a unique sequence $\big(\epsilon_n\big)_{n\geqslant 1} = \big(\epsilon_n(x)\big)_{n\geqslant 1}$ with the following properties.

(i) $\epsilon_n \in \{0,1\}$, $\forall n$. Set

(ii) For any $n \in \mathbb{N}$

$$\sum_{k=1}^{n} \frac{\epsilon_k}{2^k} \leqslant x < \frac{1}{2^n} + \sum_{k=1}^{n} \frac{\epsilon_k}{2^k}.$$

Determine the sequence $\big(\epsilon_n(x)\big)_{n\geqslant 1}$ when $x = \frac{1}{3}$. □

Exercise 4.38. We consider the power series

$$\sum_{n\geqslant 0} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots .$$

Suppose that there exists $C > 0$ such that $|a_n| \leqslant C$, $\forall n$. Show that the radius of convergence of the series

$$\sum_{n\geqslant 0} a_n x^n$$

is $\geqslant 1$. □

Exercise 4.39. Suppose that $(a_n)_{n\in\mathbb{N}}$ is a sequence of integers such that $0 \leqslant a_n \leqslant 9$ for any $n \in \mathbb{N}$, i.e.,

$$a_n \in \{0, 1, 2, \dots, 9\}, \quad \forall n \in \mathbb{N}.$$

Show that the series

$$\sum_{n\geqslant 1} a_n 10^{-n} = \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots$$

is convergent.

(b) Compute the sum of the above series in the two special special cases

$$a_n = 7, \quad \forall n \in \mathbb{N},$$

and

$$a_n = \begin{cases} 1, & n \text{ is odd} \\ 2, & n, \text{ is even.} \end{cases}$$

In each case, express the sum in decimal form.

(c) Prove that for any $x \in [0, 1]$ there exists a sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ such that

$$a_n \in \{0, 1, 2, \dots, 9\}, \quad \forall n \in \mathbb{N},$$

and

$$x = \sum_{n\geqslant 1} a_n 10^{-n}.$$

□

Exercise 4.40. Prove Corollary 4.7.4. □

Exercise 4.41. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. To any real numbers a, b such that $a < b$ we associate the sequences $(S_k(a,b))_{k\in\mathbb{N}}$ and $(T_k(a,b))_{k\in\mathbb{N}}$ in $\mathbb{N} \cup \{\infty\}$ defined inductively as follows

$$S_1(a,b) := \inf\{n \geqslant 1;\;\; x_n \leqslant a\}, \;\; T_1(a,b) := \inf\{\, n \geqslant S_1(a,b);\;\; x_n \geqslant b\,\},$$

$$S_{k+1}(a,b) := \inf\{n \geqslant T_k(a,b);\;\; x_n \leqslant a\,\}, \;\; T_{k+1}(a,b) := \inf\{\, n \geqslant S_k(a,b);\;\; x_n \geqslant b\,\},$$

where we set $\inf \varnothing = \infty$. We set

$$U_n(a,b) := \#\{k \leqslant n;\;\; T_k(a,b) \leqslant n\,\}.$$

(a) Prove that for any $a, b \in \mathbb{R}$, $a < b$, the sequence $(U_n(a,b))_{n \in bN}$ is nondecreasing. Set

$$U_\infty(a,b) := \lim_{n\to\infty} U_n(a,b).$$

(b) Prove that the following statements are equivalent.

(i) The sequence (x_n) has a limit as $n \to \infty$.

(ii) For any $a, b \in \mathbb{Q}$ such that a, b we have $U_\infty(a,b) < \infty$.

□

Exercise 4.42. Consider two series of real numbers $\sum_{n\geqslant 0} a_n$ and $\sum_{n\geqslant 0} b_n$. For each nonnegative integer n define

$$c_n := a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0 = \sum_{k=0}^{n} a_k b_{n-k}$$

Prove that the if the series $\sum_{n\geqslant 0} a_n$ and $\sum_{n\geqslant 0} b_n$ are *absolutely* convergent, then the series

$$\sum_{n\geqslant 0} c_n$$

is *absolutely* convergent and its sum is the product of the sums of the series $\sum_{n\geqslant 0} a_n$ and $\sum_{n\geqslant 0} b_n$, i.e.,

$$\lim_{n\to\infty} \sum_{k=0}^{n} c_n = \left(\lim_{n\to\infty} \sum_{j=0}^{n} a_j \right) \cdot \left(\lim_{n\to\infty} \sum_{k=0}^{n} b_k \right).$$

The series $\sum_{n\geqslant 0} c_n$ constructed above is called the *Cauchy product* of the series $\sum_{n\geqslant 0} a_n$ and $\sum_{n\geqslant 0} b_n$.

Hint: Consider first the special case $a_n, b_n \geqslant 0$, $\forall n$. Set

$$A_n := \sum_{j=0}^{n} a_j, \;\; B_n := \sum_{k=0}^{n} b_k, \;\; C_n = \sum_{\ell=0}^{n} c_\ell.$$

Prove that

$$\lim_{n\to\infty} (C_n - A_n B_n) = 0.$$

□

Exercise 4.43. Let $(a_n)_{n\geqslant 0}$ and $(b_n)_{n\geqslant 0}$ be two sequences of real numbers. For any nonnegative integer n we set

$$B_n := b_0 + b_1 + \cdots + b_n, \;\; C_n = a_0 b_0 + a_1 b_1 + \cdots + a_n b_n.$$

(a)(Abel's trick) Show that, for any $n \in \mathbb{N}$, we have

$$C_n = a_n B_n - \sum_{k=1}^{n-1} (a_{k+1} - a_k) B_k. \tag{4.9.1}$$

(b) Show that if the series

$$\sum_{n\geqslant 0} b_n$$

is convergent and the sequence $(a_n)_{n\geqslant 0}$ is monotone and bounded, then the series

$$\sum_{n\geqslant 0} a_n b_n$$

is convergent. □

Exercise 4.44. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. Prove that the following are equivalent.

(i) The series

$$\sum_{n\in\mathbb{N}} a_n$$

is absolutely convergent.

(ii) For any *injection* $\varphi : \mathbb{N} \to \mathbb{N}$ the series

$$\sum_{n\in\mathbb{N}} a_{\varphi(n)}$$

is convergent.

(iii) For any *bijection* $\sigma : \mathbb{N} \to \mathbb{N}$ the series

$$\sum_{n\in\mathbb{N}} a_{\sigma(n)}$$

is convergent and its sum is independent of σ.

□

4.10. Exercises for extra credit

Exercise* 4.1. Fix rational numbers a, b such that $1 < a < b$.

(a) Prove that

$$\lim_{n\to\infty} \frac{(2n)^b}{(2n+1)^a} = \infty.$$

(b) Prove that the series

$$\frac{1}{1^a} + \frac{1}{2^b} + \frac{1}{3^a} + \frac{1}{4^b} + \cdots$$

is convergent. □

Exercise* 4.2. Let (a_n) be a convergent sequence of real numbers. Form the new sequence (c_n) defined by the rule

$$c_n := \frac{a_1 + \cdots + a_n}{n}$$

Show that (c_n) is convergent and

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} a_n.$$ □

Exercise* 4.3. Let the two given sequences

$$a_0, a_1, a_2, \ldots,$$

$$b_0, b_1, b_2, \ldots$$

satisfy the conditions

$$b_n > 0, \;\; \forall n \geqslant 0, \tag{4.10.1a}$$

$$b_0 + b_1 + b_2 + \cdots + b_n + \cdots = \infty, \tag{4.10.1b}$$

$$\lim_{n\to\infty} \frac{a_n}{b_n} = s. \tag{4.10.1c}$$

Prove that

$$\lim_{n\to\infty} \frac{a_0 + a_1 + \cdots + a_n}{b_0 + b_1 + \cdots + b_n} = s. \qquad \square$$

Exercise* 4.4. Suppose that $(p_n)_{n\geqslant 1}$ is a sequence of *positive* real numbers, and $(x_n)_{n\geqslant 1}$ is a sequence of real numbers. For $n \in \mathbb{N}$ we set

$$b_n := p_1 + \cdots + p_n, \;\; s_n := x_1 + \cdots + x_n.$$

Suppose that

$$\lim_{n\to\infty} b_n = \infty.$$

Prove that if the series

$$\sum_{n\geqslant 1} \frac{x_n}{b_n}$$

is convergent, then

$$\lim_{n\to\infty} \frac{s_n}{b_n} = 0. \qquad \square$$

Exercise* 4.5. Suppose that the sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ satisfies the subadditivity condition

$$a_{m+n} \leqslant a_m + a_n, \;\; \forall m, n \in \mathbb{N}.$$

Prove that

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\in\mathbb{N}} \frac{a_n}{n}. \qquad \square$$

Exercise* 4.6. Let $(x_n)_{n\geqslant 0}$ be a sequence of nonzero real numbers such that

$$x_n^2 - x_{n+1}x_{n-1} = 1, \;\; \forall n \in \mathbb{N}.$$

Prove that there exists $a \in \mathbb{R}$ such that

$$x_{n+1} = ax_n - x_{n-1}, \;\; \forall n \in \mathbb{N}. \qquad \square$$

Exercise* 4.7. Suppose that a sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ satisfies

$$0 < a_n < a_{2n} + a_{2n+1}, \;\; \forall n \in \mathbb{N}.$$

Prove that the series $\sum_{n\geqslant 1} a_n$ is divergent. □

Exercise* 4.8. Suppose that $(x_n)_{n\in\mathbb{N}}$ is a sequence of positive real numbers such that the series

$$\sum_{n\in\mathbb{N}} x_n$$

is convergent and its sum is S. Prove that for any bijection $\varphi : \mathbb{N} \to \mathbb{N}$ the series

$$\sum_{n\in\mathbb{N}} x_{\varphi(n)}$$

is also convergent and its sum is also S. □

Exercise* 4.9. Suppose that the series of real numbers

$$\sum_{n\in\mathbb{N}} x_n$$

is convergent, but *not absolutely convergent.* Prove that for any real number S there exists a bijection $\varphi : \mathbb{N} \to \mathbb{N}$ such that the series

$$\sum_{n\in\mathbb{N}} x_{\varphi(n)}$$

is convergent and its sum is S. □

Exercise* 4.10. Suppose that $(a_n)_{n\geqslant 1}$ is a decreasing sequence of positive real numbers that converges to 0 and satisfies the inequalities

$$a_n \leqslant a_{n+1} + a_{n^2}, \quad \forall n \geqslant 1.$$

Prove that the series

$$\sum_{n\geqslant 1} a_n$$

is divergent. □

Chapter 5

Limits of functions

5.1. Definition and basic properties

Let X be a nonempty subset of $\mathbb{R}$. A real number c is called a *cluster point* of X if there exists a sequence (x_n) of real numbers with the following properties.

(i) $x_n \in X$, $\forall n \in \mathbb{N}$.

(ii) $x_n \neq c$, $\forall n \in \mathbb{N}$.

(iii) $\lim_n x_n = c$.

Example 5.1.1. (a) If $A = (0, 1)$, then 0 and 1 are cluster points of A, although they are not in A. Indeed, the sequence $a_n = \frac{1}{n+1}$, $n \in \mathbb{N}$ consists of elements of $(0, 1)$ and $a_n \to 0$. Similarly, the sequence $b_n = 1 - \frac{1}{n+1}$ consists of points in $(0, 1)$ and $b_n \to 1$. Observe that every point in $(0, 1)$ is also a cluster point of $(0, 1)$.

(b) Any real number is a cluster point of the set $\mathbb{Q}$ of rational numbers. □

Definition 5.1.2. Let $X \subset \mathbb{R}$. Suppose that c is a cluster point of X and $f : X \to \mathbb{R}$ is a real valued function defined on X. We say that the limit of f at c is the real number A, and we write this

$$\lim_{x \to c} f(x) = A,$$

if the following holds:

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 : \;\; \forall x \in X : \;\; 0 < |x - c| < \delta \Rightarrow |f(x) - A| < \varepsilon. \tag{5.1.1}$$

□

An alternate viewpoint. Recall that a *neighborhood* of a point a is an open interval that contains a inside. For example, the open interval $(0, 3)$ is a neighborhood of 1. We denote by $\mathcal{N}_a$ the collection of all neighborhoods of a. Thus, a statement of the form $U \in \mathcal{N}_a$ signifies that U is an open interval that contains a. A *symmetric* *neighborhood* of a is a neighborhood of the form $(a-\delta, a+\delta)$, where δ is some positive number. Observe that

$$x \in (a-\delta, a+\delta) \Longleftrightarrow \operatorname{dist}(a,x) < \delta \Longleftrightarrow |x-a| < \delta.$$

Thus, to describe a symmetric neighborhood of a, it suffices to indicate a positive real number δ, and then the symmetric neighborhood is described by the condition $\operatorname{dist}(x,a) < \delta$. We denote by $\mathcal{S}\mathcal{N}_a$ the collection of symmetric neighborhoods of a. Clearly, any symmetric neighborhood of a is also a neighborhood of a so that

$$\mathcal{S}\mathcal{N}_a \subset \mathcal{N}_a.$$

A *deleted neighborhood* of a is a set obtained from a neighborhood of a by removing the point a. For example

$$(0,2)\backslash\{1\} = (0,1) \cup (1,2)$$

is a deleted neighborhood of 1. We denote by $\mathcal{N}^*_a$ the collection of all deleted neighborhoods of a. A *symmetric deleted neighborhood* of a is a deleted neighborhood of the form

$$(a-r, a+r)\backslash\{a\} = (a-r, a) \cup (a, a+r).$$

We denote by $\mathcal{S}\mathcal{N}^*_a$ the collection of deleted symmetric neighborhoods of a. Clearly

$$\mathcal{S}\mathcal{N}^*_a \subset \mathcal{S}^*_a.$$

Observe that the definition (5.1.1) is equivalent with the following statement

$$\forall U \in \mathcal{S}\mathcal{N}_a \;\; \exists V \in \mathcal{S}\mathcal{N}^*_c \;\; \forall x \in X: \;\; x \in V \Rightarrow f(x) \in U. \tag{5.1.2}$$

Indeed, we can rephrase (5.1.1) in the following equivalent way: for any symmetric neighborhood U of A of the form $(A-\varepsilon, A+\varepsilon)$, there exists a deleted symmetric neighborhood V of c of the form $(c-\delta, c+\delta)\backslash\{c\}$ such that for any $x \in V$ we have $f(x) \in U$. That is precisely the content of (5.1.2).

The proof of the next result is left to you as an exercise.

Proposition 5.1.3. *Let $f: X \to \mathbb{R}$ be a function defined on a set $X \subset \mathbb{R}$ and c a cluster point of X. Then the following statements are equivalent.*

(i) $\lim_{x\to c} f(x) = A$, *i.e., f satisfies (5.1.1) or (5.1.2).*

(ii)
$$\forall U \in \mathcal{N}_A, \;\; \exists V \in \mathcal{N}^*_c \;\; \textit{such that} \;\; \forall x \in X: x \in V \Rightarrow f(x) \in U. \tag{5.1.3}$$

□

The following very useful result reduces the study of limits of functions to the study of a concept we are already familiar with, namely the concept of limits of sequences.

Theorem 5.1.4. *Let c be a cluster point of the set $X \subset \mathbb{R}$ and $f: X \to \mathbb{R}$ a real valued function on X. The following statements are equivalent.*

(i) $\lim_{x\to c} f(x) = A \in \mathbb{R}$.

(ii) *For any sequence $(x_n)_{n\in\mathbb{N}}$ in $X\backslash\{c\}$ such that $x_n \to c$, we have $\lim_n f(x_n) = A$.*

Proof. (i) $\Rightarrow$ (ii). We know that $\lim_{x\to c} f(x) = A$ and we have to show that if (x_n) is a sequence in $X\backslash\{c\}$ that converges to c then the sequence $\big(f(x_n)\big)$ converges to A. In other words, given the above sequence (x_n) we have to show that

$$\forall \varepsilon > 0 \;\; \exists N = N(\varepsilon) \;\; \forall n \in \mathbb{N}: \;\; n > N(\varepsilon) \Rightarrow |f(x_n) - A| < \varepsilon.$$

Let $\varepsilon > 0$. We deduce from (5.1.1) that there exists $\delta(\varepsilon) > 0$ such that

$$\forall x \in X: \;\; 0 < |x-c| < \delta \Rightarrow |f(x) - A| < \varepsilon. \tag{5.1.4}$$

Since $x_n \to c$, there exists $N = N(\delta(\varepsilon))$ such that

$$0 < |x_n - c| < \delta, \ \ \forall n > N.$$

Using (5.1.4) we deduce that for any $n > N(\delta(\varepsilon))$ we have $|f(x_n) - A| < \varepsilon$. This proves the implication (i) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (i) We know that for any sequence (x_n) in $X\backslash\{c\}$ that converges to c, the sequence $(\, f(x_n)\,)$ converges to A and we have to prove (5.1.1), i.e.,

$$\forall \varepsilon > 0 \ \ \exists \delta = \delta(\varepsilon) > 0 : \ \ \forall x \in X : \ \ 0 < |x - c| < \delta \Rightarrow |f(x) - A| < \varepsilon. \tag{5.1.5}$$

We argue by contradiction and we assume that (5.1.5) is false, so that its opposite is true, i.e.,

$$\exists \varepsilon_0 > 0 : \ \ \forall \delta > 0, \ \ \exists x = x(\delta) \in X, \ \ 0 < |x(\delta) - c| < \delta \ \text{ and } |f(\, x(\delta)\,) - A| \geqslant \varepsilon_0. \tag{5.1.6}$$

From (5.1.6) we deduce that for any δ of the form $\delta = \frac{1}{n}$, $n \in \mathbb{N}$, there exists $x_n = x(1/n) \in X$ such that

$$0 < |x_n - c| < \frac{1}{n} \ \wedge \ |f(\, x_n\,) - A| \geqslant \varepsilon_0.$$

We have thus produced a sequence (x_n) in X such that

$$0 < \operatorname{dist}(x_n, c) < \frac{1}{n} \ \wedge \ \operatorname{dist}(f(x_n), A) \geqslant \varepsilon_0.$$

Thus, (x_n) is a sequence in $X\backslash\{c\}$ that converges to c, but the sequence $(\, f(x_n)\,)$ does not converge to A. □

Using Proposition 4.3.1 we obtain the following immediate consequence.

Corollary 5.1.5. *Let $f, g : X \to \mathbb{R}$ be two functions defined on the same subset $X \subset \mathbb{R}$ and c a cluster point of X. Suppose additionally that*

$$\lim_{x\to c} f(x) = A \ \ \textit{and} \ \ \lim_{x\to c} g(x) = B.$$

Then the following hold.

(i)
$$\lim_{x\to c}\big(\, f(x) + g(x)\,\big) = A + B, \ \ \lim_{x\to c} \lambda f(x) = \lambda A, \ \ \forall \lambda \in \mathbb{R}.$$

(ii)
$$\lim_{x\to c} f(x)g(x) = AB.$$

(iii) *If $B \neq 0$ and $g(x) \neq 0$, $\forall x \in X$, then*

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{A}{B}.$$

□

Example 5.1.6. (a) Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x$. Then for any $c \in \mathbb{R}$ we have

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

(b) Let $m \in \mathbb{N}$ and define $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^m$. Corollary 5.1.5 implies that

$$\lim_{x\to c} f(x) = \lim_{x\to c} x^m = c^m.$$

Thus,

$$\lim_{x\to 3} x^2 = 3^2 = 9.$$

(c) Let $m \in \mathbb{N}$ and define $f : (0,\infty) \to \mathbb{R}$, $f(x) = x^{-m} = \frac{1}{x^m}$. Corollary 5.1.5 implies that for any $c > 0$ we have

$$\lim_{x\to c} x^{-m} = \lim_{x\to c} \frac{1}{x^m} = c^{-m}.$$

(d) Let $m, k \in \mathbb{N}$ and define $f : (0,\infty) \to \mathbb{R}$, $f(x) = x^{\frac{m}{k}}$. We want to show that for any $c > 0$ we have

$$\lim_{x\to c} x^{\frac{m}{k}} = c^{\frac{m}{k}}. \tag{5.1.7}$$

We rely on Theorem 5.1.4. Suppose that (x_n) is a sequence of positive numbers such that $x_n \to c$ and $x_n \neq c$, $\forall n$. We have to show that

$$\lim_n x_n^{\frac{m}{k}} = c^{\frac{m}{k}}.$$

Using Exercise 4.15, we deduce that

$$\lim_n x_n^{\frac{1}{k}} = c^{\frac{1}{k}}.$$

Thus,

$$\lim_n x_n^{\frac{m}{k}} = \lim_n \big(x_n^{\frac{1}{k}} \big)^m = \big(c^{\frac{1}{k}} \big)^m = c^{\frac{m}{k}}.$$

Thus,

$$\lim_{x\to c} x^r = c^r, \;\; \forall r \in \mathbb{Q}, \;\; r > 0.$$

The above equality obviously holds if $r = 0$. If $r < 0$, then $x^{-r} = \frac{1}{x^r}$ and we deduce

$$\lim_{x\to c} x^r = c^r, \;\; \forall c > 0, \; r \in \mathbb{Q}. \tag{5.1.8}$$

□

Proposition 5.1.7. *Let $f, g : X \to \mathbb{R}$ be two functions defined on the same subset $X \subset \mathbb{R}$. Suppose that c is a cluster point of X and*

$$\lim_{x\to c} f(x) = A, \;\; \lim_{x\to c} g(x) = B \;\; \text{and} \;\; A < B.$$

Then there exists a $\delta_0 > 0$ such that $f(x) < g(x)$, $\forall x \in X$, $0 < |x - c| < \delta_0$.

Proof. Fix a positive number ε such that $3\varepsilon < B - A$. In other words, ε is smaller than one third of the distance from A to B. In particular, $A + \varepsilon < B - \varepsilon$ because

$$B - \varepsilon - (A + \varepsilon) = B - A - 2\varepsilon > 3\varepsilon - 2\varepsilon > 0.$$

Since $\lim_{x\to c} f(x) = A$, there exists $\delta = \delta_f(\varepsilon) > 0$ such that

$$\forall x \in X: \quad 0 < |x - c| < \delta_f \Rightarrow A - \varepsilon < f(x) < A + \varepsilon.$$

Since $\lim_{x\to c} g(x) = B$, there exists $\delta = \delta_g(\varepsilon) > 0$ such that

$$\forall x \in X: \quad 0 < |x - c| < \delta_g \Rightarrow B - \varepsilon < f(x) < B + \varepsilon.$$

Let $\delta_0 < \min\{\delta_f, \delta_g\}$ and define

$$U := (c - \delta_0, c + \delta_0).$$

If $x \in U \cap X$, $x \neq c$, then

$$0 < |x - c| < \delta_0 < \min\{\delta_f, \delta_g\} \Rightarrow f(x) < A + \varepsilon < B - \varepsilon < g(x).$$

□

5.2. Exponentials and logarithms

In this section we want to give a meaning to the exponential a^x where a is a positive real number and x is an arbitrary real number. The case $a = 1$ is trivial: we define $1^x = 1$, $\forall x \in \mathbb{R}$.

We consider next the case $a > 1$. In Exercise 3.15 we defined a^r for any $r \in \mathbb{Q}$ and we showed that

$$a^{r_1+r_2} = a^{r_1} \cdot a^{r_2}, \quad a^{r_1-r_2} = \frac{a^{r_1}}{a^{r_2}}, \quad \big(a^{r_1}\big)^{r_2} = a^{r_1 r_2}, \quad \forall r_1, r_2 \in \mathbb{Q}. \tag{5.2.1}$$

We will use these facts to define a^x for any $x \in \mathbb{R}$. This will require several auxiliary results.

Lemma 5.2.1. *If $a > 1$, then for any rational numbers r_1, r_2 we have*

$$r_1 < r_2 \Rightarrow a^{r_1} < a^{r_2}.$$

Proof. We will use the fact that if $x, y > 0$ and $n \in \mathbb{N}$, then

$$x < y \Longleftrightarrow x^n < y^n.$$

Since $a > 1$ we deduce that $a^{\frac{1}{n}} > 1$ because

$$\big(a^{\frac{1}{n}}\big)^n = a > 1 = 1^n.$$

Thus,

$$a^{\frac{m}{n}} > 1, \quad \forall m, n \in \mathbb{N}$$

that is,

$$a^r > 1, \quad \forall r \in \mathbb{Q}, \quad r > 0.$$

Suppose that $r_1 < r_2$. Then the above inequality implies that

$$\frac{a^{r_2}}{a^{r_1}} \overset{(5.2.1)}{=} a^{r_2 - r_1} > 1$$

because $r = r_2 - r_1$ is a positive rational number. □

Lemma 5.2.2. *Let $a > 1$ and $r_0 \in \mathbb{Q}$. Then*

$$\lim_{\mathbb{Q} \ni r \to r_0} a^r = a^{r_0}.$$

Proof. We first consider the case $r_0 = 0$, i.e., we first prove that

$$\lim_{\mathbb{Q} \ni r \to 0} a^r = 1. \tag{5.2.2}$$

We have to prove that, given $\varepsilon > 0$, we can find $\delta = \delta(\varepsilon) > 0$ such that

$$0 < |r| < \delta \text{ and } r \in \mathbb{Q} \Rightarrow |a^r - 1| < \varepsilon.$$

Observe first that Exercise 4.12 implies that

$$\lim_{n \to \infty} a^{\frac{1}{n}} = \lim_{n \to \infty} a^{-\frac{1}{n}} = 1.$$

In particular, this implies that there exists $n_0 = n_0(\varepsilon) > 0$ such that, for all $n \geqslant n_0$, we have

$$1 - \varepsilon < a^{-\frac{1}{n}} < a^{\frac{1}{n}} < 1 + \varepsilon.$$

We set $\delta(\varepsilon) = \frac{1}{n_0(\varepsilon)}$. If $0 < |r| < \delta(\varepsilon)$ and $r \in \mathbb{Q}$, then $-\frac{1}{n_0(\varepsilon)} < r < \frac{1}{n_0(\varepsilon)}$ and we deduce from Lemma 5.2.1 that

$$1 - \varepsilon < a^{-\frac{1}{n_0(\varepsilon)}} < a^r < a^{\frac{1}{n_0(\varepsilon)}} < 1 + \varepsilon \Rightarrow 1 - \varepsilon < a^r < 1 + \varepsilon \Rightarrow |a^r - 1| < \varepsilon.$$

This proves (5.2.2). To deal with the general case, let $r_0 \in \mathbb{Q}$. If r_n is a sequence of rational numbers $r_n \to r_0$, then

$$a^{r_n} = a^{r_0} a^{r_n - r_0}.$$

Since $r_n - r_0 \to 0$, we deduce from (5.2.2) that $a^{r_n - r_0} \to 1$ and thus, $a^{r_n} = a^{r_0} a^{r_n - r_0} \to a^{r_0}$. The conclusion now follows from Theorem 5.1.4. □

Proposition 5.2.3. *Let $a > 1$ and $x \in \mathbb{R}$. We set*

$$\mathbb{Q}_{<x} := \big\{ r \in \mathbb{Q}, \;\; r < x \big\}, \;\; \mathbb{Q}_{>x} := \big\{ r \in \mathbb{Q}, \;\; r > x \big\}$$

$$s_x = \sup_{r \in \mathbb{Q}_{<x}} a^r, \;\; i_x = \inf_{r \in \mathbb{Q}_{>x}} a^r.$$

Then $s_x = i_x$. Moreover, if x is rational, then $s_x = i_x = a^x$.

Proof. Observe first that the set $\{a^r;\ \ r \in \mathbb{Q}_{<x}\}$ is bounded above. Indeed, if we choose a rational number $R > x$, then Lemma 5.2.1 implies that $a^r < a^R$ for any rational number $r < x$. A similar argument shows that the set $\{a^r;\ \ r \in \mathbb{Q}_{>x}\}$ is bounded below and we have

$$s_x \leqslant i_x.$$

Observe that for any rational numbers r_1, r_2 such that $r_1 < x < r_2$, we have

$$a^{r_1} \leqslant s_x \leqslant i_x \leqslant a^{r_2}.$$

Hence,

$$1 \leqslant \frac{i_x}{s_x} \leqslant \frac{a^{r_2}}{s_x} \leqslant \frac{a^{r_2}}{a^{r_1}} = a^{r_2 - r_1}.$$

Now choose two sequences $(r'_n) \subset \mathbb{Q}_{<x}$ and $(r''_n) \subset \mathbb{Q}_{>x}$ such that $r'_n \to x$ and $r''_n \to x$.[1] Then

$$1 \leqslant \frac{s_x}{i_x} \leqslant a^{r''_n - r'_n}.$$

If we let $n \to \infty$ and observe that $r''_n - r'_n \to 0$, we deduce from Lemma 5.2.2 that

$$1 \leqslant \frac{s_x}{i_x} \leqslant \lim_{n\to\infty} a^{r''_n - r'_n} = 1 \Rightarrow s_x = i_x.$$

If $x \in \mathbb{Q}$, then the sequences r'_n and r''_nabove converge to x. Invoking Lemma 5.2.2 we deduce

$$s_x = \lim_n a^{r'_n} = a^x = \lim_n a^{r''_n} = i_x.$$

□

Definition 5.2.4. For any $a > 1$ and $x \in \mathbb{R}$ we set

$$\boxed{a^x := \sup\big\{a^r;\ \ r \in \mathbb{Q},\ \ r < x\big\} = \inf\big\{a^r;\ \ r \in \mathbb{Q},\ \ r > x\big\}}.$$

If $b \in (0,1)$, then $\frac{1}{b} > 1$ and we set

$$b^x := \left(\frac{1}{b}\right)^{-x}.$$

□

Lemma 5.2.5. *Let $a > 1$. If $x < y$, then $a^x < a^y$.*

Proof. We can find rational numbers r_1, r_2 such that

$$x < r_1 < r_2 < y.$$

Then $r_1 \in \mathbb{Q}_{>x}$ and $r_2 \in \mathbb{Q}_{<y}$ so that

$$a^x \leqslant a^{r_1} < a^{r_2} \leqslant a^y.$$

□

Lemma 5.2.6. *Let $a > 1$ and $x \in \mathbb{R}$. If the sequence $(r_n) \subset \mathbb{Q}_{<x}$ converges to x, then*

$$\lim_{n\to\infty} a^{r_n} \to a^x.$$

[1] The existence of such sequences was left to you as Exercise 4.13.

Proof. We have

$$a^x = \sup_{r \in \mathbb{Q}_{<x}} a^r.$$

Thus, for any $\varepsilon > 0$, there exists $r_\varepsilon \in \mathbb{Q}_{<x}$ such that

$$a^x - \varepsilon < a^{r_\varepsilon} \leqslant a^x.$$

Since $r_n \to x$ and $r_n \in \mathbb{Q}_{<x}$, we deduce that there exists $N = N(\varepsilon)$ such that, $\forall n > N(\varepsilon)$ we have $r_\varepsilon < r_n < x$. We deduce that for all $n > N(\varepsilon)$, we have

$$a^x - \varepsilon < a^{r_\varepsilon} < a^{r_n} < a^x.$$

□

Lemma 5.2.7. *Let $a > 0$ and $x, y > 0$. Then*

$$a^x \cdot a^y = a^{x+y}.$$

Proof. Choose sequences $(r_n') \subset \mathbb{Q}_{<x}$ and $(r_n'') \subset \mathbb{Q}_{<y}$ such that $r_n' \to x$ and $r_n'' \to y$. Lemma 5.2.6 implies that

$$a^{r_n'} \to a^x \quad \wedge \quad a^{r_n''} \to a^y.$$

Hence,

$$\lim_n a^{r_n' + r_n''} = \lim_n \big(a^{r_n'} \cdot a^{r_n''} \big) = \big(\lim_n a^{r_n'} \big) \cdot \big(\lim_n a^{r_n''} \big) = a^x \cdot a^y.$$

Now observe that $r_n' + r_n'' \in \mathbb{Q}_{<x+y}$ and $r_n' + r_n'' \to x + y$. Lemma 5.2.6 implies

$$\lim_n a^{r_n' + r_n''} = a^{x+y}.$$

□

The proofs of our next two results are left to you as an exercise.

Lemma 5.2.8. *Let $a > 0$ and $x \in \mathbb{R}$. Then for any sequence of real numbers (x_n) such that $x_n \to x$ we have*

$$\lim_{n \to \infty} a^{x_n} = a^x.$$

□

Lemma 5.2.9. *Suppose that $a, b > 0$. Then for any $x \in \mathbb{R}$ we have*

$$a^x \cdot b^x = (ab)^x. \tag{5.2.3}$$

□

Definition 5.2.10. Let $X \subset \mathbb{R}$ and $f : X \to \mathbb{R}$ be a real valued function defined on X.

(i) The function f is called *increasing* if

$$\forall x_1, x_2 \in X \ \ (x_1 < x_2) \Rightarrow \big(f(x_1) < f(x_2) \big).$$

(ii) The function f is called *decreasing* if

$$\forall x_1, x_2 \in X \ \ (x_1 < x_2) \Rightarrow \big(f(x_1) > f(x_2) \big).$$

(iii) The function f is called *nondecreasing* if

$$\forall x_1, x_2 \in X \ \ (x_1 < x_2) \Rightarrow \big(f(x_1) \leqslant f(x_2) \big).$$

(iv) The function f is called *nonincreasing* if

$$\forall x_1, x_2 \in X \ \ (x_1 < x_2) \Rightarrow \big(\, f(x_1) \geqslant f(x_2)\,\big).$$

(v) The function is called *strictly monotone* if it is either increasing or decreasing. It is called *monotone* if it is either nondecreasing or nonincreasing.

□

Theorem 5.2.11. *Let $a > 0$, $a \neq 1$. Consider the function $f_a : \mathbb{R} \to (0, \infty)$ given by $f(x) = a^x$. Then the following hold.*

(i) $a^{x+y} = a^x \cdot a^y$, $\forall x, y \in \mathbb{R}$.

(ii) $(a^x)^y = a^{xy}$, $\forall x, y \in \mathbb{R}$.

(iii) *The function f_a is increasing if $a > 1$, and decreasing if $a < 1$.*

(iv) *The function f is bijective.*

(v) *For any sequence of real numbers (x_n) such that $x_n \to x$ we have*

$$\lim_{n \to \infty} a^{x_n} = a^x.$$

Proof. Part (v) above is Lemma 5.2.8. We thus have to prove (i)-(iv). We consider first the case $a > 1$. The equality (i) is Lemma 5.2.7. The statement (iii) follows from Lemma 5.2.5.

We first prove (ii) in the special case $y \in \mathbb{Q}$. Choose a sequence $r_n \in \mathbb{Q}$ such that $r_n \to x$, $r_n \neq x$. Then (5.2.1) implies

$$(a^{r_n})^y = a^{r_n y}.$$

Clearly $r_n y \to xy$ and Lemma 5.2.8 implies that

$$\lim_n a^{r_n y} = a^{xy}.$$

On the other hand, y is rational and $a^{r_n} \to a^x$ and using (5.1.8) we deduce that

$$\lim_n \big(\, a^{r_n}\,\big)^y = \big(\, a^x\,\big)^y.$$

Thus,

$$\big(\, a^x\,\big)^y = \lim_n \big(\, a^{r_n}\,\big)^y = \lim_n a^{r_n y} = a^{xy}, \ \ \forall x \in \mathbb{R}, \ \ y \in \mathbb{Q}. \tag{5.2.4}$$

Now fix $x, y \in \mathbb{R}$ and choose a sequence of *rational* numbers $y_n \to y$, $y_n \neq y$. Then

$$\big(\, a^x\,\big)^{y_n} \overset{(5.2.4)}{=} a^{xy_n}, \ \ \forall n.$$

Using Lemma 5.2.8, we deduce

$$\big(\, a^x\,\big)^y = \lim_n \big(\, a^x\,\big)^{y_n} = \lim_n a^{xy_n} = a^{xy}, \ \ \forall x, y \in \mathbb{R}.$$

This proves (ii).

To prove (iv) observe that f_a is injective because it is increasing. (We recall that we are working under the assumption $a > 1$.) To prove surjectivity, fix $y \in (0, \infty)$. We have to show that there exists $x \in \mathbb{R}$ such that $a^x = y$. Define

$$S := \big\{ s \in \mathbb{R}; \ \ a^s \leqslant y \,\big\}.$$

Observe first that $S \neq \varnothing$. Indeed

$$\lim_n a^{-n} = \lim_n \frac{1}{a^n} = 0$$

so that there exists $n_0 \in \mathbb{N}$ such that $a^{-n_0} < y$, i.e., $-n_0 \in S$. Observe that S is also bounded above. Indeed

$$\lim_n a^n = \infty.$$

Hence there exists $n_1 \in \mathbb{N}$ such that $a^{n_1} > y$. If $x \geqslant n_1$, then $a^x \geqslant a^{n_1} > y$ so that $S \cap [n_1, \infty) = \varnothing$ and thus $S \subset (-\infty, n_1)$ and therefore n_1 is an upper bound for S. Set

$$x := \sup S.$$

Note that if $x' > x$, then $a^{x'} \geqslant y$. Indeed, if $a^{x'} < y$ then for any $s < x'$ we have $a^s < a^{x'} < y$ and thus $(-\infty, x'] \subset S$. This contradicts the fact that x is an upper bound for S.

Consider now two sequences $s_n' \to x$ and $s_n'' \to x$ where $s_n' < x$ and $s_n'' > x$ then

$$a^{s_n'} \leqslant y \leqslant a^{s_n''}, \quad \forall n.$$

Letting $n \to \infty$ in the above inequalities we obtain, from Lemma 5.2.8, that

$$a^x \leqslant y \leqslant a^x \Longleftrightarrow a^x = y.$$

The case $a < 1$ follows from the case $a > 1$ by observing that

$$a^x = \left(\frac{1}{a}\right)^{-x}.$$

□

Definition 5.2.12. Let $a \in (0, \infty)$, $a \neq 1$. The bijective function

$$\mathbb{R} \ni x \mapsto a^x \in (0, \infty)$$

is called the *exponential function with base a*. Its inverse is called the *logarithm to base a* and it is a function

$$\log_a : (0, \infty) \to \mathbb{R}.$$

When $a = e =$ the Euler number, we will refer to $\log_e$ as the *natural logarithm* and we will use the simpler notation log or ln. Also, we will use the notation lg for $\log_{10}$. □

We have depicted below the graphs of the functions a^x and $\log_a x$ for $a = 2$ and $a = \frac{1}{2}$.

The meaning of the logarithm function answers the following question: given $a, y > 0$, $a \neq 1$, to what power do we need to raise a in order to obtain y? The answer: we need to raise a to the power $\log_a y$ in order to get y. Equivalently, $\log_a$ is uniquely determined by the following two fundamental identities

$$\log_a a^x = x \;\text{ and }\; a^{\log_a y} = y, \quad \forall x \in \mathbb{R}, \quad y > 0.$$

For example, $\log_2 8 = 3$ because $2^3 = 8$. Similarly $\lg 10,000 = 4$ since $10^4 = 10,000$.

Theorem 5.2.13. *Let $a > 0$, $a \neq 1$. Then the following hold.*

(i) *For any $y_1, y_2 > 0$ we have*

$$\log_a (y_1 y_2) = \log_a y_1 + \log_a y_2, \quad \log_a \frac{y_1}{y_2} = \log_a y_1 - \log_a y_2.$$

(ii) $\log_a y^\alpha = \alpha \log_a y$, $\forall y > 0$, $\alpha \in \mathbb{R}$.

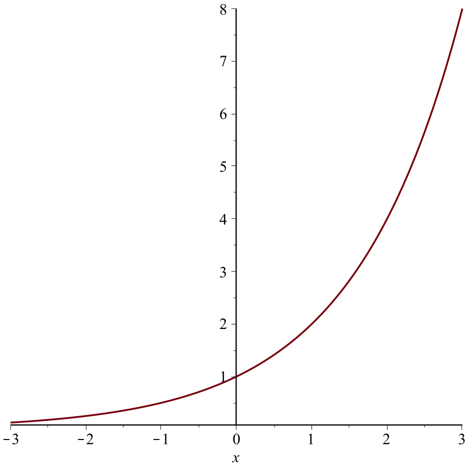

Figure 5.1. *The graph of* 2^x.

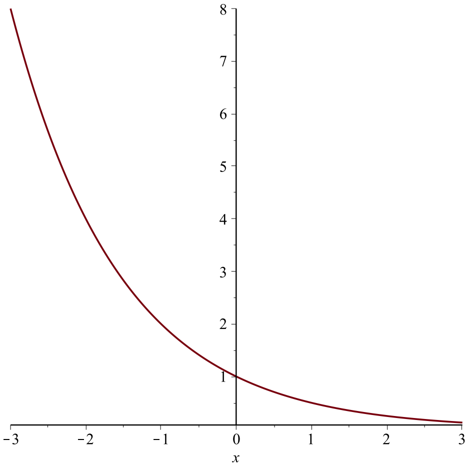

Figure 5.2. *The graph of* $\left(\frac{1}{2}\right)^x$.

(iii) *If* $b > 0$ *and* $b \neq 1$*, then*

$$\log_b y = \frac{\log_a y}{\log_a b}, \quad \forall y > 0.$$

(iv) *If* $a > 1$*, then the function* $y \mapsto \log_a y$ *is increasing, while if* $a \in (0, 1)$*, then the function* $y \mapsto \log_a y$ *is decreasing.*

(v) *If* $y > 0$*, then for any sequence of positive numbers* (y_n) *that converges to* y *we have*

$$\lim_{n\to\infty} \log_a y_n = \log_a y.$$

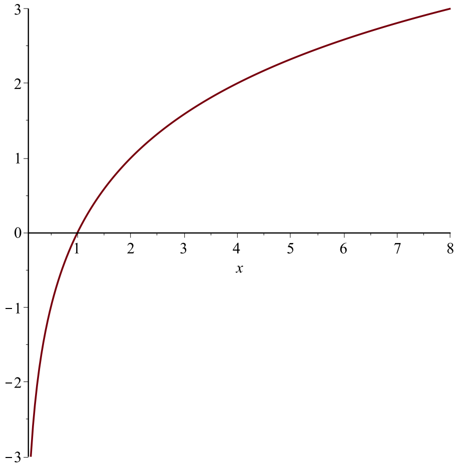

Figure 5.3. *The graph of* $\log_2 x$.

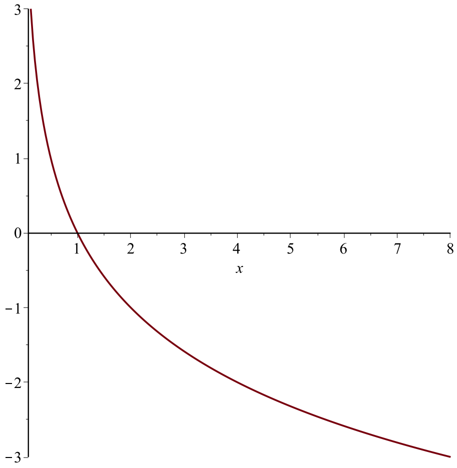

Figure 5.4. *The graph of* $\log_{1/2} x$.

Proof. (i) Let $y_1, y_2 > 0$. Set $x_1 = \log_a y_1$, $x_2 = \log_a y_2$, i.e., $a^{x_1} = y_1$ and $y_2 = a^{x_2}$. We have to show that

$$\log_a(y_1y_2) = x_1 + x_2, \ \ \log_a \frac{y_1}{y_2} = x_1 - x_2.$$

We have

$$y_1y_2 = a^{x_1}a^{x_2} = a^{x_1+x_2} \Rightarrow \log_a(y_1y_2) = \log_a a^{x_1+x_2} = x_1 + x_2,$$
$$\frac{y_1}{y_2} = \frac{a^{x_1}}{a^{x_2}} = a^{x_1-x_2} \Rightarrow \log_a \frac{y_1}{y_2} = \log_a a^{x_1-x_2} = x_1 - x_2.$$

(ii) Let $x \in \mathbb{R}$ such that $a^x = y$, i.e., $\log_a y = x$. We have to prove that

$$\log_a y^{\alpha} = \alpha x.$$

We have

$$y^{\alpha} = (a^x)^{\alpha} = a^{\alpha x} \Rightarrow \log_a y^{\alpha} = \log_a a^{\alpha x} = \alpha x.$$

(iii) Let $\beta, x, t \in \mathbb{R}$ such that $a^{\beta} = b$, $y = a^x = b^t$. Then

$$y = b^t = (a^{\beta})^t = a^{t\beta} = a^x.$$

Hence,

$$\log_a y = x = t\beta = (\log_b y)(\log_a b) \Rightarrow \log_b y = \frac{\log_a y}{\log_a b}.$$

(iv) Assume first that $a > 1$. Consider the numbers $y_2 > y_1 > 0$, and set

$$x_1 := \log_a y_1, \ \ x_2 = \log_a y_2.$$

We have to show that $x_2 > x_1$. We argue by contradiction. If $x_1 \geqslant x_2$, then

$$y_1 = a^{x_1} \geqslant a^{x_2} = y_2 \Rightarrow y_1 \geqslant y_2.$$

This contradiction proves the statement (iv) in the case $a > 1$. The case $a \in (0,1)$ is dealt with in a similar fashion.

(v) Assume first that $a > 1$ so that the function $y \mapsto \log_a y$ is increasing. Since $y_n \to y$, we deduce that

$$\frac{y_n}{y} \to 1.$$

Hence, for any $\varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that

$$\forall n > N(\varepsilon) : \frac{y_n}{y_0} \in (a^{-\varepsilon}, a^{\varepsilon}).$$

Hence, $\forall n > N(\varepsilon)$

$$-\varepsilon = \log_a a^{-\varepsilon} < \underbrace{\log_a\left(\frac{y_n}{y_0}\right)}_{=\log_a y_n - \log_a y_0} < \log_a a^{\varepsilon} = \varepsilon \Longleftrightarrow |\log_a y_n - \log_a y_0| < \varepsilon.$$

□

Theorem 5.2.14. *Fix a real number s and consider $f : (0,\infty) \to (0,\infty)$ given by $f(x) = x^s$. Then for any $c > 0$ any sequence of positive numbers (x_n), and any sequence of real numbers (s_n) such that $x_n \to c$, and $s_n \to s$, we have*

$$\boxed{\lim_{n\to\infty} x_n^{s_n} = c^s.}$$

Proof. Set

$$y_n := \log x_n^{s_n} = s_n \log x_n.$$

Theorem 5.2.13(v) implies that

$$\lim_n y_n = (\lim_n s_n)(\lim_n \log x_n) = s \log c.$$

Using Theorem 5.2.11(v), we deduce that

$$\lim_n e^{y_n} = e^{s \log c} = (e^{\log c})^s = c^s.$$

Now observe that

$$e^{y_n} = e^{\log x_n^{s_n}} = x_n^{s_n}.$$

This proves Theorem 5.2.14. □

5.3. Limits involving infinities

Suppose that we are given a subset $X \subset \mathbb{R}$ and a function $f : X \to \mathbb{R}$.

Definition 5.3.1. Let c be a cluster point of X.

(a) We say that the limit of f as $x \to c$ is ∞, and we write this

$$\lim_{x\to c} f(x) = \infty$$

if for any $M > 0$, $\exists \delta = \delta(M) > 0$ such that

$$\forall x \in X \ \ \big(0 < |x - c| < \delta \Rightarrow f(x) > M \big).$$

(b) We say that the limit of f as $x \to c$ is $-\infty$, and we write this

$$\lim_{x\to c} f(x) = -\infty$$

if for any $M > 0$, $\exists \delta = \delta(M) > 0$ such that

$$\forall x \in X \ \ \big(0 < |x - c| < \delta \Rightarrow f(x) < -M \big).$$

□

We have the following version of Proposition 5.1.3. The proof is left to you.

Proposition 5.3.2. *Let $f : X \to \mathbb{R}$ be a function defined on a set $X \subset \mathbb{R}$ and c a cluster point of X. Then the following statements are equivalent.*

(i) $\lim_{x\to c} f(x) = \infty$, *i.e., f satisfies (5.1.1) or (5.1.2).*

(ii)

$$\forall M > 0, \ \ \exists V \in \mathcal{N}_c^* \ \text{ such that } \ \forall x \in X : x \in V \Rightarrow f(x) \in (M, \infty). \tag{5.3.1}$$

□

Arguing as in the proof of Theorem 5.1.4 we obtain the following result. The details are left to you.

Theorem 5.3.3. *Let c be a cluster point of the set $X \subset \mathbb{R}$ and $f : X \to \mathbb{R}$ a real valued function on X. The following statements are equivalent.*

(i) $\lim_{x\to c} f(x) = \infty \in \mathbb{R}$.

(ii) *For any sequence $(x_n)_{n\in\mathbb{N}}$ in $X\backslash\{c\}$ such that $x_n \to c$, we have $\lim_n f(x_n) = \infty$.*

□

Observe that if $X \subset \mathbb{R}$ is not bounded above, then for any $M > 0$ the intersection $X \cap (M, \infty)$ is nonempty, i.e., for any number $M > 0$ there exists at least one number $x \in X$ such that $x > M$. Equivalently, this means that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of numbers in X such that

$$\lim_{n\to\infty} x_n = \infty.$$

Definition 5.3.4. Suppose $X \subset \mathbb{R}$ is a subset not bounded above and $f : X \to \mathbb{R}$ is a real function defined on X.

(a) We say that the limit of f as $x \to \infty$ is the real number A, and we write this $\lim_{x\to\infty} f(x) = A$, if

$$\forall \varepsilon > 0 \;\; \exists M = M(\varepsilon) > 0 \;\; \forall x \in X \;\; (x > M \Rightarrow |f(x) - A| < \varepsilon).$$

(b) We say that the limit of f as $x \to \infty$ is ∞, and we write this $\lim_{x\to\infty} f(x) = \infty$, if

$$\forall C > 0 \;\; \exists M = M(C) > 0 \;\; \forall x \in X \;\; (x > M \Rightarrow f(x) > C).$$

(c) We say that the limit of f as $x \to \infty$ is $-\infty$, and we write this $\lim_{x\to\infty} f(x) = -\infty$, if

$$\forall C > 0 \;\; \exists M = M(C) > 0 \;\; \forall x \in X \;\; (x > M \Rightarrow f(x) < -C). \qquad \square$$

Observe that if $X \subset \mathbb{R}$ is not bounded below, then for any $M > 0$ the intersection $X \cap (-\infty, -M)$ is nonempty, i.e., for any number $M > 0$ there exists at least one number $x \in X$ such that $x < -M$. Equivalently, this means that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of numbers in X such that

$$\lim_{n\to\infty} x_n = -\infty.$$

Definition 5.3.5. Suppose $X \subset \mathbb{R}$ is a subset not bounded below and $f : X \to \mathbb{R}$ is a real function defined on X.

(a) We say that the limit of f as $x \to -\infty$ is the real number A, and we write this $\lim_{x\to-\infty} f(x) = A$, if

$$\forall \varepsilon > 0 \;\; \exists M = M(\varepsilon) > 0 \;\; \forall x \in X \;\; (x < -M \Rightarrow |f(x) - A| < \varepsilon).$$

(b) We say that the limit of f as $x \to -\infty$ is ∞, and we write this $\lim_{x\to-\infty} f(x) = \infty$, if

$$\forall C > 0 \;\; \exists M = M(C) > 0 \;\; \forall x \in X \;\; (x < -M \Rightarrow f(x) > C).$$

(c) We say that the limit of f as $x \to -\infty$ is $-\infty$, and we write this $\lim_{x\to-\infty} f(x) = -\infty$, if

$$\forall C > 0 \;\; \exists M = M(C) > 0 \;\; \forall x \in X \;\; (x < -M \Rightarrow f(x) < -C). \qquad \square$$

The limits involving infinities have an alternate description involving sequences. Thus, if $X \subset \mathbb{R}$ is not bounded above and $f : X \to \mathbb{R}$ is a real function defined on X, then the equality

$$\lim_{x\to\infty} f(x) = A$$

can be given a characterization as in Theorem 5.1.4. More precisely, it means that for any sequence of real numbers $x_n \in X$ such that $x_n \to \infty$, the sequence $f(x_n)$ converges to A.

Example 5.3.6. (a) We want to prove that

$$\boxed{\lim_{x\to\infty}\left(1 + \frac{1}{x}\right)^x = e.} \tag{5.3.2}$$

We will use the fundamental result in Example 4.4.2 which states that the sequence

$$x_n := \left(1+\frac{1}{n}\right)^n, \quad n \in \mathbb{N},$$

converges to the Euler number e. In particular, we deduce that

$$\lim_{n\to\infty}\left(1+\frac{1}{n+1}\right)^n = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n+1} = e. \tag{5.3.3}$$

Recall that for any real number x we denote by $\lfloor x\rfloor$ the integer part of the real number x, i.e., the largest integer which is $\leqslant x$. Thus $\lfloor x\rfloor$ is an integer and

$$\lfloor x\rfloor \leqslant x < \lfloor x\rfloor + 1.$$

For $x \geqslant 1$ we have

$$1 \leqslant \lfloor x\lfloor\leqslant x \leqslant \lfloor x\rfloor + 1$$

and we deduce

$$1+\frac{1}{\lfloor x\rfloor+1} \leqslant 1+\frac{1}{x} \leqslant 1+\frac{1}{\lfloor x\rfloor}.$$

In particular, we deduce that

$$\left(1+\frac{1}{\lfloor x\rfloor+1}\right)^{\lfloor x\rfloor} \leqslant \left(1+\frac{1}{x}\right)^{\lfloor x\rfloor} \leqslant \left(1+\frac{1}{x}\right)^{x} \leqslant \left(1+\frac{1}{\lfloor x\rfloor}\right)^{x} \leqslant \left(1+\frac{1}{\lfloor x\rfloor}\right)^{\lfloor x\rfloor+1}. \tag{5.3.4}$$

From (5.3.3) we deduce that for any $\varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that

$$\left(1+\frac{1}{n+1}\right)^n, \quad \left(1+\frac{1}{n}\right)^{n+1} \in (e-\varepsilon, e+\varepsilon), \quad \forall n > N(\varepsilon).$$

If $x > N(\varepsilon)+1$, then $\lfloor x\rfloor > N(\varepsilon)$ and we deduce from the above that

$$e-\varepsilon < \left(1+\frac{1}{\lfloor x\rfloor+1}\right)^{\lfloor x\rfloor} < e+\varepsilon \;\text{ and }\; e-\varepsilon < \left(1+\frac{1}{\lfloor x\rfloor}\right)^{\lfloor x\rfloor+1} < e+\varepsilon.$$

The inequalities (5.3.4) now imply that for $x > N(\varepsilon)+1$ we have

$$e-\varepsilon < \left(1+\frac{1}{\lfloor x\rfloor+1}\right)^{\lfloor x\rfloor} \leqslant \left(1+\frac{1}{x}\right)^{x} \leqslant \left(1+\frac{1}{\lfloor x\rfloor}\right)^{\lfloor x\rfloor+1} < e+\varepsilon.$$

This proves (5.3.2).

(b) We want to prove that

$$\boxed{\lim_{x\to-\infty}\left(1+\frac{1}{x}\right)^{x} = e.} \tag{5.3.5}$$

We will prove that for any sequence of nonzero real numbers (x_n) such that $x_n \to -\infty$, we have

$$\lim_{n}\left(1+\frac{1}{x_n}\right)^{x_n} = e.$$

Consider the new sequence $y_n := -x_n$. Clearly $y_n \to \infty$. We have

$$\left(1+\frac{1}{x_n}\right)^{x_n} = \left(1-\frac{1}{y_n}\right)^{-y_n} = \left(\frac{y_n-1}{y_n}\right)^{-y_n} = \left(\frac{y_n}{y_n-1}\right)^{y_n}.$$

Now set $z_n := y_n - 1$ so that $y_n = z_n + 1$ and

$$\left(\frac{y_n}{y_n-1}\right)^{y_n} = \left(\frac{z_n+1}{z_n}\right)^{z_n+1} = \left(1+\frac{1}{z_n}\right)^{z_n+1} = \left(1+\frac{1}{z_n}\right)^{z_n} \times \left(1+\frac{1}{z_n}\right).$$

Clearly $z_n \to \infty$ so that

$$\lim_n \left(1+\frac{1}{z_n}\right) = 1.$$

Invoking (5.3.2) we deduce

$$\lim_{n\to\infty} \left(1+\frac{1}{z_n}\right)^{z_n} = e.$$

Hence,

$$\lim_n \left(1+\frac{1}{x_n}\right)^{x_n} = \lim_n \left(1+\frac{1}{z_n}\right)^{z_n} \times \lim_n \left(1+\frac{1}{z_n}\right) = e.$$

This proves (5.3.5). □

5.4. One-sided limits

Suppose $X \subset \mathbb{R}$ is a set of real numbers. For any $c \in \mathbb{R}$ we define

$$X_{<c} := \big\{ x \in X;\;\; x < c \big\} = X \cap (-\infty, c),\;\; X_{>c} := \big\{ x \in X;\;\; x > c \big\} = X \cap (c, \infty).$$

Definition 5.4.1. Let $f : X \to \mathbb{R}$ and $c \in \mathbb{R}$. We say that L is the *left limit* of f at c, and we write this

$$L = \lim_{x \nearrow c} f(x) = \lim_{x\to c-} f(x),$$

if

- c is a cluster point of $X_{<c}$ and
- for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\forall x \in X : \;\; x \in (c-\delta, c) \Rightarrow |f(x) - L| < \varepsilon.$$

We say that R is the *right limit* of f at c, and we write this

$$R = \lim_{x \searrow c} f(x) = \lim_{x\to c+} f(x),$$

if

- c is a cluster point of $X_{>c}$ and
- for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\forall x \in X : \;\; x \in (c, c+\delta) \Rightarrow |f(x) - R| < \varepsilon.$$

□

The next result follows immediately from Theorem 5.1.4. The details are left to you.

Theorem 5.4.2. *Let $f : X \to \mathbb{R}$ be a real valued function defined on the set $X \subset \mathbb{R}$. Fix $c \in \mathbb{R}$.*

(a) Suppose that c is a cluster point of $X_{<c}$ and $L \in \mathbb{R}$. Then the following statements are equivalent.

(i)
$$\lim_{x \nearrow c} f(x) = L.$$

(ii) *For any sequence of real numbers (x_n) in X such that $x_n \to c$ and $x_n < c$ $\forall n$ we have*
$$\lim_n f(x_n) = L.$$

(iii) *For any nondecreasing sequence of real numbers (x_n) in X such that $x_n \to c$ and $x_n < c$ $\forall n$ we have*
$$\lim_n f(x_n) = L.$$

(b) Suppose that c is a cluster point of $X_{>c}$ and $L \in \mathbb{R}$. Then the following statements are equivalent.

(i)
$$\lim_{x \searrow c} f(x) = L.$$

(ii) *For any sequence of real numbers (x_n) in X such that $x_n \to c$ and $x_n > c$ $\forall n$ we have*
$$\lim_n f(x_n) = L.$$

(iii) *For any nonincreasing sequence of real numbers (x_n) in X such that $x_n \to c$ and $x_n > c$ $\forall n$ we have*
$$\lim_n f(x_n) = L.$$

□

The next result describes one of the reasons why the one-sided limits are useful. Its proof is left to you as an exercise.

Theorem 5.4.3. *Consider three real numbers $a < c < b$, a real valued function*
$$f : (a, b)\backslash\{c\} \to \mathbb{R}.$$
and suppose that $A \in [-\infty, \infty]$. Then the following statements are equivalent.

(i)
$$\lim_{x \to c} f(x) = A.$$

(ii)
$$\lim_{x \nearrow c} f(x) = \lim_{x \searrow c} f(x).$$

□

5.5. Some fundamental limits

In this section we present a collection of examples that play a fundamental role in the development of real analysis.

Example 5.5.1. We want to prove that

$$\boxed{\lim_{x\to 0}\bigl(\,1+x\,\bigr)^{\frac{1}{x}}=e.} \tag{5.5.1}$$

We invoke Theorem 5.4.3, so we will prove that

$$\lim_{x\searrow 0}\bigl(\,1+x\,\bigr)^{\frac{1}{x}}=\lim_{x\nearrow 0}\bigl(\,1+x\,\bigr)^{\frac{1}{x}}=e.$$

We prove first the equality

$$\lim_{x\searrow 0}\bigl(\,1+x\,\bigr)^{\frac{1}{x}}=e.$$

We have to prove that if (x_n) is a sequence of *positive* numbers such that $x_n\to 0$, then

$$\lim_n\bigl(\,1+x_n\,\bigr)^{\frac{1}{x_n}}=e.$$

Set

$$y_n:=\frac{1}{x_n}.$$

Then $y_n\to\infty$ and

$$\bigl(\,1+x_n\,\bigr)^{\frac{1}{x_n}}=\left(1+\frac{1}{y_n}\right)^{y_n},$$

and, according to (5.3.3), we have

$$\left(1+\frac{1}{y_n}\right)^{y_n}=e.$$

The equality

$$\lim_{x\nearrow 0}\bigl(\,1+x\,\bigr)^{\frac{1}{x}}=e.$$

is proved in a similar fashion invoking (5.3.5) instead of (5.3.3). □

Example 5.5.2. We have ($\log=\log_e$)

$$\boxed{\lim_{x\to 0}\frac{\log\bigl(\,1+x\,\bigr)}{x}=1.} \tag{5.5.2}$$

Indeed, consider a sequence of nonzero numbers (x_n) such that $x_n\to 0$. Set

$$y_n=(1+x_n)^{\frac{1}{x_n}}.$$

From (5.5.2) we deduce that $y_n \to e$. Using Theorem 5.2.13(v), we deduce that $\log y_n \to \log e = 1$. □

Example 5.5.3. We have

$$\boxed{\lim_{x\to 0} \frac{e^x-1}{x} = 1.} \tag{5.5.3}$$

Let $x_n \to 0$. Set $y_n := e^{x_n}$ so that $x_n = \log y_n$ and $y_n \to e^0 = 1$. Next, set $h_n := y_n - 1$ so that $h_n \to 0$. Then

$$\frac{e^{x_n}-1}{x_n} = \frac{y_n-1}{\log y_n} = \frac{h_n}{\log(1+h_n)} = \frac{1}{\frac{\log(1+h_n)}{h_n}} \xrightarrow{(5.5.2)} 1.$$ □

Example 5.5.4. Suppose that $\alpha \in \mathbb{R}$, $\alpha \neq 0$. We have

$$\boxed{\lim_{x\to 0} \frac{(1+x)^\alpha - 1}{x} = \alpha.} \tag{5.5.4}$$

Let $x_n \to 0$. Then

$$(1+x_n)^\alpha = e^{\alpha \log(1+x_n)}.$$

Set $y_n := \alpha \log(1+x_n)$ so that $y_n \to 0$. Then

$$\frac{(1+x_n)^\alpha - 1}{x_n} = \frac{e^{y_n}-1}{y_n} \cdot \frac{y_n}{x_n} = \frac{e^{y_n}-1}{y_n} \cdot \frac{\alpha \log(1+x_n)}{x_n}.$$

Using (5.5.3) we deduce

$$\frac{e^{y_n}-1}{y_n} \to 1,$$

and using (5.5.2) we deduce

$$\frac{\alpha \log(1+x_n)}{x_n} \to \alpha.$$

This shows that

$$\frac{(1+x_n)^\alpha - 1}{x_n} \to \alpha.$$ □

Here is a typical application of the equality (5.5.1).

Example 5.5.5. Let us compute

$$\lim_{x\to\infty} \left(1 + \frac{x}{x^2+1}\right)^{2x}.$$

For any sequence $x_n \to \infty$ we have to compute

$$\lim_{n\to\infty} \left(1 + \frac{x_n}{x_n^2+1}\right)^{2x_n}.$$

Set

$$y_n := \frac{x_n}{x_n^2+1}.$$

Note that $y_n \to 0$ as $n \to \infty$ so that

$$e = \lim_{y\to 0}(1+y)^{\frac{1}{y}} = \lim_{n\to\infty}(1+y_n)^{\frac{1}{y_n}}.$$

We first seek to express $2x_n$ in the form

$$2x_n = \frac{s_n}{y_n} \Longleftrightarrow s_n = 2x_n y_n = \frac{2x_n^2}{x_n^2+1}.$$

Note that $s_n \to 2$ as $n \to \infty$. We deduce

$$\left(1+\frac{x_n}{x_n^2+1}\right)^{2x_n} = (1+y_n)^{\frac{s_n}{y_n}} = \Big((1+y_n)^{\frac{1}{y_n}}\Big)^{s_n},$$

so that

$$\lim_{n\to\infty}\left(1+\frac{x_n}{x_n^2+1}\right)^{2x_n} = \lim_{n\to\infty}\Big((1+y_n)^{\frac{1}{y_n}}\Big)^{s_n}$$

(use Theorem 5.2.14)

$$= \Big(\lim_n (1+y_n)^{\frac{1}{y_n}}\Big)^{\lim_n s_n} = e^2.$$

□

5.6. Trigonometric functions: a less than completely rigorous definition

Recall that the Cartesian product $\mathbb{R}^2 := \mathbb{R}\times\mathbb{R}$ is called the Cartesian plane and can be visualized as an Euclidean plane equipped with two perpendicular coordinate axes, the x-axis and the y-axis; see Figure 5.5. We can locate a point P in this plane if we can locate its projections P_x and P_y respectively, on the x- and the y-axis respectively; see Figure 5.5. The locations of these projections are indicated by two numbers, the x-coordinate and the y-coordinate respectively, of P. The point with coordinates $(0,0)$ is called the origin and it is denoted by O.

The trigonometric circle is the circle of radius 1 centered at the origin; see Figure 5.5. More precisely, a point with coordinates (x,y) lies on this circle if and only if

$$x^2+y^2=1. \tag{5.6.1}$$

Additionally, we agree that this circle is given an *orientation*, i.e., a prescribed way of traveling around it. In mathematics, the agreed upon orientation is *counterclockwise orientation* indicated by the arrow along the circle in Figure 5.5.

The starting point of the trigonometric circle is the point S with coordinates $(1,0)$. It can alternatively be described as the intersection of the circle with the positive side of the x-axis. The length[2] of the upper semi-circle is a positive number know by its famous name, π. In particular, the total length of the circle is 2π.

Suppose that we start at the point S and we travel along the circle, in the counterclockwise direction a distance $\theta \geqslant 0$. We denote by P the final point of this journey. The coordinates of this point depend only on the distance θ traveled. The x-coordinate of P

[2] We have surreptitiously avoided explaining what length means.

is denoted by $\cos\theta$, and the y-coordinate of P is denoted by $\sin\theta$. The equality (5.6.1) implies that

$$\cos^2\theta + \sin^2\theta = 1, \ \ \forall\theta \geqslant 0. \tag{5.6.2}$$

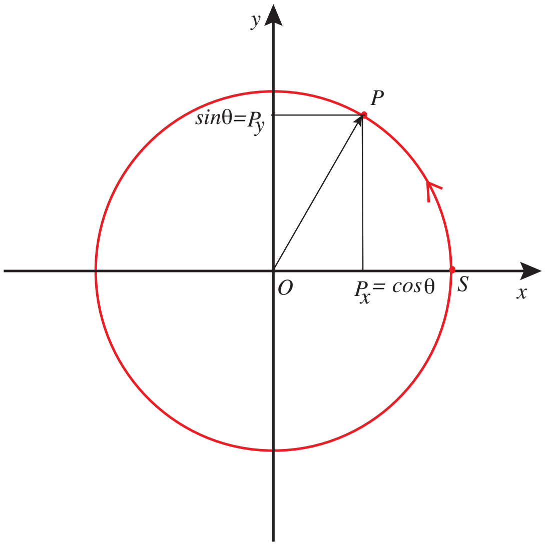

Figure 5.5. *The trigonometric circle. The distance of the journey from S to P in the counterclockwise direction is θ.*

Observe that if we continue our journey from P in the counterclockwise direction for a distance 2π then we are back at P. This shows that

$$\cos(\theta + 2\pi) = \cos\theta, \ \ \sin(\theta + 2\pi) = \sin\theta, \ \ \forall\theta \geqslant 0. \tag{5.6.3}$$

We can define $\cos\theta$ and $\sin\theta$ for negative θ's as well. Suppose that $\theta = -\phi$, $\phi \geqslant 0$. If we start at S and travel along the circle in the *clockwise* direction a distance ϕ, then we reach a point Q. By definition, its coordinates are $\cos(-\phi)$ and $\sin(-\phi)$; see Figure 5.6.

From the description it is easily seen that

$$\cos(-\phi) = \cos\phi, \ \ \sin(-\phi) = -\sin\phi, \ \ \forall\phi \geqslant 0. \tag{5.6.4}$$

We have thus constructed two functions

$$\cos, \sin : \mathbb{R} \to \mathbb{R},$$

called *trigonometric functions*. Their graphs are depicted in Figure 5.7 and 5.8.

Let us record a few important values of these functions.

We list below some of the more elementary, but very important, properties of the trigonometric functions sin and cos.

$$\cos^2 x + \sin^2 x = 1, \ \ \forall x \in \mathbb{R}. \tag{5.6.5a}$$

$$\cos(x + 2\pi) = \cos x, \ \ \sin(x + 2\pi) = \sin x, \ \ \forall x \in \mathbb{R}. \tag{5.6.5b}$$

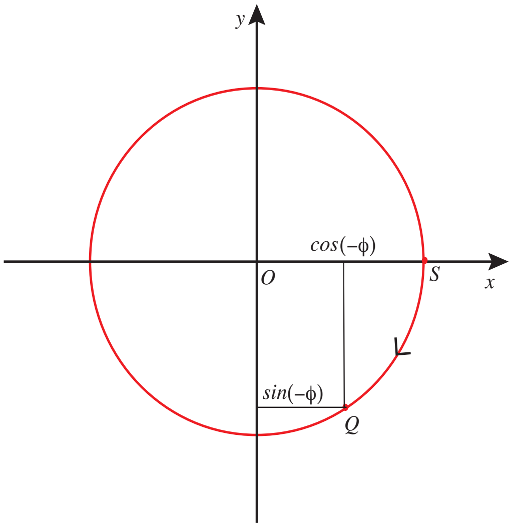

Figure 5.6. *The trigonometric circle. The distance of the journey from S to Q in the* ***clockwise*** *direction is ϕ.*

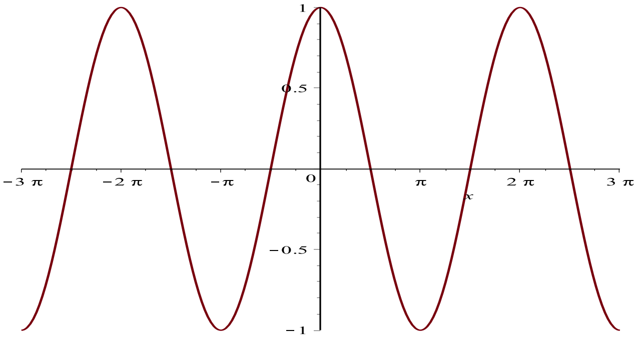

Figure 5.7. *The graph of* $\cos x$.

Table 5.1. Some important values of trig functions

θ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	π	2π
$\cos\theta$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	-1	1
$\sin\theta$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1	0	0

$$\cos(-x) = \cos x, \quad \sin(-x) = -\sin(x), \quad \forall x \in \mathbb{R}. \tag{5.6.5c}$$

$$\cos(x+\pi) = -\cos(x), \quad \sin(x+\pi) = -\sin(x), \quad \forall x \in \mathbb{R}, \tag{5.6.5d}$$

$$\sin\left(x + \frac{\pi}{2}\right) = \cos x, \quad \forall x \in \mathbb{R}. \tag{5.6.5e}$$

$$|\cos x| \leqslant 1, \quad |\sin x| \leqslant 1, \quad \forall x \in \mathbb{R}. \tag{5.6.5f}$$

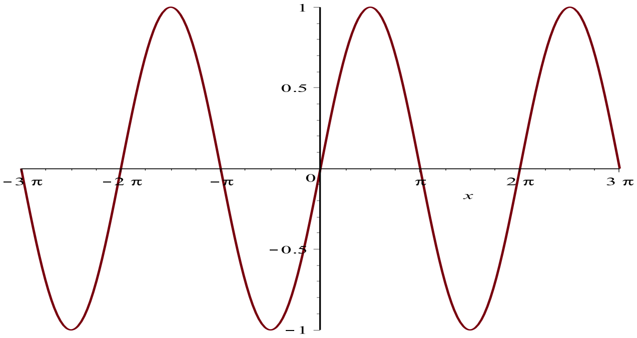

Figure 5.8. *The graph of* $\sin x$.

$$\cos x > 0,\quad \forall x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \text{ and } \sin x > 0,\quad \forall x \in (0, \pi). \tag{5.6.5g}$$

$$\cos x = 0 \Longleftrightarrow x \text{ is an odd multiple of } \tfrac{\pi}{2},\quad \sin x = 0 \Longleftrightarrow x \text{ is a multiple of } \pi. \tag{5.6.5h}$$

Definition 5.6.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a real valued function defined on the real axis $\mathbb{R}$.

(i) The function f is called *even* if

$$f(-x) = f(x),\quad \forall x \in \mathbb{R}.$$

(ii) The function f is called *odd* if

$$f(-x) = -f(x),\quad \forall x \in \mathbb{R}.$$

(iii) Suppose P is a positive real number. We say that f is P-*periodic* if

$$f(x + P) = f(x),\quad \forall x \in \mathbb{R}.$$

(iv) The function f is called periodic if there exists $P > 0$ such that f is P-periodic. Such a number P is called a *period* of f.

□

We see that the functions $\cos x$ and $\sin x$ are 2π-periodic, $\cos x$ is even, and $\sin x$ is odd.

In applications, we often rely on other trigonometric functions derived from sin and cos. We define

$$\tan x = \frac{\sin x}{\cos x},\quad \text{whenever } \cos x \neq 0,$$
$$\cot x = \frac{\cos x}{\sin x},\quad \text{whenever } \sin x \neq 0.$$

The graphs of $\tan x$ and $\cot x$ are depicted in Figure 5.9 and 5.10.

Example 5.6.2. We want to outline a geometric explanation for an important limit.

$$\boxed{\lim_{x \to 0} \frac{\sin x}{x} = 1.} \tag{5.6.6}$$

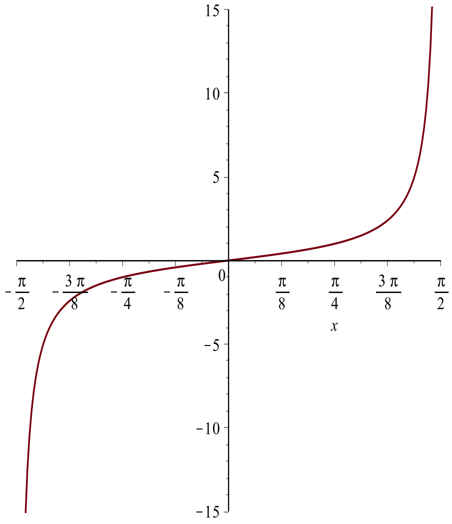

Figure 5.9. *The graph of* $\tan x$ *for* $x \in (-\pi/2, \pi/2)$.

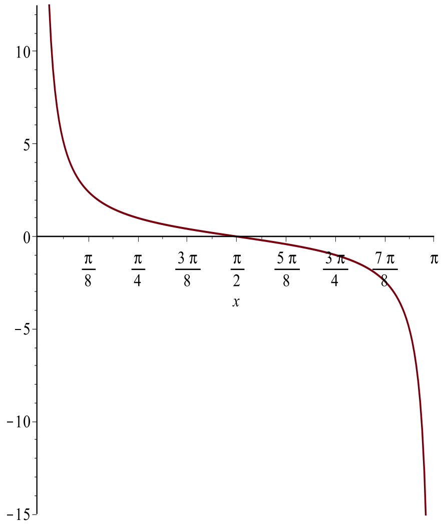

Figure 5.10. *The graph of* $\cot x$ *for* $x \in (0, \pi)$.

We will prove that

$$\lim_{x \nearrow 0} \frac{\sin x}{x} = \lim_{x \searrow 0} \frac{\sin x}{x} = 1.$$

Since

$$\frac{\sin x}{x} = \frac{\sin(-x)}{-x}$$

it suffices to prove only that

$$\lim_{x\searrow 0}\frac{\sin x}{x}=1. \tag{5.6.7}$$

This will follow immediately from the following fundamental inequalities

$$\theta\cos^2\theta\leqslant\sin\theta\leqslant\theta,\;\;\forall 0<\theta<\frac{\pi}{2}. \tag{5.6.8}$$

Let us temporarily take for granted these inequalities and show how they imply (5.6.7).

Observe that (5.6.8) implies that

$$0\leqslant\sin\theta\leqslant\theta,\;\;\forall 0<\theta<\frac{\pi}{2}.$$

The Squeezing Principle shows that

$$\lim_{\theta\searrow 0}\sin\theta=0. \tag{5.6.9}$$

This shows that the limit

$$\lim_{\theta\searrow 0}\frac{\sin\theta}{\theta}$$

is a bad limit of the type $\frac{0}{0}$. We can rewrite (5.6.8) as

$$1-\sin^2\theta=\cos^2\theta\leqslant\frac{\sin\theta}{\theta}\leqslant 1. \tag{5.6.10}$$

The equality (5.6.9) shows that

$$\lim_{\theta\searrow 0}(1-\sin^2\theta)=1.$$

The equality (5.6.8) now follows by applying the Squeezing Principle to the inequalities (5.6.10).

"Proof" of (5.6.8). Fix θ, $0<\theta<\frac{\pi}{2}$. We denote by P the point on the trigonometric circle reached from S by traveling a distance θ in the counterclockwise direction; see Figure 5.11. Denote by Q the projection of P onto the x-axis. We have

$$|OQ|=\cos\theta,\;\;|PQ|=\sin\theta.$$

Denote by M the intersection of the line OP with the circle centered at O and radius $|OP|=\cos\theta$. We distinguish three regions in Figure 5.12: the circular sector (OQM), the triangle $\triangle OSP$, and the circular sector (OSP). Clearly

$$(OQM)\subset\triangle OSP\subset(OSP)$$

so that we obtain inequalities between their areas[3]

$$\text{area}\,(OQM)\leqslant\text{area}\,\triangle OSP\leqslant\text{area}\,(OSP)$$

The area of a circular sector is[4]

$$\frac{1}{2}\times\text{square of the radius of the sector}\times\text{the size of the angle of the sector}. \tag{5.6.11}$$

We have

$$\text{area}\,(OQM)=\frac{1}{2}|OQ|^2\theta=\frac{\theta\cos^2\theta}{2},\;\;\text{area}\,(OSP)=\frac{1}{2}|OS|^2\theta=\frac{\theta}{2},$$
$$\text{area}\,\triangle OSP=\frac{1}{2}|PQ|\times|OS|=\frac{1}{2}\sin\theta.$$

[3] At this point we do not have a rigorous definition of the area of a planar region.

[4] The equality (5.6.11) needs a justification

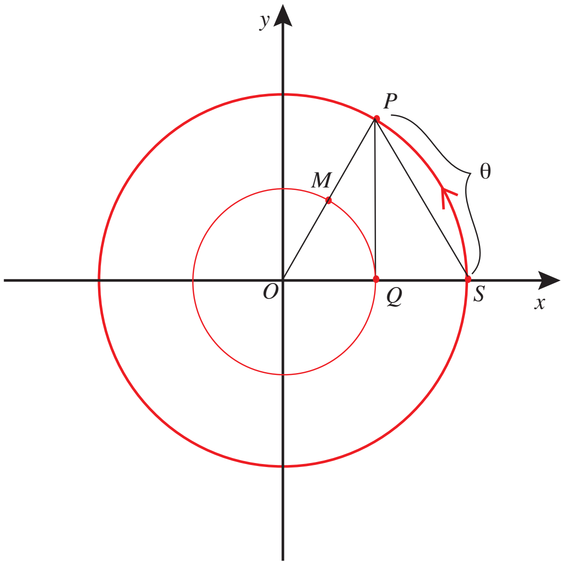

Figure 5.11. *The trigonometric circle. The distance of the journey from S to P in the* ***counterclockwise*** *direction is θ.*

Hence,

$$\frac{\theta\cos^2\theta}{2} \leqslant \frac{1}{2}\sin\theta \leqslant \frac{\theta}{2}. \qquad \square$$

5.7. Useful trig identities.

We list here a few important trigonometric identities that we will use in the future.

$$\sin(x \pm y) = \sin x\cos y \pm \sin y\cos x, \quad \cos(x \pm y) = \cos x\cos y \mp \sin x\sin y. \tag{5.7.1a}$$

$$\sin 2x = 2\sin x\cos x, \quad \cos 2x = \cos^2 x - \sin^2 x. \tag{5.7.1b}$$

$$\frac{1+\cos x}{2} = \cos^2(x/2), \quad \frac{1-\cos x}{2} = \sin^2(x/2). \tag{5.7.1c}$$

$$\cos x\cos y = \frac{1}{2}\big(\cos(x-y)+\cos(x+y)\big), \quad \sin x\sin y = \frac{1}{2}\big(\cos(x-y)-\cos(x+y)\big). \tag{5.7.1d}$$

$$\sin x\cos y = \frac{1}{2}\big(\sin(x+y)+\sin(x-y)\big) \tag{5.7.1e}$$

$$\tan(x \pm y) = \frac{\tan x + \tan y}{1 \mp \tan x\tan y}. \tag{5.7.1f}$$

5.8. Landau's notation

Let $c \in [-\infty, \infty]$ and consider two real valued functions f, g defined on the same set $X \subset \mathbb{R}$ which admits c as a cluster point. We say that

$$f(x) = O\big(\, g(x)\,\big) \;\text{ as } x \to c \tag{5.8.1}$$

if there exists a positive constant C and a neighborhood U of c such that

$$\forall x \in X,\;\; x \in (X \cap U)\backslash\{c\} \Rightarrow |f(x)| \leqslant C|g(x)|.$$

For example,

$$\frac{x}{x^2+1} = O\left(\frac{1}{x}\right) \;\text{ as }\; x \to \infty.$$

We say that

$$f(x) = o\big(\, g(x)\,\big) \;\text{ as } x \to c, \tag{5.8.2}$$

if, for any $\varepsilon > 0$, there exists a neighborhood U_ε of c such that

$$\forall x \in X,\;\; x \in U_\varepsilon\backslash\{c\} \Rightarrow |f(x)| \leqslant \varepsilon|g(x)|.$$

If $g(x) \neq 0$ for any x in a neighborhood U of c, then

$$f(x) = o\big(\, g(x)\,\big) \;\text{ as } x \to c \Longleftrightarrow \lim_{x\to c}\frac{f(x)}{g(x)} = 0.$$

Loosely speaking, this means that $f(x)$ is much, much smaller than $g(x)$ as x approaches c. For example,

$$e^{-x} = o(x^{-25}) \;\text{ as }\; x \to \infty,$$

and

$$x^3 = o(x^2) \;\text{ as }\; x \to 0.$$

However

$$x^2 = o(x^3) \;\text{ as }\; x \to \infty.$$

Finally, we say that f is *similar* to $g(x)$ as $x \to c$, and we write this

$$f(x) \sim g(x) \;\text{ as } x \to c$$

if

$$\lim_{x\to c}\frac{f(x)}{g(x)} = 1.$$

For example

$$x^3 - 39x^2 + 17 \;\sim\; x^3 + 3x^2 + 2x + 1 \;\text{ as }\; x \to \infty,$$

and

$$e^x - 1 \sim x \;\text{ as }\; x \to 0.$$

5.9. Exercises

Exercise 5.1. Prove that any real number is a cluster point of the set of rational numbers. □

Exercise 5.2. Prove Proposition 5.1.3. □

Exercise 5.3 (Squeezing principle)**.** Let $f, g, h : X \to \mathbb{R}$ be three functions defined on the same subset $X \subset \mathbb{R}$ and c a cluster point of X. Suppose that U is a deleted neighborhood of c such that

$$f(x) \leqslant h(x) \leqslant g(x), \;\; \forall x \in U \cap X.$$

Show that if

$$\lim_{x\to c} f(x) = A = \lim_{x\to c} g(x),$$

then

$$\lim_{x\to c} h(x) = A.$$

□

Exercise 5.4. Consider a subset $X \subset \mathbb{R}$, a function $f : X \to \mathbb{R}$, and a cluster point c of the set X. Prove that the following statements are equivalent.

(i) The limit $\lim_{x\to c} f(x)$ exists and it is finite.

(ii) For any sequence $(x_n)_{n\in\mathbb{N}} \subset X$ such that $x_n \to c$ and $x_n \neq c$, $\forall n \in \mathbb{N}$ the sequence $\big(f(x_n) \big)_{n\in\mathbb{N}}$ is convergent.

□

Exercise 5.5. Let $I \subset \mathbb{R}$ be an interval and $f : I \to \mathbb{R}$ a function. Suppose that f is a *Lipschitz function*, i.e., there exists a constant L such that

$$|f(x) - f(y)| \leqslant L|x - y|, \;\; \forall x, y \in I.$$

Show that for any $y \in Y$ we have

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

□

Exercise 5.6. We already know that the series

$$\sum_{n\geqslant 1} \frac{1}{n^s}$$

converges for any *rational number* $s > 1$. Prove that it converges for any *real number* $s > 1$. □

Exercise 5.7. (a) Prove that for any $n \in \mathbb{N}$ we have

$$\lim_{x\to\infty} \frac{1}{x^n} = 0.$$

(b) Let $k \in \mathbb{N}$ and consider the function $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}$, $f(x) = \frac{1}{x^{2k}}$. Show that

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

□

Exercise 5.8. Fix a natural number n. Consider the polynomial

$$P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Show that

$$\lim_{x\to\infty} P(x) = \infty, \quad \lim_{x\to-\infty} P(x) = \begin{cases} \infty, & n \text{ is even} \\ -\infty, & n \text{ is odd}. \end{cases}$$ □

Exercise 5.9. Consider two convergent sequences of real numbers $(x_n)_{n\geqslant 0}$, $(y_n)_{n\geqslant 0}$. We set

$$x := \lim_{n\to\infty} x_n, \quad y := \lim_{n\to\infty} y_n.$$

Show that if $x_n > 0$, $\forall n \geqslant 0$ and $x > 0$ then

$$\lim_{n\to\infty} x_n^{y_n} = x^y.$$

Hint. Use the same strategy as in the proof of Theorem 5.2.14. □

Exercise 5.10. Prove that

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x, \quad \forall x \in \mathbb{R}.$$

Hint: Use the result in Example 5.3.6 and Theorem 5.2.14. □

Exercise 5.11. Fix an arbitrary number $a > 1$.

(a) Prove that for any $x > 1$ we have

$$a^x \geqslant a^{\lfloor x\rfloor} \geqslant 1 + (a-1)\lfloor x\rfloor + \binom{\lfloor x\rfloor}{2}(a-1)^2.$$

(b) Prove that

$$\lim_{x\to\infty}\frac{x}{a^x} = 0, \quad \lim_{x\to\infty}\frac{a^x}{x} = \infty.$$

Hint. Use (a) and Example 4.3.4.

(c) Let $r > 0$. Prove that

$$\lim_{x\to\infty}\frac{x^r}{a^x} = 0.$$

Hint. Reduce to (b).

(d) Prove that

$$\lim_{x\to\infty}\frac{\log_a x}{x} = 0.$$

Hint. Reduce to (c). □

Exercise 5.12. Fix a positive real number s and consider the function $f : (0,\infty) \to \mathbb{R}$, $f(x) = x^s$.

(a) Show that f is an increasing function.

(b) Show that

$$\lim_{x \searrow 0} x^s = 0.$$

Exercise 5.13. Let $a, b \in \mathbb{R}$, $a < b$. Prove that if $f : (a, b) \to \mathbb{R}$ is a nondecreasing function and $x_0 \in (a, b)$, then the one sided limits

$$\lim_{x \nearrow x_0} f(x) \quad \text{and} \quad \lim_{x \searrow x_0} f(x)$$

exist and

$$\lim_{x \nearrow x_0} f(x) = \sup_{x < x_0} f(x), \quad \lim_{x \searrow x_0} f(x) = \inf_{x > x_0} f(x).$$ □

Exercise 5.14. Consider the function

$$f : \mathbb{R} \backslash \{0\} \to \mathbb{R}, \quad f(x) = x \sin\left(\frac{1}{x}\right).$$

Prove that

$$\lim_{x \to 0} f(x) = 0.$$ □

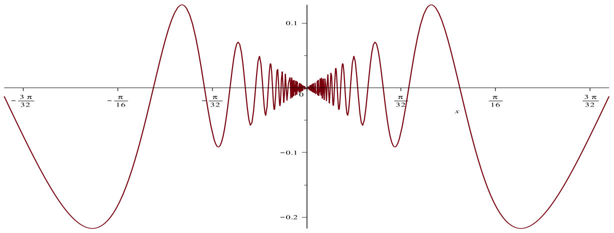

Figure 5.12. *The graph of* $x \sin(1/x)$ *for* $|x| < \pi/10$.

Exercise 5.15. Consider $f : [0, \infty) \to \mathbb{R}$,

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

Show that

$$f(x) = O(1) \ \text{ as } x \to \infty.$$

Above, we used Landau's notation introduced in section 5.8. □

Exercise 5.16. (a) Prove Lemma 5.2.8.

Hint. The case $a = 1$ is trivial. In the case $a > 1$ show that there exists a sequence of positive rational numbers (r_n) such that

$$-r_n \leqslant x_n - x \leqslant r_n, \quad \forall n.$$

Now use Lemma 5.2.2, Lemma 5.2.5, and the Squeezing Principle to conclude. The case $a < 1$ follows from the case $a > 1$.

(b) Prove the equality (5.2.3).

Hint. First prove that (5.2.3) holds for any $x \in \mathbb{Q}$. Then conclude using Lemma 5.2.8. □

Exercise 5.17. Prove Proposition 5.3.2. □

Exercise 5.18. Prove Theorem 5.3.3. □

Exercise 5.19. Prove Theorem 5.4.2. □

Exercise 5.20. Prove Theorem 5.4.3. □

5.10. Exercises for extra credit

Exercise* 5.1. Consider the sequence $(x_n)_{n\geqslant 0}$ defined by

$$x_0 = 0, \quad x_{n+1} = \sqrt{\frac{1+x_n}{2}}, \quad \forall n \geqslant 0.$$

(a) Prove that

$$x_n = \cos\frac{\pi}{2^{n+1}}, \quad \forall n \geqslant 0.$$

(b) Prove that

$$\lim_{n\to\infty}(x_1 \cdot x_2 \cdots x_n) = \frac{2}{\pi}.$$ □

Exercise* 5.2. Suppose that

$$\sum_{n\geqslant 0} a_n$$

is a convergent series of real numbers. We denote by a its sum.

(i) Show that for any $x \in (-1,1)$ the series

$$\sum_{n\geqslant 0} a_n x^n$$

is convergent. For $x \in (-1,1)$ we denote by $A(x)$ the sum of the above series.

(ii) Prove that

$$\lim_{x\nearrow 1} A(x) = a.$$

□

Chapter 6

Continuity

6.1. Definition and examples

The concept of continuity is a fundamental mathematical concept with a wide range of applications.

Definition 6.1.1. Suppose that $X \subset \mathbb{R}$ and $f : X \to \mathbb{R}$ is a real valued function defined on X. We say that the function f is *continuous at a point $x_0 \in X$* if

$$\forall \varepsilon > 0 \ \exists \delta = \delta(\varepsilon) > 0 \ \text{such that} \ \forall x \in X \ \ |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon.$$

We say that the function f is *continuous (on X)* if it is continuous at every point $x_0 \in X$. □

Arguing as in the proof of Theorem 5.1.4 we obtain the following very useful alternate characterization of continuity. The details are left to you as an exercise.

Theorem 6.1.2. *Let $X \subset \mathbb{R}$, $x_0 \in X$, and $f : X \to \mathbb{R}$ a real valued function on X. The following statements are equivalent.*

(i) *The function f is continuous at x_0.*

(ii) *For any sequence $(x_n)_{n\in\mathbb{N}}$ in X such that $x_n \to x_0$, we have $\lim_n f(x_n) = f(x_0)$.*

□

We have the following useful consequence which relates the concept of continuity to the concept of limit. Its proof is left to you as an exercise.

Corollary 6.1.3. *Let $X \subset \mathbb{R}$ and $f : X \to \mathbb{R}$. Suppose that $x_0 \in X$ is a cluster point of X. Then the following statements are equivalent.*

(i) *The function f is continuous at x_0.*

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

□

We have already encountered many examples of continuous functions.

Example 6.1.4. (a) Let $k \in \mathbb{N}$. Then the function

$$f : \mathbb{R} \to \mathbb{R}, \;\; f(x) = x^k, \;\; \forall x \in \mathbb{R},$$

is continuous on its domain $\mathbb{R}$. Indeed, if $x_0 \in \mathbb{R}$ and $(x_n)_{n \in \mathbb{R}}$ is a sequence of real numbers such that $x_n \to x_0$, then Proposition 4.3.1 implies that

$$x_n^k \to x_0^k,$$

thus proving the continuity of f at an arbitrary point $x_0 \in \mathbb{R}$.

(b) A similar argument shows that if $k \in \mathbb{N}$, then the function

$$f : \mathbb{R} \backslash \{0\} \to \mathbb{R}, \;\; f(x) = \frac{1}{x^k}, \;\; \forall x \in \mathbb{R} \backslash \{0\},$$

is continuous.

(c) Fix $s \in \mathbb{R}$. Then the function

$$f : (0, \infty) \to \mathbb{R}, \;\; f(x) = x^s, \;\; \forall x > 0,$$

is continuous. Indeed, this follows by invoking Theorems 5.2.14 and 6.1.2.

(d) Let $a > 0$. Then the functions

$$f : \mathbb{R} \to (0, \infty), \;\; f(x) = a^x,$$

and

$$g : (0, \infty) \to \mathbb{R}, \;\; g(x) = \log_a x,$$

are continuous on their domains. Indeed, the continuity of f follows from Lemma 5.2.8, while the continuity of g follows from Theorem 5.2.13.

(e) The trigonometric functions

$$\sin, \cos : \mathbb{R} \to \mathbb{R}$$

are continuous.

Let us first prove that these functions are continuous at $x_0 = 0$. The continuity of sin at $x_0 = 0$ follows immediately from (5.6.9) and Corollary 6.1.3. To prove the continuity of cos at $x_0 = 0$ we have to show that if (x_n) is a sequence of real numbers such that $x_n \to 0$, then $\cos x_n \to \cos 0 = 1$.

Let (x_n) be a sequence of real numbers converging to zero. Then

$$\cos^2 x_n = 1 - \sin^2 x_n$$

and we deduce that

$$\lim_n \cos^2 x_n = 1 - \sin^2 x_n = \lim_n (1 - \sin^2 x_n) = 1.$$

Since $x_n \to 0$, we deduce that there exists $N_0 > 0$ such that $|x_n| < \frac{\pi}{2}$, $\forall n > N_0$. The inequalities (5.6.5g) imply that

$$\cos x_n > 0, \;\; \forall n >_0,$$

so that,

$$\cos x_n = \sqrt{1 - \sin^2 x_n}, \;\; \forall n > N_0.$$

Exercise 4.15 now implies that

$$\lim_n \cos x_n = \sqrt{\lim_n (1 - \sin^2 x_n)} = \sqrt{1} = \cos 0.$$

We can now prove the continuity of sin and cos at an arbitrary point x_0. Suppose that x_n is a sequence of real numbers such that $x_n \to x_0$. We have to show that

$$\lim_n \sin x_n = \sin x_0 \;\text{ and }\; \lim_n \cos x_n = \cos x_0.$$

We set $h_n = x_n - x_0$, so that, $x_n = x_0 + h$. Then

$$\sin x_n = \sin(x_0 + h_n) \overset{(5.7.1a)}{=} \sin x_0 \cos h_n + \sin h_n \cos x_0$$

and

$$\cos x_n = \cos(x_0 + h_n) \overset{(5.7.1a)}{=} \cos x_0 \cos h_n - \sin x_0 \sin h_n.$$

Observe that $h_n \to 0$ and, since sin and cos are continuous at 0, we have $\sin h_n \to 0$ and $\cos h_n \to 1$. We deduce

$$\lim_n \sin x_n = \lim_n \sin(x_0 + h_n) = \sin x_0 \lim_n \cos h_n + \cos x_0 \lim_n \sin h_n = \sin x_0$$

and

$$\lim_n \cos x_n = \lim_n \cos(x_0 + h_n) = \cos x_0 \lim_n \cos h_n - \sin x_0 \lim_n \sin h_n = \cos x_0.$$

(f) Recall that a function $f : X \to \mathbb{R}$, $X \subset \mathbb{R}$ is called *Lipschitz* if

$$\exists L > 0 : \;\; \forall x_1, x_2 \in X \;\; |f(x_1) - f(x_2)| \leqslant L|x_1 - x_2|.$$

Observe that a Lipschitz function is necessarily continuous. Indeed, if $x_0 \in X$ and (x_n) is a sequence in X such that $x_n \to x_0$ then

$$|f(x_n) - f(x_0)| \leqslant L|x_n - x_0| \to 0,$$

and the squeezing principle implies that $f(x_n) \to f(x_0)$.

Observe that the absolute value function $f : \mathbb{R} \to [0, \infty)$, $f(x) = |x|$ is Lipschitz because of the following elementary inequality (see Exercise 4.5)

$$|f(x) - f(y)| = \big|\, |x| - |y| \,\big| \leqslant |x - y|, \;\; \forall x, y \in \mathbb{R}. \tag{6.1.1}$$

Thus the absolute value function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ is a continuous function. □

Proposition 6.1.5. *Let $X \subset \mathbb{R}$, $c \in \mathbb{R}$, and suppose that $f, g : X \to \mathbb{R}$ are two continuous functions. Then the functions*

$$f + g, cf, f \cdot g : X \to \mathbb{R},$$

are continuous. Additionally, if $\forall x \in X$ $g(x) \neq 0$, then the function

$$\frac{f}{g} : X \to \mathbb{R}$$

is also continuous.

Proof. This is an immediate consequence of Proposition 4.3.1 and Theorem 6.1.2. □

Example 6.1.6. Polynomial function $p : \mathbb{R} \to \mathbb{R}$ defined by

$$p(x) = c_n x^n + \cdots + c_1 x + c_0,$$

$n \in \mathbb{Z}$, $n \geqslant 0$, $c_0, \ldots, c_n \in \mathbb{R}$ are continuous. For example, the function $p(x) = x^3 - 2x + 5$, $x \in \mathbb{R}$, is continuous on $\mathbb{R}$.

We can easily get more complicated examples. Thus, the function $(x^3 - 2x + 5) \sin x$, $x \in \mathbb{R}$, is continuous, the function $e^x + e^{-x}$, $x \in \mathbb{R}$, is continuous and nowhere zero, so the quotient

$$\frac{(x^3 - 2x + 5) \sin x}{e^x + e^{-x}}, \quad x \in \mathbb{R}$$

is also continuous on $\mathbb{R}$. □

Proposition 6.1.7. *Suppose that $X, Y \subset \mathbb{R}$ and that $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ are continuous functions such that*

$$f(X) \subset Y.$$

Then the composition $g \circ f : X \to \mathbb{R}$, $g \circ f(x) = g(f(x))$, $\forall x \in X$ is also a continuous function.

Proof. Theorem 6.1.2 implies that we have to prove that for any $x_0 \in X$ and any sequence (x_n) in X such that $x_n \to x_0$ we have

$$g(f(x_n)) \to g(f(x_0)).$$

Set $y_0 := f(x_0)$, $y_n := f(x_n)$. Since f is continuous at x_0, Theorem 6.1.2 shows that $f(x_n) \to f(x_0)$, i.e., $y_n \to y_0$. Since g is continuous at y_0, Theorem 6.1.2 implies that $g(y_n) \to g(y_0)$, i.e.,

$$g(f(x_n)) \to g(f(x_0)).$$

□

Example 6.1.8. Consider the continuous functions

$$f, g, h : \mathbb{R} \to \mathbb{R}, \quad f(x) = \sin x, \quad g(x) = e^x, \quad h(x) = |x|.$$

Then $g \circ f(x) = e^{\sin x}$ is continuous on $\mathbb{R}$, and so is the function $f \circ g(x) = \sin e^x$. Similarly $f \circ h(x) = \sin |x|$ is a continuous function on $\mathbb{R}$. □

Definition 6.1.9. Let $X \subset \mathbb{R}$ be a set of real numbers and $f : X \to \mathbb{R}$ a real valued function on X.

(a) The sequence of functions $f_n : X \to \mathbb{R}$, $n \in \mathbb{N}$ is said to converge *pointwisely* to the function $f : X \to \mathbb{R}$ if

$$\lim_{n \to \infty} f_n(x) = f(x), \quad \forall x \in X,$$

i.e.,

$$\forall \varepsilon > 0, \quad \forall x \in X \;\; \exists N = N(\varepsilon, x): \quad \forall n > N(\varepsilon, x) \;\; |f_n(x) - f(x)| < \varepsilon. \tag{6.1.2}$$

(b) The sequence of functions $f_n : X \to \mathbb{R}, \quad n \in \mathbb{N}$ is said to *converge uniformly* to the function $f : X \to \mathbb{R}$ if

$$\forall \varepsilon > 0 \;\; \exists N = N(\varepsilon) > 0 \text{ such that } \forall n > N(\varepsilon), \quad \forall x \in X: \quad |f_n(x) - f(x)| < \varepsilon. \tag{6.1.3}$$

□

Theorem 6.1.10 (Continuity of uniform limits)**.** *Let $X \subset \mathbb{R}$ be a set of real numbers. If the sequence of* continuous *functions $f_n : X \to \mathbb{R}$, $n \in \mathbb{N}$, converges uniformly to the function $f : X \to \mathbb{R}$, then the limit function f is also continuous on X.*

Proof. We have to prove that given $x_0 \in X$ the function f is continuous at x_0, i.e., we have to show that

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 \;\; \forall x \in X \;\; |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon. \tag{6.1.4}$$

Let $\varepsilon > 0$. The uniform convergence implies that

$$\exists N(\varepsilon) > 0: \quad \forall x \in X, \quad \forall n > N(\varepsilon) \;\; |f_n(x) - f(x)| < \frac{\varepsilon}{3}. \tag{6.1.5}$$

Fix $n_0 > N(\varepsilon)$. The function f_{n_0} is continuous at x_0 and thus

$$\exists \delta(\varepsilon) > 0 \;\; \forall x \in X: \quad |x - x_0| < \delta(\varepsilon) \Rightarrow |f_{n_0}(x) - f_{n_0}(x_0)| < \frac{\varepsilon}{3}. \tag{6.1.6}$$

We deduce that if $|x - x_0| < \delta(\varepsilon)$, then

$$|f(x) - f(x_0)| \leqslant |f(x) - f_{n_0}(x)| + |f_{n_0}(x) - f_{n_0}(x_0)| + |f_{n_0}(x_0) - f(x_0)|. \tag{6.1.7}$$

From (6.1.5) we deduce that since $n_0 > N(\varepsilon)$ we have

$$|f(x) - f_{n_0}(x)|, \quad |f_{n_0}(x_0) - f(x_0)| < \frac{\varepsilon}{3}, \quad \forall x \in X.$$

From (6.1.6) we deduce that if $|x - x_0| < \delta(\varepsilon)$, then

$$|f_{n_0}(x) - f_{n_0}(x_0)| < \frac{\varepsilon}{3}.$$

Using these facts in (6.1.7) we deduce that if $|x - x_0| < \delta(\varepsilon)$, then

$$|f(x) - f(x_0)| < \varepsilon.$$

□

6.2. Fundamental properties of continuous functions

In this section we will discuss several fundamental properties of continuous functions, which hopefully will explain the usefulness of the concept of continuity.

Theorem 6.2.1. *Suppose that c is an arbitrary real number, $X \subset \mathbb{R}$ and $f : X \to \mathbb{R}$ is a function continuous at $x_0 \in X$.*

(a) If $x_0 \in X$ satisfies $f(x_0) < c$, then there exists $\delta > 0$ such that

$$\forall x \in X, \quad |x - x_0| < \delta \Rightarrow f(x) < c.$$

In other words, if $f(x_0) < c$, then for any $x \in X$ sufficiently close to x_0 we also have $f(x) < c$.

(b) If $x_0 \in X$ satisfies $f(x_0) > c$, then there exists $\delta > 0$ such that

$$\forall x \in X, \quad |x - x_0| < \delta \Rightarrow f(x) > c.$$

In other words, if $f(x_0) > c$, then for any $x \in X$ sufficiently close to x_0 we also have $f(x) > c$.

Proof. Fix $\varepsilon_0 > 0$, such that $f(x_0) + \varepsilon_0 < c$. (For example, we can choose $\varepsilon_0 = \frac{1}{2}(c - f(x_0)\,)$.)

The continuity of f at x_0 (Definition 6.1.1) implies that there exists $\delta_0 > 0$ such that for any $x \in X$ satisfying $|x - x_0| < \delta_0$ we have

$$|f(x) - f(x_0)| < \varepsilon_0,$$

so that

$$f(x_0) - \varepsilon_0 < f(x) < f(x_0) + \varepsilon_0 < c.$$

□

Corollary 6.2.2. *Suppose that $X \subset \mathbb{R}$, $x_0 \in X$ and $f : X \to \mathbb{R}$ is a continuous function such that $f(x_0) \neq 0$. Then there exists $\delta > 0$ such that*

$$\forall x \in X \quad (\,|x - x_0| < \delta \Rightarrow f(x) \neq 0\,).$$

In other words, if $f(x_0) \neq 0$, then for any $x \in X$ sufficiently close to x_0 we also have $f(x) \neq 0$.

Proof. Consider the function $g : X \to \mathbb{R}$, $g(x) = |f(x)|$. The function g is continuous because it is the composition of the absolute-value-function with the continuous function f. Additionally, $|g(x_0)| > 0$. The desired conclusion now follows from Theorem 6.2.1 (b). □

To state and prove our next result we need to make a small digression. Recall that the *Completeness Axiom* states that if the set $X \subset \mathbb{R}$ is *bounded above*, then it admits a least upper bound which is a *real number* denoted by $\sup X$. If the set X is not bounded above, then we define $\sup X := \infty$. Thus, we have given a meaning to $\sup X$ *for any* subset $X \subset \mathbb{R}$. Moreover,

$$\sup X < \infty \Longleftrightarrow \text{the set } X \text{ is bounded above.}$$

Similarly, we define $\inf X = -\infty$ for any set X that is not bounded below. Thus we have given a meaning to $\inf X$ *for any* subset $X \subset \mathbb{R}$. Moreover,

$$\inf X > -\infty \Longleftrightarrow \text{the set } X \text{ is bounded below.}$$

Lemma 6.2.3. *(a) If Y is a set of real numbers and $M = \sup Y \in (-\infty, \infty]$, then there exists an increasing sequence of real numbers $(M_n)_{n\geqslant 1}$ and a sequence (y_n) in Y such that*

$$M_n \leqslant y_n \leqslant M, \;\; \forall n, \;\; \lim_n M_n = M.$$

(b) If Y is a set of real numbers and $m = \inf Y \in [-\infty, \infty)$, then there exists a decreasing sequence of real numbers $(m_n)_{n\geqslant 1}$ and a sequence (y_n) in Y such that

$$m \leqslant y_n \leqslant m_n, \;\; \forall n, \;\; \lim_n m_n = m.$$

Proof. We prove only (a). The proof of (b) is very similar and it is left to you as an exercise. We distinguish two cases.

A. $M < \infty$. Since M is the least upper bound of Y, for any $n > 0$ there exists $y_n \in X$ such that

$$M - \frac{1}{n} \leqslant y_n \leqslant M.$$

The sequences (y_n) and $M_n = M - \frac{1}{n}$ have the desired properties.

B. $M = \infty$. Hence, the set Y is not bounded above. Thus, for any $n \in \mathbb{N}$ there exists $y_n \in Y$ such that $y_n \geqslant n$. The sequences (y_n) and $M_n = n$ have the desired properties. □

Theorem 6.2.4 (Weierstrass)**.** *Consider a continuous real valued function f defined on a **closed and bounded** interval $[a, b]$, i.e., $f : [a, b] \to \mathbb{R}$. Then the following hold.*

(i)
$$M := \sup\{\, f(x); \;\; x \in [a, b] \,\} < \infty.$$

(ii) $\exists x^* \in [a, b]$ *such that* $f(x^*) = M$.

(iii)
$$m := \inf\{\, f(x); \;\; x \in [a, b] \,\} > -\infty.$$

(iv) $\exists x_* \in [a, b]$ *such that* $f(x_*) = m$.

Proof. We prove only (i) and (ii). The proofs of statements (iii) and (iv) are similar. Denote by Y the range of the function f,

$$Y = \{\, f(x); \;\; x \in [a, b] \,\}.$$

Hence $M = \sup Y$. From Lemma 6.2.3 we deduce that there exists a sequence (y_n) in Y and an increasing sequence (M_n) such that

$$M_n \leqslant y_n \leqslant M, \;\; \lim_n M_n = M.$$

The Squeezing Principle implies that

$$\lim_n y_n = M. \tag{6.2.1}$$

Since y_n is in the range of f there exists $x_n \in [a,b]$ such that $f(x_n) = y_n$. The sequence (x_n) is obviously bounded because it is contained in the bounded interval $[a,b]$. The Bolzano-Weierstrass Theorem (Theorem 4.4.8) implies that (x_n) admits a subsequence (x_{n_k}) which converges to some number x^*

$$\lim_k x_{n_k} = x^*.$$

Since $a \leqslant x_{n_k} \leqslant b$, $\forall k$, we deduce that $x^* \in [a,b]$. The continuity of f implies that

$$\lim_k y_{n_k} = \lim_k f(x_{n_k}) = f(x^*).$$

On the other hand,

$$\lim_k y_{n_k} = \lim_n y_n \overset{(6.2.1)}{=} M.$$

Hence

$$M = f(x^*) < \infty.$$

□

Definition 6.2.5. Let $f : X \to \mathbb{R}$ be a function defined on a nonempty set $X \subset \mathbb{R}$.

(a) A point $\boldsymbol{x_*} \in X$ is called a *global minimum* of f if

$$f(x_*) \leqslant f(x), \;\; \forall x \in X.$$

(b) A point $\boldsymbol{x^*} \in X$ is called a *global maximum* of f if

$$f(x) \leqslant f(x^*), \;\; \forall x \in X.$$

□

We can rephrase Theorem 6.2.4 as follows.

Corollary 6.2.6. *A continuous function $f : [a,b] \to \mathbb{R}$ admits a global minimum and a global maximum.* □

Remark 6.2.7. The conclusions of Theorem 6.2.4 *do not necessarily hold* for continuous functions defined on *non-closed* intervals. Consider for example the continuous function

$$f : (0,1] \to \mathbb{R}, \;\; f(x) = \frac{1}{x}.$$

Note that $f(1/n) = n$, $\forall n \in \mathbb{N}$ so that

$$\sup\{\, f(x); \;\; x \in (0,1] \,\} = \infty.$$

□

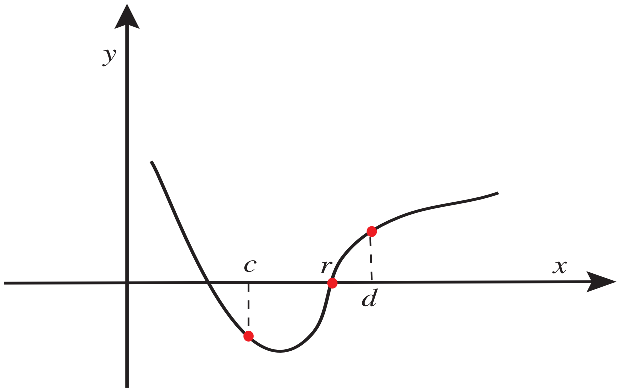

Figure 6.1. *If the graph of a continuous functions has points both below and above the x-axis, then the graph must intersect the x-axis.*

Theorem 6.2.8 (The intermediate value theorem). *Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function and $c, d \in [a,b]$ are real numbers such that*

$$c < d \ \text{ and } \ f(c) \cdot f(d) < 0.$$

Then there exists a real number $r \in (c,d)$ such that $f(r) = 0$.

Proof. We distinguish two cases: $f(c) < 0$ or $f(c) > 0$. We discuss only the case $f(c) < 0$ depicted in Figure 6.1. The second case follows from the first case applied to the continuous function $-f$. Observe that the assumption $f(c)f(d) < 0$ implies that if $f(c) < 0$, then $f(d) > 0$.

Consider the set

$$X := \{ x \in [c,d]; \ \ f(x) < 0 \}.$$

Clearly X is nonempty because $c \in X$. By construction, the set X is bounded above by d. Define

$$r := \sup X.$$

We will prove that $f(r) = 0$. Since $r = \sup X$, we deduce from Lemma 6.2.3 that there exists a sequence (x_n) in X such that $x_n \to r$ as $n \to \infty$. The function f is continuous at r so that

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

On the other hand $f(x_n) < 0$, for any n because $x_n \in X$. Hence $f(r) \leqslant 0$. In particular $r \neq d$ because $f(d) > 0$.

Figure 6.2. *The function f would be negative on $[r, r+\delta]$ if $f(r)$ were negative.*

To prove that $f(r) = 0$ it suffices to show that $f(r) \geqslant 0$. We argue by contradiction and we assume that $f(r) < 0$. Theorem 6.2.1 implies that there exists $\delta > 0$ such that if $x \in [a, b]$ and $|x - r| < \delta$, then $f(x) < 0$. Thus $f(x) < 0$ for any $x \in [a, b] \cap [r, r + \delta]$; Figure 6.2.

Choose $h > 0$ such that

$$h < \min\{\, \delta, \operatorname{dist}(r, d) \,\}.$$

Then $r + h \in [r, d]$ and $r + h \in [r, r + \delta]$. Hence $r + h \in [c, d]$ and $f(r + h) < 0$ so that $r + h \in X$. This contradicts the fact that $r = \sup X$. □

The Intermediate Value Theorem has many useful consequences. We present a few of them.

Corollary 6.2.9. *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function, $y_0 \in \mathbb{R}$ and $c \leqslant d$ are real numbers in the interval $[a, b]$ such that*

- *either $f(c) \leqslant y_0 \leqslant f(d)$, or*
- *$f(c) \geqslant y_0 \geqslant f(d)$.*

Then there exists $x_0 \in [c, d]$ such that $f(x_0) = y_0$.

Proof. If $f(c) = y_0$ or $f(d) = y_0$, then there is nothing to prove so we assume that $f(c), f(d) \neq y_0$. Consider the function $g : [a, b] \to \mathbb{R}$, $g(x) = f(x) - y_0$. Then $g(c)g(d) < 0$, and the Intermediate Value Theorem implies that there exists $x_0 \in (c, d)$ such that $g(x_0) = 0$, i.e., $f(x_0) = y_0$. □

Corollary 6.2.10. *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function and $c < d$ are real numbers in the interval $[a, b]$ such that*

$$f(x) \neq 0, \;\; \forall x \in (c, d).$$

Then the function f does not change sign in the interval (c, d), i.e., either

$$f(x) > 0, \;\; \forall x \in (c, d),$$

or

$$f(x) < 0, \;\; \forall x \in (c, d).$$

Proof. If f did change sign in the interval (c, d), then we could find two numbers $c', d' \in (c, d)$ such that $f(c') < 0$ and $f(d') > 0$. The Intermediate Value Theorem will then imply that f must equal zero at some point r situated between c' and d'. This would contradict the assumptions on f. □

Corollary 6.2.11. *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function,*

$$M = \sup\{\, f(x); \;\; x \in [a, b] \,\}, \;\; m = \inf\{\, f(x); \;\; x \in [a, b] \,\}.$$

Then the range of the function f is the interval $[m, M]$.

Proof. Observe first that

$$m \leqslant f(x) \leqslant M, \quad \forall x \in [a,b].$$

This shows that the range of f is contained in the interval $[m,M]$. Let us now prove the opposite inclusion, i.e., $[m,M]$ is contained in the range of f. More precisely, we need to show that for any $y_0 \in [m,M]$ there exists $x_0 \in [a,b]$ such that $f(x_0) = y_0$.

Observe first that Weierstrass' Theorem 6.2.4 implies that m, M belong to the range of f. In particular, there exist $c, d \in [a,b]$ such that $f(c) = m$ and $f(d) = M$. In particular,

$$f(c) \leqslant y_0 \leqslant f(d).$$

Corollary 6.2.9 implies that there exists a number x_0 situated between c and d such that $f(x_0) = y_0$. □

Corollary 6.2.12. *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuous function such that*

$$\lim_{x\to\infty} f(x) \in (0,\infty] \quad \lim_{x\to-\infty} f(x) \in [-\infty, 0).$$

Then there exists $r \in \mathbb{R}$ such that $f(r) = 0$. □

The proof of this corollary is left to you as an exercise.

Corollary 6.2.13. *Suppose that $a < b$ and $f : [a,b] \to \mathbb{R}$ is a continuous function. Then the following statements are equivalent,*

(i) *The function f is injective.*

(ii) *The function f is strictly monotone; see Definition 5.2.10(v).*

Proof. The implication (ii) $\Rightarrow$ (i) is immediate. Indeed, suppose $x_1, x_2 \in [a,b]$ and $x_1 \neq x_2$. One of the numbers x_1, x_2 is smaller than the other and we can assume $x_1 < x_2$. If f is strictly increasing, then $f(x_1) < f(x_2)$, thus $f(x_1) \neq f(x_2)$. If f is strictly decreasing, then $f(x_1) > f(x_2)$ and again we conclude that $f(x_1) \neq f(x_2)$.

Let us now prove (i) $\Rightarrow$ (ii). Since $a < b$ and f is injective we deduce that either $f(a) < f(b)$, or $f(a) > f(b)$. We discuss only the first situation, $f(a) < f(b)$. The second case follows from the first case applied to the continuous injective function $g = -f$. We will prove in several steps that f is strictly increasing.

Step 1. Suppose that $d \in [a,b)$ is such that $f(d) < f(b)$. Then

$$f(d) < f(c), \quad \forall c \in (d,b). \tag{6.2.2}$$

We argue by contradiction. Assume that there exists $c \in (d,b)$ such that $f(c) \leqslant f(d)$. Since f is injective and $d \neq c$ we deduce $f(d) \neq f(c)$ so that $f(c) < f(d)$; see Figure 6.3.

We observe that on the interval $[c,b]$ the function f has values both $< f(d)$ and $> f(d)$ because

$$f(c) < f(d) < f(b).$$

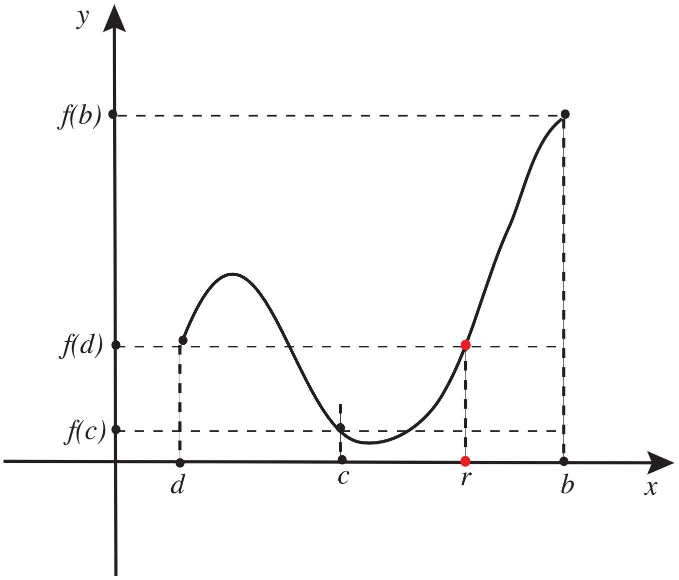

Figure 6.3. *A continuous injective function has to be monotone.*

The Intermediate Value Theorem implies that there must exist a point r in the interval (c, b) such that $f(r) = f(d)$; see Figure 6.3. This contradicts the injectivity of f and completes the proof of Step 1.

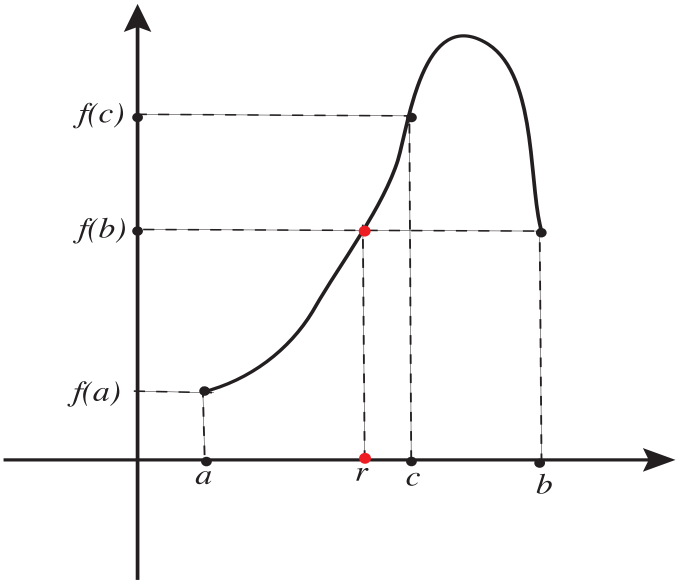

Figure 6.4. *A continuous injective function has to be monotone.*

Step 2. We will show that

$$f(c) < f(b), \quad \forall c \in (a,b). \tag{6.2.3}$$

Again we argue by contradiction. Assume that there exists $c \in (a,b)$ such that $f(c) \geqslant f(b)$. Since f is injective, $f(c) > f(b)$; Figure 6.4.

We observe that on the interval $[a,c]$ the function f has values both $< f(b)$ and $> f(b)$ because

$$f(c) > f(b) > f(a).$$

The Intermediate Value Theorem implies that there must exist a point r in the interval (a,c) such that $f(r) = f(b)$; see Figure 6.4. This contradicts the injectivity of f and completes the proof of Step 2.

Step 3. Suppose that $d < d'$ are points in the interval (a,b). We want to show that $f(d) < f(d')$. Note that since $d \in (a,b)$ we deduce from (6.2.2) and (6.2.3) that $f(a) < f(d) < f(b)$. Since $d' \in (d,b)$ and $f(d) < f(b)$ we deduce from Step 1 that $f(d) < f(d')$. □

Example 6.2.14. Consider the function

$$\sin : \big[-\pi/2, \pi/2\,\big] \to \mathbb{R}.$$

Using the trigonometric-circle definition of sin we deduce that the above function is strictly increasing. Note that

$$\sin(-\pi/2) = -1 = \min_{x\in\mathbb{R}} \sin x, \quad \sin(\pi/2) = 1 = \max_{x\in\mathbb{R}} \sin x.$$

Using Corollary 6.2.11 we deduce that the range of this function is $[-1,1]$ so that the resulting function

$$\sin[-\pi/2, \pi/2] \to [-1,1]$$

is bijective. Its inverse is the function

$$\arcsin : [-1,1] \to [-\pi/2, \pi/2].$$

We want to emphasize that, by construction, the range of arcsin is $[-\pi/2, \pi/2]$.

Similarly, the function

$$\cos : [0,\pi] \to \mathbb{R}$$

is strictly decreasing and its range is $[-1,1]$. Its inverse is the function

$$\arccos : [-1,1] \to [0,\pi].$$ □

Finally, consider the function

$$\tan : (-\pi/2, \pi/2) \to \mathbb{R}.$$

Exercise 6.15 asks you to prove that the above function is bijective. Its inverse is the function

$$\arctan : \mathbb{R} \to (-\pi/2, \pi/2).$$ □

6.3. Uniform continuity

We want to discuss a more subtle concept of continuity that will play an important role in our investigation of integrability.

Definition 6.3.1. Suppose that X is a nonempty subset of the real axis and $f : X \to \mathbb{R}$ is a real valued function defined on X. The *oscillation* of the function f on the set $S \subset X$ is the quantity

$$\operatorname{osc}(f, S) := \sup_{s \in S} f(s) - \inf_{s \in S} f(s) \in [0, \infty]. \qquad \square$$

Let us observe that

$$\operatorname{osc}(f, S) = \sup_{s', s'' \in S} |f(s') - f(s'')|. \tag{6.3.1}$$

Exercise 6.14 asks you to prove this equality.

Definition 6.3.2. Let $J \subset \mathbb{R}$ be an interval and $f : J \to \mathbb{R}$ a function. We say that f is *uniformly continuous* on J if, for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that, for any closed interval $I \subset J$ of length $\ell(I) \leqslant \delta$, we have

$$\operatorname{osc}(f, I) \leqslant \varepsilon. \qquad \square$$

Remark 6.3.3. The uniform continuity of $f : J \to \mathbb{R}$ can be alternatively characterized by the following quantized statement

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 \;\text{ such that }\; \forall x, y \in J \;\; |x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon. \qquad \square$$

Proposition 6.3.4. *Let $J \subset \mathbb{R}$ be an interval and $f : J \to \mathbb{R}$ a function. If f is uniformly continuous, then f is continuous at any point $x_0 \in J$.*

Proof. Let $x_0 \in J$. We have to prove that $\forall \varepsilon > 0$ there exists $\delta > 0$ such that

$$\forall x \;\; |x - x_0| \leqslant \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon.$$

Since f is uniformly continuous, there exists $\delta_0 = \delta_0(\varepsilon) > 0$ such that, for any interval $I \subset J$ of length $\leqslant \delta_0(\varepsilon)$ we have $\operatorname{osc}(f, I) < \varepsilon$. Consider now the interval

$$I_{x_0} := \left\{ x \in J; \;\; |x - x_0| < \frac{\delta_0}{2} \right\}.$$

Clearly I_{x_0} has length $< \delta_0$ so that $\operatorname{osc}(f, I_{x_0}) < \varepsilon$. In particular (6.3.1) implies that for any $x \in I_{x_0}$ we have

$$|f(x) - f(x_0)| < \varepsilon.$$

Hence

$$|x - x_0| < \delta(\varepsilon) := \frac{\delta_0(\varepsilon)}{2} \Rightarrow x \in I_{x_0} \Rightarrow |f(x) - f(x_0)| < \varepsilon$$

$\square$

Theorem 6.3.5 (Uniform Continuity)**.** *Suppose that $a < b$ are two real numbers and $f : [a, b] \to \mathbb{R}$ is a continuous function. Then f is uniformly continuous, i.e., for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for any interval $I \subset [a, b]$ of length $\ell(I) \leqslant \delta$ we have*

$$\operatorname{osc}(f, I) \leqslant \varepsilon.$$

Proof. We have to prove that

$$\forall \varepsilon > 0 \;\; \exists \delta > 0 \;\; \forall I \subset [a, b] \;\text{ interval},\;\; \ell(I) \leqslant \delta \Rightarrow \operatorname{osc}(f, I) \leqslant \varepsilon.$$

We argue by contradiction and we assume that the opposite is true

$$\exists \varepsilon_0 > 0 \;\; \forall \delta > 0 \;\; \exists I = I_\delta \subset [a, b] \;\text{ interval},\;\; \ell(I_\delta) \leqslant \delta \wedge \operatorname{osc}(f, I_\delta) > \varepsilon_0.$$

We deduce that ***for any*** $n \in \mathbb{N}$ there exists a closed interval $I_n = [a_n, b_n] \subset [a, b]$ of length $\leqslant \frac{1}{n}$ such that

$$\operatorname{osc}(f, [a_n, b_n]) > \varepsilon_0. \tag{6.3.2}$$

Since the length of $[a_n, b_n]$ is $\leqslant \frac{1}{n}$ we deduce

$$a_n < b_n \leqslant a_n + \frac{1}{n}.$$

The Bolzano-Weierstrass Theorem 4.4.8 implies that the sequence (a_n) admits a convergent subsequence (a_{n_k}). We set

$$a_* := \lim_{k\to\infty} a_{n_k}.$$

Since $a \leqslant a_n \leqslant b$, we deduce $a_* \in [a, b]$. Since

$$a_{n_k} < b_{n_k} \leqslant a_{n_k} + \frac{1}{n_k}$$

we deduce from the Squeezing Principle that

$$\lim_{k\to\infty} b_{n_k} = \lim_{k\to\infty} a_{n_k} = a_*.$$

On the other hand, since $a_* \in [a, b]$, the function f is continuous at a_*. Thus there exists $\delta > 0$ such that

$$|x - a_*| < \delta \Rightarrow |f(x) - f(a_*)| < \frac{\varepsilon_0}{4}.$$

In other words,

$$\operatorname{dist}(x, a_*) < \delta \Rightarrow f(a_*) - \frac{\varepsilon_0}{4} < f(x) < f(a_*) + \frac{\varepsilon_0}{4}.$$

Since $a_{n_k}, b_{n_k} \to a_*$ there exists k_0 such that

$$[a_{n_{k_0}}, b_{n_{k_0}}] \subset (a_* - \delta, a_* + \delta) \Rightarrow f(a_*) - \frac{\varepsilon_0}{4} < f(x) < f(a_*) + \frac{\varepsilon_0}{4}, \;\; \forall x \in [a_{n_{k_0}}, b_{n_{k_0}}].$$

Thus

$$f(a_*) - \frac{\varepsilon_0}{4} \leqslant \inf_{x\in[a_{n_{k_0}}, b_{n_{k_0}}]} f(x) \leqslant \sup_{x\in[a_{n_{k_0}}, b_{n_{k_0}}]} f(x) \leqslant f(a_*) + \frac{\varepsilon_0}{4}.$$

This shows that

$$\operatorname{osc}\big(\,f,\ [a_{n_{k_0}}, b_{n_{k_0}}]\,\big) \leqslant \left(f(a_*) + \frac{\varepsilon_0}{4}\right) - \left(\,f(a_*) - \frac{\varepsilon_0}{4}\right) = \frac{\varepsilon_0}{2}.$$

This contradicts (6.3.2) and completes the proof of the theorem. □

Remark 6.3.6. The above result is no longer valid for continuous functions defined on *non-closed* or *unbounded* intervals. Consider for example the continuous function $f : (0,1) \to \mathbb{R}$, $f(x) = \frac{1}{x}$. For each $n \in \mathbb{N}$, $n > 1$ we define

$$I_n = \left[\frac{1}{n+1}, \frac{1}{n}\right].$$

Since f is decreasing we deduce that

$$\sup_{x \in I_n} f(x) = f\left(\frac{1}{n+1}\right) = n+1, \quad \inf_{x \in I_n} f(x) = f\left(\frac{1}{n}\right) = n$$

so that $\operatorname{osc}(f, I_n) = 1$. On the other hand, $\ell(I_n) = \frac{1}{n(n+1)} \to 0$ as $n \to \infty$. We have thus produced arbitrarily short intervals over which the oscillation is 1.

Exercise 6.13 describes an example of continuous function over an *unbounded* interval that is *not* uniformly continuous on that interval. □

6.4. Exercises

Exercise 6.1. Prove Theorem 6.1.2. □

Exercise 6.2. Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are two continuous functions such that $f(q) = g(q)$, $\forall q \in \mathbb{Q}$. Prove that $f(x) = g(x)$, $\forall x \in \mathbb{R}$.
Hint. You may want to invoke Proposition 3.4.4. □

Exercise 6.3. Prove Corollary 6.1.3. □

Exercise 6.4. Prove the inequality (6.1.1). □

Exercise 6.5. Suppose that $f, g : [a, b] \to \mathbb{R}$ are continuous functions.

(a) Prove that the function $|f|$ continuous.

(b) Prove that for any $x \in [a, b]$ we have

$$\max\{\, f(x), g(x)\,\} = \frac{1}{2}\big(f(x) + g(x) + |f(x) - g(x)|\,\big).$$

(c) Prove that the function $h : [a, b] \to \mathbb{R}$, $h(x) = \max\{f(x), g(x)\}$ is continuous. □

Exercise 6.6 (Weierstrass)**.** Suppose that X is a nonempty set of real numbers, $f_n : X \to \mathbb{R}$, $n \in \mathbb{N}$, is a sequence of functions, and $f : X \to \mathbb{R}$ a function on X. Suppose that for any $n \in \mathbb{N}$ we have

$$M_n := \sup_{x \in X} |f_n(x) - f(x)| < \infty.$$

Prove that the following statements are equivalent.

(i) The sequence (f_n) converges uniformly to f on X.

(ii) $\lim_{n\to\infty} M_n = 0$.

□

Exercise 6.7 (Weierstrass)**.** Consider a sequence of functions $f_n : [a, b] \to \mathbb{R}$, $n \geqslant 0$, where a, b are real numbers $a < b$. Suppose that there exists a sequence of positive real numbers $(c_n)_{n\geqslant 0}$ with the following properties.

(i) $|f_n(x)| \leqslant c_n$, $\forall n \geqslant 0$, $\forall x \in [a, b]$.

(ii) The series $\sum_{n\geqslant 0} c_n$ is convergent.

(a) Prove that for any $x \in [a, b]$, the series of real numbers $\sum_{n\geqslant 0} f_n(x)$ is absolutely convergent. Denote by $s(x)$ its sum.

(b) Denote by $s_n(x)$ the n-th partial sum

$$s_n(x) = f_0(x) + f_1(x) + \cdots + f_n(x)$$

Prove that the sequence of functions $s_n : [a, b] \to \mathbb{R}$ converges uniformly on $[a, b]$ to the function $s : [a, b] \to \mathbb{R}$ defined in (a).

Hint. Use Exercise 6.6. □

Exercise 6.8. Consider the power series

$$\sum_{n \geqslant 0} a_n x^n, \quad a_n \in \mathbb{R}. \tag{6.4.1}$$

Suppose that for some $R > 0$ the series

$$\sum_{n \geqslant 0} a_n R^n$$

is absolutely convergent.

(a) Prove that the series (6.4.1) converges absolutely for any $x \in [-R, R]$. Denote by $s(x)$ its sum.

(b) Denote by $s_n(x)$ the n-th partial sum

$$s_n(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Prove that the resulting sequence of functions $s_n : [-R, R] \to \mathbb{R}$ converges uniformly to $s(x)$. Conclude that the function $s(x)$ is continuous on $[-R, R]$.

Hint. Use the results in Exercise 6.7. □

Exercise 6.9. Consider the sequence of functions

$$f_n : [0, 1] \to \mathbb{R}, \quad f_n(x) = x^n, \quad n \in \mathbb{N}.$$

(a) Prove that for any $x \in [0, 1]$ the sequence $(f_n(x))_{n \in \mathbb{N}}$ is convergent. Compute its limit $f(x)$.

(b) Given $n \in \mathbb{N}$ compute

$$\sup_{x \in [0,1]} |f_n(x) - f(x)|.$$

(c) Prove that the sequence of functions $f_n(x)$ does *not* converge uniformly to the function $f(x)$ defined in (a). □

Exercise 6.10 (Cauchy)**.** Suppose $X \subset \mathbb{R}$ is a nonempty set of real number and $f_n : X \to \mathbb{R}$ is a sequence of real valued functions defined on X. Prove that the following statements are equivalent.

(i) There exists a function $f : X \to \mathbb{R}$ such that the sequence $f_n : X \to \mathbb{R}$ converges uniformly on X to $f : X \to \mathbb{R}$.

(ii) $\forall \varepsilon > 0$, $\exists N = N(\varepsilon) \in \mathbb{N}$ such that

$$\forall n, m > N(\varepsilon), \quad \forall x \in X : \quad |f_n(x) - f_m(x)| < \varepsilon.$$

□

Exercise 6.11. (a) Prove Corollary 6.2.12.

(b) Let $f(x)$ be a polynomial of *odd* degree. Prove that there exists $r \in \mathbb{R}$ such that $f(r) = 0$. □

Exercise 6.12. Suppose that $f : [0,1] \to [0,1]$ is a continuous function. Prove that there exists $c \in [0,1]$ such that $f(c) = c$. Can you give a geometric interpretation of this result? □

Exercise 6.13. (a) Find the oscillation of the function $f : [0,\infty) \to \mathbb{R}$, $f(x) = x^2$, over an interval $[a,b] \subset (0,\infty)$.

(b) Prove that for any $n \in \mathbb{N}$ one can find an interval $[a,b] \subset [0,\infty)$ of length $\leqslant \frac{1}{n}$ over which the oscillation of f is $\geqslant 1$. □

Exercise 6.14. (a) Suppose that $f : X \to \mathbb{R}$ is a function defined on a set X, and $Y \subset X$. Prove that

$$\operatorname{osc}(f,X) = \sup_{x',x''\in X} |f(x') - f(x'')| \;\text{ and }\; \operatorname{osc}(f,Y) \leqslant \operatorname{osc}(f,X).$$

(b) Consider a function $f : (a,b) \to \mathbb{R}$. Prove that f is continuous at a point $x_0 \in (a,b)$ if and only if

$$\lim_{\delta\searrow 0} \operatorname{osc}\big(f, [x_0-\delta, x_0+\delta]\big) = 0.$$

(c) Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function. Prove that

$$\operatorname{osc}(f,\ (a,b)\,) = \operatorname{osc}(f,\ [a,b]\,).$$

Exercise 6.15. Consider the function

$$f : (-\pi/2, \pi/2) \to \mathbb{R},\;\; f(x) = \tan x = \frac{\sin x}{\cos x}.$$

Prove that f is strictly increasing and

$$\lim_{x\to\pm\pi/2} f(x) = \pm\infty.$$

Conclude that f is bijective. □

Exercise 6.16 (Kuratowski)**.** Suppose that $f_n : [0,1] \to \mathbb{R}$, $n \in \mathbb{N}$ is a sequence of continuous functions and $f : [0,1] \to \mathbb{R}$ is a *continuous* function such that

$$\forall x \in [0,1],\;\; \lim_{n\to\infty} f_n(x) = f(x).$$

Prove that the following statements are equivalent.

(i) For any convergent sequence (x_n) of points in $[0,1]$

$$\lim_{n\to\infty} f_n(x_n) = f_n\big(\lim_{n\to\infty} x_n\,\big).$$

(ii) The sequence of functions f_n converges *uniformly* to f on $[0, 1]$.

□

Exercise 6.17. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a function such that

$$\lim_{x \to \pm\infty} g(x) = \mp\infty.$$

Prove that there exists no *continuous* function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f\big(f(x) \big) = g(x).$$

Hint. Argue by contradiction. Suppose that there is such a function f. Prove that the limits of f at $\pm\infty$ exist and then show that this leads to a contradiction. □

6.5. Exercises for extra credit

Exercise* 6.1. Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function. For any $x \in [a, b]$ we define

$$m(x) = \inf_{t \in [a,x]} f(x), \quad M(x) = \sup_{t \in [a,x]} f(x).$$

Prove that the functions $x \mapsto m(x)$ and $x \mapsto M(x)$ are continuous. □

Exercise* 6.2. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function, continuous at 0 and satisfying

$$f(0) = 0, \quad f(1) = 1,$$

$$f(x + y) = f(x) + f(y), \quad \forall x \in \mathbb{R}.$$

Prove that $f(x) = x$, $\forall x \in \mathbb{R}$. □

Exercise* 6.3. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function continuous at 0 and satisfying the following properties.

(i) $f(x) > 0$, $\forall x \in \mathbb{R}$.

(ii) $f(x + y) = f(x)f(y)$, $\forall x, y \in \mathbb{R}$.

Set $a := f(1)$. Prove that $f(x) = a^x$, $\forall x \in \mathbb{R}$. □

Exercise* 6.4. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function satisfying the following conditions

$$f(x + y) = f(x) + f(y), \quad \forall x, y \in \mathbb{R}. \tag{6.5.1a}$$

$$f(xy) = f(x)f(y), \quad \forall x, y \in \mathbb{R}. \tag{6.5.1b}$$

$$f(1) \neq 0. \tag{6.5.1c}$$

Prove that $f(x) = x$, $\forall x \in \mathbb{R}$. □

Exercise* 6.5. Suppose that $f : [0,1] \to \mathbb{R}$ is a continuous function. For $n \in \mathbb{R}$ define

$$f_n : [0,1] \to \mathbb{R}$$

by setting

$$f_n(x) := \begin{cases} f(0), & \text{if } x = 0 \\ \min\big\{\, f(x);\ \frac{k-1}{n} \leqslant x \leqslant \frac{k}{n} \,\big\}, & \text{if } \frac{k-1}{n} < x \leqslant \frac{k}{n}, \quad k = 1, \dots, n. \end{cases}$$

Prove that the sequence of functions (f_n) converges *uniformly* to the function f on $[0,1]$.□

Chapter 7

Differential calculus

7.1. Linear approximation and derivative

The differential calculus is one of the most consequential scientific discoveries in the history of mankind. Surprisingly, this revolutionary theory is based on a very simple principle: often one can learn nontrivial things about complicated objects by approximating them with simpler ones.

In the case at hand, the complicated object is a function $f : (a, b) \to \mathbb{R}$ and one would like to understand its behavior near a point $x_0 \in (a, b)$. To achieve this, we try to approximate f with a simpler function, and the linear functions are the simplest nontrivial candidates.

Definition 7.1.1. Suppose that I is an interval[a] on the real axis, $f : I \to \mathbb{R}$ is a function and $x_0 \in I$. A *linear approximation* or *linearization* of f at x_0 is a linear function

$$L : \mathbb{R} \to \mathbb{R},\ \ L(x) = b + m(x - x_0)$$

such that

$$L(x_0) = f(x_0) \tag{7.1.1a}$$

$$f(x) - L(x) = o(x - x_0)\ \text{ as } x \to x_0. \tag{7.1.1b}$$

Above, we used Landau's symbol o defined in (5.8.2) signifying that

$$\lim_{x \in I,\, x \to x_0} \frac{f(x) - L(x)}{x - x_0} = 0.$$

The function is said to be *linearizable* at x_0 if it admits a linearization at x_0. □

[a] The interval I could be closed, could be open, could be neither, could be bounded or not.

Suppose that L is a linearization of the function $f : I \to \mathbb{R}$ at x_0. By (7.1.1a), the value of L at x_0 is equal to the value of f at x_0, $L(x_0) = f(x_0)$. On the other hand

$$L(x_0) = b + m(x_0 - x_0) = b$$

and we deduce that $L(x)$ has the form

$$L(x) = f(x_0) + m(x - x_0).$$

The linear function $L(x)$ is meant to approximate the function $f(x)$ for x not too far for x_0. The *error* of this linear approximation of $f(x)$ is the difference $r(x) = f(x) - L(x)$ which by definition is $o(x - x_0)$ as $x \to x_0$. In less rigorous terms, $r(x)$ is a tiny fraction of $(x - x_0)$ when x is close to x_0. Note that

$$f(x) - L(x) = f(x) - \big(m(x - x_0) + f(x_0)\big) = f(x) - f(x_0) - m(x - x_0),$$

$$\frac{f(x) - L(x)}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0} - m.$$

Since

$$0 = \lim_{x \to x_0} \frac{f(x) - L(x)}{x - x_0} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} - m$$

we deduce that

$$m = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}. \tag{7.1.2}$$

Thus if f is linearizable at x_0, then there exists a *unique* linearization $L(x)$ described by

$$L(x) = f(x_0) + m(x - x_0),$$

where the slope m is given by (7.1.2).

Definition 7.1.2. Suppose that I is an interval of the real axis, $f : I \to \mathbb{R}$ is a function and $x_0 \in I$.

(i) We say that f is *differentiable at* x_0 if the limit (7.1.2)

$$\lim_{\substack{x \to x_0 \\ x \in I}} \frac{f(x) - f(x_0)}{x - x_0} \tag{7.1.3}$$

exists and it is finite. If this is the case, we denote the limit by $f'(x_0)$ or $\frac{df}{dx}|_{x=x_0}$ and we will refer to it as the *derivative* of f at x_0.

(ii) We say that f is differentiable on I if it is differentiable at *any* point $x \in I$. The function $f' : I \to \mathbb{R}$ that assigns to $x \in I$ the derivative $f'(x)$ of f at x is called the *derivative of the function f on the interval I.* □

Remark 7.1.3. In concrete computations it is often convenient to describe the derivative of f at x_0 as the limit

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

This is obtained from (7.1.3) if we denote by h the "displacement" $x - x_0$. With this notation we have $x = x_0 + h$ and

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x_0 + h) - f(x_0)}{h}. \qquad \square$$

The next result summarizes the observations we have made so far.

Proposition 7.1.4. *Suppose that I is an interval of the real axis, $f : I \to \mathbb{R}$ is a function and $x_0 \in I$. Then the following statements are equivalent.*

(i) *The function f is differentiable at x_0.*

(ii) *The function f is linearizable at x_0.*

(iii) *The function f is differentiable at x_0 and the function $L(x) = f(x_0)+f'(x_0)(x-x_0)$ is the linearization of f at x_0, i.e.,*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + r(x), \;\; \lim_{x\to x_0} \frac{r(x)}{|x - x_0|} = 0. \tag{7.1.4}$$

□

We should perhaps give a geometric interpretation to the linear approximation of f at x_0. The graph of f is the curve

$$G_f := \Big\{ \big(x, f(x)\big) \in \mathbb{R}^2;\;\; x \in (a, b) \Big\}.$$

The point $x_0 \in I$ determines a point $P_0 = (x_0, f(x_0))$ on the curve G_f; see Figure 7.1.

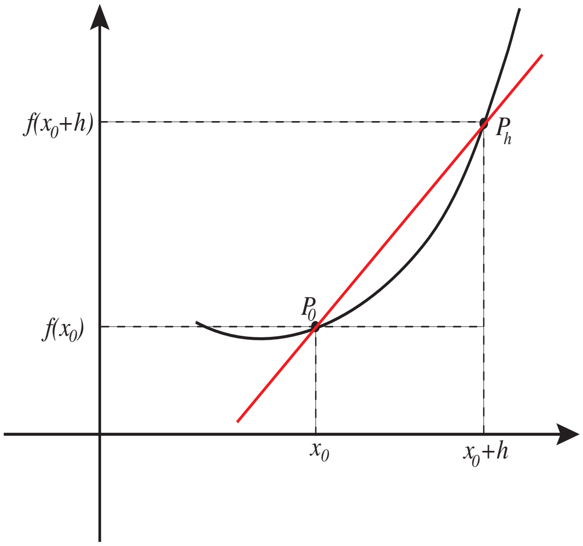

Figure 7.1. *A tangent line to the graph of a function is a limit of secant lines.*

The graph of a linear function $L(x)$ is a line in the plane and since we are interested in approximating the behavior of f near x_0 it makes sense to look only at lines $\ell_{P_0,P}$ determined by two points P_0, P on the graph G_f. Since we are interested only in the behavior of f *near* x_0, we may assume that the point P is not too far from P_0. Thus we assume that the coordinates of P are $(\,x_0+h, f(x_0+h)\,)$, where h is very small.

In more concrete terms, we look at the lines ℓ_{P_0,P_h} determined by the two points

$$P_0 := (\,x_0, f(x_0)\,),\ \ P_h := (\,x_0+h, f(x_0+h)\,),$$

where h very small. The slope of the line ℓ_{P_0,P_h} is

$$m(h) := \frac{f(x_0+h)-f(x_0)}{(x_0+h)-x_0} = \frac{f(x_0+h)-f(x_0)}{h},$$

so its equation is

$$y - f(x_0) = m(h)(x-x_0).$$

This is the graph of the linear function

$$L_{x_0,h}(x) = f(x_0) + m(h)(x-x_0).$$

Suppose that as $h \to 0$ the line ℓ_{P_0,P_h} stabilizes to some limiting position. This limit line goes through the point P_0 and therefore its position is determined by its slope

$$\lim_{h\to 0} m(h) = \lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h} = \lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}.$$

We see that this limit exists and it is finite if and only if f is differentiable at x_0. In this case, the limit line is the graph of the linear approximation of f at x_0.

Definition 7.1.5. Suppose $I \subset \mathbb{R}$ is an interval of the real axis and $f : I \to \mathbb{R}$ is a function differentiable at x_0. The *tangent line to the graph* of f at x_0 is the graph of the linearization of f at x_0. □

Remark 7.1.6. (a) The quantities

$$\frac{f(x)-f(x_0)}{x-x_0},\ \ \frac{f(x_0+h)-f(x_0)}{h}$$

are called *difference quotients* of f at x_0. You should think of such a difference quotient as measuring the average rate of change of the quantity f over the interval $[x_0, x]$.

In physics, the numerator $f(x)-f(x_0)$ is denoted by Δf while the denominator is denoted Δx. The symbol Δ is shorthand for "*variation of*". Thus

$$\frac{df}{dx} = \lim_{\Delta x\to 0}\frac{\Delta f}{\Delta x}.$$

From the equality

$$f'(x) = \frac{df}{dx}$$

we deduce formally

$$df = f'(x)dx. \tag{7.1.5}$$

The expression $f'(x)dx$ is called the *differential* of f and as the above equality suggests, it is denoted by df.

(b) Often a function $f : [a, b] \to \mathbb{R}$ has a physical meaning. For example, the interval $[a, b]$ can signify a stretch of highway between mile a and mile b and $f(x)$ could be the temperature at mile x and thus it is measured in $^\circ F$. The difference quotient

$$\frac{f(x) - f(x_0)}{x - x_0}$$

has a different meaning. The numerator $f(x) - f(x_0)$ describes the change in temperature from mile x_0 to mile x and it is again measured in $^\circ F$, while the numerator $x - x_0$ is the "distance" (could be negative) from mile x_0 to mile x and thus it is measured in miles. We deduce that the quotient is measured in different units, degrees-per-mile, and should be viewed as the average rate of change in temperature per mile. When $x \to x_0$ we are measuring the rate of change in temperature over shorter and shorter stretches of highway. For this reason, the limit $f'(x_0)$ is sometimes referred to as the *infinitesimal rate of change.* □

The differentiability of a function at a point x_0 imposes restrictions on the behavior of the function near that point. Our next elementary result describes one such restriction. Its proof is left to you as an exercise.

Proposition 7.1.7. *Suppose I is an interval of the real axis $\mathbb{R}$ and $f : I \to \mathbb{R}$ is a function that is differentiable at a point $x_0 \in I$. Then f is continuous at x_0, i.e.,*

$$\lim_{I \ni x \to x_0} f(x) = f(x_0). \qquad \square$$

Remark 7.1.8. The converse of the above result is not true. There exist continuous functions $f : [0, 1] \to \mathbb{R}$ which are nowhere differentiable. For example, the function

$$f : [0, 1] \to \mathbb{R}, \;\; f(t) = \sum_{n=0}^{\infty} \frac{\cos(5^n t)}{2^n},$$

is continuous and nowhere differentiable. Its graph, depicted in Figure 7.2, may convince you of the validity of this claim. The rigorous proof of this fact is rather ingenious and for details and generalizations we refer to [21]. □

Suppose that $I \subset \mathbb{R}$ is an interval and $f : I \to \mathbb{R}$ is a differentiable function. We say that f is *twice* differentiable if its derivative f', viewed as a function $f' : I \to \mathbb{R}$, is also differentiable. The *second derivative* of f denoted by f'' or $\frac{d^2 f}{dx^2}$ is the derivative of f'

$$f'' := \frac{d}{dx}(f').$$

Recursively, for any natural number $n > 1$, we say that f is *n-times differentiable* if its derivative is $(n - 1)$-times differentiable. The n-th derivative of f is the function

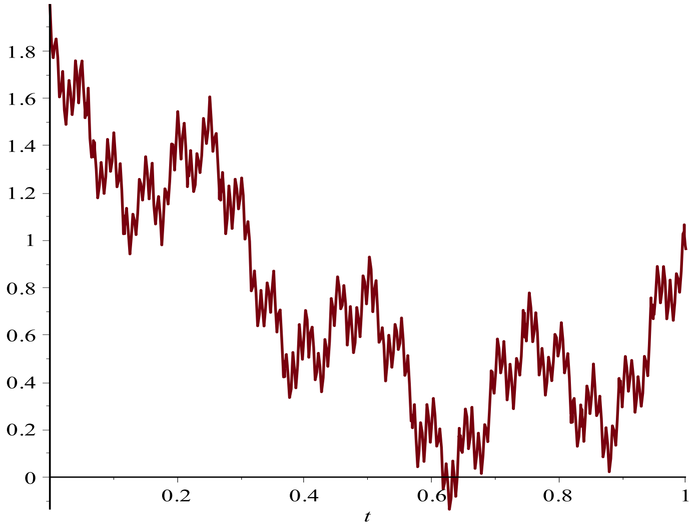

Figure 7.2. *Weierstrass's example of continuous, nowhere differentiable function.*

$f^{(n)} : I \to \mathbb{R}$ defined recursively as

$$f^{(n)} := \frac{d}{dx}\big(f^{(n-1)} \big).$$

Often we will use the alternate notation $\frac{d^n f}{dx^n}$ to denote the n-th derivative of f.

Definition 7.1.9. Let $I \subset \mathbb{R}$ be an interval.

(i) We denote by $C^0(I)$ the set consisting of all the continuous functions $f : I \to \mathbb{R}$.

(ii) If n is a natural number, then we denote by $C^n(I)$ the space of functions $f : I \to \mathbb{R}$ which are
- n-times differentiable and
- the n-th derivative $f^{(n)}$ is a continuous function.

We will refer to the functions in $C^n(I)$ as C^n*-functions.*

(iii) We denote by $C^\infty(I)$ the space of functions $I \to \mathbb{R}$ which are infinitely many times differentiable. We will refer to such functions as *smooth.*

□

7.2. Fundamental examples

In this section we describe a *very important* collection of differentiable functions.

Example 7.2.1 (Constant functions)**.** Suppose that $f : \mathbb{R} \to \mathbb{R}$ is the function which is identically equal to a fixed real number c,

$$f(x) = c, \quad \forall x \in \mathbb{R}.$$

Then f is differentiable and $f'(x) = 0$, $\forall x \in \mathbb{R}$. Indeed, for any $x_0 \in \mathbb{R}$

$$\frac{f(x_0 + h) - f(x_0)}{h} = 0, \quad \forall h \neq 0. \qquad \square$$

Example 7.2.2 (Monomials)**.** Suppose that $n \in \mathbb{N}$ and consider the *monomial function* $\mu_n : \mathbb{R} \to \mathbb{R}$, $\mu_n(x) = x^n$. Then μ_n is differentiable on $\mathbb{R}$ and its derivative is

$$\mu_n'(x) = nx^{n-1}, \quad \forall x \in \mathbb{R} \Longleftrightarrow \frac{d}{dx}(x^n) = nx^{n-1}. \tag{7.2.1}$$

To prove this claim we investigate the difference quotients of μ_n at $x_0 \in \mathbb{R}$. We have

$$\mu_n(x_0 + h) - \mu_n(x_0) = (x_0 + h)^n - x_0^n$$

(use Newton's binomial formula (3.2.4))

$$= x_0^n + \binom{n}{1} x_0^{n-1} h + \binom{n}{2} x_0^{n-2} h^2 + \cdots + \binom{n}{n} h^n - x_0^n$$

$$= h\left(\binom{n}{1} x_0^{n-1} + \binom{n}{2} x_0^{n-2} h + \cdots + \binom{n}{n} h^{n-1} \right),$$

so that

$$\frac{\mu_n(x_0 + h) - \mu_n(x_0)}{h} = \binom{n}{1} x_0^{n-1} + \binom{n}{2} x_0^{n-2} h + \cdots + \binom{n}{n} h^{n-1}.$$

Now observe that

$$\mu_n'(x_0) = \lim_{h \to 0} \frac{\mu_n(x_0 + h) - \mu_n(x_0)}{h}$$

$$= \lim_{h \to 0} \left(\binom{n}{1} x_0^{n-1} + \binom{n}{2} x_0^{n-2} h + \cdots + \binom{n}{n} h^{n-1} \right) = \binom{n}{1} x_0^{n-1} = n x_0^{n-1}.$$

For example

$$(x^2)' = 2x, \quad d(x^2) = 2xdx. \qquad \square$$

Example 7.2.3 (Power functions)**.** Fix a real number α and consider the power function

$$f : (0, \infty) \to \mathbb{R}, \quad f(x) = x^\alpha.$$

Then f is differentiable and its derivative is

$$\boxed{f'(x) = \alpha x^{\alpha - 1}, \quad \forall x > 0 \Longleftrightarrow \frac{d}{dx}(x^\alpha) = \alpha x^{\alpha - 1}}. \tag{7.2.2}$$

To prove this claim we investigate the difference quotients of $f(x)$ at $x_0 \in (0, \infty)$. We have

$$(x_0 + h)^\alpha - x_0^\alpha = \left(x_0 \left(1 + \frac{h}{x_0} \right) \right)^\alpha - x_0^\alpha = x_0^\alpha \left(\left(1 + \frac{h}{x_0} \right)^\alpha - 1 \right),$$

$$\frac{f(x_0 + h) - f(x_0)}{h} = x_0^\alpha \frac{\left(1 + \frac{h}{x_0} \right)^\alpha - 1}{h} = x_0^\alpha \frac{\left(1 + \frac{h}{x_0} \right)^\alpha - 1}{x_0 \frac{h}{x_0}}$$

$$= x_0^{\alpha-1}\frac{\left(1+\frac{h}{x_0}\right)^{\alpha}-1}{\frac{h}{x_0}}.$$

We set $t := \frac{h}{x_0}$ and we observe that $t \to 0$ as $h \to 0$ and

$$\frac{f(x_0+h)-f(x_0)}{h} = x_0^{\alpha-1}\frac{(1+t)^{\alpha}-1}{t}.$$

Invoking the fundamental limit (5.5.4) we deduce

$$\lim_{t\to 0}\frac{(1+t)^{\alpha}-1}{t} = \alpha$$

so that

$$f'(x_0) = \lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h} = \alpha x_0^{\alpha-1}.$$

Note that if $\alpha = \frac{1}{2}$, then $f(x) = \sqrt{x}$ and we deduce

$$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}},\ \ d(\sqrt{x}) = \frac{dx}{2\sqrt{x}}. \tag{7.2.3}$$

□

Example 7.2.4 (The exponential function)**.** Consider the exponential function

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

This function is differentiable and its derivative is

$$f'(x) = e^x,\ \ \forall x\in\mathbb{R}\Longleftrightarrow\frac{d}{dx}(e^x) = e^x. \tag{7.2.4}$$

To prove this claim we investigate the difference quotients of f at $x_0\in\mathbb{R}$. We have

$$f(x_0+h)-f(x_0) = e^{x_0+h}-e^{x_0} = e^{x_0}(e^h-1),$$

$$\frac{f(x_0+h)-f(x_0)}{h} = e^{x_0}\frac{e^h-1}{h}.$$

On the other hand, the fundamental limit (5.5.3) implies that

$$\lim_{h\to 0}\frac{e^h-1}{h} = 1.$$

Hence

$$f'(x_0) = \lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h} = e^{x_0}.$$

These computations show that the exponential function is smooth, i.e., infinitely many times differentiable and

$$\boxed{\frac{d^n}{dx^n}e^x = e^x,\ \ d(e^x) = e^x dx}. \tag{7.2.5}$$

□

Example 7.2.5 (The natural logarithm)**.** Consider the natural logarithm

$$f:(0,\infty)\to\mathbb{R},\;\; f(x)=\ln x=\log x.$$

Then f is differentiable and its derivative is

$$\boxed{f'(x)=\frac{1}{x},\;\;\forall x>0 \Longleftrightarrow \frac{d}{dx}(\ln x)=\frac{1}{x},\;\; d(\ln x)=\frac{dx}{x}}. \tag{7.2.6}$$

To prove this claim we investigate the difference quotients of f at $x_0>0$. We have

$$f(x_0+h)-f(x_0)=\ln(x_0+h)-\ln x_0=\ln\left(x_0\Big(1+\frac{h}{x_0}\Big)\right)-\ln x_0$$

$$=\ln x_0+\ln\Big(1+\frac{h}{x_0}\Big)-\ln x_0=\ln\Big(1+\frac{h}{x_0}\Big),$$

$$\frac{f(x_0+h)-f(x_0)}{h}=\frac{\ln\Big(1+\frac{h}{x_0}\Big)}{h}=\frac{\ln\Big(1+\frac{h}{x_0}\Big)}{x_0\frac{h}{x_0}}=\frac{1}{x_0}\frac{\ln\Big(1+\frac{h}{x_0}\Big)}{\frac{h}{x_0}}.$$

We set $t=\frac{h}{x_0}$ and we conclude from above that

$$\frac{f(x_0+h)-f(x_0)}{h}=\frac{1}{x_0}\frac{\ln(1+t)}{t}.$$

Note that t goes to zero when $h\to 0$. We can now invoke (5.5.2) to conclude that

$$\lim_{t\to 0}\frac{\ln(1+t)}{t}=1.$$

This proves

$$\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=\frac{1}{x_0}.$$

□

Example 7.2.6 (Trigonometric functions)**.** The trigonometric functions

$$\sin,\cos:\mathbb{R}\to\mathbb{R}$$

are differentiable and

$$\boxed{\frac{d}{dx}(\sin x)=\cos x,\;\;\frac{d}{dx}(\cos x)=-\sin x}. \tag{7.2.7}$$

Fix $x_0\in\mathbb{R}$. We have

$$\sin(x_0+h)-\sin x_0\overset{(5.7.1a)}{=}\sin x_0\cos h+\cos x_0\sin h-\sin x_0$$

$$=\sin x_0\big(\cos h-1\big)+\cos x_0\sin h=-2\sin^2(h/2)\sin x_0+\cos x_0\sin h.$$

Hence

$$\frac{\sin(x_0+h)-\sin x_0}{h}=-2\sin x_0\frac{\sin^2(h/2)}{h}+\cos x_0\frac{\sin h}{h}.$$

$$=-\sin x_0\frac{\sin^2(h/2)}{\frac{h}{2}}+\cos x_0\frac{\sin h}{h}=-\frac{h}{2}\sin x_0\left(\frac{\sin(\frac{h}{2})}{\frac{h}{2}}\right)^2+\cos x_0\frac{\sin h}{h}.$$

From the fundamental identity (5.6.6) we deduce that

$$\lim_{t\to 0}\frac{\sin t}{t}=1.$$

Hence

$$\lim_{h\to 0}\frac{h}{2}\sin x_0\left(\frac{\sin(\frac{h}{2})}{\frac{h}{2}}\right)^2=0,\quad \lim_{h\to 0}\cos x_0\frac{\sin h}{h}=\cos x_0,$$

and thus

$$\lim_{h\to 0}\frac{\sin(x_0+h)-\sin x_0}{h}=\cos x_0.$$

The equality

$$\lim_{h\to 0}\frac{\cos(x_0+h)-\cos x_0}{h}=-\sin x_0$$

is proved in a similar fashion and the details are left to you as an exercise. □

7.3. The basic rules of differential calculus

In the previous section we have computed the derivatives of a few important functions. In this section we describe a few basic rules which will allow us to easily compute the derivatives of almost any function.

Theorem 7.3.1 (Arithmetic rules of differentiation)**.** *Suppose that $I\subset\mathbb{R}$ is an interval, and $f,g:I\to\mathbb{R}$ are two functions differentiable at x_0. Then the following hold.*

Addition. *The sum $f+g$ is differentiable at x_0 and*

$$\boxed{(f+g)'(x_0)=f'(x_0)+g'(x_0).}$$

Scalar multiplication. *If c is a real number, then the function cf is differentiable at x_0 and*

$$\boxed{(cf)'(x_0)=cf'(x_0).}$$

Product. *The product $f\cdot g$ is differentiable at x_0 and its derivative is given by the* product rule *or* Leibniz rule

$$\boxed{(f\cdot g)'(x_0)=f'(x_0)g(x_0)+f(x_0)g'(x_0).}$$

Quotient. *If $g(x_0)\neq 0$, then there exists $\delta>0$ such that*

$$\forall x\in I\ \ |x-x_0|<\delta\Rightarrow g(x)\neq 0.$$

Set

$$I_{x_0,\delta}:=\big\{x\in I;\ \ |x-x_0|<\delta\,\big\}.$$

The quotient $\frac{f}{g}$ is a well defined function on $I_{x_0,\delta}$ which is differentiable at x_0 and its derivative at x_0 is determined by the quotient rule

$$\boxed{\left(\frac{f}{g}\right)'(x_0)=\frac{f'(x_0)g(x_0)-f(x_0)g'(x_0)}{g(x_0)^2}.}$$

Proof. Addition. We have

$$\begin{aligned}\frac{(f+g)(x_0+h)-(f+g)(x_0)}{h} &= \frac{f(x_0+h)-f(x_0)+g(x_0+h)-g(x_0)}{h}\\ &= \frac{f(x_0+h)-f(x_0)}{h}+\frac{g(x_0+h)-g(x_0)}{h}.\end{aligned}$$

Hence

$$\begin{aligned}\lim_{h\to 0}\frac{(f+g)(x_0+h)-(f+g)(x_0)}{h} &= \lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}+\lim_{h\to 0}\frac{g(x_0+h)-g(x_0)}{h}\\ &= f'(x_0)+g'(x_0).\end{aligned}$$

Scalar multiplication. We have

$$\frac{(cf)(x_0+h)-(cf)(x_0)}{h}=c\frac{f(x_0+h)-f(x_0)}{h}$$

so that

$$\lim_{h\to 0}\frac{(cf)(x_0+h)-(cf)(x_0)}{h}=c\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=cf'(x_0).$$

Product. We have

$$\begin{aligned}(f\cdot g)(x_0+h)-(f\cdot g)(x_0) &= f(x_0+h)g(x_0+h)-f(x_0)g(x_0)\\ &= f(x_0+h)g(x_0+h)-f(x_0)g(x_0+h)+f(x_0)g(x_0+h)-f(x_0)g(x_0)\\ &= \big(f(x_0+h)-f(x_0)\big)g(x_0+h)+f(x_0)\big(g(x_0+h)-g(x_0)\big),\end{aligned}$$

so that

$$\frac{(f\cdot g)(x_0+h)-(f\cdot g)(x_0)}{h}=\frac{\big(f(x_0+h)-f(x_0)\big)}{h}g(x_0+h)+f(x_0)\frac{\big(g(x_0+h)-g(x_0)\big)}{h}.$$

Since g is differentiable at x_0 it is also continuous at x_0 by Proposition 7.1.7. Hence

$$\lim_{h\to 0}g(x_0+h)=g(x_0),\quad \lim_{h\to 0}f(x_0)\frac{\big(g(x_0+h)-g(x_0)\big)}{h}=f(x_0)g'(x_0).$$

Since f is differentiable at x_0 we deduce

$$\begin{aligned}&\lim_{h\to 0}\frac{\big(f(x_0+h)-f(x_0)\big)}{h}g(x_0+h)\\ &=\lim_{h\to 0}\frac{\big(f(x_0+h)-f(x_0)\big)}{h}\cdot\lim_{h\to 0}g(x_0+h)=f'(x_0)g(x_0).\end{aligned}$$

Hence

$$\lim_{h\to 0}\frac{(f\cdot g)(x_0+h)-(f\cdot g)(x_0)}{h}=f'(x_0)g(x_0)+f(x_0)g'(x_0).$$

Quotient. The function g is differentiable at x_0, thus continuous at this point. From Theorem 6.2.1 we deduce that there exists $\delta>0$ such that

$$\forall x\in I,\quad |x-x_0|<\delta\Rightarrow g(x)\neq 0.$$

For $|h| < \delta$ such that $x_0 + h \in I$ we have

$$\left(\frac{1}{g}\right)(x_0+h) - \left(\frac{1}{g}\right)(x_0) = \frac{1}{g(x_0+h)} - \frac{1}{g(x_0)} = \frac{g(x_0)-g(x_0+h)}{g(x_0)g(x_0+h)}$$

so that

$$\frac{\left(\frac{1}{g}\right)(x_0+h) - \left(\frac{1}{g}\right)(x_0)}{h} = \frac{g(x_0)-g(x_0+h)}{h}\frac{1}{g(x_0)g(x_0+h)}$$

Hence

$$\left(\frac{1}{g}\right)'(x_0) = \lim_{h\to 0}\frac{\left(\frac{1}{g}\right)(x_0+h) - \left(\frac{1}{g}\right)(x_0)}{h}$$

$$= \lim_{h\to 0}\frac{g(x_0)-g(x_0+h)}{h}\cdot\lim_{h\to 0}\frac{1}{g(x_0)g(x_0+h)} = -\frac{g'(x_0)}{g(x_0)^2}.$$

To compute the derivative of $\frac{f}{g}$ at x_0 we use the product rule. We have

$$\frac{f}{g} = f\cdot\frac{1}{g} \Rightarrow \left(\frac{f}{g}\right)'(x_0) = \left(f\cdot\frac{1}{g}\right)'(x_0)$$

$$= f'(x_0)\frac{1}{g(x_0)} + f(x_0)\left(\frac{1}{g}\right)'(x_0)$$

$$= f'(x_0)\frac{1}{g(x_0)} - f(x_0)\frac{g'(x_0)}{g(x_0)^2} = \frac{f'(x_0)g(x_0)-f(x_0)g'(x_0)}{g(x_0)^2}.$$

□

Example 7.3.2. Let us see how the above rules work on some simple examples.

(a) Consider the polynomial function

$$p(x) = 5 - 3x^2 + 7x^5,\ \ x\in\mathbb{R}.$$

From the scalar multiplication rule and the Examples 7.2.1, 7.2.2 we deduce that each of the functions 5, $-3x^2$ and $7x^5$ is differentiable and the addition rule implies that their sum is differentiable as well. We deduce

$$p'(x) = (5)' + (-3x^2)' + (7x^5)' = -6x + 35x^4.$$

(b) From the equalities

$$\frac{d}{dx}(\sin x) = \cos x,\ \ \frac{d}{dx}(\cos x) = -\sin x$$

and the scalar multiplication rule we deduce that the trigonometric functions are smooth and we have

$$\frac{d^2}{dx^2}\sin x = -\sin x,\ \ \frac{d^2}{dx^2}\cos x = -\cos x,$$
$$\frac{d^4}{dx^4}\sin x = \sin x,\ \ \frac{d^4}{dx^4}\cos x = \cos x.$$

(c) If a is a positive real number, then

$$\log_a x = \frac{\ln x}{\ln a}$$

and we deduce

$$(\log_a x)' = \frac{1}{x \ln a}. \tag{7.3.1}$$

(d) If n is a natural number, then the function

$$f : \mathbb{R}\backslash\{0\} \to \mathbb{R},\ \ f(x) = \frac{1}{x^n} = x^{-n}$$

is differentiable by the quotient rule and we have

$$(x^{-n})' = \left(\frac{1}{x^n}\right)' = -\frac{nx^{n-1}}{x^{2n}} = -\frac{n}{x^{n+1}} = -nx^{-n-1}.$$

(e) From the quotient rule we deduce

$$\frac{d}{dx}\tan x = \frac{d}{dx}\left(\frac{\sin x}{\cos x}\right) = \frac{\cos^2 x + \sin^2 x}{\cos x^2} = \frac{1}{\cos^2 x} = 1 + \tan^2 x.$$

Thus

$$(\tan x)' = 1 + \tan^2 x = \frac{1}{\cos^2 x}. \tag{7.3.2}$$

(f) Using the product rule we deduce

$$\frac{d}{dx}(e^x \sin x) = e^x \sin x + e^x \cos x.$$

The above simple rules are unfortunately not powerful enough to allow us to compute the derivative of simple functions such as $e^{\sqrt{x}}$, $x > 0$ or $\sqrt{2 + \sin x}$. For this we need a more powerful technology. □

Theorem 7.3.3 (Chain Rule)**.** *Let I, J be two nontrivial intervals of the real axis. Suppose that we are given two functions $u : I \to \mathbb{R}$ and $f : J \to \mathbb{R}$ and a point $x_0 \in I$ with the following properties.*

(i) *The range of the function u is contained in the interval J, i.e., $u(I) \subset J$.*

(ii) *The function u is differentiable at x_0.*

(iii) *The function f is differentiable at $u_0 := u(x_0)$.*

Then the composition

$$f \circ u : I \to \mathbb{R},\ \ f \circ u(x) = f\big(u(x)\big)$$

is differentiable at x_0 and

$$\boxed{(f \circ u)'(x_0) = f'(u_0)u'(x_0).}$$

Proof. Let us begin by giving a flawed proof. We have

$$\frac{f(u(x))-f(u(x_0))}{x-x_0}=\frac{f(u(x))-f(u(x_0))}{u(x)-u(x_0)}\cdot\frac{u(x)-u(x_0)}{x-x_0}.$$

Since u is differentiable at x_0 we have

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

Thus

$$\begin{aligned}&\lim_{x\to x_0}\frac{f(u(x))-f(u(x_0))}{u(x)-u(x_0)}\cdot\frac{u(x)-u(x_0)}{x-x_0}\\ &=\left(\lim_{u(x)\to u(x_0)}\frac{f(u(x))-f(u(x_0))}{u(x)-u(x_0)}\right)\cdot\left(\lim_{x\to x_0}\frac{u(x)-u(x_0)}{x-x_0}\right)\\ &=f'(u(x_0))u'(x_0)\end{aligned}$$

et voilà, we're done!

Unfortunately the above argument has one serious flaw. More precisely it is possible that $u(x)=u(x_0)$ for infinitely many values of x close to x_0. The quotient

$$\frac{f(u(x))-f(u(x_0)}{u(x)-u(x_0)}$$

is ill-defined and thus the above argument is meaningless. Although problematic, the above argument displays the strategy of the proof. We need a bit of technical contortionism to avoid the problem of vanishing denominators. The details follow below.

Since f is differentiable at u_0 we deduce that it is linearly approximable at x_0. From (7.1.4) we deduce that

$$f(u)=f(u_0)+f'(u_0)(u-u_0)+r(u),\ \ r(u)=o(u-u_0)\ \text{ as } u\to u_0.$$

Recall that the equality

$$r(u)=o(u-u_0)\ \text{ as } u\to u_0$$

signifies that

$$\lim_{u\to u_0}\frac{r(u)}{u-u_0}=0. \tag{7.3.3}$$

In particular, we deduce that

$$f\big(\,u(x)\,\big)-f\big(\,u(x_0)\,\big)=f\big(\,u(x)\,\big)-f(\,u_0\,)=f'(u_0)\big(\,u(x)-u(x_0)\,\big)+r(\,u(x)\,)$$

$$\frac{f\big(\,u(x)\,\big)-f\big(\,u(x_0)\,\big)}{x-x_0}=f'(u_0)\frac{\big(\,u(x)-u(x_0)\,\big)}{x-x_0}+\frac{r(u(x))}{x-x_0}.$$

Observe that if we prove that

$$\lim_{x\to x_0}\frac{r(\,u(x)\,)}{x-x_0}=0, \tag{7.3.4}$$

then we deduce

$$\lim_{x\to x_0}\frac{f\big(\,u(x)\,\big)-f\big(\,u(x_0)\,\big)}{x-x_0}=f'(u_0)\lim_{x\to x_0}\frac{\big(\,u(x)-u(x_0)\,\big)}{x-x_0}=f'(u_0)u'(x_0)$$

which is the claim of the theorem.

Why do we expect (7.3.4) to be true? We have $r(u(x)) = o(u(x) - u_0)$, i.e., $r(u(x))$ is a tiny fraction of $u(x) - u_0$ if $u(x)$ is close to x. When x is close to x_0, then $u(x)$ is close to u_0 so $r(u(x))$ is a tiny fraction of $u(x) - u_0$ when x is close to x_0.

On the other hand, when x is close to x_0 we have

$$\begin{aligned} u(x) - u_0 = u'(x_0)(x - x_0) + o(x - x_0) &= u'(x_0)(x - x_0) + \text{tiny fraction of } x - x_0 \\ &= (x - x_0)(u'(x_0) + \text{tiny number}). \end{aligned}$$

Thus when x is close to x_0 the remainder $r(u(x))$ is a tiny fraction of $(x-x_0)(u'(x_0)+\text{tiny number})$ which in turn is obviously a tiny fraction of $(x - x_0)$. The precise proof is presented below.

To prove (7.3.4) it suffices to show that

$$\forall \hbar > 0 \;\; \exists d = d(\hbar) > 0 : \;\; |x - x_0| < d(\hbar) \Rightarrow \frac{|r(u(x))|}{|x - x_0|} \leqslant \hbar. \tag{7.3.5}$$

The function u is differentiable at x_0 and it is linearizable at this point. Hence

$$u(x) - u_0 = u'(x_0)(x - x_0) + \rho(x), \;\; \rho(x) = o(x - x_0) \;\text{ as } x \to x_0.$$

Since $\rho(x) = o(x - x_0)$ as $x \to x_0$ we deduce that there exists a small $\gamma > 0$ such that

$$|x - x_0| < \gamma \Rightarrow |\rho(x)| \leqslant |x - x_0|.$$

Hence, for $|x - x_0| < \gamma$ we have

$$\begin{aligned} |u(x) - u_0| &= |u'(x_0)(x - x_0) + \rho(x)| \\ &\leqslant |u'(x_0)||x - x_0| + |\rho(x)| \leqslant (|u'(x_0)| + 1)|x - x_0|. \end{aligned}$$

If we set $C := |u'(x_0)| + 1 > 0$, then we deduce

$$|x - x_0| < \gamma \Rightarrow |u(x) - u_0| \leqslant C|x - x_0|. \tag{7.3.6}$$

Note that (7.3.3) implies that

$$\forall \hbar > 0 \;\; \exists \varepsilon(\hbar) > 0 : \;\; |u - u_0| < \varepsilon(\hbar) \Rightarrow |r(u)| \leqslant \hbar|u - u_0|. \tag{7.3.7}$$

Observe that (7.3.6) implies

$$|x - x_0| < \delta(\hbar) := \min\Big\{\gamma, \frac{\varepsilon(\hbar)}{c}\Big\} \Rightarrow |u(x) - u_0| \leqslant C|x - x_0| < \varepsilon(\hbar).$$

Using this in (7.3.7) we deduce that

$$|x - x_0| < \delta(\hbar) \Rightarrow |u(x) - u_0| < \varepsilon(\hbar) \overset{(7.3.7)}{\Rightarrow} |r(\,u(x)\,)| \leqslant \hbar|u(x) - u_0| \overset{(7.3.6)}{\leqslant} C\hbar|x - x_0|.$$

We have thus proved that

$$\forall \hbar > 0 \;\; \exists \delta(\hbar) > 0 : \;\; |x - x_0| < \delta(\hbar) \Rightarrow \frac{|r(\,u(x)\,)|}{|x - x_0|} \leqslant C\hbar.$$

If we set

$$d(\hbar) := \delta(\hbar/C)$$

we obtain (7.3.5).

□

Remark 7.3.4. Since the chain rule is without a doubt the key rule in differential calculus it is perhaps appropriate to pause and provide a bit of intuition behind it. The classical point of view on this formula is in our view the most intuitive.

Before the modern concept of function (late 19th century) functions were regarded as quantities that depend on other quantities. In the chain rule we deal with three quantities denoted by x, u, f. The quantity u depends on the quantity x thus giving us the function $u = u(x)$. The quantity f depends on the quantity u thus giving us the function $f = f(u)$. Since u also depends on x, we deduce that through u as intermediary the function f also depends on x, thus giving us the composition $f \circ u$.

The derivative of $f \circ u$ with respect to x measures the rate of change in the quantity f per unit of change in x. The classics would denote this rate of change by $\frac{df}{dx}$ instead of the more complete, but more cumbersome[1] $\frac{df\circ u}{dx}$. The quantity $\frac{df}{du}$ denotes the rate of change in f per unit of change in u, The quantity $\frac{du}{dx}$ is defined in a similar fashion and the chain rule takes the simpler form

$$\boxed{\frac{df}{dx} = \frac{df}{du} \cdot \frac{du}{dx}}. \tag{7.3.8}$$

A less rigorous but more intuitive way of phrasing the above equality is

$$\frac{\text{change in } f}{\text{change in } x} = \frac{\text{change in } f}{\text{change in } u} \cdot \frac{\text{change in } u}{\text{change in } x}.$$

□

Let us see the chain rule at work in some simple examples.

Example 7.3.5. (a) Consider the function

$$\sin\sqrt{x}, \quad x > 0.$$

It is the composition of the two functions

$$f(u) = \sin u, \quad u(x) = \sqrt{x}.$$

Then

$$\frac{d}{dx}\sin\sqrt{x} = \frac{df}{du} \cdot \frac{du}{dx} = (\cos u) \cdot \frac{1}{2\sqrt{x}} = \frac{\cos\sqrt{x}}{2\sqrt{x}}.$$

(b) Consider the function 2^x. We have

$$2^x = (e^{\ln 2})^x = e^{(\ln 2)x}.$$

It is the composition of two functions

$$f(u) = e^u, \quad u(x) = (\ln 2)x.$$

Then

$$\frac{d}{dx}2^x = \frac{df}{du} \cdot \frac{du}{dx} = e^u(\ln 2) = e^{(\ln 2)x}\ln 2 = 2^x \ln 2.$$

[1] The concept of composition of function was not clearly defined given that the concept of function was nebulous.

More generally, if a is a positive real number then

$$\frac{d}{dx}a^x = a^x \ln a. \tag{7.3.9}$$

Observe that for any $\lambda \in \mathbb{R}$ we have

$$\frac{d}{dx}e^{\lambda x} = \lambda e^{\lambda x},$$

and we conclude inductively that

$$\frac{d^n}{dx^n}e^{\lambda x} = \lambda^n e^{\lambda x}, \quad \forall n \in \mathbb{N}. \tag{7.3.10}$$

(c) Consider now a trickier situation. Let $f : (0, \infty) \to \mathbb{R}$ be given by $f(x) = x^x$. We want to prove that f is differentiable and then compute its derivative. We set

$$g(x) = \ln f(x) = x \ln x.$$

Clearly g is differentiable since it is the product of differentiable functions. From the equality

$$f(x) = e^{g(x)}$$

we deduce that f is also differentiable because it is the composition of differentiable functions. Using the chain rule we deduce

$$f'(x) = e^{g(x)}g'(x) = (x^x)g'(x) = x^x(\ln x + 1).$$

□

Theorem 7.3.6 (Inverse function rule)**.** *Suppose that I, J are two intervals of the real axis and $u : I \to J$ is a bijective function satisfying the following properties.*

(i) *The function u is differentiable at the point $x_0 \in I$.*

(ii) $u'(x_0) \neq 0$.

(iii) *The inverse function u^{-1} is continuous at $y_0 = u(x_0)$.*

Then the inverse function u^{-1} is differentiable at $y_0 = u(x_0)$ and

$$(u^{-1})'(y_0) = \frac{1}{u'(x_0)}.$$

Proof. Since u is bijective we deduce that for any $y \in J$, there exists a unique $x = x(y)$ in I such that $u(x) = y$. More precisely $x(y) = u^{-1}(y)$. Since u^{-1} is continuous at y_0 we have

$$\lim_{y \to y_0} x(y) = x(y_0) = x_0.$$

Then

$$\frac{u^{-1}(y) - u^{-1}(y_0)}{y - y_0} = \frac{x - x_0}{u(x) - u(x_0)} = \frac{1}{\frac{u(x)-u(x_0)}{x-x_0}}.$$

so that

$$\lim_{y\to y_0}\frac{u^{-1}(y)-u^{-1}(y_0)}{y-y_0}=\lim_{x\to x_0}\frac{1}{\frac{u(x)-u(x_0)}{x-x_0}}=\frac{1}{u'(x_0)}.$$

□

Example 7.3.7. The inverse function rule is a bit tricky to use. We discuss a few classical examples.

(a) Consider the function

$$u:(-\pi/2,\pi/2)\to(-1,1),\ \ u(x)=\sin x.$$

This function is bijective, differentiable, and the derivative $u'(x)=\cos x$ is nowhere zero. Its inverse is the continuous function

$$\arcsin:(-1,1)\to(-\pi/2,\pi/2).$$

We have

$$\frac{d}{du}\arcsin u=\frac{1}{u'(x)}=\frac{1}{\cos x},\ \ u=\sin x.$$

Observe that on the interval $(-\pi/2,\pi/2)$ the function $\cos x$ is positive so that

$$\cos x=\sqrt{1-\sin^2 x}=\sqrt{1-u^2}.$$

Hence

$$\boxed{\frac{d}{du}\arcsin u=\frac{1}{\sqrt{1-u^2}},\ \ \forall u\in(-1,1)}. \tag{7.3.11}$$

A similar argument shows that

$$\frac{d}{du}\arccos u=-\frac{1}{\sqrt{1-u^2}},\ \ \forall u\in(-1,1). \tag{7.3.12}$$

(b) Consider the bijective differentiable function

$$u:(-\pi/2,\pi/2)\to\mathbb{R},\ \ u(x)=\tan x.$$

Its inverse is the function $\arctan:\mathbb{R}\to(-\pi/2,\pi/2)$. It is continuous and

$$\frac{d}{du}\arctan u=\frac{1}{u'(x)}=\frac{1}{(\tan x)'},\ \ u=\tan x.$$

Using the equality $(\tan x)'=1+\tan^2 x$ we deduce

$$\boxed{\frac{d}{du}\arctan u=\frac{1}{1+\tan^2 x}=\frac{1}{1+u^2}}. \tag{7.3.13}$$

□

7.4. Fundamental properties of differentiable functions

The first fundamental result concerning differentiable functions is Fermat's Principle. Before we formulate it we need to introduce a new concept.

Definition 7.4.1. Suppose that $f : I \to \mathbb{R}$ is a function defined on an interval $I \subset \mathbb{R}$.

(i) A point $x_0 \in I$ is said to be a *local minimum* of f if there exists $\delta > 0$ with the following property

$$\forall x \in I,\ \ |x - x_0| < \delta \Rightarrow f(x) \geqslant f(x_0).$$

The point x_0 is called a *strict local minimum* if there exists $\delta > 0$ with the following property

$$\forall x \in I,\ \ 0 < |x - x_0| < \delta \Rightarrow f(x) > f(x_0).$$

(ii) A point $x_0 \in I$ is said to be a *local maximum* of f if there exists $\delta > 0$ with the following property

$$\forall x \in I,\ \ |x - x_0| < \delta \Rightarrow f(x) \leqslant f(x_0).$$

The point x_0 is called a *strict local maximum* if there exists $\delta > 0$ with the following property

$$\forall x \in I,\ \ 0 < |x - x_0| < \delta \Rightarrow f(x) < f(x_0).$$

(iii) A point $x_0 \in I$ is said to be a *(strict) local extremum* of f if it is either a (strict) local minimum, or a (strict) local maximum.

□

Theorem 7.4.2 (Fermat's Principle)**.** *Consider a function $f : [a, b] \to \mathbb{R}$ which is differentiable on the open interval (a, b). Suppose that x_0 is a local extremum of f* situated in the interior, *$x_0 \in (a, b)$. Then $f'(x_0) = 0$. In geometric terms, at an* interior *local extremum, the tangent line to the graph has zero slope, i.e., it is horizontal.*

Proof. Assume for simplicity that x_0 is a local minimum; see Figure 7.4. Since x_0 is in the *interior* of the interval $[a, b]$ we can find $\delta > 0$ such that

$$(x_0 - \delta, x_0 + \delta) \subset (a, b) \ \text{ and }\ f(x_0) \leqslant f(x),\ \ \forall x \in (x_0 - \delta, x_0 + \delta).$$

We have

$$\lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) = \lim_{x \nearrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Note that

$$x \in (x_0, x_0 + \delta) \Rightarrow f(x) - f(x_0) \geqslant 0 \wedge x - x_0 > 0 \Rightarrow \frac{f(x) - f(x_0)}{x - x_0} \geqslant 0 \Rightarrow$$

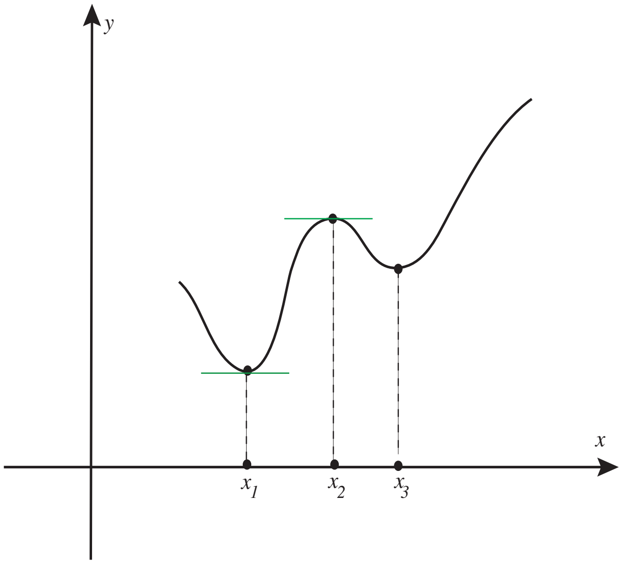

Figure 7.3. *The points x_1 and x_3 are local minima, while the point x_2 is a local maximum.*

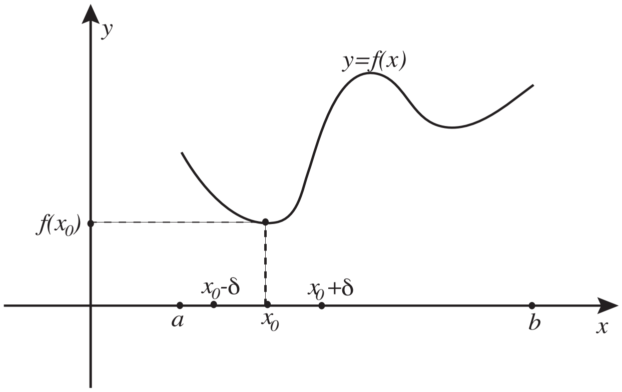

Figure 7.4. *The point x_0 is an interior local minimum.*

$$\Rightarrow \lim_{x \searrow x_0} \frac{f(x)-f(x_0)}{x-x_0} \geqslant 0 \Rightarrow f'(x_0) \geqslant 0.$$

Similarly

$$x \in (x_0-\delta, x_0) \Rightarrow f(x)-f(x_0) \geqslant 0 \wedge x-x_0 < 0 \Rightarrow \frac{f(x)-f(x_0)}{x-x_0} \leqslant 0 \Rightarrow$$

$$\Rightarrow \lim_{x \nearrow x_0} \frac{f(x)-f(x_0)}{x-x_0} \leqslant 0 \Rightarrow f'(x_0) \leqslant 0.$$

This proves that $f'(x_0)=0$. □

Remark 7.4.3. The importance of Fermat's Principle is difficult to overestimate. Locating the local extrema of a function is a problem with a huge number of applications beyond theoretical mathematics. Fermat's Principle states that the local extrema of a differentiable function $f : [a, b] \to \mathbb{R}$ are very special points: they are either endpoints of the interval, or points where the derivative of f vanishes.

This principle reduces the search of extrema to a set much much smaller than the interval $[a, b]$. Instead of looking for the needle in a haystack, we're looking for a needle hidden in a small matchbox. There is a caveat: the matchbox could be locked and it may take some ingenuity to unlock it. □

Definition 7.4.4. Suppose that $f : I \to \mathbb{R}$ is a differentiable function defined on an interval $I \subset \mathbb{R}$. A point $x_0 \in I$ is called a *critical* or *stationary* point of f if $f'(x_0) = 0$. □

We can thus rephrase Fermat's Principle as saying that ***interior*** *local extrema must be critical points*. We want to point out that not all critical points are necessarily local extrema. For example the point $x_0 = 0$ of $f(x) = x^3$, $x \in \mathbb{R}$, is a critical point of f. However it is not a local extremum because

$$f(x) > f(0)\ \ \forall x > 0\ \wedge\ \ f(x) < f(0)\ \ \forall x < 0.$$

Fermat's Principle has several fundamental consequences. We describe a few of them.

Theorem 7.4.5 (Rolle). *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function that is also differentiable on the open interval (a, b). If $f(a) = f(b)$, then there exists $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

Proof. According to Weierstrass' Theorem 6.2.4 there exist $x_*, x^* \in [a, b]$ such that

$$f(x_*) = \inf_{x \in [a,b]} f(x),\ \ f(x^*) = \sup_{x \in [a,b]} f(x). \tag{7.4.1}$$

We distinguish two cases.

1. $f(x_*) = f(x^*)$. We deduce from (7.4.1) that f is the constant function $f(x) = f(x_*)$, $\forall x \in [a, b]$. In particular $f'(x) = 0$, $\forall x \in (a, b)$, proving the claim in the theorem.

2. $f(x_*) < f(x^*)$. Thus x_* and x^* cannot simultaneously be endpoints of the interval $[a, b]$ because $f(a) = f(b)$. Hence at least one of the points x_* or x^* is located in the interior of the interval. Suppose for x_* is that point. Then x_* is a local minimum of f located in the interior of (a, b). Fermat's Principle implies that $f'(x_*) = 0$. □

Theorem 7.4.6 (Lagrange's Mean Value Theorem)**.** *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function that is also differentiable on the open interval (a, b). Then there* *exists a point $\xi \in (a,b)$ such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Geometrically this signifies that somewhere on the graph of f there exists a point so that the tangent to the graph at that point is parallel to the line connecting the endpoints of the graph of f; see Figure 7.5.

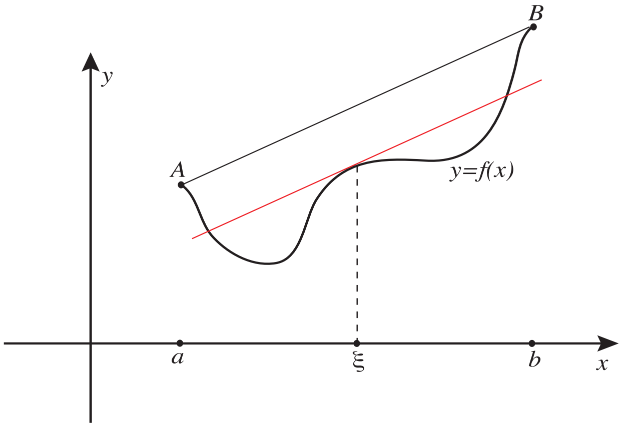

Figure 7.5. *The geometric interpretation of Theorem 7.4.6.*

Proof. We set

$$m := \frac{f(b) - f(a)}{b - a}.$$

The line passing through the points $A = (a, f(a))$ and $B = (b, f(b))$ has slope m and is the graph of the linear function

$$L(x) = m(x - a) + f(a).$$

Observe that

$$L(a) = f(a), \;\; L(b) = f(b), \;\; L'(x) = m, \;\; \forall x.$$

Define

$$g : [a,b] \to \mathbb{R}, \;\; g(x) = f(x) - L(x).$$

Note that g is continuous on $[a,b]$ and differentiable on (a,b). Moreover

$$g(a) = f(a) - L(a) = 0 = f(b) - L(b) = g(b).$$

Rolle's theorem implies that there exists $\xi \in (a,b)$ such that

$$0 = g'(\xi) = f'(\xi) - L'(\xi) = f'(\xi) - m \Rightarrow f'(\xi) = m.$$

□

Remark 7.4.7. In the Mean Value Theorem the requirement that f be continuous on the ***closed*** interval $[a,b]$ is essential and does not follow from the requirement that f be differentiable on the ***open*** interval (a,b).

In the theorem we have tacitly assumed that $a < b$. The result continues to be true even when $a > b$ because

$$\frac{f(b)-f(a)}{b-a} = \frac{f(a)-f(b)}{a-b}.$$

In this case ξ is a point in the open interval with endpoints a and b. □

Corollary 7.4.8. *Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function that is also differentiable on the open interval (a,b). Then the following statements are equivalent.*

(i) *The function f is constant.*

(ii) $f'(x) = 0$, $\forall x \in (a,b)$.

Proof. The implication (i) ⇒ (ii) is immediate since the derivative of a constant function is 0.

To prove the implication (ii) ⇒ (i) we argue by contradiction. Suppose that there exist $x_0, x_1 \in [a,b]$, such that $x_0 < x_1$ and $f(x_0) \neq f(x_1)$. The Mean Value Theorem implies that there exists $\xi \in (x_0, x_1)$ such that

$$f'(\xi) = \frac{f(x_1)-f(x_0)}{x_1-x_0} \neq 0.$$

□

Corollary 7.4.9. *Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function that is differentiable on (a,b). If $f'(x) \neq 0$ for any (a,b), then f is injective.*

Proof. If $x_0, x_1 \in [a,b]$ and $x_0 \neq x_1$, say $x_0 < x_1$, then the Mean Value Theorem implies that there exists $\xi \in (x_0, x_1)$ such that

$$f(x_1) - f(x_0) = f'(\xi)(x_1 - x_0) \neq 0.$$

This proves the injectivity of f. □

Corollary 7.4.10. *Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function that is differentiable on (a,b). Then the following statements are equivalent.*

(i) *The function f is nondecreasing.*

(ii) $f'(x) \geqslant 0$, $\forall x \in (a,b)$.

Also, the following statements are equivalent.

(iii) *The function f is nonincreasing.*

(iv) $f'(x) \leqslant 0$, $\forall x \in (a,b)$.

Proof. (i) $\Rightarrow$ (ii). Let $x_0 \in (a, b)$. Then for $h > 0$ we have $f(x_0 + h) - f(x_0) \geqslant 0$ so that

$$\frac{f(x_0+h)-f(x_0)}{h} \geqslant 0 \Rightarrow f'(x_0) = \lim_{h \searrow 0} \frac{f(x_0+h)-f(x_0)}{h} \geqslant 0.$$

(ii) $\Rightarrow$ (i). Suppose that $x_0, x_1 \in [a, b]$ are such that $x_0 < x_1$. The Mean Value Theorem implies that there exists $\xi \in (x_0, x_1)$ such that

$$f'(\xi) = \frac{f(x_1)-f(x_0)}{x_1-x_0} \Rightarrow f(x_1) - f(x_0) = f'(\xi)(x_1 - x_0) \geqslant 0.$$

□

Remark 7.4.11. If in the above result we replace (ii) with the stronger condition

$$f'(x) > 0, \;\; \forall x \in (a, b),$$

then we obtain a stronger conclusion namely that f is (strictly) increasing. This follows by coupling Corollary 7.4.10 with Corollary 7.4.9. □

Example 7.4.12. (a) We want to prove that

$$e^x \geqslant x + 1, \;\; \forall x \in \mathbb{R}. \tag{7.4.2}$$

To this aim consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x - (x + 1)$. This function is differentiable and $f'(x) = e^x - 1$.

We see that the derivative is positive on $(0, \infty)$ and negative on $(-\infty, 0)$. Hence f is increasing on $(0, \infty)$ and thus $f(x) > f(0) = 0$, $\forall x > 0$ and $f(x) > 0$, $\forall x \in (-\infty, 0)$. In other words,

$$e^x - (x + 1) \geqslant 0, \;\; \forall x \in \mathbb{R},$$

which is (7.4.2).

(b) We want to prove that

$$x \geqslant \sin x, \;\; \forall x \geqslant 0. \tag{7.4.3}$$

Consider the function $f : [0, \infty) \to \mathbb{R}$, $f(x) = x - \sin x$. This function is differentiable and

$$f'(x) = 1 - \cos x \geqslant 0, \;\; \forall x \geqslant 0.$$

Hence f is nondecreasing and thus

$$x - \sin x = f(x) \geqslant f(0) = 0, \;\; \forall x \geqslant 0.$$

(c) We want to prove that

$$\cos x \geqslant 1 - \frac{x^2}{2}, \;\; \forall x \in \mathbb{R}. \tag{7.4.4}$$

Consider the function

$$f : \mathbb{R} \to \mathbb{R}, \;\; f(x) = \cos x - \left(1 - \frac{x^2}{2}\right), \;\; \forall x \in \mathbb{R}.$$

We have to prove that $f(x) \geqslant 0$, $\forall x \in \mathbb{R}$. We observe that f is an even function, i.e., $f(-x) = f(x)$, $\forall x \in \mathbb{R}$ so it suffices to show that $f(x) \geqslant 0$, $\forall x \geqslant 0$. Note that f is differentiable and

$$f'(x) = -\sin x + x \overset{(7.4.3)}{\geqslant} 0 \;\; \forall x \geqslant 0.$$

Thus f is nondecreasing on the interval $[0, \infty)$ and we conclude that $f(x) \geqslant f(0) = 0$, $\forall x \geqslant 0$. □

Example 7.4.13 (Young's inequality)**.** Suppose that $p \in (1, \infty)$. Define $q \in (1, \infty)$ by $\frac{1}{p} + \frac{1}{q} = 1$, i.e., $q = \frac{p}{p-1}$. Consider $f : (0, \infty) \to \mathbb{R}$

$$f(x) = x^{\alpha} - \alpha x + \alpha - 1, \;\; \alpha := \frac{1}{p}.$$

We want to prove that $f(x) \leqslant 0$, $\forall x > 0$. We have

$$f'(x) = \alpha x^{\alpha-1} - \alpha = \alpha(x^{\alpha-1} - 1) = \alpha\left(\frac{1}{x^{1-\alpha}} - 1\right).$$

Observe that $f'(x) = 0$ if and only if $x = 1$. Moreover $f'(x) < 0$ for $x > 1$ and $f'(x) > 0$ for $x < 1$ because $1 - \alpha = 1 - \frac{1}{p} > 0$. Thus the function f increases on $(0, 1)$ and decreases on $(1, \infty)$ so that

$$0 = f(1) \geqslant f(x) \;\; \forall x > 0.$$

Thus

$$x^{\alpha} - \alpha x \leqslant 1 - \alpha = 1 - \frac{1}{p} > 0 = \frac{1}{q}.$$

If we choose $a, b > 0$ and we set $x = \frac{a}{b}$ we deduce

$$\left(\frac{a}{b}\right)^{\frac{1}{p}} - \frac{1}{p}\left(\frac{a}{b}\right)^{\frac{1}{p}+\frac{1}{q}} \leqslant \frac{1}{q} \Rightarrow \left(\frac{a}{b}\right)^{\frac{1}{p}} \leqslant \frac{1}{p}\left(\frac{a}{b}\right)^{\frac{1}{p}+\frac{1}{q}} + \frac{1}{q}.$$

Multiplying both sides by $b = b^{\frac{1}{p}+\frac{1}{q}}$ we deduce

$$a^{\frac{1}{p}} b^{\frac{1}{q}} \leqslant \frac{a}{p} + \frac{b}{q}, \;\; \forall a, b > 0. \tag{7.4.5}$$

If we set $u := a^{\frac{1}{p}}$, $v := b^{\frac{1}{q}}$ then we can rewrite the above inequality in the commonly encountered form

$$uv \leqslant \frac{u^p}{p} + \frac{v^q}{q}, \;\; \forall u, v > 0, \;\; p, q > 1, \;\; \frac{1}{p} + \frac{1}{q} = 1. \tag{7.4.6}$$

The last inequality is known as *Young's inequality.* □

Corollary 7.4.14. *Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function that is twice differentiable on (a, b). Let $x_0 \in (a, b)$ be a critical point of f, i.e., $f'(x_0) = 0$. Then the following hold.*

(i) *If $f''(x_0) > 0$, then x_0 is a strict local minimum of f.*

(ii) *If $f''(x_0) < 0$, then x_0 is a strict local maximum of f.*

Proof. We prove only (i). Part (ii) follows by applying (i) to the new function $-f$. Suppose that

$$f'(x_0) = 0,\ \ f''(x_0) > 0.$$

We have to prove that there exists $\delta > 0$ such that

$$0 < |x - x_0| < \delta \Rightarrow f(x) > f(x_0).$$

We have

$$\lim_{x \searrow x_0} \frac{f'(x)}{x - x_0} = \lim_{x \searrow x_0} \frac{f'(x) - f'(x_0)}{x - x_0} = f''(x_0) > 0.$$

Thus there exists $\delta_1 > 0$ such that,

$$x \in (x_0, x_0 + \delta_1) \Rightarrow \frac{f'(x)}{x - x_0} > 0 \Rightarrow f'(x) > 0.$$

The Mean Value Theorem implies that for any $x \in (x_0, x_0 + \delta_1)$ there exists $\xi \in (x_0, x)$ such that

$$f(x) - f(x_0) = f'(\xi)(x - x_0).$$

Since $\xi \in (x_0, x_0 + \delta_1)$ we have $f'(\xi) > 0$ and thus $f'(\xi)(x - x_0) > 0$.

Similarly

$$\lim_{x \nearrow x_0} \frac{f'(x)}{x - x_0} = \lim_{x \nearrow x_0} \frac{f'(x) - f'(x_0)}{x - x_0} = f''(x_0) > 0.$$

Thus there exists $\delta_2 > 0$ such that,

$$x \in (x_0 - \delta_2, x_0) \Rightarrow \frac{f'(x)}{x - x_0} > 0 \rightarrow f'(x) < 0.$$

Hence if $x \in (x_0 - \delta_2, x_0)$, then the Mean Value Theorem implies that there exists $\eta \in (x, x_0) \subset (x_0 - \delta_2, x_0)$ such that

$$f(x) - f(x_0) = f'(\eta)(x - x_0) > 0.$$

If we let $\delta := \min(\delta_1, \delta_2)$, then we deduce

$$0 < |x - x_0| < \delta \Rightarrow f(x) > f(x_0).$$

□

Example 7.4.15. Here is a simple application of the above corollary. Fix a positive number a. Consider the function

$$f : [0, a] \to \mathbb{R},\ \ f(x) = x(a - x)^2.$$

We want to find the maximum possible value of this function. It is achieved either at one of the end points $0, a$ or at some interior point x_0. Note that $f(0) = f(a) = 0$ and $f(x) \geqslant 0$, $\forall x \in [0, a]$, so there must exist an interior maximum which must be a critical point. To find the critical points of f we need to solve the equation $f'(x) = 0$. We have

$$f'(x) = (a - x)^2 - 2x(a - x) = x^2 - 2ax + a^2 - 2ax + 2x^2 = 3x^2 - 4ax + a^2.$$

The discriminant of the quadratic equation $3x^2 - 4ax + a^2 = 0$ is

$$\Delta = 16a^2 - 12a^2 = 4a^2 > 0$$

Thus this quadratic equation has two roots

$$x_\pm = \frac{4a \pm 2a}{6} = a, \frac{a}{3}.$$

Only one of these roots is in the interval $(0, a)$, namely $\frac{a}{3}$. Note that

$$f''(x) = 6x - 4a, \;\; f''(a/3) = 2a - 4a < 0.$$

Thus $a/3$ is the unique maximum point of f, and thus it is absolute maximum point. We have

$$f(x) \leqslant f(a/3) = \frac{4a^3}{27}, \;\; \forall x \in [0, a]. \qquad \square$$

Theorem 7.4.16 (Cauchy's finite increment theorem)**.** *Suppose that $f, g : [a, b] \to \mathbb{R}$ are two continuous functions that are differentiable on (a, b). Then there exists $\xi \in (a, b)$ such that*

$$f'(\xi)\big(\, g(b) - g(a)\,\big) = g'(\xi)\big(\, f(b) - f(a)\,\big). \tag{7.4.7}$$

In particular, if $g'(t) \neq 0$ for any $t \in (a, b)$, then $g(b) \neq g(a)$ and

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}. \tag{7.4.8}$$

Proof. Consider the function $F : [a, b] \to \mathbb{R}$ defined by

$$F(x) = f(x) \underbrace{\big(\, g(b) - g(a)\,\big)}_{=:\Delta_g} - g(x) \underbrace{\big(\, f(b) - f(a)\,\big)}_{=:\Delta_f}, \;\; \forall x \in [a, b].$$

This function is continuous on $[a, b]$ and differentiable on (a, b). Moreover

$$F(b) - F(a) = \big(\, f(b)\Delta_g - g(b)\Delta_f\,\big) - \big(\, f(a)\Delta_g - g(a)\Delta_f\,\big)$$
$$= \big(\, f(b) - f(a)\,\big)\Delta_g + \big(\, g(a) - g(b)\,\big)\Delta_f = 0.$$

Rolle's theorem implies that there exists $\xi \in (a, b)$ such that $F'(\xi) = 0$. This proves (7.4.7). To obtain (7.4.8) we observe that the assumption $g'(t) \neq 0$ for any $t \in (a, b)$ implies that g is injective and thus $g(b) \neq g(a)$. Dividing both sides of (7.4.7) by $g(b) - g(a)$ we deduce (7.4.8). □

Remark 7.4.17. In the above theorem we have tacitly assumed that $a < b$. The result continues to be true even when $a > b$ because

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f(a) - f(b)}{g(a) - g(b)}.$$

In this case ξ is a point in the open interval with endpoints a and b. □

If $f : I \to \mathbb{R}$ is a function differentiable on the interval I, then its derivative $f' : I \to \mathbb{R}$ need not be continuous. However, the derivative is very close to being continuous in the sense that it satisfies the *intermediate value property*, just like continuous functions do.

Theorem 7.4.18 (Darboux)**.** *Suppose that I is an interval of the real axis and $f : I \to \mathbb{R}$ is a differentiable function. Then the derivative f' satisfies the intermediate value property: given $a, b \in I$, $a < b$, and a number γ strictly between $f'(a)$ and $f'(b)$, there exists a number $\xi \in (a, b)$ such that $f'(\xi) = \gamma$.* □

Exercise 7.1 will guide you toward a proof of this theorem which is also a consequence of Fermat's principle.

7.5. Table of derivatives

Table 7.1 summarizes the derivatives of the most frequently encountered functions.

$f(x)$	$f'(x)$
x^n, $(x \in \mathbb{R},\ n \in \mathbb{N})$	nx^{n-1}
x^{-n} $(x \neq 0,\ n \in \mathbb{N})$	$-nx^{-n-1}$
x^α, $(\alpha \in \mathbb{R},\ x > 0)$	$\alpha x^{\alpha-1}$
$\sqrt{x}$, $(x > 0)$	$\frac{1}{2\sqrt{x}}$
$\ln x$	$1/x$
e^x, $(x \in \mathbb{R})$	e^x
a^x, $(a > 0,\ x \in \mathbb{R})$	$a^x \ln a$
$\sin x$, $(x \in \mathbb{R})$	$\cos x$
$\cos x$, $(x \in \mathbb{R})$	$-\sin x$
$\tan x$, $(\cos x \neq 0)$	$1 + \tan^2 x = \frac{1}{\cos^2 x}$
$\arcsin x$, $x \in (-1, 1)$	$\frac{1}{\sqrt{1-x^2}}$
$\arccos x$, $x \in (-1, 1)$	$-\frac{1}{\sqrt{1-x^2}}$
$\arctan x$, $(x \in \mathbb{R})$	$\frac{1}{1+x^2}$
$\sinh x$, $(x \in \mathbb{R})$	$\cosh x$
$\cosh x$, $(x \in \mathbb{R})$	$\sinh x$

Table 7.1. Table of derivatives.

The *hyperbolic functions* $\sinh x$ and $\cosh x$ are defined by the equalities

$$\cosh x := \frac{e^x + e^{-x}}{2}, \quad \sinh x = \frac{e^x - e^{-x}}{2}.$$

The function sinh is called the *hyperbolic sine* while the function cosh is called the *hyperbolic cosine*.

7.6. Exercises

Exercise 7.1. Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$.

(i) Sketch the graph of f.

(ii) Show that f is not differentiable at 0.

(iii) Show that f is differentiable at any point $x_0 \neq 0$ and then compute the derivative of f at x_0.

□

Exercise 7.2. Prove Proposition 7.1.7. □

Exercise 7.3. Imitate the strategy in Example 7.2.6 to prove

$$\lim_{h\to 0} \frac{\cos(x_0 + h) - \cos x_0}{h} = -\sin x_0.$$

Hint. You need to use the trigonometric identities (5.7.1a) and (5.7.1c). □

Exercise 7.4. Consider the function $f : (-\pi/2, \pi/2) \to \mathbb{R}$, $f(x) = \tan x$. Write the equation of the tangent line to the graph of f at the point $(\,\pi/4, f(\pi/4)\,)$. □

Exercise 7.5. Suppose that the functions $f, g : I \to \mathbb{R}$ are n-times differentiable. Prove that their product $f \cdot g$ is also n-times differentiable and satisfies the generalized product rule

$$\frac{d^n}{dx^n}(fg) = \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)} g^{(k)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}, \tag{7.6.1}$$

where we defined $f^{(0)} := f$, $g^{(0)} = g$.

Hint. Argue by induction on n. At some point you need to use the Pascal formula (3.2.5),

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1},$$

also used in the proof of Newton's binomial formula (3.2.4). □

Exercise 7.6. Let n be a natural number. A real number r is said to be a *root* of order n of a polynomial $P(x)$ if there exists a polynomial $Q(x)$ with the following properties:

- $P(x) = (x - r)^n Q(x)$, $\forall x \in \mathbb{R}$.
- $Q(r) \neq 0$.

(a) Prove that if $n > 1$ and r is a root of $P(x)$ of order n, then r is also a root of order $(n-1)$ of $P'(x)$.

(b) Prove that for any natural numbers $k < n$ the real numbers ±1 are roots of order $(n-k)$ of the polynomial

$$\frac{d^k}{dx^k}(x^2-1)^n.$$

(c) For any natural number n we define the n-th *Legendre polynomial* to be

$$P_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}\big(x^2-1\big)^n.$$

Use (7.6.1) to compute $P_n(\pm 1)$. □

Exercise 7.7. Consider the continuous function $f : [0,\infty) \to \mathbb{R}$, $f(x) = \sqrt{x}$. Show that f is not differentiable at 0. □

Exercise 7.8. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & |x| \geqslant 1 \\ e^{-T(x)}, & |x| < 1, \end{cases} \quad \text{where } \; T(x) = \frac{1}{1-x^2}, \;\; \forall |x| < 1.$$

(a) Set

$$F_n(x) := \frac{d^n}{dx^n}\big(\, e^{-T(x)}\,\big), \;\; \forall |x| < 1.$$

Prove by induction that for any $n \in \mathbb{N}$ there exists a polynomial $P_n(x)$ and a natural number k_n such that

$$F_n(x) = P_n(x)T(x)^{k_n}e^{-T(x)}, \;\; \forall |x| < 1.$$

Hint. Observe that

$$T'(x) = 2xT(x)^2.$$

(b) Prove that f is a smooth function, i.e., infinitely many times differentiable.

Hint. Prove by induction that

$$f^{(n)}(x) = \begin{cases} 0, & |x| \geqslant 1, \\ F_n(x), & |x| < 1. \end{cases}$$

For the inductive step observe that for $|x| < 1$ we have

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

$$\frac{f^{(n)}(x) - f^{(n)}(1)}{x-1} = \frac{F_n(x)}{x-1} = -(x+1)T(x)F_n(x) = (x+1)P_n(x)T(x)^{k_n+1}e^{-T(x)},$$

$$\frac{F_n(x)}{x+1} = (x-1)T(x)F_n(x) = -(x-1)P_n(x)T(x)^{k_n+1}e^{-T(x)}.$$

Then

$$\begin{aligned} \lim_{x\nearrow 1}\frac{f^{(n)}(x) - f^{(n)}(1)}{x-1} &= -\lim_{x\nearrow 1}\frac{F_n(x)}{x-1} = \Big(\lim_{x\nearrow 1}(x+1)P_n(x)\Big)\cdot\Big(\lim_{x\nearrow 1}T(x)^{k_n+1}e^{-T(x)}\Big) \\ &= -2P_n(1)\Big(\lim_{x\nearrow 1}T(x)^{k_n+1}e^{-T(x)}\Big). \end{aligned}$$

Now observe that

$$\lim_{x\nearrow 1}T(x) = \infty.$$

Use the result in Exercise 5.11 (b) to deduce

$$\lim_{x\nearrow 1} T(x)^{k_n+1}e^{-T(x)} = 0.$$

□

Exercise 7.9. [2] Fix a natural number n and real numbers p, q.

(a) Prove that for any $t \in \mathbb{R}$ we have

$$np(tp+q)^{n-1} = \sum_{k=1}^{n} k\binom{n}{k}t^{k-1}p^kq^{n-k},$$

$$n(n-1)p^2(tp+q)^{n-2} = \sum_{k=2}^{n} k(k-1)\binom{n}{k}t^{k-2}p^kq^{n-k}.$$

Hint. Consider the function

$$f_n : \mathbb{R} \to \mathbb{R},\ \ f_n(t) = \big(tp+q\big)^n.$$

Compute the derivatives $f_n'(t)$, $f_n''(t)$. Then describe $f_n(t)$ using Newton's binomial formula and compute the same derivatives using the new description of $f_n(t)$.

(b) For any integer k, $0 \leqslant k \leqslant n$, and any $x \in \mathbb{R}$ set $w_k(x) := \binom{n}{k}x^k(1-x)^{n-k}$. Use part (a) to prove that for any $x \in \mathbb{R}$

$$1 = \sum_{k=0}^{n} w_k(x) \tag{7.6.2a}$$

$$nx = \sum_{k=0}^{n} kw_k(x), \tag{7.6.2b}$$

$$n(n-1)x^2 = \sum_{k=0}^{n} k(k-1)w_k(x) = \sum_{k=0}^{n} k^2w_k(x) - nx, \tag{7.6.2c}$$

$$nx(1-x) = \sum_{k=0}^{n}(k-nx)^2w_k(x). \tag{7.6.2d}$$

Hint. Use the results in (a) in the special case $p = x$, $q = 1-x$, $t = 1$. □

Exercise 7.10. Find the extrema and the intervals on which the following functions are increasing.

(i) $f(x) = \sqrt{x} - 2\sqrt{x+2}$, $x > 0$.

(ii) $g(x) = \frac{x}{x^2+1}$, $x \in \mathbb{R}$.

□

[2] The results in this exercise are very useful in probability theory.

Exercise 7.11. Suppose that $f : [a, b] \to \mathbb{R}$ is continuous and differentiable on (a, b). Show that if

$$\lim_{x \to a} f'(x) = A,$$

then f is differentiable at a and $f'(a) = A$. □

Exercise 7.12. Prove that if $f : I \to \mathbb{R}$ is a differentiable function defined on an interval I, and the derivative f' is bounded on I, then f is a Lipschitz function, i.e.,

$$\exists L > 0, \;\; \forall x, y \in I \;\; |f(x) - f(y)| \leqslant L|x - y|.$$ □

Exercise 7.13. Use the Mean Value Theorem to prove that

$$|\sin(x) - \sin(y)| \leqslant |x - y|, \;\; \forall x, y \in \mathbb{R}.$$ □

Exercise 7.14. Fix a real number λ and suppose that $u : \mathbb{R} \to \mathbb{R}$ is a differentiable function satisfying the differential equation

$$u'(t) = \lambda u(t), \;\; \forall t \in \mathbb{R}.$$

Prove that there exists a constant $c \in \mathbb{R}$ such that $u(t) = ce^{\lambda t}$, $\forall t \in \mathbb{R}$.

Hint. Show that the function $f(t) = e^{-\lambda t} u(t)$, $t \in \mathbb{R}$ is constant. □

Exercise 7.15. Suppose that b, c are real numbers and $u, v : \mathbb{R} \to \mathbb{R}$ are twice differentiable functions satisfying the differential equation

$$u''(t) + bu'(t) + cu(t) = 0 = v''(t) + bv'(t) + cv(t), \;\; \forall t \in \mathbb{R}.$$

Define the *Wronskian* to be the function

$$W(t) = u(t)v'(t) - u'(t)v(t), \;\; t \in \mathbb{R}.$$

Prove that

$$W'(t) + bW(t) = 0$$

and deduce that

$$W(t) = W(0)e^{-bt}.$$

Hint. You may want to use Exercise 7.14. □

Exercise 7.16. (a) Suppose that $u : \mathbb{R} \to \mathbb{R}$ is a twice differentiable function satisfying the differential equation

$$u''(t) + u(t) = 0, \;\; \forall t \in \mathbb{R}. \tag{7.6.3}$$

Prove that

$$u'(t)^2 + u(t)^2 = u'(0)^2 + u(0)^2, \;\; \forall t \in \mathbb{R}.$$

(b) Suppose that $u, v : \mathbb{R} \to \mathbb{R}$ are twice differentiable functions satisfying the differential equation (7.6.3), i.e.,

$$u''(t) + u(t) = 0 = v''(t) + v(t), \;\; \forall t \in \mathbb{R}.$$

Show that the difference $w(t) = u(t) - v(t)$ also satisfies the differential equation (7.6.3). Use part (a) to prove that if $u(0) = v(0)$ and $u'(0) = v'(0)$, then $u(t) = v(t)$, $\forall t \in \mathbb{R}$.

(c) Can you think of a function $u : \mathbb{R} \to \mathbb{R}$ satisfying (7.6.3) and the initial conditions

$$u(0) = 0, \;\; u'(0) = 1?$$ □

Exercise 7.17. (a) Prove that for any real number $\alpha \geqslant 1$ and any $x > -1$ we have

$$(1+x)^\alpha \geqslant 1 + \alpha x.$$

(b) Prove by induction that for any natural number n and any $x \geqslant 0$ we have

$$1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} \leqslant e^x.$$

Hint. Have a look at Example 7.4.12. □

Exercise 7.18. Prove that

$$\sin x \geqslant x - \frac{x^3}{6}, \quad \forall x \geqslant 0.$$

Hint. Have a look at Example 7.4.12. □

Exercise 7.19. Prove that the function

$$f : (0, \infty) \to \mathbb{R}, \quad f(x) = \left(1 + \frac{1}{x}\right)^x$$

is increasing. □

Exercise 7.20. Use Lagrange's mean value theorem to show that for any $x > 0$ we have

$$\frac{1}{x+1} < \ln(x+1) - \ln x < \frac{1}{x}.$$

Conclude that

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} > \ln(n+1), \quad \forall n \in \mathbb{N}.$$ □

Exercise 7.21. Fix a real number $s \in (0,1)$. Prove that for any $x > 0$ we have

$$(1+x)^{1-s} - x^{1-s} < \frac{1-s}{x^s}.$$

Conclude that

$$1 + \frac{1}{2^s} + \cdots + \frac{1}{n^s} > \frac{1}{1-s}\big((n+1)^{1-s} - 1\big).$$ □

Exercise 7.22. Find the maximum possible volume of an open rectangular box that can be obtained from a square sheet of cardboard with a 6 ft side by cutting squares at each of the corners and bending up the ends of the resulting cross-like figure; see Figure 7.6. □

Exercise 7.23. Prove that among all the rectangles with given perimeter P the square has the largest area. □

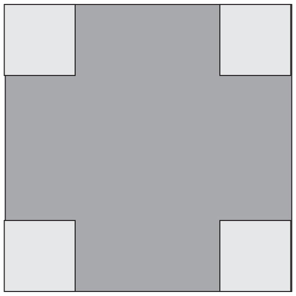

Figure 7.6. *Cutting out a box.*

Exercise 7.24. Suppose that $f : [-1, 1] \to \mathbb{R}$ is a differentiable function.

(a) Prove that if f is even, i.e., $f(x) = f(-x)$, $\forall x \in [-1, 1]$, then $f'(x)$ is odd, $f'(-x) = -f'(x)$, $\forall x \in [-1, 1]$. In particular, $f'(0) = 0$.

(b) Prove that if f is odd, then f' is even. □

Exercise 7.25. Fix a natural number n and suppose that $f : (a, b) \to \mathbb{R}$ is a $2n$-times differentiable function. Prove the following statements.

(a) If $x_0 \in (a, b)$ satisfies

$$f'(x_0) = \cdots = f^{(2n-1)}(x_0) = 0, \;\; f^{(2n)}(x_0) > 0,$$

then x_0 is a strict local minimum of f.

(b) If $x_0 \in (a, b)$ satisfies

$$f'(x_0) = \cdots = f^{(2n-1)}(x_0) = 0, \;\; f^{(2n)}(x_0) < 0,$$

then x_0 is a strict local maximum of f.

Hint. Use proof of Corollary 7.4.14 as inspiration and prove (in case (a)) that there exists $\delta > 0$ such that for $x \in (x_0, x_0 + \delta)$ we have $f^{(k)}(x) > 0, \forall k = 1, \dots, 2n - 1$ and for $x \in (x_0 - \delta, x_0)$ we have $f^{(k)}(x) < 0$, $\forall k = 1, \dots, 2n - 1$. □

7.7. Exercises for extra credit

Exercise* 7.1 (Intermediate value property of derivatives)**.** Suppose that $f : [a, b] \to \mathbb{R}$ is a differentiable function.

(a) Prove that if $f'(a) < 0 < f'(b)$, then there exists $\xi \in (a, b)$ such that $f'(\xi) = 0$.

Hint. Think Fermat.

(b) More generally, prove that if $f'(a) < f'(b)$ and $m \in (f'(a), f'(b))$, then there exists $\xi \in (a, b)$ such that $f'(\xi) = m$. □

Exercise* 7.2. Suppose $f_n : [a,b] \to \mathbb{R}$, $n \in \mathbb{N}$, is a sequence of differentiable functions functions with the following properties.

(i) The sequence of derivatives $f_n' : [a,b] \to \mathbb{R}$ converge that converges *uniformly* to a function $g : [a,b] \to \mathbb{R}$.

(ii) The sequence $f_n : [a,b] \to \mathbb{R}$ converges *pointwisely* to a function $f : [a,b] \to \mathbb{R}$.

Prove that the following hold.

(a) The sequence $f_n : [a,b] \to \mathbb{R}$ converges *uniformly* to $f : [a,b] \to \mathbb{R}$.

(b) The function f is differentiable and $f' = g$, i.e., the sequence $f_n' : [a,b] \to \mathbb{R}$ converges uniformly to f'.

Hint. Use Exercise 6.10 and the Mean Value Theorem. □

Exercise* 7.3. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuous function such that

$$\lim_{h \searrow 0} \frac{f(x+2h) - f(x+h)}{h} = 0, \quad \forall x \in \mathbb{R}.$$

Prove that f is a constant function.

Hint. Argue by contradiction and assume there exist a, b such that $f(a) \neq f(b)$, say $f(a) < f(b)$. Consider the function $g(x) = f(x) - mx$, $m := \frac{f(b)-f(a)}{b-a}$. Note that $g(a) = g(b)$ and

$$\lim_{h \searrow 0} \frac{g(x+2h) - g(x+h)}{h} = -m < 0,$$

and prove that g admits a local maximum in $[a,b)$. □

Exercise* 7.4. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a C^2-function, i.e., twice differentiable and the second derivative is continuous. Show that if the functions f and $f^{(2)}$ are bounded on $\mathbb{R}$, then so is the function f'. □

Exercise* 7.5 (Bernstein)**.** Let $f : [0,1] \to \mathbb{R}$ be a continuous function. For any $n \in \mathbb{N}$ we denote by $B_n^f(x)$ the n-th Bernstein polynomial determined by f,

$$B_n(x) = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k}.$$

(a) Show that for any $x \in [0,1]$ we have

$$f(x) - B_n^f(x) = \sum_{k=0}^{n} \big(f(x) - f(k/n) \big) \binom{n}{k} x^k (1-x)^k.$$

(b) Show that for any $\delta \in (0,1)$ and $x \in [0,1]$ we have

$$\sum_{|k/n - x| \geqslant \delta} \binom{n}{k} x^k (1-x)^k \leqslant \sum_{k=0} \frac{(k-nx)^2}{n^2\delta^2} \leqslant \frac{x(1-x)}{n\delta^2}.$$

(c) Use (a) and (b) to prove that as $n \to \infty$ the sequence $(B_n^f(x))$ converges to $f(x)$ uniformly in $x \in [0,1]$.

Hint. Use the equalities in Exercise 7.9. □

Chapter 8

Applications of differential calculus

8.1. Taylor approximations

The concept of derivative is based on the idea of approximation. Thus, if $f : I \to \mathbb{R}$ is a differentiable function and $x_0 \in I$, then the linearization of f at x_0,

$$L(x) = f(x_0) + f'(x_0)(x - x_0),$$

is a good approximation for $f(x)$ when x is not too far from x_0. More precisely, the error

$$r(x) = f(x) - L(x)$$

is $o(x - x_0)$, much much smaller than $|x - x_0|$, which itself is small when x is close to x_0. In this section we want to refine and improve this observation.

Definition 8.1.1. Suppose that $f : I \to \mathbb{R}$ is an n-times differentiable function defined on an interval I. For $x_0 \in I$ we define the *degree n Taylor polynomial* of f at x_0 to be

$$T_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

Often the Taylor polynomial of f at $x_0 = 0$ is referred to as the *Maclaurin polynomial.*

If $f : I \to \mathbb{R}$ is a smooth function, then the series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

is called the *Taylor series* or *Taylor expansion* of the smooth function f at the point x_0. Note that if f is a polynomial, then the Taylor series is a finite sum coinciding with the Taylor polynomial of the same degree of f. □

Example 8.1.2. (a) Consider a differentiable function $f : I \to \mathbb{R}$. Then the degree 1 Taylor polynomial of f at x_0 is

$$T_1(x) = f(x_0) + f'(x_0)(x - x_0).$$

Thus, $T_1(x)$ is the linearization of f at x_0.

(b) Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$. We know that $f^{(n)}(x) = e^x$, $\forall n \in \mathbb{N}$, $x \in \mathbb{R}$ and we deduce that

$$f^{(k)}(0) = 1, \ \ \forall k \in \mathbb{N}.$$

In particular, the degree n Taylor polynomial of e^x at $x_0 = 0$ is

$$T_n(x) = 1 + \frac{x}{1!} + \cdots + \frac{x^n}{n!}.$$

The Taylor series of e^x at $x_0 = 0$ is

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

(c) Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin x$. We have

$$f^{(4k)}(x) = \sin x, \ \ f^{(4k+1)}(x) = \cos x, \ \ f^{(4k+2)}(x) = -\sin x, \ \ f^{(4k+3)}(x) = -\cos x, \ \ \forall k \geqslant 0,$$

$$f^{(4k)}(0) = 0, \ \ f^{(4k+1)}(0) = 1, \ \ f^{(4k+2)}(0) = 0, \ \ f^{(4k+3)}(0) = -1.$$

We deduce that the Taylor polynomials of $\sin x$ at $x_0 = 0$ are

$$T_1(x) = f(0) + \frac{f'(0)}{1!}x = x,$$

$$T_2(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 = x,$$

$$T_3(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 = x - \frac{x^3}{6},$$

$$T_n(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots .$$

The Taylor series of $\sin x$ at $x_0 = 0$ is

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

(d) Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos x$. We have

$$f^{(4k)}(x) = \cos x, \ \ f^{(4k+1)}(x) = -\sin x, \ \ f^{(4k+2)}(x) = -\cos x, \ \ f^{(4k+3)}(x) = \sin x, \ \ \forall k \geqslant 0$$

$$f^{(4k)}(0) = 1, \ \ f^{(4k+1)}(0) = 0, \ \ f^{(4k+2)}(0) = -1, \ \ f^{(4k+3)}(0) = 0.$$

We deduce that the Taylor polynomials of $\cos x$ at $x_0 = 0$ are

$$T_1(x) = f(0) + \frac{f'(0)}{1!}x = 1,$$

$$T_2(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 = 1 - \frac{x^2}{2!},$$

$$T_3(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 = 1 - \frac{x^2}{2!},$$

$$T_n(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots .$$

The Taylor series of $\cos x$ at $x_0 = 0$ is

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

(e) Fix a real number α and define $f : (0, \infty) \to \mathbb{R}$, $f(x) = x^{\alpha}$. We have

$$f'(x) = \alpha x^{\alpha-1},\;\; f^{(2)}(x) = \alpha(\alpha-1)x^{\alpha-2},\;\; f^{(k)}(x) = \alpha(\alpha-1)\cdots(\alpha-(k-1)\,)x^{\alpha-k}.$$

We deduce that

$$f^{(k)}(1) = \alpha(\alpha-1)\cdots(\alpha-(k-1)\,)$$

and thus the degree n Taylor polynomial of x^{α} at $x_0 = 1$ is

$$T_n(x) = 1 + \frac{\alpha}{1!}(x-1) + \frac{\alpha(\alpha-1)}{2!}(x-1)^2 + \cdots + \frac{\alpha(\alpha-1)\cdots(\alpha-(n-1)\,)}{n!}(x-1)^n.$$

The coefficients of the above polynomial coincide with the binomial coefficients if α is a natural number. For this reason, for any $\alpha \in \mathbb{R}$ we introduce the notation

$$\binom{\alpha}{0} = 1,\;\; \binom{\alpha}{n} = \frac{\alpha(\alpha-1)\cdots(\alpha-(n-1)\,)}{n!},\;\; n \in \mathbb{N}.$$

The degree n Taylor polynomial of x^{α} at $x_0 = 1$ can then be described in the more compact form

$$T_n(x) = \sum_{k=0}^{n} \binom{\alpha}{k}(x-1)^{\alpha-k}.$$

□

Remark 8.1.3. The degree n Taylor polynomial of a function f at a point x_0 is the unique polynomial of degree $\leqslant n$ such that

$$T_n(x_0) = f(x_0),\;\; T_n'(x_0) = f'(x_0),\;\; T_n^{(k)}(x_0) = f^{(k)}(x_0),\;\; \forall k = 1, \dots, n.$$

Exercise 8.1 asks you to prove this fact. □

Example 8.1.2 shows that the degree 1 Taylor polynomial of a differentiable function at a point x_0 is the linear approximation of f at x_0, and we know that it provides a very good approximation for $f(x)$ if x is near x_0. The next result states that the same is true for the higher degree Taylor polynomials.

Theorem 8.1.4 (Taylor approximation)**.** *Suppose that $f : [a, b] \to \mathbb{R}$ is $(n+1)$-times differentiable, $n \in \mathbb{N}$. Fix $x_0 \in [a, b]$. We form the degree n Taylor polynomial of f at x_0*

$$T_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

and we consider the remainder (or error)

$$R_n(x_0,x) = f(x) - T_n(x), \quad x \in [a,b].$$

Fix $x \in [a,b]$, $x \neq x_0$, and a continuous function $\varphi : [x_0,x] \to \mathbb{R}$ which is differentiable on (x_0,x) and $\varphi'(t) \neq 0$, $\forall t \in (x_0,x)$. (Here we are deliberately a bit negligent and we think of $[x_0,x]$ as the closed interval with endpoints x_0, x, even in the case $x_0 > x$.)

Then there exists ξ in the open interval with endpoints x_0 and x such that

$$R_n(x_0,x) = \frac{\varphi(x)-\varphi(x_0)}{n!\varphi'(\xi)} f^{(n+1)}(\xi)(x-\xi)^n. \tag{8.1.1}$$

Proof. Consider the function $F : [x_0,x] \to \mathbb{R}$ given by

$$F(t) = f(x) - \left(f(t) + \frac{f'(t)}{1!}(x-t) + \cdots + \frac{f^{(n)}(t)}{n!}(x-t)^n \right), \quad \forall t \in [x_0,x].$$

Note that $F(x) = 0$, $F(x_0) = R_n(x_0,x)$. From Cauchy's finite increment theorem, Theorem 7.4.16, we deduce that there exists ξ in the interval (x_0,x) such that

$$\frac{F(x)-F(x_0)}{\varphi(x)-\varphi(x_0)} = \frac{F'(\xi)}{\varphi'(\xi)}.$$

Now observe that

$$\begin{aligned} -F'(t) &= f'(t) + \left(\frac{f''(t)}{1!}(x-t) - \frac{f'(t)}{1!}\right) + \left(\frac{f^{(3)}(t)}{2!}(x-t)^2 - \frac{f^{(2)}(t)}{1!}(x-t)\right) \\ &\quad + \cdots + \left(\frac{f^{(n+1)}(t)}{n!}(x-t)^n - \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}\right) \\ &= \frac{f^{(n+1)}(t)}{n!}(x-t)^n. \end{aligned}$$

Thus

$$-\frac{R_n(x_0,x)}{\varphi(x)-\varphi(x_0)} = \frac{F(x)-F(x_0)}{\varphi(x)-\varphi(x_0)} = \frac{F'(\xi)}{\varphi'(\xi)} = -\frac{f^{(n+1)}(t)(x-\xi)^n}{n!\varphi'(\xi)}$$

The last equality clearly implies (8.1.1). □

If we let $\varphi(t) = (x-t)^{n+1}$ in the above theorem, we obtain the following important consequence.

Corollary 8.1.5 (Lagrange remainder formula)**.** *There exists $\xi \in (x_0,x)$ such that*

$$\boxed{f(x) - T_n(x) = R_n(x_0,x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-x_0)^{n+1}.} \tag{8.1.2}$$

Proof. We have $\varphi(x) = 0$ and $\varphi(x)-\varphi(x_0) = -(x-x_0)^{n+1}$, $\varphi'(\xi) = -(n+1)(x-\xi)^n$. □

Remark 8.1.6. Let us explain how this works in applications. Suppose that $f : [a, b] \to \mathbb{R}$ is $(n+1)$-times differentiable and $x_0 \in [a, b]$. The degree n Taylor polynomial of f at x_0 is

$$T_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

It is convenient to introduce the notation $h = x - x_0$ so that $x = x_0 + h$ and we deduce

$$T_n(x_0 + h) = f(x_0) + \frac{f'(x_0)}{1!}h + \cdots + \frac{f^{(n)}(x_0)}{n!}h^n.$$

If h is sufficiently small, then $T_n(x_0 + h)$ is an approximation for $f(x_0 + h)$. The error of this approximation is given by the remainder $R_n(x_0, x) = f(x_0 + h) - T_n(x_0 + h)$. This remainder really depends only on the difference $h = x - x_0$ and, to emphasize this fact, we will write $R_n(x_0, h)$ instead of $R_n(x_0, x)$ in the argument below. Also, for simplicity, we will denote by $(x_0, x_0 + h)$ the open interval with endpoints x_0 and $x_0 + h$. (Note that $x_0 + h < x_0$ when $h < 0$.)

The Lagrange remainder formula tells us that there exists $\xi \in (x_0, x_0 + h)$

$$R_n(x_0, h) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) h^{n+1},$$

If we define

$$M_{n+1}(x_0, h) := \sup_{\xi \in [x_0, x_0 + h]} |f^{(n+1)}(\xi)|,$$

then we deduce

$$|\, R_n(x_0, h)\, | \leqslant \frac{M_{n+1}(x_0, h)|h|^{n+1}}{(n+1)!}. \tag{8.1.3}$$

If the right-hand side of the above inequality is small, then the error has to be small. The above result implies that

$$\big|\, f(x) - T_n(x) \,\big| = O\big(\, |x - x_0|^{n+1} \,\big) \;\text{ as } x \to x_0, , \tag{8.1.4}$$

where O is Landau's symbol defined in (5.8.1). □

Example 8.1.7. Let us show how the above remark works in a rather concrete case. Suppose $f(x) = \sin x$. We use Taylor approximations of $\sin x$ at $x_0 = 0$. For example, the degree 4 Taylor polynomial of $\sin x$ at $x_0 = 0$ is

$$T_4(h) = \sin(0) + \frac{\cos(0)}{1!}h - \frac{\sin(0)}{2!}h^2 - \frac{\cos(0)}{3!}h^3 + \frac{\sin(0)}{4!}h^4 = h - \frac{h^3}{3!} = h - \frac{h^3}{6}.$$

We have

$$\sin h \approx h - \frac{h^3}{6}.$$

To estimate the error of this approximation we use (8.1.2). The 5th derivative of $\sin x$ is $\cos x$ so that $|\cos \xi| \leqslant 1$, $\forall x \in \mathbb{R}$. We deduce from (8.1.2) that for some ξ between 0 and x we have

$$\left|\sin h - \left(h - \frac{h^3}{6}\right)\right| = \frac{|\cos \xi|}{5!}h^5 \leqslant \frac{|h|^5}{5!} = \frac{|h|^5}{120}.$$

If for example $|h| \leqslant \frac{1}{2}$, then

$$\frac{|h|^5}{120} \leqslant \frac{1}{32 \cdot 120} = \frac{1}{3840} < \frac{1}{10^3}.$$

Thus for $|h| \leqslant \frac{1}{2}$ the expression $h - \frac{h^3}{6}$ approximates $\sin h$ up to two decimals. For example

$$0.5 - (0.5)^3/6 = 0.47916... \Rightarrow \sin 0.5 = 0.47...$$

If $h = \frac{1}{4}$, then

$$\frac{|h|^5}{120} = \frac{1}{4^5 \cdot 120} = \frac{1}{1024 \cdot 120} = \frac{1}{122880} \leqslant \frac{1}{10^5},$$

and $0.25 - (0.25)^3/6$ computes $\sin(0.25)$ up to four decimals. Thus

$$0.25 - (0.25)^3/6 = 0.248666... \Rightarrow \sin(0.25) = 0.2486....$$

In Figure 8.1 we have depicted side-by-side the graph of $\sin(x)$ for $|x| \leqslant 10$ and the graph of $T_7(x)$, its degree 7 Taylor approximation at $x_0 = 0$. While $T_7(x)$ takes large values for $|x|$ large, it matches very well the graph of $\sin x$ on the interval $[-3, 3]$. □

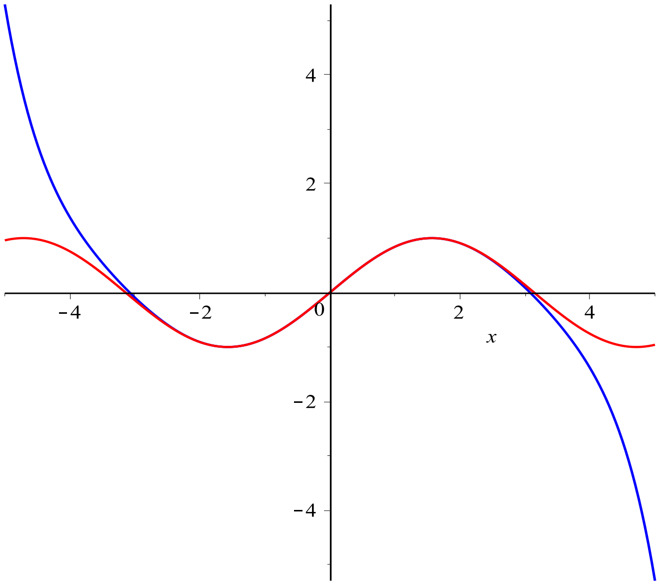

Figure 8.1. *The graphs of* $\sin x$ *and its degree 7 Taylor approximation at the origin.*

Here is a nice consequence of Corollary 8.1.5.

Corollary 8.1.8. *For any $x \in \mathbb{R}$ we have*

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{8.1.5}$$

Note that for $x = 1$ the above equality specializes to (4.6.6).

Proof. Observe that for any natural number n the partial sum

$$s_n(x) = 1 + \frac{x}{1!} + \cdots + \frac{x^n}{n!}$$

is the n-th Taylor polynomial of e^x at $x_0 = 0$. Corollary 8.1.5 implies that there exists a real number ξ_n between 0 and x such that

$$e^x - s_n(x) = e^{\xi_n} \frac{x^{n+1}}{(n+1)!}.$$

Observe that since $-|x| \leqslant \xi_n \leqslant |x|$ we have $e^{\xi_n} \leqslant e^{|x|}$ so that

$$\big| e^x - s_n(x) \big| \leqslant e^{|x|} \frac{|x|^{n+1}}{(n+1)!}. \tag{8.1.6}$$

From (4.2.8) we deduce that

$$\lim_{n\to\infty} \frac{|x|^{n+1}}{(n+1)!} = 0.$$

The Squeezing Principle then implies that

$$\lim_{n\to\infty} \big| e^x - s_n(x) \big| = 0.$$

□

Remark 8.1.9. The above proof shows a bit more namely that for any $R > 0$, the partial sums $s_n(x)$ converge to e^x *uniformly* on $[-R, R]$. Indeed, if $x \in [-R, R]$ so that $|x| \leqslant R$, then (18.4.51) implies that

$$\big| e^x - s_n(x) \big| \leqslant e^R \frac{R^{n+1}}{(n+1)!}, \quad \forall |x| \leqslant R.$$

Note that the right-hand side of the above inequality is independent of x and converges to 0 as $n \to \infty$ according to (4.2.8). Weierstrass criterion in Exercise 6.6 implies the claimed uniform convergence. □

8.2. L'Hôpital's rule

Differential calculus is also very useful in dealing with singular limits such as $\frac{0}{0}$, $\frac{\infty}{\infty}$.

Proposition 8.2.1 (L'Hôpital's Rule). *Let $a, b \in [-\infty, \infty]$, $a < b$. Suppose that the differentiable functions $f, g : (a, b) \to \mathbb{R}$ satisfy the following conditions.*

(i) $g'(x) \neq 0$, $\forall x \in (a, b)$.

(ii)

$$\lim_{x \nearrow b} \frac{f'(x)}{g'(x)} = A \in [-\infty, \infty].$$

(iii) *Either*

$$\lim_{x\nearrow b} f(x) = \lim_{x\nearrow b} g(x) = 0, \tag{iii$_0$}$$

or

$$\lim_{x\nearrow b} g(x) = \pm\infty. \tag{iii$_\infty$}$$

Then

$$\lim_{x\nearrow b} \frac{f(x)}{g(x)} = A.$$

Proof. Let us first observe that (i) and Rolle's Theorem imply that g is injective. Hence, there exists $a' \in [a, b)$ such that $g(x) \neq 0$, $\forall x \in (a', b)$. Without any loss of generality we can assume that $a = a'$ since we are interested in the behavior of f, g near b. We have to prove that for any sequence $x_n \in (a, b)$ such that $\lim x_n = b$ we have

$$\lim_{n\to\infty} \frac{f(x_n)}{g(x_n)} = A.$$

Fix one such sequence $(x_n)_{n\in\mathbb{N}}$. At this point we want to invoke the following auxiliary fact whose proof we postpone.

Lemma 8.2.2. *There exists a sequence* (y_n) *in* (a, b) *such that* $x_n \neq y_n$, $\forall n$, $\lim_{n\to\infty} y_n = b$ *and*

$$\lim \left(\frac{|f(y_n)| + |g(y_n)|}{|g(x_n)|} \right) = 0. \qquad \square$$

Choose a sequence (y_n) as in the above lemma so that

$$\lim_{n\to\infty} \frac{f(y_n)}{g(x_n)} = \lim_{n\to\infty} \frac{g(y_n)}{g(x_n)} = 0.$$

From Cauchy's Finite Increment Theorem 7.4.16 we deduce that there exists $\xi_n \in (x_n, y_n)$ such that

$$r_n = \frac{f(x_n) - f(y_n)}{g(x_n) - g(y_n)} = \frac{f'(\xi_n)}{g'(\xi_n)}.$$

Since $x_n \to b$ we deduce $\xi_n \to b$ so that

$$\lim_{n\to\infty} r_n = \lim_{n\to\infty} \frac{f'(\xi_n)}{g'(\xi_n)} = A. \tag{8.2.1}$$

On the other hand, for any n we have

$$r_n = \frac{f(x_n) - f(y_n)}{g(x_n) - g(y_n)} = \frac{f(x_n) - f(y_n)}{g(x_n)\big(1 - \frac{g(y_n)}{g(x_n)}\big)} = \frac{\frac{f(x_n)}{g(x_n)} - \frac{f(y_n)}{g(x_n)}}{1 - \frac{g(y_n)}{g(x_n)}}.$$

We deduce

$$\frac{f(x_n)}{g(x_n)} - \frac{f(y_n)}{g(x_n)} = r_n \left(1 - \frac{g(y_n)}{g(x_n)} \right) \Rightarrow \frac{f(x_n)}{g(x_n)} = \frac{f(y_n)}{g(x_n)} + r_n \left(1 - \frac{g(y_n)}{g(x_n)} \right).$$

Hence

$$\lim_{n\to\infty}\frac{f(x_n)}{g(x_n)} = \underbrace{\lim_{n\to\infty}\frac{f(y_n)}{g(x_n)}}_{=0} + \bigl(\lim_{n\to\infty} r_n\bigr)\cdot\underbrace{\lim_{n\to\infty}\left(1-\frac{g(y_n)}{g(x_n)}\right)}_{=1}$$

$$= \lim_{n\to\infty} r_n \overset{(8.2.1)}{=} A.$$

All there is left to do is prove Lemma 8.2.2.

Proof of Lemma 8.2.2 We consider two cases.

1. Suppose that (iii_0) holds, i.e.,

$$\lim_{x\nearrow b} f(x) = \lim_{\nearrow b} g(x) = 0.$$

Then for any n we can find $y_n \in (x_n, b)$ such that

$$|f(y_n)| + |g(y_n)| < \frac{1}{n}|g(x_n)|.$$

so that

$$\frac{|f(y_n)| + |g(y_n)|}{|g(x_n)|} < \frac{1}{n},\quad \forall n,$$

and thus

$$\lim_{n\to\infty}\frac{|f(y_n)| + |g(y_n)|}{|g(x_n)|} = 0.$$

---

2. Suppose that (iii_∞) holds, i.e.,

$$\lim_{n\to\infty} g(x_n) = \pm\infty.$$

For $t \in (a, b)$ we set $h(t) := |f(t)| + |g(t)|$. We construct inductively an increasing sequence of natural numbers (n_k) as follows.

A. Since $|g(x_n)| \to \infty$ there exists $n_0 \in \mathbb{N}$ such that

$$|g(x_n)| > h(x_1),\quad \forall n \geqslant n_0.$$

B. Since $|g(x_n)| \to \infty$, for any $k \in \mathbb{N}$, $k > 1$, we can find $n_k \in \mathbb{N}$ such that $n_k > n_{k-1}$ and

$$|g(x_n)| > 2^k h(x_{n_{k-1}})\ \ \forall n \geqslant n_k. \tag{8.2.2}$$

Now define y_n by setting

$$y_n := \begin{cases} x_1, & 1 \leqslant n < n_1\\ x_{n_{k-1}}, & n_k \leqslant n < n_{k+1},\ k \in \mathbb{N}.\end{cases}$$

Observe that for $n \in [n_k, n_{k+1})$ we have

$$\frac{h(y_n)}{|g(x_n)|} = \frac{|h(x_{n_{k-1}})|}{g(x_n)} \overset{(8.2.2)}{<} \frac{1}{2^k}.$$

This proves that

$$\lim_{n\to\infty}\frac{h(y_n)}{|g(x_n)|} = 0.$$ □

□

Remark 8.2.3. Proposition 8.2.1 has a counterpart involving the left limit $\lim_{x\searrow a}$. Its statement is obtained from the statement of Proposition 8.2.1 by globally replacing the limit at b with the limit at a. The proof is entirely similar. □

Example 8.2.4. (a) We want to compute

$$\lim_{x\to 0}\frac{1-\cos x}{x^2}.$$

According to L'Hôpital's theorem we have

$$\lim_{x\to 0}\frac{1-\cos x}{x^2}=\lim_{x\to 0}\frac{(1-\cos x)'}{(x^2)'}=\lim_{x\to 0}\frac{\sin x}{2x}=\frac{1}{2}.$$

(b) Consider the function $f:(0,\infty)\to\mathbb{R}$, $f(x)=x^x$. We want to investigate the limit

$$\lim_{x\to 0}x^x.$$

Formally the limit ought to be 0^0, but we do not know what 0^0 means. Consider a new function

$$g(x)=\ln x^x=x\ln x,\;\;x>0.$$

In this case we have

$$\lim_{x\to 0+}g(x)=0\cdot(-\infty)$$

which is a degenerate limit. We rewrite

$$g(x)=\frac{\ln x}{\frac{1}{x}}$$

and we observe that in this case

$$\lim_{x\to 0+}g(x)=-\frac{\infty}{\infty}$$

which suggests trying L'Hôpital's rule. We have

$$(\ln x)'=\frac{1}{x},\;\;(1/x)'=-\frac{1}{x^2}$$

and

$$\frac{1/x}{-1/x^2}=-x\to 0\;\text{ as }x\to 0+.$$

Hence

$$\lim_{x\to 0+}g(x)=0\Rightarrow\lim_{x\to 0+}f(x)=e^0=1.$$ □

8.3. Convexity

In this section we discuss in some detail a concept that has found many useful applications.

8.3.1. Basic facts about convex functions.

We begin with a simple geometric observation.

Proposition 8.3.1. *Let $x, x_1, x_2 \in \mathbb{R}$, $x_1 < x_2$. The following statements are equivalent.*

(i) $x \in [x_1, x_2]$.

(ii) *There exist $t_1, t_2 \geqslant 0$ such that $t_1 + t_2 = 1$ and $x = t_1x_1 + t_2x_2$.*

Proof. (i) ⇒ (ii) Suppose $x \in [x_1, x_2]$. We set

$$t_1 := \frac{x_2 - x}{x_2 - x_1}, \quad t_2 := \frac{x - x_1}{x_2 - x_1}. \tag{8.3.1}$$

Since $x_1 \leqslant x \leqslant x_2$ we deduce that $t_1, t_2 \geqslant 0$. We observe that

$$t_1 + t_2 = \frac{x_2 - x}{x_2 - x_1} + \frac{x - x_1}{x_2 - x_1} = \frac{x_2 - x_1}{x_2 - x_1} = 1,$$

and

$$t_1x_1 + t_2x_2 = \frac{x_1(x_2 - x) + x_2(x - x_1)}{x_2 - x_1} = \frac{x_2x - x_1x}{x_2 - x_1} = x. \tag{8.3.2}$$

(ii) ⇒ (i) Suppose that there exist $t_1, t_2 \geqslant 0$ such that $t_1 + t_2 = 1$ and $x = t_1x_1 + t_2x_2$. We have

$$x - x_1 = (t_1 - 1)x_1 + t_2x_2 = -t_2x_1 + t_2x_2 = t_2(x_2 - x_1) \geqslant 0,$$
$$x_2 - x = (1 - t_2)x_2 - t_1x_1 = t_1x_2 - t_1x_1 = t_1(x_2 - x_1) \geqslant 0.$$

Hence $x_1 \leqslant x \leqslant x_2$. □

Remark 8.3.2. The point $t_1x_1 + t_2x_2$ is interpreted as the center of mass of a system of two particles, one located at x_1 and of mass t_1 and the other located at x_2 and of mass t_2.

In general, given n particles of masses $m_1, \dots, m_n$ respectively located at $x_1, \dots, x_n$, then the *center of mass* of this system is the point

$$\bar{x} = \frac{m_1x_1 + \cdots + m_nx_n}{m_1 + \cdots + m_n}.$$

Note that if we define

$$t_k := \frac{m_k}{m_1 + m_2 + \cdots + m_n}, \quad k = 1, 2, \dots, n,$$

then

$$t_1 + t_2 + \cdots + t_n = 1 \text{ and } \bar{x} = t_1x_1 + \cdots + t_nx_n.$$

Thus, a point x lies between x_1 and x_2 if and only if it is the center of mass of a system of particles located at x_1 and x_2. □

Given a function $f : (a, b) \to \mathbb{R}$ and $x_1, x_2 \in (a, b)$, $x_1 < x_2$, we denote by $L^f_{x_1,x_2}$ the linear function whose graph contains the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of f. The slope of this line is

$$m = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

and thus the equation of this line is

$$L^f_{x_1,x_2}(x) = f(x_1) + m(x - x_1) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1)$$

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

Hence

$$L^f_{x_1,x_2}(x) = \frac{x_2 - x}{x_2 - x_1}f(x_1) + \frac{x - x_1}{x_2 - x_1}f(x_2). \tag{8.3.3}$$

Above we recognize the numbers t_1, t_2 defined in (8.3.1).

Proposition 8.3.3. *Consider a function $f : (a, b) \to \mathbb{R}$ and $x_1, x_2 \in (a, b)$, $x_1 < x_2$. Denote by $L^f_{x_1,x_2}$ the linear function whose graph contains the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of f. The following statements are equivalent.*

$$f(x) \leqslant L^f_{x_1,x_2}(x), \quad \forall x \in [x_1, x_2]. \tag{8.3.4a}$$

$$f(x) \leqslant \frac{x_2 - x}{x_2 - x_1}f(x_1) + \frac{x - x_1}{x_2 - x_1}f(x_2), \quad \forall x \in [x_1, x_2]. \tag{8.3.4b}$$

$$\forall t_1, t_2 \geqslant 0 \;\; \textit{such that} \;\; t_1 + t_2 = 1 \;\; f(t_1x_1 + t_2x_2) \leqslant t_1f(x_1) + t_2f(x_2), \tag{8.3.4c}$$

Proof. The equivalence (8.3.4a) $\Longleftrightarrow$ (8.3.4b) follows from (8.3.3). The equivalence (8.3.4b) $\Longleftrightarrow$ (8.3.4c) follows from (8.3.1) and (8.3.2).

□

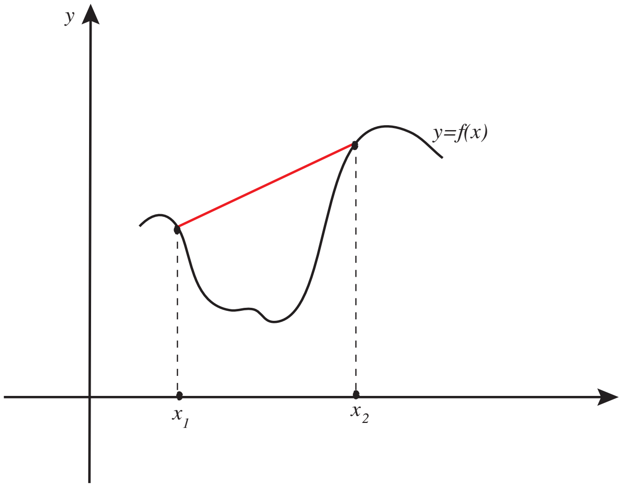

Figure 8.2. *The graph lies below the chord.*

Remark 8.3.4. The part of the graph of $L^f_{x_1,x_2}$ over the interval $[x_1, x_2]$ is called the *chord* of the graph of f determined by the interval $[x_1, x_2]$. The condition (8.3.4a) is equivalent to saying that the part of the graph of f corresponding to the interval $[x_1, x_2]$ lies below the chord of the graph determined by this interval; see Figure 8.2. □

Definition 8.3.5. Let $f : I \to \mathbb{R}$ be a real valued function defined on an interval I.

(i) The function f is called *convex* if, for any $x_1, x_2 \in I$, and any $t_1, t_2 \geqslant 0$ such that $t_1 + t_2 = 1$, we have

$$f(t_1x_1 + t_2x_2) \leqslant t_1f(x_1) + t_2f(x_2).$$

(ii) The function f is called *concave* if, for any $x_1, x_2 \in I$, and any $t_1, t_2 \geqslant 0$ such that $t_1 + t_2 = 1$, we have

$$f(t_1x_1 + t_2x_2) \geqslant t_1f(x_1) + t_2f(x_2).$$

□

Remark 8.3.6. (a) From Propositions 8.3.1 and 8.3.3 we deduce that a function $f : I \to \mathbb{R}$ is convex if and only if, for any interval $[x_1, x_2] \subset I$, the part of the graph of f determined by the interval $[x_1, x_2]$ is below the chord of the graph determined by this interval. It is concave if the graph is above the chords.

(b) Observe that if $t_1 = 1$ and $t_2 = 0$ we have $t_1x_1 + t_2x_2 = x_1$ and $t_1f(x_1) + t_2f(x_2) = f(x_1)$ and thus

$$f(t_1x_1 + t_2x_2) \leqslant t_1f(x_1) + t_2f(x_2).$$

is automatically satisfied. A similar thing happens when $t_1 = 0$ and $t_2 = 1$. Thus the definition of convexity is equivalent to the weaker requirement that for any $x_1, x_2 \in I$, and any *positive* t_1, t_2 such that $t_1 + t_2 = 1$, we have

$$f(t_1x_1 + t_2x_2) \leqslant t_1f(x_1) + t_2f(x_2).$$

(c) Observe that a function f is concave if and only if $-f$ is convex.

(d) In many calculus texts, convex functions are called *concave-up* and concave functions are called *concave-down.* □

Before we can give examples of convex functions we need to produce simple criteria for recognizing when a function is convex.

Proposition 8.3.1 implies that a function $f : I \to \mathbb{R}$ is convex if and only if for any $x_1, x_2 \in I$ and any $x \in (x_1, x_2)$ we have

$$f(x) \leqslant \frac{x_2 - x}{x_2 - x_1}f(x_1) + \frac{x - x_1}{x_2 - x_1}f(x_2)$$

Since

$$1 = \frac{x_2 - x}{x_2 - x_1} + \frac{x - x_1}{x_2 - x_1},$$

we deduce that f is convex if and only if

$$\begin{aligned} f(x)\left(\frac{x_2-x}{x_2-x_1}+\frac{x-x_1}{x_2-x_1}\right) &\leqslant \frac{x_2-x}{x_2-x_1}f(x_1)+\frac{x-x_1}{x_2-x_1}f(x_2) \\ \iff \frac{x_2-x}{x_2-x_1}\big(f(x)-f(x_1)\big) &\leqslant \frac{x-x_1}{x_2-x_1}\big(f(x_2)-f(x)\big) \\ \iff (x_2-x)\big(f(x)-f(x_1)\big) &\leqslant (x-x_1)\big(f(x_2)-f(x)\big) \\ \iff \frac{f(x)-f(x_1)}{x-x_1} &\leqslant \frac{f(x_2)-f(x)}{x_2-x}. \end{aligned}$$

We have thus proved the following result.

Corollary 8.3.7. *Let $f : I \to \mathbb{R}$ be a function defined on the interval $I \subset \mathbb{R}$. The following statements are equivalent.*

(i) *The function f is convex.*

(ii) *For any $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$ we have*

$$\frac{f(x)-f(x_1)}{x-x_1} \leqslant \frac{f(x_2)-f(x)}{x_2-x}.$$

□

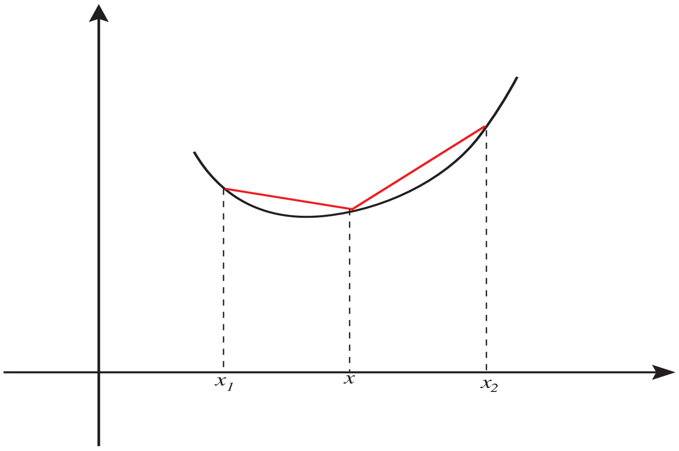

Figure 8.3. *Chords of the graph of a convex function become more inclined as they move to the right.*

Let us observe that $\frac{f(x)-f(x_1)}{x-x_1}$ is the slope of the chord determined by $[x_1, x]$ while $\frac{f(x_2)-f(x)}{x_2-x}$ is the slope of the chord determined by $[x, x_2]$. The above result states that f is convex if and only if for any $x_1 < x < x_2$ the chord determined by $[x_1, x]$ has a smaller inclination than the chord determined by $[x, x_2]$; see Figure 8.3.

Corollary 8.3.8. *Suppose that $f : I \to \mathbb{R}$ is a convex function. Then*

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_4) - f(x_3)}{x_4 - x_3}, \quad \forall x_1, x_2, x_3, x_4 \in I, \quad x_1 < x_2 < x_3 < x_4.$$

Proof. From Corollary 8.3.7 we deduce that the slope of the chord determined by $[x_1, x_2]$ is smaller than the slope of the chord determined by $[x_2, x_3]$ which in turn is smaller than the slope of the chord determined by $[x_3, x_4]$; see Figure 8.4. In other words,

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2} \leqslant \frac{f(x_4) - f(x_3)}{x_4 - x_3}.$$

□

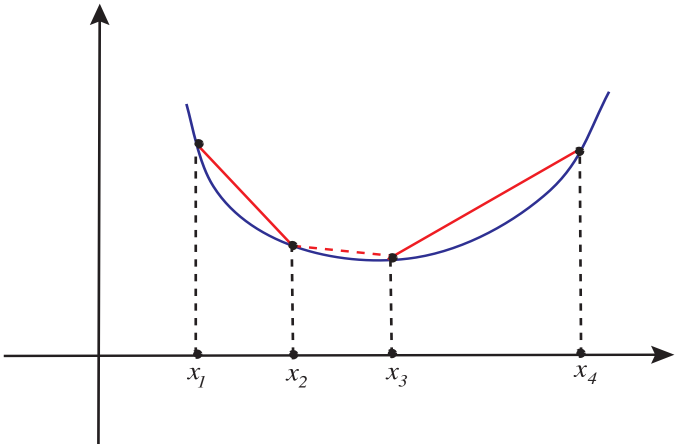

Figure 8.4. *Chords of the graph of a convex function become more inclined as they move to the right.*

Corollary 8.3.9. *Suppose that $f : I \to \mathbb{R}$ is a differentiable function. Then the following statements are equivalent.*

(i) *The function f is convex.*

(ii) *The derivative f' is a nondecreasing function.*

Proof. (ii) ⇒ (i) In view of Corollary 8.3.7 we have to prove that for any $x_1 < x_2 < x_3 \in I$ we have

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

From Lagrange's Mean Value theorem we deduce that there exist $\xi_1 \in (x_1, x_2)$ and $\xi_2 \in (x_2, x_3)$ such that

$$f'(\xi_1) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \quad f'(\xi_2) = \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

Since f' is nondecreasing and $\xi_1 < x_2 < \xi_2$, we deduce $f'(\xi_1) \leqslant f'(\xi_2)$.

(i) ⇒ (ii) We know that f is convex and we have to prove that f' is nondecreasing, i.e.,

$$x_1 < x_2 \Rightarrow f'(x_1) \leqslant f'(x_2).$$

For $h > 0$ sufficiently small, $h < \frac{1}{2}(x_2 - x_1)$, we have

$$x_1 < x_1 + h < x_2 - h < x_2.$$

From Corollary 8.3.8 we deduce that slope of the chord determined by $[x_1, x_1 + h]$ is smaller than the slope of the chord determined by $[x_2 - h, x_2]$, that is,

$$\frac{f(x_1 + h) - f(x_1)}{h} \leqslant \frac{f(x_2) - f(x_2 - h)}{h} = \frac{f(x_2 - h) - f(x_2)}{-h}.$$

Hence

$$f'(x_1) = \lim_{h \to 0+} \frac{f(x_1 + h) - f(x_1)}{h} \leqslant \lim_{h \to 0+} \frac{f(x_2 - h) - f(x_2)}{-h} = f'(x_2).$$

□

Since a differentiable function is nondecreasing iff its derivative is nonnegative, we deduce the following useful result.

Corollary 8.3.10. *Suppose that $f : I \to \mathbb{R}$ is a twice differentiable function. Then the following statements are equivalent.*

(i) *The function f is convex.*

(ii) *The second derivative f'' is nonnegative, $f''(x) \geqslant 0$, $\forall x \in I$.*

□

Since a function is concave if and only if $-f$ is convex we deduce the following result.

Corollary 8.3.11. *Suppose that $f : I \to \mathbb{R}$ is a twice differentiable function. Then the following statements are equivalent.*

(i) *The function f is concave.*

(ii) *The second derivative f'' is nonpositive, $f''(x) \leqslant 0$, $\forall x \in I$.*

□

Example 8.3.12. The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$ is convex since $f''(x) = e^x > 0$ for any $x \in \mathbb{R}$. The function $f : (0, \infty) \to \mathbb{R}$, $f(x) = \ln x$ is concave since

$$f'(x) = \frac{1}{x}, \quad f''(x) = -\frac{1}{x^2} < 0, \quad \forall x > 0.$$

Fix $\alpha \in \mathbb{R}$ and consider the power function

$$p : (0, \infty) \to \mathbb{R}, \quad p(x) = x^\alpha.$$

Then

$$p''(x) = \alpha(\alpha-1)x^{\alpha-2}.$$

Note that if $\alpha(\alpha-1) > 0$ this function is convex, if $\alpha(\alpha-1) < 0$ this function is concave, and if $\alpha = 0$ or $\alpha = 1$ this function is both convex and concave. Thus, the function $\sqrt{x}$ is concave, while the function $\frac{1}{\sqrt{x}} = x^{-\frac{1}{2}}$ is convex. □

8.3.2. Some classical applications of convexity. We start with a simple geometric consequence of convexity.

Proposition 8.3.13. *Suppose that $f : I \to \mathbb{R}$ is a differentiable convex function. Then the graph of f lies above any tangent to the graph; see Figure 8.5. If additionally f' is strictly increasing, then any tangent to the graph intersects the graph at a unique point.*

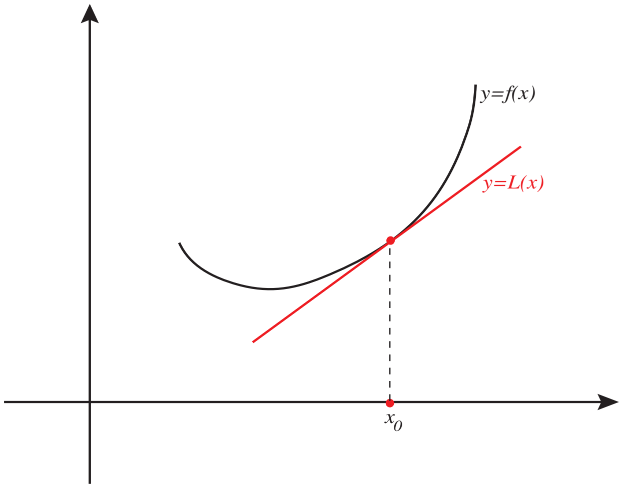

Figure 8.5. *The graph of a convex function lies above any of its tangents.*

Proof. Let $x_0 \in I$. The tangent to the graph of f at the point $(x_0, f(x_0))$ is the graph of the linearization of f at x_0 which is the function

$$L(x) = f(x_0) + f'(x_0)(x - x_0).$$

We have to prove that

$$f(x) - L(x) \geqslant 0, \quad \forall x \in I.$$

We have

$$f(x) - L(x) = f(x) - f(x_0) - f'(x_0)(x - x_0).$$

Suppose $x \neq x_0$. Lagrange's Mean Value Theorem implies that there exists ξ between x_0 and x such that $f(x) - f(x_0) = f'(\xi)(x - x_0)$. Hence

$$f(x) - L(x) = f'(\xi)(x - x_0) - f'(x_0)(x - x_0) = \big(f'(\xi) - f'(x_0)\big)(x - x_0).$$

We distinguish two cases.

1. $x > x_0$. Then $\xi > x_0$ and $(x - x_0) > 0$. Since f is convex, f' is increasing and thus $f'(\xi) \geqslant f'(x_0)$ so that

$$\big(\, f'(\xi) - f'(x_0)\,\big)(x - x_0) \geqslant 0.$$

Clearly if f' is strictly increasing, then $f'(\xi) > f'(x_0)$ and $\big(\, f'(\xi) - f'(x_0)\,\big)(x - x_0) > 0$.

2. $x < x_0$. Then $\xi < x_0$ and $(x - x_0) < 0$. Since f is convex, f' is increasing and thus $f'(\xi) \leqslant f'(x_0)$ so that

$$\big(\, f'(\xi) - f'(x_0)\,\big)(x - x_0) \geqslant 0.$$

Clearly if f' is strictly increasing, then $f'(\xi) < f'(x_0)$ and $\big(\, f'(\xi) - f'(x_0)\,\big)(x - x_0) > 0$ □

Example 8.3.14 (Newton's Method)**.** We want to describe an ingenious method devised by Isaac Newton[1] for approximating the solutions of an equation $f(x) = 0$.

Suppose that $f : (a, b) \to \mathbb{R}$ is a C^2-function such that

$$f'(x),\quad f''(x) > 0,\ \ \forall x \in (a, b). \tag{8.3.5}$$

Suppose $z_0 \in (a, b)$ satisfies

$$f(z_0) = 0.$$

The condition (8.3.5) implies that f is strictly increasing and thus z_0 is the unique solution of the equation $f(x) = 0$. Newton's method described one way of constructing very accurate approximations for z_0.

Here is roughly the principle behind the method. Pick an arbitrary point $x_0 \in (z_0, b)$. The linearization $L(x)$ of f at x_0 is an approximation for $f(x)$ so, intuitively, the solution of the equation $L(x) = 0$ ought to approximate the solution of the equation $f(x) = 0$. Denote by $Z(x_0)$ the solution of the equation $L(x) = 0$, i.e., the point where the tangent to the graph of f at $(x_0, f(x_0))$ intersects the horizontal axis; see Figure 8.6.

More precisely, we have $L(x) = f(x_0) + f'(x_0)(x - x_0)$ and thus,

$$L(x) = 0 \Longleftrightarrow f'(x_0)(x - x_0) = -f(x_0) \Longleftrightarrow x - x_0 = -\frac{f(x_0)}{f'(x_0)}$$

$$\Longleftrightarrow x = Z(x_0) = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Key Remark. *The point $Z(x_0)$ lies between z_0 and x_0, $z_0 < Z(x_0) < x_0$. In particular, $Z(x_0)$ is closer to z_0 than x_0.*

Clearly $Z(x_0) < x_0$ because $L(x_0) = f(x_0) > 0 = L(Z(x_0))$ and the linear function $L(x)$ is increasing. The assumption (8.3.5) implies that f is convex and f' is strictly increasing. Proposition 8.3.13 implies that the tangent lies below the graph, i.e.,

$$f\big(\, Z(x_0)\,\big) > L\big(\, Z(x_0)\,\big) = 0 = f(z_0).$$

[1] Isaac Newton (1642-1726) was an English mathematician and physicist who is widely recognized as one of the most influential scientists of all time and a key figure in the scientific revolution; see Wikipedia.

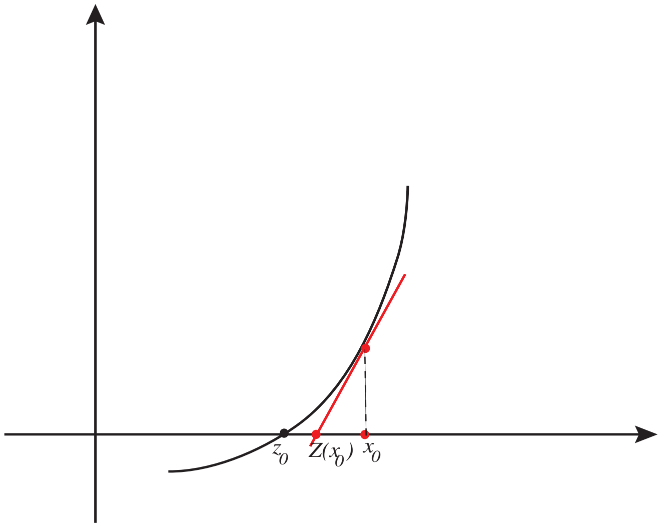

Figure 8.6. *The geometry behind Newton's method.*

Since f is strictly increasing we deduce $Z(x_0) > z_0$.

The correspondence $x_0 \mapsto Z(x_0)$ is thus a map $(z_0, b) \to (z_0, b)$ with the property that $z_0 < Z(x_0) < x_0$, $\forall x_0 \in (z_0, b)$.

We iterate this procedure. We set $x_1 = Z(x_0)$ so that $z_0 < x_1 < x_0$. Define next $x_2 = Z(x_1)$ so that $z_0 < x_2 < x_1$ and inductively

$$x_{n+1} := Z(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \geqslant 0. \tag{8.3.6}$$

The above discussion shows that the sequence (x_n) is strictly decreasing and bounded below by z_0. It is therefore convergent and we set $\bar{x} = \lim x_n$. Observe that $\bar{x} \geqslant z_0$. Letting $n \to \infty$ in (8.3.6) and taking into account the continuity of f and f' we deduce

$$\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})} \Rightarrow \frac{f(\bar{x})}{f'(\bar{x})} = 0 \Rightarrow f(\bar{x}) = 0.$$

Since z_0 is the unique solution of the equation $f(x) = 0$ we deduce $\bar{x} = z_0$. Thus the sequence generated by Newton's iteration (8.3.6) converges to the unique zero of f.

Remarkably, the above sequence (x_n) converges to z_0 extremely quickly. Taylor's formula with Lagrange remainder implies that for any n there exists $\xi_n \in (z_0, x_n)$ such that

$$0 = f(z_0) = f(x_n) + f'(x_n)(z_0 - x_n) + \frac{1}{2} f''(\xi_n)(z_0 - x_n)^2.$$

Hence

$$0 = \frac{f(x_n)}{f'(x_n)} + z_0 - x_n + \frac{f''(\xi_n)}{2f'(x_n)}(z_0 - x_n)^2 \Rightarrow \underbrace{\frac{f(x_n)}{f'(x_n)} + z_0 - x_n}_{= z_0 - x_{n+1}} = -\frac{f''(\xi_n)}{2f'(x_n)}(z_0 - x_n)^2$$

Hence

$$(z_0 - x_{n+1}) = -\frac{f''(\xi_n)}{2f'(x_n)}(z_0 - x_n)^2.$$

If we denote by ε_n the error, $\varepsilon_n := x_n - z_0$ we deduce

$$\varepsilon_{n+1} = \frac{f''(\xi_n)}{2f'(x_n)}\varepsilon_n^2. \tag{8.3.7}$$

Thus, the error at the $(n+1)$ -th step is roughly the square of the error at the n-th step. If e.g. the error ε_n is < 0.01, then we expect $\varepsilon_{n+1} < (0.1)^2 = 0.0001$.

Let us see how this works in a simple case. Let k be a natural number $\geqslant 2$. Consider the function

$$f : (0,\infty) \to \mathbb{R},\ \ f(x) = x^k - 2.$$

Then

$$f'(x) = kx^{k-1},\ \ f''(x) = k(k-1)x^{k-2}$$

so the assumption (8.3.5) is satisfied. The unique solution of the equation $f(x) = 0$ is the number $\sqrt[k]{2}$ and Newton's method will produce approximations for this number.

We first need to choose a number $x_0 > \sqrt[k]{2}$. How do we do this when we do not know what the number $\sqrt[k]{2}$ is?

Observe we have to choose a number x_0 such that $f(x_0) > f(\sqrt[k]{2}) = 0$, or equivalently,

$$x_0^k > 2.$$

Let's pick $x_0 = \frac{3}{2}$. Then

$$\left(\frac{3}{2}\right)^k \geqslant \left(\frac{3}{2}\right)^2 = \frac{9}{4} > 2.$$

Note also that $f(1) = 1^k - 2 = -1 < 0$ so that

$$1 < \sqrt[k]{2} < \frac{3}{2}$$

and thus the error

$$\varepsilon_0 = x_0 = \sqrt[k]{2} < \frac{1}{2}.$$

In this case we have

$$Z(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^k - 2}{kx^{k-1}} = \frac{k-1}{k}x + \frac{2}{kx^{k-1}}.$$

Observe that for $k = 2$ we have

$$Z(x) = \frac{x}{2} + \frac{1}{x}.$$

and the recurrence $x_{n+1} = Z(x_n)$ takes the form

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$$

Above, we recognize the recurrence that we have investigated earlier in Example 4.4.4.

For $k = 3$ the recurrence $x_{n+1} = Z(x_n)$ takes the form

$$x_{n+1} = \frac{2x_n}{3} + \frac{2}{3x_n^2}, \;\; x_0 = 1.5.$$

We have

$$x_1 = 1.296296..., \;\; x_2 = 1.260932..., \;\; x_3 = 1.25992186...,$$
$$x_4 = 1.25992104..., \;\; x_5 = 1.25992104...$$

Note that, as predicted theoretically, this sequence displays a very rapid stabilization. Thus

$$\sqrt[3]{2} \approx 1.25992.....$$

We can independently confirm the above claim by observing that

$$(1.25992)^3 = 1.999995. \qquad \square$$

Theorem 8.3.15 (Jensen's inequality). *Suppose that $f : I \to \mathbb{R}$ is a convex function defined on an interval I. Then for any $n \in \mathbb{N}$, any $x_1, \dots, x_n \in I$ and any $t_1, \dots, t_n \geqslant 0$ such that*

$$t_1 + \cdots + t_n = 1$$

we have $t_1x_1 + \cdots + t_nx_n \in I$ and

$$f\big(t_1x_1 + \cdots + t_nx_n\big) \leqslant t_1f(x_1) + \cdots + t_nf(x_n). \tag{8.3.8}$$

Proof. We argue by induction on n. For $n = 1$ the inequality is trivially true, while for $n = 2$ it is the definition of convexity. We assume that the inequality is true for n and we prove it for $n + 1$.

Let $x_0, \dots, x_n \in I$ and $t_0, \dots, t_n \geqslant 0$ such that

$$t_0 + \cdots + t_n = 1.$$

We have to prove that

$$f(t_0x_0 + t_1x_1 + t_2x_2 + \cdots + t_nx_n) \leqslant t_0f(x_0) + t_1f(x_1) + t_2f(y_2) + \cdots + t_nf(y_n). \tag{8.3.9}$$

If one of the numbers $t_0, t_1, \dots, t_n$ is zero, then the above inequality reduces to the case n. We can therefore assume that $t_0, t_1, \dots, t_n > 0$. Consider now the real numbers

$$s_1 := t_0 + t_1, \;\; s_2 := t_2, \dots, s_n := t_n,$$

$$y_1 := \frac{t_0}{t_0 + t_1}x_0 + \frac{t_1}{t_0 + t_1}x_1, \;\; y_2 := x_2, \dots, y_n := x_n.$$

Note that

$$s_1, s_2, \dots, s_n \geqslant 0 \;\text{ and }\; s_1 + \cdots + s_n = 1$$

and since

$$\frac{t_0}{t_0 + t_1} + \frac{t_1}{t_0 + t_1} = 1$$

the point y_1 lies between x_0 and x_1 and thus in the interval I. From the induction assumption we deduce

$$s_1y_1 + \cdots + s_ny_n \in I,$$

and

$$f(s_1y_1 + \cdots + s_ny_n) \leqslant s_1f(y_1) + s_2f(y_2) + \cdots + s_nf(y_n)$$

$$= (t_0+t_1)f\left(\frac{t_0}{t_0+t_1}x_0+\frac{t_1}{t_0+t_1}x_1\right)+s_2f(y_2)+\cdots+s_nf(y_n).$$

Now observe that

$$s_1y_1+\cdots+s_ny_n = (t_0+t_1)\left(\frac{t_0}{t_0+t_1}x_0+\frac{t_1}{t_0+t_1}x_1\right)+t_2y_2+\cdots+t_ny_n$$
$$= t_0x_0+t_1x_1+t_2x_2+\cdots+t_nx_n$$

and since f is convex

$$f\left(\frac{t_0}{t_0+t_1}x_0+\frac{t_1}{t_0+t_1}x_1\right) \leqslant \frac{t_0}{t_0+t_1}f(x_0)+\frac{t_1}{t_0+t_1}f(x_1)$$

so that

$$(t_0+t_1)f\left(\frac{t_0}{t_0+t_1}x_0+\frac{t_1}{t_0+t_1}x_1\right) \leqslant t_0f(x_0)+t_1f(x_1).$$

Putting together all of the above we deduce (8.3.9). □

Corollary 8.3.16. *If $f : I \to \mathbb{R}$ is a convex function defined on an interval I, then for any $n \in \mathbb{N}$ and any $x_1,\dots,x_n \in I$ we have*

$$\boxed{f\left(\frac{x_1+\cdots+x_n}{n}\right) \leqslant \frac{f(x_1)+\cdots+f(x_n)}{n}.} \tag{8.3.10}$$

Proof. Use (8.3.8) in which $t_1 = t_2 = \cdots = t_n = \frac{1}{n}$. □

Corollary 8.3.17. *Suppose that $g : I \to \mathbb{R}$ is a concave function defined on an interval I. Then for any $n \in \mathbb{N}$, any $x_1,\dots,x_n \in I$ and any $t_1,\dots,t_n \geqslant 0$ such that*

$$t_1+\cdots+t_n = 1$$

we have $t_1x_1+\cdots+t_nx_n \in I$ and

$$g\big(t_1x_1+\cdots+t_nx_n\big) \geqslant t_1g(x_1)+\cdots+t_ng(x_n). \tag{8.3.11}$$

In particular,

$$\boxed{g\left(\frac{x_1+\cdots+x_n}{n}\right) \geqslant \frac{g(x_1)+\cdots+g(x_n)}{n}.} \tag{8.3.12}$$

Proof. Apply Theorem 8.3.15 to the convex function $f = -g$. □

Corollary 8.3.18 (AM-GM inequality)**.** *For any natural number n and any positive real numbers $x_1,\dots,x_n$ we have*

$$\boxed{\big(x_1\cdots x_n\big)^{\frac{1}{n}} \leqslant \frac{x_1+\cdots+x_n}{n}}. \tag{8.3.13}$$

The left-hand side of the above inequality is called the geometric mean *(GM) of the numbers $x_1,\dots,x_n$, while the right-hand side is called the* arithmetic mean *(AM) of the same numbers.*

Proof. Consider the function

$$f : (0,\infty) \to \mathbb{R},\quad f(x) = \ln x.$$

This function is concave and (8.3.12) implies that

$$\ln\left(\frac{x_1+\cdots+x_n}{n}\right) \geqslant \frac{\ln x_1+\cdots+\ln x_n}{n}.$$

Exponentiating this inequality we deduce

$$\frac{x_1+\cdots+x_n}{n} = e^{\ln(\frac{x_1+\cdots+x_n}{n})}$$

$$\geqslant e^{\frac{\ln x_1+\cdots+\ln x_n}{n}} = e^{\frac{\ln(x_1\cdots x_n)}{n}} = (x_1\cdots x_n)^{\frac{1}{n}}.$$

□

Corollary 8.3.19 (Hölder's inequality). *Fix a real number $p>1$ and define $q>1$ by the equality*

$$\frac{1}{q} = 1-\frac{1}{p} = \frac{p-1}{p}.$$

Then for any natural number n and any nonnegative real numbers $a_1,\dots,a_n$, $b_1,\dots,b_n$ we have

$$a_1b_1+\cdots+a_nb_n \leqslant \big(a_1^p+\cdots+a_n^p\big)^{\frac{1}{p}}\big(b_1^q+\cdots+b_n^q\big)^{\frac{1}{q}}, \tag{8.3.14}$$

or, using the summation notation,

$$\boxed{\sum_{k=1}^n a_kb_k \leqslant \left(\sum_{i=1}^n a_i^p\right)^{\frac{1}{p}}\left(\sum_{j=1}^n b_j^q\right)^{\frac{1}{q}}.} \tag{8.3.15}$$

Proof. Since $p>1$, the function $f:[0,\infty)\to\mathbb{R}$, $f(x)=x^p$, is convex. We define

$$B := b_1^q+\cdots+b_n^q,$$

$$t_k := \frac{b_k^q}{B},\quad k=1,\dots,n,$$

$$x_k := a_kb_k^{-\frac{1}{p-1}}B,\quad k=1,\dots,n.$$

Observe that $t_k \geqslant 0$, $\forall k$ and

$$t_1+\cdots+t_k = 1.$$

Using Jensen's inequality (8.3.10) we deduce that

$$\big(t_1x_1+\cdots+t_nx_n\big)^p \leqslant t_1x_1^p+\cdots+t_nx_n^p.$$

Observe that

$$\big(t_1x_1+\cdots+t_nx_n\big)^p = \left(a_1b_1^{q-\frac{1}{p-1}}+\cdots+a_nb_n^{q-\frac{1}{p-1}}\right)^p$$

$(q - \frac{1}{p-1} = 1)$

$$= (\, a_1 b_1 + \cdots + a_n b_n \,)^p \,.$$

Similarly

$$t_1 x_1^p + \cdots + t_n x_n^p = \frac{b_1^q}{B} a_1^p b_1^{-\frac{p}{p-1}} B^p + \cdots + \frac{b_n^q}{B} a_n^p b_n^{-\frac{p}{p-1}} B^p$$

$(q - \frac{p}{p-1} = 0)$

$$= B^{p-1}\big(\, a_1^p + \cdots + a_n^p \,\big).$$

Hence

$$(\, a_1 b_1 + \cdots + a_n b_n \,)^p \leqslant B^{p-1} \,(\, a_1^p + \cdots + a_n^p \,)$$

so that

$$\begin{aligned} a_1 b_1 + \cdots + a_n b_n &\leqslant B^{\frac{p-1}{p}} \,(\, a_1^p + \cdots + a_n^p \,)^{\frac{1}{p}} \\ &= (\, a_1^p + \cdots + a_n^p \,)^{\frac{1}{p}} (\, b_1^q + \cdots + b_n^q \,)^{\frac{1}{q}}. \end{aligned}$$

□

If in Hölder's inequality we let $p = 2$, then $q = 2$, and we obtain the following important result.

Corollary 8.3.20 (Cauchy-Schwarz inequality)**.** *For any natural number n and any real numbers $x_1, \dots, x_n$, $y_1, \dots, y_n$ we have*

$$\boxed{\left| \sum_{k=1}^n x_k y_k \right| \leqslant \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}} \left(\sum_{j=1}^n y_j^2 \right)^{\frac{1}{2}} .} \tag{8.3.16}$$

Proof. We define

$$a_k = |x_k|, \quad b_k = |y_k|, \quad k = 1, \dots, n.$$

Note that $a_k^2 = x_k^2$, $b_k^2 = y_k^2$. Using Hölder's inequality with $p = q = 2$ we deduce

$$\sum_{k=1}^n |x_k y_k| \leqslant \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}} \left(\sum_{j=1}^n y_j^2 \right)^{\frac{1}{2}} .$$

Now observe that

$$\left| \sum_{k=1}^n x_k y_k \right| \leqslant \sum_{k=1}^n |x_k y_k|.$$

□

Corollary 8.3.21 (Minkowski's inequality)**.** *For any real number $p \in [1, \infty)$, any natural number n, and any real numbers $x_1, \dots, x_n$, $y_1, \dots, y_n$ we have*

$$\boxed{\left(\sum_{k=1}^n |x_k + y_k|^p \right)^{\frac{1}{p}} \leqslant \left(\sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}} + \left(\sum_{k=1}^n |y_k|^p \right)^{\frac{1}{p}} .} \tag{8.3.17}$$

Proof. We set

$$X := \left(\sum_{k=1}^n |x_k|^p\right)^{\frac{1}{p}},\ \ Y := \left(\sum_{k=1}^n |y_k|^p\right)^{\frac{1}{p}},\ \ Z := \left(\sum_{k=1}^n |x_k+y_k|^p\right)^{\frac{1}{p}}.$$

Clearly $X, Y, Z \geqslant 0$. We have to prove that $Z \leqslant X + Y$. This inequality is obviously true if $Z = 0$ so we assume that $Z > 0$. Note that we have

$$\begin{aligned} Z^p &= \sum_{k=1}^n |x_k+y_k|^p = \sum_{k=1}^n \underbrace{|x_k+y_k|}_{\leqslant |x_k|+|y_k|}\,|x_k+y_k|^{p-1} \\ &\leqslant \sum_{k=1}^n |x_k|\,|x_k+y_k|^{p-1} + \sum_{k=1}^n |y_k|\,|x_k+y_k|^{p-1}. \end{aligned} \tag{8.3.18}$$

This proves (8.3.17) in the special case $p = 1$ so in the sequel we assume that $p > 1$. Let $q = \frac{p}{p-1}$ so that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Using Hölder's inequality we deduce that for any $k = 1, \dots, n$ we deduce

$$\sum_{k=1}^n |x_k|\,|x_k+y_k|^{p-1} \leqslant \underbrace{\left(\sum_{k=1}^p |x_k|^p\right)^{\frac{1}{p}}}_{X}\ \underbrace{\left(\sum_{k=1}^n |x_k+y_k|^p\right)^{\frac{p-1}{p}}}_{Z^{p-1}},$$

$$\sum_{k=1}^n |y_k|\,|x_k+y_k|^{p-1} \leqslant \underbrace{\left(\sum_{k=1}^p |y_k|^p\right)^{\frac{1}{p}}}_{Y}\ \underbrace{\left(\sum_{k=1}^n |x_k+y_k|^p\right)^{\frac{p-1}{p}}}_{Z^{p-1}}.$$

Using the last two inequalities in (8.3.18) we deduce

$$Z^p \leqslant (X+Y)Z^{p-1} \overset{Z>0}{\Rightarrow} Z \leqslant X + Y.$$

□

Remark 8.3.22. Minkowski's inequality has a very useful interpretation. For a natural number n we denote by $\mathbb{R}^n$ the n-dimensional Euclidean space whose points are called (n-dimensional) *vectors* and are defined to be n-tuples

$$\boldsymbol{x} = (x_1, \dots, x_n),\ \ x_i \in \mathbb{R},\ \ 1 \leqslant i \leqslant n.$$

The space $\mathbb{R}^n$ has a rich algebraic structure. We mention here two operations. One is the addition of vectors. Given

$$\boldsymbol{x} = (x_1, \dots, x_n),\ \ \boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$$

we define their sum $\boldsymbol{x} + \boldsymbol{y}$ to be the vector

$$\boldsymbol{x} + \boldsymbol{y} := (x_1 + y_1, \dots, x_n + y_n).$$

Another is the multiplication by a scalar. Given

$$\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n, \quad t \in \mathbb{R},$$

we define

$$t\boldsymbol{x} := (tx_1, \dots, tx_n).$$

For $p \in [1, \infty)$ and $\boldsymbol{x} \in \mathbb{R}^n$ we set

$$\|\boldsymbol{x}\|_p := \left(\sum_{k=1}^n |x_k|^p\right)^{\frac{1}{p}}.$$

Note that

$$\begin{gathered} \|t\boldsymbol{x}\|_p = |t|\,\|\boldsymbol{x}\|_p, \quad \forall t \in \mathbb{R}, \quad \boldsymbol{x} \in \mathbb{R}^n, \\ \|\boldsymbol{x}\|_p \geqslant 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \\ \|\boldsymbol{x}\|_p = 0 \Longleftrightarrow \boldsymbol{x} = (0, 0, \dots, 0). \end{gathered} \tag{8.3.19}$$

Minkowski's inequality is then equivalent to the *triangle inequality*

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \leqslant \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{8.3.20}$$

A function $\mathbb{R}^n \to \mathbb{R}$ that associates to a vector $\mathbb{R}$ a real number $\|\boldsymbol{x}\|$ satisfying (8.3.19) and (8.3.20) is called a *norm* on $\mathbb{R}^n$. Minkowski's inequality can be interpreted as saying that for any $p \in [1, \infty)$ the correspondence

$$\mathbb{R}^n \ni \boldsymbol{x} \mapsto \|\boldsymbol{x}\|_{\boldsymbol{p}} \in [0, \infty),$$

defines a norm on $\mathbb{R}^n$.

Note that (8.3.20) implies that for any $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ we have

$$\|\boldsymbol{u} - \boldsymbol{w}\|_p \leqslant \|\boldsymbol{u} - \boldsymbol{v}\|_p + \|\boldsymbol{v} - \boldsymbol{w}\|_p, \tag{8.3.21}$$

since

$$\underbrace{(\boldsymbol{u} - \boldsymbol{v})}_{\boldsymbol{x}} + \underbrace{(\boldsymbol{v} - \boldsymbol{w})}_{\boldsymbol{y}} = \underbrace{(\boldsymbol{u} - \boldsymbol{w})}_{\boldsymbol{x}+\boldsymbol{y}}. \qquad \square$$

8.4. How to sketch the graph of a function

Differential calculus can be quite useful in producing sketches of the graphs of functions. Instead of giving a detailed description of the steps that need to be taken to produce a sketch of a graph, we will outline a few general principles and illustrate them on a few examples.

In sketching the graph of a function $f(x)$, one needs to look at certain distinguishing features.

- Locate the intersections of f with the coordinate axes, if possible.
- Locate, if possible, the critical points of f, i.e., the points x such that $f'(x) = 0$.
- Locate the intervals where f is increasing and the intervals where f is decreasing, if possible.

- Locate the intervals where f is convex, and the intervals where f is concave, if possible. The endpoints of such intervals are found among the solutions of the equation.
$$f''(x) = 0.$$
Sometimes solving this equation explicitly may not be possible.
- Locate the asymptotes, if any.

Example 8.4.1 (Cubic polynomials)**.** Consider an arbitrary cubic polynomial
$$p : \mathbb{R} \to \mathbb{R},\;\; p(x) = x^3 + a_2x^2 + a_1x + a_0,$$
where a_0, a_1, a_2 are given real numbers. We would like to describe the general appearance of the graph of p and analyze how it depends on the coefficients a_0, a_1, a_2. Observe first that
$$\lim_{x\to\pm\infty} p(x) = \pm\infty.$$
The graph intersects the y-axis at $y = a_0$. The intersection with the x-axis is difficult to find because the equation $p(x) = 0$ is difficult to solve. Instead, we will try to find the critical points of $p(x)$ i.e., the solutions of the equation $p'(x) = 0$.
$$3x^2 + 2a_2x + a_1 = 0. \tag{8.4.1}$$
The function $p'(x)$ has a global minimum achieved at the point μ defined by the equation
$$p''(\mu) = 0 \iff 6\mu + 2a_2 = 0 \iff \mu = -\frac{a_2}{3}.$$
The function $p'(x)$ is decreasing on the interval $(-\infty, \mu]$ and increasing on $[\mu, \infty)$. Thus $p(x)$ is concave on $(-\infty, \mu]$ and convex on $[\mu, \infty)$. The point μ is an inflection point of p.

The general theory of quadratic equations tells us that (8.4.1) can have zero, one or two solutions depending on whether $\Delta = 4a_2^2 - 12a_1$ is negative, zero or positive. These situations are depicted in Figure 8.7.

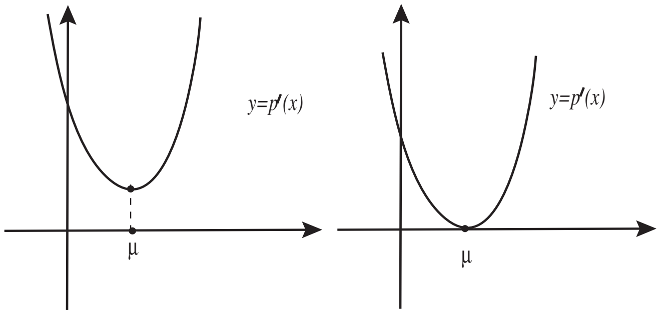

Figure 8.7. $\Delta = 4a_2^2 - 12a_1 \leqslant 0$.

If p has no critical points, as in the left-hand side of Figure 8.7, then $p'(x) > 0$ for any $x \in \mathbb{R}$. This shows that p is increasing. Similarly, if p has a single critical point, then again

$p(x)$ is increasing. In both cases, the graph of p looks like the left-hand side of Figure 8.9.

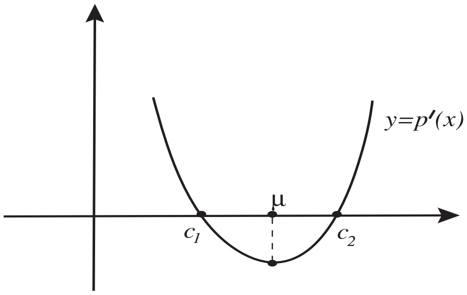

Figure 8.8. $\Delta = 4a_2^2 - 12a_1 > 0$.

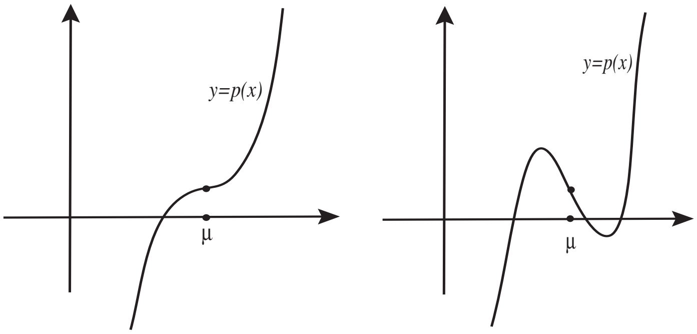

Figure 8.9. *The graph of* $y = x^3 + a_2x^2 + a_1x + a_0$.

If $p(x)$ has two critical points $c_1 < c_2$, then $p'(x) < 0$ on (c_1, c_2) and positive on $(-\infty, c_1) \cup (c_2, \infty)$; see Figure 8.8. The point c_1 is a local max of p and c_2 is a local min of p. The inflection point μ is the midpoint of the interval $[c_1, c_2]$. The graph of p is depicted on the right-hand side of Figure 8.9. □

Example 8.4.2. Consider the function

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

We have not specified its domain so it is understood to consist of all the x for which the fraction

$$\frac{x^2+1}{x^2-3x+2}$$

is well defined. The only problems are the points where the denominator vanishes,

$$x^2 - 3x + 2 = 0 \Longleftrightarrow x = 1 \quad \vee \quad x = 2.$$

Thus the domain is

$$(-\infty, 1) \cup (1, 2) \cup (2, \infty).$$

The points 1 and 2 are also points where the vertical asymptotes could be located. We will investigate this issue later.

We have

$$f'(x) = \frac{(2x)(x^2-3x+2)-(x^2+1)(2x-3)}{(x^2-3x+2)^2} = \frac{2x^3-6x^2+4x-(2x^3-3x^2+2x-3)}{(x^2-3x+2)^2}$$
$$= \frac{-3x^2+2x+3}{(x^2-3x+2)^2}.$$

The derivative vanishes when $3x^2-2x-3=0$. The roots of this quadratic polynomial are

$$\frac{2\pm\sqrt{4+36}}{6} = \frac{2\pm\sqrt{40}}{6} = \frac{2\pm 2\sqrt{10}}{6} = \frac{1\pm\sqrt{10}}{3}.$$

One root is obviously negative. Since $3<\sqrt{10}<4$ we deduce

$$1 < \frac{1+\sqrt{10}}{3} < \frac{5}{3} < 2.$$

The intersection with the y-axis is obtained by computing $f(0)=\frac{1}{2}$. There is no intersection with the x axis since the numerator does not vanish. We have already detected several remarkable points

$$-\infty,\ c_1 = \frac{1-\sqrt{10}}{3},\ 1,\ c_2 = \frac{1+\sqrt{10}}{3},\ 2,\ \infty.$$

Observe that

$$\lim_{x\to\pm\infty} \frac{x^2+1}{x^2-3x+2},$$

so the horizontal line $y=1$ is a horizontal asymptote for $f(x)$ at $\pm\infty$. We do not investigate the second derivative because it requires a substantial amount of work, with little payoff.

Table 8.1 organizes the information we have collected. The exclamation signs indicate that the corresponding functions are not defined at those points. As x approaches 1 from the left, the function $f(x)$ is increasing and

$$\lim_{x\to 1-} f(x) = \infty.$$

x	$-\infty$		c_1		1		c_2		2		∞
(x^2-3x+2)	∞	$++$	$+$	$++$	0	$---$	$-$	$--$	0	$++$	∞
$-3x^2+2x+3$	$-\infty$	$--$	0	$++$	$+$	$++$	0	$--$	$-$	$--$	$-\infty$
$f'(x)$	$-$	$--$	0	$++$	!	$++$	0	$--$	!	$--$	$-$
$f(x)$	1	↘	min	↗	!	↗	max	↘	!	↘	1

Table 8.1. *Organizing all the relevant data.*

Similarly, the table shows

$$\lim_{x\to 1+} f(x) = -\infty, \quad \lim_{x\to 2-} f(x) = -\infty, \quad \lim_{x\to 2+} f(x) = \infty.$$

This shows that the vertical lines $x = 1$ and $x = 2$ are asymptotes of $f(x)$. Figure 8.10 contains a sketch of the graph of the function $f(x)$.

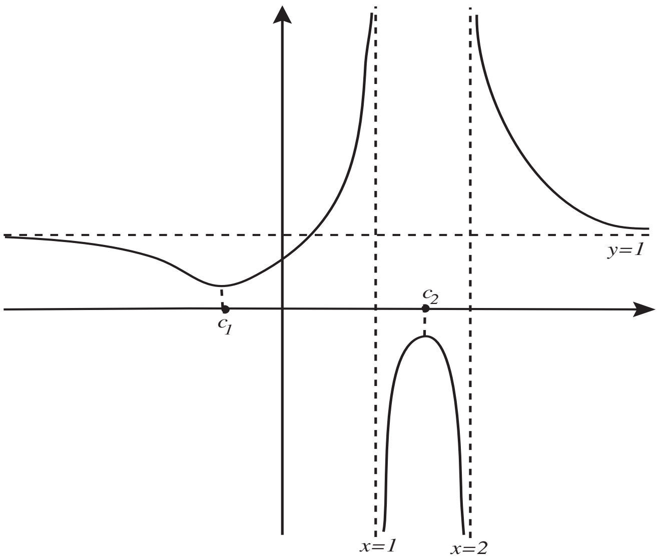

Figure 8.10. *The graph of* $\frac{x^2+1}{x^2-3x+2}$.

□

Some functions admit *inclined asymptotes.*

Definition 8.4.3. (a) The line $y = mx + b$ is the asymptote of $f(x)$ at ∞ if

$$\lim_{x\to\infty} \frac{f(x)}{x} = m \text{ and } \lim_{x\to\infty}(f(x) - mx) = b.$$

(b) The line $y = mx + b$ is the asymptote of $f(x)$ at $-\infty$ if

$$\lim_{x\to-\infty} \frac{f(x)}{x} = m \text{ and } \lim_{x\to-\infty}(f(x) - mx) = b.$$ □

Example 8.4.4. The function

$$f(x) = \frac{x^5 + 2x^4 + 3x^3 + 4x + 5}{x^4 + 1}$$

admits an inclined asymptote $y = mx + b$ as $x \to \infty$. The slope m can be found from the equality

$$m = \lim_{x\to\infty} \frac{f(x)}{x} = 1,$$

and b can be found from the equality

$$b = \lim_{x\to\infty}\big(\, f(x) - x\,\big) = \lim_{x\to\infty} \frac{x^5 + 2x^4 + 3x^3 + 4x + 5 - x(x^4+1)}{x^4+1}$$
$$= \lim_{x\to\infty} \frac{2x^4 + 3x^3 + 3x + 5}{x^4+1} = 2. \qquad \square$$

8.5. Antiderivatives

Definition 8.5.1. Suppose that $f : I \to \mathbb{R}$ is a function defined on an interval $I \subset \mathbb{R}$. A function $F : I \to \mathbb{R}$ is called an *antiderivative* or *primitive* of f on I if F is differentiable, and

$$F'(x) = f(x), \;\; \forall x \in I. \qquad \square$$

Example 8.5.2. (a) The function x^2 is an antiderivative of $2x$ on $\mathbb{R}$. Similarly, the function $\sin x$ is an antiderivative of $\cos x$ on $\mathbb{R}$. □

Observe that if $F(x)$ is an antiderivative of a function $f(x)$ on an interval I, then for any constant $C \in \mathbb{R}$ the function $F(x) + C$ is also an antiderivative of $f(x)$ on I. The converse is also true.

Proposition 8.5.3. *If F_1, F_2 are antiderivatives of the function $f : I \to \mathbb{R}$, then $F_1 - F_2$ is constant.*

Proof. Observe that $(F_1 - F_2)' = F_1' - F_2' = f - f = 0$ and Corollary 7.4.8 implies that $F_1 - F_2$ is constant on I. □

Definition 8.5.4. Given a function $f : I \to \mathbb{R}$ we denote by $\int f(x)dx$ the *collection* of all the antiderivatives of f on I. Usually $\int f(x)dx$ is referred to as the *indefinite integral* of f. □

For example,

$$\int \cos x\, dx = \sin x + C, \;\; \int 2x\, dx = x^2 + C.$$

Table 8.2 describes the antiderivatives of some basic functions.

Note that if $f :\to \mathbb{R}$ is a differentiable function, then f is an antiderivative of f' so that

$$\int f'(x)dx = f(x) + c. \tag{8.5.1}$$

Observing that $f'(x)dx = df$ we rewrite the above equality as

$$\int df = f + C. \tag{8.5.2}$$

In general, the computation of an antiderivative is a more challenging task that cannot always be completed. There are a few tricks and a few classes of functions for which this

$f(x)$	$\int f(x)dx$
x^n, $(x \in \mathbb{R},\ n \in \mathbb{Z},\ n \geqslant 0)$	$\frac{x^{n+1}}{(n+1)} + C$
$\frac{1}{x^n}$ $(x \neq 0,\ n \in \mathbb{N},\ n > 1)$	$-\frac{1}{(n-1)x^{n-1}} + C$
x^α, $(\alpha \in \mathbb{R},\ \alpha \neq -1,\ x > 0)$	$\frac{x^{\alpha+1}}{\alpha+1} + C$
$1/x,\ x \neq 0$	$\ln\lvert x\rvert + C$
e^x, $(x \in \mathbb{R})$	$e^x + C$
$\sin x$, $(x \in \mathbb{R})$	$-\cos x + C$
$\cos x$, $(x \in \mathbb{R})$	$\sin x + C$
$1/\cos^2 x$	$\tan x + C$
$\frac{1}{\sqrt{1-x^2}},\ x \in (-1,1)$	$\arcsin x + C$
$\frac{1}{1+x^2},\ x \in \mathbb{R}$	$\arctan x + C$
$\int \frac{1}{\sqrt{x^2 \pm 1}} dx,\ x^2 \pm 1 > 0$	$\ln\big\lvert x + \sqrt{x^2 \pm 1}\,\big\rvert + C$

Table 8.2. Table of integrals.

task is feasible. We will spend the remainder of this section discussing a few frequently encountered techniques for computing antiderivatives.

Proposition 8.5.5 (Linearity)**.** *Suppose $f, g : I \to \mathbb{R}$ and $a, b \in \mathbb{R}$. If $F, G : I \to \mathbb{R}$ are antiderivatives of f and respectively g on I, then $aF + bG$ is an antiderivative of $af + bg$ on I. We write this in condensed form*

$$\int (af + bg)dx = a\int f dx + b\int g dx.$$

Proof.

$$(aF + bG)' = aF' + bG' = af + bg.$$

□

Example 8.5.6.

$$\int (3+5x+7x^2)dx = 3\int dx + 5\int xdx + 7\int x^2dx = 3x + \frac{5}{2}x^2 + \frac{7}{3}x^3 + C. \qquad \square$$

Proposition 8.5.7 (Integration by parts)**.** *Suppose that $f, g : I \to \mathbb{R}$ are two differentiable functions. If the function $f(x)g'(x)$ admits antiderivatives on I, then so does the function $f'(x)g(x)$ and moreover*

$$\boxed{\int f(x)g'(x)dx = f(x)g(x) - \int g(x)f'(x)dx}. \tag{8.5.3}$$

Proof. The function $(fg)' = f'g + fg'$ admits antiderivatives and thus the difference

$$(fg)' - fg' = f'g$$

admits antiderivatives. Moreover,

$$fg = \int (fg)'dx = \int (f'g + fg')dx = \int f'gdx + \int fg'dx \Rightarrow \int fg'dx = fg - \int gf'dx.$$

$\square$

Let us observe that we can rewrite (8.5.3) in the simpler form

$$\boxed{\int fdg = fg - \int gdf}. \tag{8.5.4}$$

Example 8.5.8. (a) We can use integration by parts to find the antiderivatives of $\ln x$, $x > 0$. We have

$$\int \ln xdx = (\ln x)x - \int xd(\ln x) = x\ln x - \int x\frac{dx}{x}$$
$$= x\ln x - \int dx = x\ln x - x + C.$$

(b) For $a \in \mathbb{R}$ consider the indefinite integrals

$$\boxed{I_a = \int e^{ax}\cos x\,dx, \quad J_a = \int e^{ax}\sin x\,dx}.$$

We have

$$I_a = \int e^{ax}d(\sin x) = e^{ax}\sin x - \int \sin xd(e^{ax}) = e^{ax}\sin x - \int ae^{ax}\sin xdx$$
$$= e^{ax}\sin x - aJ_a.$$

Similarly we have

$$J_a = \int e^{ax}d(-\cos x) = -e^{ax}\cos x + \int \cos xd(e^{ax}) = -e^{ax}\cos x + a\int e^{ax}\cos xdx$$
$$= -e^{ax}\cos x + aI_a.$$

We deduce

$$I_a = e^{ax}\sin x - a(-e^{ax}\cos x + aI_a) = e^{ax}\sin x + ae^{ax}\cos x - a^2 I_a,$$

so that

$$(a^2+1)I_a = e^{ax}\sin x + ae^{ax}\cos x,$$

which shows that

$$\boxed{I_a = \frac{1}{a^2+1}\big(\,e^{ax}\sin x + ae^{ax}\cos x\,\big) + C}. \tag{8.5.5}$$

From this we deduce

$$J_a = aI_a - e^{ax}\cos x = \frac{a}{a^2+1}\big(\,e^{ax}\sin x + ae^{ax}\cos x\,\big) - e^{ax}\cos x + C,$$

so that

$$\boxed{J_a = \frac{1}{a^2+1}\big(\,ae^{ax}\sin x - e^{ax}\cos x\,\big) + C}. \tag{8.5.6}$$

(c) For any nonnegative integer n we consider the indefinite integral

$$\boxed{I_n = \int x^n e^x dx}.$$

Note that

$$I_0 = \int e^x dx = e^x + c.$$

In general, we have

$$I_{n+1} = \int x^{n+1} d(e^x) = x^{n+1}e^x - \int e^x d(x^{n+1}) = x^{n+1}e^x - (n+1)\int x^n e^x dx$$

so that

$$\boxed{I_{n+1} = x^{n+1}e^x - (n+1)I_n,\ \ \forall n = 0,1,2,\dots}. \tag{8.5.7}$$

If we let $n=0$ in the above equality we deduce

$$I_1 = xe^x - I_0 = xe^x - e^x + C, \tag{8.5.8}$$

Using $n=1$ in (8.5.7) we obtain

$$I_2 = x^2e^x - 2I_1 = x^2e^x - 2xe^x + 2e^x + C.$$

This suggests that in general $I_n = P_n(x)e^x + C$, where $P_n(x)$ is a polynomial of degree n. For example,

$$P_0(x) = 1,\ \ P_1(x) = (x-1),\ \ P_2(x) = x^2 - 2x + 2.$$

The equality (8.5.7) shows that

$$P_{n+1}(x) = x^{n+1} - (n+1)P_n(x),\ \ \forall n = 0,1,2,\dots\ . \tag{8.5.9}$$

(d) Let us now explain how to compute the integrals

$$\boxed{A_n = \int \frac{dx}{(x^2+1)^n}}.$$

Note that

$$A_1 = \int \frac{dx}{x^2+1} = \arctan x + C.$$

In general

$$A_n = \int (x^2+1)^{-n} dx = x(x^2+1)^{-n} - \int x d\Big((x^2+1)^{-n} \Big)$$

$$= \frac{x}{(x^2+1)^n} - \int x \frac{-2nx}{(x^2+1)^{n+1}} dx = \frac{x}{(x^2+1)^n} + 2n \int \frac{x^2}{(x^2+1)^{n+1}} dx$$

$$= \frac{x}{(x^2+1)^n} + 2n \int \frac{x^2+1-1}{(x^2+1)^{n+1}} dx$$

$$= \frac{x}{(x^2+1)^n} + 2n \int \frac{1}{(x^2+1)^n} dx - 2n \int \frac{1}{(x^2+1)^{n+1}} dx$$

$$= \frac{x}{(x^2+1)^n} + 2nA_n - 2nA_{n+1}.$$

Hence

$$A_n = \frac{x}{(x^2+1)^n} + 2nA_n - 2nA_{n+1},$$

so that

$$2nA_{n+1} = \frac{x}{(x^2+1)^n} + (2n-1)A_n,$$

and thus

$$\boxed{A_{n+1} = \frac{1}{2n} \frac{x}{(x^2+1)^n} + \frac{(2n-1)}{2n} A_n}. \tag{8.5.10}$$

For example,

$$\int \frac{1}{(x^2+1)^2} dx = \frac{1}{2} \frac{x}{x^2+1} + \frac{1}{2} \arctan x + C. \qquad \square$$

Proposition 8.5.9 (Integration by substitution)**.** *Suppose that $u : I \to J$ and $f : J \to \mathbb{R}$ are differentiable functions. Then the function $f'(u(x))u'(x)$ admits antiderivatives on I and*

$$\int f'(u(x))u'(x)dx = \int f'(u)du = \int df = f(u) + C, \quad u = u(x). \tag{8.5.11}$$

Proof. The chain formula shows that $f'(u(x))u'(x)$ is the derivative of $f(u(x))$ so that $f(u(x))$ is an antiderivative of $f'(u(x))u'(x)$. □

Example 8.5.10. (a) To find an antiderivative of xe^{x^2} we use the change in variables $u = x^2$. Then

$$du = 2xdx \Rightarrow xdx = \frac{du}{2}$$

so that

$$\int e^{x^2} xdx = \int e^u \frac{du}{2} = \frac{1}{2} e^u + C = \frac{1}{2} e^{x^2} + C.$$

(b) Let us compute an antiderivative of $\tan x = \frac{\sin x}{\cos x}$ on an interval I where $\cos x \neq 0$. We distinguish two cases.

1. $\cos x > 0$ on I. We make the change in variables $u = \cos x$ so that $u > 0$, and $du = -\sin x dx$. We have

$$\int \frac{\sin x}{\cos x} dx = -\int \frac{du}{u} = -\ln u + C = -\ln \cos x + C = -\ln|\cos x| + C.$$

2. $\cos x < 0$ on I. We make the change in variables $v = -\cos x$ so that $v > 0$ and $dv = \sin x dx$. We have

$$\int \frac{\sin x}{\cos x} dx = \int \frac{dv}{-v} = -\ln v + C = -\ln(-\cos x) + C = -\ln|\cos x| + C.$$

Thus, in either case we have

$$\int \tan x dx = -\ln|\cos x| + C. \tag{8.5.12}$$

(c) To compute the integral

$$\int (ax+b)^n dx, \quad n \in \mathbb{N}, \quad a > 0,$$

we make the change in variables $u = ax + b$. Then $du = adx$ so that $dx = \frac{1}{a}du$ and we have

$$\int (ax+b)^n dx = \frac{1}{a}\int u^n du = \frac{1}{a(n+1)}u^{n+1} + C = \frac{1}{a(n+1)}(ax+b)^{n+1} + C.$$

(d) To compute the integral

$$\int \frac{1}{(ax+b)^n} dx, \quad a \neq 0, n \in \mathbb{N}$$

we again make the change in variables $u = (ax + b)$ and we deduce

$$\int \frac{1}{(ax+b)^n} dx = \frac{1}{a}\int \frac{1}{u^n} du = C + \begin{cases} \frac{1}{a}\ln|u|, & n = 1 \\ \\ \frac{1}{a(1-n)u^{n-1}}, & n > 1. \end{cases}, \quad u = ax + b.$$

(e) To compute the integral

$$\boxed{B_n := \int \frac{x}{(x^2+1)^n} dx}.$$

We make the change in variables $u = x^2 + 1$. Then $du = 2xdx$ so that $xdx = \frac{1}{2}du$ and thus

$$\int \frac{x}{(x^2+1)^n} dx = \frac{1}{2}\int \frac{1}{u^n} du = C + \frac{1}{2} \times \begin{cases} \ln u, & n = 1 \\ \\ \frac{1}{(1-n)u^{n-1}}, & n > 1. \end{cases}, \quad u = x^2 + 1.$$

(f) The integrals of the form

$$\boxed{\int (\sin x)^m (\cos x)^{2k+1} dx, \quad k, m \in \mathbb{Z}_{\geqslant 0}},$$

are found using the change in variables $u = \sin x$. Then

$$du = \cos x dx, \quad (\cos x)^{2k+1}dx = (\cos^2 x)^k \cos x dx = (1-\sin^2 x)^k d(\sin x) = (1-u^2)^k du$$

and

$$\int (\sin x)^m (\cos x)^{2k+1} dx = \int u^m (1-u^2)^k du.$$

Similarly, the integrals of the form

$$\int (\cos x)^m (\sin x)^{2k+1} dx, \quad m, k \in \mathbb{Z}_{\geqslant 0},$$

are found using the change in variables $v = \cos x$. Then

$$\int (\cos x)^m (\sin x)^{2k+1} dx = -\int v^m (1-v^2)^k dv.$$

(g) The integrals of the form

$$\boxed{\int (\sin x)^{2m} (\cos x)^{2k} dx, \quad k, m \in \mathbb{Z}_{\geqslant 0}}$$

are a bit trickier to compute. There are two possible strategies.

One strategy is based on the trigonometric identities

$$\boxed{\sin^2 x = \frac{1-\cos 2x}{2}, \quad \cos^2 x = \frac{1+\cos 2x}{2}}.$$

Using the change in variables $u = 2x$, so that

$$du = 2dx \Rightarrow dx = \frac{1}{2}du$$

we deduce

$$\int (\sin x)^{2m} (\cos x)^{2k} dx = \frac{1}{2^{m+k+1}} \int (1-\cos u)^m (1+\cos u)^k du.$$

The last integral involves <u>*smaller*</u> powers in $\cos u$. For example

$$\int \cos^4 x dx = \int \left(\frac{1+\cos u}{2}\right)^2 \frac{du}{2}, \quad u = 2x,$$

$$= \frac{1}{8}\int (1 + 2\cos u + \cos^2 u) du = \frac{1}{8}u + \frac{1}{4}\sin u + \frac{1}{8}\int \cos^2 u du$$

$(v = 2u = 4x)$

$$= \frac{1}{8}u + \frac{1}{4}\sin u + \frac{1}{8}\int \left(\frac{1+\cos v}{2}\right)\frac{dv}{2} = \frac{1}{8}u + \frac{1}{4}\sin u + \frac{1}{32}\int (1+\cos v) dv$$

$$= \frac{1}{4}x + \frac{1}{4}\sin(2x) + \frac{v}{32} + \frac{1}{32}\sin v + C$$

$$= \frac{x}{4} + \frac{1}{4}\sin(2x) + \frac{x}{8} + \frac{1}{32}\sin(4x) + C = \frac{3}{8}x + \frac{1}{4}\sin(2x) + \frac{1}{32}\sin(4x) + C.$$

One other possible strategy is to use the change in variables $u = \tan x$. Then

$$\cos^2 x = \frac{1}{1+\tan^2 x} = \frac{1}{1+u^2},\quad \sin^2 x = \cos^2 x \tan^2 x = \frac{\tan^2 x}{1+\tan^2 x} = \frac{u^2}{1+u^2}$$

$$du = d(\tan x) = (1+\tan^2 x)dx = (1+u^2)dx \Rightarrow dx = \frac{du}{1+u^2}.$$

We deduce

$$\int (\sin x)^{2m}(\cos x)^{2k}dx = \int \left(\frac{u^2}{1+u^2}\right)^m \left(\frac{1}{1+u^2}\right)^k \frac{du}{1+u^2}$$
$$= \int \frac{u^{2m}}{(1+u^2)^{m+k+1}}du.$$

Thus we need to know how to compute integrals of the form

$$\boxed{J(m,n) = \int \frac{u^{2m}}{(1+u^2)^n}du,\quad 0 \leqslant m < n,\quad m,n \in \mathbb{Z}}.$$

Observe first that when $m = 0$ the integrals $J(0,n)$ coincide with the integrals A_n of (8.5.10). The general case can be gradually reduced to the case $J(0,n)$ by observing that

$$J(m,n) = \int \frac{u^{2m} + u^{2m-2} - u^{2m-2}}{(1+u^2)^n}du = \int \frac{u^{2m-2}(1+u^2)}{(1+u^2)^n} - \int \frac{u^{2m-2}}{(1+u^2)^n}$$
$$= \int \frac{u^{2m-2}}{(1+u^2)^{n-1}} - J(m-1,n)$$

so that

$$\boxed{J(m,n) = J(m-1,n-1) - J(m-1,n)}. \qquad \square$$

The examples discussed above will allow us to describe a procedure for computing the antiderivatives of any *rational function*, i.e., a function $f(x)$ of the form

$$f(x) = \frac{P(x)}{Q(x)}$$

where $P(x)$ and $Q(x)$ are polynomials. Theoretically, the procedure works for any rational function, but the practical implementation can lead to complex computations. Such computation is possible because any rational function can be written as a sum of rational functions of the following simpler types.

Type I.

$$ax^n,\quad a \in \mathbb{R},\quad n = 0,1,2,\dots.$$

Type II.

$$\frac{a}{(x-r)^n},\quad c,r \in \mathbb{R},\quad n \in \mathbb{N}.$$

Type III.

$$\frac{bx+c}{\big((x-r)^2+a^2\big)^n},\quad a,b,c,r \in \mathbb{R},\quad n \in \mathbb{N}.$$

If the degree of the numerator $P(x)$ is smaller than the degree of the denominator $Q(x)$, then only the Type II and Type III functions appear in the decomposition of $\frac{P(x)}{Q(x)}$. The functions of Type II and III are also known as *partial fractions* or *simple fractions*.

Actually finding the decomposition of a rational function as a sum of simple fractions requires a substantial amount of work and it is not very practical for more complicated rational functions. For this reason we will not discuss this technique in great detail.

The primitives of a function of Type I are known. More precisely

$$\int ax^n dx = \frac{a}{n+1}x^{n+1} + C.$$

The primitives of the functions of Type II where computed in Example 8.5.10(e). To deal with the Type III functions we make a change in variables

$$x - r = at \Longleftrightarrow x = at + r.$$

Then

$$dx = adt,\ \ bx + c = b(at + r) + \beta = abt + rb + c,$$
$$(x-r)^2 + a^2 = a^2t^2 + a^2 = a^2(t^2+1),$$

so that

$$\int \frac{bx+c}{\big((x-r)^2+a^2\big)^n}dx = \int \frac{abt+rb+c}{a^{2n}(t^2+1)^n}adt$$
$$= \frac{b}{a^{2n-2}}\int \frac{t}{(t^2+1)^n}dt + \frac{rb+c}{a^{2n-1}}\int \frac{1}{(t^2+1)^n}dt.$$

The computation of integral

$$\int \frac{1}{(t^2+1)^n}dt$$

is described in (8.5.10), while the computation of the integral

$$\int \frac{t}{(t^2+1)^n}dt$$

as described in Example 8.5.10(e).

Let us illustrate this strategy on a simple example.

Example 8.5.11. Consider the rational function

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

Let us observe that

$$x^2 + 2x + 2 = (x+1)^2 + 1^2.$$

The function admits a decomposition of the form

$$\frac{1}{(x-1)^2(x^2+2x+2)} = f(x) = \frac{A_1}{x-1} + \frac{A_2}{(x-1)^2} + \frac{B_1x + C_1}{x^2+2x+2}.$$

Multiplying both sides by $(x-1)^2(x^2+2x+2)$ we deduce that for any $x \in \mathbb{R}$ we have

$$\begin{aligned} 1 &= A_1(x-1)(x^2+2x+2) + A_2(x^2+2x+2) + (B_1x+C_1)(x-1)^2. \\ &= A_1(x^3+2x^2+2x-x^2-2x-2) + A_2(x^2+2x+2) + (B_1x+C_1)(x^2-2x+1) \\ &= A_1(x^3+x^2-2) + A_2(x^2+2x+2) + (B_1x^3-2B_1x^2+B_1x+C_1x^2-2C_1x+C_1) \\ &= (A_1+B_1)x^3 + (A_1+A_2-2B_1+C_1)x^2 + (2A_2+B_1-2C_1)x - 2A_1 + 2A_2 + C_1. \end{aligned}$$

This implies

$$\begin{cases} A_1+B_1 &= 0 \\ A_1+A_2-2B_1+C_1 &= 0 \\ 2A_2+B_1-2C_1 &= 0 \\ -2A_1+2A_2+C_1 &= 1. \end{cases}$$

From the first equality we deduce $A_1 = -B_1$ and using this in the last three equalities above we deduce

$$\begin{cases} A_2-3B_1+C_1 &= 0 \\ 2A_2+B_1-2C_1 &= 0 \\ 2A_2+2B_1+C_1 &= 1. \end{cases}$$

From the first equality we deduce $A_2 = 3B_1 - C_1$. Using this in the last two equalities we deduce

$$\begin{cases} 7B_1-4C_1 &= 0 \\ 8B_1-C_1 &= 1. \end{cases}$$

Hence,

$$\frac{7}{4}B_1 = C_1 = 8B_1 - 1 \Rightarrow \frac{25}{4}B_1 = 1 \Rightarrow B_1 = \frac{4}{25},\;\; C_1 = \frac{7}{25} \Rightarrow A_1 = -\frac{4}{25},$$

$$A_2 = 3B_1 - C_1 = \frac{12}{25} - \frac{7}{25} = \frac{1}{5}.$$

Hence

$$\frac{1}{(x-1)^2(x^2+2x+2)} = -\frac{4}{25(x-1)} + \frac{1}{5(x-1)^2} + \frac{4x+7}{25\big((x+1)^2+1^2\big)}. \qquad \square$$

Example 8.5.12 (First order linear differential equations)**.** A quantity u that depends on time can be viewed as a function

$$u : I \to \mathbb{R},\;\; t \mapsto u(t),$$

where $I \subset \mathbb{R}$ is a time interval. We say that u satisfies a *linear first order differential equation* if u is differentiable and it satisfies an equality of the form

$$u'(t) + r(t)u(t) = f(t),\;\; \forall t \in I, \tag{8.5.13}$$

where $r, f : I \to \mathbb{R}$ are some given functions. Solving a differential equation such as (8.5.13) means finding all the differentiable functions $u : I \to \mathbb{R}$ satisfying the above equality. Let us look at some special examples.

(a) If $r(t) = 0$ for any $t \in I$, then (8.5.13) has the simpler form $u'(t) = f(t)$, so that $u(t)$ must be an antiderivative of $f(t)$.

(b) The general case. Suppose that $r(t)$ admits antiderivatives on I. The differential equation (8.5.13) is solved as follows.

Step 1. Choose one antiderivative $R(t)$ of $r(t)$, i.e., a function $R(t)$ such that $R'(t) = r(t)$.

Step 2. Multiply both sides of (8.5.13) by $e^{R(t)}$. We obtain the equality

$$e^{R(t)}u'(t) + e^{R(t)}r(t)u(t) = f(t)e^{R(t)}.$$

Now observe that the left-hand side of the above equality is the derivative of $e^{R(t)}u(t)$,

$$\big(\, e^{R(t)}u(t)\,\big)' = e^{R(t)}u'(t) + e^{R(t)}R'(t)u(t) = e^{R(t)}u'(t) + e^{R(t)}r(t)u(t) = f(t)e^{R(t)}.$$

This shows that $e^{R(t)}u(t)$ is an antiderivative of $f(t)e^{R(t)}$.

Step 3. Find one antiderivative $G(t)$ of $f(t)e^{R(t)}$. We deduce that there exists a constant $C \in \mathbb{R}$ such that

$$e^{R(t)}u(t) = G(t) + C \Rightarrow u(t) = e^{-R(t)}G(t) + Ce^{-R(t)}.$$

Take for example the equation

$$u'(t) + 2tu(t) = t.$$

In this case

$$r(t) = 2t, \;\; f(t) = t.$$

We can choose $R(t) = t^2$ and we have

$$\frac{d}{dt}\big(e^{t^2}u(t)\,\big) = e^{t^2}u'(t) + 2te^{t^2}u(t) = e^{t^2}t,$$

so that

$$e^{t^2}u(t) = \int e^{t^2}t dt = \frac{1}{2}\int e^{t^2}d(t^2) = \frac{1}{2}e^{t^2} + C$$
$$\Rightarrow u(t) = e^{-t^2}\left(C + \frac{1}{2}e^{t^2}\right) = Ce^{-t^2} + \frac{1}{2}. \qquad \square$$

8.6. Exercises

Exercise 8.1. Let $n \in \mathbb{N}$, $x_0, c_0, c_1, \dots, c_n \in \mathbb{R}$ and

$$P(x) = c_0 + \frac{c_1}{1!}(x-x_0) + \frac{c_2}{2!}(x-x_0)^2 + \cdots + \frac{c_n}{n!}(x-x_0)^n = \sum_{k=0}^{n} \frac{c_k}{k!}(x-x_0)^k.$$

(a) Prove that for any $k = 0, 1, 2, \dots, n$ we have

$$P^{(k)}(x_0) = c_k.$$

(b) Prove that if $Q(x) = q_0 + q_1 x + \cdots q_n x^n$ is a polynomial of degree $\leqslant n$ such that

$$Q^{(k)}(x_0) = c_k, \;\; \forall k = 0, 1, 2, \dots, n,$$

then $Q(x) = P(x)$, $\forall x \in \mathbb{R}$.

Hint. Consider the difference $D(x) = P(x) - Q(x)$, observe that

$$D^{(k)}(x_0) = 0, \;\; \forall k = 0, 1, 2, \dots, n,$$

and conclude from the above that $D(x) = 0$, $\forall x \in \mathbb{R}$. To reach this conclusion write

$$D(x) = d_0 + d_1 x + \cdots + d_n x^n,$$

and observe first that $D^{(n)}(x) = n! d_n$, $\forall x \in \mathbb{R}$. □

Exercise 8.2. Suppose that $a, b \in \mathbb{R}$, $b \geqslant 0$ and consider $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1+ax^2}{1+bx^2}.$$

Find the degree 4 Taylor polynomial of f at $x_0 = 0$. For which values of a, b does this polynomial coincide with the degree 4 Taylor polynomial of $\cos x$ at $x_0 = 0$?

Hint. To simplify the computations of the derivatives of f at 0 use the following trick. Let $N(x) = 1 + ax^2$ be the numerator of the fraction, $D(x) = 1 + bx^2$ be the denominator. Then

$$N(0) = D(0) = 1, \;\; N'(0) = D'(0) = 0, \;\; N''(0) = 2a, \;\; D''(0) = 2b, \tag{8.6.1}$$

$$N^{(k)}(x) = D^{(k)}(x) = 0, \;\; \forall k \geqslant 3, \;\; x \in \mathbb{R}. \tag{8.6.2}$$

We have

$$N(x) = D(x)f(x), \;\; N'(x) = D'(x)f(x) + D(x)f'(x),$$
$$2a = N''(x) = D''(x)f(x) + 2D'(x)f'(x) + D(x)f''(x),$$
$$0 = N^{(n)}(x) \overset{(7.6.1)}{=} \sum_{k=0}^{n} \binom{n}{k} D^{(k)}(x) f^{(n-k)}(x) \overset{(8.6.2)}{=} \sum_{k=0}^{2} \binom{n}{k} D^{(k)}(x) f^{(n-k)}(x)$$
$$= D(x)f^{(n)}(x) + nD'(x)f^{(n-1)}(x) + \frac{n(n-1)}{2} D''(x) f^{(n-2)}(x), \;\; \forall n > 2.$$

We deduce

$$f(0) = D(0)f(0) = N(0) = 1, \;\; f'(0) = D(0)f'(0) = N'(0) - D'(0)f(0) \overset{(8.6.1)}{=} 0,$$
$$f''(0) = D(0)f''(0) = N''(0) - 2D'(0)f'(0) - D''(0)f(0) \overset{(8.6.1)}{=} N''(0) - D''(0)f(0) = 2a - 2b,$$
$$f^{(n)}(0) = D(0)f^{(n)}(0) = -nD'(0)f^{(n-1)}(0) - \frac{n(n-1)}{2} D''(0) f^{(n-2)}(0)$$
$$\overset{(8.6.1)}{=} -bn(n-1)f^{(n-2)}(0), \;\; n > 2.$$

□

Exercise 8.3. Use the inequality $2 < e < 3$ and the strategy outlined in Remark 8.1.6 to show that

$$\left| e^h - \left(1 + \frac{h}{1!} + \cdots + \frac{h^n}{n!}\right) \right| \leqslant \frac{3|h|^{n+1}}{(n+1)!}, \quad \forall |h| \leqslant 1.$$

□

Exercise 8.4. Using Example 8.1.7 as a guide, compute $\cos 1$ up to two decimals. □

Exercise 8.5. Approximate $\sqrt[3]{8.1}$ using the degree 3 Taylor polynomial of $f(x) = \sqrt[3]{x}$ at $x_0 = 8$. Estimate the error of this approximation using the Lagrange estimate (8.1.3). □

Exercise 8.6. Find the Taylor series of the function

$$f(x) = \frac{1}{1-x}, \quad x \neq 1$$

at $x_0 = 0$. For which values of x is this series convergent? □

Exercise 8.7. Prove that the Taylor series of $\ln(1-x)$ at $x_0 = 0$ is

$$-\sum_{n=1}^{\infty} \frac{x^n}{n}.$$

and then show that this series converges to $\ln(1-x)$ for any $x \in (-1, \frac{1}{2})$.

Hint. Use Corollary 8.1.5.[2] □

Exercise 8.8. (a) Prove that the Taylor series of $\sin x$ at $x_0 = 0$,

$$\sum_{k \geqslant 0} (-1)^k \frac{x^{2k+1}}{(2k+1)!},$$

is absolutely convergent for any $x \in \mathbb{R}$ and its sum is $\sin x$. Show that the convergence is uniform on any interval $[-R, R]$.

(b) Prove that the Taylor series of $\cos x$ at $x_0 = 0$,

$$\sum_{k \geqslant 0} (-1)^k \frac{x^{2k}}{(2k)!}$$

is absolutely convergent and for any $x \in \mathbb{R}$ and its sum is $\cos x$. Show that the convergence is uniform on any interval $[-R, R]$.

Hint. Use Corollary 8.1.5. □

Exercise 8.9. Find

$$\lim_{x \to \infty} x \left[\frac{1}{e} - \left(\frac{x}{x+1} \right)^x \right].$$

□

[2] The Taylor series of $\ln(1-x)$ at $x_0 = 0$ converges to $\ln(1-x)$ *for all* $|x| < 1$. However, the Lagrange remainder formula is not strong enough to prove this. We need a different remainder formula (9.6.18) to prove this stronger statement. For details see Example 9.6.10.

Exercise 8.10. Using the fact that the function $\ln : (0,\infty) \to \mathbb{R}$ is concave prove *Young's inequality*: if $p, q \in (1,\infty)$ are such that

$$\frac{1}{p} + \frac{1}{q} = 1,$$

then

$$xy \leqslant \frac{x^p}{p} + \frac{y^q}{q}, \;\; \forall x, y > 0. \tag{8.6.3}$$

□

Exercise 8.11. Use the AM-GM inequality to prove that if $x \in \mathbb{R}$, $n, m \in \mathbb{N}$ and $-x < n < m$, then

$$\Bigl(1 + \frac{x}{n}\Bigr)^n \leqslant \Bigl(1 + \frac{x}{m}\Bigr)^m .$$ □

Exercise 8.12. Let $x_1, \dots, x_n > 0$.

(i) Prove that

$$x_1^2 + \cdots + x_n^2 + \frac{1}{(x_1 \cdots x_n)^2} \geqslant n + 1.$$

(ii) Prove that

$$\sum_{1 \leqslant i < j \leqslant n} x_i x_j + \sum_{k=1}^n \frac{1}{x_k^{n-1}} \geqslant \frac{n(n+1)}{2}.$$

□

Exercise 8.13. Suppose that $a < b$ are two real numbers and $f : (a,b) \to \mathbb{R}$ is a convex function.

(a) Prove that for any $x_1 < x_2 < x_3 \in (a,b)$ we have

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_1)}{x_3 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

Hint. Give a geometric interpretation to this statement and then think geometrically.

(b) Suppose that $x_0 \in (a,b)$. Prove that the one-sided limits

$$m_{\pm}(x_0) = \lim_{h \to 0\pm} \frac{f(x_0 + h) - f(x_0)}{h}$$

exist, are finite and $m_-(x_0) \leqslant m_+(x_0)$.

(c) Suppose $x_0 \in (a,b)$ and $m_{\pm}(x_0)$ are as above. Fix $m \in [m_-(x_0), m_+(x_0)]$. Show that

$$f(x) \geqslant f(x_0) + m(x - x_0), \;\; \forall x \in (a,b).$$

Can you give a geometric interpretation of this fact?

(d) Prove that $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ is convex. For $x_0 := 0$, compute the numbers $m_{\pm}(x_0)$ defined as in (b). □

Exercise 8.14. [3] Suppose that $f : [0,1] \to [0,\infty)$ is a C^2-function satisfying the following additional properties.

(i) $f'(x) \geqslant 0$, $\forall x \in [0,1]$.

(ii) $f''(x) > 0$, $\forall x \in (0,1)$.

(iii) $f(1) = 1$, $f'(1) > 1$ and $f(0) > 0$.

Prove that the following hold.

(a) $f(x) \in [0,1]$, $\forall x \in [0,1]$.

(b) If $x_0 \in (0,1)$ is a *fixed point* of f, i.e., $f(x_0) = x_0$, then $f'(x_0) < 1$.

Hint. Argue by contradiction. Use the Mean Value Theorem with the quotient

$$\frac{f(1) - f(x_0)}{1 - x_0}.$$

(c) The function f has a *unique* fixed point x_* located in the *open* interval $(0,1)$.

Hint. Argue by contradiction. Suppose that f has two fixed points $x_* < y_*$. in $(0,1)$. Use the Mean Value Theorem for the quotient

$$\frac{f(y_*) - f(x_*)}{y_* - x_*}$$

and reach a contradiction using (b).

(d) Fix $s \in (0,1)$ and consider the sequence (x_n) defined by the recurrence

$$x_0 = s, \quad x_{n+1} = f(x_n), \quad \forall n \geqslant 0.$$

Prove that

$$\lim_n x_n = x_*,$$

where x_* is the unique fixed point of f located in the interval $(0,1)$.

Hint. The sequence is bounded since it lies in $[0,1]$. Show that the sequence is monotone and the limit lies in $(0,1)$. □

Exercise 8.15. Prove that for any $n \in \mathbb{N}$ and any numbers $x_1, x_2, \dots, x_n \geqslant 0$ we have

$$\left(\frac{x_1 + \cdots + x_n}{n}\right)^2 \leqslant \frac{x_1^2 + \cdots + x_n^2}{n}.$$

Hint. Use the Cauchy-Schwarz inequality. □

Exercise 8.16. Suppose that $(x_n)_{n\geqslant 1}$ is a sequence of real numbers > 0. We set

$$P_n^+ := \prod_{k=1}^n (1 \pm x_k), \quad S_n = \sum_{k=1}^n x_k.$$

[3] The results in this exercise are particularly useful in probability theory in the investigation of the so called *branching processes.*

(i) Prove that

$$1 + S_n \leqslant P_n^+ \leqslant e^{S_n}.$$

Hint. Prove first the case $n = 1$.

(ii) Prove that P_n^+ is convergent iff S_n is convergent.

(iii) Assume that $x_n \neq 1$, $\forall n$. Prove that P_n^- converges to a *nonzero* limit iff S_n is convergent. **Hint.** Use the equality $\lim_{x\searrow 0} \frac{\log(1-x)}{x} = 1$, where $\log = \ln$.

□

Exercise 8.17. Consider the *Gauss bell*, i.e., the function

$$\gamma : \mathbb{R} \to \mathbb{R}, \;\; \gamma(x) = e^{-\frac{x^2}{2}}.$$

(a) Prove that for any $n \in \mathbb{N}$ there exists a polynomial $H_n(x)$ of degree n such that

$$\gamma^{(n)}(x) = (-1)^n H_n(x)\gamma(x).$$

(The polynomial $H_n(x)$ is called the *degree n Hermite polynomial.*)

(b) Prove that

$$H_{n+1}(x) = xH_n(x) - H_n'(x), \;\; \forall n \in \mathbb{N}.$$

(c) Compute $H_1(x)$, $H_2(x)$, $H_3(x)$.

(d) Find the intervals of convexity and concavity of $\gamma(x)$.

(e) Sketch the graph of the function $\gamma(x)$. □

Exercise 8.18. Consider the *hyperbolic functions*

$$\cosh, \sinh : \mathbb{R} \to \mathbb{R}, \;\; \cosh x = \frac{e^x + e^{-x}}{2}, \;\; \sinh(x) = \frac{e^x - e^{-x}}{2}, \;\; \forall x \in \mathbb{R}.$$

(cosh=hyperbolic cosine, sinh= hyperbolic sine)

(i) Prove that

$$\cosh' x = \sinh x, \;\; \sinh' x = \cosh x,$$

$$\cosh^2 x - \sinh^2 x = 1, \;\; \cosh^2 x + \sinh^2 x = \cosh(2x), \;\; \forall x \in \mathbb{R}.$$

(ii) Find the Taylor series of $\cosh x$ and $\sinh x$ at $x_0 = 0$ and prove that they converge to $\cosh x$ and respectively $\sinh x$ for any $x \in \mathbb{R}$.

(iii) Prove that

$$\forall x \neq 0, \;\; \cosh(x) > 1 + \frac{x^2}{2} \;\text{ and }\; \frac{\sinh(x)}{x} > 1 + \frac{x^2}{6}.$$

(iv) Prove that the function sinh is bijective and then find its inverse.

(v) Sketch the graphs of cosh and sinh.

□

Exercise 8.19. Consider the function

$$f:(e,\infty)\to\mathbb{R},\quad f(x)=\log\big(\log x\big)$$

where e is Euler's number and log denotes the natural logarithm.

(i) Prove that f is concave and $f(x)>0$, $\forall x>e$.

(ii) Prove that for any $x,y>e$

$$\big((\log x)(\log y)\big)^{\frac{1}{2}}\leqslant\log\Big(\frac{x+y}{2}\Big).$$

□

Exercise 8.20. Compute

$$\int xe^{2x}dx,\quad \int xe^{2x}\cos x dx,\int xe^{2x}\sin x dx,\quad \int\sin^3 x\cos^2 x dx.$$ □

Exercise 8.21. Compute

$$\int\frac{1}{(4+x^2)^5}dx$$

by reducing it to the computation in Example 8.5.8(d). □

Exercise 8.22. Compute

$$\int(\cos x)^{11}dx.$$ □

Exercise 8.23. Using the strategy outlined in Example 8.5.12 find the function $u(t)$, $v(t)$ and $f(t)$ satisfying the differential equations

$$u'(t)+2u(t)=t,\quad v'(t)-v(t)=\cos t,$$

$$f'(t)-(\tan t)f(t)=t,\quad -\frac{\pi}{2}<t<\frac{\pi}{2}.$$ □

Exercise 8.24. Suppose that we are given a huge container containing 200 liters of pure water. In this container, starting at $t=0$, we continuously add 10 liters of salted water per minute containing 1.5 grams of salt per liter and, at the same time, the container is leaking salt-water mixture at a constant rate of 10 liters per minute. Denote by $m(t)$ the amount of salt (in grams) contained in the mixture after t minutes from the start.

(a) Prove that $m(t)$ satisfies the differential equation

$$\frac{dm}{dt}=15-\frac{m(t)}{20}.$$

(b) Recalling that initially there was no salt in the water, i.e., $m(0)=0$, find $m(t)$ for any $t>0$. □

8.7. Exercises for extra credit

Exercise* 8.1. Suppose that $f : (0,\infty) \to \mathbb{R}$ is a differentiable function such that

$$\lim_{x\to\infty} \big(f(x) + f'(x) \big) = 0.$$

Show that

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} f'(x) = 0. \qquad \square$$

Exercise* 8.2. (a) Prove that for any $n \in \mathbb{N}$ and any real numbers $a, r > 0$ we have

$$a^{\frac{n}{n+1}} \leqslant \frac{1}{r}\left(\frac{r^{n+1}}{n+1} + \frac{na}{n+1} \right).$$

Hint: Use Young's inequality (8.6.3).

(b) Prove that if $\sum_{n\geqslant 0} a_n$ is a convergent series of positive numbers, then so is $\sum_{n\geqslant 0} a_n^{\frac{n}{n+1}}$.$\square$

Exercise* 8.3. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a C^3-function. Prove that there exists $a \in \mathbb{R}$ such that

$$f(a) \cdot f'(a) \cdot f''(a) \cdot f'''(a) \geqslant 0. \qquad \square$$

Exercise* 8.4. Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function satisfying

$$f\left(\frac{x+y}{2} \right) \leqslant \frac{f(x)+f(y)}{2}, \quad \forall x, y \in [a,b].$$

Prove that f is convex. $\square$

Exercise* 8.5. Suppose that $f : (a,b) \to \mathbb{R}$ is a convex function. Prove that f is continuous. $\square$

Exercise* 8.6. Show that for any positive real numbers a, b, c we have

$$a + b + c \leqslant \frac{a^3}{bc} + \frac{b^3}{ac} + \frac{c^3}{ab}. \qquad \square$$

Exercise* 8.7. Fix a natural number n and positive real numbers $x_1, \dots, x_n$. For any $\alpha > 0$ we set

$$M_\alpha(x_1, \dots, x_n) := \left(\frac{x_1^\alpha + \cdots + x_n^\alpha}{n} \right)^{\frac{1}{\alpha}}.$$

(a) Show that

$$M_\alpha(x_1, \dots, x_n) \leqslant M_\beta(x_1, \dots, x_n), \quad \forall 0 < \alpha < \beta.$$

(b) Compute

$$\lim_{\alpha\to 0+} M_\alpha(x_1, \dots, x_n). \qquad \square$$

Exercise* 8.8. (a) Prove that for any $n \in \mathbb{N}$ the equation $x^n + x = 1$ has a unique positive solution x_n.

(b) Prove that

$$\lim_{n\to\infty} x_n = 1. \qquad \square$$

Chapter 9

Integral calculus

9.1. The integral as area: a first look

The Riemann integral is a very complicated infinite summation process that is often required when we want to compute areas or volumes of more irregular regions.

By way of motivation, let us consider a famous problem first solved by Archimedes by other means. Consider the arc of parabola in Figure 9.1 given by the equation

$$y = x^2, \;\; 0 \leqslant x \leqslant 1.$$

We would like to compute the area of the region R between the x-axis, the parabola and the vertical line $x = 1$.

Let us observe that we do not have a precise definition of the concept of area. We only have an intuitive belief that

(i) the area of a rectangle is width × length, and

(ii) the area of a union of rectangles that intersect only along edges should be the sum of the area of the rectangles. We will refer to such regions as *simple type* regions.

We proceed by approximating R by a region of simple type. We subdivide the interval $[0, 1]$ into N equal parts, where N is a very large natural number. We obtain the points

$$x_0 = 0, \;\; x_1 = \frac{1}{N}, \;\; x_2 = \frac{2}{N}, \dots, x_N = \frac{N}{N}.$$

For each $k = 1, 2, \dots, N$ we denote by R_k the very thin slice of R of width $\frac{1}{N}$ delimited by the vertical lines $x = x_{k-1}$ and $x = x_k$. We have thus decomposed R into N thin slices

$R_1, \dots, R_N$ and

$$\operatorname{area}(R) = \sum_{k=1}^{n} \operatorname{area}(R_k) = \operatorname{area}(R_1) + \cdots + \operatorname{area}(R_N).$$

Now observe that the slice R_k contains a thin rectangle $\underline{R}_k$ of height $f(x_{k-1})$ and is contained in a thin rectangle $\overline{R}_k$ of height $f(x_k)$; see Figure 9.1.

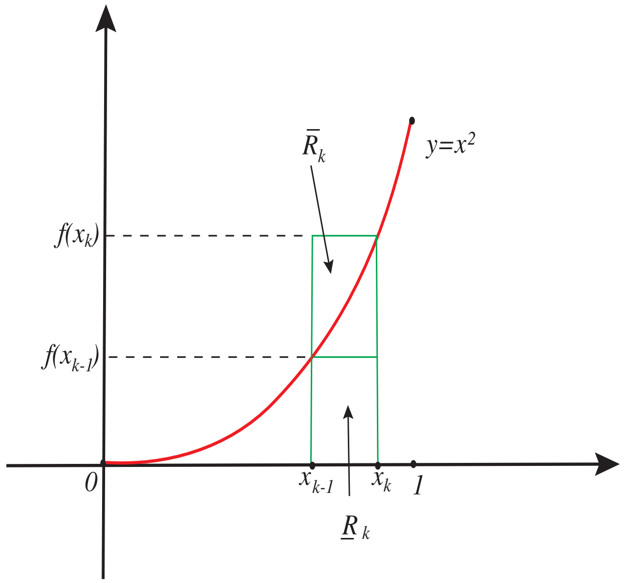

Figure 9.1. *Computing the area underneath an arc of parabola.*

Thus

$$f(x_{k-1}) \times (x_k - x_{k-1}) = \operatorname{area}(\underline{R}_k) \leqslant \operatorname{area}(R_k) \leqslant \operatorname{area}(\overline{R}_k) = f(x_k) \times (x_k - x_{k-1}).$$

Since $f(x_k) = \frac{k^2}{N^2}$ and $x_k - x_{k-1} = \frac{1}{N}$ we deduce

$$\frac{(k-1)^2}{N^3} \leqslant \operatorname{area}(R_k) \leqslant \frac{k^2}{N^3},$$

and thus

$$\underbrace{\sum_{k=1}^{N} \frac{(k-1)^2}{N^3}}_{=:L_N} \leqslant \underbrace{\sum_{k=1}^{N} \operatorname{area}(R_k)}_{=\operatorname{area}(R)} \leqslant \underbrace{\sum_{k=1}^{N} \frac{k^2}{N^3}}_{=:U_N}. \tag{9.1.1}$$

Thus

$$L_N \leqslant \operatorname{area}(R) \leqslant U_N. \tag{9.1.2}$$

Observe that

$$L_N = \frac{0^2}{N^3} + \frac{1^2}{N^3} + \cdots + \frac{(N-1)^2}{N^3} = \frac{1^2 + 2^2 + \cdots + (N-1)^2}{N^3},$$

$$U_N = \frac{1^2}{N^3} + \cdots + \frac{(N-1)^2}{N^3} + \frac{N^2}{N^3} = \frac{1^2 + 2^2 + \cdots + N^2}{N^3},$$

so that

$$U_N - L_N = \frac{N^2}{N^3} = \frac{1}{N}.$$

For N very large, the difference $U_N - L_N$ is very small and thus the sequence (L_N) converges if and only if the sequence (U_N) converges. Moreover, the inequality (9.1.2) shows that the common limit of these sequences, if it exists, must be equal to the area of R. To compute the limit of U_N we use the following famous identity whose proof is left to you as an exercise.

$$1^2 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}. \tag{9.1.3}$$

We deduce that

$$U_N = \frac{N(N+1)(2N+1)}{6N^3} = \frac{1}{6}\frac{N}{N}\frac{N+1}{N}\frac{2N+1}{N} \to \frac{2}{6} \text{ as } N \to \infty.$$

Thus

$$\operatorname{area}(R) = \frac{1}{3}.$$

This example describes the bare bones of the process called *integration*. As this simple example suggests, the integration it involves a sophisticated infinite summation and a bit of good fortune, in the guise of (9.1.3), that allowed us to actually compute the result of this infinite summation.

We will spend the rest of this chapter describing rigorously and in great generality this process and we will show that in a large number of cases we can cleverly create our good fortune and succeed in carrying out explicit computations of the limits of infinite summations involved.

9.2. The Riemann integral

The process sketched in the previous section can be carried out in greater generality. We present the quite involved details in this section.

Definition 9.2.1 (Partitions)**.** Fix an interval $[a,b]$, $a < b$.

(a) A *partition* $\boldsymbol{P}$ of $[a,b]$ is a finite collection of points $x_0, x_1, \dots, x_n$ of the interval such that

$$a = x_0 < x_1 < \cdots < x_n = b.$$

The natural number n is called the *order* of the partition, while the points $x_0, \dots, x_n$ are called the *nodes* of the partition. The intervals

$$[x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n]$$

are called the *intervals of the partition*. The interval $[x_{k-1}, x_k]$ is called the k-th interval of the partition and it is denoted by $I_k(\boldsymbol{P})$. Its length is denoted by $\Delta_k(\boldsymbol{P})$ or Δx_k. The largest of these lengths is called the *mesh size* of the partition and it is denoted by $\|\boldsymbol{P}\|$,

$$\boxed{\|\boldsymbol{P}\| := \max_{1\leqslant k\leqslant n}(x_k - x_{k-1}) = \max_{1\leqslant k\leqslant n}\Delta_k(\boldsymbol{P})}.$$

We denote by $\mathcal{P}_{[a,b]}$ the collection of all partitions of the interval $[a, b]$.

(b) A *sample* of a partition $\boldsymbol{P}$ of order n is a collection $\underline{\xi}$ consisting of n points $\xi_1, \dots, \xi_n$ such that

$$\xi_k \in I_k(\boldsymbol{P}),\ \ \forall k = 1, \dots, n.$$

The point ξ_k is called the *sample point* of the interval $I_k(\boldsymbol{P})$. We denote by $\mathcal{S}(\boldsymbol{P})$ the collection of all possible samples of the partition $\boldsymbol{P}$.

(c) A *sampled partition* of the interval $[a, b]$ is a pair $(\boldsymbol{P}, \underline{\xi})$, where $\boldsymbol{P}$ is a partition of $[a, b]$ and $\underline{\xi} \in \mathcal{S}(\boldsymbol{P})$ is a sample of $\boldsymbol{P}$. □

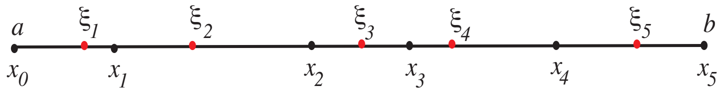

Figure 9.2. *A sampled partition of order* 5 *of an interval* $[a, b]$. *Its longest interval is* $[x_1.x_2]$ *so its mesh size is* $(x_2 - x_1)$.

Example 9.2.2. Any compact interval $[a, b]$ has a natural partition $\boldsymbol{U}_n$ of order n corresponding to a subdivision of $[a, b]$ into n subintervals of order n. More precisely, $\boldsymbol{U}_n$ is defined by the points

$$x_0 = a,\ \ x_1 = a + \frac{1}{n}(b-a),\ \ x_k = a + \frac{k}{n}(b-a),\ \ k = 0, 1, \dots, n.$$

The partition $\boldsymbol{U}_n$ is called the *uniform partition of order* n of $[a, b]$. Note that

$$\|\boldsymbol{U}_n\| = \frac{b-a}{n}.$$ □

Definition 9.2.3. Let $f : [a, b] \to \mathbb{R}$ be a function defined on the closed and bounded interval $[a, b]$. Given a partition $\boldsymbol{P} = (x_0 < \cdots < x_n)$ of $[a, b]$, and a sample $\underline{\xi}$ of $\boldsymbol{P}$, we define the *Riemann*[1] *sum* of f associated to the sampled partition $(P, \underline{\xi})$ to be the number

$$\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) = \sum_{k=1}^{n} f(\xi_k)\Delta_k(\boldsymbol{P}) = \sum_{k=1}^{n} f(\xi_k)\Delta x_k = \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}).$$ □

As depicted in Figure 9.3, each term $f(\xi_k)(x_k - x_{k-1})$ in a Riemann sum is equal to the area of a "thin" rectangle of width $\Delta x_k = (x_k - x_{k-1})$, and height given by the

[1] Named after Bernhardt Riemann (1826-1866) German mathematician who made lasting and revolutionary contributions to analysis, number theory, and differential geometry; see Wikipedia.

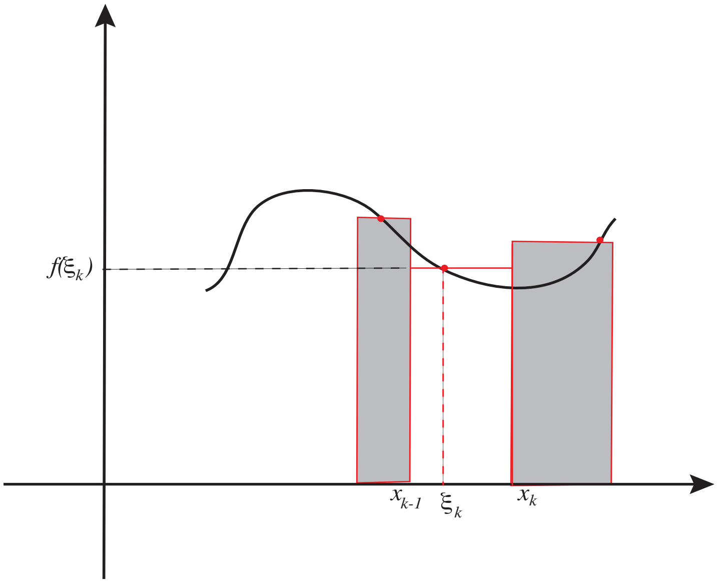

Figure 9.3. *The term* $f(\xi_k)\Delta x_k$ *is the area of a rectangle.*

altitude of the point on the graph of f determined by the sample point $\xi_k \in [x_{k-1}, x_k]$. The Riemann sum is therefore the area of the region formed by putting side by side each of these thin rectangles. The hope is that the area of this rather jagged looking region is an approximation for the area of the region under the graph of f. The next definition makes this intuition precise.

Definition 9.2.4. Suppose that $f : [a,b] \to \mathbb{R}$ is a function defined on the *closed and bounded* interval $[a,b]$. We say that f is *Riemann integrable on* $[a,b]$ if there exists a real number I with the following property: for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that, *for any partition* $\boldsymbol{P}$ of $[a,b]$ with mesh size $\|\boldsymbol{P}\| < \delta$, *and any sample* $\underline{\xi}$ of $\boldsymbol{P}$ we have

$$\big|\, I - \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \,\big| < \varepsilon.$$

Equivalently, as a quantified statement, the above reads

$$\begin{aligned} \exists I \in \mathbb{R},\ \ \forall \varepsilon > 0,\ \ \exists \delta = \delta(\varepsilon) > 0,\ \ \forall \boldsymbol{P} \in \mathcal{P}_{[a,b]},\ \ \forall \underline{\xi} \in \mathcal{S}(\boldsymbol{P}) : \\ \|\boldsymbol{P}\| < \delta \Rightarrow \big|\, I - \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \,\big| < \varepsilon. \end{aligned} \tag{9.2.1}$$

We will denote by $\mathcal{R}[a,b]$ the collection of all Riemann integrable functions $f : [a,b] \to \mathbb{R}$. □

Suppose that $f : [a,b] \to \mathbb{R}$ is Riemann integrable. For any $n \in \mathbb{N}$ we fix a sample $\underline{\xi}^{(n)}$ of $\boldsymbol{U}_n$, the uniform partition of order n of $[a,b]$. If I is *any* real number satisfying (9.2.1), then from the equality

$$\lim_{n\to\infty} \|\boldsymbol{U}_n\| = 0$$

we deduce that

$$I = \lim_{n\to\infty} \boldsymbol{S}\big(\, f, \boldsymbol{U}_n, \underline{\xi}^{(n)}\,\big).$$

Since a convergent sequence has a *unique* limit, we deduce that there exists precisely one real number I satisfying (9.2.1). This real number is called the *Riemann integral* of f on $[a, b]$ and it is denoted by

$$\int_a^b f(x)dx.$$

It bears repeating the definition of $\int_a^b f(x)dx$.

*The Riemann integral of f over $[a, b]$, when it exists, is the **unique real number** $\int_a^b f(x)dx$ with the following property: for any $\varepsilon > 0$ there exists $= \delta = \delta(\varepsilon) > 0$ such that for any partition $\boldsymbol{P}$ of $[a, b]$ with mesh $\|\boldsymbol{P}\| < \delta$, and for any sample $\underline{\xi}$ of $\boldsymbol{P}$, the Riemann sum $\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi})$ is within ε of $\int_a^b f(x)dx$, i.e.,*

$$\left| \int_a^b f(x)dx - \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \right| < \varepsilon.$$

We can *loosely* rephrase this as follows

$$\int_a^b f(x)dx = \lim_{\substack{\|\boldsymbol{P}\|\to 0,\\ \underline{\xi}\in\mathcal{S}(\boldsymbol{P})}} \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}). \tag{9.2.2}$$

Example 9.2.5. Consider the constant function $f : [a, b] \to \mathbb{R}$, $f(x) = C$, for all $x \in [a, b]$ where C is a fixed real number. Note that for any sampled partition of order n $(\boldsymbol{P}, \underline{\xi})$ of $[a, b]$ we have

$$\begin{aligned} \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) &= f(\xi_1)(x_1 - x_0) + f(\xi_2)(x_2 - x_1) + \cdots + f(\xi_n)(x_n - x_{n-1}) \\ &= C(x_1 - x_0) + C(x_2 - x_1) + \cdots + C(x_n - x_{n-1}) \\ &= C\Big(\,(x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1})\,\Big) = C(x_n - x_0) = C(b - a). \end{aligned}$$

This shows that the constant function is integrable and

$$\int_a^b C dx = C(b - a).$$

□

It is natural to ask if there exist Riemann integrable functions more complicated than the constant functions. The next section will address precisely this issue. We will see that indeed, the world of integrable functions is very large. Until then, let us observe that not any function is Riemann integrable.

Proposition 9.2.6. *Suppose that $f : [a, b] \to \mathbb{R}$ is a Riemann integrable function. Then f is bounded, i.e.,*

$$-\infty < \inf_{x\in[a,b]} f(x) < \sup_{x\in[a,b]} f(x) < \infty.$$

Proof. We argue by contradiction. Suppose that $f:[a,b]\to\mathbb{R}$ is Riemann integrable and unbounded above, i.e.,

$$\sup_{x\in[a,b]} f(x)=\infty.$$

For any $n\in\mathbb{N}$ consider the uniform partition $\boldsymbol{U}_n$ of $[a,b]$. Then there exists $k=k(n)$ such that f is unbounded the interval $I_k=I_{k(n)}$ of this partition. For $j\neq k$ fix an arbitrary sample point $\xi_j\in I_j$. Since f is not bounded above on I_k, there exists $\xi_k\in I_k$ such that

$$f(\xi_k)>\frac{n}{\Delta x_k}-\sum_{j\neq k}f(\xi_j)\frac{\Delta x_j}{\Delta x_k}\Longleftrightarrow f(\xi_k)\Delta x_k+\sum_{j\neq k}f(\xi_j)\Delta x_j>n.$$

We obtain a sample $\underline{\xi}^{(n)}$ of $\boldsymbol{U}_n$ and for this sample we have

$$\boldsymbol{S}\big(f,\boldsymbol{U}_n,\underline{\xi}^{(n)}\big)=f(x_k)\Delta x_k+\sum_{j\neq k}f(\xi_j)\Delta x_j>n,\ \ \forall n\in\mathbb{N}.$$

The Riemann integrability of f implies that the sequence of Riemann sums $\boldsymbol{S}\big(f,\boldsymbol{U}_n,\underline{\xi}^{(n)}\big)$ is convergent. This contradicts the last inequality which states that this sequence is unbounded. □

The above result shows that the function

$$f:[0,1]\to\mathbb{R},\ \ f(x)=\begin{cases}0, & x=0,\\ \frac{1}{\sqrt{x}}, & x\in(0,1],\end{cases}$$

is not Riemann integrable because it is not bounded.

9.3. Darboux sums and Riemann integrability

To be able to construct examples of integrable functions we need a criterion for recognizing such functions, more flexible than the definition. Fortunately there is one such criterion due to Gaston Darboux. To formulate it we need to introduce several new concepts.

Definition 9.3.1. Suppose that $f:[a,b]\to\mathbb{R}$ is a *bounded* function defined on the closed and bounded interval $[a,b]$. For any partition $\boldsymbol{P}$ of $[a,b]$ of order n we set

$$\boldsymbol{S}^*(f,\boldsymbol{P}):=\sum_{k=1}^{n}\sup_{x\in I_k(\boldsymbol{P})}f(x)\Delta x_k,$$

$$\boldsymbol{S}_*(f,\boldsymbol{P}):=\sum_{k=1}^{n}\inf_{x\in I_k(\boldsymbol{P})}f(x)\Delta x_k,$$

$$\omega(f,\boldsymbol{P}):=\sum_{k=1}\operatorname{osc}(f,I_k)\Delta x_k,$$

where

- $I_k=I_k(\boldsymbol{P})$ is the k-th interval of the partition $\boldsymbol{P}$,
- Δx_k is the length of I_k,

- $\operatorname{osc}(f, I_k)$ denotes the oscillation of f on I_k.

The quantity $\boldsymbol{S}^*(f, \boldsymbol{P})$ is called *the upper Darboux*[2] *sum* of the function f determined by the partition $\boldsymbol{P}$, while $\boldsymbol{S}_*(f, \boldsymbol{P})$ is called *the lower Darboux sum* of the function f determined by the partition $\boldsymbol{P}$. We will refer to $\omega(f, \boldsymbol{P})$ as the *mean oscillation of* f along $\boldsymbol{P}$. □

Proposition 9.3.2. *If $f : [a, b] \to \mathbb{R}$ is a bounded function, then for any partition $\boldsymbol{P}$ of $[a, b]$ and any sample $\underline{\xi}$ of $\boldsymbol{P}$ we have*

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}), \tag{9.3.1a}$$

$$\omega(f, \boldsymbol{P}) = \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}). \tag{9.3.1b}$$

Proof. Suppose that $\boldsymbol{P}$ is a partition of order n of $[a, b]$ and $\underline{\xi}$ is a sample of $\boldsymbol{P}$. For $k = 1, \dots, n$ we denote by I_k the k-the interval of $\boldsymbol{P}$ and we set

$$M_k := \sup_{x \in I_k} f(x), \quad m_k := \inf_{x \in I_k} f(x).$$

Then $M_k - m_k = \operatorname{osc}(f, I_k)$ and

$$\begin{aligned} \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}) &= \big(M_1 \Delta x_1 + \cdots + M_n \Delta x_n \big) - \big(m_1 \Delta x_1 + \cdots + m_n \Delta x_n \big) \\ &= (M_1 - m_1)\Delta x_1 + \cdots + (M_n - m_n)\Delta x_n \\ &= \operatorname{osc}(f, I_1)\Delta x_1 + \cdots + \operatorname{osc}(f, I_n)\Delta x_n = \omega(f, \boldsymbol{P}). \end{aligned}$$

This proves (9.3.1b). If $\underline{\xi}$ is a sample of $\boldsymbol{P}$, then

$$m_k \Delta x_k \leqslant f(\xi_k)\Delta x_k \leqslant M_k \Delta x_k, \quad \forall k = 1, \dots, n,$$

so that

$$\sum_{k=1}^n m_k \Delta x_k \leqslant \sum_{k=1}^n f(\xi_k)\Delta x_k \leqslant \sum_{k=1}^n M_k \Delta x_k.$$

This proves (9.3.1a). □

Corollary 9.3.3. *If $f : [a, b] \to \mathbb{R}$ is a bounded function then for any partition P of $[a, b]$ and for any samples $\underline{\xi}, \underline{\xi}'$ of $\boldsymbol{P}$ we have*

$$\big| \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) - \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}') \big| \leqslant \omega(f, \boldsymbol{P}).$$

Proof. According to (9.3.1a) the Riemann sums $\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi})$, $\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}')$ are both contained in the interval $[\boldsymbol{S}_*(f, \boldsymbol{P}), \boldsymbol{S}^*(f, \boldsymbol{P})]$ so the distance between them must be smaller than the length of this interval which is equal to $\omega(f, \boldsymbol{P})$ according to (9.3.1b). □

[2] Named after Gaston Darboux (1842-1917) French mathematician who made several important contributions to geometry and mathematical analysis; see Wikipedia.

Proposition 9.3.4. *Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function and $\boldsymbol{P}$ is a partition of $[a,b]$. If $\boldsymbol{P}'$ is a partition of $[a,b]$ obtained from $\boldsymbol{P}$ by adding one extra node x' in the interior of some interval of $\boldsymbol{P}$, then*

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}).$$

Thus, by adding a node the upper Darboux sums decrease, while the lower Darboux sums increase.

Proof. The inequality (9.3.1a) shows that $\boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}')$. Suppose that the extra node x' is contained in (x_{k-1},x_k). We set

$$M_k := \sup_{x\in I_k} f(x), \;\; m_k := \inf_{x\in I_k} f(x).$$

Then

$$\boldsymbol{S}_*(f,\boldsymbol{P}') = \sum_{j<k} m_j\Delta x_j + \underbrace{\inf_{x\in[x_{k-1},x']} f(x)}_{\geqslant m_k}\,(x'-x_{k-1}) + \underbrace{\inf_{[x',x_k]} f(x)}_{\geqslant m_k}\,(x_k-x') + \sum_{\ell>k} m_\ell\Delta x_\ell$$

$$\geqslant \sum_{j<k} m_j\Delta x_j + \underbrace{m_k(x'-x_{k-1}) + m_k(x_k-x')}_{=m_k(x_k-x_{k-1})} + \sum_{\ell>k} m_\ell\Delta x_\ell$$

$$= \sum_{j<k} m_j\Delta x_j + m_k\Delta x_k + \sum_{\ell>k} m_\ell\Delta x_\ell = \sum_{i=1}^n m_i\Delta x_i = \boldsymbol{S}_*(f,\boldsymbol{P}).$$

The inequality

$$\boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P})$$

is proved in a similar fashion.

□

Definition 9.3.5. Given two partitions $\boldsymbol{P},\boldsymbol{P}'$ of $[a,b]$, we say that $\boldsymbol{P}'$ is a *refinement* of $\boldsymbol{P}$, and we write this $\boldsymbol{P}' \succ \boldsymbol{P}$, if $\boldsymbol{P}'$ is obtained from $\boldsymbol{P}$ by adding a few more nodes. □

Since the addition of nodes increases lower Darboux sums and decreases upper Darboux sums we deduce the following result.

Proposition 9.3.6. *Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function and $\boldsymbol{P},\boldsymbol{P}'$ are partitions of $[a,b]$. If $\boldsymbol{P}' \succ \boldsymbol{P}$, then*

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}).$$ □

Corollary 9.3.7. *Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function and $\boldsymbol{P},\boldsymbol{P}'$ are partitions of $[a,b]$. If $\boldsymbol{P}' \succ \boldsymbol{P}$,*

$$\omega(f,\boldsymbol{P}') \leqslant \omega(f,\boldsymbol{P}). \tag{9.3.2}$$

Proof. From (9.3.3) we deduce

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}),$$

so that,

$$\omega(f,\boldsymbol{P}') = \boldsymbol{S}^*(f,\boldsymbol{P}') - \boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}) - \boldsymbol{S}_*(f,\boldsymbol{P}) = \omega(f,\boldsymbol{P}).$$

□

Given two partitions $\boldsymbol{P},\boldsymbol{P}'$ of $[a,b]$ we denote by $\boldsymbol{P}\vee\boldsymbol{P}'$ the partition whose set of nodes is the union of the sets of nodes of the partitions $\boldsymbol{P}$ and $\boldsymbol{P}'$. Clearly $\boldsymbol{P}\vee\boldsymbol{P}'$ is a refinement of both $\boldsymbol{P}$ and $\boldsymbol{P}'$. From Proposition 9.3.6 we deduce the following important consequence.

Corollary 9.3.8. *Suppose that $f:[a,b]\to\mathbb{R}$ is a bounded function and $\boldsymbol{P}_0,\boldsymbol{P}_1$ are partitions of $[a,b]$. Then*

$$\boldsymbol{S}_*(f,\boldsymbol{P}_1) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}_0\vee\boldsymbol{P}_1) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}_0\vee\boldsymbol{P}_1) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}_0). \tag{9.3.3}$$

□

The above corollary shows that if $f:[a,b]\to\mathbb{R}$ is a bounded function, then the set

$$\big\{\boldsymbol{S}^*(f,\boldsymbol{P});\ \ \boldsymbol{P}\in\mathcal{P}_{[a,b]}\big\}$$

is bounded below. Indeed, if we denote by $\boldsymbol{U}_1$ the uniform partition of order 1 of $[a,b]$, then (9.3.3) shows that

$$\boldsymbol{S}_*(f,\boldsymbol{U}_1) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}),\ \ \forall\boldsymbol{P}\in\mathcal{P}_{[a,b]}.$$

We set

$$\boldsymbol{S}^*(f) := \inf\big\{\boldsymbol{S}^*(f,\boldsymbol{P});\ \ \boldsymbol{P}\in\mathcal{P}_{[a,b]}\big\}.$$

Similarly, the set

$$\big\{\boldsymbol{S}_*(f,\boldsymbol{P});\ \ \boldsymbol{P}\in\mathcal{P}_{[a,b]}\big\}$$

is bounded above and we define

$$\boldsymbol{S}_*(f) := \sup\big\{\boldsymbol{S}_*(f,\boldsymbol{P});\ \ \boldsymbol{P}\in\mathcal{P}_{[a,b]}\big\}.$$

Proposition 9.3.9. *If $f:[a,b]\to\mathbb{R}$ is a bounded function, then*

$$\boldsymbol{S}_*(f) \leqslant \boldsymbol{S}^*(f). \tag{9.3.4}$$

Proof. From (9.3.3) we deduce that $\forall\boldsymbol{P}_0,\boldsymbol{P}_1\in\mathcal{P}_{[a,b]}$ we have

$$\boldsymbol{S}_*(f,\boldsymbol{P}_1) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}_0) \Rightarrow \boldsymbol{S}_*(f,\boldsymbol{P}_1) \leqslant \inf_{\boldsymbol{P}_0}\boldsymbol{S}^*(f,\boldsymbol{P}_0) = \boldsymbol{S}^*(f)$$

$$\Rightarrow \boldsymbol{S}_*(f) = \sup_{\boldsymbol{P}_1}\boldsymbol{S}_*(f,\boldsymbol{P}_1) \leqslant \boldsymbol{S}^*(f).$$

□

Definition 9.3.10. Let $f : [a,b] \to \mathbb{R}$ be a *bounded* function.

(a) The numbers $\boldsymbol{S}_*(f)$ and respectively $\boldsymbol{S}^*(f)$ are called the *lower* and respectively *upper Darboux integrals* of f.

(b) The function f is called *Darboux integrable* if $\boldsymbol{S}_*(f) = \boldsymbol{S}^*(f)$. □

Theorem 9.3.11 (Riemann-Darboux)**.** *Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function. Then the following statements are equivalent.*

(i) *The function f is Riemann integrable.*

(ii) *The function f is Darboux integrable, i.e., $\boldsymbol{S}_*(f) = \boldsymbol{S}^*(f)$.*

(iii) $\inf_{\boldsymbol{P}} \omega(f, \boldsymbol{P}) = 0$, *i.e.,*

$$\forall \varepsilon > 0,\ \ \exists \boldsymbol{P}_\varepsilon \in \mathcal{P}_{[a,b]} :\ \ \omega(f, \boldsymbol{P}_\varepsilon) < \varepsilon. \tag{$\boldsymbol{\omega}_0$}$$

(iv) $\lim_{\|\boldsymbol{P}\| \to 0} \omega(f, \boldsymbol{P}) = 0$, *i.e.,*

$$\forall \varepsilon > 0\ \ \exists \delta = \delta(\varepsilon) > 0\ \ \forall \boldsymbol{P} \in \mathcal{P}_{[a,b]} :\ \ \|\boldsymbol{P}\| < \delta \Rightarrow \omega(f, \boldsymbol{P}) < \varepsilon. \tag{$\boldsymbol{\omega}$}$$

Proof. We will prove these equivalences using the following logical successions

$$(iii) \Longleftrightarrow (ii),\ \ (iv)\ \Rightarrow\ (iii),\ \ (iv) \Longleftrightarrow (i),\ \ (iii)\ \Rightarrow\ (iv).$$

(iii) ⇒ (ii). For any $\varepsilon > 0$ we can find a partition $\boldsymbol{P}_\varepsilon$ such that $\omega(f, \boldsymbol{P}_\varepsilon) < \varepsilon$. Now observe that

$$\boldsymbol{S}_*(f, \boldsymbol{P}_\varepsilon) \leqslant \boldsymbol{S}_*(f) \leqslant \boldsymbol{S}^*(f) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}_\varepsilon),$$

and

$$\boldsymbol{S}^*(f, \boldsymbol{P}_\varepsilon) - \boldsymbol{S}_*(f, \boldsymbol{P}_\varepsilon) = \omega(f, \boldsymbol{P}_\varepsilon) < \varepsilon.$$

Hence

$$0 \leqslant \boldsymbol{S}^*(f) - \boldsymbol{S}_*(f) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}_\varepsilon) - \boldsymbol{S}_*(f, \boldsymbol{P}_\varepsilon) < \varepsilon,\ \ \forall \varepsilon > 0,$$

so that

$$\boldsymbol{S}_*(f) = \boldsymbol{S}^*(f).$$

(ii) ⇒ (iii). We know that $\boldsymbol{S}_*(f) = \boldsymbol{S}^*(f)$. Denote by $\boldsymbol{S}(f)$ this common value. Since

$$\boldsymbol{S}(f) = \boldsymbol{S}_*(f) = \sup_{\boldsymbol{P}} \boldsymbol{S}_*(f, \boldsymbol{P}),$$

we deduce that for any $\varepsilon > 0$ there exists a partition P_ε^- such that

$$\boldsymbol{S}(f) - \frac{\varepsilon}{2} < \boldsymbol{S}_*(f, \boldsymbol{P}_\varepsilon^-) \leqslant \boldsymbol{S}(f).$$

Since

$$\boldsymbol{S}(f) = \boldsymbol{S}^*(f) = \inf_{\boldsymbol{P}} \boldsymbol{S}^*(f, \boldsymbol{P}),$$

we deduce that for any $\varepsilon > 0$ there exists a partition P_ε^+ such that

$$\boldsymbol{S}(f) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}_\varepsilon^+) < \boldsymbol{S}(f) + \frac{\varepsilon}{2}.$$

Hence

$$\boldsymbol{S}(f)-\frac{\varepsilon}{2}<\boldsymbol{S}_*(f,\boldsymbol{P}_\varepsilon^-)\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}_\varepsilon^+)<\boldsymbol{S}(f)+\frac{\varepsilon}{2}.$$

Now set $\boldsymbol{P}_\varepsilon := \boldsymbol{P}_\varepsilon^- \vee \boldsymbol{P}_\varepsilon^+$. We deduce from (9.3.3) that

$$\boldsymbol{S}(f)-\frac{\varepsilon}{2}<\boldsymbol{S}_*(f,\boldsymbol{P}_\varepsilon^-)\leqslant\boldsymbol{S}_*(f,\boldsymbol{P}_\varepsilon)\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}_\varepsilon)\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}_\varepsilon^+)<\boldsymbol{S}(f)+\frac{\varepsilon}{2}.$$

This proves that

$$\omega(f,\boldsymbol{P}_\varepsilon)=\boldsymbol{S}^*(f,\boldsymbol{P}_\varepsilon)-\boldsymbol{S}_*(f,\boldsymbol{P}_\varepsilon)<\varepsilon.$$

(iv) ⇒ (iii). This is obvious.

(iv) ⇒ (i). From the above we deduce that (iv) ⇒ (ii) ∧ (iii) so $\boldsymbol{S}_*(f)=\boldsymbol{S}^*(f)$. We set

$$\boldsymbol{S}(f):=\boldsymbol{S}_*(f)=\boldsymbol{S}^*(f).$$

We will show that f is integrable and its Riemann integral is $\boldsymbol{S}(f)$.

Fix $\varepsilon>0$. According to ($\boldsymbol{\omega}$), there exists $\delta=\delta(\varepsilon)>0$ such that for any partition $\boldsymbol{P}$ of $[a,b]$ satisfying $\|\boldsymbol{P}\|<\delta$ we have

$$\omega(f,\boldsymbol{P})<\varepsilon.$$

Given a partition $\boldsymbol{P}$ such that $\|\boldsymbol{P}\|<\delta$ and $\underline{\xi}$ a sample of $\boldsymbol{P}$ we have

$$\boldsymbol{S}_*(f,\boldsymbol{P})\leqslant\boldsymbol{S}(f)\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}),$$
$$\boldsymbol{S}_*(f,\boldsymbol{P})\leqslant\boldsymbol{S}(f,\boldsymbol{P},\xi)\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}).$$

Thus both numbers $\boldsymbol{S}(f)$ and $\boldsymbol{S}(f,\boldsymbol{P},\xi)$ lie in the interval $[\boldsymbol{S}_*(f,\boldsymbol{P}),\boldsymbol{S}^*(f,\boldsymbol{P})]$ of length $\omega(f,\boldsymbol{P})<\varepsilon$. Hence

$$\big|\,\boldsymbol{S}(f,\boldsymbol{P},\xi)-\boldsymbol{S}(f)\,\big|<\varepsilon,\ \ \forall\|P\|<\delta(\varepsilon),\ \ \forall\underline{\xi}\in\mathcal{S}(\boldsymbol{P}).$$

This proves that f is Riemann integrable.

(i) ⇒ (iv). We have to prove that if f is Riemann integrable, then f satisfies ($\boldsymbol{\omega}$). We first need an auxiliary result.

Lemma 9.3.12. *Suppose that $f:[a,b]\to\mathbb{R}$ is a bounded function. Then, for any partition $\boldsymbol{P}$ of $[a,b]$ we have*

$$\boldsymbol{S}_*(f,\boldsymbol{P})=\inf_{\underline{\xi}\in\mathcal{S}(\boldsymbol{P})}\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}),$$

$$\boldsymbol{S}^*(f,\boldsymbol{P})=\sup_{\underline{\xi}\in\mathcal{S}(\boldsymbol{P})}\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}).$$

In other words, for any $\varepsilon>0$, and any partition $\boldsymbol{P}$ of $[a,b]$, there exist samples $\underline{\xi}'$ and $\underline{\xi}''$ of $\boldsymbol{P}$ such that

$$\boldsymbol{S}_*(f,P)\leqslant\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}')<\boldsymbol{S}_*(f,P)+\varepsilon,$$
$$\boldsymbol{S}^*(f,\boldsymbol{P})-\varepsilon<\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'')\leqslant\boldsymbol{S}^*(f,\boldsymbol{P}).$$

In particular

$$\omega(f,\boldsymbol{P}) = \boldsymbol{S}^*(f,\boldsymbol{P}) - \boldsymbol{S}_*(f,\boldsymbol{P}) = \sup_{\underline{\xi}\in\mathcal{S}(\boldsymbol{P})} \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}) - \inf_{\underline{\xi}\in\mathcal{S}(\boldsymbol{P})} \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}). \tag{9.3.5}$$

Proof. We prove only the statement involving lower sums. The proof of the statement involving upper sums is similar. Denote by n the order of $\boldsymbol{P}$ and by I_k the k-th interval of $\boldsymbol{P}$ and, as usual, we set

$$m_k = \inf_{x\in I_k} f(x).$$

In particular, there exists $\xi_k' \in I_k$ such that

$$m_k \leqslant f(\xi_k') < m_k + \frac{\varepsilon}{b-a}.$$

The collection $\underline{\xi}' = (\xi_k')_{1\leqslant k\leqslant n}$ is a sample of $\boldsymbol{P}$ satisfying

$$m_k(x_k - x_{k-1}) \leqslant f(\xi_k')(x_k - x_{k-1}) < m_k(x_k - x_{k-1}) + \frac{\varepsilon}{b-a}(x_k - x_{k-1}).$$

Hence

$$\boldsymbol{S}_*(f,\boldsymbol{P}) = \sum_{k=1}^n m_k(x_k - x_{k-1}) \leqslant \underbrace{\sum_{k=1}^n f(\xi_k')(x_k - x_{k-1})}_{=\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}')}$$

$$< \underbrace{\sum_{k=1}^n m_k(x_k - x_{k-1})}_{=\boldsymbol{S}_*(f,\boldsymbol{P})} + \frac{\varepsilon}{b-a}\underbrace{\sum_{k=1}^n (x_k - x_{k-1})}_{=(b-a)} = \boldsymbol{S}_*(f,P) + \varepsilon.$$

□

We can now complete the proof of ($\boldsymbol{\omega}$). Since f is Riemann integrable, there exists $S_f \in \mathbb{R}$ such that, for any $\varepsilon > 0$ we can find $\delta = \delta(\varepsilon) > 0$ with the property that for any partition $\boldsymbol{P}$ with mesh size $\|\boldsymbol{P}\| < \delta$ and any sample $\underline{\xi}$ of $\boldsymbol{P}$ we have

$$\big|\, S_f - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}) \,\big| < \frac{\varepsilon}{4}. \tag{9.3.6}$$

According to Lemma 9.3.12 we can find samples $\underline{\xi}'$ and $\underline{\xi}''$ such that

$$\big|\, \boldsymbol{S}_*(f,\boldsymbol{P}) - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}') \,\big|,\ \ \big|\, \boldsymbol{S}^*(f,\boldsymbol{P}) - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'') \,\big| < \frac{\varepsilon}{4}. \tag{9.3.7}$$

If $\|\boldsymbol{P}\| < \delta$, then

$$\omega(f,\boldsymbol{P}) = \big|\boldsymbol{S}^*(f,\boldsymbol{P}) - \boldsymbol{S}_*(f,\boldsymbol{P})\big|$$

$$\leqslant \big|\, \boldsymbol{S}_*(f,\boldsymbol{P}) - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}') \,\big| + \big|\, \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}') - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'') \,\big| + \big|\, \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'') - \boldsymbol{S}^*(f,\boldsymbol{P}) \,\big|$$

$$\overset{(9.3.7)}{<} \frac{\varepsilon}{4} + \big|\, \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}') - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'') \,\big| + \frac{\varepsilon}{4}$$

$$\leqslant \frac{\varepsilon}{2} + \big|\, \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}') - S_f \,\big| + \big|\, S_f - \boldsymbol{S}(f,\boldsymbol{P},\underline{\xi}'') \,\big| \overset{(9.3.6)}{<} \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.$$

(iii) ⇒ (iv). We have to show that if f satisfies ($\boldsymbol{\omega}_0$), then it also satisfies ($\boldsymbol{\omega}$). We need the following auxiliary result.

Lemma 9.3.13. *Suppose that $\boldsymbol{P}_0 = \{a = z_0 < z_1 < \cdots < z_{n_0} = b\}$ is a partition of $[a,b]$ of order n_0. Denote by λ_0 the length of the shortest intervals of the partition $\boldsymbol{P}_0$, i.e.,*

$$\lambda_0 := \min_{1 \leqslant j \leqslant n_0} (z_j - z_{j-1}).$$

For any partition $\boldsymbol{P}$ such that $\|\boldsymbol{P}\| < \lambda_0$ we have

$$\omega(f, \boldsymbol{P}) \leqslant (n_0 - 1)\|\boldsymbol{P}\| \operatorname{osc}(f, [a,b]) + \omega(f, \boldsymbol{P}_0). \tag{9.3.8}$$

Proof. Denote by $I_1, \dots, I_{n_0}$ the intervals of $\boldsymbol{P}_0$. Denote by n the order of $\boldsymbol{P}$, and by $J_1, \dots, J_n$ the intervals of $\boldsymbol{P}$. We will denote by $\ell(J_k)$ the length of J_k and by $\ell(I_j)$ the length of I_j

Since $\ell(J_k) \leqslant \ell(I_j)$, $\forall j = 1, \dots, n_0$, $k = 1, \dots, n$ we deduce that the intervals J_k of $\boldsymbol{P}$ are of only the following two types.

Type 1. The interval J_k is contained in an interval I_j of $\boldsymbol{P}_0$.

Type 2. The interval J_k contains in the *interior* a node $z_{j(k)}$ of $\boldsymbol{P}_0$.

We denote by $\mathcal{J}^1$ the collection of Type 1 intervals of $\boldsymbol{P}$, and by $\mathcal{J}^2$ the collection of Type 2 intervals of $\boldsymbol{P}$. We remark that $\mathcal{J}^2$ could be empty. Moreover, for any node z_j of $\boldsymbol{P}_0$ there exists at most one Type 2 interval of $\boldsymbol{P}$ that contains z_j in the interior. Thus $\mathcal{J}^2$ consist of at most $n_0 - 1$ intervals, i.e., its cardinality $|\mathcal{J}^2|$ satisfies

$$|\mathcal{J}^2| \leqslant n_0 - 1.$$

We have

$$\omega(f, \boldsymbol{P}) = \sum_{k=1}^{n} \operatorname{osc}(f, J_k)\ell(J_k) = \underbrace{\sum_{j_k \in \mathcal{J}^1} \operatorname{osc}(f, J_k)\ell(J_k)}_{=:S_1} + \underbrace{\sum_{J_k \in \mathcal{J}^2} \operatorname{osc}(f, J_k)\ell(J_k)}_{=:S_2}.$$

We now estimate S_1 from above

$$S_1 = \sum_{j=1}^{n_0} \left(\sum_{J_k \subset I_j} \operatorname{osc}(f, J_k)\ell(J_k) \right)$$

($\operatorname{osc}(f, J_k) \leqslant \operatorname{osc}(f, I_j)$ whenever $J_k \subset I_j$)

$$\leqslant \sum_{j=1}^{n_0} \left(\sum_{J_k \subset I_j} \operatorname{osc}(f, I_j)\ell(J_k) \right) = \sum_{j=1}^{n_0} \operatorname{osc}(f, I_j) \underbrace{\left(\sum_{J_k \subset I_j} \ell(J_k) \right)}_{\leqslant \ell(I_j)} \leqslant \sum_{j=1}^{n_0} \operatorname{osc}(f, I_j)\ell(I_j) = \omega(f, \boldsymbol{P}_0).$$

Now observe that if J_k is a Type 2 interval of $\boldsymbol{P}$, then $\ell(J_k) \leqslant \|\boldsymbol{P}\|$ and $\operatorname{osc}(f, J_k) \leqslant \operatorname{osc}(f, [a,b])$. Hence

$$S_2 \leqslant \sum_{J_k \in \mathcal{J}^2} \operatorname{osc}(f, [a,b])\|\boldsymbol{P}\| \leqslant |\mathcal{J}^2| \operatorname{osc}(f, [a,b])\|\boldsymbol{P}\| \leqslant (n_0 - 1) \operatorname{osc}(f, [a,b])\|\boldsymbol{P}\|.$$

Hence

$$\omega(f, \boldsymbol{P}) = S_1 + S_2 \leqslant (n_0 - 1) \operatorname{osc}(f, [a,b])\|\boldsymbol{P}\| + \omega(f, \boldsymbol{P}_0).$$

□

Returning to our implication ($\boldsymbol{\omega}_0$) $\Rightarrow$ ($\boldsymbol{\omega}$), we observe that ($\boldsymbol{\omega}_0$) implies that for any $\varepsilon > 0$ there exists a partition $\boldsymbol{P}_\varepsilon$ such that

$$\omega(f, \boldsymbol{P}_\varepsilon) < \frac{\varepsilon}{2}.$$

Denote by n_ε the order of $\boldsymbol{P}_\varepsilon$ and by $x_0 < x_1 < \cdots < x_{n_\varepsilon}$ the nodes of $\boldsymbol{P}_\varepsilon$. We set

$$\lambda_\varepsilon := \min_{1 \leqslant j \leqslant n_\varepsilon} (x_j - x_{j-1}).$$

Now choose $\delta = \delta(\varepsilon) > 0$ such that

$$\delta < \lambda_\varepsilon \;\text{ and }\; (n_\varepsilon - 1)\operatorname{osc}(f, [a,b])\delta < \frac{\varepsilon}{2} \iff \delta < \min\left(\lambda_\varepsilon, \;\; \frac{\varepsilon}{2(n_\varepsilon - 1)\operatorname{osc}(f, [a,b])} \right).$$

If $\boldsymbol{P}$ is an arbitrary partition of $[a,b]$ such that $\|\boldsymbol{P}\| < \delta(\varepsilon)$, then Lemma 9.3.13 implies that

$$\omega(f, \boldsymbol{P}) \leqslant (n_\varepsilon - 1)\operatorname{osc}(f, [a,b])\delta + \omega(f, \boldsymbol{P}_\varepsilon) < \varepsilon.$$

This proves that f satisfies ($\boldsymbol{\omega}$) and completes the proof of the Riemann-Darboux Theorem.

□

We record here for later use a direct consequence of the above proof.

Corollary 9.3.14. *Suppose that $f : [a,b] \to \mathbb{R}$ is a Riemann integrable function. Then*

$$\int_a^b f(x)dx = \boldsymbol{S}_*(f) = \boldsymbol{S}^*(f). \tag{9.3.9}$$

In particular,

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \int_a^b f(x)dx \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}'), \;\; \forall \boldsymbol{P}, \boldsymbol{P}' \in \mathcal{P}_{[a,b]}. \tag{9.3.10}$$

□

9.4. Examples of Riemann integrable functions

We are now going to collect the reward for the effort we spent proving the Riemann-Darboux theorem.

Proposition 9.4.1. *Any continuous function $f : [a,b] \to \mathbb{R}$ is Riemann integrable.*

Proof. We will use the Riemann-Darboux theorem to prove the claim. Note first that the Weierstrass Theorem 6.2.4 shows that f is bounded.

To prove that f satisfies ($\boldsymbol{\omega}$) we rely on the Uniform Continuity Theorem 6.3.5. According to this theorem, for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for any interval $I \subset [a,b]$ of length $< \delta$ we have

$$\operatorname{osc}(f, I) < \frac{\varepsilon}{b-a}.$$

If $\boldsymbol{P}$ is any partition of $[a,b]$ of order n and mesh size $\|\boldsymbol{P}\| < \delta(\varepsilon)$, then for any interval I_k of $\boldsymbol{P}$ we have

$$\operatorname{osc}(f, I_k) < \frac{\varepsilon}{b-a}.$$

Hence

$$\omega(f,\boldsymbol{P}) = \sum_{k=1}^{n} \operatorname{osc}(f,I_k)\Delta x_k < \frac{\varepsilon}{b-a} \underbrace{\sum_{k=1}^{n} \Delta x_k}_{=(b-a)} = \varepsilon.$$

This shows that f satisfies ($\boldsymbol{\omega}$) and thus it is Riemann integrable. □

Example 9.4.2. The function $f:[0,1]\to\mathbb{R}$, $f(x)=x^2$ is continuous and thus integrable. Thus

$$\int_0^1 x^2 dx = \lim_{N\to\infty} \boldsymbol{S}_*(f,\boldsymbol{U}_N),$$

where $\boldsymbol{U}_N$ denote the uniform partition of order N of $[0,1]$. Since f is nondecreasing we deduce that $\boldsymbol{S}_*(f,\boldsymbol{U}_N)$ coincides with the sum L_N defined in (9.1.1). As explained in Section 9.1 the sum L_N converges to $\frac{1}{3}$ as $N\to\infty$. □

Proposition 9.4.3. *Any nondecreasing function $f:[a,b]\to\mathbb{R}$ is Riemann integrable.*

Proof. Clearly f is bounded since $f(a)\leqslant f(x)\leqslant f(b)$, $\forall x\in[a,b]$. If $\boldsymbol{P}$ is any partition of $[a,b]$ of order n, then for an interval $I_k=[x_{k-1},x_k]$ of this partition we have

$$\operatorname{osc}(f,I_k) = f(x_k)-f(x_{k-1}),$$

$$\operatorname{osc}(f,I_k)\Delta x_k \leqslant \operatorname{osc}(f,I_k)\|\boldsymbol{P}\| = \|\boldsymbol{P}\|\big(f(x_k)-f(x_{k-1})\big)$$

so that

$$\omega(f,\boldsymbol{P}) = \sum_{k=1}^{n}\operatorname{osc}(f,I_k)\Delta x_k \leqslant \|\boldsymbol{P}\|\sum_{k=1}^{n}\big(f(x_k)-f(x_{k-1})\big) = \|\boldsymbol{P}\|\big(f(b)-f(a)\big).$$

This shows that f satisfies ($\boldsymbol{\omega}$) since

$$\lim_{\|\boldsymbol{P}\|\to 0}\|\boldsymbol{P}\|\big(f(b)-f(a)\big)=0.$$

□

Proposition 9.4.4. *Suppose that $f:[a,b]\to\mathbb{R}$ is a bounded function which is continuous on (a,b). Then f is Riemann integrable.*

Proof. We will prove that f satisfies ($\boldsymbol{\omega}_0$). Fix $\varepsilon>0$ and choose a positive real number $d(\varepsilon)$ such that

$$\operatorname{osc}(f,[a,b])d(\varepsilon) < \frac{\varepsilon}{4}. \tag{9.4.1}$$

Denote by J_ε the compact interval $J_\varepsilon := [a+d(\varepsilon), b-d(\varepsilon)]$; see Figure 9.4.

The restriction of f to J_ε is continuous. The Uniform Continuity Theorem 6.3.5 implies that there exists $\delta = \delta(\varepsilon) < d(\varepsilon)$ with the property that for any interval $I \subset J_\varepsilon$ of length $\ell(I) < \delta(\varepsilon)$ we have

$$\operatorname{osc}(f, I) < \frac{\varepsilon}{2(b-a)}. \tag{9.4.2}$$

Consider a partition $\boldsymbol{P}_\varepsilon$ of order n of J_ε satisfying $\|\boldsymbol{P}\| < \delta(\varepsilon)$. We denote by I_k, $k = 1\dots, n$, the intervals of $\boldsymbol{P}_\varepsilon$; see Figure 9.4. We set

$$I_* := [a, a + d(\varepsilon)], \;\; I^* = [b - d(\varepsilon), b].$$

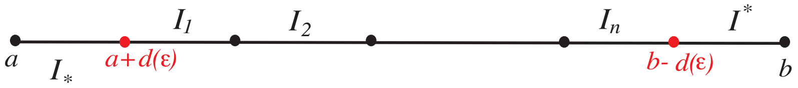

Figure 9.4. *Isolating the possible points of discontinuity of f.*

The collection of intervals

$$I_*, I_1, \dots, I_n, I^*$$

defines a partition $\widehat{\boldsymbol{P}}_\varepsilon$ of $[a, b]$; see Figure 9.4. We have

$$\omega(f, \widehat{\boldsymbol{P}}_\varepsilon) = \underbrace{\operatorname{osc}(f, I_*)\ell(I_*)}_{=:T_*} + \underbrace{\sum_{k=1}^{n} \operatorname{osc}(f, I_k)\ell(I_k)}_{=:T} + \underbrace{\operatorname{osc}(f, I^*)\ell(I^*)}_{=:T^*}.$$

Note that

$$\ell(I_*) = \ell(I^*) = d(\varepsilon).$$

so that

$$T_* = \operatorname{osc}(f, I_*)d(\varepsilon) \leqslant \operatorname{osc}(f, [a, b])d(\varepsilon) \overset{(9.4.1)}{<} \frac{\varepsilon}{4},$$

$$T^* = \operatorname{osc}(f, I^*)d(\varepsilon) \leqslant \operatorname{osc}(f, [a, b])d(\varepsilon) \overset{(9.4.1)}{<} \frac{\varepsilon}{4}.$$

Moreover,

$$T = \sum_{k=1}^{n} \operatorname{osc}(f, I_k)\ell(I_k) \overset{(9.4.2)}{<} \frac{\varepsilon}{2(b-a)} \sum_{k=1}^{n} \ell(I_k) = \frac{\varepsilon}{2(b-a)}(b-a) = \frac{\varepsilon}{2}.$$

Hence,

$$\omega(f, \widehat{\boldsymbol{P}}_\varepsilon) = T_* + T + T^* < \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon.$$

This proves that f satisfies ($\boldsymbol{\omega}_0$) and thus it is Riemann integrable. □

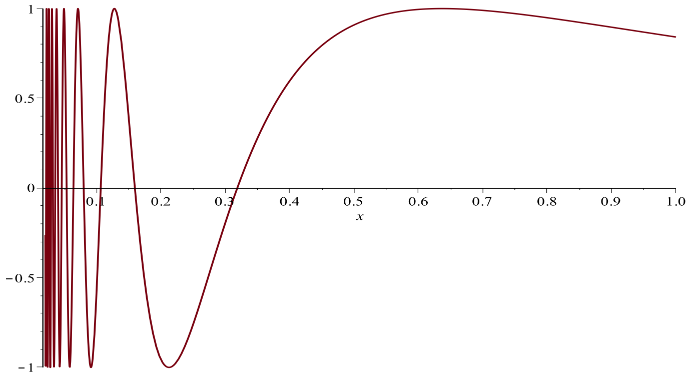

Figure 9.5. *A wildly oscillating, yet Riemann integrable function.*

Remark 9.4.5. Proposition 9.4.4 has some surprising nontrivial consequences. For example, it shows that the wildly oscillating function (see Figure 9.5)

$$f : [0,1] \to \mathbb{R}, \quad f(x) = \begin{cases} \sin\left(\frac{1}{x}\right), & x \in (0,1], \\ 0, & x = 0, \end{cases}$$

is Riemann integrable. □

Proposition 9.4.6. *Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function and $c \in (a,b)$. The following statements are equivalent.*

(i) *The function f is Riemann integrable on $[a,b]$.*

(ii) *The restrictions of $f|_{[a,c]}$ and $f|_{[c,b]}$ of f to $[a,c]$ and $[c,b]$ are Riemann integrable functions.*

Moreover, if f satisfies either one of the two equivalent conditions above, then

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx. \tag{9.4.3}$$

Proof. (i) ⇒ (ii). Suppose that f is Riemann integrable on $[a,b]$. Given a partition $\boldsymbol{P}'$ of $[a,c]$ and a partition $\boldsymbol{P}''$ of $[c,b]$ we obtain a partition $\boldsymbol{P}' * \boldsymbol{P}''$ of $[a,b]$ whose set of nodes is the union of the sets of nodes of $\boldsymbol{P}'$ and $\boldsymbol{P}''$. Note that

$$\|\boldsymbol{P}' * \boldsymbol{P}''\| \leqslant \max\{\, \|\boldsymbol{P}'\|,\ \|\boldsymbol{P}''\|\,\},$$

and

$$\omega(\, f, \boldsymbol{P}' * \boldsymbol{P}''\,) = \omega(\, f|_{[a,c]}, \boldsymbol{P}'\,) + \omega\big(\, f|_{[c,b]}, \boldsymbol{P}'\,\big).$$

Since f is Riemann integrable on $[a,b]$, it satisfies the property ($\boldsymbol{\omega}$) so, for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that, for any partition $\boldsymbol{P}$ of $[a,b]$ with mesh size $\|\boldsymbol{P}\| < \delta(\varepsilon)$, we have

$$\omega(f, \boldsymbol{P}) < \varepsilon.$$

If the partitions $\boldsymbol{P}'$ and $\boldsymbol{P}''$ satisfy

$$\max\big\{ \|\boldsymbol{P}'\|,\ \|\boldsymbol{P}''\| \big\} < \delta(\varepsilon),$$

then $\|\boldsymbol{P}' * \boldsymbol{P}''\| < \delta(\varepsilon)$ so that

$$\omega(f|_{[a,c]}, \boldsymbol{P}') + \omega(f|_{[c,b]}, \boldsymbol{P}'') = \omega(f, \boldsymbol{P}' * \boldsymbol{P}'') < \varepsilon.$$

This shows that both restrictions $f|_{[a,c]}$ and $f|_{[c,b]}$ satisfy ($\boldsymbol{\omega}$) and thus are Riemann integrable.

(ii) $\Rightarrow$ (i). We will prove that if $f|_{[a,c]}$ and $f|_{[c,b]}$ are Riemann integrable, then f is integrable on $[a, b]$. We invoke Theorem 9.3.11. It suffices to show that f satisfies ($\boldsymbol{\omega}_0$). Fix $\varepsilon > 0$. We have to prove that there exists a partition $\boldsymbol{P}_\varepsilon$ of $[a, b]$ such that $\omega(f, \boldsymbol{P}_\varepsilon) < \varepsilon$.

Since $f|_{[a,c]}$ and $f|_{[c,b]}$ are Riemann integrable, they satisfy ($\boldsymbol{\omega}_0$), and we deduce that there exist partitions $\boldsymbol{P}'_\varepsilon$ of $[a, c]$, and $\boldsymbol{P}''_\varepsilon$ of $[c, b]$ such that

$$\omega(f, \boldsymbol{P}'_\varepsilon),\ \ \omega(f, \boldsymbol{P}''_\varepsilon) < \frac{\varepsilon}{2}.$$

Then $\boldsymbol{P}_\varepsilon = \boldsymbol{P}'_\varepsilon * \boldsymbol{P}''_\varepsilon$ is a partition of $[a, b]$, and

$$\omega(f, \boldsymbol{P}_\varepsilon) = \omega(f, \boldsymbol{P}'_\varepsilon) + \omega(f, \boldsymbol{P}''_\varepsilon) < \varepsilon.$$

To prove (9.4.3) assume that f satisfies both (i) and (ii). Denote by $\boldsymbol{U}'_n$ the uniform partition of order n of $[a, c]$ and by $\boldsymbol{U}''_n$ the uniform partition of order n of $[c, b]$. Set

$$\boldsymbol{P}_n := \boldsymbol{U}'_n * \boldsymbol{U}''_n.$$

Note that

$$\|\boldsymbol{P}_n\| = \max\big(\|\boldsymbol{U}'_n\|,\ \|\boldsymbol{U}''_n\| \big) \to 0 \ \text{ as } n \to \infty. \tag{9.4.4}$$

Denote by $\underline{\xi}'_n$ the midpoint sample of $\boldsymbol{U}'_n$, and by $\underline{\xi}''_n$ the midpoint sample of $\boldsymbol{U}''_n$. Then $\underline{\xi}_n := \underline{\xi}_n \cup \underline{\xi}''_n$ is the midpoint sample of $\boldsymbol{P}_n$. We have

$$\boldsymbol{S}(f, \boldsymbol{P}_n, \underline{\xi}_n) = \boldsymbol{S}(f, \boldsymbol{U}'_n, \underline{\xi}'_n) + \boldsymbol{S}(f, \boldsymbol{U}''_n, \underline{\xi}''_n). \tag{9.4.5}$$

From (i), (9.4.4), and (9.2.2) we deduce that

$$\lim_{n\to\infty} \boldsymbol{S}(f, \boldsymbol{P}_n, \underline{\xi}_n) = \int_a^b f(x)dx.$$

From (ii), (9.4.4), and (9.2.2) we deduce that

$$\lim_{n\to\infty} \boldsymbol{S}(f, \boldsymbol{U}'_n, \underline{\xi}'_n) = \int_a^c f(x)dx,$$

$$\lim_{n\to\infty} \boldsymbol{S}(f, \boldsymbol{U}''_n, \underline{\xi}''_n) = \int_c^b f(x)dx.$$

The equality (9.4.3) now follows from the above three equalities after letting $n \to \infty$ in (9.4.5). □

Applying Proposition 9.4.6 iteratively we deduce the following consequence.

Corollary 9.4.7. *Suppose that $f : [a, b] \to \mathbb{R}$ is a bounded function and*

$$\boldsymbol{P} = (a = x_0 < x_1 < \cdots < x_n = b)$$

is a partition of $[a, b]$. Then the following statements are equivalent.

(i) *The function f is Riemann integrable on $[a, b]$.*

(ii) *For any $k = 1, \dots, n$ the restriction of f to $[x_{k-1}, x_k]$ is Riemann integrable.*

Moreover, if any of the above two equivalent conditions is satisfied, then

$$\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^b f(x)dx. \tag{9.4.6}$$

□

Corollary 9.4.8. *If $f : [a, b] \to \mathbb{R}$ is a bounded function and $D \subset [a, b]$ is a finite set such that f is continuous at any point in $[a, b] \setminus D$, then f is Riemann integrable.*

Proof. We add to D the endpoints a, b if they are not contained in D and we obtain a partition $\boldsymbol{P}$ of $[a, b]$ such that f is continuous in the *interior* of any interval $[x_{k-1}, x_k]$ of $\boldsymbol{P}$. Proposition 9.4.4 implies that f is Riemann integrable on each of the intervals $[x_{k-1}, x_k]$ and Corollary 9.4.7 implies that f is integrable on $[a, b]$. □

Proposition 9.4.9. *If $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable, then for any constants $\alpha, \beta \in \mathbb{R}$ the sum $\alpha f + \beta g : [a, b] \to \mathbb{R}$ is also Riemann integrable and*

$$\int_a^b \big(\alpha f(x) + \beta g(x)\big)dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx. \tag{9.4.7}$$

Proof. We will show that $\alpha f + \beta g$ satisfies the definition of Riemann integrability, Definition 9.2.4. Observe first that if $(\boldsymbol{P}, \underline{\xi})$ is a sampled partition of $[a, b]$, then

$$\boldsymbol{S}\big(\alpha f + \beta g, \boldsymbol{P}, \xi\big) = \alpha \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) + \beta \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}). \tag{9.4.8}$$

Indeed, if the partition $\boldsymbol{P}$ is

$$\boldsymbol{P} = \{a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b\},$$

and the sample $\underline{\xi}$ is $\underline{\xi} = (\xi_k)_{1 \leqslant k \leqslant n}$, then

$$\boldsymbol{S}\big(\alpha f + \beta g, \boldsymbol{P}, \xi\big) = \sum_k \big(\alpha f(\xi_k) + \beta g(\xi_k)\big)\Delta x_k = \sum_k \alpha f(\xi_k)\Delta x_k + \sum_k \beta g(\xi_k)\Delta x_k$$

$$= \alpha \sum_k f(\xi_k)\Delta x_k + \beta \sum_k g(\xi_k)\Delta x_k = \alpha \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) + \beta \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}).$$

Set

$$K := (|\alpha| + |\beta| + 1).$$

Fix $\varepsilon > 0$. Since f is Riemann integrable, there exists $\delta_1 = \delta_1(\varepsilon) > 0$ such that, $\forall \boldsymbol{P} \in \mathcal{P}_{[a,b]}$, $\forall \underline{\xi} \in \mathcal{S}(\boldsymbol{P})$ we have

$$\|\boldsymbol{P}\| < \delta_1 \Rightarrow \left| \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) - \int_a^b f(x)dx \right| < \frac{\varepsilon}{K}. \tag{9.4.9}$$

Since g is Riemann integrable, there exists $\delta_2 = \delta_2(\varepsilon) > 0$ such that, $\forall \boldsymbol{P} \in \mathcal{P}_{[a,b]}$, $\forall \underline{\xi} \in \mathcal{S}(\boldsymbol{P})$ we have

$$\|\boldsymbol{P}\| < \delta_2 \Rightarrow \left| \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}) - \int_a^b g(x)dx \right| < \frac{\varepsilon}{K}. \tag{9.4.10}$$

Set

$$\delta = \delta(\varepsilon) := \min\big(\,\delta_1(\varepsilon), \delta_2(\varepsilon)\,\big), \;\; S := \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx.$$

Let $\boldsymbol{P} \in \mathcal{P}_{[a,b]}$ be an arbitrary partition such that $\|\boldsymbol{P}\| < \delta$. Then for any sample $\underline{\xi} \in \mathcal{S}(\boldsymbol{P})$ we have

$$|\boldsymbol{S}(\alpha f + \beta g, \boldsymbol{P}, \xi) - S| \overset{(9.4.8)}{=} \left| \alpha\Big(\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) - \int_a^b f(x)dx \Big) + \beta\Big(\boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}) - \int_a^b g(x)dx \Big) \right|$$

$$\leqslant |\alpha| \cdot \left| \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) - \int_a^b f(x)dx \right| + |\beta| \cdot \left| \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}) - \int_a^b g(x)dx \right|$$

(use (9.4.9) and (9.4.10))

$$\leqslant |\alpha| \frac{\varepsilon}{K} + |\beta| \frac{\varepsilon}{K} = \frac{|\alpha| + |\beta|}{|\alpha| + |\beta| + 1}\varepsilon < \varepsilon.$$

This proves that $\alpha f + \beta g$ is Riemann integrable and

$$\int_a^b f(x)dx = S = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx.$$

□

Corollary 9.4.10. *Suppose that $f, g : [a, b] \to \mathbb{R}$ are two functions such that*

$$f(x) = g(x), \;\; \forall x \in (a, b).$$

If f is Riemann integrable, then so is g and, moreover,

$$\int_a^b f(x)dx = \int_a^b g(x)dx. \tag{9.4.11}$$

Proof. Consider the difference $h : [a, b] \to \mathbb{R}$, $h(x) = g(x) - f(x)$, $\forall x \in [a, b]$. Note that h is bounded on $[a, b]$ and continuous on (a, b) because $h(x) = 0$, $\forall x \in (a, b)$. Using Proposition 9.4.4 we deduce that h is Riemann integrable on $[a, b]$. Since $g = f + h$, we deduce from Proposition 9.4.9 that g is Riemann integrable on $[a, b]$ and

$$\int_a^b g(x)dx = \int_a^b f(x)dx + \int_a^b h(x)dx.$$

Thus, to prove (9.4.11) we have to show that

$$\int_a^b h(x)dx = 0.$$

To do this, denote by $\boldsymbol{U}_n$ the uniform partition of order n of $[a,b]$, and denote by $\underline{\xi}^{(n)}$ the sample of $\boldsymbol{U}_n$ consisting of the midpoints of the intervals of $\boldsymbol{U}_n$. Then

$$\boldsymbol{S}(h, \boldsymbol{U}_n, \underline{\xi}^{(n)}) = 0.$$

Since h is Riemann integrable, we have

$$\int_a^b h(x)dx = \lim_{n\to\infty} \boldsymbol{S}(h, \boldsymbol{U}_n, \underline{\xi}^{(n)}) = 0.$$

□

Example 9.4.11. We say that a function $f : [a,b] \to \mathbb{R}$ is *piecewise constant* if there exists a partition

$$\boldsymbol{P} = (a = x_0 < x_1 < \cdots < x_n = b)$$

and constants $c_1, \dots, c_n$ such that for any $k = 1, \dots, n$ the restriction of f to the open interval (x_{k-1}, x_k) is the constant function c_k. From the above corollary we deduce that f is Riemann integrable on each of the intervals $[x_{k-1}, x_k]$. Moreover, the computation in Example 9.2.5 implies that

$$\int_{x_{k-1}}^{x_k} f(t)dt = c_k(x_k - x_{k-1}).$$

Corollary 9.4.7 implies that f is Riemann integrable on $[a,b]$ and

$$\int_a^b f(x)dx = c_1(x_1 - x_0) + \cdots + c_n(x_n - x_{n-1}).$$

□

Proposition 9.4.12. *Suppose that $f : [a,b] \to \mathbb{R}$ is a Riemann integrable function, J is an interval containing the range of f and $G : J \to \mathbb{R}$ is a Lipschitz function. Then $G \circ f : [a,b] \to \mathbb{R}$ is Riemann integrable.*

Proof. Fix a positive constant L such that

$$|G(y_1) - G(y_2)| \leqslant L|y_1 - y_2|, \quad \forall y_1, y_2 \in J.$$

Observe that for any $X \subset [a,b]$ and any $x', x'' \in X$ we have

$$\big| G \circ f(x') - G \circ f(x'') \big| \leqslant L|f(x') - f(x'')|.$$

Hence

$$\operatorname{osc}(G \circ f, X) = \sup_{x', x'' \in X} \big| G \circ f(x') - G \circ f(x'') \big| \leqslant L \sup_{x', x'' \in X} |f(x') - f(x'')| = L \operatorname{osc}(f, X).$$

We deduce as in the proof of Proposition 9.4.9 that for any partition $\boldsymbol{P}$ of $[a,b]$ we have

$$\omega(G \circ f, \boldsymbol{P}) \leqslant L\omega(f, \boldsymbol{P}).$$

Since f is Riemann integrable we deduce that

$$\lim_{\|\boldsymbol{P}\|\to 0} \omega(f, \boldsymbol{P}) = 0$$

so that

$$\lim_{\|\boldsymbol{P}\|\to 0} \omega(G\circ f, \boldsymbol{P}) = 0.$$

□

Corollary 9.4.13. *Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then f^2 is also Riemann integrable on $[a, b]$.*

Proof. Since f is Riemann integrable it is bounded so its range is contained in some interval $[-M, M]$, $M > 0$. The function $G : [-M, M] \to \mathbb{R}$, $G(x) = x^2$ is Lipschitz on this interval because for any $x, y \in [-M, M]$ we have

$$|G(x) - G(y)| = |x^2 - y^2| = |x + y| \cdot |x - y| \leqslant (|x| + |y|)|x - y| \leqslant 2M|x - y|.$$

Proposition 9.4.12 implies that $G \circ f = f^2$ is Riemann integrable. □

Corollary 9.4.14. *If $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable, then so is their product fg.*

Proof. The function $f+g$ is integrable according to Proposition 9.4.9. Invoking Corollary 9.4.13 we deduce that the functions $(f + g)^2, f^2, g^2$ are Riemann integrable. Proposition 9.4.9 now implies that the function

$$\frac{1}{2}\Big((f+g)^2 - f^2 - g^2 \Big) = \frac{1}{2}\big(f^2 + g^2 + 2fg - f^2 - g^2 \big) = fg$$

is Riemann integrable. □

Corollary 9.4.15. *Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then the function $|f|$ is also Riemann integrable.*

Proof. The function $G : \mathbb{R} \to \mathbb{R}$, $G(y) = |y|$ is Lipschitz so the function $G \circ f = |f|$ is Riemann integrable. □

☞ ***A very useful convention.*** We denoted the Riemann integral of a function $f : [a, b] \to \mathbb{R}$ with the symbol

$$\int_a^b f(x)dx,$$

where the lower endpoint a is at the bottom of the integral sign $\int$ and the upper endpoint b is at the top of the integral sign. We define

$$\int_b^a f(x)dx := -\int_a^b f(x)dx, \;\; \int_a^a f(x)dx = 0.$$

There are several arguments in favor of this convention. For example, we can rewrite (9.5.3) as

$$f(\xi)=\frac{1}{b-a}\int_a^b f(x)dx=\frac{1}{a-b}\int_b^a f(x)dx. \tag{9.4.12}$$

This formulation will be especially useful when we do not know whether $a<b$ or $b<a$. The above equality says that it does not matter.

Another advantage comes from the following additivity identity.

$$\int_a^c f(x)dx=\int_a^b f(x)dx+\int_b^c f(x)dx,\quad \forall a,b,c\in\mathbb{R}. \tag{9.4.13}$$

If $a<b<c$, then (9.4.13) is an immediate consequence of Corollary 9.4.7. When the numbers a,b,c are situated in a different order, the identity (9.4.13) is still a consequence of Corollary 9.4.7, but in a more roundabout way. For example, if $a=0$, $b=2$ and $c=1$, then

$$\int_0^1 f(x)dx=\int_0^2 f(x)dx-\int_1^2 f(x)dx=\int_0^2 f(x)dx+\int_2^1 f(x)dx. \qquad \square$$

9.5. Basic properties of the Riemann integral

Now that we have seen how the concept of integrability interacts with the basic arithmetic operations on functions we want to discuss a few simple techniques for estimating Riemann integrals. All these techniques are based on the following simple result.

Proposition 9.5.1 (Positivity). *Suppose that* $f:[a,b]\to\mathbb{R}$ *is Riemann integrable and* $f(x)\geqslant 0$ *for any* $x\in[a,b]$. *Then*

$$\int_a^b f(x)dx\geqslant 0.$$

Proof. Denote by $\boldsymbol{U}_1$ the partition of $[a,b]$ consisting of a single interval. Then

$$0\leqslant\Big(\inf_{x\in[a,b]}f(x)\Big)(b-a)=\boldsymbol{S}_*(f,\boldsymbol{U}_1)\overset{(9.3.10)}{\leqslant}\int_a^b f(x)dx.$$

$\square$

Corollary 9.5.2 (Monotonicity). *If* $f,g:[a,b]\to\mathbb{R}$ *are Riemann integrable functions and* $f(x)\leqslant g(x)$, $\forall x\in[a,b]$, *then*

$$\int_a^b f(x)dx\leqslant\int_a^b g(x)dx.$$

Proof. The function $(g - f)$ is integrable and nonnegative so

$$\int_a^b g(x)dx - \int_a^b f(x)dx = \int_a^b (g(x) - f(x))dx \geqslant 0.$$

□

Corollary 9.5.3. *If $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then*

$$\boxed{\left| \int_a^b f(x)dx \right| \leqslant \int_a^b |f(x)|\, dx.} \tag{9.5.1}$$

Proof. We know that

$$f(x) \leqslant |f(x)| \;\text{ and }\; -f(x) \leqslant |f(x)|, \;\; \forall x \in [a, b].$$

Hence

$$\int_a^b f(x)dx \leqslant \int_a^b |f(x)|dx \;\text{ and }\; -\int_a^b f(x)dx \leqslant \int_a^b |f(x)|dx.$$

The last two inequalities imply (9.5.1).

□

Corollary 9.5.4. *Suppose that $f : [a, b] \to \mathbb{R}$ is a Riemann integrable function. We set*

$$m := \inf_{x \in [a,b]} f(x), \;\; M = \sup_{x \in [a,b]} f(x).$$

Then

$$m(b - a) \leqslant \int_a^b f(x)dx \leqslant M(b - a).$$

Proof. We have

$$m \leqslant f(x) \leqslant M, \;\; \forall x \in [a, b],$$

so that

$$m(b - a) = \int_a^b m dx \leqslant \int_a^b f(x)dx \leqslant \int_a^b M dx = M(b - a).$$

□

Definition 9.5.5. If $f : [a, b] \to \mathbb{R}$ is a Riemann integrable function, then the quantity

$$\frac{1}{b - a} \int_a^b f(x)dx$$

is called the *average value* of f, or the *mean* of f, or the *expectation* of f and we denote it by $\operatorname{Mean}(f)$. □

We see that we can rephrase the inequality in Corollary 9.5.4 as

$$\inf_{x\in[a,b]} f(x) \leqslant \operatorname{Mean}(f) \leqslant \sup_{x\in[a,b]} f(x). \tag{9.5.2}$$

Theorem 9.5.6 (Integral Mean Value Theorem)**.** *Suppose that $f:[a,b]\to\mathbb{R}$ is a **continuous** function. Then there exists $\xi\in[a,b]$ such that*

$$f(\xi) = \operatorname{Mean}(f),$$

i.e.,

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)dx. \tag{9.5.3}$$

Proof. Let

$$m := \inf_{x\in[a,b]} f(x),\quad M = \sup_{x\in[a,b]} f(x).$$

Then (9.5.2) implies that $\operatorname{Mean}(f)\in[m,M]$.

On the other hand, since f is continuous we deduce from Weierstrass' Theorem 6.2.4 that there exist $x_*, x^*\in[a,b]$ such that

$$f(x_*) = m,\quad f(x^*) = M.$$

Since $\operatorname{Mean}(f)\in[f(x_*), f(x^*)]$ we deduce from the Intermediate Value Theorem that there exists ξ in the interval $[x_*, x^*]$ such that $f(\xi) = \operatorname{Mean}(f)$. □

Theorem 9.5.7. *Suppose that $f:[a,b]\to\mathbb{R}$ is a Riemann integrable function. We define*

$$F:[a,b]\to\mathbb{R},\quad F(x) := \int_a^x f(t)dt.$$

Then the following hold.

(i) *The function F is Lipschitz. In particular, F is continuous.*

(ii) *If the function f is continuous, then the function $F(x)$ is differentiable on $[a,b]$ and*

$$F'(x) = f(x),\quad \forall x\in[a,b].$$

In other words, $F(x)$ is an antiderivative of f, more precisely the unique antiderivative on $[a,b]$ such that $F(a) = 0$.

Proof. (i) We set

$$M := \sup_{x\in[a,b]} |f(x)|.$$

If $x, y\in[a,b]$, $x<y$, then

$$|F(x)-F(y)| = |F(y)-F(x)| = \left|\int_a^y f(t)dt - \int_a^x f(t)dt\right|$$

$$= \left| \int_x^y f(t)dt \right| \leqslant \int_x^y |f(t)|dt \leqslant \int_x^y M dt = M(y-x) = M|x-y|.$$

This proves that F is Lipschitz.

(ii) We have to prove that if $x_0 \in [a,b]$, then

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

Using (9.4.13) we deduce

$$F(x)-F(x_0) = \int_a^x f(t)dt - \int_a^{x_0} f(t)dt = \int_{x_0}^x f(t)dt$$

so that we have to show that

$$\lim_{x\to x_0} \frac{1}{x-x_0} \int_{x_0}^x f(t)dt = f(x_0).$$

In other words, we have to prove that for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\forall x \in [a,b],\ \ 0 < |x-x_0| < \delta \Rightarrow \left| \frac{1}{x-x_0} \int_{x_0}^x f(t)dt - f(x_0) \right| < \varepsilon. \tag{9.5.4}$$

Since f is continuous at x_0, given $\varepsilon > 0$ we can find $\delta = \delta(\varepsilon) > 0$ such that

$$\forall x \in [a,b],\ \ |x-x_0| < \delta \Rightarrow |f(x)-f(x_0)| < \varepsilon.$$

On the other hand, invoking the continuity of f again, we deduce from the Integral Mean Value Theorem that, for any $x \neq x_0$, there exists ξ_x between x_0 and x such that

$$f(\xi_x) = \frac{1}{x-x_0} \int_{x_0}^x f(t)dt.$$

In particular, if $|x-x_0| < \delta$, then $|\xi_x - x_0| < \delta$, and thus

$$\left| \frac{1}{x-x_0} \int_{x_0}^x f(t)dt - f(x_0) \right| = |f(\xi_x) - f(x_0)| < \varepsilon.$$

□

9.6. How to compute a Riemann integral

To this day, the best method of computing by hand Riemann integrals is the fundamental theorem of calculus.

Theorem 9.6.1 (The Fundamental Theorem of Calculus: Part 1)**.** *Suppose that $f : [a,b] \to \mathbb{R}$ is a function satisfying the following two conditions.*

(i) *The function f is Riemann integrable.*

(ii) *The function f admits antiderivatives on $[a,b]$.*

If $F : [a,b] \to \mathbb{R}$ is an antiderivative of f, then

$$\boxed{\int_a^x f(t)dt = F(x) - F(a), \quad \forall x \in (a,b].} \tag{9.6.1}$$

In particular,

$$\boxed{\int_a^b f(t)dt = F(t)\Big|_{t=a}^{t=b} := F(b) - F(a).} \tag{9.6.2}$$

Proof. Fix $x \in (a,b]$. Denote by $\boldsymbol{U}_n$ the uniform partition of $[a,x]$ of order n. Since f is Riemann integrable we deduce that *for any choices of samples* $\underline{\xi}^{(n)}$ of $\boldsymbol{U}_n$ we have

$$\int_a^x f(t)dt = \lim_{n\to\infty} \boldsymbol{S}\big(\, f, \boldsymbol{U}_n, \underline{\xi}^{(n)}\,\big).$$

The miracle is that for any n we can cleverly choose a sample

$$\underline{\xi}^{(n)} = (\xi_1^n, \dots, \xi_n^n)$$

of $\boldsymbol{U}_n$ such that the Riemann sum $\boldsymbol{S}\big(\, f, \boldsymbol{U}_n, \underline{\xi}^{(n)}\,\big)$ has an extremely simple form. Here are the details.

The k-th node of $\boldsymbol{U}_n$ is $x_k^n = a + \frac{k}{n}(x-a)$ and the k-th interval is $I_k = [x_{k-1}^n, x_k^n]$. The function F is differentiable on the closed interval $[a,b]$ and, in particular, it is continuous on $[a,b]$. We can invoke Lagrange's Mean Value Theorem to conclude that, for any $k = 1, \dots, n$, there exists $\xi_k^n \in (x_{k-1}^n, x_k^n)$ such that

$$f(\xi_k^n) = F'(\xi_k^n) = \frac{F(x_k^n) - F(x_{k-1}^n)}{x_k^n - x_{k-1}^n},$$

i.e.,

$$f(\xi_k^n)(x_k^n - x_{k-1}^n) = F(x_k^n) - F(x_{k-1}^n).$$

The collection $(\xi_1^n, \dots, \xi_n^n)$ is a sample $\underline{\xi}^{(n)}$ of the partition $\boldsymbol{U}_n$. The associated Riemann sum satisfies

$$\begin{aligned}\boldsymbol{S}\big(\, f, \boldsymbol{U}_n, \underline{\xi}^{(n)}\,\big) &= f(\xi_1^n)(x_1^n - x_0^n) + f(\xi_2^n)(x_2^n - x_1^n) + \cdots + f(\xi_n^n)(x_n^n - x_{n-1}^n)\\ &= F(x_1^n) - F(x_0^n) + F(x_2^n) - F(x_1^n) + \cdots + F(x_n^n) - F(x_{n-1}^n)\end{aligned}$$

(the above is a telescopic sum!!!)

$$= F(x_n^n) - F(x_0^n) = F(x) - F(a).$$

Thus the sequence of Riemann sums $\boldsymbol{S}(f, \boldsymbol{U}_n, \underline{\xi}^{(n)})$ is constant, equal to $F(x) - F(a)$. Hence

$$\int_a^x f(t)dt = \lim_{n\to\infty} \boldsymbol{S}\big(\, f, \boldsymbol{U}_n, \underline{\xi}^{(n)}\,\big) = F(x) - F(a).$$

The equality (9.6.2) follows from (9.6.1) by letting $x = b$. □

Corollary 9.6.2 (The Fundamental Theorem of Calculus: Part 2)**.** *Suppose that*

$$f:[a,b]\to\mathbb{R}$$

is a continuous function. Then f admits antiderivatives on $[a,b]$ and, if $F(x)$ is any antiderivative of f on $[a,b]$, then

$$\int_a^b f(x)dx = F\Big|_a^b := F(b)-F(a),\ \ F(x)=F(a)+\int_a^x f(t)dt,\ \ \forall x\in[a,b]. \tag{9.6.3}$$

Proof. The fact that f admits antiderivatives follows from Theorem 9.5.7(b). The rest follows from Theorem 9.6.1. □

Remark 9.6.3. (a) Theorem 9.6.1 shows that the computation of Riemann integral of a function can be reduced to the computation of the antiderivatives of that function, if they exist. As we have seen in the previous chapter, for many classes of continuous function this computation can be carried out successfully in a *finite number* of purely algebraic steps.

If we ponder for a little bit, the equality (9.6.2) is a truly remarkable result. The left-hand side of (9.6.2) is a Riemann integral defined by a very laborious limiting process which involves *infinitely many and computationally very punishing steps.* The right-hand side of (9.6.2) involves computing the values of an antiderivative at two points. Often this can be achieved in *finitely many arithmetic steps*!

The attribute *fundamental* attached to Theorem 9.6.1 is fully justified: it describes a *finite-time* shortcut to an *infinite-time* process.

(b) Both assumptions (i) and (ii) are needed in Theorem 9.6.1! Indeed, there exist functions that satisfy (i) but not (ii), and there exist functions satisfying (ii), but not (i). Their constructions are rather ingenious and we refer to [**19**] for more details. Note that the continuous functions automatically satisfy both (i) and (ii). □

Example 9.6.4. For $k\in\mathbb{N}$ consider the continuous function $f:[0,1]\to\mathbb{R}$, $f(x)=x^k$. The function $F(x)=\frac{1}{k+1}x^{k+1}$ is an antiderivative of f and (9.6.3) implies

$$\int_0^1 x^k dx = \left(\frac{1}{k+1}x^{k+1}\right)\Big|_0^1 = \frac{1}{k+1}.$$

In particular, for $k=2$ we deduce

$$\int_0^1 x^2 dx = \frac{1}{3}.$$

This agrees with the elementary computations in Section 9.1. □

The techniques for computing antiderivatives can now be used for computing Riemann integrals. As we have seen, there are basically two methods for computing antiderivatives:

integration by parts, and change of variables. These lead to two basic techniques for computing Riemann integrals. In applications most often one needs to use a blend of these techniques to compute a Riemann integral.

9.6.1. Integration by parts. We state a special case that covers most of the concrete situations.

Proposition 9.6.5. *Suppose that $u, v : [a, b] \to \mathbb{R}$ are two C^1 functions, i.e., they are differentiable and have continuous derivatives. Then uv' and $u'v$ are Riemann integrable and*

$$\boxed{\int_a^b u(x)v'(x)dx = u(x)v(x)\Big|_a^b - \int_a^b v(x)u'(x)dx}. \tag{9.6.4}$$

Proof. The functions $u'v$ and uv' are continuous since they are products of continuous functions. In particular these functions are integrable, and we have

$$\begin{aligned}\int_a^b u'(x)v(x)dx + \int_a^b u(x)v'(x)dx &= \int_a^b \big(\, u'(x)v(x) + u(x)v'(x)\,\big)dx\\ &= \int_a^b (uv)'(x)dx \overset{(9.6.2)}{=} u(x)v(x)\Big|_a^b.\end{aligned}$$

The equality (9.6.4) is now obvious. □

Remark 9.6.6. The integration-by-parts formula (9.6.4) is often written in the shorter form

$$\boxed{\int_a^b udv = uv\Big|_a^b - \int_a^b vdu.} \tag{9.6.5}$$

Observing that

$$uv\Big|_b^a = u(a)v(a) - u(b)v(b) = -\big(\, u(b)v(b) - u(a)v(a)\,\big) = -uv\Big|_a^b,$$

we deduce that

$$\int_b^a udv = uv\Big|_b^a - \int_b^a vdu,$$

even though the upper limit of integration a is smaller than the lower limit of integration b. □

Example 9.6.7. For any nonnegative integers m, n we set

$$I_{m,n} = \int_{-1}^1 (x-1)^m(x+1)^n dx. \tag{9.6.6}$$

This integral is theoretically computable because $(x-1)^m(x+1)^n$ is a polynomial. Its precise form is obtained via Newton's binomial formula and the final result is rather complicated. For example

$$(x-1)^2(x+1)^3 = (x^2-2x+1)(x^3+3x^2+3x+1) = x^5 + x^4 - 2\,x^3 - 2\,x^2 + x + 1.$$

In general, we need to multiply the two polynomials in the right-hand side of (9.6.6) to obtain the explicit form of $(x-1)^m(x+1)^n$. This is an elaborate process which becomes increasingly more complex as the powers m and n increase. However, an ingenious usage of the integration-by-parts trick leads to a much simpler way of computing $I_{m,n}$.

Let us first observe that

$$(x+1)^n = \frac{1}{n+1}\frac{d}{dx}(x+1)^{n+1},$$

from which we deduce

$$I_{0,n} = \int_{-1}^{1}(x+1)^n dx = \frac{1}{n+1}(x+1)^{n+1}\Big|_{-1}^{1} = \frac{2^{n+1}}{n+1}. \tag{9.6.7}$$

Observe now that if $m > 0$, then

$$\begin{aligned} I_{m,n} &= \int_{-1}^{1}(x-1)^m(x+1)^n dx = \frac{1}{n+1}\int_{-1}^{1}(x-1)^m\frac{d}{dx}(x+1)^{n+1}dx \\ &= \underbrace{\frac{1}{n+1}(x-1)^m(x+1)^{n+1}\Big|_{-1}^{1}}_{=0} - \frac{m}{n+1}\int_{-1}^{1}(x-1)^{m-1}(x+1)^{n+1}dx. \end{aligned}$$

We obtain in this fashion the recurrence relation

$$I_{m,n} = -\frac{m}{n+1}I_{m-1,n+1}, \quad \forall m > 0, \quad n \geqslant 0. \tag{9.6.8}$$

If $m-1>0$, then we can continue this process and we deduce

$$I_{m-1,n+1} = -\frac{m-1}{n+2}I_{m-2,n+2} \Rightarrow I_{m,n} = \frac{m(m-1)}{(n+1)(n+2)}I_{m-2,n+2}.$$

Iterating this procedure we conclude that

$$\begin{aligned} I_{m,n} &= (-1)^m\frac{m(m-1)\cdots 2\cdot 1}{(n+1)(n+2)\cdots(n+m-1)(n+m)}I_{0,n+m} \\ &= (-1)^m\frac{m!}{(n+1)\cdots(n+m)}I_{0,n+m} = (-1)^m\frac{1}{\binom{n+m}{m}}I_{0,n+m}. \end{aligned}$$

Invoking (9.6.7) we deduce

$$\boxed{I_{m,n} = (-1)^m\frac{1}{\binom{n+m}{m}}\cdot\frac{2^{n+m+1}}{(n+m+1)}}. \tag{9.6.9}$$

When $m = n$ we have

$$I_{n,n} = \int_{-1}^{1}(x-1)^n(x+1)^n dx = \int_{-1}^{1}(x^2-1)^n dx$$

and we conclude that

$$\boxed{\int_{-1}^{1}(x^2-1)^n dx = I_{n,n} = \frac{(-1)^n}{\binom{2n}{n}}\cdot\frac{2^{2n+1}}{(2n+1)}}. \tag{9.6.10}$$

□

Example 9.6.8 (Wallis' formula)**.** For nonnegative integer n we set

$$I_n := \int_0^{\frac{\pi}{2}} (\sin x)^n \, dx.$$

Note that

$$I_0 = \frac{\pi}{2}, \;\; I_1 = \int_0^{\frac{\pi}{2}} \sin x \, dx = (-\cos x)\Big|_{x=0}^{x=\frac{\pi}{2}} = 1.$$

In general, for $n > 0$, we have

$$I_{n+1} = \int_0^{\frac{\pi}{2}} (\sin x)^n d(-\cos x) = \underbrace{(\sin x)^n(-\cos x)\Big|_{x=0}^{x=\frac{\pi}{2}}}_{=0} + \int_0^{\frac{\pi}{2}} \cos x \, d(\sin x)^n$$

$$= n\int_0^{\frac{\pi}{2}} (\sin x)^{n-1} \cos^2 x \, dx = n\int_0^{\frac{\pi}{2}} (\sin x)^{n-1}(1-\sin^2 x)\, dx = nI_{n-1} - nI_{n+1}.$$

Hence

$$I_{n+1} = nI_{n-1} - nI_{n+1}$$

so that

$$(n+1)I_{n+1} = nI_{n-1}, \;\; I_{n+1} = \frac{n}{n+1} I_{n-1}. \tag{9.6.11}$$

We deduce

$$I_2 = \frac{1}{2} I_0 = \frac{1}{2}\frac{\pi}{2}, \;\; I_4 = \frac{3}{4} I_2 = \frac{3}{4}\frac{1}{2}\frac{\pi}{2},$$

and, in general,

$$\boxed{I_{2n} = \int_0^{\frac{\pi}{2}} (\sin x)^{2n} dx = \frac{2n-1}{2n} \cdots \frac{3}{4}\frac{1}{2}\frac{\pi}{2}}. \tag{9.6.12}$$

Similarly,

$$I_3 = \frac{2}{3} I_1 = \frac{2}{3}, \;\; I_5 = \frac{4}{5} I_3 = \frac{4}{5}\frac{2}{3},$$

and, in general,

$$\boxed{I_{2n+1} = \int_0^{\frac{\pi}{2}} (\sin x)^{2n+1} dx = \frac{2n}{2n+1} \cdots \frac{4}{5}\frac{2}{3}}. \tag{9.6.13}$$

If we introduce the notation

$$(2k)!! := 2 \cdot 4 \cdot 6 \cdots (2k), \;\; (2k-1)!! := 1 \cdot 3 \cdot 5 \cdots (2k-1), \tag{9.6.14}$$

then we can rewrite the equalities (9.6.12) and (9.6.13) in a more compact form

$$\boxed{I_{2j} = \frac{\pi}{2}\frac{(2j-1)!!}{(2j)!!}, \;\; I_{2j-1} = \frac{(2j-2)!!}{(2j-1)!!}}. \tag{9.6.15}$$

Since $\sin x \in [0,1]$, $\forall x \in [0,\pi/2]$, we deduce

$$(\sin x)^{n+1} \leqslant (\sin x)^n, \quad \forall x \in [0,\pi/2],$$

and thus,

$$I_{n+1} \leqslant I_n, \quad \forall n \in \mathbb{N}.$$

We deduce

$$\frac{2n}{2n+1} \overset{(9.6.11)}{=} \frac{I_{2n+1}}{I_{2n-1}} \leqslant \frac{I_{2n+1}}{I_{2n}} \leqslant 1.$$

From the above equalities we deduce

$$\lim_{n\to\infty} \frac{I_{2n+1}}{I_{2n}} = 1.$$

Using (9.6.12) and (9.6.13) we deduce

$$\frac{I_{2n+1}}{I_{2n}} = \frac{2}{\pi} \cdot \frac{1}{2n+1} \cdot \frac{2^2 4^2 \cdots (2n)^2}{1^2 3^2 \cdots (2n-1)^2}.$$

This implies the celebrated *Wallis' formula*

$$\boxed{\frac{\pi}{2} = \frac{\pi}{2} \lim_{n\to\infty} \frac{I_{2n+1}}{I_{2n}} = \lim_{n\to\infty} \frac{2^2 4^2 \cdots (2n)^2}{1^2 3^2 \cdots (2n-1)^2} \cdot \frac{1}{2n+1}.} \tag{9.6.16}$$

Later on, we will need an equivalent version of the above equality, namely

$$\boxed{\frac{\pi}{2} = \frac{\pi}{2} \lim_{n\to\infty} \frac{2n+1}{2n} \frac{I_{2n+1}}{I_{2n}} = \lim_{n\to\infty} \frac{2^2 4^2 \cdots (2n)^2}{1^2 3^2 \cdots (2n-1)^2} \cdot \frac{1}{2n}.} \tag{9.6.17}$$

□

Let us discuss another simple but useful application of the integration-by-parts trick.

Proposition 9.6.9 (Integral remainder formula). *Let $n \in \mathbb{N}$ and suppose that $f : [a,b] \to \mathbb{R}$ is a C^{n+1}-function, i.e., $(n+1)$-times differentiable and the $(n+1)$-th derivative is continuous. If $x_0 \in [a,b]$ and $T_n(x)$ is the degree-n Taylor polynomial of f at x_0,*

$$T_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n,$$

then the remainder $R_n(x) := f(x) - T_n(x)$ admits the integral representation

$$\boxed{R_n(x) = \frac{1}{n!}\int_{x_0}^x f^{(n+1)}(t)(x-t)^n dt, \quad \forall x \in [a,b]}. \tag{9.6.18}$$

Proof. Fix $x \neq x_0$. We have

$$f(x) - f(x_0) = \int_{x_0}^x f'(t)dt = -\int_{x_0}^x f'(t)\frac{d}{dt}(x-t)\, dt$$

$$= -\Big(f'(t)(x-t)\Big)\Big|_{t=x_0}^{t=x} + \int_{x_0}^x f''(t)(x-t)dt$$

$$
\begin{aligned}
&= f'(x_0)(x-x_0) - \int_{x_0}^{x} f''(t)\frac{d}{dt}\left(\frac{1}{2}(x-t)^2\right)dt\\
&= f'(x_0)(x-x_0) - \left(\frac{1}{2}f''(t)(x-t)^2\right)\Big|_{t=x_0}^{t=x} + \frac{1}{2}\int_{x_0}^{x} f^{(3)}(t)(x-t)^2dt\\
&= f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 - \frac{1}{3!}\int_{x_0}^{x} f^{(3)}(t)\frac{d}{dt}(x-t)^3dt\\
&= f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 - \frac{1}{3!}\big(f^{(3)}(t)(x-t)^3\big)\Big|_{t=x_0}^{t=x} + \frac{1}{3!}\int_{x_0}^{x} f^{(4)}(t)(x-t)^3dt\\
&= f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \frac{f^{(3)}(x_0)}{3!}(x-x_0)^3 + \frac{1}{3!}\int_{x_0}^{x} f^{(4)}(t)(x-t)^3dt\\
&= f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \frac{f^{(3)}(x_0)}{3!}(x-x_0)^3 - \frac{1}{4!}\int_{x_0}^{x} f^{(4)}(t)\frac{d}{dt}(x-t)^4dt\\
&= \cdots\cdots\cdots\cdots\cdots =\\
&= f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^ndt.
\end{aligned}
$$

Thus

$$
\begin{aligned}
f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n\\
&\quad + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^ndt\\
&= T_n(x) + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^ndt.
\end{aligned}
$$

This proves (9.6.18). □

Example 9.6.10. Let us show how we can use the integral remainder formula to strengthen the result in Exercise 8.7. Consider the function $f:(-1,1)\to\mathbb{R}$, $f(x)=\ln(1-x)$. Since

$$
f'(x) = -\frac{1}{1-x} = (x-1)^{-1},\quad f''(x) = \frac{d}{dx}(x-1)^{-1} = -(x-1)^{-2},
$$

$$
f^{(3)}(x) = -\frac{d}{dx}(x-1)^{-2} = 2(x-1)^{-3},\ldots
$$

$$
f^{(n)}(x) = (-1)^{n-1}(n-1)!(x-1)^{-n},\quad \forall n\in\mathbb{N}
$$

we deduce that

$$
f(0)=0,\quad f^{(n)}(0) = (-1)^n(n-1)!(-1)^{-n} = -(n-1)!,\quad \forall n\in\mathbb{N},
$$

and thus, the Taylor series of f at $x_0=0$ is

$$
-\sum_{k=1}^{\infty}\frac{x^k}{k}.
$$

We denote by $T_n(x)$ the degree n Taylor polynomial of $f(x)$ at $x_0 = 0$,

$$T_n(x) = -\sum_{k=1}^{n} \frac{x^k}{k} = -x - \frac{x^2}{2} - \cdots - \frac{x^n}{n!}.$$

We want to prove that this series converges to $\ln(1-x)$ for any $x \in [-1,1)$. To do this we have to show that

$$\lim_{n\to\infty} |f(x) - T_n(x)| = 0, \;\; \forall x \in [-1,1).$$

We need to estimate the remainder $R_n(x) = f(x) - T_n(x)$. We distinguish two cases.

1. $x \in [0,1)$. Using the integral remainder formula (9.6.18) we deduce

$$R_n(x) = \frac{1}{n!}\int_0^x f^{(n+1)}(t)(x-t)^n dt = (-1)^n \int_0^x (t-1)^{-n-1}(x-t)^n dt.$$

Hence

$$|R_n(x)| = \int_0^x \frac{(x-t)^n}{(1-t)^{n+1}} dt.$$

Observe that for $t \in [0,x]$ we have $1-t \geqslant 1-x > 0$ so that, for any $t \in [0,x]$ we have

$$(1-t)^{n+1} \geqslant (1-t)^n(1-x) > 0, \Longleftrightarrow 0 < \frac{1}{(1-t)^{n+1}} \leqslant \frac{1}{1-x} \cdot \frac{1}{(1-t)^n}.$$

Hence

$$|R_n(x)| \leqslant \frac{1}{1-x}\int_0^x \left(\frac{x-t}{1-t}\right)^n dt.$$

Now consider the function

$$g : [0,x] \to \mathbb{R}, \;\; g(t) = \frac{x-t}{1-t}.$$

We have

$$g'(t) = \frac{-(1-t) + (x-t)}{(1-t)^2} = \frac{x-1}{(1-t)^2} < 0.$$

Hence

$$0 = g(x) \leqslant g(t) \leqslant g(0) = x, \;\; \forall t \in [0,x],$$

and thus

$$|R_n(x)| \leqslant \frac{1}{1-x}\int_0^x g(t)^n dt \leqslant \frac{1}{1-x}\int_0^x x^n dt = \frac{x^{n+1}}{1-x}.$$

We deduce

$$|R_n(x)| \leqslant \frac{x^{n+1}}{1-x}, \;\; \forall x \in [0,1),$$

so that

$$\lim_{n\to\infty} R_n(x) = \lim_{n\to\infty} \frac{x^{n+1}}{1-x} = 0, \;\; \forall x \in [0,1).$$

2. $x \in [-1,0)$. We estimate $R_n(x)$ using the Lagrange remainder formula. Hence, there exists $\xi \in (x,0)$ such that

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1} = (-1)^n \frac{n!(\xi-1)^{-(n+1)}}{(n+1)!}x^{n+1} = (-1)^n \frac{1}{(n+1)(\xi-1)^{n+1}}x^{n+1}.$$

Hence, since $\xi \in (x,0)$, we have $|\xi - 1| = |\xi| + 1$ and

$$|R_n(x)| = \frac{|x|^{n+1}}{(n+1)(1+|\xi|)^{n+1}} \leqslant \frac{|x|^{n+1}}{n+1}.$$

Since $|x| \leqslant 1$ we deduce

$$\lim_{n\to\infty} |R_n(x)| = 0, \;\; \forall x \in [-1,0).$$

We have thus proved that

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}, \;\; \forall x \in [-1,1).$$

Note in particular that

$$f(-1) = \ln 2 = -\sum_{n=1}^{\infty} \frac{(-1)^n}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots. \tag{9.6.19}$$

□

9.6.2. Change of variables. The change of variables in the Riemann integral is very similar to the integration-by-substitution trick used in the computation of antiderivatives, but it has a few peculiarities. There are two versions of the change in variables formula.

Proposition 9.6.11 (Change in variables formula: version 1, $t = \phi(x)$)**.** *Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function and $\phi : [\alpha,\beta] \to [a,b]$ is a C^1-function. Then the function $f\big(\phi(x)\big)\phi'(x)$ is integrable on $[\alpha,\beta]$ and*

$$\boxed{\int_\alpha^\beta f\big(\phi(x)\big)\phi'(x)dx = \int_{\phi(\alpha)}^{\phi(\beta)} f(t)dt.} \tag{9.6.20}$$

Proof. Since f is continuous it admits antiderivatives. Fix an antiderivative F of f. The chain rule shows that $F\big(\phi(x)\big)$ is an antiderivative of the continuous function $f\big(\phi(x)\big)\phi'(x)$. The Fundamental Theorem of Calculus then shows

$$\int_\alpha^\beta f\big(\phi(x)\big)\phi'(x)dt = F\big(\phi(x)\big)\Big|_{x=\alpha}^{x=\beta} = F\big(\phi(\beta)\big) - F\big(\phi(\alpha)\big) = \int_{\phi(\alpha)}^{\phi(\beta)} f(t)dt.$$

□

We can relax the continuity assumption of f, but to do so we need to make an additional assumption of the nature of the change in variables, $t = \phi(x)$.

Proposition 9.6.12 (Change in variables formula: version 2, $x = \varphi(t)$)**.** *Suppose that $f : [a,b] \to \mathbb{R}$ is a Riemann integrable function and $\varphi : [\alpha,\beta] \to [a,b]$ is a C^1-function such that*

$$\varphi'(t) \neq 0, \;\; \forall t \in (\alpha,\beta).$$

Then $f\big(\varphi(t)\big)\varphi'(t)$ is Riemann integrable on $[\alpha,\beta]$ and

$$\boxed{\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx = \int_\alpha^\beta f\big(\varphi(t)\big)\varphi'(t)dt.} \tag{9.6.21}$$

Proof. Set

$$M := \sup_{t\in[\alpha,\beta]} |\varphi'(t)|.$$

Note that $M > 0$. Since $|\varphi'(t)|$ is continuous, Weierstrass' theorem implies that $M < \infty$. Since $\varphi'(t) \neq 0$ for any $t \in (\alpha,\beta)$ we deduce from the Intermediate Value Theorem that

$$\text{either } \;\varphi'(t) > 0, \;\; \forall t \in (\alpha,\beta) \;\text{ or; }\; \varphi'(t) < 0, \;\; \forall t \in (\alpha,\beta).$$

Thus, either φ is strictly increasing and its range is $[\varphi(\alpha),\varphi(\beta)]$, or φ is strictly decreasing and its range is $[\varphi(\beta),\varphi(\alpha)]$. We need to discuss each case separately, but we will present the details only for the first case and leave the details for the second case for you as an exercise. In the sequel we will assume that φ is increasing and thus

$$0 < \varphi'(t) \leqslant M, \;\; \forall t \in (\alpha,\beta).$$

For simplicity we set

$$g(t) := f\big(\varphi(t)\big)\varphi'(t), \;\; t \in [\alpha,\beta].$$

We will show that g is Riemann integrable on $[\alpha,\beta]$ and its Riemann integral is given by the left-hand side of (9.6.21). We will need the following technical result.

Lemma 9.6.13. *For any partition $\boldsymbol{P}$ of $[\alpha,\beta]$, there exists a partition $\boldsymbol{P}_\varphi$ of $[\varphi(\alpha),\varphi(\beta)]$ and samples $\underline{\xi}$ of $\boldsymbol{P}$ and $\underline{\eta}$ of $\boldsymbol{P}_\varphi$ such that*

$$\|\boldsymbol{P}_\varphi\| \leqslant M\|\boldsymbol{P}\|, \tag{9.6.22a}$$

$$\boldsymbol{S}(\boldsymbol{P}, g, \underline{\xi}) = \boldsymbol{S}(\boldsymbol{P}_\varphi, f, \underline{\xi}_\varphi). \tag{9.6.22b}$$

Let us first show that Lemma 9.6.13 implies that g is Riemann integrable and satisfies (9.6.21). Fix $\varepsilon > 0$. The function f is Riemann integrable on $[\varphi(\alpha),\varphi(\beta)]$ and thus there exists $\delta_0 = \delta_0(\varepsilon) > 0$ such that for any partition $\boldsymbol{Q}$ of $[\varphi(\alpha),\varphi(\beta)]$ and any sample $\underline{\eta}$ of $\boldsymbol{Q}$ we have

$$\left| \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx - \boldsymbol{S}(f, \boldsymbol{Q}, \underline{\eta}) \right| < \varepsilon. \tag{9.6.23}$$

Set

$$\delta = \delta(\varepsilon) := \frac{1}{M}\delta_0(\varepsilon).$$

For any partition $\boldsymbol{P}$ of $[\alpha,\beta]$ of mesh $\|\boldsymbol{P}\| < \delta(\varepsilon)$, and any sample $\underline{\xi}$ of $\boldsymbol{P}$, the sampled partition $(\boldsymbol{P}_\varphi, \underline{\xi}_\varphi)$ of $[\varphi(\alpha),\varphi(\beta)]$ associated to $(\boldsymbol{P}, \underline{\xi})$ by Lemma 9.6.13 satisfies

$$\boldsymbol{P}_\varphi\| < M\delta(\varepsilon) = \delta_0(\varepsilon) \;\text{ and }\; \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}) = \boldsymbol{S}(f, \boldsymbol{P}_\varphi, \underline{\xi}_\varphi).$$

We deduce that

$$\left| \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx - \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}) \right| = \left| \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx - \boldsymbol{S}(f, \boldsymbol{P}_\varphi, \underline{\xi}_\varphi) \right| \overset{(9.6.23)}{<} \varepsilon.$$

This proves that $g(t)$ is integrable on $[\alpha,\beta]$ and its integral is equal to $\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx$.

Proof of Lemma 9.6.13. Consider a partition $\boldsymbol{P} = (\alpha = t_0 < t_1 < \cdots < t_n = \beta)$ of $[\alpha, \beta]$. For $k = 0, 1, \dots, x_n$ we set

$$x_k := \varphi(t_k).$$

Since φ is increasing we have

$$x_{k-1} < x_k, \;\; \forall k = 1, \dots, n.$$

Thus

$$\varphi(\alpha) = x_0 < x_1 < \cdots < x_n = \varphi(\beta)$$

is a partition of $[\varphi(\alpha), \varphi(\beta)]$ that we denote by $\boldsymbol{P}_\varphi$. Note that

$$x_k - x_{k-1} = \varphi(t_k) - \varphi(t_{k-1}).$$

Lagrange's Mean Value theorem implies that there exists $\xi_k \in (t_{k-1}, t_k)$ such that

$$x_k - x_{k-1} = \varphi(t_k) - \varphi(t_{k-1}) = \varphi'(\xi_k)(t_k - t_{k-1}).$$

In particular, this shows that

$$|x_{k-1} - x_k| = |\varphi'(\xi_k)| \cdot |t_k - t_{k-1}| \leqslant M|t_k - t_{k-1}|, \;\; \forall k = 1, \dots, k.$$

Hence

$$\|\boldsymbol{P}_\varphi\| \leqslant M\|\boldsymbol{P}\|.$$

This proves (9.6.22a).

Set $\eta_k := \varphi(\xi_k)$. Note that since φ is increasing we have $\eta_k \in (x_{k-1}, x_k)$. The collection $\underline{\xi} = (\xi_1, \dots, \xi_k)$ is a sample of $\boldsymbol{P}$, and the collection $\underline{\eta} = (\eta_1, \dots, \eta_n)$ is a sample of $\boldsymbol{P}_\varphi$. Observe that

$$f(\eta_k)(x_k - x_{k-1}) = f\big(\varphi(\xi_k)\big)\varphi'(\xi_k)(t_k - t_{k-1}) = g(\xi_k)(t_k - t_{k-1}).$$

Thus

$$\boldsymbol{S}(f, \boldsymbol{P}_\varphi, \underline{\eta}) = \sum_{k=1}^n f(\eta_k)(x_k - x_{k-1}) = \sum_{k=1}^n g(\xi_k)(t_k - t_{k-1}) = \boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}).$$

This proves (9.6.22b) and completes the proof of Proposition 9.6.12. □

Remark 9.6.14. In concrete examples, the right-hand sides of the equalities (9.6.20) and (9.6.21) are quantities that we know how to compute. The left-hand sides are the unknown quantities whose computations are sought. For this reason these two equalities play different roles in applications. □

Example 9.6.15. (a) Suppose that we want to compute

$$\int_{-1}^2 \cos(x^2) x dx = \frac{1}{2}\int_{-1}^2 \cos(x^2) d(x^2).$$

We make the change of variables $t = x^2$. Note that $x = -1 \Rightarrow t = 1$, $x = 2 \Rightarrow t = 4$ and we deduce

$$\int_{-1}^2 \cos(x^2) x dx \overset{(9.6.20)}{=} \frac{1}{2}\int_1^4 \cos t \, dt = \frac{\sin 4 - \sin 1}{2}.$$

Note that in this case (9.6.21) is not applicable.

(b) Suppose that we want to compute

$$\int_0^{\frac{\pi}{2}} e^{\sin x} \cos x \, dx = \int_0^{\frac{\pi}{2}} e^{\sin x} d(\sin x).$$

We make the change in variables $t = \sin x$. Note that $x = 0 \Rightarrow t = 0$, $x = \frac{\pi}{2} \Rightarrow t = 1$ and we deduce

$$\int_0^{\frac{\pi}{2}} e^{\sin x} \cos x dx \overset{(9.6.20)}{=} \int_0^1 e^t dt = e^t \Big|_{t=0}^{t=1} = e - 1.$$

(c) Suppose we want to compute

$$\int_{-1}^1 \sqrt{1 - x^2} dx$$

We make a change of variables $x = \sin t$ so that $dx = d(\sin t) = \cos t dt$. Note that

$$x = -1 \Rightarrow t = -\frac{\pi}{2}, \quad x = 1 \Rightarrow t = \frac{\pi}{2},$$

and $\cos t > 0$ when $t \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Hence

$$\sqrt{1 - x^2} = \sqrt{1 - \sin^2 t} = \sqrt{\cos^2 t} = \cos t, \quad -\frac{\pi}{2} \leqslant t \leqslant \frac{\pi}{2}.$$

We deduce

$$\int_{-1}^1 \sqrt{1 - x^2} dx \overset{(9.6.21)}{=} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2 t dt = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1 + \cos 2t}{2} dt$$
$$= \frac{1}{2}\left(\frac{\pi}{2} + \frac{\pi}{2}\right) + \frac{1}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos 2t dt = \frac{\pi}{2} + \frac{1}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos 2t dt.$$

To compute the last integral we use the change in variables $u = 2t$ so that $dt = \frac{1}{2} du$,

$$t = -\frac{\pi}{2} \Rightarrow u = -\pi, \quad t = \frac{\pi}{2} \Rightarrow u = \pi.$$

Hence

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos 2t \, dt = \frac{1}{2} \int_{-\pi}^{\pi} \cos u \, du = \frac{1}{2}\big(\sin \pi - \sin(-\pi) \big) = 0.$$

We conclude that

$$\int_{-1}^1 \sqrt{1 - x^2} dx = \frac{\pi}{2}. \tag{9.6.24}$$

Let us observe that this equality provides a way of approximating $\frac{\pi}{2}$ by using Riemann sums to approximate the integral in the left-hand side. If we use the uniform partition $\boldsymbol{U}_{200}$ of order 200 of $[-1, 1]$ and as sample $\underline{\xi}$ the right endpoints of the intervals of the partition, then we deduce

$$\pi \approx 2\boldsymbol{S}\big(\sqrt{1 - x^2}, \boldsymbol{U}_{200}, \underline{\xi} \big) \approx 3.14041.....$$

If we use the uniform partition of order $2,000$ and a similar sample, then we deduce

$$\pi \approx 2\boldsymbol{S}\big(\sqrt{1 - x^2}, \boldsymbol{U}_{2,000}, \underline{\xi} \big) \approx 3.14157.....$$

(d) Suppose that we want to compute the integral.

$$\int_1^e \frac{\ln x}{x} dx.$$

We make the change in variables $x = e^t$ and we observe that $dx = e^t dt$,

$$x = 1 \Rightarrow t = 0, \;\; x = e \Rightarrow t = 1.$$

The derivative $\frac{dx}{dt} = e^t$ is everywhere positive and we deduce

$$\int_1^e \frac{\ln x}{x} dx \overset{(9.6.21)}{=} \int_0^1 \frac{\ln e^t}{e^t} e^t dt = \int_0^1 t dt = \frac{t^2}{2}\Big|_{t=0}^{t=1} = \frac{1}{2}. \qquad \square$$

Example 9.6.16 (Stirling's formula)**.** In many applications we need to have a simpler way of understanding the size of $n!$ for n very large. This is what Stirling's formula accomplishes.

More precisely we want to prove the refined inequalities

$$\boxed{1 < \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} < 1 + \frac{1}{4n}, \;\; \forall n \in \mathbb{N}}. \tag{9.6.25}$$

The inequalities (9.6.25) imply the classical *Stirling formula*

$$\boxed{n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \;\; \text{as } n \to \infty}, \tag{9.6.26}$$

where we recall that the notation $x_n \sim y_n$ as $n \to \infty$ (read x_n *is asymptotic to* y_n *as* $n \to \infty$) signifies

$$\lim_{n\to\infty} \frac{x_n}{y_n} = 1.$$

To prove the inequalities (9.6.25) we follow the very nice approach in [**15**, §2.6].

We set

$$F_n := \ln n! = \ln(1) + \ln(2) + \cdots + \ln(n) = \ln(2) + \cdots + \ln(n)$$

and we aim to find accurate approximations for F_n. We will find these by providing rather sharp approximations for the integral

$$I_n := \int_1^n \ln x dx = \big(\, x\ln x - x \,\big)\Big|_{x=1}^{x=n} = n\ln n - n + 1.$$

To see why such an integral might be relevant observe that

$$\ln\left(\frac{n}{e}\right)^n = n\ln n - n.$$

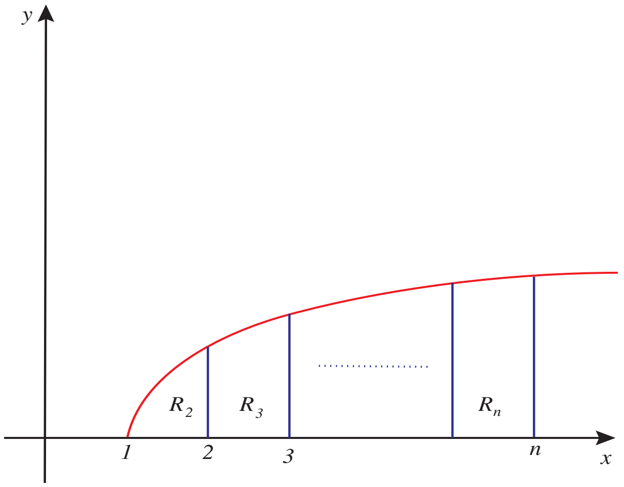

Figure 9.6. *Computing the area underneath* $\ln x$, $x \in [1,n]$.

Observe that I_n is the area below the graph of $\ln x$ and above the interval $[1,n]$ on the x axis. For $k =, 2, 3, \dots, n$ we denote by R_k the region below the graph of $\ln x$ and above the interval $[k-1,k]$ on the x-axis; see Figure 9.6. Then

$$I_n = \text{Area}\,(R_2) + \cdots + \text{Area}\,(R_n).$$

We will provide lower and upper estimates for I_n by producing lower and upper estimates for the areas of the regions R_k. To produce these bounds for the area of R_k we will take adavantage of the fact that $\ln x$ is concave so its graph lies above any chord and below any tangent.

Denote by p_k the point on the graph of $\ln x$ corresponding to $x = k$, i.e., $p_k = (k, \ln k)$. Due to the concavity of $\ln x$ the region R_k contains the trapezoid A_k determined by the chord connecting the points p_{k-1} and p_k; see Figure 9.7.

Denote by q_k the point on the graph of $\ln x$ above the midpoint of the interval $[k-1,k]$, i.e., $q_k = (k - 1/2, \ln(k-1/2)\,)$. The tangent to the graph of $\ln x$ at q_k determines a trapezoid B_k that contains the region R_k Hence

$$\underbrace{\text{Area}\,(R_k) - \text{Area}\,(A_k)}_{=:s_k} < \text{Area}\,(B_k) - \text{Area}\,(A_k).$$

Hence

$$I_n = \sum_{k=2}^n \text{Area}\,(R_k) = \sum_{k=2}^n \text{Area}\,(A_k) + \underbrace{\sum_{k=2}^n s_k}_{=:S_n}. \tag{9.6.27}$$

Observe that

$$\text{Area}\,(A_k) = \frac{1}{2}\big(\ln(k-1) + \ln k\,\big),$$

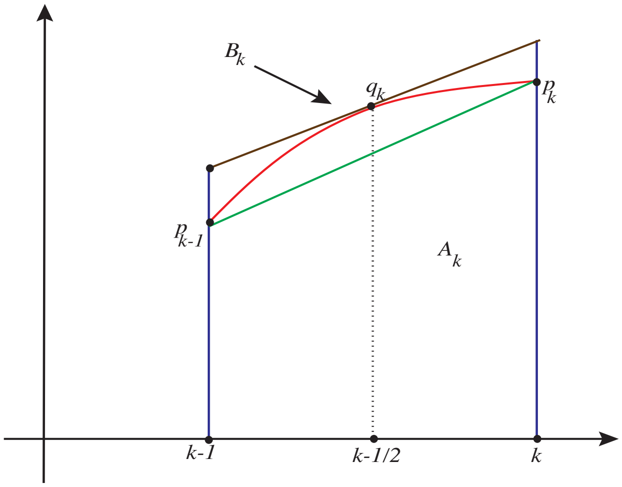

Figure 9.7. *Approximating the region R_k by trapezoids.*

so

$$\begin{aligned}\text{Area}\,(A_2)+\cdots+\text{Area}\,(A_n) &= \frac{1}{2}\log 2+\frac{1}{2}\bigl(\ln 2+\ln 3\bigr)+\cdots+\frac{1}{2}\bigl(\ln(n-1)+\ln n\bigr)\\ &= \ln 2+\ln 3+\cdots+\ln n-\frac{1}{2}\ln n=\ln n!-\frac{1}{2}\ln n.\end{aligned}$$

Using this in (9.6.27) we deduce

$$I_n+\frac{1}{2}\ln n=\ln n!+S_n,$$

Recalling that $I_n=n\ln n-n+1$ we deduce

$$n\ln n-n+\frac{1}{2}\ln n+1-S_n=\ln n!,$$

or, equivalently

$$n!=C_n\sqrt{n}\left(\frac{n}{e}\right)^n,\;\; C_n=e^{1-S_n}. \tag{9.6.28}$$

To progress further we need to gain some information about S_n. Observing that

$$\text{Area}\,(B_k)=\ln\left(k-\frac{1}{2}\right)$$

we deduce

$$\begin{aligned}s_k<\text{Area}\,(B_k)-\text{Area}\,(A_k) &= \ln\left(k-\frac{1}{2}\right)-\frac{1}{2}\bigl(\,\ln(k-1)+\ln k\bigr)\\ &= \frac{1}{2}\ln\left(\frac{k-\frac{1}{2}}{k-1}\right)-\frac{1}{2}\ln\left(\frac{k}{k-\frac{1}{2}}\right)\\ &= \frac{1}{2}\ln\left(1+\frac{1}{2k-2}\right)-\frac{1}{2}\ln\left(1+\frac{1}{2k-1}\right)\end{aligned}$$

$$< \frac{1}{2}\ln\left(1+\frac{1}{2k-2}\right) - \frac{1}{2}\ln\left(1+\frac{1}{2k}\right)$$

We deduce

$$S_n = \sum_{k=2}^{n} s_k < \frac{1}{2}\sum_{k=2}^{n}\left(\ln\left(1+\frac{1}{2k-2}\right) - \frac{1}{2}\ln\left(1+\frac{1}{2k}\right)\right)$$

(the last sum is a telescoping sum)

$$= \frac{1}{2}\ln\left(1+\frac{1}{2}\right) - \frac{1}{2}\ln\left(1+\frac{1}{2n}\right) < \frac{1}{2}\ln\frac{3}{2}.$$

This shows that the sequence S_n is bounded above. Since this sequence is obviously increasing, we deduce that (S_n) is convergent. We denote by S its limit. Since the sequence S_n is increasing, the sequence $C_n = e^{1-S_n}$ is decreasing and converges to $C = e^{1-S}$. Using this in(9.6.28) we deduce

$$n! = C_n\sqrt{n}\left(\frac{n}{e}\right)^n > C\sqrt{n}\left(\frac{n}{e}\right)^n. \tag{9.6.29}$$

Observe next that

$$\frac{C_n}{C} = e^{S-S_n}$$

and

$$S - S_n = \sum_{k>n} s_k < \frac{1}{2}\sum_{k>n}\left(\ln\left(1+\frac{1}{2k-2}\right) - \frac{1}{2}\ln\left(1+\frac{1}{2k}\right)\right)$$

$$= \frac{1}{2}\ln\left(1+\frac{1}{2n}\right) = \ln\left(1+\frac{1}{2n}\right)^{\frac{1}{2}}$$

Hence

$$\frac{C_n}{C} < \left(1+\frac{1}{2k}\right)^{\frac{1}{2}} < 1+\frac{1}{4n} \Rightarrow C_n < C\left(1+\frac{1}{4n}\right).$$

We deduce

$$C\sqrt{n}\left(\frac{n}{e}\right)^n < n! = C_n\sqrt{n}\left(\frac{n}{e}\right)^n < C\left(1+\frac{1}{4n}\right)\sqrt{n}\left(\frac{n}{e}\right)^n. \tag{9.6.30}$$

It remains to determine the constant C. We set

$$P_n := \sqrt{n}\left(\frac{n}{e}\right)^n.$$

From (9.6.30) we deduce

$$n! \sim CP_n \;\text{ as }\; n\to\infty. \tag{9.6.31}$$

To obtain C from the above equality we rely on Wallis' formula (9.6.17) which states that

$$\frac{\pi}{2} = \lim_{n\to\infty}\frac{2^2 4^2\cdots(2n)^2}{1^2 3^2\cdots(2n-1)^2}\cdot\frac{1}{2n}.$$

Now observe that

$$\frac{2^2 4^2\cdots(2n)^2}{1^2 3^2\cdots(2n-1)^2}\cdot\frac{1}{2n} = \frac{(n!)^2 2^{2n}}{1^2 3^2\cdots(2n-1)^2}\cdot\frac{1}{2n} = \frac{(n!)^4 2^{4n}}{(\,(2n)!\,)^2(2n)}.$$

Hence

$$\sqrt{\frac{\pi}{2}} = \lim_{n\to\infty}\frac{(n!)^2 2^{2n}}{(2n)!\sqrt{2n}},$$

i.e.,

$$\sqrt{\pi} = \lim_{n\to\infty}\frac{(n!)^2 2^{2n}}{(2n)!\sqrt{n}} = \lim_{n\to\infty}\frac{C^2P_n^2 2^{2n}}{CP_{2n}\sqrt{n}}\cdot\frac{\left(\frac{n!}{CP_n}\right)^2 2^{2n}}{\frac{(2n)!}{CP_{2n}}} \overset{(9.6.31)}{=} \lim_{n\to\infty}\frac{C^2P_n^2 2^{2n}}{CP_{2n}\sqrt{n}} = C\lim_{n\to\infty}\frac{P_n^2 2^{2n}}{P_{2n}\sqrt{n}}.$$

Now observe that

$$P_n = \sqrt{n}\left(\frac{n}{e}\right)^n \Rightarrow P_n^2 = \frac{n^{2n+1}}{e^{2n}},\ \ P_{2n} = \sqrt{2n}\frac{(2n)^{2n}}{e^{2n}} = 2^{2n}\sqrt{2n}\frac{n^{2n}}{e^{2n}},$$

and thus

$$\frac{P_n^2 2^{2n}}{P_{2n}\sqrt{n}} = \frac{2^{2n}\frac{n^{2n+1}}{e^{2n}}}{2^{2n}\sqrt{2n}\cdot\frac{n^{2n}}{e^{2n}}\cdot\sqrt{n}} = \frac{1}{\sqrt{2}}.$$

Hence

$$\sqrt{\pi} = C\lim_{n\to\infty}\frac{P_{2n}\sqrt{n}}{2^{2n}P_n^2} = \frac{C}{\sqrt{2}} \Rightarrow C = \sqrt{2\pi}.$$

The inequalities (9.6.30) with $C = \sqrt{2\pi}$ are precisely the inequalities (9.6.25) that we wanted to prove. □

9.7. Improper integrals

The Riemann integral is an operation defined for certain *bounded* functions defined on *bounded* intervals. Sometimes, even when one or both of these boundedness requirements are violated we can still give a meaning to an integral. Before we proceed with rigorous definitions it is helpful to look at some guiding examples.

Example 9.7.1. (a) Let $\alpha \in (0,1)$ and consider the function

$$f:(0,1]\to\mathbb{R},\ \ f(x) = \frac{1}{x^\alpha}.$$

This function is continuous on $(0,1]$, but it is not bounded on this interval because

$$\lim_{x\to 0+}\frac{1}{x^\alpha} = \infty.$$

It is however continuous on any compact interval $[\varepsilon, 1]$ and so it is Riemann integrable on such an interval. Note that

$$\int_\varepsilon^1 x^{-\alpha}dx = \frac{x^{1-\alpha}}{1-\alpha}\bigg|_\varepsilon^1 = \frac{1}{1-\alpha}(1-\varepsilon^{1-\alpha}).$$

Since $1-\alpha > 0$ we deduce that $\varepsilon^{1-\alpha}\to 0$ as $\varepsilon\searrow 0$ and thus

$$\lim_{\varepsilon\searrow 0}\int_\varepsilon^1 x^{-\alpha}dx = \frac{1}{1-\alpha}.$$

We can define the *improper Riemann* integral of $x^{-\alpha}$ over $[0,1]$ to be

$$\int_0^1 x^{-\alpha}dx := \lim_{\varepsilon\searrow 0}\int_\varepsilon^1 x^{-\alpha}dx = \frac{1}{1-\alpha}.$$

(b) Let $p>1$ and consider the function $g:[1,\infty)\to\mathbb{R}$, $g(x)=\frac{1}{x^p}$. The function g is bounded

$$0 < g(x) \leqslant 1,\ \ \forall x\geqslant 1$$

but it is defined on the unbounded interval $[1,\infty)$. It is integrable on any interval $[1,L]$ and we have

$$\int_1^L x^{-p}dx = \frac{x^{1-p}}{1-p}\Big|_1^L = \frac{1}{1-p}(L^{1-p}-1).$$

Since $1-p<0$ we deduce that $L^{1-p}\to 0$ as $L\to\infty$ and thus

$$\lim_{L\to\infty}\int_1^L x^{-p}dx = -\frac{1}{1-p} = \frac{1}{p-1}.$$

We define the *improper Riemann* integral of x^{-p} over $[1,\infty)$ to be

$$\int_1^\infty x^{-p}dx := \lim_{L\to\infty}\int_1^L x^{-p}dx = \frac{1}{p-1}.$$ □

The above examples gave meaning to integrals of *functions that are not defined on compact intervals*. Such integrals are called *improper*.

Definition 9.7.2 (Improper integrals)**.** (a) Let $-\infty < a < \omega \leqslant \infty$. Given a function $f:[a,\omega)\to\mathbb{R}$ we say that the *improper integral*

$$\int_a^\omega f(x)\,dx$$

is *convergent* if

- the restriction of f to any interval $[a,x]\subset[a,\omega)$ is Riemann integrable and,
- the limit
$$\lim_{x\nearrow\omega}\int_a^x f(t)\,dt$$
exists and it is finite.

When these happen we set

$$\int_a^\omega f(x)dx := \lim_{x\nearrow\omega}\int_a^x f(t)\,dt.$$

(b) Let $-\infty\leqslant\omega<b<\infty$. Given a function $f:(\omega,b]\to\mathbb{R}$ we say that the *improper integral*

$$\int_\omega^b f(x)dx$$

is *convergent* if

- the restriction of f to any interval $[x,b] \subset (\omega, b]$ is Riemann integrable and
- the limit
$$\lim_{x \searrow \omega} \int_x^b f(t)\,dt$$
exists and it is finite.

When these happen we set
$$\int_\omega^b f(x)dx := \lim_{x \searrow \omega} \int_x^b f(t)\,dt. \qquad \square$$

Remark 9.7.3. (a) We can rephrase the conclusion of Example 9.7.1(a) by saying that the integral
$$\int_0^1 \frac{1}{x^\alpha} dx$$
is convergent if $\alpha \in (0,1)$. Example 9.7.1(b) shows that the integral
$$\int_1^\infty \frac{1}{x^p} dx$$
is convergent if $p > 1$.

(b) In the sequel, in order to keep the presentation within bearable limits, we will state and prove results only for the improper integrals of type (a) in Definition 9.7.2. These involve functions that have a "problem" at the *upper* endpoint ω of their domain: either that endpoint is infinite, or the function "explodes" as x approaches ω.

These results have obvious counterparts for the integrals of type (b) in Definition 9.7.2 that involve functions that have a "problem" at the *lower* endpoint of their domain. Their statements and proofs closely mimic the corresponding ones for type (a) integrals. □

Example 9.7.4. For any $a, b \in \mathbb{R}$, $a < b$, the improper integrals
$$\int_a^b \frac{1}{(x-a)^\alpha} dx, \quad \int_a^b \frac{1}{(b-x)^\alpha} dx$$
are convergent for $\alpha < 1$ and divergent if $\alpha \geqslant 1$. Indeed, if $\alpha \neq 1$ we have
$$\int_{a+\varepsilon}^b \frac{1}{(x-a)^\alpha} dx = \frac{1}{1-\alpha}(x-a)^{1-\alpha}\Big|_{x=a+\varepsilon}^{a=b} = \frac{1}{1-\alpha}\big((b-a)^{1-\alpha} - \varepsilon^{1-\alpha}\big),$$
If $\alpha = 1$ we have
$$\int_{a+\varepsilon}^b \frac{1}{(x-a)} dx = \ln(x-a)\Big|_{x=a+\varepsilon}^{x=b} = \ln(b-a) - \ln\varepsilon.$$
These computations show that
$$\lim_{\varepsilon \searrow 0} \int_{a+\varepsilon}^b \frac{1}{(x-a)^\alpha} dx = \begin{cases} \frac{1}{1-\alpha}(b-a)^{1-\alpha}, & \alpha < 1, \\ \infty, & \alpha \geqslant 1. \end{cases}$$

The convergence of the integral

$$\int_a^b \frac{1}{(b-x)^\alpha}dx$$

is analyzed in a similar fashion.

(b) The integral

$$\int_1^\infty \frac{1}{x^p}dx,\ \ p \in \mathbb{R}.$$

is convergent for $p > 1$ and divergent if $p \leqslant 1$.

Indeed, if $p \neq 1$, then

$$\int_1^L x^{-p}dx = \frac{1}{1-p}x^{1-p}\Big|_{x=1}^{x=L} = \frac{1}{1-p}\big(L^{1-p}-1\big).$$

Now observe that

$$\lim_{L\to\infty} L^{1-p} = \begin{cases} 0, & p > 1,\\ \infty, & p < 1.\end{cases}$$

When $p = 1$, we have

$$\int_1^L \frac{1}{x}dx = \ln L \to \infty \ \text{ as } \ L \to \infty.$$

Similarly, the integral

$$\int_{-\infty}^{-1} \frac{1}{|x|^p}dx$$

converges for $p > 1$ and diverges for $p \leqslant 1$. □

We have the following immediate result whose proof is left to you as an exercise.

Proposition 9.7.5. *Let $-\infty < a < \omega \leqslant \infty$ and $f_1, f_2 : [a,\omega) \to \mathbb{R}$ be functions that are Riemann integrable on each of the intervals $[a,x]$, $x \in (a,\omega)$.*

(a) If $t_1, t_2 \in \mathbb{R}$, and the improper integrals

$$\int_a^\omega f_i(x)dx,\ \ i = 1,2$$

are convergent, then the integral

$$\int_a^\omega \big(t_1 f_1(x) + t_2 f_2(x)\big)dx$$

is convergent, and

$$\int_a^\omega \big(t_1 f_1(x) + t_2 f_2(x)\big)dx = t_1\int_a^\omega f_1(x)dx + t_2\int_a^\omega f_2(x)dx.$$

(b) Let $b \in (a,\omega)$. The improper integral

$$\int_a^\omega f_1(x)dx$$

is convergent if and only if the improper integral

$$\int_b^\omega f_1(x)dx$$

is convergent. Moreover, when these integrals are convergent we have

$$\int_a^\omega f_1(x)dx = \int_a^b f_1(x)dx + \int_b^\omega f_1(x)dx. \tag{9.7.1}$$

□

Theorem 9.7.6 (Cauchy). *Let $-\infty < a < \omega \leqslant \infty$ and suppose that $f : [a, \omega) \to \mathbb{R}$ is a function which is Riemann integrable on each of the intervals $[a, x] \subset [a, \omega)$. Then the following statements are equivalent.*

(i) *The integral $\int_a^\omega f(t)dt$ is convergent.*

(ii) *For any $\varepsilon > 0$ there exists $c = c(\varepsilon) \in (a, \omega)$ such that*

$$\forall x, y : \;\; x, y \in (c(\varepsilon), \omega) \Rightarrow \left| \int_x^y f(t)dt \right| < \varepsilon.$$

Proof. We set

$$I(x) := \int_a^x f(t)dt, \;\; \forall x \in [a, \omega).$$

(i) ⇒ (ii). We know that the limit

$$I_\omega := \lim_{x \to \omega} I(x)$$

exists and it is finite. Let $\varepsilon > 0$. There exists $c = c(\varepsilon) \in [a, \omega)$ such that

$$\forall x, y : \;\; x, y \in (c, \omega) \Rightarrow |I(x) - I_\omega| < \frac{\varepsilon}{2} \text{ and } |I(y) - I_\omega| < \frac{\varepsilon}{2}.$$

Observe that for any $x, y \in (c, \omega)$ we have

$$\left| \int_x^y f(t)dt \right| = \big| I(y) - I(x) \big| \leqslant \big| I(y) - I_\omega \big| + \big| I_\omega - I(x) \big| < \varepsilon.$$

This proves (ii).

(ii) ⇒ (i). We know that for any $\varepsilon > 0$ there exists $c = c(\varepsilon) \in [a, \omega)$ such that

$$\forall x < y : \;\; x, y \in (\, c(\varepsilon), \omega \,) \Rightarrow \left| \int_x^y f(t)dt \right| < \frac{\varepsilon}{2}. \tag{9.7.2}$$

Choose a sequence (x_n) in $[a, \omega)$ such that

$$\lim_n x_n = \omega.$$

We deduce that for any $\varepsilon > 0$ there exists $N = N(\varepsilon)$ such that

$$\forall n : \;\; n > N(\varepsilon) \Rightarrow x_n \in (\, c(\varepsilon), \omega).$$

Hence, for any $m, n > N(\varepsilon)$ we have

$$|I(x_m) - I(x_n)| < \frac{\varepsilon}{2} < \varepsilon, \;\; \forall m, n > N(\varepsilon) \tag{9.7.3}$$

proving that the sequence $(I(x_n))$ is Cauchy, thus convergent. Set

$$J := \lim_n I(x_n).$$

We will show that

$$\lim_{x \to \omega} I(x) = J.$$

Letting $m \to \infty$ in (9.7.3) we deduce that for any $\varepsilon > 0$ and any $n > N(\varepsilon)$ we have

$$x_n \in (\, c(\varepsilon), \omega) \;\text{ and }\; |J - I(x_n)| \leqslant \frac{\varepsilon}{2}. \tag{9.7.4}$$

Let $x \in (\, c(\varepsilon), \omega)$ and $n > N(\varepsilon/2)$. Then $x, x_n \in (\, c(\varepsilon), \omega)$ and (9.7.2) implies that

$$|I(x_n) - I(x)| < \frac{\varepsilon}{2} \tag{9.7.5}$$

We deduce

$$|I(x) - J| \leqslant |\, I(x) - I(x_n) \,| + |\, I(x_n) - J \,| \overset{(9.7.4),(9.7.5)}{<} \varepsilon, \;\; \forall x \in (c(\varepsilon), \omega).$$

This proves (i). □

Corollary 9.7.7 (Comparison Principle). *Let $-\infty < a < \omega \leqslant \infty$ and suppose that $f, g : [a, \omega) \to \mathbb{R}$ are two real functions satisfying the following properties.*

(i) *For any $x \in [a, \omega)$ the restrictions of f, g to $[a, x]$ are Riemann integrable.*

(ii) *$\exists b \in [a, \omega)$, such that $0 \leqslant f(x) \leqslant g(x)$, $\forall x \in [b, \omega)$.*

Then

$$\int_a^\omega g(x)dx \;\; \textit{is convergent} \;\Rightarrow\; \int_a^\omega f(x)dx \;\; \textit{is convergent.}$$

Proof. Since the improper integral

$$\int_a^\omega g(x)dx$$

is convergent we deduce from Proposition 9.7.5(b) that the integral

$$\int_b^\omega g(x)dx$$

is also convergent. Theorem 9.7.6 shows that for any $\varepsilon > 0$ there exists $c(\varepsilon) \in [b, \omega)$ such that

$$\forall x < y : \;\; x, y \in (c(\varepsilon), \omega) \Rightarrow \int_x^y g(t)dt = \left| \int_x^y g(t)dt \right| < \varepsilon.$$

Using the assumption (i) we deduce that

$$\forall x < y: \quad x,y \in (c(\varepsilon),\omega) \Rightarrow \left|\int_x^y f(t)dt\right| = \int_x^y f(t)dt \leqslant \int_x^y g(t)dt.$$

We can invoke Theorem 9.7.6 to conclude that the integral

$$\int_b^\omega f(x)dx$$

is convergent. Proposition 9.7.5(b) now implies that

$$\int_a^\omega f(x)dx$$

is convergent. □

Remark 9.7.8. Using the logical tautology

$$p \Rightarrow q \longleftrightarrow \neg q \Rightarrow \neg p,$$

we see that if f and g are as in Corollary 9.7.7, then

$$\int_a^\omega f(x)dx \;\text{ is divergent } \Rightarrow \int_a^\omega g(x)dx \;\text{ is divergent.}$$ □

Corollary 9.7.9. *Let $-\infty < a < \omega \leqslant \infty$ and suppose that $f,g:[a,\omega)\to\mathbb{R}$ are two real functions satisfying the following properties.*

(i) $\exists b \in [a,\omega)$, *such that* $f(x) \geqslant 0$ *and* $g(x) > 0$, $\forall x \in [b,\omega)$.

(ii) *There exists $C \geqslant 0$ such that*

$$\lim_{x\to\omega}\frac{f(x)}{g(x)} = C.$$

(iii) *For any $x \in [a,\omega)$ the restrictions of f and g to $[a,x]$ are Riemann integrable.*

Then

$$\int_a^\omega g(x)dx \;\textit{ is convergent } \Rightarrow \int_a^\omega f(x)dx \;\textit{ is convergent.}$$

Proof. The integral

$$\int_a^\omega (C+1)g(x)dx$$

is convergent.

The assumption (ii) implies that there exists $b_0 \in (b,\omega)$ such that

$$f(x) < (C+1)g(x), \quad \forall x \in (b_0,\omega).$$

We can now invoke Corollary 9.7.7 to reach the desired conclusion. □

Example 9.7.10. (a) Consider the continuous function

$$f:[1,\infty)\to\mathbb{R},\quad f(x)=\frac{x+2}{4x^3+3x^2+2x+1}$$

Note that $f(x)\geqslant 0$ for any $x\in[1,\infty)$. To decide the convergence of the integral

$$\int_1^\infty f(x)dx$$

we compare $f(x)$ with the function $g:[1,\infty)\to\mathbb{R}$, $g(x)=\frac{1}{x^2}$. Observe that

$$\frac{f(x)}{g(x)}=\frac{x^3+2x^2}{4x^3+3x^2+2x+1}\to\frac{1}{4}\ \text{ as } x\to\infty$$

Since

$$\int_1^\infty \frac{1}{x^2}dx$$

is convergent we deduce from Corollary 9.7.9 that the integral

$$\int_1^\infty f(x)dx$$

is also convergent.

(b) Consider the function

$$f:(0,1]\to\mathbb{R},\quad f(x)=\frac{\sin\sqrt{x}}{x}.$$

Note that

$$\lim_{x\searrow 0} f(x)=\lim_{x\searrow 0}\frac{\sin\sqrt{x}}{\sqrt{x}}\frac{1}{\sqrt{x}}=\infty.$$

In particular, $f(x)>0$ for $x>0$ small. Since

$$\frac{f(x)}{\frac{1}{\sqrt{x}}}=\frac{\sin\sqrt{x}}{\sqrt{x}}\to 1\ \text{ as } x\searrow 0$$

and the improper integral

$$\int_0^1\frac{1}{\sqrt{x}}dx$$

is convergent, we deduce from Corollary 9.7.9 that the improper integral $\int_0^1 f(x)dx$ is also convergent.

(c) Consider the function $f:[0,\infty)\to\mathbb{R}$, $f(x)=xe^{-x^2}$. Note that $f(x)\geqslant 0$, $\forall x$ and

$$\frac{f(x)}{\frac{1}{x^2}}=x^3e^{-x^2}=\frac{x^2}{e^{x^2}}\to 0\ \text{ as } x\to\infty.$$

Thus the integral

$$\int_0^\infty xe^{-x^2}dx$$

is convergent. To evaluate this integral we begin by evaluating the integrals

$$\int_0^L xe^{-x^2}dx$$

where $L \to \infty$. We use the change in variables $u = x^2$ so that $du = 2xdx$

$$x = 0 \Rightarrow u = 0, \;\; x = L \Rightarrow u = L^2$$

and we deduce

$$\int_0^L xe^{-x^2}dx = \frac{1}{2}\int_0^L e^{-x^2}(2xdx) = \frac{1}{2}\int_0^{L^2} e^{-u}du = \frac{1}{2}\Big(-e^{-u}\Big)\Big|_{u=0}^{u=L^2} = \frac{1}{2}(1-e^{-L^2}).$$

Now observe that

$$\lim_{L\to\infty}\frac{1}{2}(1-e^{-L^2}) = \frac{1}{2},$$

so that

$$\int_0^\infty xe^{-x^2}dx = \frac{1}{2}.$$

So far we have investigated improper integrals of function that had a problem at ω, one of the endpoints of its domain: either $\omega = \infty$, or the function “explodes” as it approaches ω. Sometime we need to deal with functions that have problems at both endpoints of its domain. The next example explains how to proceed in this case.

(d) Consider the function

$$f : (-1, 1) \to \mathbb{R}, \;\; f(x) = \frac{1}{\sqrt{(1-x^2)}}.$$

To decide the convergence of the integral

$$\int_{-1}^1 f(x)dx,$$

we must first locate the sources of the possible problems. We note that $f(x)$ “explodes” as $x \to \pm 1$, i.e.,

$$\lim_{x\to\pm 1} f(x) = \infty.$$

We split the integral into two parts,

$$I_{-1} = \int_{-1}^0 f(x)dx, \;\; I_1 = \int_0^1 f(x)dx.$$

Each of the above integrals has only one problem point and, if both integrals are convergent, then the original integral will be convergent if and only if both integrals above are convergent and, when this happens, we have

$$\int_{-1}^1 f(x)dx = \int_{-1}^0 f(x)dx + \int_0^1 f(x)dx.$$

Now observe that

$$f(x) = \frac{1}{\sqrt{(1-x)(1+x)}}.$$

The term $(1-x)$ is responsible for the bad behavior near $x=1$, while the term $(1+x)$ is responsible for the bad behavior near $x=-1$.

From Example 9.7.4 we deduce that both integrals

$$\int_{-1}^{0}\frac{1}{\sqrt{1+x}}dx,\;\;\int_{0}^{1}\frac{1}{\sqrt{1-x}}$$

are convergent. Observe next that

$$\lim_{x\to-1}\frac{f(x)}{\frac{1}{\sqrt{1+x}}}=\lim_{x\to-1}\frac{\frac{1}{\sqrt{(1-x)(1+x)}}}{\frac{1}{\sqrt{1+x}}}=\lim_{x\to-1}\frac{1}{\sqrt{1-x}}=\frac{1}{\sqrt{2}},$$

$$\lim_{x\to 1}\frac{f(x)}{\frac{1}{\sqrt{1-x}}}=\lim_{x\to 1}\frac{\frac{1}{\sqrt{(1-x)(1+x)}}}{\frac{1}{\sqrt{1-x}}}=\lim_{x\to 1}\frac{1}{\sqrt{1+x}}=\frac{1}{\sqrt{2}}.$$

Using Corollary 9.7.9 we now deduce that both integrals $I_{\pm 1}$ are convergent. In particular, we deduce that the improper integral

$$\int_{-1}^{1}f(x)dx$$

is convergent. We can actually compute it. Let $-1<a<0<b<1$. We have

$$\int_a^b\frac{1}{\sqrt{1-x^2}}dx=\arcsin x\Big|_{x=a}^{x=b}=\arcsin b-\arcsin a.$$

Note that

$$\lim_{b\nearrow 1}\arcsin b=\arcsin 1=\frac{\pi}{2},\;\;\lim_{a\searrow -1}\arcsin a=\arcsin(-1)=-\frac{\pi}{2}$$

so that

$$\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}dx=\frac{\pi}{2}-\Big(-\frac{\pi}{2}\Big)=\pi. \tag{9.7.6}$$

Definition 9.7.11. Let $-\infty<a<\omega\leqslant\infty$ and $f:[a,\omega)\to\mathbb{R}$ a function that is Riemann integrable on any interval $[a,x]$, $x\in(a,\omega)$. We say that the improper integral

$$\int_a^{\omega}f(x)dx$$

is *absolutely convergent* if the improper integral

$$\int_a^{\omega}|f(x)|dx$$

is convergent. □

The next result is very similar to Theorem 4.6.13.

Proposition 9.7.12. *Let $-\infty < a < \omega \leqslant \infty$ and $f : [a,\omega) \to \mathbb{R}$ a function that is Riemann integrable on any interval $[a,x]$, $x \in (a,\omega)$. Then*

$$\int_a^\omega f(x)dx \;\; \textit{absolutely convergent} \;\Rightarrow\; \int_a^\omega f(x)dx \;\; \textit{convergent.}$$

Proof. We rely on Cauchy's Theorem 9.7.6. Since the integral

$$\int_a^\omega |f(x)|dx$$

is convergent we deduce from Cauchy's theorem that for any $\varepsilon > 0$ there exists $c(\varepsilon) \in (a,\omega)$ such that

$$\forall x,y;\;\; x,y \in (c(\varepsilon),\omega) \Rightarrow \left|\int_x^y |f(t))|dt\right| < \varepsilon.$$

On the other hand, (9.5.1) shows that

$$\left|\int_x^y f(t)dt\right| \leqslant \left|\int_x^y |f(t)|dt\right|$$

and we deduce that

$$\forall x,y,\;\; x,y \in (c(\varepsilon),\omega) \Rightarrow \left|\int_x^y f(t))dt\right| < \varepsilon.$$

Cauchy's theorem now implies that

$$\int_a^\omega f(x)dx$$

is convergent. □

The comparison principle Corollary 9.7.7 yields a comparison principle involving absolute convergence.

Corollary 9.7.13 (Comparison Principle)**.** *Let $-\infty < a < \omega \leqslant \infty$ and suppose that $f,g : [a,\omega) \to \mathbb{R}$ are two real functions satisfying the following properties.*

(i) $\exists b \in [a,\omega)$, *such that* $|f(x)| \leqslant |g(x)|$, $\forall x \in [b,\omega)$.

(ii) *For any $x \in [a,\omega)$ the restrictions of f,g to $[a,x]$ are Riemann integrable.*

Then

$$\int_a^\omega g(x)dx \;\; \textit{is absolutely convergent} \;\Rightarrow\; \int_a^\omega f(x)dx \;\; \textit{is absolutely convergent.}$$ □

Example 9.7.14. Consider the function

$$f : [1,\infty) \to \mathbb{R},\;\; f(x) = \frac{\sin x}{x^2}.$$

Note that

$$|f(x)| \leqslant \frac{1}{x^2}, \ \ \forall x \geqslant 1$$

and since $\int_1^\infty \frac{1}{x^2}dx$ is convergent we deduce that $\int_a^\infty f(x)dx$ is absolutely convergent. □

9.7.1. Euler's Gamma function. For every $x > 0$ we set

$$\boxed{\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt}. \tag{9.7.7}$$

For each fixed $x > 0$ this improper integral is convergent. To see this we split the above integral into two parts

$$I_0 = \int_0^1 t^{x-1}e^{-t}dt, \ \ I_\infty = \int_1^\infty t^{x-1}e^{-t}dt.$$

To prove the convergence of I_0 we observe that

$$0 < t^{x-1}e^{-t} \leqslant t^{x-1} \ \ \forall t \in (0,1].$$

Since $x - 1 > -1$ the improper integral

$$\int_0^1 t^{x-1}dt$$

is convergent. The Comparison Principle then implies that I_0 is also convergent.

To prove the convergence of I_∞ we observe that and as $t \to \infty$ the function $t^{x-1}e^{-t}$ decays to zero faster, than any power t^{-n}, $n \in \mathbb{N}$. In particular

$$\lim_{t\to\infty} \frac{t^{x-1}e^{-t}}{t^{-2}}dt = 0.$$

Since the integral

$$\int_1^\infty t^{-2}dt$$

is convergent we deduce from the Comparison Principle that I_∞ is convergent as well.

The resulting function

$$(0,\infty) \ni x \mapsto \Gamma(x) \in (0,\infty)$$

is called *Euler's Gamma function*

Observe that

$$\Gamma(1) = \int_0^\infty e^{-t}dt = \left(-e^{-t}\right)\Big|_{t=0}^{t=\infty} = 1, \tag{9.7.8}$$

and, for $x > 0$,

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}dt = -\int_0^\infty t^x d(e^{-t}) = -\underbrace{\left(t^x e^{-t}\right)\Big|_{t=0}^{\infty}}_{=0} + x\underbrace{\int_0^\infty t^{x-1}e^{-t}dt}_{=\Gamma(x)} = x\Gamma(x).$$

so that

$$\boxed{\Gamma(x+1) = x\Gamma(x), \ \ \forall x > 0.} \tag{9.7.9}$$

From (9.7.8) and (9.7.9) we deduce inductively

$$\Gamma(2)=1\Gamma(1)=1,\;\;\Gamma(3)=2\Gamma(2)=2,\;\;\Gamma(4)=3\Gamma(3)=3\cdot 2=3!,\dots,$$

$$\boxed{\Gamma(n)=(n-1)!,\;\;\forall n\in\mathbb{N}.} \tag{9.7.10}$$

Fix $\lambda>0$. In the definition

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

we make the change of variables $t=\lambda s$ we deduce

$$\Gamma(x)=\int_0^\infty \lambda^{x-1}s^{x-1}e^{-\lambda s}\lambda ds=\lambda^x\int_0^\infty s^{x-1}e^{-\lambda s}ds,$$

so that

$$\boxed{\frac{\Gamma(x)}{\lambda^x}=\int_0^\infty s^{x-1}e^{-\lambda s}ds,\;\;\forall x,\lambda>0.} \tag{9.7.11}$$

9.8. Length, area and volume

The concept of integral is involved in the definition of important geometric quantities such length, area and volume. Their definition in the most general context is quite involved and we restrict ourselves to special cases that still have a wide range of applications.

9.8.1. Length. We will define the length of special curves in the plane, namely the curves defined by the graphs of differentiable functions.

Definition 9.8.1. Suppose that $-\infty\leqslant a<b\leqslant\infty$ and $f:(a,b)\to\mathbb{R}$ is a C^1-function. We say that its graph has *finite length* if the integral

$$\int_a^b\sqrt{1+f'(x)^2}dx$$

is convergent. The value of this integral is then declared to be the *length of the graph* Γ_f *of* f. We write this

$$\operatorname{length}(\Gamma_f)\int_a^b\sqrt{1+f'(x)^2}dx. \tag{9.8.1}$$

□

Here is the intuition behind the definition. If we are located at the point $(x_0,y_0)=(x_0,f(x_0))$ on the graph of f and we move a tiny bit, from x_0 to x_0+dx, then the rise, that is the change in altitude is

$$dy=\frac{dy}{dx}\cdot dx=f'(x_0)dx.$$

The Pythagorean theorem then shows that the distance covered along the graph is approximately

$$\sqrt{dx^2+dy^2}=\sqrt{dx^2+f'(x_0)^2dx^2}=\sqrt{1+f'(x_0)^2}\,dx.$$

The total distance traveled along the graph, i.e., the length of the trip is obtained by summing all these infinitesimal distances

$$\int_a^b \sqrt{1+f'(x)^2}\,dx.$$

The next examples support the validity of the proposed formula for the length.

Example 9.8.2. Consider two points in the plane, P_1 with coordinates (x_1, y_1) and P_2 with coordinates (x_2, y_2). Assume moreover that $x_1 < x_2$; see Figure 9.8. We want to compute the length $|P_1P_2|$ of the line segment connecting P_1 to P_2.

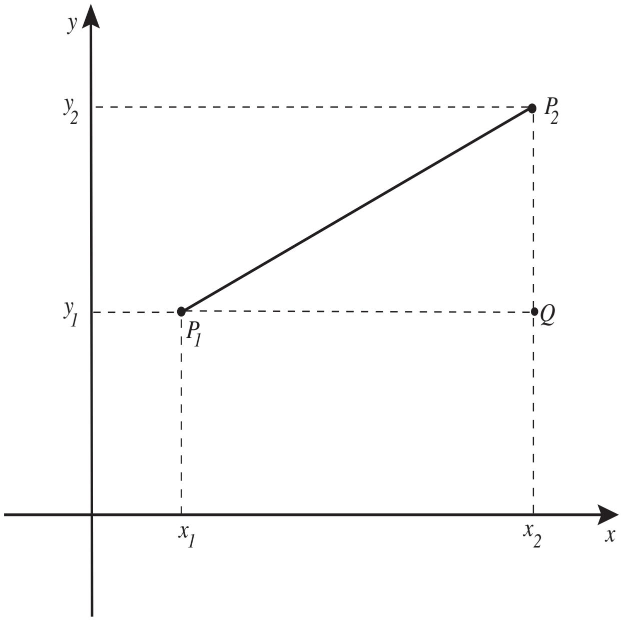

Figure 9.8. *Computing the length of a line segment.*

Pythagoras' theorem shows that

$$|P_1P_2|^2 = |P_1Q|^2 + |QP_2|^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2. \tag{9.8.2}$$

Let us show that the formula proposed in Definition 9.8.1 yields the same result.

The line determined by the points P_1, P_2 has slope

$$m := \frac{y_2 - y_1}{x_2 - x_1},$$

and thus it is described by the equation

$$y = m(x - x_1) + y_1.$$

In other words, the line segment is the graph of the linear function

$$f : [x_1, x_2] \to \mathbb{R}, \quad f(x) = m(x - x_1) + y_1.$$

Note that $f'(x) = m$, $\forall x \in [x_1, x_2]$, and according to Definition 9.8.1, we have

$$|P_1P_2| = \int_{x_1}^{x_2} \sqrt{1 + f'(x)^2}dx = \int_{x_1}^{x_2} \sqrt{1+m^2}dx = \sqrt{1+m^2}(x_2 - x_1).$$

Hence

$$|P_1P_2|^2 = (1+m^2)(x_2-x_1)^2 = \left(1 + \frac{(y_2-y_1)^2}{(x_2-x_1)^2}\right)(x_2-x_1)^2 = (x_2-x_1)^2 + (y_2-y_1)^2.$$

This agrees with the Pythagorean prediction (9.8.2). □

Example 9.8.3. Consider the function

$$f : (-1,1) \to \mathbb{R}, \quad f(x) = \sqrt{1-x^2}.$$

The graph of this function is the upper half-circle of radius 1 centered at the origin; see Figure 9.9. Indeed, a point (x, y) on this circle satisfies

$$x^2 + y^2 = 1, \quad y \geqslant 0 \Longleftrightarrow y = \sqrt{1-x^2}.$$

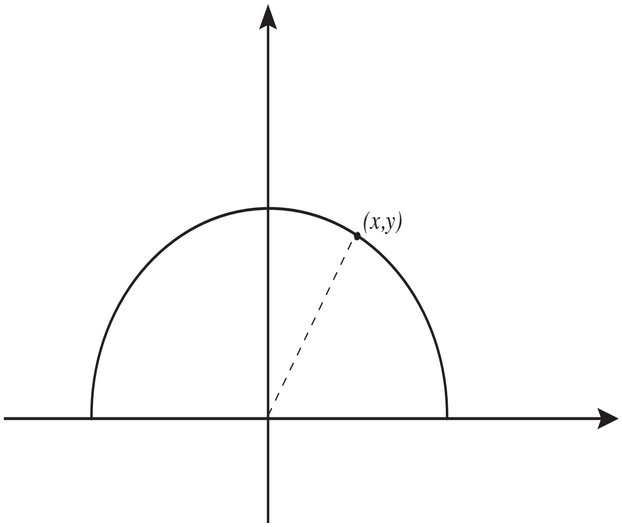

Figure 9.9. *Computing the length of a half-circle.*

The function $f(x)$ is differentiable on $(-1, 1)$ and we have

$$f'(x) = -\frac{x}{\sqrt{1-x^2}}, \quad 1 + f'(x)^2 = 1 + \frac{x^2}{1-x^2} = \frac{1}{1-x^2}, \quad \forall x \in (-1,1). \tag{9.8.3}$$

Hence the length of this semi-circle is

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\, dx \overset{(9.7.6)}{=} \pi.$$ □

We can define the length of more complicated curves.

Definition 9.8.4. Let $-\infty < a < b \leqslant \infty$. A *continuous* function $(a,b) \to \mathbb{R}$ is called *piecewise* C^1 if there exist points $x_1, \dots, x_n \in (a,b)$ such that

$$a < x_1 < x_2 < \cdots < x_n < b$$

and the function f is C^1 on each of the subintervals

$$(a, x_1), \ \ (x_1, x_2), \dots, (x_n, b).$$

The length of its graph is then given by

$$\int_a^{x_1} \sqrt{1 + f'(x)^2}\, dx + \int_{x_1}^{x_2} \sqrt{1 + f'(x)^2}\, dx + \cdots + \int_{x_n}^{b} \sqrt{1 + f'(x)^2}\, dx.$$

Above, some of the integrals could be improper and for the length to be finite these integrals have to be convergent. □

9.8.2. Area. A region D of the Cartesian plane $\mathbb{R}^2$ is said to be of *simple type* with respect to the x-axis if there exists an interval I and functions

$$F, C : I \to \mathbb{R}$$

such that

$$F(x) \leqslant C(x), \ \ \forall x \in I,$$

and

$$(x,y) \in D \iff x \in I \ \wedge \ F(x) \leqslant y \leqslant C(x).$$

The function F is called the *floor* of the region D, while the function C is called the *ceiling* of the region; see Figure 9.10

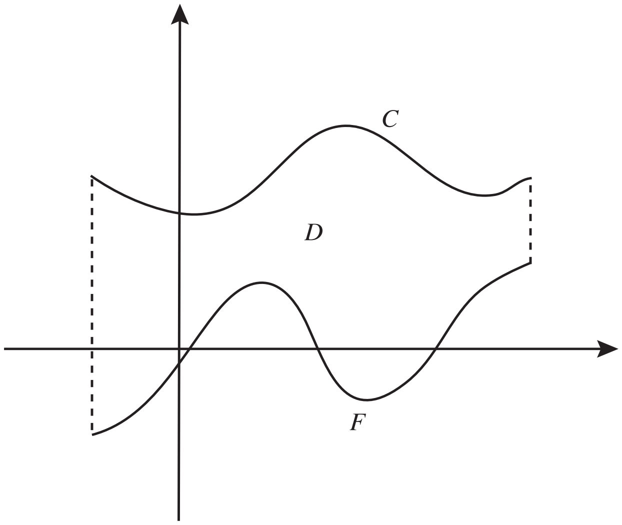

Figure 9.10. *A planar region of simple type with respect to the x-axis.*

The *area* of the region D is given by the improper integral

$$\operatorname{Area}(D) := \int_I \big(C(x) - F(x) \big)\, dx,$$

whenever this integral is well defined[3] and convergent.

A region D of the cartesian plane $\mathbb{R}^2$ is said to be of *simple type* with respect to the y-axis if there exists an interval J and functions

$$L, R : J \to \mathbb{R}$$

such that

$$L(y) \leqslant R(y), \quad \forall y \in J$$

and

$$(x, y) \in R \Longleftrightarrow y \in J \ \wedge \ L(y) \leqslant x \leqslant R(y).$$

The function L is called the *left wall* of the region D, while the function R is called the *right wall* of the region; see Figure 9.11.

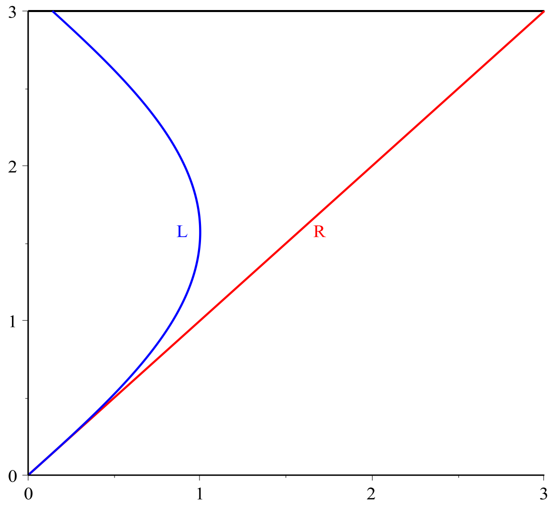

Figure 9.11. *A planar region of simple type with respect to the y-axis,* $\sin y \leqslant x \leqslant y$, $0 \leqslant y \leqslant 3$.

The *area* of the region D is given by the improper integral

$$\operatorname{Area}(D) := \int_J \big(R(y) - L(y) \big)\, dy,$$

whenever this integral is well defined

[3] The integral is well defined if the function $C(x) - F(x)$ is Riemann integrable on any compact interval $[\alpha, \beta] \subset (a, b)$.

Remark 9.8.5. (a) We swept under the rug a rather subtle fact. A region in the plane can be simultaneously simple type with respect to the x-axis, and simple type with respect to the y-axis. In such situations there are two possible ways of computing the area and they'd better produce the same result. This is indeed the case, but the proof in general is quite complicated, and the best approach relies on the concept of multiple integrals.

To see that this is not merely a theoretical possibility, consider the region (see Figure 9.12)

$$R = \big\{ (x,y) \in \mathbb{R}^2;\ \ x \in [0,1],\ \ x^2 \leqslant y \leqslant x \big\}.$$

The above description shows that R is a region of simple type with respect to the x-axis. However, R can be given the alternate description as a region of simple type with respect to the y-axis,

$$R = \big\{ (x,t) \in \mathbb{R}^2;\ \ y \in [0,1],\ \ y \leqslant x \leqslant \sqrt{y} \big\}.$$

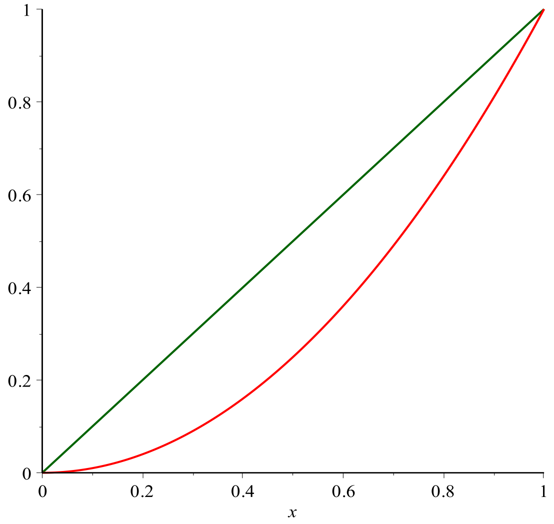

Figure 9.12. *A planar region that simple type with respect to both axes:* $x^2 \leqslant y \leqslant x$, $0 \leqslant x \leqslant 1$.

If we use the first description we deduce

$$\operatorname{Area}(R) = \int_0^1 (x - x^2)dx = \Big(\frac{x^2}{2} - \frac{x^3}{3}\Big)\Big|_{x=0}^{x=1} = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

If we use the second description we deduce

$$\operatorname{Area}(R) = \int_0^1 (\sqrt{y} - y)dy = \Big(\frac{2x^{3/2}}{3} - \frac{x^2}{2}\Big)\Big|_{x=0}^{x=1} = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}.$$

Many regions in the plane decompose into finitely many simple type regions that have overlaps only along boundary curves. For such a region, the area is defined as the sum of the areas of the simple-type sub-regions it decomposes into. This raises an even trickier question: why is the answer independent of the procedure we use to decompose the region into simple-type sub-regions? To answer this question one needs the full apparatus of multiple integrals.

(b) Let us observe that a simple-type region can have finite area, even if it is unbounded. Consider for example the region between the x-axis and the graph of the function

$$g : [0, \infty) \to \mathbb{R}, \quad g(x) = e^{-x}.$$

The area of this region is

$$\int_0^\infty e^{-x} dx = \big(-e^{-x}\big)\Big|_0^\infty = -e^{-\infty} - (-1) = 0 + 1 = 1. \qquad \square$$

9.8.3. Solids of revolution. Suppose that we are given an open interval (a, b) and a function

$$g : (a, b) \to \mathbb{R}$$

called a *generatrix* such that $g(x) \geqslant 0$, $\forall x \in (a, b)$. If we rotate the graph of g about the x-axis we get a surface of revolution Σ_g that surrounds a solid of revolution S_g; see Figure 9.13.

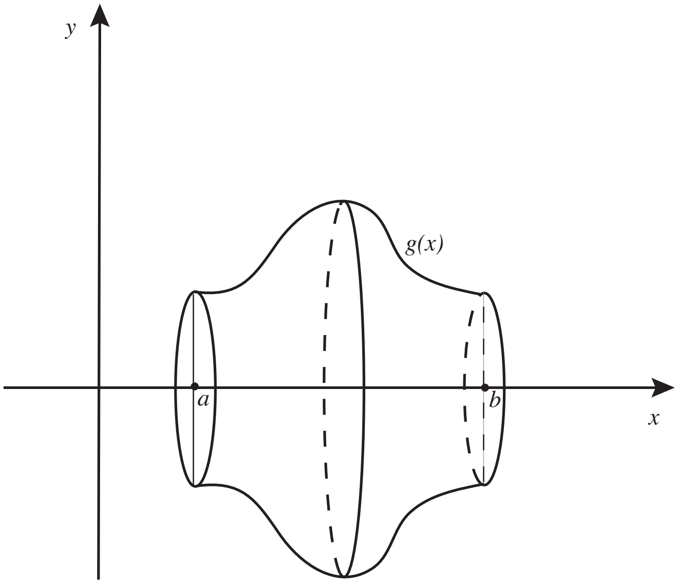

Figure 9.13. *A surface of revolution.*

The *area* of the surface of revolution Σ_g is given by the improper integral

$$\boxed{\operatorname{area}(\Sigma_g) := 2\pi \int_a^b g(x)\sqrt{1 + g'(x)^2} dx}, \tag{9.8.4}$$

whenever the integral is well defined. The *volume* of the solid of revolution S_g is given by the improper integral

$$\boxed{\operatorname{vol}(S_g) := \pi \int_a^b g(x)^2 dx}, \tag{9.8.5}$$

whenever the integral is well defined.

Example 9.8.6. (a) Suppose that the generatrix is the function $g : (-1,1) \to \mathbb{R}$, $g(x) = \sqrt{1-x^2}$. Its graph is the upper half-circle of radius 1 depicted in Figure 9.9. When we rotate this half-circle about the x-axis, the surface of revolution obtained is a sphere Σ_g of radius 1 that surrounds a solid ball S_g of radius 1.

The computations in (9.8.3) show that

$$\sqrt{1+g'(x)^2} = \frac{1}{\sqrt{1-x^2}}$$

so that

$$g(x)\sqrt{1+g'(x)^2} = 1.$$

We deduce that the area of the unit sphere is

$$2\pi \int_{-1}^{1} g(x)\sqrt{1+g'(x)^2} dx = 2\pi \in_{-1}^{1} dx = 4\pi.$$

The volume of the unit ball is

$$\pi \int_{-1}^{1} g(x)^2 dx = \pi \int_{-1}^{1} (1-x^2) dx = \pi \left(x\Big|_{-1}^{1} - \frac{x^3}{3}\Big|_{-1}^{1} \right) = \pi\Big(2 - \frac{2}{3}\Big) = \frac{4\pi}{3}.$$

These equalities confirm the classical formulæ taught in elementary solid geometry.

(b)

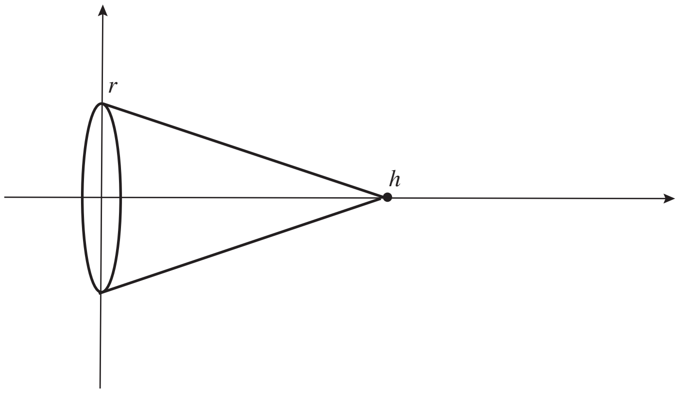

Figure 9.14. *A cone.*

Consider the cone depicted in Figure 9.14. It is obtained by rotating a line segment about the x-axis, more precisely, the line segment connecting the point $(0,r)$ on the y-axis with the point $(h,0)$ on the x-axis. Here $h, r > 0$.

This line segment lies on the line with slope $m = -r/h$ and y-intercept r. In other words, this line is given by the equation

$$g(x) = -\frac{r}{h}x + r.$$

Observe that

$$g'(x) = -\frac{r}{h}, \quad \sqrt{1+g'(x)^2} = \frac{\sqrt{h^2+r^2}}{h},$$
$$g(x)\sqrt{1+g'(x)^2} = \frac{h^2+r^2}{h}\Big(-\frac{r}{h}x+r\Big).$$

We deduce that the area of this cone (excluding its base) is

$$\begin{aligned} 2\pi\int_0^h \frac{\sqrt{h^2+r^2}}{h}\Big(-\frac{r}{h}x+r\Big)dx &= 2\pi\frac{\sqrt{h^2+r^2}}{h}\int_0^h\Big(-\frac{r}{h}x+r\Big)dx \\ &= 2\pi r\frac{\sqrt{h^2+r^2}}{h}\int_0^h dx - 2\pi r\frac{\sqrt{h^2+r^2}}{h^2}\int_0^h xdx \\ &= 2\pi r\sqrt{h^2+r^2} - \pi r\sqrt{h^2+r^2} = \pi r\sqrt{h^2+r^2}. \end{aligned}$$

This agrees with the known formulæ in solid geometry.

The volume of the cone is

$$\pi\int_0^h\Big(-\frac{r}{h}x+r\Big)^2dx = \frac{\pi r^2}{h^2}\int_0^h(h-x)^2dx = \frac{\pi r^2}{h^2}\times\frac{h^3}{3} = \frac{\pi r^2 h}{3}.$$

(c) Let $\alpha \in (\frac{1}{2}, 1)$ and consider the function

$$g : [1,\infty) \to \mathbb{R}, \quad g(x) = \frac{1}{x^\alpha}.$$

The surface of revolution obtained by rotating the graph of g about the x-axis has the bugle shape in Figure 9.15

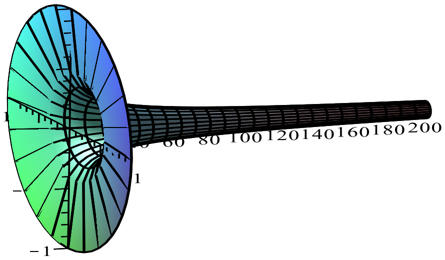

Figure 9.15. *An infinite bugle.*

The volume of this bugle is

$$\pi \int_1^\infty g(x)^2 dx = \pi \int_1^\infty \frac{1}{x^{2\alpha}} dx.$$

Since $2\alpha > 1$, the above integral is convergent and in fact

$$\pi \int_1^\infty \frac{1}{x^{2\alpha}} dx = \frac{\pi}{2\alpha - 1}.$$

On the other hand, the area of the bugle is

$$\begin{aligned} 2\pi \lim_{L\to\infty} \int_1^L g(x)\sqrt{1+g'(x)^2} dx &\geqslant 2\pi \lim_{L\to\infty} \int_1^L g(x) dx \\ = 2\pi \lim_{L\to\infty} \int_1^L \frac{1}{x^\alpha} dx &= 2\pi \lim_{L\to\infty} \left(\frac{x^{1-\alpha}}{1-\alpha}\right)\Big|_1^L = \infty, \end{aligned}$$

because $\alpha < 1$. This is surprising: you need a finite amount of water to fill the bugle, but an infinite amount of paint if you want to paint it!!! □

9.9. Exercises

Exercise 9.1. Prove by induction the equality (9.1.3). □

Exercise 9.2. Consider the function $f : [0,4] \to \mathbb{R}$, $f(x) = x^2$, and the partition

$$\boldsymbol{P} = (0,\ 0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4)$$

of the interval $[0,4]$.

(a) Find the mesh size $\|\boldsymbol{P}\|$ of $\boldsymbol{P}$.

(b) Compute the Riemann sum $\boldsymbol{S}(f,\boldsymbol{P},\underline{\xi})$ when the sample $\underline{\xi}$ consists of the right endpoints of the subintervals of $\boldsymbol{P}$. □

Exercise 9.3. (a) Suppose that $f,g : [a,b] \to \mathbb{R}$ are two functions. Prove that for any sampled partition $(\boldsymbol{P},\underline{\xi})$ of $[a,b]$ and for any real numbers α,β we have

$$\boldsymbol{S}\big(\alpha f + \beta g, \boldsymbol{P}, \underline{\xi}\big) = \alpha \boldsymbol{S}\big(f, \boldsymbol{P}, \underline{\xi}\big) + \beta \boldsymbol{S}\big(g, \boldsymbol{P}, \underline{\xi}\big).$$ □

(b) Let $f : [a,b] \to \mathbb{R}$. Prove that the following statements are equivalent.

(i) The function f is *not* Riemann integrable.

(ii) There exists ε_0 such that, for any $n \in \mathbb{N}$ there exist sampled partitions $(\boldsymbol{P}_n,\underline{\xi}^n)$ and $(\boldsymbol{Q}_n,\underline{\zeta}^n)$ satisfying

$$\|\boldsymbol{P}_n\|,\ \|\boldsymbol{Q}_n\| < \frac{1}{n} \ \text{ and } \ \big|\, \boldsymbol{S}(f,\boldsymbol{P}_n,\underline{\xi}^n) - \boldsymbol{S}(f,\boldsymbol{Q}_n,\underline{\zeta}^n) \,\big| > \varepsilon_0.$$

Hint. For the implication $(i) \Rightarrow (ii)$ choose partitions $\boldsymbol{P}_n$, $\boldsymbol{Q}_n$ such that $\|\boldsymbol{P}_n\|$, $\|\boldsymbol{Q}_n\| < \frac{1}{n}$ and $S_*(f,\boldsymbol{P}_n) \to \boldsymbol{S}_*(f)$ and $\boldsymbol{S}(f,\boldsymbol{Q}_n) \to S^*(f)$. For the implication (ii) ⇒ (i) use the Riemann-Darboux theorem and the equality (9.3.5). □

Exercise 9.4. Consider the function $f : [-2,2] \to \mathbb{R}$, $f(x) = x^2$ and the partition

$$\boldsymbol{P} = (-2,\ -1.5,\ -1\ ,-0.5,\ 0,\ 0.5,\ 1,\ 1.5,\ 2)$$

of $[-2,2]$.

(a) Compute the upper and lower Darboux sums $\boldsymbol{S}^*(f,\boldsymbol{P})$, $\boldsymbol{S}_*(f,\boldsymbol{P})$.

(b) Compute $\omega(f,\boldsymbol{P})$. □

Exercise 9.5. Suppose that $f : [0,1] \to \mathbb{R}$ is a C^1-function, i.e., it is differentiable on $[0,1]$ and the derivative is continuous. We set

$$M := \sup_{x\in[0,1]} |f'(x)|.$$

(a) Suppose that $I \subset [0,1]$ is an interval of length δ. Show that

$$\operatorname{osc}(f,I) \leqslant M\delta.$$

(b) For $n \in \mathbb{N}$ we denote by $\boldsymbol{U}_n$ the uniform partition of order n of $[0,1]$. Show that

$$\omega(f,\boldsymbol{U}_n) \leqslant \frac{M}{n}, \quad \forall n \in \mathbb{N}.$$

(c) Fix $n \in \mathbb{N}$ and a sample $\underline{\xi}$ of $\boldsymbol{U}_n$. Show that

$$\left| \int_0^1 f(x)dx - \boldsymbol{S}(f, \boldsymbol{U}_n, \underline{\xi}) \right| \leqslant \frac{M}{n}.$$

□

Exercise 9.6. Let $a > 0$ and assume that $f : [-a, a] \to \mathbb{R}$ is a Riemann integrable function.

(a) Prove that if f is an *odd* function, i.e., $f(-x) = -f(x)$, $\forall x \in [-a, a]$, then

$$\int_{-a}^{a} f(x)dx = 0.$$

(b) Prove that if f is an *even* function, i.e., $f(-x) = f(x)$, $\forall x \in [-a, a]$, then

$$\int_{-a}^{a} f(x)dx = 2\int_0^a f(x)dx.$$

□

Exercise 9.7. (a) Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are two Lipschitz functions. Show that the composition $f \circ g$ is also Lipschitz.

(b) Suppose that the function $g : [a, b] \to \mathbb{R}$ is Riemann integrable and the function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz. Prove that $f \circ g$ is Riemann integrable.

(c) Suppose that the function $g : [a, b] \to \mathbb{R}$ is Riemann integrable and the function $f : \mathbb{R} \to \mathbb{R}$ is C^1, i.e., differentiable with continuous derivative. Prove that $f \circ g$ is Riemann integrable. □

Exercise 9.8. Suppose that the functions $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable. Let $p, q > 1$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

(a) Prove that the functions $|f|^p$ and $|g|^q$ are Riemann integrable.

(b) Prove that

$$\int_a^b |f(x)g(x)|dx \leqslant \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q dx\right)^{\frac{1}{q}}.$$

(c) Prove that

$$\left(\int_a^b |f(x) + g(x)|^p dx\right)^{\frac{1}{p}} \leqslant \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}} + \left(\int_a^b |g(x)|^p dx\right)^{\frac{1}{p}}.$$

Hint. Approximate the integrals by Riemann sums and then use the inequalities (8.3.14) and (8.3.17). □

Exercise 9.9. (a) Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function such that $f(x) \geqslant 0$, $\forall x \in [a, b]$. Prove that

$$\int_a^b f(x)dx = 0 \Longleftrightarrow f(x) = 0, \quad \forall x \in [a, b].$$

□

(b) Show that for any $a < b$ there exists a continuous function $u : \mathbb{R} \to \mathbb{R}$ such that $u(x) > 0$, $\forall x \in (a, b)$, and $u(x) = 0$ $\forall x \in \mathbb{R} \backslash (a, b)$.

Hint. Think of a function u whose graph looks like a roof.

(c) Suppose that $f : [0, 1] \to \mathbb{R}$ is a continuous function such that

$$\int_0^1 f(x)u(x)dx = 0,$$

for any continuous function $u : [0, 1] \to \mathbb{R}$ such that $u(0) = u(1) = 0$. Prove that $f(x) = 0$, $\forall x \in [0, 1]$.

Hint. Argue by contradiction. Suppose that there exists $x_0 \in [0, 1]$ such that $f(x_0) \neq 0$, say $f(x_0) > 0$. Reach a contradiction using Theorem 6.2.1, and the facts (a), (b) above. □

Exercise 9.10. Suppose that $f_n : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, is a sequence of Riemann integrable functions that converges *uniformly* on $[a, b]$ to the function $f : [a, b] \to \mathbb{R}$. We set

$$d_n := \sup_{x \in [a,b]} |f(x) - f_n(x)|.$$

(a) (Compare with Exercise 6.6.) Prove that

$$\lim_{n \to \infty} d_n = 0.$$

(b) Let $X \subset [a, b]$ be a nonempty subset of $[a, b]$. Prove that, for any $n \in \mathbb{N}$, we have

$$\operatorname{osc}(f, X) \leqslant \operatorname{osc}(f_n, X) + 2d_n.$$

(c) Prove that, for any partition $\boldsymbol{P}$ of $[a, b]$, and any $n \in \mathbb{N}$, we have

$$\omega(f, \boldsymbol{P}) \leqslant \omega(f_n, \boldsymbol{P}) + 2d_n(b - a).$$

(d) Prove that f is Riemann integrable and

$$\lim_{n \to \infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx.$$

□

Exercise 9.11. (a) Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous and convex function. Prove that

$$\frac{1}{b - a} \int_a^b f(x)dx \leqslant \frac{f(a) + f(b)}{2}.$$

(b) Use (a) to show that for any $x > y > 0$ we have

$$\frac{1}{2y} \ln \frac{x + y}{x - y} \leqslant \frac{x}{x^2 - y^2}.$$

□

Exercise 9.12. Consider the function $f : [0, 1] \to \mathbb{R}$, $f(x) = \frac{1}{x+1}$.

(a) Compute $\int_0^1 f(x)dx$.

(b) For $n \in \mathbb{N}$ we denote by $\boldsymbol{U}_n$ the uniform partition of order n of $[0,1]$ and by $\underline{\xi}^{(n)}$ the sample of $\boldsymbol{U}_n$ given by

$$\underline{\xi}_k^{(n)} = \frac{k}{n}, \;\; k = 1, \dots, n.$$

Describe explicitly the Riemann sum $\boldsymbol{S}(f, \boldsymbol{U}_n, \underline{\xi}^{(n)})$.

(c) Use parts (a) and (b) to compute the limit in Exercise 4.22. □

Exercise 9.13. Use Riemann sums for an appropriate Riemann integrable function to compute the limit

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n+1}} + \frac{1}{\sqrt{n+2}} + \cdots + \frac{1}{\sqrt{2n}} \right)$$ □

Exercise 9.14. Fix a natural number k.

(a) Prove that for any $n \in \mathbb{N}$ we have

$$1^k + 2^k + \cdots + (n-1)^k \leqslant \int_0^n x^k dx \leqslant 1^k + 2^k + \cdots + n^k.$$

(b) Use (a) to prove that

$$\lim_{n\to\infty} \frac{1^k + 2^k + \cdots + n^k}{n^{k+1}} = \frac{1}{k+1}.$$ □

Exercise 9.15. Consider the function

$$F : [0, \infty) \to \mathbb{R}, \;\; F(x) = \int_0^{\sqrt{x}} e^{\frac{t^2}{2}} dt.$$

Show that $F(x)$ is differentiable on $(0, \infty)$ and then compute $F'(x)$, $x > 0$. □

Exercise 9.16. Suppose $f_n : [a,b] \to \mathbb{R}$, $n \in \mathbb{N}$, is a sequence of C^1-functions with the following properties.

(i) The sequence of derivatives $f_n' : [a,b] \to \mathbb{R}$ converges *uniformly* to a function $g : [a,b] \to \mathbb{R}$.

(ii) The sequence $f_n : [a,b] \to \mathbb{R}$ converges *pointwisely* to a function $f : [a,b] \to \mathbb{R}$.

Prove that the following hold.

(a) The sequence $f_n : [a,b] \to \mathbb{R}$ converges *uniformly* to $f : [a,b] \to \mathbb{R}$.

Hint. Define $G : [a,b] \to \mathbb{R}$, $G(x) = f(a) + \int_a^x g(t)dt$. (The function g is continuous since it is a uniform limit of continuous functions.) Since f_n' is continuous, the Fundamental Theorem of Calculus shows that

$$f_n(x) = f_n(a) + \int_a^x f_n'(t)dt.$$

Then

$$f_n(x) - G(x) = f_n(a) - f(a) + \int_a^x \big(f_n'(t) - g(t) \big) dt.$$

Use the above equality to show that the sequence f_n converges uniformly on $[a,b]$ to G. Argue next that $G = f$.

(b) The function f is C^1 and $f' = g$, i.e., the sequence $f_n' : [a,b] \to \mathbb{R}$ converges uniformly to f'. □

Exercise 9.17. Let $L > 0$. Suppose that the power series with real coefficients

$$a_0 + a_1 x + a_2 x^2 + \cdots$$

converges absolutely for any $|x| < L$. For every $x \in (-L, L)$ we denote by $s(x)$ the sum of the above series.

(a) Show that the function $x \mapsto s(x)$ is continuous on $(-L, L)$ and, for any $R \in (0, L)$, we have

$$\int_0^R s(x)dx = a_0 R + \frac{a_1}{2}R^2 + \frac{a_2}{3}R^3 + \cdots .$$

Hint. Use the Exercises 6.8 and 9.10.

(b) Prove that the power series

$$a_1 + 2a_2 x + 3a_3 x^2 + \cdots$$

also converges absolutely for any $|x| < L$.

(c) Prove that $s(x)$ is differentiable on $(-L, L)$ and that

$$s'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \cdots , \quad \forall |x| < L.$$

Hint. Use the Exercises 6.8, 9.16. □

Exercise 9.18. Consider the power series

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots ,$$

and respectively,

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots .$$

(a) Prove that the above series converge absolutely for any $x \in \mathbb{R}$. Denote their sums by $a(x)$ and respectively $b(x)$.

(b) Show that the functions $a, b : \mathbb{R} \to \mathbb{R}$ are differentiable and satisfy the equalities

$$a'(x) = b(x), \quad b'(x) = -a(x).$$

Hint. Use Exercise 9.17.

(c) Show that $a(x)$ is the unique solution of the differential equation

$$a''(x) + a(x) = 0, \quad \forall x \in \mathbb{R}$$

satisfying the condition $a(0) = 0$, $a'(0) = 1$. (Compare with Exercise 7.16.) □

Exercise 9.19. Consider the function

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \frac{1}{1 + x^2}.$$

(a) Prove that

$$f(x) = \sum_{n=0}^{\infty} (-1)^n x^{2n}, \quad \forall |x| < 1.$$

(b) Conclude from (a) that the Taylor series of $f(x)$ at $x_0 = 0$ is

$$1 - x^2 + x^4 - x^6 + \cdots .$$

Hint. Use Exercise 9.17.

(c) Deduce from (a) that

$$\arctan x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots , \;\; \forall |x| < 1. \qquad \square$$

Exercise 9.20. (a) Suppose that $f, w : [a, b] \to \mathbb{R}$ are two continuous functions satisfying the following properties.

(i) The function f is continuous.

(ii) The function w is Riemann integrable and nonnegative, i.e., $w(x) \geqslant 0$, $\forall x \in [a, b]$.

(iii) The integral

$$W := \int_a^b w(x)dx$$

is strictly positive.

We set

$$m := \inf_{x \in [a,b]} f(x), \;\; M := \sup_{x \in [a,b]} f(x).$$

Show that

$$m \leqslant \frac{1}{W} \int_a^b f(x)w(x)dx \leqslant M,$$

and then conclude that there exists a point ξ in the *open* interval (a, b) such that

$$f(\xi) = \frac{1}{W} \int_a^b f(x)w(x)dx. \qquad \square$$

(b) Use the result in (a) to show that the Integral Remainder Formula (9.6.18) implies the Lagrange Remainder Formula (8.1.2). □

Exercise 9.21. Consider the function $f : [0, 2] \to \mathbb{R}$, $f(x) = 1 - |x - 1|$, $\forall x \in [0, 2]$.

(a) Sketch the graph of f.

(b) Compute $\int_0^2 f(x)dx$. □

Exercise 9.22. For any natural number n we define the n-th *Legendre polynomial* to be

$$P_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} \big(x^2 - 1\big)^n.$$

We set $P_0(x) = 1$, $\forall x$.

(a) Compute $P_1(x), P_2(x), P_3(x)$.

(b) Compute

$$\int_{-1}^{1} P_1(x)^2dx,\ \ \int_{-1}^{1} P_2(x)^2dx,\ \ \int_{-1}^{1} P_3(x)^2dx,\ \ \int_{-1}^{1} P_1(x)P_2(x)dx,$$

(c) Use integration-by-parts to compute

$$\int_{-1}^{1} P_m(x)P_n(x)dx,\ \ \int_{-1}^{1} P_n(x)^2dx,\ \ m,n\in\mathbb{N},\ \ m\neq n.$$

Hint. You may want to use the results in Exercise 7.6 and Example 9.6.7. □

Exercise 9.23. Fix an integer k. Use Stirling's formula (9.6.26) to compute

$$\lim_{n\to\infty}\frac{\sqrt{2n}}{2^{2n}}\binom{2n}{n+k},\ \ \lim_{n\to\infty}\frac{\sqrt{2n+1}}{2^{2n+1}}\binom{2n+1}{n+k}.$$ □

Exercise 9.24. (a) For any integer $n\geqslant 0$ compute the numbers

$$\int_0^1 \sin^2(2\pi nt)dt\ \ \int_0^1 \cos^2(2\pi nt)dt.$$

(b) Consider the function

$$f:[0,1]\to\mathbb{R},\ \ f(x)=\frac{1}{2}-\left|x-\frac{1}{2}\right|.$$

Sketch its graph and then compute

$$\int_0^1 f^2(x)dx.$$

(c) Let f be as above. For any integer $n\geqslant 0$ compute the numbers

$$a_n=\int_0^1 f(x)\cos(2\pi nx)dx,\ \ b_n=\int_0^1 f(x)\sin(2\pi nx)dx.$$

(d) With a_n, b_n as in (c) prove that the series

$$\sum_{n\geqslant 1}(a_n^2+b_n^2)$$

is convergent.[4]

Hint. When computing the above integrals it is convenient to use the change in variables $u=x-\frac{1}{2}$, some of the trig identities in Section 5.6 and Exercise 9.6. □

Exercise 9.25. Compute the area of the region depicted in Figure 9.11. □

[4] A nontrivial result in the theory of Fourier series shows that

$$\int_0^1 f^2(x)dx=a_0^2+2\sum_{n\geqslant 1}(a_n^2+b_n^2).$$

Exercise 9.26. Prove Proposition 9.7.5. □

Exercise 9.27. Consider the function

$$f : [0, 2] \to \mathbb{R}, \quad f(x) = \max\{2 - x, x^2\}.$$

(a) Sketch the graph of the function.

(b) Compute the area of the region between the x-axis and the graph of f.

(c) Show that the function f is piecewise C^1 and then compute the length of its graph. □

Exercise 9.28. Prove that for any $a \in (-1, 0)$ and any $b > 0$ the integrals

$$\int_0^1 t^a |\ln t|^b dt, \quad \int_1^\infty t^{a-1} |\ln t|^b dt$$

are convergent. □

Exercise 9.29. Prove that the Gamma function $\Gamma : (0, \infty) \to (0, \infty)$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

is continuous.

Hint. Fix $t > 0$ and then use Lagrange's mean value theorem for the function $f : (0, \infty) \to \mathbb{R}$, $f(x) = t^x$. Then use Exercise 9.28 to conclude. □

Exercise 9.30. Suppose that $f : [0, \infty) \to (0, \infty)$ is a decreasing function. Prove that the following statements are equivalent.

(i) The improper integral

$$\int_0^\infty f(x) dx$$

is convergent.

(ii) The series

$$f(0) + f(1) + f(2) + \cdots$$

is convergent.

□

9.10. Exercises for extra credit

Exercise* 9.1. Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function and $\Phi : \mathbb{R} \to \mathbb{R}$ is a convex continuous[5] function. Prove *Jensen's inequality*

$$\Phi\left(\frac{1}{b-a}\int_a^b f(x)dx\right) \leqslant \frac{1}{b-a}\int_a^b \Phi\big(f(x)\big)dx. \tag{9.10.1}$$

[5] The continuity assumption is redundant since any convex function $\mathbb{R} \to \mathbb{R}$ is automatically continuous.

□

Exercise* 9.2. Show that the improper integrals

$$\int_0^\infty \frac{\sin x}{x}dx, \quad \int_0^\infty \sin(x^2)dx$$

are convergent. □

Exercise* 9.3. Construct a continuous function $f : [0, \infty) \to \mathbb{R}$ satisfying the following properties.

(i) $f(x) \geqslant 0$, $\forall x \geqslant 0$.

(ii) $\sup_{x \geqslant 0} f(x) = \infty$.

(iii) The integral $\int_0^\infty f(x)dx$ is convergent.

□

Exercise* 9.4. Suppose that $f : [0, \infty) \to \mathbb{R}$ is a C^2-function satisfying the following conditions

(i) $f'(0) = 0$.

(ii)

$$\lim_{x\to\infty} \frac{1}{\ln x}\big(f(x) + f'(x) \big) = 0.$$

Prove that for any $\alpha \in (0, 1)$ the integral $\int_0^\infty \frac{f'(x)}{x^\alpha}dx$ is convergent. □

Exercise* 9.5. Suppose that $f : [1, \infty) \to \mathbb{R}$ is differentiable, the derivative $f' : [0, \infty) \to \mathbb{R}$ is increasing and

$$\lim_{x\to\infty} f'(x) = 0.$$

(For example $f(x) = \frac{1}{x}$ or $f(x) = \ln x$.) Prove that the sequence

$$S_n := \frac{1}{2}f(1) + f(2) + \cdots + f(n-1) + \frac{1}{2}f(n) - \int_1^n f(x)dx$$

is convergent and, if S is its limit, then for any $n \in \mathbb{N}$ we have

$$\frac{f'(n)}{n} < \frac{1}{2}f(1) + f(2) + \cdots + f(n-1) + \frac{1}{2}f(n) - \int_1^n f(x)dx - S < 0.$$ □

Exercise* 9.6. Suppose that $f : [1, \infty) \to \mathbb{R}$ is differentiable, the derivative $f' : [0, \infty) \to \mathbb{R}$ is increasing and

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

(For example, $f(x) = x^\alpha$, $\alpha > 1$.) Prove that there resists a constant $C > 0$ such for any $n \in \mathbb{N}$ we have

$$\left| \frac{1}{2}f(1) + f(2) + \cdots + f(n-1) + \frac{1}{2}f(n) - \int_1^n f(x)dx \right| \leqslant C|f'(n)|.$$ □

Exercise* 9.7. (a) Suppose that $f : [a,b] \to [0,\infty)$ is a Riemann integrable function. For any natural numbers $k \leqslant n$ we set

$$\delta_n := \frac{b-a}{n}, \;\; f_{n,k} = f\big(a + k\delta_n\big).$$

Prove that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f_{n,k} = \frac{1}{b-a}\int_a^b f(x)dx,$$

$$\lim_{n\to\infty}\left(\prod_{k=1}^{n} f_{n,k}\right)^{\frac{1}{n}} = \exp\left(\frac{1}{b-a}\int_a^b f(x)dx\right), \;\; \exp(x) := e^x.$$

(b) Fix real numbers $c, r > 0$. Denote by A_n, and respectively G_n, the arithmetic, and respectively geometric, mean of the numbers

$$c + r, c + 2r, \dots, c + nr.$$

Prove that

$$\lim_{n\to\infty}\frac{G_n}{A_n} = \frac{2}{e}.$$

□

Chapter 10

Complex numbers and some of their applications

10.1. The field of complex numbers

It is well known that there exists no real number x such that $x^2 = -1$ because $x^2 \geqslant 0 > -1$, $\forall x \in \mathbb{R}$. Following L. Euler, we introduce an imaginary number $\boldsymbol{i}$ with the property that

$$\boldsymbol{i}^2 = -1. \tag{10.1.1}$$

Sometimes we write $\boldsymbol{i} = \sqrt{-1}$. The number $\boldsymbol{i}$ is called the *imaginary unit.* This bold and somewhat arbitrary move raises some troubling questions.

Can we really do this? Yes, we just did, by fiat. Where does the "number" $\boldsymbol{i}$ come from? As its name suggests, it comes from our imagination. Can't we get into some sort of trouble? This vaguely formulated question is the more serious one, but let's just admit that we won't get in any trouble. This can be argued rigorously, but requires more advanced mathematics that did not even exist during Euler's time. It took more than a century to settle this issue. During that time mathematicians found convincing semi-rigorous arguments that this construction leads to no contradictions. As Euler and his followers, we will take it on faith that this construction won't lead us to shaky grounds.

What can we do with $\boldsymbol{i}$? Following Gauss, we define the *complex numbers.* These are quantities of the form

$$z := x + y\boldsymbol{i}, \;\; x, y \in \mathbb{R}.$$

The *real part* of the complex number z is

$$\mathbf{Re}\, z := x,$$

while its *imaginary part* is

$$\mathbf{Im}\, z := y.$$

The set of all the complex numbers is denoted by $\mathbb{C}$. .

The reason we are referring to the quantities $a+b\boldsymbol{i}$ as *numbers* is because we can operate with them, much like we do with real numbers. First, we can add complex numbers. If

$$z_1 := x_1 + y_1\boldsymbol{i},\ \ z_2 = x_2 + y_2\boldsymbol{i},$$

then we define

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2)\boldsymbol{i}.$$

This operation satisfies the same properties as the addition of real numbers, namely the Axioms 1-4 in Section 2.1. Note that the real numbers are special examples of complex numbers: they are the complex numbers whose imaginary part is zero.

We can also multiply complex numbers in a natural way, taking (10.1.1) into account. Thus

$$\begin{aligned}(x_1 + y_1\boldsymbol{i})(x_2 + y_2\boldsymbol{i}) &= x_1x_2 + x_1y_2\boldsymbol{i} + y_1x_2\boldsymbol{i} + y_1y_2\boldsymbol{i}^2\\ &= (x_1x_2 - y_1y_2) + (x_1y_2 + y_1x_2)\boldsymbol{i}.\end{aligned}$$

One can check that this multiplication is commutative, associative, and distributive with respect to the above addition operation. Moreover, the real number 1 acts as a multiplicative unit for this operation as well, and every nonzero real number z has an inverse. The construction of the inverse requires a bit of ingenuity.

To a complex number $z = x + y\boldsymbol{i}$ we associate its conjugate

$$\bar{z} = x - y\boldsymbol{i}.$$

Observe that

$$z\bar{z} = (x + y\boldsymbol{i})(x - y\boldsymbol{i}) = x^2 - (y\boldsymbol{i})^2 = x^2 + y^2.$$

The quantity $\sqrt{x^2 + y^2}$ is called the *norm* of the complex number z and it is denoted by $|z|$,

$$|z| := \sqrt{x^2 + y^2}.$$

Thus

$$\bar{z}z = z\bar{z} = |z|^2.$$

In particular, if $z \neq 0$, then $|z| \neq 0$ and we have

$$\frac{1}{|z|^2}\bar{z} \cdot z = z \cdot \frac{1}{|z|^2}\bar{z} = 1.$$

Thus

$$z^{-1} = \frac{1}{z} = \frac{\bar{z}}{|z|^2}. \tag{10.1.2}$$

The operation of conjugation interacts well with the operations of addition and multiplication introduced above. More precisely,

$$\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2,\ \ \overline{z_1z_2} = \bar{z}_1\bar{z}_2,\ \ \forall z_1, z_2 \in \mathbb{C}. \tag{10.1.3}$$

Moreover

$$|z_1 z_2| = |z_1| \cdot |z_2|, \quad \forall z_1, z_2 \in \mathbb{C}. \tag{10.1.4}$$

The simple proofs of these equalities are left to you as an exercise.

10.1.1. The geometric interpretation of complex numbers. The complex numbers have a very useful geometric interpretation. More precisely, we identify the complex number $z = x + y\boldsymbol{i}$ with the point $Z = (x, y)$ in the Cartesian plane $\mathbb{R}^2$. In turn we can identify the point Z with its position vector $\overrightarrow{OZ}$. For this reason we will often refer to $\mathbb{C}$ as the *complex plane*.

Given two complex numbers $z_1 = x_1 + y_1\boldsymbol{i}$, $z_2 = x_2 + y_2\boldsymbol{i}$ represented in the plane by the position vectors $\overrightarrow{OZ_1}$ and $\overrightarrow{OZ_2}$, then their sum $z_3 = (x_1+x_2)+(y_1+y_2)\boldsymbol{i}$ is represented in the plane by the point Z_3 with position vector

$$\overrightarrow{OZ_3} = \overrightarrow{OZ_1} + \overrightarrow{OZ_2},$$

where the addition of vectors is performed via the parallelogram rule; see Figure 10.1.

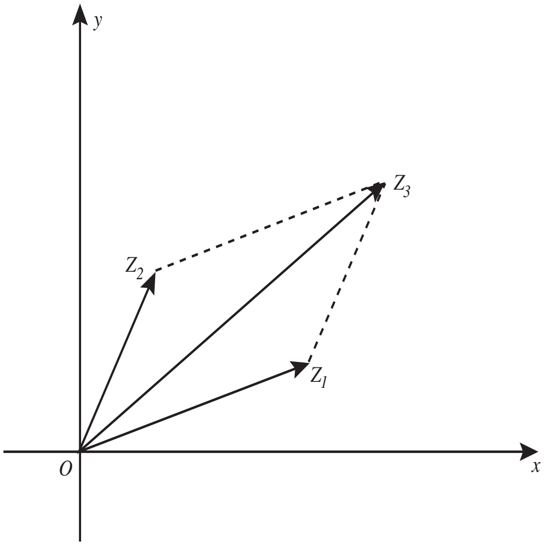

Figure 10.1. *The geometric interpretation of the sum of complex numbers.*

If the complex number $z = x + y\boldsymbol{i}$ is described by the point $Z = (x, y)$ in $\mathbb{R}^2$, then its conjugate $\bar{z} = x - y\boldsymbol{i}$ is represented by the point $Z^- = (x, -y)$, the reflection of Z in the x-axis; see Figure 10.2. Note that the norm $|z| = \sqrt{x^2 + y^2}$ is equal to the length of the vector $\overrightarrow{OZ}$,

$$|z| = \big|\overrightarrow{OZ}\big|.$$

The vector $\overrightarrow{OZ}$ makes an angle θ with the x-axis measured in a counterclockwise fashion, starting on the x-axis. Measured in radians, it can be any number in $[0, 2\pi)$. This angle is called the *argument* of the complex number z and it is denoted by $\arg z$.

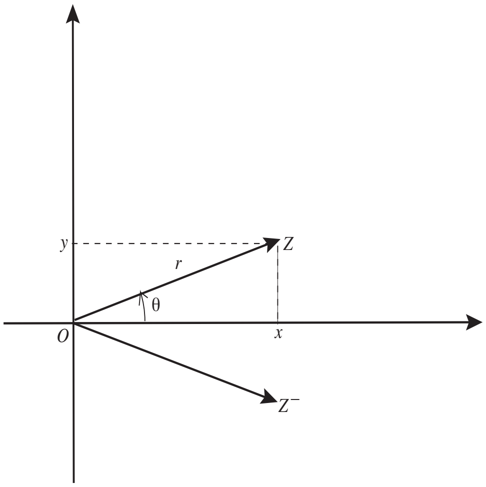

Figure 10.2. *The geometric interpretation of the conjugation of complex numbers.*

Denote by r the norm of z

$$r = |z| = \sqrt{x^2 + y^2}.$$

From Figure 10.2 we deduce that the coordinates (x, y) of Z can be expressed in terms of r and θ via the equalities

$$x = r\cos\theta, \quad y = r\sin\theta,$$

so that

$$z = r\cos\theta + r\sin\theta\boldsymbol{i} = r(\cos\theta + \boldsymbol{i}\sin\theta), \quad r = |z|, \quad \theta = \arg z. \tag{10.1.5}$$

The equality (10.1.5) is usually referred to as the *trigonometric representation* of the complex number $z = x + y\boldsymbol{i}$.

Suppose that we have two complex numbers z_1, z_2 with trigonometric representations

$$z_k = r_k(\cos\theta_k + \boldsymbol{i}\sin\theta_k), \quad r_k \geqslant 0, \quad k = 1, 2.$$

Then

$$\mathbf{Re}\, z_k = r_k\cos\theta_k, \quad \mathbf{Im}\, z_k = r_k\sin\theta_k.$$

Moreover

$$\begin{aligned} z_1z_2 &= (r_1r_2)(\cos\theta_1 + \boldsymbol{i}\sin\theta_1)(\cos\theta_2 + \boldsymbol{i}\sin\theta_2) \\ &= r_1r_2\Big\{\underbrace{\big(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2\big)}_{=\cos(\theta_1+\theta_2)} + \boldsymbol{i}\underbrace{\big(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2\big)}_{=\sin(\theta_1+\theta_2)}\Big\}. \end{aligned}$$

We have thus proved that

$$r_1(\cos\theta_1 + \boldsymbol{i}\sin\theta_1) \times r_2(\cos\theta_2 + \boldsymbol{i}\sin\theta_2) = r_1r_2\big(\cos(\theta_1+\theta_2) + \boldsymbol{i}\sin(\theta_1+\theta_2)\big). \tag{10.1.6}$$

Applying the above equality iteratively we obtain the celebrated *Moivre's formula*

$$\big(\cos\theta + \boldsymbol{i}\sin\theta\big)^n = \cos(n\theta) + \boldsymbol{i}\sin(n\theta), \quad \forall n \in \mathbb{N}, \quad \theta \in \mathbb{R}. \tag{10.1.7}$$

If we combine Moivre's formula with Newton's binomial formula we can obtain many interesting consequences. We have

$$\cos n\theta + \boldsymbol{i}\sin n\theta = \sum_{k=0}^{n}\binom{n}{k}\boldsymbol{i}^k(\cos\theta)^k(\sin\theta)^{n-k}.$$

Separating the real and imaginary parts in the right-hand side of the above equality taking into account that

$$\boldsymbol{i}^2 = -1,\ \ \boldsymbol{i}^3 = -\boldsymbol{i},\ \ \boldsymbol{i}^4 = 1,$$

we deduce

$$\cos n\theta = (\cos\theta)^n - \binom{n}{2}(\cos\theta)^{n-2}(\sin\theta)^2 + \binom{n}{4}(\cos\theta)^n(\sin\theta)^4 - \cdots \tag{10.1.8a}$$

$$\sin n\theta = \binom{n}{1}(\cos\theta)^{n-1}\sin\theta - \binom{n}{3}(\cos\theta)^{n-3}(\sin\theta)^3 + \cdots\,. \tag{10.1.8b}$$

For example,

$$\cos 2\theta = \cos^2\theta - \sin^2\theta,\ \ \sin 2\theta = 2\sin\theta\cos\theta,$$

$$\cos 3\theta = \cos^3\theta - 3\cos\theta\sin^2\theta,\ \ \sin 3\theta = 3\cos^2\theta\sin\theta - \sin^3\theta,$$

$$\cos 4\theta = \cos^4\theta - \binom{4}{2}\cos^2\theta\sin^2\theta + \sin^4\theta = \cos^4\theta - 6\cos^2\theta\sin^2\theta + \cos^4\theta,$$

$$\sin 4\theta = 4\cos^3\theta\sin\theta - 4\cos\theta\sin^3\theta.$$

Example 10.1.1. Consider the complex number

$$z = \cos\frac{\pi}{4} + \boldsymbol{i}\sin\frac{\pi}{4} = \frac{1}{\sqrt{2}}(1+\boldsymbol{i}).$$

For any $n \in \mathbb{N}$ we have

$$z^{8n} = \cos 2n\pi + \boldsymbol{i}\sin 2n\pi = 1.$$

On the other hand we have

$$z^{8n} = \frac{1}{2^{4n}}(1+\boldsymbol{i})^{8n}$$

so that

$$2^{4n} = (1+\boldsymbol{i})^{4n} = \sum_{k=0}^{8n}\binom{8n}{k}\boldsymbol{i}^k.$$

Isolating the real and imaginary parts in the right-hand side and equating them with the real and imaginary parts in the left-hand side we deduce

$$2^{4n} = \binom{8n}{0} - \binom{8n}{2} + \binom{8n}{4} - \cdots\,,$$

$$0 = \binom{8n}{1} - \binom{8n}{3} + \binom{8n}{5} - \cdots\,.$$

□

Example 10.1.2. Fix a natural number $n \geqslant 2$. Observe that the numbers

$$\zeta_k = \cos\Big(\frac{2\pi}{k}n\Big) + \boldsymbol{i}\sin\Big(\frac{2\pi}{n}\Big),\ \ k = 0, 1, \dots, n-1$$

satisfy the equation

$$\zeta_k^n = 1,\ \ \forall k.$$

Conversely, if z is a complex number such that $z^n = 1$, then we deduce

$$|z|^n = 1 \Rightarrow |z| = 1,$$

and thus there exists $\theta \in [0, 2\pi)$ such that

$$z = \cos\theta + \boldsymbol{i}\sin\theta.$$

Using Moivre's formula we deduce $\cos n\theta = 1$ and $\sin n\theta = 0$ which is possible if and only if $n\theta$ is a multiple of 2π. Thus θ can only be one of the numbers

$$\frac{2\pi k}{n},\ \ k = 0, 1, \dots, n-1.$$

In other words $z^n = 1$ if and only if z is equal to one of the numbers ζ_k. For this reason the numbers ζ_k are called *the n-th roots of unity.* □

10.2. Analytic properties of complex numbers

Most of the analysis we developed for real numbers carries over to complex numbers. The next result is crucial in this endeavor.

Proposition 10.2.1. *(a) For any complex numbers z_1, z_2 we have*

$$|z_1 + z_2| \leqslant |z_1| + |z_2|. \tag{10.2.1}$$

(b) if $z = x + y\boldsymbol{i} \in \mathbb{C}$ then

$$\frac{1}{2}(|x| + |y|) \leqslant |z| = \sqrt{x^2 + y^2} \leqslant |x| + |y|. \tag{10.2.2}$$

Proof. (a) Let

$$z_1 = x_1 + y_1\boldsymbol{i},\ \ z_2 = x_2 + y_2\boldsymbol{i}.$$

Then

$$|z_1| = \sqrt{x_1^2 + y_1^2},\ \ |z_2| = \sqrt{x_2^2 + y_2^2}.$$

The Cauchy-Schwarz inequality, Corollary 8.3.20, implies that

$$x_1x_2 + y_1y_2 \leqslant \left(\sqrt{x_1^2 + y_1^2}\right) \cdot \left(\sqrt{x_2^2 + y_2^2}\right) = |z_1| \cdot |z_2|.$$

We have

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2)\boldsymbol{i},$$
$$|z_1 + z_2|^2 = (x_1 + y_1)^2 + (x_2 + y_2)^2$$
$$= x_1^2 + y_1^2 + 2x_1y_1 + x_2^2 + y_2^2 + 2x_2y_2 = |z_1|^2 + |z_2|^2 + 2(x_1y_1 + 2x_2y_2)$$

$$\leqslant |z_1|^2 + |z_2|^2 + 2|z_1| \cdot |z_2| = (|z_1| + |z_2|)^2.$$

This proves (10.2.1).

(b) Observe that

$$(|x| + |y|)^2 = |x|^2 + |y|^2 + 2|x| \cdot |y| \geqslant |x|^2 + |y|^2 = x^2 + y^2.$$

This shows that

$$|x| + |y| \geqslant \sqrt{x^2 + y^2}.$$

On the other hand,

$$0 \leqslant (|x| - |y|)^2 = |x|^2 + |y|^2 - 2|xy| \Rightarrow 2|xy| \leqslant x^2 + y^2$$

$$\Rightarrow (|x| + |y|)^2 = |x|^2 + |y|^2 + 2|x| \cdot |y| \leqslant 2(x^2 + y^2) \Rightarrow \frac{1}{\sqrt{2}}(|x| + |y|) \leqslant \sqrt{x^2 + y^2}.$$

This proves (10.2.2). □

Definition 10.2.2. We define the distance between two complex numbers z_1, z_2 to be the nonnegative real number

$$\operatorname{dist}(z_1, z_2) := |z_1 - z_2|.$$ □

Corollary 10.2.3 (The triangle inequality)**.** *For any $z_1, z_2, z_3 \in \mathbb{C}$ we have*

$$\operatorname{dist}(z_1, z_3) \leqslant \operatorname{dist}(z_1, z_2) + \operatorname{dist}(z_2, z_3).$$

Proof. We have

$$\operatorname{dist}(z_1, z_3) = |z_1 - z_3| = |(z_1 - z_2) + (z_2 - z_3)|$$

$$\overset{(10.2.1)}{\leqslant} |z_1 - z_2| + |z_2 - z_3| = \operatorname{dist}(z_1, z_2) + \operatorname{dist}(z_2, z_3).$$

□

Definition 10.2.4. (a) Let $z_0 \in \mathbb{C}$ and $r > 0$. The *open disk* of center z_0 and radius r is the set

$$D_r(z_0) := \big\{ z \in \mathbb{C};\ \ \operatorname{dist}(z, z_0) < r \big\}.$$

(b) A subset $\mathcal{O} \subset \mathbb{C}$ is called *open* if for any $z_0 \in \mathcal{O}$ there exists $\varepsilon > 0$ such that

$$D_\varepsilon(z_0) \subset \mathcal{O}.$$ □

(c) A set $X \subset \mathbb{C}$ is called *closed* if the complement $\mathbb{C}\backslash X$ is open.

(d) A set $X \subset \mathbb{C}$ is called *bounded* if there exists $R > 0$ such that

$$X \subset D_R(0) \iff |z| < R,\ \ \forall z \in X.$$ □

Definition 10.2.5. (a) We say that a sequence of complex numbers $(z_n)_{n\geqslant 1}$ is *bounded* if the sequence of norms $(|z_n|)_{n\geqslant 1}$ is bounded as a sequence of real numbers.

(b) We say that a sequence of complex numbers $(z_n)_{n\geqslant 1}$ *converges to the complex number* z_*, and we denote this

$$\lim_n z_n = z_*,$$

if the sequence of nonnegative real numbers $\operatorname{dist}(z_n, z_*)$ converges to 0, i.e.,

$$\forall \varepsilon > 0 \;\; \exists N = N(\varepsilon) > 0 \;\text{ such that }\; \forall n \;\; (n > N(\varepsilon) \Rightarrow |z_n - z_*| < \varepsilon). \qquad \square$$

Proposition 10.2.6. *Suppose that $(z_n)_{n\geqslant 1}$ is a sequence of complex numbers. We set $x_n = \mathbf{Re}\, z_n$, $y_n = \mathbf{Im}\, z_n$. The following statements are equivalent.*

(i) *The sequence (z_n) converges to the complex number $z_* = x_* + y_* \boldsymbol{i}$.*

(ii) *The sequences of* real *numbers $(x_n)_{n\geqslant 1}$ and $(y_n)_{n\geqslant 1}$ converge to x_* and respectively y_*.*

Proof. (i) ⇒ (ii). From the first part of (10.2.2) we deduce that

$$\frac{1}{2}\left(|x_n - x_*| + |y_n - y_*|\right) \leqslant |z_n - z_*|.$$

Since $\lim_n z_n = z_*$ we deduce $\lim_n |z_n - z_*| = 0$ and the Squeezing Principle implies

$$\lim_n \left(|x_n - x_*| + |y_n - y_*|\right) = 0.$$

The last equality implies (ii).

(ii) ⇒ (i). From the second part of (10.2.2) we deduce that

$$|z_n - z_*| \leqslant |x_n - x_*| + |y_n - y_*|.$$

The assumption (ii) implies that

$$\lim_n \big(\, |x_n - x_*| + |y_n - y_*| \,\big) = 0.$$

From this we conclude that $\lim_n |z_n - z_*| = 0$, which is the statement (i). □

Corollary 10.2.7. *If the sequence of complex numbers $(z_n)_{n\geqslant 1}$ converges to z, then*

$$\lim_n |z_n| = |z|.$$

Proof. Let $x_n := \mathbf{Re}\, z_n$ and $y_n := \mathbf{Im}\, z_n$, $x = \mathbf{Re}\, z$, $y := \mathbf{Im}\, z$. Then

$$\lim_n z_n = z \Rightarrow \lim_n x_n = x \;\;\wedge\;\; \lim_n y_n = y$$

$$\Rightarrow \lim_n (x_n^2 + y_n^2) = x^2 + y^2 \Rightarrow \lim_n \sqrt{x_n^2 + y_n^2} = \sqrt{x^2 + y^2} \iff \lim_n |z_n| = |z|.$$

□

Corollary 10.2.8. *Any convergent sequence of complex numbers is bounded.*

Proof. Given a convergent sequence of complex numbers, the associated sequence of norms is convergent according to Corollary 10.2.7. The sequence of norms is thus a convergent sequence of *real* numbers, hence bounded according to Proposition 4.2.12. □

Example 10.2.9. Suppose z is a complex number such that $|z| < 1$. Then

$$\lim_n z^n = 0.$$

We have to show that the sequence of nonnegative numbers $|z^n|$ goes to zero as $n \to \infty$. We set $r : |z|$ and we observe that

$$|z^n| \overset{(10.1.4)}{=} |z|^n = r^n.$$

As shown in Example 4.2.10

$$|r| < 1 \Rightarrow \lim_n r^n = 0 \Rightarrow \lim_n z^n = 0. \qquad \square$$

The convergent sequences of complex numbers satisfy many of the same properties of convergent sequences of real numbers. We summarize these facts in our next result whose proof is left to you as an exercise.

Proposition 10.2.10 (Passage to the limit)**.** *Suppose that $(a_n)_{n\geqslant 1}$ and $(b_n)_{n\geqslant 1}$ are two convergent sequences of complex numbers,*

$$a := \lim_{n\to\infty} a_n, \;\; b = \lim_{n\to\infty} b_n.$$

The following hold.

(i) *The sequence $(a_n + b_n)_{n\geqslant 1}$ is convergent and*

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n = a + b.$$

(ii) *If $\lambda \in \mathbb{C}$ then*

$$\lim_{n\to\infty}(\lambda a_n) = \lambda \lim_{n\to\infty} a_n = \lambda a.$$

(iii)

$$\lim_{n\to\infty}(a_n \cdot b_n) = \big(\lim_{n\to\infty} a_n\big)\cdot\big(\lim_{n\to\infty} b_n\big) = ab.$$

(iv) *Suppose that $b \neq 0$. Then there exists $N_0 > 0$ such that $b_n \neq 0$, $\forall N > N_0$, and*

$$\lim_{n\to\infty}\frac{a_n}{b_n} = \frac{a}{b}. \qquad \square$$

Definition 10.2.11. A sequence of complex numbers $(z_n)_{n\geqslant 1}$ is called *Cauchy* if

$$\forall \varepsilon > 0 \;\; \exists N = N(\varepsilon) > 0 \;\text{ such that }\; \forall m, n \;\; (m, n > N(\varepsilon) \Rightarrow |z_m - z_n| < \varepsilon). \qquad \square$$

The concept of Cauchy sequence of complex numbers is closely related to the notion of Cauchy sequence of real numbers. We state this in a precise form in our next result. Its proof is very similar to the proof of Proposition 10.2.6 and we leave the details to you as an exercise.

Proposition 10.2.12. *Suppose that $(z_n)_{n\geqslant 1}$ is a sequence of complex numbers. We set $x_n := \mathbf{Re}\, z_n$, $y_n := \mathbf{Im}\, z_n$. The following statements are equivalent.*

(i) *The sequence $(z_n)_{n\geqslant 1}$ is Cauchy.*

(ii) *The sequences of* real *numbers $(x_n)_{n\geqslant 1}$ and $(y_n)_{n\geqslant 1}$ are Cauchy.*

□

Definition 10.2.13. The *series* associated to a sequence $(z_n)_{n\geqslant 0}$ of complex numbers is the new sequence $(s_n)_{n\geqslant 0}$ defined by the *partial sums*

$$s_0 = z_0,\ \ s_1 = z_0 + z_1,\ \ s_2 = z_0 + z_1 + z_2, \dots, s_n = \sum_{j=0}^{n} a_j\ \dots .$$

The series associated to the sequence $(z_n)_{n\geqslant 0}$ is denoted by the symbol

$$\sum_{n\geqslant 0}^{\infty} z_n \quad or \quad \sum_{n\geqslant 0} z_n$$

The series is called *convergent* if the sequence of partial sums $(s_n)_{n\geqslant 0}$ is convergent. The limit $\lim_{n\to\infty} s_n$ is called *the sum* of the series. We will use the notation

$$\sum_{n\geqslant 0} a_n = S$$

to indicate that the series is convergent and its sum is the real number S. □

Example 10.2.14. The geometric series

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots$$

is convergent for any complex number z of norm $|z| < 1$. Indeed, its n-th partial sum is

$$s_n = 1 + z + \cdots + z^n = \frac{1 - z^{n+1}}{1 - z}.$$

If $|z| < 1$, then we deduce from Example 10.2.9 and Proposition 10.2.10 that

$$\lim_n s_n = \lim_n \frac{1 - z^{n+1}}{1 - z} = \frac{1}{1 - z}.$$

This shows that the series is convergent and its sum is

$$1 + z + z^2 + \cdots + z^n + \cdots = \frac{1}{1 - z},\ \ \forall |z| < 1. \tag{10.2.3}$$

□

Proposition 10.2.15. *If the series of complex numbers*

$$\sum_{n\geqslant 0} z_n$$

is convergent, then its terms converge to zero, $\lim_n z_n = 0$.

Proof. Denote by s the sum of the series and by s_n its n-th partial sum,

$$s_n = z_0 + z_1 + \cdots + z_n.$$

Then $z_n = s_n - s_{n-1}$ and

$$\lim_n z_n = \lim_n (s_n - s_{n-1}) = \lim_n s_n - \lim_n s_{n-1} = s - s = 0.$$

□

Example 10.2.16. The geometric series

$$1 + z + z^2 + \cdots$$

is divergent if $|z| \geqslant 1$. Indeed, we have

$$|z^n| = |z|^n$$

and

$$\lim_n |z|^n = \begin{cases} 1, & |z| = 1, \\ \infty, & |z| > 1. \end{cases}$$

This shows that when $|z| \geqslant 1$ the sequence (z^n) does not converge to zero and thus, according to Proposition 10.2.15, the geometric series cannot be convergent. □

Definition 10.2.17. A series of complex numbers

$$\sum_{n\geqslant 0} z_n$$

is called *absolutely convergent* if the series of *nonnegative real numbers*

$$\sum_{n\geqslant 0} |z_n|$$

is convergent. □

Proposition 10.2.18. *If the series of complex numbers* $\sum_{n\geqslant 0} z_n$ *is absolutely convergent, then it is also convergent.*

Proof. We mimic the proof of Theorem 4.6.13. Denote by s_n the n-th partial sum of the series $\sum_{n\geqslant 0} z_n$ and by $\hat{s}_n$ the n-th partial sum of the series $\sum_{n\geqslant 0} |z_n|$,

$$s_n = z_0 + \cdots + z_n, \quad \hat{s}_n = |z_0| + \cdots + |z_n|.$$

For $n > m$ we have

$$s_n - s_m = z_{m+1} + \cdots + z_n, \quad \hat{s}_n - \hat{s}_m = |z_{m+1}| + \cdots + |z_n|$$

Using (10.2.1) we deduce

$$|s_n - s_m| = |z_{m+1} + \cdots + z_n| \leqslant |z_{m+1}| + \cdots + |z_n| = \hat{s}_n - \hat{s}_m = |\hat{s}_n - \hat{s}_m|. \tag{10.2.4}$$

Since the series $\sum_{n \geqslant 0} |z_n|$ is convergent we deduce that the sequence of partial sums $(\hat{s}_n)_{n \geqslant 0}$ is Cauchy. Hence, for any $\varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that for any $n > m > N(\varepsilon)$ we have

$$|\hat{s}_n - \hat{s}_m| < \varepsilon.$$

Using (10.2.4) we deduce that for any $n > m > N(\varepsilon)$ we have

$$|s_n - s_m| < \varepsilon.$$

This shows that the sequence (s_n) is Cauchy and thus convergent according to Proposition 10.2.12. □

The above result reduces the problem of deciding the absolute convergence of a series of complex numbers to deciding whether a series of nonnegative *real* numbers is convergent. We have investigated this issue in Section 4.6. We mention here one useful convergence test.

Corollary 10.2.19 (Ratio test)**.** *Suppose that*

$$z_0 + z_1 + z_2 + \cdots$$

is a series of complex numbers such that

$$L = \lim_n \frac{|z_{n+1}|}{|z_n|}$$

exists, $L \in [0, \infty]$. *Then the following hold.*

(i) *If* $L < 1$, *then the series* $\sum_{n \geqslant 0} z_n$ *is absolutely convergent.*

(ii) *If* $L > 1$, *then the series is divergent.*

Proof. (i) The ratio test Corollary 4.6.15 implies that the series of positive real numbers

$$\sum_{n \geqslant 0} |z_n|$$

is convergent.

(ii) If

$$\lim_n \frac{|z_{n+1}|}{|z_n|} > 1,$$

then $|z_{n+1}| > |z_n|$ for n sufficiently large. In particular, the sequence (z_n) does not converge to 0 and thus the series $\sum_{n \geqslant 0} z_n$ is divergent. □

10.3. Complex power series

A complex *power series* is a series of the form

$$s(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots = \sum_{n \geqslant 0} a_n z^n,$$

where z and the numbers $a_0, a_1, \dots$ are complex. The number z should be viewed as a quantity that is allowed to vary, while the numbers $a_0, a_1, \dots$ should be viewed as fixed quantities. As such they are called the *coefficients* of the power series. Note that for different choices of z we obtain different series.

Example 10.3.1. Consider for example the power series

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

The coefficients of this power series are

$$a_0 = 1, \;\; a_1 = -2, \;\; a_2 = 2^2, \dots, a_n = (-2)^n, \dots$$

Note that we can rewrite the above series as

$$s(z) = 1 + (-2z) + (-2z)^2 + (-2z)^3 + \cdots = \sum_{n \geqslant 0} (-2z)^n.$$

If we make the substitution $\zeta := -2z$ we can further rewrite

$$s(z) = 1 + \zeta + \zeta^2 + \cdots .$$

We know that this series is absolutely convergent for $|\zeta| > 1$ and divergent for $|\zeta| > 1$. In other words, the power series $s(z)$ converges absolutely if $|z| < \frac{1}{2}$ and diverges if $|z| > \frac{1}{2}$. Note that the set of complex numbers z such that $|z| < \frac{1}{2}$ is the open disk of center 0 and radius $\frac{1}{2}$. □

Proposition 10.3.2. *Consider a complex power series*

$$s(z) = \sum_{n \geqslant 0} a_n z^n.$$

*(a) If for some $z_0 \neq 0$ the series $s(z_0)$ converges*absolutely*, then for any $z \in \mathbb{C}$ such that $|z| \leqslant |z_0|$ the series $s(z)$ converges absolutely.*

(b) If for some $z_0 \neq 0$ the series $s(z_0)$ is convergent, not necessarily absolutely, then for any $z \in \mathbb{C}$ such that $|z| < |z_0|$, the series $s(z)$ converges absolutely*.*

Proof. (a) Since $|z| \leqslant |z_0|$ we deduce that

$$|a_n z^n| \leqslant |a_n z_0^n|, \;\; \forall n \geqslant 0.$$

The desired conclusion now follows from the comparison principle.

(b) Since $s(z_0)$ converges we deduce that

$$\lim_n a_n z_0^n = 0.$$

In particular, we deduce that the sequence $(a_n z_0^n)$ is bounded, i.e., there exists $C > 0$ such that

$$|a_n z_0^n| \leqslant C, \quad \forall n \geqslant 0.$$

We set

$$r := \left|\frac{z}{z_0}\right| = \frac{|z|}{|z_0|} < 1.$$

We observe that

$$|a_n z^n| = |a_n z_0^n| \frac{|z|^n}{|z_0|^n} \leqslant Cr^n.$$

Since $r < 1$ we deduce that the geometric series

$$\sum_{n \geqslant 0} Cr^n$$

is convergent and the comparison principle implies that the series

$$\sum_{n \geqslant 0} |a_n z^n|$$

is also convergent. □

Consider a complex power series

$$s(z) = \sum_{n \geqslant 0} a_n z^n$$

We consider the set

$$\mathcal{R} = \big\{ r \geqslant 0; \;\; \exists z \in \mathbb{C} \text{ such that } \; |z| = r, \;\; s(z) \text{ is convergent} \,\big\} \subset \mathbb{R}.$$

Note that the set $\mathcal{R}$ is not empty because $0 \in \mathbb{R}$. Next observe that Proposition 10.3.2(b) implies that if $r_0 \in \mathcal{R}$, then $[0, r_0) \subset \mathcal{R}$. We set

$$R := \sup \mathcal{R} \in [0, \infty].$$

Proposition 10.3.2 shows that $s(z)$ converges absolutely for any $|z| < R$, and diverges for $|z| > R$. The number $R \in [0, \infty]$ is called the *radius of convergence* of the power series $s(z)$.

Example 10.3.3 (Complex exponential)**.** Consider the power series

$$E(z) = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots = \sum_{n \geqslant 0} \frac{1}{n!} z^n.$$

This series is absolutely convergent for any $z \in \mathbb{C}$ because the series of positive numbers

$$\sum_{n \geqslant 0} \frac{|z|^n}{n!}$$

is convergent for any z. Thus the radius of convergence of this power series is ∞. For simplicity we will denote by $E(z)$ the sum of the series $E(z)$.

Observe that for a real number x the sum of the series $E(x)$ is e^x; see Exercise 8.7. We write this

$$E(x) = e^x, \quad \forall x \in \mathbb{R}. \tag{10.3.1}$$

The properties of the exponential show that

$$E(x+y) = e^{x+y} = e^x e^y = E(x)E(y), \quad \forall x, y \in \mathbb{R}. \tag{10.3.2}$$

A more general result is true, namely,

$$E(z+\zeta) = E(z)E(\zeta), \quad \forall z, \zeta \in \mathbb{C}. \tag{10.3.3}$$

To prove the above equality we denote by $E_n(z)$ the n-th partial sum of the series $E(z)$,

$$E_n(z) = 1 + \frac{z}{1!} + \cdots + \frac{z^n}{n!}.$$

The equality (10.3.3) is equivalent to the equality

$$\lim_n \big(E_{2n}(z+\zeta) - E_{2n}(z)E_{2n}(\zeta) \big) = 0. \tag{10.3.4}$$

Fix a real number $M > 1$ such that

$$|z|, \ |\zeta| < M.$$

We have

$$E_{2n}(z+\zeta) = \sum_{m=0}^{2n} \frac{1}{m!}(z+\zeta)^m = \sum_{m=0}^{2n} \frac{1}{m!} \sum_{j=0}^{m} \binom{m}{j} z^{m-j}\zeta^j$$

$$= \sum_{m=0}^{2n} \frac{1}{m!} \sum_{j=0}^{m} \frac{m! z^{m-j}\zeta^j}{(m-j)!j!} = \sum_{m=0}^{2n} \sum_{j=0}^{m} \frac{z^{m-j}\zeta^j}{(m-j)!j!}$$

$(k := m-j)$

$$= \sum_{m=0}^{2n} \sum_{\substack{j+k=m\\ j,k\geqslant 0}} \frac{z^k\zeta^j}{k!j!} = \sum_{\substack{j+k\leqslant 2n\\ j,k\geqslant 0}} \frac{z^k\zeta^j}{k!j!}.$$

Similarly we have

$$E_{2n}(z)E_{2n}(\zeta) = \left(\sum_{k=0}^{2n} \frac{z^k}{k!} \right) \left(\sum_{j=0}^{2n} \frac{\zeta^j}{j!} \right) = \sum_{0\leqslant j,k\leqslant 2n} \frac{z^k\zeta^j}{k!j!}.$$

We deduce

$$|E_{2n}(z+\zeta) - E_{2n}(z)E_{2n}(\zeta)| = \left| \sum_{\substack{j+k>2n\\ 0\leqslant j,k\leqslant 2n}} \frac{z^k\zeta^j}{k!j!} \right|$$

$$\leqslant \sum_{\substack{j+k>2n\\ 0\leqslant j,k\leqslant 2n}} \frac{|z|^k|\zeta|^j}{k!j!} \leqslant M^{4n} \sum_{\substack{j+k>2n\\ 0\leqslant j,k\leqslant 2n}} \frac{1}{k!j!} \leqslant \frac{M^{4n}}{n!} \sum_{\substack{j+k>2n\\ 0\leqslant j,k\leqslant 2n}} 1 \leqslant \frac{4n^2M^{4n}}{n!}.$$

From (4.2.8) we deduce that

$$\lim_n \frac{4n^2M^{4n}}{n!} \to 0.$$

Because of the equalities (10.3.1) and (10.3.3), for any $z \in \mathbb{C}$ we set

$$e^z := E(z) = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} \cdots . \tag{10.3.5}$$

Suppose that in (10.3.5) the number z is purely imaginary, $z = \boldsymbol{i}t$, $t \in \mathbb{R}$. We deduce the celebrated *Euler's formula*

$$\begin{aligned} e^{\boldsymbol{i}t} &= 1 + \frac{\boldsymbol{i}t}{1!} + \frac{\boldsymbol{i}^2t^2}{2!} + \frac{\boldsymbol{i}^3t^3}{3!} + \cdots \\ &= \left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots\right) + \boldsymbol{i}\left(t - \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots\right) \\ &= \cos t + \boldsymbol{i}\sin t. \end{aligned} \tag{10.3.6}$$

If we let $t = \pi$ in the above equality we deduce

$$e^{\boldsymbol{i}\pi} = \cos\pi + \boldsymbol{i}\sin\pi = -1$$

i.e.,

$$e^{\boldsymbol{i}\pi} + 1 = 0. \tag{10.3.7}$$

The last very compact equality describes a deep connection between the five most important numbers in science: $0, 1, e, \pi, \boldsymbol{i}$. □

10.4. Exercises

Exercise 10.1. Prove the equalities (10.1.3) and (10.1.4). □

Exercise 10.2. (a) Consider the complex numbers

$$z_1 = 4 + 5\boldsymbol{i}, \;\; z_2 = 5 + 12\boldsymbol{i}.$$

Compute

$$z_1 z_2, \;\; |z_2|, \;\; \frac{z_1}{z_2}.$$

(b) Show that if

$$z = \frac{1}{2}(1 + \sqrt{3}\boldsymbol{i}),$$

then

$$z^2 + z + 1 = \bar{z}^2 + \bar{z} + 1 = 0, \;\; z^3 = \bar{z}^3 = 1.$$ □

Exercise 10.3. (a) Prove that if $z \in \mathbb{C}$, then

$$z^5 = 1 \wedge z \neq 1 \iff z^4 + z^3 + z^2 + z + 1 = 0 \iff z^2 + z + 1 + \frac{1}{z} + \frac{1}{z^2} = 0.$$

(b) Suppose that z satisfies the above equation, $z^4 + z^3 + z^2 + z + 1 = 0$. We set

$$\zeta := z + \frac{1}{z}.$$

Prove that

$$z^2 + \frac{1}{z^2} = \zeta^2 - 2,$$

and

$$\zeta^2 + \zeta - 1 = 0. \tag{10.4.1}$$

(c) Find the two roots ζ_1, ζ_2 of the quadratic equation (10.4.1).

(d) If ζ_1, ζ_2 are as above, find all the complex numbers z such that

$$z + \frac{1}{z} = \zeta_1 \;\; \vee \;\; z + \frac{1}{z} = \zeta_2.$$

(e) Use (d) to compute $\cos(2\pi/5)$, $\sin(2\pi/5)$. □

Exercise 10.4. (a) Let $z_0 \in \mathbb{C}$ and $r > 0$. Prove that the open disk $D_r(z_0)$ is an open set in the sense of Definition 10.2.4(b).

(b) Prove that if $\mathcal{O}_1, \mathcal{O}_2 \subset \mathbb{C}$ are open sets, then so are the sets $\mathcal{O}_1 \cap \mathcal{O}_2$, $\mathcal{O}_1 \cup \mathcal{O}_2$.

(c) Consider the set

$$S := \big\{ z \in \mathbb{C}; \;\; \mathbf{Im}\, z = 0, \;\; \mathbf{Re}\, z \in [0, 1] \big\}.$$

Draw a picture of S and then prove that it is a closed set in the sense of Definition 10.2.4(c). □

Exercise 10.5. Let S be a subset of the complex plane, $S \subset \mathbb{C}$. Prove that the following statements are equivalent.

(i) The set S is closed.

(ii) For any sequence $(z_n)_{n \geqslant 1}$ of points in S, $z_n \in S$, $\forall n$, if the sequence converges to z_*, then $z_* \in S$.

□

Exercise 10.6. Use the ideas in the proof of Proposition 10.2.6 to prove Proposition 10.2.12. □

Exercise 10.7. Prove Proposition 10.2.10 by imitating the proof of Proposition 4.3.1. □

Chapter 11

The geometry and topology of Euclidean spaces

The calculus of one-real-variable functions has a several-variable counterpart. To state and prove these results we need an appropriate language. The goal of this chapter is to introduce the terminology and the concepts required to make the jump into higher dimensions.

11.1. Basic affine geometry

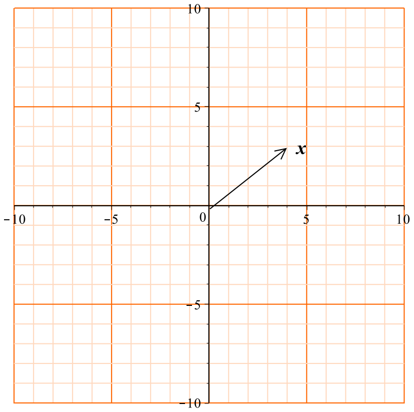

Figure 11.1. *The point $\boldsymbol{x} \in \mathbb{R}^2$ with (Cartesian) coordinates $(4, 3)$ is identified with the vector that starts at the origin and ends at $\boldsymbol{x}$.*

Let $n \in \mathbb{N}$. The *canonical n-dimensional real Euclidean space* is the Cartesian product

$$\mathbb{R}^n := \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}}.$$

The elements of $\mathbb{R}^n$ are called (n-dimensional) *vectors* or *points* and they are n-tuples of real numbers

$$\boldsymbol{x} := \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}. \tag{11.1.1}$$

Above, the real numbers $x^1, \dots, x^n$ are called the *Cartesian coordinates* of the vector $\boldsymbol{x}$; see Figure 11.1.

☞ Several comments are in order. First, note that we represent the vector as a (vertical) *column*. To remind us of this, we use the *superscript* notation x^i rather than the *subscript* notation x_i. There are several other good reasons for this choice of notation, but explaining them is difficult at this time. This choice is part of a larger collection of conventions sometimes referred to as the *Einstein's conventions*. For now, accept and use this convention as a very good idea with a nebulous payoff that will reveal itself once your mathematical background is a bit more sophisticated.

For typographical reasons it is inconvenient to work with tall columns of numbers of the type appearing in (11.1.1) so we will use the notation $[x^1, \dots, x^n]^\top$ or $(x^1, \dots, x^n)$ to denote the *column* in the right-hand side of (11.1.1).

Also, when we refer to a point $\boldsymbol{x} \in \mathbb{R}^n$ as a vector we secretly think of $\boldsymbol{x}$ as the tip of an arrow that starts at the origin and ends at $\boldsymbol{x}$; see Figure 11.1.

The attribute *Euclidean space* attached to the set $\mathbb{R}^n$ refers to the additional structure this set is equipped with. First of all, $\mathbb{R}^n$ has a structure of *vector space*[1]. More precisely, it is equipped with two operations, *addition* and *multiplication by scalars* satisfying certain properties.

The addition is a function $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ that associates to a pair of vectors $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n$ a third vector, its *sum* $\boldsymbol{x} + \boldsymbol{y} \in \mathbb{R}^n$, defined as follows: if

$$\boldsymbol{x} = \big(x^1, \dots, x^n\big), \;\; \boldsymbol{y} = \big(y^1, \dots, y^n\big),$$

then

$$\boldsymbol{x} + \boldsymbol{y} := \big(x^1 + y^1, \dots, x^n + y^n\big) \in \mathbb{R}^n.$$

The multiplication-by-scalars operation associates to a pair $(\lambda, \boldsymbol{x})$ consisting of a real number (or scalar) λ and a vector $\boldsymbol{x} \in \mathbb{R}^n$, a new vector denoted by $\lambda\boldsymbol{x}$ (or $\lambda \cdot \boldsymbol{x}$) and

[1] As we progress in this course I will assume increased knowledge of linear algebra. I recommend [40] as a linear algebra source very appropriate for the goals of this course.

defined as follows: if $\boldsymbol{x} = \big(x^1, x^2, \dots, x^n\big)$, then

$$\lambda\boldsymbol{x} := \big(\lambda x^1, \dots, \lambda x^n\big) \in \mathbb{R}^n.$$

These operations satisfy the following properties.[2]

(i) (Associativity) For any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$

$$(\boldsymbol{x}+\boldsymbol{y})+\boldsymbol{z} = \boldsymbol{x}+(\boldsymbol{y}+\boldsymbol{z}).$$

(ii) (Commutativity) For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\boldsymbol{x}+\boldsymbol{y} = \boldsymbol{y}+\boldsymbol{x}.$$

(iii) (Neutral or identity element) The vector $\boldsymbol{0} := (0, \dots, 0) \in \mathbb{R}^n$ has the property: $\forall \boldsymbol{x} \in \mathbb{R}^n$ we have

$$\boldsymbol{0}+\boldsymbol{x} = \boldsymbol{x}+\boldsymbol{0} = \boldsymbol{x}.$$

(iv) (Inverse or opposite element) For any $\boldsymbol{x} = \big(x_1, \dots, x_n\big) \in \mathbb{R}^n$, the vector

$$-\boldsymbol{x} := \big(-x_1, \dots, -x_n\big)$$

has the property:

$$\boldsymbol{x}+(-\boldsymbol{x}) = (-\boldsymbol{x})+\boldsymbol{x} = \boldsymbol{0}.$$

(v) (Distributivity with respect to vector addition) For any $\lambda \in \mathbb{R}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\lambda(\boldsymbol{x}+\boldsymbol{y}) = \lambda\boldsymbol{x}+\lambda\boldsymbol{y}.$$

(vi) (Distributivity with respect to the scalar addition) For any $\lambda, \mu \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^n$

$$(\lambda+\mu)\boldsymbol{x} = \lambda\boldsymbol{x}+\mu\boldsymbol{x}, \;\; (\lambda\mu)\boldsymbol{x} = \lambda(\mu\boldsymbol{x}).$$

(vii) For any $\boldsymbol{x} \in \mathbb{R}^n$,

$$1 \cdot \boldsymbol{x} = \boldsymbol{x}.$$

Note that $0 \cdot \boldsymbol{x} = \boldsymbol{0}$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.

Definition 11.1.1. The canonical or natural basis of $\mathbb{R}^n$ is the set of vectors $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, where

$$\boldsymbol{e}_1 := \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \;\; \boldsymbol{e}_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, \boldsymbol{e}_n := \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{11.1.2}$$

□

[2] Compare them with the algebraic axioms of $\mathbb{R}$.

Note that if $\boldsymbol{x} = \big(x^1, \dots, x^n\big)$, then

$$\boxed{\boldsymbol{x} = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix} = x^1\boldsymbol{e}_1 + \cdots + x^n\boldsymbol{e}_n = \sum_{i=1}^n x^i\boldsymbol{e}_i}.$$

For example, we have the following equality in $\mathbb{R}^3$,

$$\begin{bmatrix} 3 \\ -4 \\ 5 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} -4\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 5\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 3\boldsymbol{e}_1 - 4\boldsymbol{e}_2 + 5\boldsymbol{e}_3. \tag{11.1.3}$$

At this point it is convenient to introduce the *Kronecker symbol* δ^i_j,

$$\delta^i_j := \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{11.1.4}$$

Using the Kronecker symbol we observe that

$$\boldsymbol{e}_k = \begin{bmatrix} \delta^1_k \\ \delta^2_k \\ \vdots \\ \delta^n_k \end{bmatrix}, \;\; \forall k = 1, \dots, n.$$

Remark 11.1.2. When $n = 2$, the coordinates x^1, x^2 are usually denoted by x and respectively y, and the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ are usually denoted by $\boldsymbol{i}$ and respectively $\boldsymbol{j}$; see Figure 11.2.

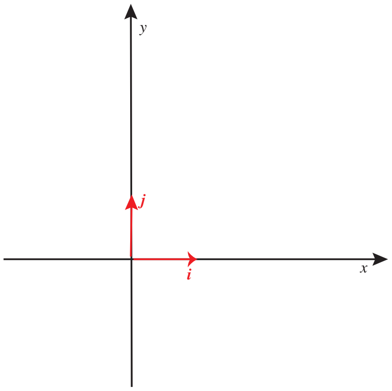

Figure 11.2. *A Cartesian coordinate system in* $\mathbb{R}^2$.

When $n = 3$, the coordinates x^1, x^2, x^3 are usually denoted by x, y and respectively z, and the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are usually denoted by $\boldsymbol{i}, \boldsymbol{j}$ and respectively $\boldsymbol{k}$; see Figure 11.3.

Thus, in $\mathbb{R}^3$, the equality (11.1.3) could be rewritten as

$$\begin{bmatrix} 3 \\ -4 \\ 5 \end{bmatrix} = 3\boldsymbol{i} - 4\boldsymbol{j} + 5\boldsymbol{k}. \qquad \square$$

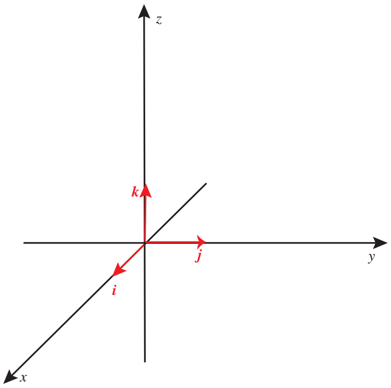

Figure 11.3. *A Cartesian coordinate system in* $\mathbb{R}^3$.

Definition 11.1.3. Two nonzero vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are called *collinear* if one is a multiple of the other, i.e., there exists $t \in \mathbb{R}$, $t \neq 0$, such that $\boldsymbol{v} = t\boldsymbol{u}$ (and thus $\boldsymbol{u} = t^{-1}\boldsymbol{v}$). □

Definition 11.1.4. Let $\boldsymbol{p}, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v} \neq \boldsymbol{0}$. The *line* in $\mathbb{R}^n$ *through* $\boldsymbol{p}$ *and in the direction* $\boldsymbol{v}$ is the set

$$\boxed{\ell_{\boldsymbol{p},\boldsymbol{v}} := \left\{ \boldsymbol{p} + t\boldsymbol{v};\ \ t \in \mathbb{R} \right\} \subset \mathbb{R}^n}. \tag{11.1.5}$$

The vector $\boldsymbol{v}$ is called a *direction vector* of the line. □

Let us point out that, if the two nonzero vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are collinear, then, for any point $\boldsymbol{p} \in \mathbb{R}^n$, the line through $\boldsymbol{p}$ in the direction $\boldsymbol{u}$ coincides with the line through $\boldsymbol{p}$ in the direction $\boldsymbol{v}$, i.e.,

$$\ell_{\boldsymbol{p},\boldsymbol{u}} = \ell_{\boldsymbol{p},\boldsymbol{v}}.$$

Exercise 11.1 asks you to prove this fact.

Observe that the line through $\boldsymbol{p}$ and in the direction $\boldsymbol{v}$ is the image of the function

$$\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^n, \quad \boldsymbol{f}(t) = \boldsymbol{p} + t\boldsymbol{v}.$$

You can think of the map $\boldsymbol{f}$ as describing the motion of a point in $\mathbb{R}^n$ so that its location at time $t \in \mathbb{R}$ is $\boldsymbol{p} + t\boldsymbol{v}$. The line $\ell_{\boldsymbol{p},\boldsymbol{v}}$ is then the *trajectory* described by this point during its motion. If

$$\boldsymbol{p} = \begin{bmatrix} p^1 \\ \vdots \\ p^n \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} v^1 \\ \vdots \\ v^n \end{bmatrix},$$

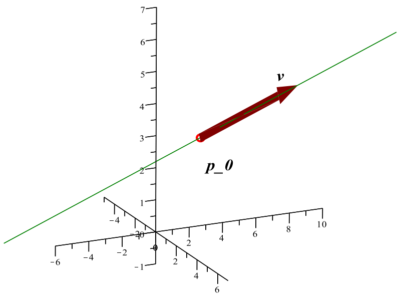

Figure 11.4. *The line through the point* $\boldsymbol{p} = [1,2,3]^{\top}$ *and in the direction* $\boldsymbol{v} = [3,4,2]^{\top}$.

then

$$\boldsymbol{p} + t\boldsymbol{v} = \begin{bmatrix} p^1 + tv^1 \\ \vdots \\ p^n + tv^n \end{bmatrix}$$

and we can describe the line through $\boldsymbol{p}$ in the direction $\boldsymbol{v}$ using the *parametric equation* or *parametrization*

$$\left\{ \begin{array}{ccc} x^1 & = & p^1 + tv^1 \\ \vdots & \vdots & \vdots \\ x^n & = & p^n + tv^n, \end{array} \right. \qquad t \in \mathbb{R}. \tag{11.1.6}$$

Above, the variable t is called the *parameter* (of the parametric equations). As t varies, the right-hand side of (11.1.6) describes the coordinates of a moving point along the line. The parametric equations (11.1.6) should be interpreted as saying that

$$\boldsymbol{x} \in \ell_{\boldsymbol{p},\boldsymbol{v}} \Longleftrightarrow \exists t \in \mathbb{R} : \;\; x^i = p^i + tv^i, \;\; \forall i = 1, \dots, n.$$

Definition 11.1.5. The lines $\ell_{\boldsymbol{0},\boldsymbol{e}_1}, \dots, \ell_{\boldsymbol{0},\boldsymbol{e}_n}$ are called the *coordinate axes* of $\mathbb{R}^n$. □

Example 11.1.6. Figures 11.2 and 11.3 depict the coordinate axes in $\mathbb{R}^2$ and respectively $\mathbb{R}^3$. □

Suppose that we are given two *distinct* points $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$. These two points determine two collinear vectors, $\boldsymbol{v} = \boldsymbol{q} - \boldsymbol{p}$ and $-\boldsymbol{v} = \boldsymbol{p} - \boldsymbol{q}$; see Figure 11.5.[3]

[3] The old-fashioned notation for the vector $\boldsymbol{q} - \boldsymbol{p}$ is $\overrightarrow{\boldsymbol{pq}}$

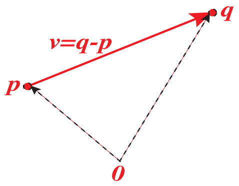

Figure 11.5. *You should think of $\boldsymbol{v}=\boldsymbol{q}-\boldsymbol{p}$ as the vector described by the arrow that starts at $\boldsymbol{p}$ and ends at $\boldsymbol{q}$.*

The distinct points $\boldsymbol{p},\boldsymbol{q}$ belong to both lines $\ell_{\boldsymbol{p},\boldsymbol{v}}$ and $\ell_{\boldsymbol{q},-\boldsymbol{v}}$. Since these two lines intersect in two distinct points they must coincide; see Exercise 11.2. Thus

$$\ell_{\boldsymbol{p},\boldsymbol{v}}=\ell_{\boldsymbol{q},-\boldsymbol{v}}.$$

This line is called the *line determined by the (distinct) points $\boldsymbol{p}$ and $\boldsymbol{q}$*, and we will denote it by $\boldsymbol{pq}$. In other words, $\boldsymbol{pq}$ is the line through $\boldsymbol{p}$ in the direction $\boldsymbol{q}-\boldsymbol{p}$,

$$\boldsymbol{pq}=\ell_{\boldsymbol{p},\boldsymbol{q}-\boldsymbol{p}}.$$

By construction, either of the vectors $\boldsymbol{q}-\boldsymbol{p}$ or $\boldsymbol{p}-\boldsymbol{q}$ is a direction vector of the line $\boldsymbol{pq}$. Observe that this line consists of all the points in $\mathbb{R}^n$ of the form

$$\boldsymbol{p}+t\boldsymbol{v}=\boldsymbol{p}+t(\boldsymbol{q}-\boldsymbol{p})=(1-t)\boldsymbol{p}+t\boldsymbol{q},\;\; t\in\mathbb{R}.$$

We thus have the important equality

$$\boxed{\boldsymbol{pq}=\Big\{(1-t)\boldsymbol{p}+t\boldsymbol{q}\in\mathbb{R}^n,\;\; t\in\mathbb{R}\Big\}=\boldsymbol{qp}}. \tag{11.1.7}$$

Example 11.1.7. Consider the points $\boldsymbol{p}=(1,2,3)$ and $\boldsymbol{q}=(4,5,6)$ in $\mathbb{R}^3$. Then the line through $\boldsymbol{p}$ and $\boldsymbol{q}$ is the subset of $\mathbb{R}^3$ described by

$$\begin{aligned}\boldsymbol{pq}&=\big\{(1-t)\cdot(1,2,3)+t\cdot(4,5,6);\;\; t\in\mathbb{R}\big\}\\&=\big\{(1+3t,2+3t,3+3t);\;\; t\in\mathbb{R}\big\}.\end{aligned}$$

Equivalently, we say that the line $\boldsymbol{pq}$ is described by the equations

$$\begin{array}{rcl}x&=&1+3t,\\y&=&2+3t,\\z&=&3+3t,\end{array}$$

$t\in\mathbb{R}$. □

Given $\boldsymbol{p},\boldsymbol{q}\in\mathbb{R}^n$, $\boldsymbol{p}\neq\boldsymbol{q}$, the line $\boldsymbol{pq}$ is the image of the function

$$\boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}:\mathbb{R}\to\mathbb{R}^n,\;\; \boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}(t)=(1-t)\boldsymbol{p}+t\boldsymbol{q}.$$

Moreover,

$$\boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}(0)=\boldsymbol{p},\;\; \boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}(1)=\boldsymbol{q}.$$

Intuitively, the function $\boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}$ describes the motion of a particle in the space $\mathbb{R}^n$ that is located at $\boldsymbol{f}_{\boldsymbol{p},\boldsymbol{q}}(t)$ at the moment of time t. The line $\boldsymbol{pq}$ is then the trajectory described by this moving particle. Note that at $t = 0$ the particle is located at $\boldsymbol{p}$ while, a second later, at $t = 1$, the particle is located at $\boldsymbol{q}$. The *line segment* connecting $\boldsymbol{p}$ to $\boldsymbol{q}$ is defined to be the portion of the trajectory described by this particle during the time interval $[0, 1]$. We denote this line segment by $[\boldsymbol{p}, \boldsymbol{q}]$ and we observe that it has the algebraic description

$$\boxed{[\boldsymbol{p}, \boldsymbol{q}] := \Big\{ (1 - t)\boldsymbol{p} + t\boldsymbol{q};\ \ t \in [0, 1] \Big\}}. \tag{11.1.8}$$

Definition 11.1.8 (Convex sets)**.** Let $n \in \mathbb{N}$. A subset $C \subset \mathbb{R}^n$ is called *convex* if for any two points in C, the segment connecting them is entirely contained in C. More formally, C is convex iff

$$\forall \boldsymbol{p}, \boldsymbol{q} \in C,\ \ [\boldsymbol{p}, \boldsymbol{q}] \subset C,$$

or, equivalently,

$$\boxed{\forall \boldsymbol{p}, \boldsymbol{q} \in C,\ \ \forall t \in [0, 1],\ \ (1 - t)\boldsymbol{p} + t\boldsymbol{q} \in C}. \qquad \square$$

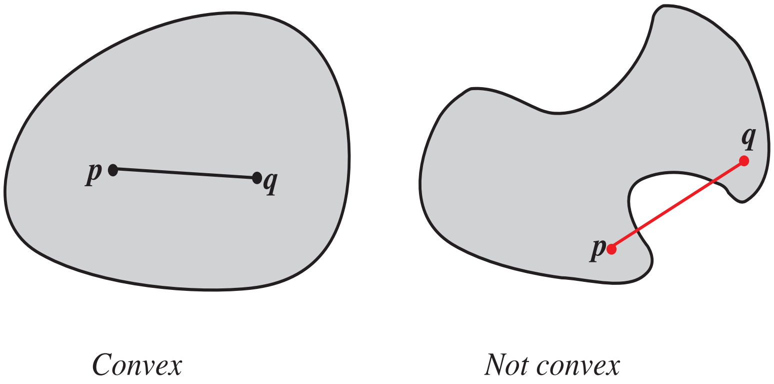

Figure 11.6. *Examples of convex and non-convex planar sets.*

Definition 11.1.9 (Linear forms)**.** A *linear form* or *linear functional* on $\mathbb{R}^n$ is a map $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ satisfying the following two properties.

(i) (Additivity.) For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have $\boldsymbol{\xi}(\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{\xi}(\boldsymbol{x}) + \boldsymbol{\xi}(\boldsymbol{y})$.

(ii) (Homogeneity.) For any $t \in \mathbb{R}$ and any $\boldsymbol{x} \in \mathbb{R}^n$ we have $\boldsymbol{\xi}(t\boldsymbol{x}) = t\boldsymbol{\xi}(\boldsymbol{x})$.

We denote by $(\mathbb{R}^n)^*$ the set of linear forms on $\mathbb{R}^n$ and we will refer to it as the *dual* of $\mathbb{R}^n$. □

☞ We want to emphasize that the linear forms are "*beasts that eat vectors and spit out numbers*".

Example 11.1.10. (a) The set $(\mathbb{R}^n)^*$ is not empty. The trivial map $\mathbb{R}^n \to \mathbb{R}$ that sends every $\boldsymbol{x}$ to 0 is a linear functional. We will denote it by $\mathbf{0}$.

(b) Consider *addition function* $\alpha : \mathbb{R}^2 \to \mathbb{R}$, $\alpha(\boldsymbol{x}) = x^1 + x^2$. Concretely, the function α "eats" a two-dimensional vector $\boldsymbol{x} = (x^1, x^2)$ and returns the sum of its coordinates. Let us verify that α is a linear form.

Indeed, we have

$$\alpha(\boldsymbol{x}+\boldsymbol{y}) = \alpha\big(\,(x^1+y^1, x^2+y^2)\,\big) = (x^1+y^1) + (x^2+y^2)$$

$$= (x^1+x^2) + (y^1+y^2) = \alpha(\boldsymbol{x}) + \alpha(\boldsymbol{y}),\ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2,$$

$$\alpha(t\boldsymbol{x}) = \alpha\big(\,(tx^1, tx^2)\,\big) = tx^1 + tx^2 = t(x^1+x^2) = t\alpha(\boldsymbol{x}),\ \ \forall t \in \mathbb{R},\ \ \boldsymbol{x} \in \mathbb{R}^2.$$

(c) For any $k = 1, \dots, n$, we define $\boldsymbol{e}^k : \mathbb{R}^n \to \mathbb{R}$ by

$$\boldsymbol{e}^k(\boldsymbol{x}) = x^k,\ \ \forall \boldsymbol{x} = (x^1, \dots, x^n) \in \mathbb{R}^n. \tag{11.1.9}$$

From the definition of the addition and multiplication by scalars we deduce immediately that the maps $\boldsymbol{e}^k$ are linear functionals. The linear forms $\boldsymbol{e}^1, \dots, \boldsymbol{e}^n$ are called the *basic linear forms on* $\mathbb{R}^n$. □

The proof of the next result is left to you as an exercise.

Proposition 11.1.11. *If $\boldsymbol{\xi}, \boldsymbol{\omega}$ are linear forms on $\mathbb{R}^n$ and t is a real number, then the sum $\boldsymbol{\xi} + \boldsymbol{\omega}$ and the multiple $t\boldsymbol{\xi}$ are linear functionals on $\mathbb{R}^n$.*[4] □

The linear forms on $\mathbb{R}^n$ have a very simple structure described in our next result.

Proposition 11.1.12. *Let $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ be a linear form. For $i = 1, \dots, n$ we set*[5]

$$\xi_i := \xi(\boldsymbol{e}_i),$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ is the canonical basis (11.1.2) of $\mathbb{R}^n$. Then,

$$\boldsymbol{\xi}(\boldsymbol{x}) = \xi_1 x^1 + \xi_2 x^2 + \cdots + \xi_n x^n = \sum_{i=1}^n \xi_i x^i,\ \ \forall \boldsymbol{x} = (x^1, \dots, x^n) \in \mathbb{R}^n. \tag{11.1.10}$$

Conversely, given any real numbers $\xi_1, \dots, \xi_n$, the linear form

$$\boldsymbol{\xi} = \xi_1 \boldsymbol{e}^1 + \cdots + \xi_n \boldsymbol{e}^n,$$

where $\boldsymbol{e}^k$ are defined by (11.1.9), satisfies (11.1.10).

[4] In modern language this signifies that the space $(\mathbb{R}^n)^*$ of linear forms on $\mathbb{R}^n$ is a vector subspace of the vector space of functions on $\mathbb{R}^n \to \mathbb{R}$.

[5] Note that here we use the subscript notation, ξ_i instead of the superscript notation ξ^i. This is part of Einstein's conventions I referred to at the beginning of this chapter.

Proof. To prove (11.1.10) let $\boldsymbol{x} = (x^1, \dots, x^n) \in \mathbb{R}^n$. Then

$$\boldsymbol{x} = x^1\boldsymbol{e}_1 + \cdots + x^n\boldsymbol{e}_n.$$

From the additivity of $\boldsymbol{\xi}$ we deduce

$$\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\xi}(x^1\boldsymbol{e}_1 + \cdots + x^n\boldsymbol{e}_n) = \boldsymbol{\xi}(x^1\boldsymbol{e}_1) + \cdots + \xi(x^n\boldsymbol{e}_n)$$

(use the homogeneity of ξ)

$$= x^1\boldsymbol{\xi}(\boldsymbol{e}_1) + \cdots + x^n\boldsymbol{\xi}(\boldsymbol{e}_n) = \xi_1 x^1 + \xi_2 x^2 + \cdots + \xi_n x^n.$$

This proves (11.1.10).

Conversely, if $\boldsymbol{\xi} = \xi_1\boldsymbol{e}^1 + \cdots + \xi_n\boldsymbol{e}^n$, then

$$\boldsymbol{\xi}(\boldsymbol{x}) = \xi_1\boldsymbol{e}^1(\boldsymbol{x}) + \cdots + \xi_n\boldsymbol{e}^n(\boldsymbol{x}) \overset{(11.1.9)}{=} \xi_1 x^1 + \xi_2 x^2 + \cdots + \xi_n x^n.$$

□

The above proposition shows that a linear form $\boldsymbol{\xi}$ on $\mathbb{R}^n$ is completely and uniquely determined by its values on the basic vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$. We will identify $\boldsymbol{\xi}$ with the *row*

$$[\xi_1, \dots, \xi_n], \;\; \xi_i = \xi(\boldsymbol{e}_i),$$

and *we will think of any length-n row of real numbers as defining a linear form on $\mathbb{R}^n$*. In the physics literature the linear forms are often referred to as *covectors*.

The basic linear forms $\boldsymbol{e}^1, \dots, \boldsymbol{e}^n$ defined in (11.1.9) are uniquely determined by the equalities

$$\boxed{\boldsymbol{e}^i(\boldsymbol{e}_j) = \delta^i_j, \;\; \forall i, j = 1, \dots, n}, \tag{11.1.11}$$

where we recall that δ^i_j is the Kronecker symbol (11.1.4).

Example 11.1.13. Suppose that $n = 4$. Then the linear form $\boldsymbol{\xi} : \mathbb{R}^4 \to \mathbb{R}$ defined by the *row* vector $[3, 5, 7, 9]$ is given by

$$\boldsymbol{\xi}(x^1, x^2, x^3, x^4) = 3x^1 + 5x^2 + 7x^3 + 9x^4, \;\; \forall (x^1, x^2, x^3, x^4) \in \mathbb{R}^4.$$

□

Definition 11.1.14. A subset H of $\mathbb{R}^n$ is called a *hyperplane* if there exists a *nonzero* linear form $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ and a real constant c such that H consists of all the points $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\boldsymbol{\xi}(\boldsymbol{x}) = c$. □

Example 11.1.15. (a) A hyperplane in $\mathbb{R}^2$ is a line in $\mathbb{R}^2$. Indeed, any linear form on $\mathbb{R}^2$ has the form

$$\xi(x, y) = ax + by$$

where a, b are fixed real numbers and x, y denote the Cartesian coordinates on $\mathbb{R}^2$. An equation of the form

$$ax + by = c$$

describes a line in $\mathbb{R}^2$. For example, the equation $-2x + y = 3$ describes the line $y = 2x + 3$, with slope 2 and y-intercept 3; see Figure 11.7.

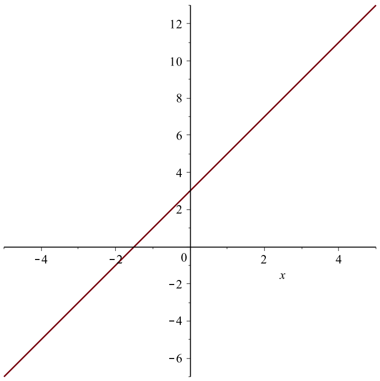

Figure 11.7. *The planar line with slope* 2 *and* y*-intercept* 3.

(b) A hyperplane in $\mathbb{R}^3$ is a plane. For example, Figure 11.8 depicts the plane $x+2y+3z = 4$.

(c) A row vector $[\xi_1, \dots, \xi_n]$ and a constant c define the hyperplane in $\mathbb{R}^n$ consisting of all the points $\boldsymbol{x} = (x^1, \dots, x^n) \in \mathbb{R}^n$ satisfying the linear equation

$$\xi_1 x^1 + \cdots + \xi_n x^n = c.$$

All the hyperplanes in $\mathbb{R}^n$ are of this form. □

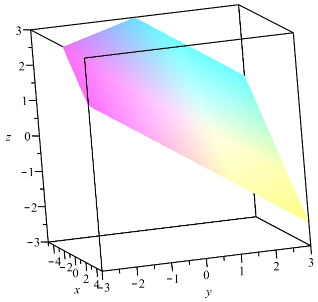

Figure 11.8. *The plane* $x + 2y + 3z = 4$.

Definition 11.1.16 (Affine subspaces)**.** (a) A nonempty subset $S \subset \mathbb{R}^n$ is called an *affine subspace* if it has the following property: for any points $\boldsymbol{p}, \boldsymbol{q} \in S$, $\boldsymbol{p} \neq \boldsymbol{q}$, the line $\boldsymbol{pq}$ is contained in S. In algebraic terms this means that S is an affine subspace if and only if, for any $\boldsymbol{p}, \boldsymbol{q} \in S$, $\boldsymbol{p} \neq \boldsymbol{q}$, and any $t \in \mathbb{R}$ we have $(1-t)\boldsymbol{p} + t\boldsymbol{q} \in S$.

(b) The subset S is called a *linear subspace* or *vector subspace* if it is an affine subspace and contains the origin. □

Example 11.1.17. (a) Any point in $\mathbb{R}^n$ is an affine subspace. The space $\mathbb{R}^n$ is obviously an affine subspace of itself.

(b) The lines and the hyperplanes in $\mathbb{R}^n$ are special examples of affine subspaces; see Exercise 11.8. When $n > 3$, there are examples of affine subspaces of $\mathbb{R}^n$ that are neither lines, nor hyperplanes.

(c) If nonempty, the intersection of two affine subspaces is an affine subspace. In particular, if two hyperplanes are not disjoint, then their intersection is an affine subspace. One can prove that if S is an affine subspace of $\mathbb{R}^n$ and $S \neq \mathbb{R}^n$, then S is the intersection of finitely many hyperplanes. □

Proposition 11.1.18. *Let S be a nonempty subset of $\mathbb{R}^n$. Then the following statements are equivalent.*

(i) *The set S is a linear subspace, i.e., it is an affine subspace of $\mathbb{R}^n$ containing the origin.*

(ii) *For any $\boldsymbol{u}, \boldsymbol{v} \in S$ and any $t \in \mathbb{R}$ we have*

$$t\boldsymbol{u} \in S \;\text{ and }\; \boldsymbol{u} + \boldsymbol{v} \in S.$$

In other words, either of the conditions (i) or (ii) above can be used as definition of a linear subspace.

Proof. (i) ⇒ (ii) We know that S is an affine subspace and $\mathbf{0} \in S$. Clearly $t\mathbf{0} = \mathbf{0}$, $\forall t \in \mathbb{R}$. For any $\boldsymbol{v} \in S$, $\boldsymbol{v} \neq 0$ and any $t \in \mathbb{R}$ we have

$$t\boldsymbol{v} = (1-t)\mathbf{0} + t\boldsymbol{v} \in S.$$

Thus, any multiple of any vector in S is also a vector in S. Thus, if $\boldsymbol{u} = \boldsymbol{v} \in S$ we have $\boldsymbol{u} + \boldsymbol{v} = 2\boldsymbol{u} \in S$. On the other hand, since S is an affine subspace, if $\boldsymbol{u}, \boldsymbol{v} \in S$, $\boldsymbol{u} \neq \boldsymbol{v}$, the vector $\boldsymbol{w} = \frac{1}{2}\boldsymbol{u} + \frac{1}{2}\boldsymbol{v}$ belongs to S. Hence the multiple $2\boldsymbol{w}$ belongs to S and therefore $\boldsymbol{u} + \boldsymbol{v} = 2\boldsymbol{w} \in S$.

(ii) ⇒ (i) Let $\boldsymbol{u} \in S$. Hence $\mathbf{0} = 0 \cdot \boldsymbol{u} \in S$. Next observe that if $\boldsymbol{u}, \boldsymbol{v} \in S$, $\boldsymbol{u} \neq \boldsymbol{v}$, and $t \in \mathbb{R}$, then

$$(1-t)\boldsymbol{u},\; t\boldsymbol{v} \in S \;\Rightarrow\; (1-t)\boldsymbol{u} + t\boldsymbol{v} \in S.$$

This proves that S is an affine subspace. □

Definition 11.1.19 (Linear operators)**.** Fix $m, n \in \mathbb{N}$. A map $A : \mathbb{R}^n \to \mathbb{R}^m$ is called *linear* or a *linear operator* if it satisfies the following two properties.

(i) (Additivity.) For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have $A(\boldsymbol{x} + \boldsymbol{y}) = A(\boldsymbol{x}) + A(\boldsymbol{y})$.

(ii) (Homogeneity.) For any $t \in \mathbb{R}$ and any $\boldsymbol{x} \in \mathbb{R}^n$ we have $A(t\boldsymbol{x}) = tA(\boldsymbol{x})$.

We denote by $\operatorname{Hom}(\mathbb{R}^n, \mathbb{R}^m)$ the set of linear operators $\mathbb{R}^n \to \mathbb{R}^m$. □

Note that $\operatorname{Hom}(\mathbb{R}^n, \mathbb{R})$ is none other than the dual of $\mathbb{R}^n$, i.e., the space $(\mathbb{R}^n)^*$ of linear functionals on $\mathbb{R}^n$. Let us mention a simplifying convention that has been universally adopted. If $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator and $\boldsymbol{x} \in \mathbb{R}^n$, then we will often use the simpler notation $A\boldsymbol{x}$ when referring to $A(\boldsymbol{x})$.

The linear operators $\mathbb{R}^n \to \mathbb{R}^m$ have a rather simple structure. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator. Denote by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ the canonical basis of $\mathbb{R}^n$ and by $x^1, \dots, x^n$ the canonical Cartesian coordinates. Similarly, we denote by $\boldsymbol{f}_1, \dots, \boldsymbol{f}_m$ the canonical basis of $\mathbb{R}^m$ and by $y^1, \dots, y^m$ the canonical Cartesian coordinates.

For any

$$\boldsymbol{x} = (x^1, \dots, x^n) = x^1\boldsymbol{e}_1 + \cdots + x^n\boldsymbol{e}_n \in \mathbb{R}^n$$

we have

$$A\boldsymbol{x} = A(x^1\boldsymbol{e}_1 + \cdots + x^n\boldsymbol{e}_n) = A(x^1\boldsymbol{e}_1) + \cdots + A(x^n\boldsymbol{e}_n) = x^1A\boldsymbol{e}_1 + \cdots + x^nA\boldsymbol{e}_n. \tag{11.1.12}$$

This shows that the operator A is completely determined by the m-dimensional vectors

$$A\boldsymbol{e}_1, \dots, A\boldsymbol{e}_n \in \mathbb{R}^m.$$

These m-dimensional vectors are described by columns of height m.

$$A\boldsymbol{e}_1 = \begin{bmatrix} A^1_1 \\ \\ A^2_1 \\ \vdots \\ A^m_1 \end{bmatrix}, \dots, A\boldsymbol{e}_j = \begin{bmatrix} A^1_j \\ \\ A^2_j \\ \vdots \\ A^m_j \end{bmatrix}, \dots, A\boldsymbol{e}_n = \begin{bmatrix} A^1_n \\ \\ A^2_n \\ \vdots \\ A^m_n \end{bmatrix}.$$

Arranging these columns one next to the other we obtain the rectangular array

$$\mathcal{M}_A = \begin{bmatrix} A^1_1 & \cdots & A^1_j & \cdots & A^1_n \\ \\ A^2_1 & \cdots & A^2_j & \cdots & A^2_n \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ A^m_1 & \cdots & A^m_j & \cdots & A^m_n \end{bmatrix}.$$

✍ We need to introduce some terminology and conventions.

- A rectangular array of numbers as above is called a *matrix*.
- The horizontal strings of numbers are called *rows*, and the vertical ones are called *columns*.
- We will denote by $\operatorname{Mat}_{m\times n}(\mathbb{R})$ the space of matrices with real entries, with m rows and n columns. The matrix $\mathcal{M}_A$ above is an $m \times n$ matrix.
- A *square matrix* is a matrix with an equal number of rows and columns. We will denote by $\operatorname{Mat}_n(\mathbb{R})$ the space of square matrices with n rows and columns.

- *The superscripts label the rows and the subscripts label the columns.* Thus, A^3_7 is the entry located at the intersection of the 3rd row with the 7th column of a matrix A.
- We denote by A_j the j-th column and by A^i the i-th row of a matrix A.

Note that a $1 \times k$ matrix is a length-k row

$$R = [r_1 \;\; r_2 \;\; \dots r_k],$$

while a $k \times 1$ matrix is a height-k column

$$C = \begin{bmatrix} c^1 \\ \vdots \\ c^k \end{bmatrix}$$

The *pairing* between a row R and a column C of the same size k is defined to be the number

$$\boxed{R \bullet C := r_1c^1 + r_2c^2 + \cdots + r_kc^k}. \tag{11.1.13}$$

If we identify rows with linear functionals, then $R \bullet C$ is the real number that we get when we feed the vector C to the linear functional defined by R.

The above discussion shows that to any linear operator $A : \mathbb{R}^n \to \mathbb{R}^m$ we can canonically associate an $m \times n$ matrix called the *matrix associated to the linear operator*. This matrix has n columns $A_1, \dots, A_n$ that describe the coordinates of the vectors $A\boldsymbol{e}_1, \dots, A\boldsymbol{e}_n$.

Using (11.1.12) we deduce

$$\begin{aligned} A\boldsymbol{x} &= x^1 A\boldsymbol{e}_1 + \cdots + x^n A\boldsymbol{e}_n \\ &= x^1 \begin{bmatrix} A^1_1 \\ \\ A^2_1 \\ \vdots \\ A^m_1 \end{bmatrix} + x^2 \begin{bmatrix} A^1_2 \\ \\ A^2_2 \\ \vdots \\ A^m_2 \end{bmatrix} + \cdots + x^n \begin{bmatrix} A^1_n \\ \\ A^2_n \\ \vdots \\ A^m_n \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} x^1A^1_1 + x^2A^1_2 + \cdots + x^nA^1_n \\ \\ x^1A^2_1 + x^2A^2_2 + \cdots + x^nA^2_n \\ \\ \vdots \\ \\ x^1A^m_1 + x^2A^m_2 + \cdots + x^nA^m_n \end{bmatrix} = \begin{bmatrix} A^1_1x^1 + A^1_2x^2 + \cdots + A^1_nx^n \\ \\ A^2_1x^1 + A^2_2x^2 + \cdots + x^nA^2_nx^n \\ \\ \vdots \\ \\ A^m_1x^1 + A^m_2x^2 + \cdots + A^m_nx^n \end{bmatrix}$$

Let us analyze a bit the above sum equality. Note that the i-th coordinate of $A\boldsymbol{x}$ is the quantity

$$\sum_{j=1}^n A^i_jx^j = A^i_1x^1 + A^i_2x^2 + \cdots + A^i_nx^n,$$

Note also that the above expression is obtained by pairing the i-th row $A^i = [A^i_1, \dots, A^i_n]$ of the matrix $\mathcal{M}_A$ with the column vector $\boldsymbol{x} = [x^1, \dots, x^n]^\top$. Thus, the vector $A\boldsymbol{x}$ in $\mathbb{R}^m$ is described by the column of height m

$$A\boldsymbol{x} = \begin{bmatrix} \sum_{j=1}^n A^1_j x^j \\ \\ \sum_{j=1}^n A^2_j x^j \\ \vdots \\ \sum_{j=1}^n A^m_j x^j \end{bmatrix} = \begin{bmatrix} A^1 \bullet \boldsymbol{x} \\ \\ A^2 \bullet \boldsymbol{x} \\ \vdots \\ A^m \bullet \boldsymbol{x} \end{bmatrix}. \tag{11.1.14}$$

The above equality shows that each component of $A\boldsymbol{x}$ is a linear functional in $\boldsymbol{x}$.

Conversely, given an $m \times n$ matrix $\mathcal{A}$, its columns $A_1, \dots, A_n$ define vectors in $\mathbb{R}^m$ and we can use these vectors to define a linear operator $L = L_{\mathcal{A}} : \mathbb{R}^n \to \mathbb{R}^m$ via the formula

$$L_{\mathcal{A}}(\boldsymbol{x}) = x^1 A_1 + \cdots + x^n A_n, \;\; \boldsymbol{x} = (x^1, \;\; x^2, \dots, x^n).$$

In particular,

$$L_{\mathcal{A}} \boldsymbol{e}_j = A_j,$$

so that the matrix associated to the operator $L_{\mathcal{A}}$ is the matrix $\mathcal{A}$ we started with. This proves the following very useful fact.

Theorem 11.1.20. *The correspondence that associates to a linear operator $\mathbb{R}^n \to \mathbb{R}^m$ its $m \times n$ matrix is a bijection between the set of linear operators $\operatorname{Hom}(\mathbb{R}^n, \mathbb{R}^m)$ and the set $\operatorname{Mat}_{m\times n}(\mathbb{R})$ of $m \times n$ matrices with real entries.* □

Because of the above bijective correspondence we will denote a linear operator and its associated matrix by the same symbol.

Proposition 11.1.21. *Let $\ell, m, n \in \mathbb{N}$. If $A : \mathbb{R}^n \to \mathbb{R}^m$ and $B : \mathbb{R}^m \to \mathbb{R}^\ell$ are linear operators, then so is their composition $BA := B \circ A : \mathbb{R}^n \to \mathbb{R}^\ell$.*

Proof. To prove the additivity of BA we choose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then

$$BA(\boldsymbol{x} + \boldsymbol{y}) = B\big(\, A(\boldsymbol{x} + \boldsymbol{y})\,\big)$$

(use the additivity of A)

$$= B\big(\, A\boldsymbol{x} + A\boldsymbol{y}\,\big)$$

(use the additivity of B)

$$= B(A\boldsymbol{x}) + B(A\boldsymbol{y}) = BA(\boldsymbol{x}) + BA(\boldsymbol{y}).$$

The homogeneity of BA is proved in a similar fashion. □

In Proposition 11.1.21 the operator A is represented by an $m \times n$ matrix $\mathcal{M}_A$ and the operator B by an $\ell \times m$ matrix $\mathcal{M}_B$

$$\mathcal{M}_A = \begin{bmatrix} A^1_1 & A^1_2 & \cdots & A^1_n \\ \\ A^2_1 & A^2_2 & \cdots & A^2_n \\ \vdots & \vdots & \ddots & \vdots \\ A^m_1 & A^m_2 & \cdots & A^m_n \end{bmatrix}, \quad \mathcal{M}_B = \begin{bmatrix} B^1_1 & B^1_2 & \cdots & B^1_m \\ \\ B^2_1 & B^2_2 & \cdots & B^2_m \\ \vdots & \vdots & \ddots & \vdots \\ B^\ell_1 & B^\ell_2 & \cdots & B^\ell_m \end{bmatrix}.$$

The operator $BA : \mathbb{R}^n \to \mathbb{R}^\ell$ is represented by an $\ell \times n$ matrix $\mathcal{M}_{BA}$ with entries $(BA)^i_j$ that we want to describe explicitly. Note that the columns of this matrix describe the coordinates of the vectors

$$B(A\boldsymbol{e}_1), \dots, B(A\boldsymbol{e}_n) \in \mathbb{R}^\ell.$$

Thus, for $i = 1, \dots, \ell$, the entry $(BA)^i_j$ denotes the i-th coordinate of the vector $B(A\boldsymbol{e}_j)$. The vector $A\boldsymbol{e}_j$ is described by the column

$$A\boldsymbol{e}_j = A_j = \begin{bmatrix} A^1_j \\ \vdots \\ A^m_j \end{bmatrix}.$$

Since $(BA)^i_j$ is the i-th coordinate of $B(A\boldsymbol{e}_j)$, we deduce from (11.1.14) with $\boldsymbol{x} = A\boldsymbol{e}_j = A_j$ that

$$\boxed{(BA)^i_j = B^i \bullet A_j}. \tag{11.1.15}$$

More explicitly, given that $B^i = [B^i_1, \dots, B^i_m]$, we deduce from (11.1.13) with $U = B^i$ and $V = A_j$ that

$$(BA)^i_j = B^i_1 A^1_j + B^i_2 A^2_j + \cdots + B^i_m A^m_j.$$

Definition 11.1.22 (Matrix multiplication[6])**.** Given two matrices

$$A \in \mathrm{Mat}_{m \times n}(\mathbb{R}) \;\text{ and }\; B \in \mathrm{Mat}_{\ell \times m}(\mathbb{R})$$

(so that the number of columns of B is equal to the number of rows of A) their *product* is the $\ell \times n$ matrix $B \cdot A$ whose (i, j) entry is the pairing of the i-th row of B with the j-th column of A,

$$\boxed{(B \cdot A)^i_j = B^i \bullet A_j = B^i_1 A^1_j + B^i_2 A^2_j + \cdots + B^i_m A^m_j}.$$ □

The next result summarizes the above discussion.

Proposition 11.1.23. *The matrix associated to the composition of two linear operators*

$$A : \mathbb{R}^n \to \mathbb{R}^m, \quad B : \mathbb{R}^m \to \mathbb{R}^\ell$$

is the product of the matrices associated to these operators,

$$\mathcal{M}_{B \circ A} = \mathcal{M}_B \cdot \mathcal{M}_A.$$ □

[6] Check the site http://matrixmultiplication.xyz/ that interactively shows you how to multiply matrices.

Remark 11.1.24. According to Theorem 11.1.20, any matrix $A \in \operatorname{Mat}_{m\times n}(\mathbb{R})$ defines a linear operator $L_A : \mathbb{R}^n \to \mathbb{R}^m$. A vector $\boldsymbol{x} \in \mathbb{R}^n$ is represented by a column, i.e., by an $n \times 1$ matrix. The product of the matrices A and $\boldsymbol{x}$ is well defined and produces an $m \times 1$ matrix $A \cdot \boldsymbol{x}$ which can also be viewed as a vector in $\mathbb{R}^m$.

When we feed the vector $\boldsymbol{x}$ to the linear operator L_A defined by A we also obtain a vector in $\mathbb{R}^m$ given by (11.1.14)

$$L_A\boldsymbol{x} = \begin{bmatrix} A^1 \bullet \boldsymbol{x} \\ \\ A^2 \bullet \boldsymbol{x} \\ \vdots \\ A^m \bullet \boldsymbol{x} \end{bmatrix}$$

The column on the right-hand side of the above equality is none other than the matrix multiplication $A \cdot \boldsymbol{x}$, i.e.,

$$L_A\boldsymbol{x} = A \cdot \boldsymbol{x}.$$

Thus, *when viewed as a linear operator, the action of a matrix on a vector coincides with the product of that matrix with the vector viewed as a matrix consisting of a single column*.

This remarkable coincidence is one of the main reasons we prefer to think of the vectors in $\mathbb{R}^n$ as *column* vectors. □

☞ **Important Convention** *In the sequel, to ease the notational burden, we will denote with the same symbol a linear operator and its associated matrix. With this convention, the equality $L_A\boldsymbol{x} = A \cdot \boldsymbol{x}$ above takes the simper form*

$$A\boldsymbol{x} = A \cdot \boldsymbol{x}. \tag{11.1.16}$$

Also, due to Proposition 11.1.23 we will use the simpler notation BA instead of $B \cdot A$ when referring to matrix multiplication.

Example 11.1.25. (a) A linear operator $\mathbb{R} \to \mathbb{R}$ corresponds to a 1×1-matrix which in turn can be identified with a number. If A is a real number, then the associated linear operator sends a real number x to the real number Ax. Thus, the real number A is the slope of the linear function $f(x) = Ax$. This simple example shows that the matrix associated to a linear operator is a sort of "generalized slope" of the linear operator.

(b) The identity operator $\mathbb{1} : \mathbb{R}^n \to \mathbb{R}^n$ is represented by the $n \times n$ diagonal matrix

$$\mathbb{1} = \mathbb{1}_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

E.g.

$$\mathbb{1}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbb{1}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Note that the (i,j) entry of $\mathbb{1}_n$ is δ^i_j, where δ^i_j is the Kronecker symbol defined in (11.1.4). The identity operator (matrix) $\mathbb{1}_n$ has the property that

$$\mathbb{1}_n A = A\mathbb{1}_n = A, \ \ \forall A \in \operatorname{Mat}_{n\times n}(\mathbb{R}).$$

We will denote by $\mathbf{0}$ a matrix whose entries are all equal to 0.

(c) The *diagonal* of a square $n \times n$ matrix A consists of the entries $A^1_1, A^2_2, \dots, A^n_n$. For example the diagonal of the 2×2 matrix

$$A = \begin{bmatrix} \boxed{1} & 2 \\ 3 & \boxed{4} \end{bmatrix}$$

consists of the boxed entries. An $n \times n$ *diagonal matrix* is a matrix of the form

$$\begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & c_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & c_n \end{bmatrix}.$$

We will denote the above matrix by $\operatorname{Diag}(c_1, \dots, c_n)$.

(d) An $n \times n$ matrix A is called *symmetric* if $A^i_j = A^j_i$, $\forall i, j = 1, \dots, n$. For example, the matrix below is symmetric.

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}.$$

(e) We can add two matrices of the same dimensions. Thus

$$(A + B)^i_j = A^i_j + B^i_j,$$

i.e., the (i,j)-entry of $A + B$ is the sum of the (i,j)-entry of A with the (i,j)-entry of B. We can also multiply a matrix A by a scalar $c \in \mathbb{R}$. The new matrix is obtained by multiplying all entries of A by the constant c. □

Example 11.1.26. The multiplication of matrices resembles in some respects the multiplication of real numbers. For example, the multiplication of matrices is associative

$$(A \cdot B) \cdot C = A \cdot (B \cdot C)$$

for any matrices $A \in \operatorname{Mat}_{k\times\ell}(\mathbb{R})$, $B \in \operatorname{Mat}_{\ell\times m}(\mathbb{R})$, $C \in \operatorname{Mat}_{m\times n}(\mathbb{R})$. It is also distributive with respect to the addition of matrices

$$A \cdot (B + C) = AB + AC, \ \ \forall A \in \operatorname{Mat}_{\ell\times m}(\mathbb{R}), \ \ B, C \in \operatorname{Mat}_{m\times n}(\mathbb{R}).$$

However, there are some important differences. Consider for example the 2×2 matrices

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \;\; B = \begin{bmatrix} 0 & 3 \\ 0 & 4 \end{bmatrix}.$$

Observe that

$$A \cdot B = \begin{bmatrix} 0 & 3+8 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 11 \\ 0 & 0 \end{bmatrix}, \;\; B \cdot A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

□

This example shows two things.

- The multiplication of matrices is *not* commutative since obviously $AB \neq BA$ in the above example.
- The product of two matrices can be zero, although none of them is zero as in example $BA = 0$ above.

Definition 11.1.27. Suppose that $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator. The *kernel* of A, denoted by $\ker A$ is the set

$$\ker A := \big\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; A\boldsymbol{x} = \boldsymbol{0} \big\} \subset \mathbb{R}^n.$$ □

We have the following useful result whose proof is left to you as an exercise.

Proposition 11.1.28. *Suppose that $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator and $S \subset \mathbb{R}^n$ is a vector subspace. Then its kernel* $\ker A$ *is a linear subspace of* $\mathbb{R}^n$ *and the image $A(S)$ of S is a vector subspace of $\mathbb{R}^m$. In particular, the range $\boldsymbol{R}(A) := A(\mathbb{R}^n)$ is a linear subspace of $\mathbb{R}^m$.* □

Example 11.1.29. Consider the 2×3 matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

As such, it defines a linear operator $A : \mathbb{R}^3 \to \mathbb{R}^2$ described by

$$\mathbb{R}^3 \ni \boldsymbol{x} = \begin{bmatrix} x^1 \\ x^2 \\ x^3 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} x^1 \\ x^2 \\ x^3 \end{bmatrix} = \begin{bmatrix} x^1 + 2x^2 + 3x^3 \\ 4x^1 + 5x^2 + 6x^3 \end{bmatrix} \in \mathbb{R}^2.$$

If $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ is the natural basis, then $A\boldsymbol{e}_1, A\boldsymbol{e}_2, A\boldsymbol{e}_3$ are described respectively by the columns A_1, A_2, A_3 of A. E.g.,

$$A\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \in \mathbb{R}^2.$$

The kernel of this operator consists of vectors $\boldsymbol{x} = (x^1, x^2, x^3) \in \mathbb{R}^3$ satisfying $A\boldsymbol{x} = 0$, i.e., the system of linear equations

$$\begin{cases} x^1 + 2x^2 + 3x^3 & = & 0 \\ 4x^1 + 5x^2 + 6x^3 & = & 0. \end{cases}$$

If we multiply the first line above by 4 and then subtract it from the second line we deduce

$$\begin{cases} x^1+2x^2+3x^3 &= 0\\ \quad -3x^2-6x^3 &= 0 \end{cases} \Longleftrightarrow \begin{cases} x^1+2x^2+3x^3 &= 0\\ \qquad x^2+2x^3 &= 0. \end{cases}$$

We deduce that

$$x^2=-2x^3,\;\; x^1=-2x^2-3x^3=x^3.$$

If we set $t:=x^3$ we deduce that $(x^1,x^2,x^3)\in\ker A$ if and only if it has the form

$$x^1=t,\;\; x^2=-2t,\;\; x^3=t,\;\; t\in\mathbb{R}.$$

Thus the kernel of A is the line through the origin with direction vector $\boldsymbol{v}=(1,-2,1)$,

$$\ker A=\ell_{\boldsymbol{0},\boldsymbol{v}}. \qquad \square$$

11.2. Basic Euclidean geometry

The space $\mathbb{R}^n$ has a considerably richer structure than the ones we have discussed in the previous section. The goal of the present section is to describe this additional structure and some of its consequences.

Definition 11.2.1 (Inner product)**.** The *canonical inner product* in $\mathbb{R}^n$ is the map $\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ that associates to a pair of vectors $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^n\times\mathbb{R}^n$ the real *number* $\langle\boldsymbol{x},\boldsymbol{y}\rangle$ defined by

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle:=\sum_{j=1}^n x^jy^j=x^1y^1+\cdots+x^ny^n. \qquad \square$$

Proposition 11.2.2. *The inner product $\langle -,-\rangle:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ satisfies the following properties.*

(i) *For any $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{R}^n$ we have*

$$\langle\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}\rangle=\langle\boldsymbol{x},\boldsymbol{z}\rangle+\langle\boldsymbol{y},\boldsymbol{z}\rangle.$$

(ii) *For any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ and any $t\in\mathbb{R}$ we have*

$$\langle t\boldsymbol{x},\boldsymbol{y}\rangle=\langle\boldsymbol{x},t\boldsymbol{y}\rangle=t\langle\boldsymbol{x},\boldsymbol{y}\rangle.$$

(iii) *For any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ we have*

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

(iv) *For any $\boldsymbol{x}\in\mathbb{R}^n$ we have $\langle\boldsymbol{x},\boldsymbol{x}\rangle\geqslant 0$ with equality if and only if $\boldsymbol{x}=\boldsymbol{0}$.*

Proof. (i) We have

$$\langle\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}\rangle=(x^1+y^1)z^1+\cdots+(x^n+y^n)z^n=(x^1z^1+\cdots+x^nz^n)+(y^1z^1+\cdots+y^nz^n)$$
$$=\langle\boldsymbol{x},\boldsymbol{z}\rangle+\langle\boldsymbol{y},\boldsymbol{z}\rangle.$$

The properties (ii) and (iii) are obvious. As for (iv), note that

$$\langle\boldsymbol{x},\boldsymbol{x}\rangle=(x^1)^2+\cdots+(x^n)^2\geqslant 0.$$

Clearly, we have equality if and only if $x^1 = \cdots = x^n = 0$.

□

Definition 11.2.3. The *Euclidean norm* or *length* of a vector $\boldsymbol{x} = [x^1, \dots, x^n]^\top \in \mathbb{R}^n$ is the nonnegative real number $\|\boldsymbol{x}\|$ defined by

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} = \sqrt{(x^1)^2 + \cdots + (x^n)^2}.$$ □

Observe that

$$\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

The Cauchy-Schwarz inequality (8.3.16) implies that for any

$$\boldsymbol{x} = [x^1, \dots, x^n]^\top, \ \boldsymbol{y} = [y^1, \dots, y^n]^\top$$

we have

$$\left| x^1y^1 + \cdots + x^ny^n \right| \leqslant \sqrt{(x^1)^2 + \cdots (x^n)^2} \cdot \sqrt{(y^1)^2 + \cdots + (y^n)^2}.$$

This can be rewritten in the more compact form

$$\boxed{\big|\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big| \leqslant \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.} \tag{11.2.1}$$

We will refer to (11.2.1) as the Cauchy-Schwarz inequality. Given the importance of this inequality we present below an alternate proof

Alternate proof of the inequality (11.2.1). The inequality (11.2.1) obviously holds if $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{y} = \boldsymbol{0}$ so it suffices to prove it in the case $\boldsymbol{x}, \boldsymbol{y} \neq 0$. Consider the function

$$f : \mathbb{R} \to \mathbb{R}, \quad f(t) = \langle t\boldsymbol{x} + \boldsymbol{y}, t\boldsymbol{x} + \boldsymbol{y}\rangle = \|t\boldsymbol{x} + \boldsymbol{y}\|^2.$$

Clearly, $f(t) \geqslant 0$ and $f(t_0) = 0$ for some $t_0 \in \mathbb{R}$ if and only if $\boldsymbol{x}, \boldsymbol{y}$ are collinear, $\boldsymbol{y} = -t_0\boldsymbol{x}$. Using Proposition 11.2.2 we deduce

$$f(t) = \langle t\boldsymbol{x} + \boldsymbol{y}, t\boldsymbol{x}\rangle + \langle t\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{y}\rangle = t\langle t\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x}\rangle + \langle t\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{y}\rangle$$

$$= t\big(\langle t\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{y}, \boldsymbol{x}\rangle\big) + t\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle$$

$$= t^2\langle \boldsymbol{x}, \boldsymbol{x}\rangle + t\langle \boldsymbol{y}, \boldsymbol{x}\rangle + t\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle = \underbrace{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}_{a}\ t^2 + \underbrace{2\langle \boldsymbol{x}, \boldsymbol{y}\rangle}_{b}\ t + \underbrace{\langle \boldsymbol{y}, \boldsymbol{y}\rangle}_{c}$$

$$= at^2 + bt + c, \quad a > 0.$$

This shows that the quadratic polynomial $at^2 + bt + c$ with $a > 0$ is nonnegative for every $t \in \mathbb{R}$. From Exercise 3.11(a) we conclude that this is possible if and only if $b^2 - 4ac \leqslant 0$, i.e.,

$$4\big|\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big|^2 - 4\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 \leqslant 0.$$

This implies $\big|\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big| \leqslant \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$. □

Remark 11.2.4. In the above argument observe that if

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$$

then $b^2 - 4ac = 0$. In particular, this implies that there exists $t \in \mathbb{R}$ such that $t\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{0}$, i.e., the vectors are collinear. Conversely, if the vectors $\boldsymbol{x}, \boldsymbol{y}$ are collinear, then clearly $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$.

The above argument proves a bit more namely

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leqslant \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

with equality if and only if one of the vectors is a multiple of the other. □

The Cauchy-Schwarz inequality implies that for any nonzero vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

$$\frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|} \in [-1, 1].$$

Thus, there exists a unique $\theta \in [0, \pi]$ such that

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

Definition 11.2.5. The *angle* between the *nonzero* vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, denoted by $\measuredangle(\boldsymbol{x}, \boldsymbol{y})$, is defined to be the unique number $\theta \in [0, \pi]$ such that

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

Thus, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{x}, \boldsymbol{y} \neq \boldsymbol{0}$, we have

$$\boxed{\cos\measuredangle(\boldsymbol{x}, \boldsymbol{y}) = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|}} \text{ and } \boxed{\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \cos\measuredangle(\boldsymbol{x}, \boldsymbol{y})}. \tag{11.2.2}$$

Classically, two nonzero vectors $\boldsymbol{x}, \boldsymbol{y}$ are orthogonal if $\measuredangle(\boldsymbol{x}, \boldsymbol{y}) = \frac{\pi}{2}$, i.e., $\cos\measuredangle(\boldsymbol{x}, \boldsymbol{y}) = 0$. The equality (11.2.2) shows that this happens iff $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0$. This justifies our next definition.

Definition 11.2.6. We say that two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are *orthogonal*, and we write this $\boldsymbol{x} \perp \boldsymbol{y}$, if $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0$. □

Example 11.2.7. If $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ is the canonical basis of $\mathbb{R}^n$ (see (11.1.2)), then

$$\|\boldsymbol{e}_1\| = \cdots = \|\boldsymbol{e}_n\| = 1,$$

and

$$\boldsymbol{e}_i \perp \boldsymbol{e}_j, \quad \forall i \neq j.$$

We can rewrite these facts in the more succinct form

$$\langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle = \delta_{ij} := \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

The collection (δ_{ij}) above is also called *Kronecker symbol.* Note that for any vector

$$\boldsymbol{x}=\begin{bmatrix} x^1\\ \vdots\\ x^n\end{bmatrix}\in\mathbb{R}^n$$

we have

$$x^i=\langle\boldsymbol{x},\boldsymbol{e}_i\rangle,\ \ \forall i=1,2,\dots,n,$$

and thus

$$\boldsymbol{x}=\langle\boldsymbol{x},\boldsymbol{e}_1\rangle\boldsymbol{e}_1+\cdots+\langle\boldsymbol{x},\boldsymbol{e}_n\rangle\boldsymbol{e}_n.$$

□

Theorem 11.2.8 (Pythagoras)**.** *If $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ and $\boldsymbol{x}\perp\boldsymbol{y}$, then*

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

Proof. We have

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

$$=\langle\boldsymbol{x},\boldsymbol{x}\rangle+\underbrace{\langle\boldsymbol{x},\boldsymbol{y}\rangle+\langle\boldsymbol{y},\boldsymbol{x}\rangle}_{=0}+\langle\boldsymbol{y},\boldsymbol{y}\rangle=\langle\boldsymbol{x},\boldsymbol{x}\rangle+\langle\boldsymbol{y},\boldsymbol{y}\rangle=\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2.$$

□

Observe that any vector $\boldsymbol{x}\in\mathbb{R}^n$ defines a linear functional

$$\boldsymbol{x}^{\downarrow}:\mathbb{R}^n\to\mathbb{R},\ \ \boldsymbol{x}^{\downarrow}(\boldsymbol{y}):=\langle\boldsymbol{x},\boldsymbol{y}\rangle.$$

We will refer to the functional $\boldsymbol{x}^{\downarrow}$ as the *dual* of $\boldsymbol{x}$. It is not hard to see that all the linear functionals on $\mathbb{R}^n$ are duals of vectors in $\mathbb{R}^n$.

Proposition 11.2.9. *Let $n\in\mathbb{N}$. Any linear functional $\boldsymbol{\xi}:\mathbb{R}^n\to\mathbb{R}$ is the dual of a unique vector in $\mathbb{R}^n$. This means that there exists a unique vector $\boldsymbol{z}\in\mathbb{R}^n$ such that $\boldsymbol{\xi}=\boldsymbol{z}^{\downarrow}$, i.e.,*

$$\boldsymbol{\xi}(\boldsymbol{x})=\langle\boldsymbol{z},\boldsymbol{x}\rangle,\ \ \forall\boldsymbol{x}\in\mathbb{R}^n. \tag{11.2.3}$$

This unique vector $\boldsymbol{z}$ is called the dual *of $\boldsymbol{\xi}$ and it is denoted by $\boldsymbol{\xi}_{\uparrow}$.*

Proof. Let $\boldsymbol{e}_1,\dots,\boldsymbol{e}_n$ be the canonical basis of $\mathbb{R}^n$. Set

$$\xi_i:=\boldsymbol{\xi}(\boldsymbol{e}_i),\ \ i=1,2,\dots,n,$$

The vector $\boldsymbol{z}=[z^1,\dots,z^n]^\top$ satisfies (11.2.3) if and only if

$$z^i=\langle\boldsymbol{z},\boldsymbol{e}_i\rangle=\boldsymbol{\xi}(\boldsymbol{e}_i)=\xi_i,\ \ i=1,2,\dots,n.$$

□

The above proof shows that, if the *linear form* $\boldsymbol{\xi}$ is described by the *row*

$$\boldsymbol{\xi} = [\xi_1, \dots, \xi_n],$$

then $\boldsymbol{\xi}_\uparrow$ is the *vector* described by the *column*

$$\boldsymbol{\xi}_\uparrow = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \Longleftrightarrow \boldsymbol{\xi}^i_\uparrow = \xi_i, \tag{11.2.4a}$$

$$\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\xi} \bullet \boldsymbol{x} = \langle \boldsymbol{\xi}_\uparrow, \boldsymbol{x} \rangle, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{11.2.4b}$$

Note that

$$(\boldsymbol{e}_i)^\downarrow = \boldsymbol{e}^i, \;\; (\boldsymbol{e}^j)_\uparrow = \boldsymbol{e}_j, \;\; \forall i, j = 1, \dots, n. \tag{11.2.5}$$

The duality operation defined above has a very simple intuitive description: it takes a row $\boldsymbol{\xi}$ and transforms into a column $\boldsymbol{\xi}_\uparrow$ with the same entries, and vice-versa, it takes a column $\boldsymbol{x}$ and transforms it into a row $\boldsymbol{x}^\downarrow$ with the same entries. E.g.,

$$[1, -2, 3]_\uparrow = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}, \;\; \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}^\downarrow = [4, 5, 6].$$

Proposition 11.2.10. *Let $n \in \mathbb{N}$ and $H \subset \mathbb{R}^n$. The following statements are equivalent.*

(i) *The subset H is a hyperplane.*

(ii) *There exists a nonzero vector $\boldsymbol{N} \in \mathbb{R}^n$ and a constant $c \in \mathbb{R}$ such that $\boldsymbol{p} \in H$ if and only if $\langle \boldsymbol{N}, \boldsymbol{p} \rangle = c$.*

Proof. (i) ⇒ (ii) Since H is a hyperplane there exists a nonzero linear functional $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ and a real number c such that

$$\boldsymbol{x} \in H \Longleftrightarrow \boldsymbol{\xi}(\boldsymbol{x}) = c.$$

Let $\boldsymbol{N} := \boldsymbol{\xi}_\uparrow$, i.e., $\langle \boldsymbol{N}, \boldsymbol{x} \rangle = \boldsymbol{\xi}(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. Then, for any $\boldsymbol{p}, \boldsymbol{q} \in H$, we have

$$\langle \boldsymbol{N}, \boldsymbol{p} \rangle = \boldsymbol{\xi}(\boldsymbol{p}) = c = \boldsymbol{\xi}(\boldsymbol{q}) = \langle \boldsymbol{N}, \boldsymbol{q} \rangle.$$

(ii) ⇒ (i) Let $\boldsymbol{\xi} := \boldsymbol{N}^\downarrow$. Then

$$\boldsymbol{p} \in H \Longleftrightarrow \langle \boldsymbol{N}, \boldsymbol{p} \rangle = c \Longleftrightarrow \boldsymbol{\xi}(\boldsymbol{p}) = c.$$

This shows that H is a hyperplane. □

Suppose that $H \subset \mathbb{R}^n$ is a hyperplane. Hence, there exist $\boldsymbol{N} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ and $c \in \mathbb{R}$ such that

$$\boldsymbol{x} \in H \Longleftrightarrow \langle \boldsymbol{N}, \boldsymbol{x} \rangle = c.$$

If $\boldsymbol{p}, \boldsymbol{q} \in H$ and $\boldsymbol{p} \neq \boldsymbol{q}$, then the direction of the line $\boldsymbol{pq}$ is given by the vector $\boldsymbol{q} - \boldsymbol{p}$. Now observe that

$$\langle \boldsymbol{N}, \boldsymbol{q} - \boldsymbol{p} \rangle = \langle \boldsymbol{N}, \boldsymbol{q} \rangle - \langle \boldsymbol{N}, \boldsymbol{p} \rangle = 0 \Rightarrow \boldsymbol{N} \perp (\boldsymbol{q} - \boldsymbol{p}).$$

Thus, the defining vector $\boldsymbol{N}$ is *perpendicular to all the lines contained in H*. We say that $\boldsymbol{N}$ is orthogonal to H, we write this $\boldsymbol{N} \perp H$ and we will to refer to $\boldsymbol{N}$ as *a normal vector* of H.

Example 11.2.11. (a) As we have mentioned earlier, any line in $\mathbb{R}^2$ is also an affine hyperplane. For example, the line given by the equation $2x + 3y = 5$ admits the vector $\boldsymbol{N} = (2,3)$ as normal vector.

(b) If $n \in \mathbb{N}$, then for any $\boldsymbol{p} \in \mathbb{R}^n$ and any $\boldsymbol{N} \in \mathbb{R}^n$, $\boldsymbol{N} \neq \boldsymbol{0}$, we denote by $H_{\boldsymbol{p},\boldsymbol{N}}$ the *hyperplane through $\boldsymbol{p}$ and normal $\boldsymbol{N}$*, i.e., the hyperplane

$$H_{\boldsymbol{p},\boldsymbol{N}} = \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ \langle \boldsymbol{N}, \boldsymbol{x} \rangle = \langle \boldsymbol{N}, \boldsymbol{p} \rangle \big\}.$$

Clearly $\boldsymbol{p} \in H$. For example if $n = 3$, $\boldsymbol{p} = (1,1,1)$ and $\boldsymbol{N} = (1,2,3)$, then

$$\langle \boldsymbol{N}, \boldsymbol{p} \rangle = 1 + 2 + 3 = 6,$$

and

$$H_{\boldsymbol{p},\boldsymbol{N}} = \big\{ (x,y,z) \in \mathbb{R}^3;\ \ x + 2y + 3z = 6 \big\}. \qquad \square$$

Example 11.2.12 (The cross product in $\mathbb{R}^3$)**.** The 3-dimensional Euclidean space $\mathbb{R}^3$ is equipped with another operation that is not available in any other dimensions. The *cross product* is the map

$$\times : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3,\ \ (\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u} \times \boldsymbol{v}$$

uniquely characterized by the following conditions

(i) $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$

$$(\boldsymbol{u} + \boldsymbol{v}) \times \boldsymbol{w} = (\boldsymbol{u} \times \boldsymbol{w}) + (\boldsymbol{v} \times \boldsymbol{w}),$$
$$\boldsymbol{w} \times (\boldsymbol{u} + \boldsymbol{v}) = (\boldsymbol{w} \times \boldsymbol{u}) + (\boldsymbol{w} \times \boldsymbol{v}).$$

(ii)

$$(t\boldsymbol{u}) \times \boldsymbol{v} = \boldsymbol{u} \times (t\boldsymbol{v}) = t(\boldsymbol{u} \times \boldsymbol{v}),\ \ \forall t \in \mathbb{R},\ \ \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3.$$

(iii)

$$\boldsymbol{u} \times \boldsymbol{v} = -(\boldsymbol{v} \times \boldsymbol{u}),\ \ \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3.$$

(iv)

$$\boldsymbol{e}_1 \times \boldsymbol{e}_2 = \boldsymbol{e}_3,\ \ \boldsymbol{e}_2 \times \boldsymbol{e}_3 = \boldsymbol{e}_1,\ \ \boldsymbol{e}_3 \times \boldsymbol{e}_1 = \boldsymbol{e}_2.$$

Note that (iii) implies that

$$\boldsymbol{u} \times \boldsymbol{u} = \boldsymbol{0},\ \ \forall \boldsymbol{u} \in \mathbb{R}^3.$$

Indeed

$$\boldsymbol{u} \times \boldsymbol{u} = -(\boldsymbol{u} \times \boldsymbol{u}) \Rightarrow 2(\boldsymbol{u} \times \boldsymbol{u}) = \boldsymbol{0} \Rightarrow \boldsymbol{u} \times \boldsymbol{u} = \boldsymbol{0}.$$

For example, if

$$\boldsymbol{u} = [1,2,3]^\top,\ \ \boldsymbol{v} = [4,5,6]^\top,$$

then

$$\boldsymbol{u} \times \boldsymbol{v} = (\boldsymbol{e}_1 + 2\boldsymbol{e}_2 + 3\boldsymbol{e}_3) \times (4\boldsymbol{e}_1 + 5\boldsymbol{e}_2 + 6\boldsymbol{e}_3)$$

$$
\begin{aligned}
&= \underbrace{\boldsymbol{e}_1 \times (4\boldsymbol{e}_1 + 5\boldsymbol{e}_2 + 6\boldsymbol{e}_3)}_{I} + \underbrace{2\boldsymbol{e}_2 \times (4\boldsymbol{e}_1 + 5\boldsymbol{e}_2 + 6\boldsymbol{e}_3)}_{II} + \underbrace{3\boldsymbol{e}_3 \times (4\boldsymbol{e}_1 + 5\boldsymbol{e}_2 + 6\boldsymbol{e}_3)}_{III}\\
&= \underbrace{5\boldsymbol{e}_1 \times \boldsymbol{e}_2 + 6\boldsymbol{e}_1 \times \boldsymbol{e}_3}_{I} + \underbrace{8\boldsymbol{e}_2 \times \boldsymbol{e}_1 + 12\boldsymbol{e}_2 \times \boldsymbol{e}_3}_{II} + \underbrace{12\boldsymbol{e}_3 \times \boldsymbol{e}_1 + 15\boldsymbol{e}_3 \times \boldsymbol{e}_2}_{III}\\
&= \underbrace{(5\boldsymbol{e}_3 - 6\boldsymbol{e}_2)}_{I} + \underbrace{(-8\boldsymbol{e}_3 + 12\boldsymbol{e}_1)}_{II} + \underbrace{(12\boldsymbol{e}_2 - 15\boldsymbol{e}_1)}_{III}\\
&= -3\boldsymbol{e}_1 + 6\boldsymbol{e}_2 - 3\boldsymbol{e}_3 = [-3, 6, -3|^{\top}.
\end{aligned}
$$

If we set $\boldsymbol{w} = \boldsymbol{u} \times \boldsymbol{v} = [-3, 6, -3]^{\top}$, then we observe that

$$\langle \boldsymbol{w}, \boldsymbol{u} \rangle = \langle \boldsymbol{w}, \boldsymbol{v} \rangle = 0.$$

We have

$$
\begin{gathered}
\|\boldsymbol{u}\| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}, \;\; \|\boldsymbol{v}\| = \sqrt{4^2 + 5^2 + 6^2} = \sqrt{77},\\
\|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\| = \sqrt{14 \cdot 77} = \sqrt{1078},\\
\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 4 + 10 + 18 = 32.
\end{gathered}
$$

If we denote by θ the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$, then we deduce

$$\cos\theta = \frac{32}{\sqrt{1078}}.$$

Hence

$$\sin^2\theta = 1 - \cos^2\theta = \frac{54}{1078}.$$

Note that

$$\|\boldsymbol{u} \times \boldsymbol{v}\| = \sqrt{3^2 + 6^2 + 3^2} = \sqrt{54},$$

This proves that

$$\boldsymbol{u}, \boldsymbol{v} \perp (\boldsymbol{u} \times \boldsymbol{v}), \;\; \|\boldsymbol{u} \times \boldsymbol{v}\| = \sqrt{54} = \sqrt{1078} \cdot \sqrt{\frac{54}{1078}} = \|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\| \sin\theta.$$

Let us observe that the quantity $\|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\| \sin\theta$ is the area of the parallelogram spanned by the vectors $\boldsymbol{u}, \boldsymbol{v}$.

The above observations are manifestations of a more general phenomenon. Given any two vectors

$$\boldsymbol{u} = [u^1, u^2, u^3]^{\top}, \;\; \boldsymbol{v} = [v^1, v^2, v^3]^{\top} \in \mathbb{R}^3,$$

then the properties(i)-(iv) show that[7]

$$\boxed{\boldsymbol{u} \times \boldsymbol{v} = (u^2v^3 - u^3v^2)\boldsymbol{e}_1 + (u^3v^1 - u^1v^3)\boldsymbol{e}_2 + (u^1v^2 - u^2v^1)\boldsymbol{e}_3}. \tag{11.2.6}$$

Using this equality one can show that $\boldsymbol{u} \times \boldsymbol{v}$ is a vector perpendicular to both $\boldsymbol{u}$ and $\boldsymbol{v}$ and its length is equal to the area of the parallelogram spanned by the vectors $\boldsymbol{u}, \boldsymbol{v}$. These facts alone almost completely determine the vector $\boldsymbol{u} \times \boldsymbol{v}$. There are two vectors with these properties, and to determine which is the cross product we need to indicate the direction or orientation of this vector. This is achieved using the *right-hand rule*.

[7] Do not try to memorize (11.2.6). Use (i)-(iv) whenever you want to compute a cross product.

☛ Align your right hand thumb with the vector $\boldsymbol{u}$ and your right hand index with the vector $\boldsymbol{v}$. If you then move the right hand middle-finger so it is perpendicular to your right-hand palm, then it will be aligned with $\boldsymbol{u}\times\boldsymbol{v}$. □

Definition 11.2.13. Suppose that $V\subset\mathbb{R}^n$ is a vector subspace. Its *orthogonal complement* is the subset

$$V^\perp := \big\{\,\boldsymbol{u}\in\mathbb{R}^n;\ \langle\boldsymbol{u},\boldsymbol{v}\rangle=0,\ \ \forall\boldsymbol{v}\in V\,\big\}.$$ □

11.3. Basic Euclidean topology

The notions of convergence and continuity on the real axis have a multidimensional counterpart. The main reason why this happens is because the Euclidean norm $\|-\|$ behaves like the absolute value on $\mathbb{R}$. Observe first that

$$\|t\boldsymbol{x}\| = |t|\cdot\|\boldsymbol{x}\|,\ \ \forall\boldsymbol{x}\in\mathbb{R}^n,\ \ t\in\mathbb{R} \tag{11.3.1a}$$

$$\|\boldsymbol{x}\|\geqslant 0,\ \ \|\boldsymbol{x}\|=0\Longleftrightarrow\boldsymbol{x}=\boldsymbol{0}, \tag{11.3.1b}$$

Additionally, and less trivially, we have the following key result.

Theorem 11.3.1 (Triangle inequality)**.** *Let $n\in\mathbb{N}$. For any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ we have*

$$\|\boldsymbol{x}+\boldsymbol{y}\|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|. \tag{11.3.2a}$$

$$\Big|\,\|\boldsymbol{x}\|-\|\boldsymbol{y}\|\,\Big|\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|. \tag{11.3.2b}$$

Proof. Observe that

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

(use the Cauchy-Schwarz inequality)

$$\leqslant\|\boldsymbol{x}\|^2+2\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|+\|\boldsymbol{y}\|^2=\big(\,\|\boldsymbol{x}\|+\|\boldsymbol{y}\|\,\big)^2.$$

Hence

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2\leqslant\big(\,\|\boldsymbol{x}\|+\|\boldsymbol{y}\|\,\big)^2.$$

This proves (11.3.2a).

Next, observe that (11.3.2a) implies

$$\|\boldsymbol{x}\|=\|\boldsymbol{y}+(\boldsymbol{x}-\boldsymbol{y})\|\leqslant\|\boldsymbol{y}\|+\|\boldsymbol{x}-\boldsymbol{y}\|\Rightarrow\|\boldsymbol{x}\|-\|\boldsymbol{y}\|\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|.$$

Similarly

$$\|\boldsymbol{y}\|=\|\boldsymbol{x}+(\boldsymbol{y}-\boldsymbol{x})\|\leqslant\|\boldsymbol{x}\|+\|(\boldsymbol{y}-\boldsymbol{x})\|=\|\boldsymbol{x}\|+\|\boldsymbol{x}-\boldsymbol{y}\|$$
$$\Rightarrow\|\boldsymbol{y}\|-\|\boldsymbol{x}\|\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|.$$

Hence

$$\pm\big(\,\|\boldsymbol{x}\|-\|\boldsymbol{y}\|\,\big)\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|.$$

This is clearly equivalent to (11.3.2b). □

Definition 11.3.2 (Euclidean distance)**.** Let $n \in \mathbb{N}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The *Euclidean distance* between the points $\boldsymbol{x}, \boldsymbol{y}$ is the nonnegative real number

$$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|. \qquad \square$$

Example 11.3.3. (a) If $n = 1$, then for any $x, y \in \mathbb{R}$ we have $\operatorname{dist}(x, y) = |x - y|$.
(b) For any $n \in \mathbb{N}$ and any $\boldsymbol{x} \in \mathbb{R}^n$ we have $\|\boldsymbol{x}\| = \operatorname{dist}(\boldsymbol{x}, \boldsymbol{0})$. □

Proposition 11.3.4. *Let $n \in \mathbb{N}$. For any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ the following hold.*

(i) $\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) \geqslant 0$ *with equality if and only if* $\boldsymbol{x} = \boldsymbol{y}$.

(ii) $\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{dist}(\boldsymbol{y}, \boldsymbol{x})$.

(iii) (Triangle inequality) $\operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) \leqslant \operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) + \operatorname{dist}(\boldsymbol{y}, \boldsymbol{z})$.

Proof. We have

$$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \big\|-(\boldsymbol{x} - \boldsymbol{y})\big\| = \|\boldsymbol{y} - \boldsymbol{x}\| = \operatorname{dist}(\boldsymbol{y}, \boldsymbol{x}) \geqslant 0.$$

Clearly

$$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) = 0 \Longleftrightarrow \|\boldsymbol{x} - \boldsymbol{y}\| = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{y}.$$

To prove (iii) note that

$$\begin{aligned} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) &= \|\boldsymbol{x} - \boldsymbol{z}\| = \|(\boldsymbol{x} - \boldsymbol{y}) + (\boldsymbol{y} - \boldsymbol{z})\| \\ &\overset{(11.3.2a)}{\leqslant} \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{y} - \boldsymbol{z}\| = \operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) + \operatorname{dist}(\boldsymbol{y}, \boldsymbol{z}). \end{aligned}$$

□

Definition 11.3.5 (Open sets)**.** Let $n \in \mathbb{N}$.

(i) For $r > 0$ and $\boldsymbol{p} \in \mathbb{R}^n$ we define the *open (Euclidean) ball of radius r and center $\boldsymbol{p}$* to be the set

$$B_r(\boldsymbol{p}) := \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \operatorname{dist}(\boldsymbol{x}, \boldsymbol{p}) < r \big\} = \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \|\boldsymbol{x} - \boldsymbol{p}\| < r \big\}. \tag{11.3.3}$$

Sometimes, when we want to emphasize the ambient space $\mathbb{R}^n$ we will use the more precise notation $B^n_r(\boldsymbol{p})$ when referring to the open ball in $\mathbb{R}^n$ of radius r and center $\boldsymbol{p}$.

(ii) A set $U \subset \mathbb{R}^n$ is called *open* (in $\mathbb{R}^n$) if, for any $\boldsymbol{p} \in U$, there exists $r > 0$ such that $B_r(\boldsymbol{p}) \subset U$.

(iii) An *open neighborhood* of $\boldsymbol{x}_0$ in $\mathbb{R}^n$ is defined to be an open subset of $\mathbb{R}^n$ that contains $\boldsymbol{x}_0$.

□

Example 11.3.6. For any real numbers $a < b$, the intervals (a, b), $(-\infty, a)$ and (a, ∞) are open subsets of $\mathbb{R}$. □

Proposition 11.3.7. *Let $n \in \mathbb{N}$. Then, for any $\boldsymbol{p} \in \mathbb{R}^n$ and any $r > 0$, the open ball $B_r(\boldsymbol{p})$ is an open subset of $\mathbb{R}^n$.*

Proof. Let $r > 0$ and $\boldsymbol{p} \in \mathbb{R}^n$. Given $\boldsymbol{q} \in B_r(\boldsymbol{p})$ let $\rho := \operatorname{dist}(\boldsymbol{p}, \boldsymbol{q})$. Note that $\rho < r$. We claim that $B_{r-\rho}(\boldsymbol{q}) \subset B_r(\boldsymbol{p})$. Indeed, if $\boldsymbol{x} \in B_{r-\rho}(\boldsymbol{q})$, then $\operatorname{dist}(\boldsymbol{q}, \boldsymbol{x}) < r - \rho$. Using the triangle inequality we deduce

$$\operatorname{dist}(\boldsymbol{p}, \boldsymbol{x}) \leqslant \operatorname{dist}(\boldsymbol{p}, \boldsymbol{q}) + \operatorname{dist}(\boldsymbol{q}, \boldsymbol{x}) < \rho + (r - \rho) = r.$$

This proves that $\boldsymbol{x} \in B_r(\boldsymbol{p})$. □

Proposition 11.3.8. *Let $n \in \mathbb{N}$. Then the following hold.*

(i) *The empty set and the whole space $\mathbb{R}^n$ are open subsets of $\mathbb{R}^n$.*

(ii) *The intersection of two open subsets of $\mathbb{R}^n$ is also an open subset of $\mathbb{R}^n$.*

(iii) *The union of a (possibly infinite) collection of open subsets of $\mathbb{R}^n$ is also an open subset of $\mathbb{R}^n$.*

Proof. The statement (i) is obvious. To prove (ii) consider two open subsets $U_1, U_2 \subset \mathbb{R}^n$. We have to show that $U_1 \cap U_2$ is open, i.e., for any $\boldsymbol{p} \in U_1 \cap U_2$ there exists $r > 0$ such that $B_r(\boldsymbol{p}) \subset U_1 \cap U_2$.

Since U_1 is open, there exists $r_1 > 0$ such that $B_{r_1}(\boldsymbol{p}) \subset U_1$. Similarly, there exists $r_2 > 0$ such that $B_{r_2}(\boldsymbol{p}) \subset U_2$. If $r = \min(r_1, r_2)$, then

$$B_r(\boldsymbol{p}) = B_{r_1}(\boldsymbol{p}) \cap B_{r_2}(\boldsymbol{p}) \subset U_1 \cap U_2.$$

(iii) Suppose that $(U_i)_{i \in I}$ is a collection of open subsets of $\mathbb{R}^n$. Denote by U their union. If $\boldsymbol{p} \in U$, then there exists a set U_{i_0} of this collection that contains $\boldsymbol{p}$. Since U_{i_0} is open, there exists $r_0 > 0$ such that

$$B_{r_0}(\boldsymbol{p}) \subset U_{i_0} \subset U.$$

This proves that U is open. □

Definition 11.3.9. Let $n \in \mathbb{N}$. For any $\boldsymbol{x} = [x^1, \dots, x^n]^\top \in \mathbb{R}^n$ we set

$$\|\boldsymbol{x}\|_\infty := \max\{\, |x^1|, \dots, |x^n| \,\}.$$

We will refer to $\|\boldsymbol{x}\|_\infty$ as the *sup-norm* of $\boldsymbol{x}$. □

Example 11.3.10. If $\boldsymbol{x} = [3, 1, -7, 5]^\top \in \mathbb{R}^4$, then

$$\|\boldsymbol{x}\|_\infty = 7 \ \text{ and } \ \|\boldsymbol{x}\| = \sqrt{9 + 1 + 49 + 25} = \sqrt{84}.$$ □

The proof of the following result is left to you as an exercise.

Proposition 11.3.11. *Let $n \in \mathbb{N}$. Then*

$$\|\boldsymbol{x} + \boldsymbol{y}\|_\infty \leqslant \|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \tag{11.3.4a}$$

$$\Big|\, \|\boldsymbol{x}\|_\infty - \|\boldsymbol{y}\|_\infty \,\Big| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \tag{11.3.4b}$$

and

$$\|\boldsymbol{x}\|_\infty \leqslant \|\boldsymbol{x}\| \leqslant \sqrt{n}\|\boldsymbol{x}\|_\infty, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{11.3.5}$$

□

Definition 11.3.12. Let $n \in \mathbb{N}$. For any $\boldsymbol{p} \in \mathbb{R}^n$ and $r > 0$ we define the *open cube* of center $\boldsymbol{p}$ and radius r to be the set

$$C_r(\boldsymbol{p}) := \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x} - \boldsymbol{p}\|_\infty < r \,\big\}.$$

□

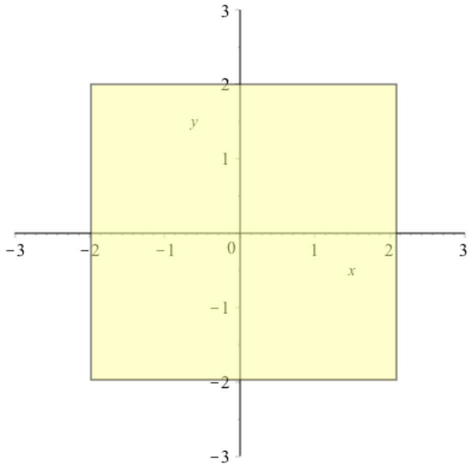

Figure 11.9. *The open cube* $C_2(\mathbf{0})$ *of radius* 2 *and center* $\mathbf{0} \in \mathbb{R}^2$.

Observe that if $\boldsymbol{p} = [p^1, \dots, p^n]^\top \in \mathbb{R}^n$ and $r > 0$ then

$$\begin{aligned} \boldsymbol{x} \in C_r(\boldsymbol{p}) &\Longleftrightarrow |x^i - p^i| < r, \quad \forall i = 1, 2, \dots, n \\ &\Longleftrightarrow x^i \in (p^i - r, p^i + r), \quad \forall i = 1, 2, \dots, n \\ &\Longleftrightarrow \boldsymbol{x} \in (p^1 - r, p^1 + r) \times (p^2 - r, p^2 + r) \times \cdots \times (p^n - r, p^n + r). \end{aligned}$$

Note that the inequality (11.3.5) implies that

$$\forall \boldsymbol{p} \in \mathbb{R}^n, \quad \forall r > 0 : \quad C_{r/\sqrt{n}}(\boldsymbol{p}) \subset B_r(\boldsymbol{p}) \subset C_r(\boldsymbol{p}). \tag{11.3.6}$$

Proposition 11.3.13. *For any $n \in \mathbb{N}$, $\boldsymbol{p} \in \mathbb{R}^n$ and $r > 0$ the open cube $C_r(\boldsymbol{p})$ is an* open *subset of* $\mathbb{R}^n$. □

The proof is left to you as an exercise.

Proposition 11.3.14. *Let $n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$. The following statements are equivalent.*

(i) *The set U is open.*

(ii) *For all $\boldsymbol{p} \in U$, $\exists r > 0$ such that $C_r(\boldsymbol{p}) \subset U$.*

□

Definition 11.3.15 (Closed sets)**.** Let $n \in \mathbb{N}$. A subset $C \subset \mathbb{R}^n$ is called *closed* (in $\mathbb{R}^n$) if its complement $\mathbb{R}^n \backslash C$ is open in $\mathbb{R}^n$. More explicitly, this means that

$$\forall \boldsymbol{p} \in \mathbb{R}^n \backslash C \;\; \exists r > 0 : \;\; B_r(\boldsymbol{p}) \subset \mathbb{R}^n \backslash C.$$ □

Example 11.3.16. (a) For any real numbers $a < b$, the intervals $[a, b]$, $(-\infty, b]$, $[b, \infty)$ are closed subsets of $\mathbb{R}$.

(b) For $\boldsymbol{p} \in \mathbb{R}^n$ and $r > 0$ we set

$$\overline{B_r(\boldsymbol{p})} := \big\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; \|\boldsymbol{x} - \boldsymbol{p}\| \leqslant r \big\}.$$

Then $\overline{B_r(\boldsymbol{p})}$ is a closed subset of $\mathbb{R}^n$, i.e., $\mathbb{R}^n \backslash \overline{B_r(\boldsymbol{p})}$ is open.

Indeed, let $\boldsymbol{q} \in \mathbb{R}^n \backslash \overline{B_r(\boldsymbol{p})}$. Thus $\|\boldsymbol{q} - \boldsymbol{p}\| > r$. Set $R = \|\boldsymbol{q} - \boldsymbol{p}\|$. We claim that

$$B_{R-r}(\boldsymbol{q}) \subset \mathbb{R}^n \backslash \overline{B_r(\boldsymbol{p})}.$$

Let $\boldsymbol{y} \in B_{R-r}(\boldsymbol{q})$. We have

$$R = \|\boldsymbol{p} - \boldsymbol{q}\| \leqslant \|\boldsymbol{p} - \boldsymbol{y}\| + \|\boldsymbol{y} - \boldsymbol{q}\| < \|\boldsymbol{p} - \boldsymbol{y}\| + R - r \Rightarrow r < \|\boldsymbol{p} - \boldsymbol{y}\|$$

$$\Rightarrow \boldsymbol{y} \in \mathbb{R}^n \backslash \overline{B_r(\boldsymbol{p})}.$$

(c) For $\boldsymbol{p} \in \mathbb{R}^n$ and $r > 0$ we set

$$\overline{C_r(\boldsymbol{p})} := \big\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; \|\boldsymbol{x} - \boldsymbol{p}\|_\infty \leqslant r \big\}.$$

Then $\overline{C_r(\boldsymbol{p})}$ is a closed subset of $\mathbb{R}^n$. To prove this fact, imitate the argument in (b) with the Euclidean norm $\| - \|$ replaced by the sup-norm $\| - \|_\infty$ and then invoke Proposition 11.3.14. □

Definition 11.3.17. The sets $\overline{B_r(\boldsymbol{p})}$ and $\overline{C_r(\boldsymbol{p})}$ are called the *closed ball and respectively closed cube* of center $\boldsymbol{p}$ and radius r. □

According to the De Morgan law (Proposition 1.3.2) the complement of a union of sets is the intersection of the complements of the sets, and the complement of an intersection of sets is the union of the complements of the sets. Invoking Proposition 11.3.8 we deduce the following result.

Proposition 11.3.18. *Let $n \in \mathbb{N}$. The following hold.*

(i) *The empty set and the whole space $\mathbb{R}^n$ are closed subsets of $\mathbb{R}^n$.*

(ii) *The union of two closed subsets of $\mathbb{R}^n$ is also a closed subset of $\mathbb{R}^n$.*

(iii) *The intersection of a (possibly infinite) collection of closed subsets of $\mathbb{R}^n$ is also a closed subset of $\mathbb{R}^n$.*

□

11.4. Convergence

The concept of convergence of sequences of real numbers has a multidimensional counterpart. In fact, the concept of convergence of a sequence of points in a Euclidean space $\mathbb{R}^n$ can be expressed in terms of the concept of convergence of sequences of real numbers.

Definition 11.4.1 (Convergent sequences)**.** Let $n \in \mathbb{N}$. A sequence $(\boldsymbol{p}_\nu)_{\nu \geqslant 1}$ of points in $\mathbb{R}^n$ is said to be *convergent* if there exists $\boldsymbol{p}_\infty$ such that the sequence of real numbers $\big(\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty)\big)_{\nu \geqslant 1}$ converges to 0,

$$\lim_{\nu \to \infty} \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) = 0.$$

More precisely, this means that $\forall \varepsilon > 0$, $\exists N = N(\varepsilon) > 0$ such that $\forall \nu > N(\varepsilon)$ we have $\|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\| < \varepsilon$. The point $\boldsymbol{p}_\infty$ is called the *limit* of the sequence $(\boldsymbol{p}_\nu)$ and we write this

$$\boldsymbol{p}_\infty = \lim_{\nu \to \infty} \boldsymbol{p}_\nu.$$

□

Note that

$$\boldsymbol{p}_\infty = \lim_{\nu \to \infty} \boldsymbol{p}_\nu \Longleftrightarrow \lim_{\nu \to \infty} \|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\| = 0 \Longleftrightarrow \lim_{\nu \to \infty} \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) = 0. \tag{11.4.1}$$

The notion of convergence can be expressed in terms of open balls because the statement "$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{p}) < \varepsilon$" is equivalent to the statement: "the point $\boldsymbol{x}$ belongs to the open ball of center $\boldsymbol{p}$ and radius ε". More precisely, we have the following result.

Proposition 11.4.2. *Let $n \in \mathbb{N}$ and $(\boldsymbol{p}_\nu)$ a sequence of points in $\mathbb{R}^n$. The following statements are equivalent.*

(i)
$$\boldsymbol{p}_\infty = \lim_{\nu \to \infty} \boldsymbol{p}_\nu.$$

(ii) *For any $\varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that, $\forall \nu > N(\varepsilon)$ we have $\boldsymbol{p}_\nu \in B_\varepsilon(\boldsymbol{p}_\infty)$.*

□

The proof of the next result is left to you as an exercise.

Proposition 11.4.3. *Let $n \in \mathbb{N}$. Consider a sequence of points in $\mathbb{R}^n$*

$$\boldsymbol{p}_\nu = \begin{bmatrix} p_\nu^1 \\ \vdots \\ p_\nu^n \end{bmatrix}, \quad \nu = 1, 2, \dots,$$

and

$$\boldsymbol{p}_\infty = \begin{bmatrix} p_\infty^1 \\ \vdots \\ p_\infty^n \end{bmatrix} \in \mathbb{R}^n.$$

The following statements are equivalent.

(i)

$$\lim_{\nu\to\infty} \boldsymbol{p}_\nu = \boldsymbol{p}_\infty$$

(ii)

$$\lim_{\nu\to\infty} \|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\|_\infty = 0.$$

(iii) *For any $i = 1, 2, \dots, n$, the i-th coordinate of $\boldsymbol{p}_\nu$ converges to the i-th coordinate of $\boldsymbol{p}_\infty$, i.e.,*

$$\lim_{\nu\to\infty} p^i_\nu = p^i_\infty, \;\; \forall i = 1, 2, \dots, n.$$

□

Example 11.4.4. The sequence of points

$$\boldsymbol{p}_\nu = \begin{bmatrix} \frac{1}{\nu} \\ \\ \frac{\nu+1}{\nu^2} \\ \\ \frac{\nu}{\nu+1} \end{bmatrix} \in \mathbb{R}^3, \;\; \nu \in \mathbb{N},$$

converges as $\nu \to \infty$ to the point

$$\boldsymbol{p}_\infty = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

since

$$\lim_{\nu\to\infty} \frac{1}{\nu} = \lim_{\nu\to\infty} \frac{\nu+1}{\nu^2} = 0, \;\; \lim_{\nu\to\infty} \frac{\nu}{\nu+1} = 1.$$

□

The following property of convergent sequences is an immediate generalization of its one-dimensional cousin Proposition 4.2.7.

Proposition 11.4.5. *If the sequence $(\boldsymbol{p}_\nu)$ of points in $\mathbb{R}^n$ converges to a point $\boldsymbol{p}$, then any subsequence of $(\boldsymbol{p}_\nu)$ converges to the same point $\boldsymbol{p}$.* □

Definition 11.4.6. A sequence $(\boldsymbol{p}_\nu)_{\nu\geqslant 1}$ in $\mathbb{R}^n$ is called *bounded* if there exists $R > 0$ such that

$$\|\boldsymbol{p}_\nu\| < R, \;\; \forall \nu \geqslant 1.$$

□

Proposition 11.4.7. *A convergent sequence of points on $\mathbb{R}^n$ is also bounded.*

Proof. Suppose that the sequence

$$\boldsymbol{p}_\nu = \begin{bmatrix} p^1_\nu \\ \vdots \\ p^n_\nu \end{bmatrix}, \;\; \nu = 1, 2, \dots$$

is convergent. According to Proposition 11.4.3, for each $i = 1, 2, \dots, n$ the sequence of coordinates (p^i_ν) is a convergent sequence of real numbers and thus, according to Proposition 4.2.12, it is bounded. Hence, there exists $C_i > 0$ such that

$$|p^i_\nu| < C_i, \;\; \forall \nu = 1, 2, \dots .$$

Set

$$C := \max(C_1, \dots, C_n).$$

Hence

$$\|\boldsymbol{p}_\nu\|_\infty = \max\big(\, |p^i_\nu|, \dots, |p^i_\nu| \, \big) < C, \;\; \forall \nu \geqslant 1.$$

Using (11.3.5) we deduce

$$\|\boldsymbol{p}_\nu\| \leqslant \sqrt{n}\|\boldsymbol{p}_\nu\|_\infty < C\sqrt{n}, \;\; \forall \nu \geqslant 1.$$

This proves that the sequence $(\boldsymbol{p}_\nu)$ is bounded. □

Proposition 11.4.8. *Let $n \in \mathbb{N}$. Suppose that $(\boldsymbol{p}_\nu)_{\nu \geqslant 1}$ and $(\boldsymbol{q}_\nu)_{\nu \geqslant 1}$ are convergent sequences of points in $\mathbb{R}^n$. Denote by $\boldsymbol{p}_\infty$ and respectively $\boldsymbol{q}_\infty$ their limits. Then the following hold.*

(i)
$$\lim_{\nu \to \infty} (\boldsymbol{p}_\nu + \boldsymbol{q}_\nu) = \boldsymbol{p}_\infty + \boldsymbol{q}_\infty.$$

(ii) *If $(t_\nu)_{\nu \geqslant 1}$ is a convergent sequence of real numbers with limit t_∞, then*
$$\lim_{\nu \to \infty} t_\nu \boldsymbol{p}_\nu = t_\infty \boldsymbol{p}_\infty.$$

(iii)
$$\lim_{\nu \to \infty} \langle \boldsymbol{p}_\nu, \boldsymbol{q}_\nu \rangle = \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\infty \rangle.$$

Proof. (i) We have

$$\begin{aligned} \operatorname{dist}\big(\boldsymbol{p}_\nu + \boldsymbol{q}_\nu, \boldsymbol{p}_\infty + \boldsymbol{q}_\infty \big) &= \| \, (\boldsymbol{p}_\nu + \boldsymbol{q}_\nu) - (\boldsymbol{p}_\infty + \boldsymbol{q}_\infty) \, \| = \|(\boldsymbol{p}_\nu - \boldsymbol{p}_\infty) + (\boldsymbol{q}_\nu - \boldsymbol{q}_\infty)\| \\ &\leqslant \|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\| + \|\boldsymbol{q}_\nu - \boldsymbol{q}_\infty\| = \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) + \operatorname{dist}(\boldsymbol{q}_\nu, \boldsymbol{q}_\infty) \to 0 \;\text{ as } \nu \to \infty. \end{aligned}$$

The claim now follows from the Squeezing Principle.

(ii) Since the sequences (t_ν) and $(\boldsymbol{p}_\nu)$ are convergent, they are also bounded and thus there exists $C > 0$ such that

$$|t_\nu|, \;\; \|\boldsymbol{p}_\nu\| < C, \;\; \forall \nu \geqslant 1.$$

We have

$$\begin{aligned} \operatorname{dist}(t_\nu \boldsymbol{p}_\nu, t_\infty \boldsymbol{p}_\infty) &= \|t_\nu \boldsymbol{p}_\nu - t_\infty \boldsymbol{p}_\infty\| = \|t_\nu \boldsymbol{p}_\nu - t_\infty \boldsymbol{p}_\nu + t_\infty \boldsymbol{p}_\nu - t_\infty \boldsymbol{p}_\infty\| \\ &\leqslant \|t_\nu \boldsymbol{p}_\nu - t_\infty \boldsymbol{p}_\nu\| + \|t_\infty \boldsymbol{p}_\nu - t_\infty \boldsymbol{p}_\infty\| = \|(t_\nu - t_\infty)\boldsymbol{p}_\nu\| + \|t_\infty(\boldsymbol{p}_\nu - \boldsymbol{p}_\infty)\| \\ &= |t_\nu - t_\infty| \cdot \|\boldsymbol{p}_\nu\| + |t_\infty| \cdot \|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\| \\ &\leqslant C|t_\nu - t_\infty| + |t_\infty| \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) \to 0 \;\text{ as } \nu \to \infty. \end{aligned}$$

(iii) Since the sequences $(\boldsymbol{p}_\nu)$ and $(\boldsymbol{q}_\nu)$ are convergent, they are also bounded and thus there exists $C > 0$ such that

$$\|\boldsymbol{p}_\nu\|, \;\; \|\boldsymbol{q}_\nu\| < C, \;\; \forall \nu \geqslant 1.$$

We have

$$\begin{aligned}
\big|\langle \boldsymbol{p}_\nu, \boldsymbol{q}_\nu\rangle - \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\infty\rangle\big| &= \big|\langle \boldsymbol{p}_\nu, \boldsymbol{q}_\nu\rangle - \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\nu\rangle + \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\nu\rangle - \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\infty\rangle\big| \\
&\leqslant \big|\langle \boldsymbol{p}_\nu, \boldsymbol{q}_\nu\rangle - \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\nu\rangle\big| + \big|\langle \boldsymbol{p}_\infty, \boldsymbol{q}_\nu\rangle - \langle \boldsymbol{p}_\infty, \boldsymbol{q}_\infty\rangle\big| \\
&= \big|\langle \boldsymbol{p}_\nu - \boldsymbol{p}_\infty, \boldsymbol{q}_\nu\rangle\big| + \big|\langle \boldsymbol{p}_\infty, \boldsymbol{q}_\nu - \boldsymbol{q}_\infty\rangle\big|
\end{aligned}$$

(use the Cauchy-Schwarz inequality)

$$\begin{aligned}
&\leqslant \|\boldsymbol{p}_\nu - \boldsymbol{p}_\infty\| \cdot \|\boldsymbol{q}_\nu\| + \|\boldsymbol{p}_\infty\| \cdot \|\boldsymbol{q}_\nu - \boldsymbol{q}_\infty\| \\
&\leqslant C \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) + \|\boldsymbol{p}_\infty\| \operatorname{dist}(\boldsymbol{q}_\nu, \boldsymbol{q}_\infty) \to 0 \;\text{ as } \nu \to \infty.
\end{aligned}$$

□

Definition 11.4.9. Let $n \in \mathbb{N}$. A sequence $(\boldsymbol{p}_\nu)_{\nu\geqslant 1}$ of points in $\mathbb{R}^n$ is called *Cauchy* or *fundamental* if $\forall \varepsilon > 0$, $\exists N = N(\varepsilon) > 0$ such that

$$\forall \nu, \mu > N(\varepsilon): \quad \operatorname{dist}(\boldsymbol{p}_\mu, \boldsymbol{p}_\nu) = \|\boldsymbol{p}_\mu - \boldsymbol{p}_\nu\| < \varepsilon.$$

□

Theorem 11.4.10 (Cauchy sequences)**.** *Let $n \in \mathbb{N}$ and consider a sequence $(\boldsymbol{p}_\nu)_{\nu\geqslant 1}$ of points in $\mathbb{R}^n$. The following statements are equivalent.*

(i) *The sequence $(\boldsymbol{p}_\nu)_{\nu\geqslant 1}$ is Cauchy.*

(ii) *The sequence $(\boldsymbol{p}_\nu)_{\nu\geqslant 1}$ converges to a point $\boldsymbol{p}_\infty \in \mathbb{R}^n$.*

Proof. (i) ⇒ (ii) Assume

$$\boldsymbol{p}_\nu = \begin{bmatrix} p_\nu^1 \\ \vdots \\ p_\nu^n \end{bmatrix}.$$

For each $i = 1, \dots, n$ and any $\mu, \nu \in \mathbb{N}$ we have

$$\big|\,p_\mu^i - p_\nu^i\,\big| = \sqrt{\big(p_\mu^i - p_\nu^i\big)^2} \leqslant \sqrt{\big(p_\mu^1 - p_\nu^1\big)^2 + \cdots + \big(p_\mu^n - p_\nu^n\big)^2} \leqslant \|\boldsymbol{p}_\mu - \boldsymbol{p}_\nu\|.$$

The above inequality shows that, for each $i = 1, \dots, n$, the sequence of *real numbers* $(p_\nu^i)_{\nu\geqslant 1}$ is Cauchy. Invoking Cauchy's Theorem 4.5.2 we deduce that, for each $i = 1, \dots, n$, the sequence $(p_\nu^i)_{\nu\geqslant 1}$ is convergent. Hence, for every $i = 1, \dots, n$, there exists $p_\infty^i \in \mathbb{R}$ such that

$$\lim_{\nu\to\infty} p_\nu^i = p_\infty^i.$$

From Proposition 11.4.3 we now deduce that

$$\lim_{\nu\to\infty} \begin{bmatrix} p_\nu^1 \\ \vdots \\ p_\nu^n \end{bmatrix} = \begin{bmatrix} p_\infty^1 \\ \vdots \\ p_\infty^n \end{bmatrix} =: \boldsymbol{p}_\infty.$$

(ii) ⇒ (i) Suppose that

$$\boldsymbol{p}_\infty = \lim_{\nu\to\infty} \boldsymbol{p}_\nu.$$

Then, $\forall \varepsilon > 0$, $\exists N = N(\varepsilon) > 0$ such that $\forall \nu > N(\varepsilon)$ we have

$$\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) < \frac{\varepsilon}{2}.$$

Then, for any $\mu, \nu > N(\varepsilon)$ we have

$$\operatorname{dist}(\boldsymbol{p}_\mu, \boldsymbol{p}_\nu) \leqslant \operatorname{dist}(\boldsymbol{p}_\mu, \boldsymbol{p}_\infty) + \operatorname{dist}(\boldsymbol{p}_\infty, \boldsymbol{p}_\nu) < \varepsilon.$$

□

Proposition 11.4.11. *Let $n \in \mathbb{N}$ and $C \subset \mathbb{R}^n$. Then the following statements are equivalent.*

(i) *The set $C \subset \mathbb{R}^n$ is closed in $\mathbb{R}^n$.*

(ii) *For any* convergent *sequence of points in C, its limit is also a point in C.*

Proof. (i) ⇒ (ii). We know that $\mathbb{R}^n \backslash C$ is open and we have to show that if $(\boldsymbol{p}_\nu)_{\nu \geqslant 1}$ is a convergent sequence of points in C, then its limit $\boldsymbol{p}_\infty$ belongs to C. We argue by contradiction. Suppose that $\boldsymbol{p}_\infty \in \mathbb{R}^n \backslash C$. Since $\mathbb{R}^n \backslash C$ is open, there exists $r > 0$ such that $B_r(\boldsymbol{p}_\infty) \subset \mathbb{R}^n \backslash C$, i.e.,

$$B_r(\boldsymbol{p}_\infty) \cap C = \varnothing.$$

This proves that, $\forall \nu \geqslant 1$, $\boldsymbol{p}_\nu \notin B_r(\boldsymbol{p}_\infty)$, i.e.,

$$\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) \geqslant r, \quad \forall \nu \geqslant 1.$$

This contradicts the fact that $\lim_{\nu\to\infty} \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_\infty) = 0$.

(ii) ⇒ (i) We have to show that $\mathbb{R}^n \backslash C$ is open. We argue by contradiction. Assume that there exists $\boldsymbol{p}_* \in \mathbb{R}^n \backslash C$ such that, $\forall r > 0$, the ball $B_r(\boldsymbol{p}_*)$ is not contained in $\mathbb{R}^n \backslash C$. Thus, for any $r > 0$ there exists $\boldsymbol{p}(r) \in B_r(\boldsymbol{p}_*) \cap C$, i.e., $\boldsymbol{p}(r) \in C$, $\operatorname{dist}(\boldsymbol{p}(r), \boldsymbol{p}_*) < r$. Thus, for any $\nu \in \mathbb{N}$, there exists $\boldsymbol{p}_\nu \in C$ such that

$$\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}_*) < \frac{1}{\nu}, \quad \forall \nu \in \mathbb{N}.$$

This shows that the sequence of points $(\boldsymbol{p}_\nu)$ in C converges to the point $\boldsymbol{p}_*$ *that is not in* C. This contradicts (ii).

□

Example 11.4.12. Any affine line in $\mathbb{R}^n$ is a closed subset. We will prove this in two different ways. Consider the line $\ell_{\boldsymbol{p},\boldsymbol{v}} \subset \mathbb{R}^n$ passing through the point $\boldsymbol{p}$ in the direction $\boldsymbol{v} \neq \boldsymbol{0}$.

1st Method. Suppose that $(\boldsymbol{q}_\nu)$ is a convergent sequence of points on this line. We denote by $\boldsymbol{q}_\infty$ its limit. We want to prove that $\boldsymbol{q}_\infty$ also lies on the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$.

To see this note first that since $\boldsymbol{q}_\nu \in \ell_{\boldsymbol{p},\boldsymbol{v}}$, there exists $t_\nu \in \mathbb{R}$ such that

$$\boldsymbol{q}_\nu = \boldsymbol{p} + t_\nu \boldsymbol{v}.$$

We deduce that for any $\mu, \nu \geqslant 1$ we have

$$\operatorname{dist}(\boldsymbol{q}_\mu, \boldsymbol{q}_\nu) = \|\boldsymbol{q}_\mu - \boldsymbol{q}_\nu\| = |t_\mu - t_\nu| \cdot \|\boldsymbol{v}\|. \Rightarrow |t_\mu - t_\nu| = \frac{1}{\|\boldsymbol{v}\|} \operatorname{dist}(\boldsymbol{q}_\mu, \boldsymbol{q}_\nu).$$

Since the sequence $(\boldsymbol{q}_\nu)$ is convergent, it is also Cauchy, and the above equality shows that the sequence (t_ν) is Cauchy as well. Hence the sequence (t_ν) is convergent in $\mathbb{R}$. If t_∞ is its limit, then Proposition 11.4.8 implies that

$$\boldsymbol{q}_\infty = \lim_{\nu \to \infty} (\boldsymbol{p} + t_\nu \boldsymbol{v}) = \boldsymbol{p} + t_\infty \boldsymbol{v} \in \ell_{\boldsymbol{p},\boldsymbol{v}}.$$

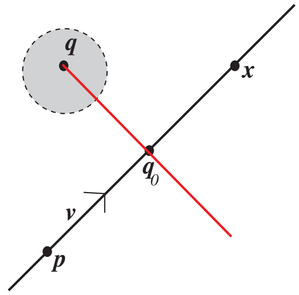

Figure 11.10. $\operatorname{dist}(\boldsymbol{q}, \boldsymbol{q}_0) \leqslant \operatorname{dist}(\boldsymbol{q}, \boldsymbol{x})$, $\forall \boldsymbol{x} \in \ell_{\boldsymbol{p},\boldsymbol{v}}$.

2nd Method. We will prove that the complement of the line is open, i.e., if $\boldsymbol{q}$ is a point outside the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$, then there exists an open ball centered at $\boldsymbol{q}$ that does not intersect the line; see Figure 11.10.

To do so, we will find the point $\boldsymbol{q}_0$ on the line closest to $\boldsymbol{q}$. Usual Euclidean geometry suggests that if $\boldsymbol{q}_0$ is such a point, then the line $\boldsymbol{q}\boldsymbol{q}_0$ should be perpendicular to $\ell_{\boldsymbol{p},\boldsymbol{v}}$; see Figure 11.10. So, instead of looking for a point on the line closest to $\boldsymbol{q}$, we will look for a point $\boldsymbol{q}_0$ such that $(\boldsymbol{q} - \boldsymbol{q}_0) \perp \boldsymbol{v}$. As we will see, such a $\boldsymbol{q}_0$ will indeed be the point on the line closest to $\boldsymbol{q}$. Observe that

$$(\boldsymbol{q} - \boldsymbol{q}_0) \perp \boldsymbol{v} \Longleftrightarrow \langle \boldsymbol{q} - \boldsymbol{q}_0, \boldsymbol{v} \rangle \Longleftrightarrow \langle \boldsymbol{q}, \boldsymbol{v} \rangle = \langle \boldsymbol{q}_0, \boldsymbol{v} \rangle.$$

Since $\boldsymbol{q}_0$ is on the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$ it has the form $\boldsymbol{q} = \boldsymbol{p} + t_0 \boldsymbol{v}$ for some real number t_0. Using this in the above equality we deduce

$$\langle \boldsymbol{q}, \boldsymbol{v} \rangle = \langle \boldsymbol{p} + t_0 \boldsymbol{v}, \boldsymbol{v} \rangle = \langle \boldsymbol{p}, \boldsymbol{v} \rangle + t_0 \langle \boldsymbol{v}, \boldsymbol{v} \rangle = \langle \boldsymbol{p}, \boldsymbol{v} \rangle + t_0 \|\boldsymbol{v}\|^2$$
$$\Rightarrow t_0 \|\boldsymbol{v}\|^2 = \langle \boldsymbol{q} - \boldsymbol{p}, \boldsymbol{v} \rangle \Rightarrow t_0 = \frac{\langle \boldsymbol{q} - \boldsymbol{p}, \boldsymbol{v} \rangle}{\|\boldsymbol{v}\|^2}.$$

Note that if $\boldsymbol{x} \in \ell_{\boldsymbol{p},\boldsymbol{q}}$, then $\boldsymbol{x} - \boldsymbol{q}_0$ is a multiple of $\boldsymbol{v}$ so $(\boldsymbol{x} - \boldsymbol{q}_0) \perp (\boldsymbol{q} - \boldsymbol{q}_0)$; see Exercise 11.2(b). Pythagoras' theorem then implies that (Figure 11.10)

$$\operatorname{dist}(\boldsymbol{q}, \boldsymbol{x})^2 = \operatorname{dist}(\boldsymbol{q}, \boldsymbol{q}_0)^2 + \operatorname{dist}(\boldsymbol{q}_0, \boldsymbol{x})^2 \geqslant \operatorname{dist}(\boldsymbol{q}, \boldsymbol{q}_0)^2.$$

Hence, if we set $r := \operatorname{dist}(\boldsymbol{q}, \boldsymbol{q}_0)$, then we deduce that $r > 0$ and $r \geqslant \operatorname{dist}(\boldsymbol{q}, \boldsymbol{x})$, $\forall \boldsymbol{x} \in \ell_{\boldsymbol{p},\boldsymbol{v}}$. In particular this shows that the ball $B_{r/2}(\boldsymbol{q})$ of radius $r/2$ and centered at $\boldsymbol{q}$ does not intersect the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$.

□

Definition 11.4.13. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$.

(i) A point $\boldsymbol{p} \in \mathbb{R}^n$ is a *cluster point* of X if, for any $\varepsilon > 0$, the ball $B_\varepsilon(\boldsymbol{p})$ contains a point in X $\underline{\text{not equal}}$ to $\boldsymbol{p}$.

(ii) A subset $S \subset X$ is called *dense* in X if, for any $\boldsymbol{x} \in X$ and any $\varepsilon > 0$, the ball $B_\varepsilon(\boldsymbol{x})$ contains a point in S.

□

Example 11.4.14. Proposition 3.4.4 shows that the set $\mathbb{Q}$ of rational numbers is dense in $\mathbb{R}$. More generally, the set $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$. □

Proposition 11.4.15. *Let $n \in \mathbb{N}$, $X \subset \mathbb{R}^n$ and $\boldsymbol{p} \in \mathbb{R}^n$. The following statements are equivalent.*

(i) *The point $\boldsymbol{p}$ is a cluster point of X.*

(ii) *There exists a sequence of points $(\boldsymbol{p}_\nu)$ in $X\backslash\{\boldsymbol{p}\}$ that converges to $\boldsymbol{p}$.*

Proof. (i) ⇒ (ii) Since $\boldsymbol{p}$ is a cluster point of X we deduce that, for any $\nu \in \mathbb{N}$, the ball $B_{1/\nu}(\boldsymbol{p})$ contains a point $\boldsymbol{p}_\nu \in X\backslash\{\boldsymbol{p}\}$. Observing that $\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}) < \frac{1}{\nu}$ we deduce that

$$\lim_{\nu\to\infty} \operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}) = 0,$$

i.e., $(\boldsymbol{p}_\nu)$ is a sequence in $X\backslash\{\boldsymbol{p}\}$ that converges to $\boldsymbol{p}$.

(ii) ⇒ (i) We know that there exists a sequence $(\boldsymbol{p}_\nu)$ in $X\backslash\{\boldsymbol{p}\}$ that converges to $\boldsymbol{p}$. Let $\varepsilon > 0$. There exists $N = N(\varepsilon) > 0$ such that $\operatorname{dist}(\boldsymbol{p}_\nu, \boldsymbol{p}) < \varepsilon$, $\forall \nu > N(\varepsilon)$. Thus the ball $B_\varepsilon(\boldsymbol{p})$ contains all the points $\boldsymbol{p}_\nu$, $\nu > N(\varepsilon)$ and none of these points is equal to $\boldsymbol{p}$.

□

11.5. Exercises

Exercise 11.1. Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$. Show that the following statements are equivalent.

(i) The vectors $\boldsymbol{u}, \boldsymbol{v}$ are collinear.

(ii) For any $\boldsymbol{p} \in \mathbb{R}^n$ the lines $\ell_{\boldsymbol{p},\boldsymbol{u}}$, $\ell_{\boldsymbol{p},\boldsymbol{v}}$ coincide, i.e., $\ell_{\boldsymbol{p},\boldsymbol{u}} = \ell_{\boldsymbol{p},\boldsymbol{v}}$, $\forall \boldsymbol{p} \in \mathbb{R}^n$.

(iii) The lines $\ell_{\boldsymbol{0},\boldsymbol{u}}$, $\ell_{\boldsymbol{0},\boldsymbol{v}}$ coincide, i.e., $\ell_{\boldsymbol{0},\boldsymbol{u}} = \ell_{\boldsymbol{0},\boldsymbol{v}}$.

□

Exercise 11.2. (a) Let $\boldsymbol{p}, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v} \neq \boldsymbol{0}$. Prove that if $\boldsymbol{q} \in \ell_{\boldsymbol{p},\boldsymbol{v}}$, then $\ell_{\boldsymbol{p},\boldsymbol{v}} = \ell_{\boldsymbol{q},\boldsymbol{v}}$.

(b) Let $\boldsymbol{p}, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v} \neq \boldsymbol{0}$. Prove that if $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \ell_{\boldsymbol{p},\boldsymbol{v}}$ and $\boldsymbol{p}_1 \neq \boldsymbol{p}_2$, then the vectors $\boldsymbol{v}$ and $\boldsymbol{u} := \boldsymbol{p}_2 - \boldsymbol{p}_1$ are collinear and $\ell_{\boldsymbol{p},\boldsymbol{v}} = \ell_{\boldsymbol{p},\boldsymbol{u}} = \ell_{\boldsymbol{p}_1,\boldsymbol{u}} = \ell_{\boldsymbol{p}_2,\boldsymbol{u}}$.

(c) Let $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{u}, \boldsymbol{v} \neq \boldsymbol{0}$. Show that if the lines $\ell_{\boldsymbol{p},\boldsymbol{u}}$ and $\ell_{\boldsymbol{q},\boldsymbol{v}}$ have two distinct points in common, then they coincide. □

Exercise 11.3. Consider the points in $\mathbb{R}^2$

$$\boldsymbol{p}_0 = (0,0),\ \ \boldsymbol{q}_0 = (1,1),\ \ \boldsymbol{p}_1 = (1,0),\ \ \boldsymbol{q}_1 = (0,1).$$

(a) Depict these points and the lines $\ell_0 = \boldsymbol{p}_0\boldsymbol{q}_0$, $\ell_1 = \boldsymbol{p}_1\boldsymbol{q}_1$ on the same planar coordinate system of the type depicted in Figure 11.2.

(b) Find the coordinates of the point where the lines ℓ_0, ℓ_1 intersect. □

Exercise 11.4. Prove Proposition 11.1.11. □

Exercise 11.5. Let $n \in \mathbb{N}$ and $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$. Prove that the following statements are equivalent.

(i) $\boldsymbol{p} \neq \boldsymbol{q}$.

(ii) There exists a linear form $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ such that $\boldsymbol{\xi}(\boldsymbol{p}) \neq \boldsymbol{\xi}(\boldsymbol{q})$.

□

Exercise 11.6. Find a parametric equation (see (11.1.6)) for the line in $\mathbb{R}^2$ described by the equation

$$x^1 + 2x^2 = 3.$$

Hint: Use the equality $x^1 = 3 - 2x^2$ to find two distinct points on this line. □

Exercise 11.7. Let $\boldsymbol{p} = (1,2,3) \in \mathbb{R}^3$ and $\boldsymbol{v} = (1,1,1) \in \mathbb{R}^3$. Find the coordinates of the point of intersection of the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$ with the hyperplane

$$3x^1 + 4x^2 + 5x^3 = 6.$$ □

Exercise 11.8. Prove that the lines and the hyperplanes in $\mathbb{R}^n$ are affine subspaces. □

Exercise 11.9. Let S be a subset of the Euclidean space $\mathbb{R}^n$, $n \in \mathbb{N}$. Prove that the following statements are equivalent.

(i) The set S is an affine subspace.

(ii) For any $k \in \mathbb{N}$, any points $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_k \in S$ and any real numbers $t_0, t_1, \dots, t_k$ such that $t_0 + t_1 + \cdots + t_k = 1$ we have

$$t_0\boldsymbol{p}_0 + t_1\boldsymbol{p}_1 + \cdots + t_k\boldsymbol{p}_k \in S.$$

Hint: The implication (ii) ⇒ (i) is immediate. To prove the opposite implication (i) ⇒ (ii) argue by induction on k. Observe that least one of the numbers $t_0, t_1, \dots, t_k$ is not equal to 1, say $t_k \neq 1$. Then $1 - t_k \neq 0$ and

$$t_0\boldsymbol{p}_0 + t_1\boldsymbol{p}_1 + \cdots + t_k\boldsymbol{p}_k = (1 - t_k)\underbrace{\left(\frac{t_0}{1-t_k}\boldsymbol{p}_1 + \cdots + \frac{t_{k-1}}{1-t_k}\boldsymbol{p}_{k-1}\right)}_{\boldsymbol{q}} + t_k\boldsymbol{p}_k.$$

Use the induction assumption to argue that $\boldsymbol{q} \in S$. Conclude using (i). □

Exercise 11.10. Prove Proposition 11.1.28. □

Exercise 11.11. Consider the linear operator $A : \mathbb{R}^3 \to \mathbb{R}^3$ characterized by the equalities

$$A\boldsymbol{e}_1 = \boldsymbol{e}_1 + 2\boldsymbol{e}_2 + 3\boldsymbol{e}_3, \quad A\boldsymbol{e}_2 = 4\boldsymbol{e}_1 + 5\boldsymbol{e}_2 + 5\boldsymbol{e}_3, \quad A\boldsymbol{e}_3 = 7\boldsymbol{e}_1 + 8\boldsymbol{e}_2 + 9\boldsymbol{e}_3,$$

where $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ is the canonical basis of $\mathbb{R}^3$.

(i) Find the 3×3 matrix associated to this linear operator.

(ii) Find the vector

$$A\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

□

Exercise 11.12. Consider the linear operator $A : \mathbb{R}^3 \to \mathbb{R}^2$ given by the matrix

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

Show that there exists a nonzero vector $\boldsymbol{v} \in \mathbb{R}^3$ such that $\ker A$ is equal to the line $\ell_{\boldsymbol{0},\boldsymbol{v}}$. □

Exercise 11.13. Suppose that $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator. Prove that the following statements are equivalent.

(i) A is injective.

(ii) $\ker A = \{\boldsymbol{0}\}$.

□

Exercise 11.14. (a) An *automorphism* of $\mathbb{R}^k$ is a *bijective* linear operator $T : \mathbb{R}^k \to \mathbb{R}^k$. Prove that if T is an automorphism of $\mathbb{R}^k$ then its inverse is also an automorphism of $\mathbb{R}^k$.

(b) A $k \times k$ matrix A is called *invertible* if and only if there exists a $k \times k$ matrix A' such that $AA' = A'A = \mathbb{1}_k$. Prove that if A is invertible, then there exists *a unique* matrix A' with these properties. This unique matrix is called the *inverse* of A and it is denoted by A^{-1}.

(c) Show that T is an automorphism of $\mathbb{R}^k$ if and only if the $k \times k$ matrix representing T is invertible. □

Exercise 11.15. Let $m, n \in \mathbb{N}$, $B \in \mathrm{Mat}_m(\mathbb{R})$, $C \in \mathrm{Mat}_n(\mathbb{R})$, $D \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ and $E \in \mathrm{Mat}_{n\times m}(\mathbb{R})$. Consider the square matrices $S, T \in \mathrm{Mat}_{m+n}(\mathbb{R})$ with block decompositions

$$S = \left[\begin{array}{cc} B & D \\ \mathbf{0}_{n\times m} & C \end{array}\right], \quad T = \left[\begin{array}{cc} B & \mathbf{0}_{m\times n} \\ E & C \end{array}\right],$$

and $\mathbf{0}_{k\times \ell}$ denotes the $k \times \ell$ matrix with all entries 0.

Show that if B, C are invertible, then so are S and T and, moreover,

$$S^{-1} = \left[\begin{array}{cc} B^{-1} & -B^{-1}DC^{-1} \\ \mathbf{0}_{n\times m} & C^{-1} \end{array}\right], \quad T^{-1} = \left[\begin{array}{cc} B^{-1} & \mathbf{0}_{m\times n} \\ -C^{-1}EB^{-1} & C^{-1} \end{array}\right].$$ □

Exercise 11.16. We say that a matrix $R \in \mathrm{Mat}_{k\times k}(\mathbb{R})$ is *nilpotent* if there exists $n \in \mathbb{N}$ such that $R^n = \mathbf{0}$. Show that if R is a $k \times k$ nilpotent matrix, then the matrix $\mathbb{1}_k - R$ is invertible.

Hint: Prove first that if $X \in \mathrm{Mat}_{k\times k}(\mathbb{R})$, then

$$\mathbb{1}_k - X^n = (\mathbb{1}_k - X)(\mathbb{1}_k + X + \cdots + X^{n-1}), \quad \forall n \in \mathbb{N}.$$ □

Exercise 11.17. Show that the space $\mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m)$ of linear operators $\mathbb{R}^n \to \mathbb{R}^m$ is a real vector space. □

Exercise 11.18. Consider the matrices

$$A = \left[\begin{array}{rrr} 1 & -2 & 3 \\ 0 & 1 & -4 \end{array}\right], \quad B = \left[\begin{array}{rr} 1 & 0 \\ -2 & 1 \\ 3 & -4 \end{array}\right].$$

(i) Compute the products AB and BA.

(ii) Show that for any vectors $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{y} \in \mathbb{R}^3$ we have

$$\langle \boldsymbol{x}, A\boldsymbol{y} \rangle = \langle B\boldsymbol{x}, \boldsymbol{y} \rangle.$$

□

Exercise 11.19. Let $m \in \mathbb{N}$, $m \geqslant 2$ and consider the $m \times m$ matrix

$$N = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Compute the powers N^k, $k \in \mathbb{N}$.

Hint: Regard N as a linear operator $\mathbb{R}^m \to \mathbb{R}^m$ and observe that

$$N\boldsymbol{e}_1 = 0, \;\; N\boldsymbol{e}_2 = \boldsymbol{e}_1, \;\; N\boldsymbol{e}_3 = \boldsymbol{e}_2, \; \dots, \; N\boldsymbol{e}_m = \boldsymbol{e}_{m-1},$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_m$ is the natural basis of $\mathbb{R}^m$. Then use the fact that the composition of two linear operators corresponds to the multiplication of the corresponding matrices. □

Exercise 11.20. For every $\alpha \in [0, 2\pi]$ we denote by $R_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ the *counterclockwise rotation* of angle α about the origin $\mathbf{0}$.

(i) Express the coordinates y^1, y^2 of $\boldsymbol{y} = R_\alpha \boldsymbol{x}$ in terms of the coordinates of $\boldsymbol{x} = [x^1, x^2]^\top$.

(ii) Show that R_α is a linear operator and compute its associated matrix. Continue to denote by R_α the associated matrix.

(iii) Given $\alpha, \beta \in [0, 2\pi]$ compute the product $R_\alpha \cdot R_\beta$.

Hint: (i) Set $r := \|\boldsymbol{x}\|$, and denote by θ the angle the vector $\boldsymbol{x}$ makes with the x^1-axis, measured conterclockwisely starting at the positive x^1-axis. Then $x^1 = r\cos\theta$, $x^2 = r\sin\theta$. Next, set $\boldsymbol{y} := R_\alpha \boldsymbol{x}$ and show that, $y^1 = r\cos(\theta+\alpha)$, $y^2 = r\sin(\theta+\alpha)$. Conclude using the trig formulæ (5.7.1a). □

Exercise 11.21. The *trace* of an $n \times n$ matrix A is the scalar denoted by $\operatorname{tr} A$ and defined as the sum of the diagonal entries of A,

$$\operatorname{tr} A := A^1_1 + \cdots + A^n_n.$$

(i) Show that if $A, B \in \operatorname{Mat}_{n\times n}(\mathbb{R})$, $c \in \mathbb{R}$, then

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

(ii) Show that if $A \in \operatorname{Mat}_{m\times n}(\mathbb{R})$ and $B \in \operatorname{Mat}_{n\times m}(\mathbb{R})$, then

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

Hint: Use (11.1.15).

(iii) Show that there do not exist matrices $A, B \in \operatorname{Mat}_{n\times n}(\mathbb{R})$ such that $AB - BA = \mathbb{1}_n$.

□

Exercise 11.22. Prove (11.2.5). □

Exercise 11.23. Suppose that $A \in \mathrm{Mat}_{m\times n}(\mathbb{R})$. Denote by $(\boldsymbol{e}_j)_{1\leqslant j\leqslant n}$ the canonical basis of $\mathbb{R}^n$ and by $(\boldsymbol{f}_i)_{1\leqslant i\leqslant m}$ the canonical basis of $\mathbb{R}^m$. Prove that

$$A^i_j = \langle \boldsymbol{f}_i, A\boldsymbol{e}_j\rangle, \;\; \forall i = 1,\dots,m, \;\; j = 1,\dots,n. \qquad \square$$

Exercise 11.24. Suppose that $A \in \mathrm{Mat}_{m\times n}(\mathbb{R})$. The *transpose* of A is the $n\times m$ matrix $A^\top$ defined by the requirement

$$(A^\top)^j_i = A^i_j, \;\; \forall i = 1,\dots,m, \;\; j = 1,\dots,n.$$

In other words, the rows of $A^\top$ coincide with the columns of A. (For example, the transpose of the matrix A in Exercise 11.18 is the matrix B in the same exercise.)

(i) Suppose that $B \in \mathrm{Mat}_{p\times m}(\mathbb{R})$. Prove that

$$(B\cdot A)^\top = (A^\top)\cdot(B^\top).$$

Hint: Check this first in the special case when $p = m = 1$, i.e., B is a matrix consisting one row of size m, and A is a matrix consisting of one column of size m. Use this special case and the equality (11.1.15) to deduce the general case.

(ii) Prove that, for any $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, we have (identifying 1×1 matrices with numbers)

$$\langle \boldsymbol{x}, A\boldsymbol{y}\rangle = \boldsymbol{x}^\top \cdot A \cdot \boldsymbol{y} = \langle A^\top \boldsymbol{x}, \boldsymbol{y}\rangle.$$

(iii) Prove that for any $y \in \mathbb{R}^n$ we have

$$\langle A^\top A\boldsymbol{y}, \boldsymbol{y}\rangle \geqslant 0.$$

(iv) Prove that an $n\times n$ matrix A is symmetric if and only if

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, A\boldsymbol{y}\rangle, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Hint: Use Exercise 11.23 and part (i) of this exercise. □

Exercise 11.25. Let $n \in \mathbb{N}$ and suppose that $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator. As usual we will continue to denote by A the associated matrix. Prove that the following statements are equivalent.

(i) $\langle A\boldsymbol{x}, A\boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

(ii) $\|A\boldsymbol{x}\| = \|\boldsymbol{x}\|$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.

(iii) $A^\top \cdot A = \mathbb{1}_n$.

An operator or matrix with any of the above three equivalent properties is called *orthogonal.* □

Exercise 11.26. Suppose that $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ is a linear functional. Show that the graph of $\boldsymbol{\xi}$, defined as

$$G_{\boldsymbol{\xi}} = \left\{ (\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R}; \;\; y = \boldsymbol{\xi}(\boldsymbol{x}) \right\},$$

is a hyperplane in $\mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$ and then find a normal vector to this hyperplane. □

Exercise 11.27. Prove Proposition 11.3.11. □

Exercise 11.28. Prove (11.3.6). □

Exercise 11.29. Prove Proposition 11.3.13.

Hint: Use (11.3.6). □

Exercise 11.30. Prove Proposition 11.3.14. □

Exercise 11.31. Prove that if $U \subset \mathbb{R}^m$ is open in $\mathbb{R}^m$ and $V \subset \mathbb{R}^n$ is open in $\mathbb{R}^n$, then $U \times V$ is open in $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$.

Hint: Use Proposition 11.3.14 and observe several things. First, if $\boldsymbol{p} \in \mathbb{R}^m$ and $\boldsymbol{q} \in \mathbb{R}^n$ then the pair $(\boldsymbol{p}, \boldsymbol{q}) \in \mathbb{R}^m \times \mathbb{R}^n$ and the Cartesian product can be identified with $\mathbb{R}^{m+n}$. Next observe that the Cartesian product $C_r(\boldsymbol{p}) \times C_r(\boldsymbol{q}) \subset \mathbb{R}^m \times \mathbb{R}^n$ can be identified with $C_r\big((\boldsymbol{p}, \boldsymbol{q})\big)$, the cube of radius r with center $(\boldsymbol{p}, \boldsymbol{q}) \in \mathbb{R}^{m+n}$. □

Exercise 11.32. Complete the proof of the claim in Example 11.3.16(c). □

Exercise 11.33. Prove Proposition 11.4.3.

Hint: Use (11.3.5). □

Exercise 11.34. (a) Prove that any finite subset of $\mathbb{R}^n$ is closed.

(b) Prove that any affine hyperplane in $\mathbb{R}^n$ is a closed subset. □

Exercise 11.35. Prove that any open subset $U \subset \mathbb{R}^n$ is the union of a (possibly infinite) family of open cubes. □

Exercise 11.36. Let $n \in \mathbb{N}$. Prove that for any $\boldsymbol{p} \in \mathbb{R}^n$ and any $r > 0$ the open Euclidean ball $B_r(\boldsymbol{p})$ and the closed Euclidean ball $\overline{B_r(\boldsymbol{p})}$ are convex sets. □

Exercise 11.37. Let $n \in \mathbb{N}$.

(a) Suppose that $(\boldsymbol{p}_\nu)$ is a sequence in $\mathbb{R}^n$ that converges to $\boldsymbol{p} \in \mathbb{R}^n$. Prove that

$$\lim_{\nu\to\infty} \|\boldsymbol{p}_\nu\| = \|\boldsymbol{p}\| \text{ and } \lim_{\nu\to\infty} \|\boldsymbol{p}_\nu\|_\infty = \|\boldsymbol{p}\|_\infty.$$

(b) Let $r > 0$. Prove that any point $\boldsymbol{x} \in \mathbb{R}^n$ such that $\|\boldsymbol{x}\| = r$ is a cluster point of the open ball $B_r(\boldsymbol{0})$.

Hint: (a) Use (11.3.2b) and (11.3.4b). □

Exercise 11.38. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. Prove that the following statements are equivalent.

(i) The set X is closed.

(ii) The set X contains all its cluster points.

□

Exercise 11.39. Prove that the set $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$.

Hint: Use Proposition 3.4.4 and Proposition 11.4.3 . □

Exercise 11.40. Let $n \in \mathbb{N}$. Consider a sequence of vectors $(\boldsymbol{x}_\nu)_{\nu\in\mathbb{N}}$ in $\mathbb{R}^n$. The *series*

$$\sum_{\nu\in\mathbb{N}} \boldsymbol{x}_\nu$$

associated to this sequence is the new sequence $(S_N)_{N\in\mathbb{N}}$ of vectors in $\mathbb{R}^n$ described by the *partial sums*

$$S_N = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_N,\quad N \in \mathbb{N}.$$

The series $\sum_{\nu\in\mathbb{N}} \boldsymbol{x}_\nu$ is called *convergent* if the sequence of partial sums S_N is convergent.

Prove that if the *series* of real numbers $\sum_{\nu\in\mathbb{N}} \|\boldsymbol{x}_\nu\|$ is convergent, then the *series* of vectors $\sum_{\nu\in\mathbb{N}} \boldsymbol{x}_\nu$ is also convergent.

Hint. It suffices to show that the sequence (S_N) is Cauchy. Define

$$S_N^* := \|\boldsymbol{x}_1\| + \cdots + \|\boldsymbol{x}_N\|,\quad N \in \mathbb{N}.$$

The *series* of real numbers $\sum_{\nu\in\mathbb{N}} \|\boldsymbol{x}_\nu\|$ is convergent and thus sequence (S_N^*) is convergent, hence Cauchy. Prove that this implies that the sequence S_N is Cauchy by imitating the proof of Absolute Convergence Theorem 4.6.13.□

Exercise 11.41 (Banach's fixed point theorem)**.** Suppose that $X \subset \mathbb{R}^n$ is a *closed* subset and $F : X \to \mathbb{R}^n$ is a map satisfying the following conditions:

$$F(\boldsymbol{x}) \in X,\quad \forall \boldsymbol{x} \in X. \tag{C_1}$$

$$\exists r \in (0,1) \text{ such that } \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in X :\ \ \|F(\boldsymbol{x}_1) - F(\boldsymbol{x}_2)\| \leqslant r\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{C_2}$$

Fix $\boldsymbol{x}_0 \in X$ and define inductively the sequence of points in X,

$$\boldsymbol{x}_1 = F(\boldsymbol{x}_0),\ \ \boldsymbol{x}_2 = F(\boldsymbol{x}_1), \dots, \boldsymbol{x}_\nu = F(\boldsymbol{x}_{\nu-1}),\ \ \forall \nu \in \mathbb{N}.$$

Prove that the following hold.

(i) For any $\nu \in \mathbb{N}$,

$$\|\boldsymbol{x}_{\nu+1} - \boldsymbol{x}_\nu\| \leqslant r^\nu \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|.$$

(ii) For any $\mu, \nu \in \mathbb{N}$, $\mu < \nu$

$$\|\boldsymbol{x}_\nu - \boldsymbol{x}_\mu\| \leqslant \frac{r^\mu(1 - r^{\nu-\mu})}{1-r}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\| \leqslant \frac{r^\mu}{1-r}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|.$$

(iii) The sequence $(\boldsymbol{x}_\nu)_{\nu\geqslant 0}$ is Cauchy.

(iv) If $\boldsymbol{x}_*$ is the limit of the sequence $(\boldsymbol{x}_\nu)_{\nu\geqslant 0}$, then $F(\boldsymbol{x}_*) = \boldsymbol{x}_*$.

(v) Show that if $\boldsymbol{p} \in X$ is a *fixed point of* F, i.e., it satisfies $F(\boldsymbol{p}) = \boldsymbol{p}$, then $\boldsymbol{p}$ must be equal to the point $\boldsymbol{x}_*$ defined above.

□

Exercise 11.42. Suppose that $S \subset \mathbb{R}^{17}$ consists of $1,234,567,890$ points and $T : S \to S$ is a map such that

$$\|T\boldsymbol{s}_1 - T\boldsymbol{s}_2\| < \|s_1 - s_2\|, \ \ \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in S, \ \ \boldsymbol{s}_1 \neq \boldsymbol{s}_2.$$

Prove that there exists $\boldsymbol{s}_* \in S$ such that $T(\boldsymbol{s}_*) = \boldsymbol{s}_*$.

Hint: Use the result in the previous exercise. □

11.6. Exercises for extra credit

Exercise* 11.1. (a) Prove that if S is an affine subspace of $\mathbb{R}^2$, then S is either a point, or a line, or the whole $\mathbb{R}^2$.

(b) Prove that if S is an affine subspace of $\mathbb{R}^3$, then S is either a point, or a line, or a plane, or the whole $\mathbb{R}^3$. □

Exercise* 11.2. Suppose that $n \in \mathbb{N}$ and $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$. Prove that the following statements are equivalent.

(i) The matrix A is invertible in the sense defined in Exercise 11.14.

(ii) There exists $B \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ such that $BA = \mathbb{1}_n$.

(iii) There exists $C \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ such that $AC = \mathbb{1}_n$.

(iv) The linear operator $\mathbb{R}^n \to \mathbb{R}^n$ defined by A is bijective.

(v) The linear operator $\mathbb{R}^n \to \mathbb{R}^n$ defined by A is injective.

(vi) The linear operator $\mathbb{R}^n \to \mathbb{R}^n$ defined by A is surjective.

Hint: You need to use the fact that $\mathbb{R}^n$ is a *finite dimensional* vector space. □

Chapter 12

Continuity

A function $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ can be viewed as transporting in some fashion the Euclidean space $\mathbb{R}^n$ into the Euclidean space $\mathbb{R}^m$. The space $\mathbb{R}^m$ is often called the *target space*. For example, a map $\boldsymbol{F} : \mathbb{R} \to \mathbb{R}^2$ "transports" the real axis $\mathbb{R}$ into a region of $\mathbb{R}^2$ that typically looks like a curve; see Figure 12.1. For this reason functions $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ are often called *transformations*, *operators*, or *maps*.

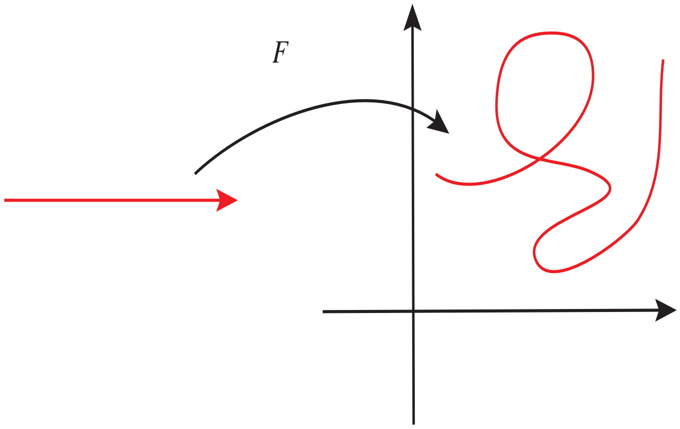

Figure 12.1. *A map* $F : \mathbb{R} \to \mathbb{R}^2$.

Suppose that $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ is a map. For any $\boldsymbol{x} \in \mathbb{R}^n$, its image $\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x})$ is a point in $\mathbb{R}^m$ and thus it is determined by a column vector

$$\boldsymbol{y} = \begin{bmatrix} y^1 \\ \vdots \\ y^m \end{bmatrix}.$$

The coordinates $y^1, \dots, y^m$ depend on the point $\boldsymbol{x}$ and thus they are described by functions

$$F^i : \mathbb{R}^n \to \mathbb{R},\ \ y^i = F^i(x^1, \dots, x^n),\ \ i = 1, \dots, m.$$

We can turn this argument on its head, and think of a collection of functions

$$F^1, \dots, F^m : \mathbb{R}^n \to \mathbb{R}$$

as defining a map $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$. Often, when working with a map $\mathbb{R}^n \to \mathbb{R}^m$ and no confusion is possible, we will dispense of the extra symbol $\boldsymbol{F}$ and describe the map in a simpler way as a collection of functions

$$y^1 = y^1(x^1, \dots, x^n), \dots, y^m = y^m(x^1, \dots, x^n).$$

Example 12.0.1. When predicting the weather (on the surface of the Earth) we need to describe several quantities: temperature (T), pressure (P) and wind velocity $V = (V^1, V^2)$. These quantities depend on the location (determined by two coordinates x^1, x^2), and the time t. We thus have a collection of 4 functions P, T, V^1, V^2 depending on 3 variables x^1, x^2, t,

$$P = P(x^1, x^2, t),\ \ V^1 = V^1(x^1, x^2, t) \text{ etc},$$

and thus we are dealing with a map $\mathbb{R}^3 \to \mathbb{R}^4$. □

Definition 12.0.2. Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. The *graph* of a map $\boldsymbol{F} : X \to \mathbb{R}^m$ is the set

$$G_{\boldsymbol{F}} := \big\{ \big(\boldsymbol{x}, \boldsymbol{y} \big) \in X \times \mathbb{R}^m;\ \ \boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x}) \big\} \subset X \times \mathbb{R}^m. \qquad \square$$

As we know, the graph of a function $f : \mathbb{R} \to \mathbb{R}$ can be visualized as a curve in $\mathbb{R}^2$. Similarly, the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ can be visualized as surface in $\mathbb{R}^3$. If we denote by x, y, z the Euclidean coordinates in $\mathbb{R}^3$, then the graph of a function of two variables $f(x, y)$ is described by the equation $z = f(x, y)$. You can think of the graph as describing a form of relief on Earth, where the altitude z at the point with coordinates (x, y) is $f(x, y)$; see e.g. Figure 12.2.

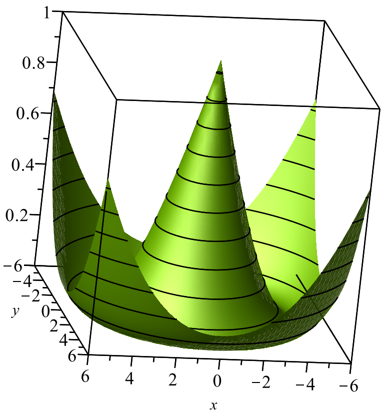

Figure 12.2. *The graph of the function* $f : [-6,6]\times[-6,6]\to\mathbb{R}$, $f(x,y) = 1-\sin\frac{\sqrt{x^2+y^2}}{3}$.

12.1. Limits and continuity

Definition 12.1.1. Let $m, n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. Suppose we are given a map $\boldsymbol{F} : X \to \mathbb{R}^m$ and a cluster point $\boldsymbol{x}_0$ of X. (*The point* $\boldsymbol{x}_0$ *need not belong to* X.)

We say that *the limit of* $\boldsymbol{F}(\boldsymbol{x})$ *when* $\boldsymbol{x}$ *approaches* $\boldsymbol{x}_0$ *is the point* $\boldsymbol{y}_0$ (in the target space $\mathbb{R}^m$) if

$$\boxed{\forall\varepsilon>0\;\;\exists\delta=\delta(\varepsilon)>0\;\;\text{such that}\;\;\forall\boldsymbol{x}\in X\backslash\{\boldsymbol{x}_0\}:\;\;\|\boldsymbol{x}-\boldsymbol{x}_0\|<\delta\Rightarrow\|\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{y}_0\|<\varepsilon}. \tag{12.1.1}$$

We will indicate this using the notation

$$\boldsymbol{y}_0 = \lim_{\boldsymbol{x}\to\boldsymbol{x}_0}\boldsymbol{F}(\boldsymbol{x}). \qquad \square$$

We have the following multidimensional counterpart of Theorem 5.1.4.

Proposition 12.1.2. *Let* $m, n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. *Suppose we are given a map* $\boldsymbol{F} : X \to \mathbb{R}^m$ *and a cluster point* $\boldsymbol{x}_0$ *of* X. *The following statements are equivalent.*

(i)

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{y}_0\in\mathbb{R}^m.$$

(ii) *For any sequence* $(\boldsymbol{x}_\nu)$ *in* $X\backslash\{\boldsymbol{x}_0\}$ *that converges to* $\boldsymbol{x}_0$ *we have*

$$\lim_{\nu\to\infty}\boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{y}_0.$$

Proof. (i) ⇒ (ii) Suppose that $(\boldsymbol{x}_\nu)$ is a sequence in $X\backslash\{\boldsymbol{x}_0\}$ that converges to $\boldsymbol{x}_0$. We have to show that, given the condition (12.1.1), the sequence $F(\boldsymbol{x}_\nu)$ converges to $\boldsymbol{y}_0$.

Let $\varepsilon > 0$. Choose $\delta(\varepsilon) > 0$ determined by (12.1.1). Since $\boldsymbol{x}_\nu \to \boldsymbol{x}_0$, there exists $N = N(\varepsilon)$ such that, for all $\nu > N(\varepsilon)$ we have $\|\boldsymbol{x}_\nu - \boldsymbol{x}_0\| < \delta(\varepsilon)$. Invoking (12.1.1) we deduce that for all $\nu > N(\varepsilon)$ we have $\|\boldsymbol{F}(\boldsymbol{x}_\nu) - \boldsymbol{y}_0\| < \varepsilon$. This proves that

$$\lim_{\nu\to\infty} \boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{y}_0.$$

(ii) ⇒ (i) We argue by contradiction. Assume that (12.1.1) is false so that

$$\exists \varepsilon_0 > 0: \;\; \forall \delta > 0,\;\; \exists \boldsymbol{x}_\delta \in X\backslash\{\boldsymbol{x}_0\}: \;\; \|\boldsymbol{x}_\delta - \boldsymbol{x}_0\| < \delta \;\text{ and }\; \|\boldsymbol{F}(\boldsymbol{x}_\delta) - \boldsymbol{y}_0\| \geqslant \varepsilon_0.$$

Thus, if we choose δ of the form $\delta = \frac{1}{\nu}$, $\nu \in \mathbb{N}$, we deduce that for any $\nu \in \mathbb{N}$ there exists $\boldsymbol{x}_\nu \in X\backslash\{\boldsymbol{x}_0\}$ such that

$$\|\boldsymbol{x}_\nu - \boldsymbol{x}_0\| < \frac{1}{\nu} \;\text{ and }\; \|F(\boldsymbol{x}_\nu) - \boldsymbol{y}_0\| \geqslant \varepsilon_0.$$

This shows that the sequence $(\boldsymbol{x}_\nu)$ in $X\backslash\{\boldsymbol{x}_0\}$ converges to $\boldsymbol{x}_0$, but the sequence $\boldsymbol{F}(\boldsymbol{x}_\nu)$ does not converge to $\boldsymbol{y}_0$. This contradicts (ii). □

Definition 12.1.3 (Continuity)**.** Let $m, n \in \mathbb{N}$, $X \subset \mathbb{R}^n$.

(i) A map $\boldsymbol{F}: X \to \mathbb{R}^m$ is said to be *continuous at* $\boldsymbol{x}_0 \in X$ if

$$\boxed{\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 \;\text{ such that }\; \forall \boldsymbol{x} \in X: \;\; \|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta \Rightarrow \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_0)\| < \varepsilon}. \tag{12.1.2}$$

(ii) A map $\boldsymbol{F}: X \to \mathbb{R}^m$ is said to be *continuous on* X if it is continuous at every point $\boldsymbol{x}_0 \in X$.

□

Proposition 12.1.4. *Let $m, n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. Consider a map*

$$\boldsymbol{F}: X \to \mathbb{R}^m, \;\; \boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} F^1(\boldsymbol{x}) \\ \vdots \\ F^m(\boldsymbol{x}) \end{bmatrix}.$$

The following statements are equivalent.

(i) *The map $\boldsymbol{F}$ is continuous at $\boldsymbol{x}_0$.*

(ii) *For any sequence $(\boldsymbol{x}_\nu)$ in X that converges to $\boldsymbol{x}_0$ we have*

$$\lim_{\nu\to\infty} \boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{F}(\boldsymbol{x}_0).$$

(iii) *The components $F^1, \dots, F^m : X \to \mathbb{R}$ are continuous at $\boldsymbol{x}_0$.*

Proof. The proof of the equivalence (i) $\Longleftrightarrow$ (ii) is identical to the proof of Proposition 12.1.2 and the details are left to the reader. The proof of the equivalence (ii) $\Longleftrightarrow$ (iii) relies on the equivalence (i) $\Longleftrightarrow$ (ii).

(ii) $\Longleftrightarrow$ (iii) According to the equivalence (i) $\Longleftrightarrow$ (ii) applied to each component F^i individually, the functions $F^1, \dots, F^m$ are continuous at $\boldsymbol{x}_0$ *if and only if*, for any sequence $(\boldsymbol{x}_\nu)$ in X that converges to $\boldsymbol{x}_0$ we have

$$\lim_{\nu\to\infty} F^i(\boldsymbol{x}_\nu) = F^i(\boldsymbol{x}_0),\ \ i = 1, 2, \dots, m.$$

Proposition 11.4.3 shows that these conditions are equivalent to

$$\lim_{\nu\to\infty} \boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{F}(\boldsymbol{x}_0).$$

In turn, this is equivalent to the continuity of $\boldsymbol{F}$ at $\boldsymbol{x}_0$. □

Example 12.1.5. The *multiplication function* $\mu : \mathbb{R}^2 \to \mathbb{R}$ given by $\mu(x, y) = xy$ is continuous. We will prove this using Proposition 12.1.4. Consider a point $\boldsymbol{p}_0 = (x_0, y_0) \in \mathbb{R}^2$.

If $\boldsymbol{p}_\nu = (x_\nu, y_\nu) \in \mathbb{R}^2$ is a sequence of points converging to $\boldsymbol{p}_0$, then $x_\nu \to x_0$ and $y_\nu \to y_0$ as $\nu \to \infty$. Hence

$$\lim_{\nu\to\infty} \mu(\boldsymbol{p}_\nu) = \lim_{\nu\to\infty} (x_\nu y_\nu) = x_0 y_0 = \mu(\boldsymbol{p}_0).$$ □

Definition 12.1.6 (Paths)**.** Let $n \in \mathbb{N}$. A *continuous path* in $\mathbb{R}^n$ is a continuous map

$$\boldsymbol{\gamma} : I \to \mathbb{R}^n,$$

where $I \subset \mathbb{R}$ is an interval. □

A path $\boldsymbol{\gamma} : I \to \mathbb{R}^n$ is completely determined by its components

$$\gamma^1, \dots, \gamma^n : I \to \mathbb{R}$$

which are continuous functions. It is convenient to think of the interval I as a *time* interval so the components γ^i are functions of time, $\gamma^i = \gamma^i(t)$. As time goes by, the point

$$\boldsymbol{\gamma}(t) = \begin{bmatrix} \gamma^1(t) \\ \vdots \\ \gamma^n(t) \end{bmatrix} \in \mathbb{R}^n$$

moves in space. Thus we can think of a path as describing the motion of a point in space during a given interval of time I. The image of a path $\boldsymbol{F} : I \to \mathbb{R}^n$ is the trajectory of this motion and it typically looks like a curve. Traditionally, a path is indicated by a system of equations

$$x^i = \gamma^i(t),\ \ i = 1, \dots, n,$$

meaning that the coordinates $x^1, \dots, x^n$ of the moving point at time t are given by the functions $\gamma^1(t), \dots, \gamma^n(t)$.

Example 12.1.7. For example, the trajectory of the path

$$\boldsymbol{\gamma} : [0, 4\pi] \to \mathbb{R}^2,\ \ \boldsymbol{\gamma}(t) = \begin{bmatrix} (t+1)\cos(2t) \\ (t+1)\sin(2t) \end{bmatrix} \in \mathbb{R}^2$$

is the *spiral* depicted in Figure 12.3. □

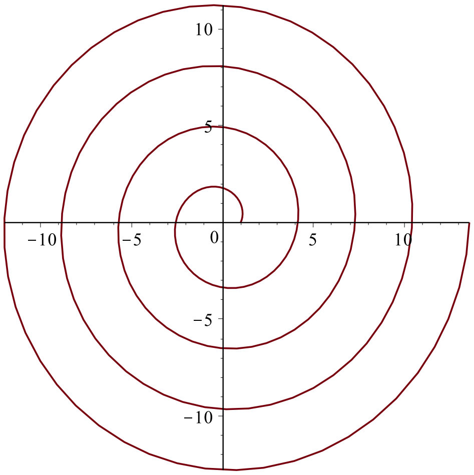

Figure 12.3. *A linear spiral* $x = (1+t)\cos 2t$, $y = (1+t)\sin 2t$, $t \in [0, 4\pi]$.

Definition 12.1.8. Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. A map $\boldsymbol{F} : X \to \mathbb{R}^m$ is called *Lipschitz* if it admits a *Lipschitz constant*, i.e., a constant $L > 0$ such that

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in X. \tag{12.1.3}$$

□

Proposition 12.1.9. *Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. Then a Lipschitz map $\boldsymbol{F} : X \to \mathbb{R}^m$ is continuous.*

Proof. Fix a Lipschitz constant $L > 0$ as in the Lipschitz condition (12.1.3). Let $\boldsymbol{x}_0 \in X$ be an arbitrary point in X. To prove that F is continuous at $\boldsymbol{x}_0$ we use Proposition 12.1.4(ii). Suppose that $(\boldsymbol{x}_\nu)$ is a sequence of points in X such that

$$\lim_{\nu \to \infty} \boldsymbol{x}_\nu = \boldsymbol{x}_0.$$

From the Lipschitz condition we deduce

$$\|\boldsymbol{F}(\boldsymbol{x}_\nu) - \boldsymbol{F}(\boldsymbol{x}_0)\| \leqslant L\|\boldsymbol{x}_\nu - \boldsymbol{x}_0\|.$$

Invoking the Squeezing Principle Proposition 4.2.8 we conclude that

$$\lim_{\nu \to \infty} \|\boldsymbol{F}(\boldsymbol{x}_\nu) - \boldsymbol{F}(\boldsymbol{x}_0)\| = 0 \Rightarrow \lim_{\nu \to \infty} \boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{F}(\boldsymbol{x}_0).$$

This proves that $\boldsymbol{F}$ is continuous at $\boldsymbol{x}_0$. □

Proposition 12.1.10. *Let $m, n \in \mathbb{N}$. The following hold.*

(i) *The norm functions*

$$\mathbb{R}^n \ni \boldsymbol{x} \mapsto \|\boldsymbol{x}\| \in \mathbb{R}, \;\; \mathbb{R}^n \ni \boldsymbol{x} \mapsto \|\boldsymbol{x}\|_\infty$$

are Lipschitz.

(ii) *Any linear form* $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ *is Lipschitz.*

(iii) *Any linear operator* $A : \mathbb{R}^n \to \mathbb{R}^m$ *is Lipschitz.*

In particular, all the maps above are continuous.

Proof. (i) Using (11.3.2b) and (11.3.4b) we deduce

$$\big|\,\|\boldsymbol{x}\| - \|\boldsymbol{y}\|\,\big| \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|, \;\; \big|\,\|\boldsymbol{x}\|_\infty - \|\boldsymbol{y}\|_\infty\,\big| \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|_\infty$$

which shows that the constant 1 is a Lipschitz constant of both functions $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$.

(ii) Let $\boldsymbol{\xi}_\uparrow$ be the dual of $\boldsymbol{\xi}$ defined in Proposition 11.2.9 . We recall that this means that $\boldsymbol{\xi}_\uparrow$ is the unique vector in $\mathbb{R}^n$ such that

$$\boldsymbol{\xi}(\boldsymbol{x}) = \langle \boldsymbol{\xi}_\uparrow, \boldsymbol{x}\rangle.$$

If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then

$$\big|\,\boldsymbol{\xi}(\boldsymbol{x}) - \boldsymbol{\xi}(\boldsymbol{y})\,\big| = \big|\,\boldsymbol{\xi}(\boldsymbol{x}-\boldsymbol{y})\,\big| = \big|\,\langle \boldsymbol{\xi}_\uparrow, \boldsymbol{x}-\boldsymbol{y}\rangle\,\big|$$

(use the Cauchy-Schwarz inequality)

$$\leqslant \|\boldsymbol{\xi}_\uparrow\| \cdot \|\boldsymbol{x}-\boldsymbol{y}\|.$$

This proves that $\boldsymbol{\xi}$ is Lipschitz, and the norm of $\|\boldsymbol{\xi}_\uparrow\|$ is a Lipschitz constant of $\boldsymbol{\xi}$. In particular,

$$\big|\,\boldsymbol{\xi}(\boldsymbol{z})\,\big| = \big|\,\boldsymbol{\xi} \bullet \boldsymbol{z}\,\big| \leqslant \|\boldsymbol{\xi}_\uparrow\| \cdot \|\boldsymbol{z}\|, \;\; \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{12.1.4}$$

(iii) As we have seen earlier, the components of $A\boldsymbol{x}$ are linear functionals in $\boldsymbol{x}$

$$A\boldsymbol{x} = \begin{bmatrix} A^1 \bullet \boldsymbol{x} \\ \vdots \\ A^m \bullet \boldsymbol{x} \end{bmatrix},$$

where $A^1, \dots, A^m$ are the rows of the $m \times n$ matrix associated to the operator A. From (12.1.4) we deduce

$$\big|\,A^i \bullet \boldsymbol{z}\,\big| \leqslant \big\|(A^i)_\uparrow\big\| \cdot \|\boldsymbol{z}\|, \;\; \forall \boldsymbol{z} \in \mathbb{R}^n, \;\; i = 1, \dots, m.$$

Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we set $\boldsymbol{z} := \boldsymbol{x} - \boldsymbol{y}$ and we have

$$A(\boldsymbol{x}-\boldsymbol{y}) = A\boldsymbol{z} = \begin{bmatrix} A^1 \bullet \boldsymbol{z} \\ \vdots \\ A^m \bullet \boldsymbol{z} \end{bmatrix}$$

so that

$$\begin{aligned} \|A(\boldsymbol{x}-\boldsymbol{y})\|^2 &= |A^1 \bullet \boldsymbol{z}|^2 + \cdots + |A^m \bullet \boldsymbol{z}|^2 \\ &\leqslant \|(A^1)_\uparrow\|^2 \cdot \|\boldsymbol{z}\|^2 + \cdots + \|(A^m)_\uparrow\|^2 \cdot \|\boldsymbol{z}\|^2 \\ &= \big(\,\|(A^1)_\uparrow\|^2 + \cdots + \|(A^m)_\uparrow\|^2\,\big)\,\|\boldsymbol{z}\|^2 \\ &= \big(\,\|(A^1)_\uparrow\|^2 + \cdots + \|(A^m)_\uparrow\|^2\,\big)\,\|\boldsymbol{x}-\boldsymbol{y}\|^2. \end{aligned}$$

□

Remark 12.1.11. (a) If $\boldsymbol{\xi}$ is a linear functional on $\mathbb{R}^n$ described by the *row* vector

$$[\xi_1, \dots, \xi_n],$$

then $\boldsymbol{\xi}_\uparrow$ is the *column* vector

$$\boldsymbol{\xi}_\uparrow = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}$$

and

$$\|\boldsymbol{\xi}_\uparrow\| = \sqrt{\xi_1^2 + \cdots + \xi_n^2} = \sqrt{\sum_{j=1}^n \xi_j^2}.$$

(b) Suppose that A is an $m \times n$ matrix with real entries. As usual, we denote by A^i the i-th row of A and by A_j the j-th column of A. The quantity

$$\sqrt{\sum_{i=1}^m \|(A^i)_\uparrow\|^2} = \sqrt{\|(A^1)_\uparrow\|^2 + \cdots + \|(A^m)_\uparrow\|^2}$$

that appears in the proof of Proposition 12.1.10(iii) is denoted by $\|A\|_{HS}$ and it is called the *Frobenius norm* or *Hilbert-Schmidt norm* of A. It can be given an alternate and more suggestive description.

Observe first that for any $i = 1, \dots, m$, the quantity $\|(A^i)_\uparrow\|^2$ is the sum of the squares of all the entries of A located on the i-th row. We deduce

$$\|A\|_{HS}^2 = \|(A^1)_\uparrow\|^2 + \cdots + \|(A^m)_\uparrow\|^2 = \text{the sum of the squares of all the entries of } A.$$

An $m \times n$ matrix A is a collection of mn real numbers and, as such, it can be viewed as an element of the Euclidean vector space $\mathbb{R}^{mn}$. We see that the Hilbert-Schmidt norm of A is none other than the Euclidean norm of A viewed as an element of $\mathbb{R}^{mn}$. In particular, if $A, B \in \operatorname{Mat}_{m\times n}(\mathbb{R})$ then

$$\|A + B\|_{HS} \leqslant \|A\|_{HS} + \|B\|_{HS}. \tag{12.1.5}$$

We can also speak of convergent sequences of matrices.

Definition 12.1.12. *A sequence (A_ν) of $m \times n$ matrices is said to converge to the $m \times n$ matrix A if*

$$\lim_{\nu\to\infty} \|A_\nu - A\|_{HS} = 0.$$

The proof of Proposition 12.1.10(iii) shows that we have the following important inequality

$$\|A \cdot \boldsymbol{x}\| \leqslant \|A\|_{HS} \cdot \|\boldsymbol{x}\|, \quad \forall A \in \operatorname{Mat}_{m\times n}(\mathbb{R}), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{12.1.6}$$

□

Corollary 12.1.13. *The* addition function $\alpha : \mathbb{R}^2 \to \mathbb{R}$, $\alpha(x, y) = (x + y)$ *is continuous.*

Proof. As shown in Example 11.1.10(b), the function α is linear and thus continuous according to Proposition 12.1.10(ii). □

Proposition 12.1.14. *Let $\ell, m, n \in \mathbb{N}$, $X \subset \mathbb{R}^\ell$ and $Y \subset \mathbb{R}^m$. If $\boldsymbol{F} : X \to \mathbb{R}^m$ and $\boldsymbol{G} : Y \to \mathbb{R}^n$ are continuous maps such that*

$$\boldsymbol{F}(X) \subset Y,$$

then the composition $\boldsymbol{G} \circ \boldsymbol{F} : X \to \mathbb{R}^n$ is also a continuous map.

Proof. Let $\boldsymbol{x}_0 \in X$ and set $\boldsymbol{y}_0 := \boldsymbol{F}(\boldsymbol{x}_0) \in Y$. We have to prove that if $(\boldsymbol{x}_\nu)$ is a sequence in X such that $\boldsymbol{x}_\nu \to \boldsymbol{x}_0$ as $\nu \to \infty$, then

$$\lim_{\nu\to\infty} \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_\nu)) = \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_0)) = \boldsymbol{G}(\boldsymbol{y}_0).$$

We set $\boldsymbol{y}_\nu := \boldsymbol{F}(\boldsymbol{x}_\nu)$. Then $\boldsymbol{y}_\nu \in Y$ and

$$\lim_{\nu\to\infty} \boldsymbol{y}_\nu = \lim_{\nu\to\infty} \boldsymbol{F}(\boldsymbol{x}_\nu) = \boldsymbol{F}(\boldsymbol{x}_0) = \boldsymbol{y}_0,$$

since F is continuous at $\boldsymbol{x}_0$. On the other hand, since G is continuous at $\boldsymbol{y}_0$ we have

$$\lim_{\nu\to\infty} \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_\nu)) = \lim_{\nu\to\infty} \boldsymbol{G}(\boldsymbol{y}_\nu) = \boldsymbol{G}(\boldsymbol{y}_0).$$

□

Corollary 12.1.15. *Suppose that $I \subset \mathbb{R}$ is an interval, $\gamma : I \to \mathbb{R}^m$ is a continuous path and $\boldsymbol{F} : \mathbb{R}^m \to \mathbb{R}^n$ is a continuous map. Then the composition $\boldsymbol{F} \circ \gamma : I \to \mathbb{R}^n$ is also a continuous path.* □

Definition 12.1.16. Let $n \in \mathbb{N}$. For any $X \subset \mathbb{R}^n$ we denote by $C(X)$ the space of continuous functions $f : X \to \mathbb{R}$. □

Corollary 12.1.17. *Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. Then, for any $f, g \in C(X)$ and any $t \in \mathbb{R}$ the functions $f + g$, $t \cdot f$ and fg are continuous.*

Proof. Consider the maps

$$P : X \to \mathbb{R}^2, \;\; P(\boldsymbol{x}) = \begin{bmatrix} f(\boldsymbol{x}) \\ g(\boldsymbol{x}) \end{bmatrix}$$

$\mu_t : \mathbb{R} \to \mathbb{R}$, $\mu_t(u) = tu$, and $\alpha, \mu : \mathbb{R}^2 \to \mathbb{R}$, $\alpha(u, v) = u + v$, $\mu(u, v) = uv$. Each of these maps is continuous and we have

$$\alpha \circ P(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}), \;\; \mu \circ P(\boldsymbol{x}) = f(\boldsymbol{x})g(\boldsymbol{x}), \;\; \mu_t \circ f(\boldsymbol{x}) = tf(\boldsymbol{x}).$$

The desired conclusion follows by invoking Proposition 12.1.14. □

Remark 12.1.18. The set $C(X)$ is nonempty since obviously the constant functions belong to $C(X)$. However, if X consists of more than one point, then X also contains nonconstant functions. For example, given $\boldsymbol{x}_0 \in X$, the function

$$d_{\boldsymbol{x}_0} : X \to \mathbb{R},\ \ d_{\boldsymbol{x}_0}(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_0\|,$$

is continuous and nonconstant since $d_{\boldsymbol{x}_0}(\boldsymbol{x}_0) = 0$ and $d_{\boldsymbol{x}_0}(\boldsymbol{x}) > 0$, $\forall \boldsymbol{x} \in X$. □

Definition 12.1.19. Let $m, n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$.

(i) The sequence of maps $\boldsymbol{F}_\nu : X \to \mathbb{R}^m$, $\nu \in \mathbb{N}$ is said to *converge pointwisely* to the map $\boldsymbol{F} : X \to \mathbb{R}^m$ if

$$\forall \boldsymbol{x} \in X \quad \lim_{\nu\to\infty} \boldsymbol{F}_\nu(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}),$$

i.e.,

$$\forall \boldsymbol{x} \in X,\ \ \forall \varepsilon > 0,\ \ \exists N = N(\varepsilon, \boldsymbol{x}) > 0 :\ \ \forall \nu > N\ \ \|\boldsymbol{F}_\nu(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x})\| < \varepsilon.$$

(ii) The sequence of maps $\boldsymbol{F}_\nu : X \to \mathbb{R}^m$, $\nu \in \mathbb{N}$ is said to *converge uniformly* to the map $\boldsymbol{F} : X \to \mathbb{R}^m$ if

$$\forall \varepsilon > 0,\ \ \exists N = N(\varepsilon) > 0 \ \text{ such that }\ \ \forall \boldsymbol{x} \in X,\ \forall \nu > N :\ \ \|\boldsymbol{F}_\nu(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x})\| < \varepsilon.$$

□

Theorem 12.1.20. *Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. Suppose that the sequence of continuous maps $\boldsymbol{F}_\nu : X \to \mathbb{R}^m$ converges uniformly to the map $\boldsymbol{F} : X \to \mathbb{R}^m$. Then the following hold.*

(i) *The sequence $(\boldsymbol{F}_\nu)$ converges pointwisely to $\boldsymbol{F}$.*

(ii) *The map $\boldsymbol{F}$ is continuous.*

□

The proof of this theorem is very similar to the proof of Theorem 6.1.10 and is left to you as an exercise.

12.2. Connectedness and compactness

In this section we discuss two very important concepts that have many applications.

12.2.1. Connectedness.

Definition 12.2.1. Let $n \in \mathbb{N}$. A subset $X \subset \mathbb{R}^n$ is called *path connected* if any two points in X can be connected by a continuous path contained in X. More precisely, this means that for any $\boldsymbol{x}_0, \boldsymbol{x}_1 \in X$, there exists a continuous path $\boldsymbol{\gamma} : [t_0, t_1] \to \mathbb{R}^n$ satisfying the following properties.

(i) $\boldsymbol{\gamma}(t) \in X$, $\forall t \in [t_0, t_1]$.

(ii) $\boldsymbol{\gamma}(t_0) = \boldsymbol{x}_0$, $\gamma(t_1) = \boldsymbol{x}_1$.

□

Remark 12.2.2. The above definition has some built-in flexibility. Note that if, for some $t_0 < t_1$, there exists a continuous path $\boldsymbol{\gamma} : [t_0, t_1] \to X$ such that $\boldsymbol{\gamma}(t_0) = \boldsymbol{x}_0$ and $\boldsymbol{\gamma}(t_1) = \boldsymbol{x}_1$, then, for any $s_0 < s_1$, there exists a continuous path $\tilde{\boldsymbol{\gamma}} : [s_0, s_1] \to X$ such that $\tilde{\boldsymbol{\gamma}}(s_0) = \boldsymbol{x}_0$ and $\tilde{\boldsymbol{\gamma}}(s_1) = \boldsymbol{x}_1$. To see this consider the linear function

$$\ell : [s_0, s_1] \to \mathbb{R}, \;\; \ell(s) = t_0 + \frac{t_1 - t_0}{s_1 - s_0}(s - s_0).$$

This function is increasing,

$$\ell(s_0) = t_0, \;\; \ell(s_1) = t_0 + \frac{t_1 - t_0}{s_1 - s_0}(s_1 - s_0) = t_0 + t_1 - t_0 = t_1.$$

Now define $\tilde{\boldsymbol{\gamma}} : [s_0, s_1] \to X$ by setting $\tilde{\boldsymbol{\gamma}}(s) = \boldsymbol{\gamma}\big(\,\ell(s)\,\big)$. Clearly

$$\tilde{\boldsymbol{\gamma}}(s_0) = \boldsymbol{\gamma}\big(\,\ell(s_0)\,\big) = \boldsymbol{\gamma}(t_0) = \boldsymbol{x}_0$$

and, similarly, $\tilde{\boldsymbol{\gamma}}(s_1) = \boldsymbol{x}_1$. □

Proposition 12.2.3. *Let $n \in \mathbb{N}$. If $X \subset \mathbb{R}^n$ is convex, then X is path connected.*

Proof. This should be intuitively very clear because in a convex set X, any two points $\boldsymbol{x}_0, \boldsymbol{x}_1$ are connected by the line segment $[\boldsymbol{x}_0, \boldsymbol{x}_1]$ which, by definition is contained in X. Formally, the argument goes as follows. Consider the continuous path

$$\boldsymbol{\gamma} : [0, 1] \to \mathbb{R}^n, \;\; \boldsymbol{\gamma}(t) = (1 - t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \;\; \forall t \in [0, 1].$$

The image (or trajectory) of this continuous path is the line segment $[\boldsymbol{x}_0, \boldsymbol{x}_1]$ which is contained in X since X is assumed convex. □

Proposition 12.2.4. *Let $X \subset \mathbb{R}$. The following statements are equivalent.*

(i) *X is path connected.*

(ii) *X is an interval.*

Proof. The implication (ii) ⇒ (i) is immediate. If X is an interval, then X is convex and thus path connected according to the previous proposition.

Assume now that X is path connected. To prove that it is an interval we have to show (see Exercise 12.12) that for any $x_0, x_1 \in X$, $x_0 < x_1$, the interval $[x_0, x_1]$ is contained in X.

Let $x_0, x_1 \in X$, $x_0 < x_1$. We have to show that if $x_0 \leqslant u \leqslant x_1$, then $u \in X$. Since X is path connected there exists a continuous path $\gamma : [t_0, t_1] \to X \subset \mathbb{R}$ such that $\gamma(t_0) = x_0$, $\gamma(t_1) = x_1$. Since $x_0 \leqslant u \leqslant x_1$ we deduce from the intermediate value property that there exists $\tau \in [t_0, t_1]$ such that $u = \gamma(\tau)$. Since $\gamma(\tau) \in X$ we deduce $u \in X$. □

12.2.2. Compactness.

Definition 12.2.5. Let $n \in \mathbb{N}$. A subset $K \subset \mathbb{R}^n$ satisfies the *Bolzano-Weierstrass property* or *BW* for brevity, if any sequence $(\boldsymbol{p}_\nu)_{\nu\in\mathbb{N}}$ of points in K contains a subsequence that converges to a point $\boldsymbol{p}$, $\underline{\textit{also in } K}$. □

Example 12.2.6. The Bolzano-Weierstrass Theorem 4.4.8 shows that intervals in $\mathbb{R}$ of the form $[a, b]$ satisfy BW. □

Proposition 12.2.7. *Let $m, n \in \mathbb{N}$. Suppose that $K \subset \mathbb{R}^m$ and $L \subset \mathbb{R}^n$ satisfy BW. Then the Cartesian product $K \times L \subset \mathbb{R}^{m+n}$ also satisfies BW.*

Proof. Let $(\boldsymbol{p}_\nu, \boldsymbol{q}_\nu) \in K \times L$, $\nu \in N$, be a sequence of points in $K \times L$. Since K satisfies BW, the sequence $(\boldsymbol{p}_\nu)$ of points in K contains a subsequence

$$(\boldsymbol{p}_{\nu_i}) = \boldsymbol{p}_{\nu_1},\ \boldsymbol{p}_{\nu_2}, \dots$$

that converges to a point $\boldsymbol{p} \in K$. Since L satisfies BW, the subsequence $(\boldsymbol{q}_{\nu_i})$ of points in L contains a sub-subsequence $(\boldsymbol{q}_{\mu_j})$ that converges to a point $\boldsymbol{q} \in L$.

The sub-subsequence $(\boldsymbol{p}_{\mu_j})$ of the subsequence $(\boldsymbol{p}_{\nu_i})$ converges to the same limit $\boldsymbol{p}$. Thus, the subsequence $(\boldsymbol{p}_{\mu_j}, \boldsymbol{q}_{\mu_j})$ of $(\boldsymbol{p}_\nu, \boldsymbol{q}_\nu)$ converges to $(\boldsymbol{p}, \boldsymbol{q}) \in K \times L$. □

Definition 12.2.8. Let $n \in \mathbb{N}$. An n-dimensional *closed box* (or *closed rectangle*) is a subset of $\mathbb{R}^n$ of the form

$$[a_1, b_1] \times \cdots \times [a_n, b_n],\ \ a_1 \leqslant b_1, \dots, a_n \leqslant b_n.$$

An *open box* in $\mathbb{R}^n$ is a set of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$.

Note that the closed cubes are special examples of closed boxes.

Corollary 12.2.9. *The closed boxes in $\mathbb{R}^n$ satisfy BW.*

Proof. We argue by induction on n. For $n = 1$ this follows from the Bolzano-Weierstrass Theorem 4.4.8. For the inductive step suppose that $B \subset \mathbb{R}^{n+1}$ is a box,

$$B = \underbrace{[a_1, b_1] \times \cdots \times [a_n, b_n]}_{=B'} \times [a_{n+1}, b_{n+1}]$$

From the induction assumption we deduce that $B' \subset \mathbb{R}^n$ satisfies BW. Proposition 12.2.7 now implies that $B = B' \times [a_{n+1}, b_{n+1}]$ satisfies BW. □

Definition 12.2.10. Let $n \in \mathbb{N}$. A set $X \subset \mathbb{R}^n$ is called *bounded* if it is contained in some box $B \subset \mathbb{R}^n$. □

Proposition 12.2.11. *Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. The following statements are equivalent.*

(i) *The set X is bounded.*

(ii) *There exists* $R > 0$ *such that*

$$\|\boldsymbol{x}\| \leqslant R, \quad \forall \boldsymbol{x} \in X. \tag{12.2.1}$$

Proof. (i) ⇒ (ii) Suppose that X is contained in the box

$$B = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

Observe that there exists $M > 0$ large enough so that

$$[a_1, b_1], \dots, [a_n, b_n] \subset [-M, M].$$

Thus, for any $\boldsymbol{x} = (x^1, \dots, x^n) \in B$, we have

$$|x^i| \leqslant M, \quad \forall i = 1, \dots, n$$

so that

$$\|\boldsymbol{x}\|^2 = |x^1|^2 + \cdots + |x^n|^2 \leqslant nM^2.$$

Hence

$$\|\boldsymbol{x}\| \leqslant M\sqrt{n}, \quad \forall \boldsymbol{x} \in B.$$

In particular, this shows that X satisfies (12.2.1).

(ii) ⇒ (i). Suppose that X satisfies (12.2.1). Thus there exists $R > 0$ such that X is contained in the closed Euclidean ball $\overline{B_R(\mathbf{0})}$ which in turn is contained in the closed cube $\overline{C_R(\mathbf{0})}$. □

Theorem 12.2.12 (Bolzano-Weierstrass)**.** *Let* $n \in \mathbb{N}$ *and* $X \subset \mathbb{R}^n$*. The following statements are equivalent.*

(i) *The set* X *satisfies* BW*.*

(ii) *The set* X *is closed and bounded.*

Proof. (i) ⇒ (ii) Assume that X satisfies BW. We have to prove that X is bounded and closed. To prove that X is closed we have to show that if $(\boldsymbol{p}_\nu)$ is a sequence of points in X that converges to some point $\boldsymbol{p} \in \mathbb{R}^n$, then $\boldsymbol{p} \in X$.

Since X satisfies BW, the sequence $(\boldsymbol{p}_\nu)$ contains a subsequence that converges to a point $\boldsymbol{p}_* \in X$. Since the limit of any subsequence is equal to the limit of the whole sequence, we deduce $\boldsymbol{p} = \boldsymbol{p}_* \in X$.

To prove that X is bounded we argue by contradiction. Thus, the condition (12.2.1) is violated. Hence, for any $\nu \in \mathbb{N}$ there exists $\boldsymbol{x}_\nu \in X$ such that $\|\boldsymbol{x}_\nu\| > \nu$. Since X satisfies BW, the sequence $(\boldsymbol{x}_\nu)$ contains a subsequence $(\boldsymbol{x}_{\nu_i})_{i \in \mathbb{N}}$ converging to $\boldsymbol{x}_* \in X$. We deduce

$$\lim_{i \to \infty} \|\boldsymbol{x}_{\nu_i}\| = \|\boldsymbol{x}_*\| < \infty.$$

This is impossible since

$$\|\boldsymbol{x}_{\nu_i}\| > \nu_i, \quad \lim_{i \to \infty} \nu_i = \infty$$

and thus

$$\lim_{i \to \infty} \|\boldsymbol{x}_{\nu_i}\| = \infty.$$

(ii) ⇒ (i) Suppose that X is closed and bounded. Since X is bounded, it is contained in a closed box B. Suppose now that $(\boldsymbol{p}_\nu)$ is a sequence of points in X. According to Corollary 12.2.9 the box B satisfies BW, so the sequence $(\boldsymbol{p}_\nu)$ contains a subsequence $(\boldsymbol{p}_{\nu_i})$ that converges to a point $\boldsymbol{p} \in B$. On the other hand the limit of any convergent sequence of points in X is a point in X. Thus the limit of the sequence $(\boldsymbol{p}_{\nu_i})_{i\in N}$ must belong to X. This shows that X satisfies BW. □

Corollary 12.2.13. *Let $n \in \mathbb{N}$. For any $R > 0$ and any $\boldsymbol{p} \in \mathbb{R}^n$ the closed ball $\overline{B_R(\boldsymbol{p})}$ and the closed cube $\overline{C_R(\boldsymbol{p})}$ satisfy BW.* □

Proof. Indeed, the closed ball $\overline{B_R(\boldsymbol{p})}$ and the closed cube $\overline{C_R(\boldsymbol{p})}$ are closed and bounded. □

Corollary 12.2.14 (Bolzano-Weierstrass)**.** *Let $n \in \mathbb{N}$. If $(\boldsymbol{x}_\nu)$ is a bounded sequence of points in $\mathbb{R}^n$, i.e.,*

$$\exists R > 0 \text{ such that } \|\boldsymbol{x}_\nu\| \leqslant R, \quad \forall \nu \in \mathbb{N},$$

then $(\boldsymbol{x}_\nu)$ contains a convergent subsequence.

Proof. The sequence $(\boldsymbol{x}_\nu)$ is contained in a closed ball which satisfies BW. □

Definition 12.2.15. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$.

(i) A (possibly infinite) collection of subsets of $\mathbb{R}^n$ is said to *cover* X if their union contains X.

(ii) The set X is said to satisfy the *weak Heine-Borel*[1] *property* (or *wHB* for brevity) if any collection of open boxes that covers X contains a finite subcollection that covers X.

(iii) The set X is said to satisfy the *Heine-Borel property* (or *HB* for brevity) if any collection of open sets that covers X contains a finite subcollection that covers X.

□

Often we use the expression "$\mathcal{U}$ *is an open cover of* X" to indicate that $\mathcal{U}$ is a collection of open sets that covers X. Given an open cover $\mathcal{U}$ of X, we define *subcover* of $\mathcal{U}$ is a subfamily of $\mathcal{U}$ that still covers X.

Example 12.2.16. The interval $(0, 1]$ does not satisfy the HB property. Indeed, the family of open sets

$$U_n := (1/n, 2), \quad n \geqslant 2,$$

covers $(0, 1]$, but no finite subfamily covers $(0, 1]$. Indeed if $U_{n_1}, \dots, U_{n_k}$ is a finite subfamily, $n_1 < \cdots < n_k$, then

$$U_{n_1} \subset \cdots \subset U_{n_k}, \quad U_{n_1} \cup \cdots \cup U_{n_k} = U_{n_k}$$

[1] Émile Borel (1871-1956) was a French mathematician and politician. As a mathematician, he was known for his founding work in the areas of measure theory and probability.

and the interval U_{n_k} does not contain $(0,1]$. □

Lemma 12.2.17. *A set satisfies wHB if and only if it satisfies HB.*

Proof. Clearly $HB \Rightarrow wHB$ so it suffices to show only that $wHB \Rightarrow HB$. Suppose that the collection $\mathcal{U}$ of open sets covers X. Each open set U in the family $\mathcal{U}$ is the union of a collection $\mathscr{C}_U$ open cubes; see Exercise 11.35.

The family $\mathscr{C}$ of all the cubes in all the collections $\mathscr{C}_U$, $U \in \mathcal{U}$ covers X. Since X satisfies wHB, there exists a finite subfamily $\mathcal{F} \subset \mathscr{C}$ that covers X. Each cube $C \in \mathcal{F}$ is contained in some open set $U = U_C$ of the family $\mathcal{U}$. It follows that the finite subfamily

$$\{\, U_C;\ \ C \in \mathcal{F} \,\} \subset \mathcal{U}$$

covers X.

□

Theorem 12.2.18 (Heine-Borel)**.** *For any $a, b \in \mathbb{R}$, the closed interval $[a, b]$ satisfies HB.*

Proof. It suffices to verify only the wHB property. Let's observe that the open boxes in $\mathbb{R}$ are the open intervals. Suppose that $\mathcal{I} := (I_\alpha)_{\alpha \in \mathcal{A}}$ is a collection of open intervals that covers $[a, b]$. We have to prove that there exists a *finite* subcollection of $\mathcal{I}$ that covers $[a, b]$. We define

$$X := \{\, x \in [a, b];\ \ [a, x] \text{ is covered by some finite subcollection of } \mathcal{I} \,\}.$$

Note first that $a \in X$ because a is contained in some interval of the family $\mathcal{I}$. Thus X is nonempty and bounded above, and therefore it admits a supremum $x^* := \sup X$. Note that $x^* \in [a, b]$. It suffices to prove that

$$x^* \in X, \tag{12.2.2a}$$

$$x^* = b. \tag{12.2.2b}$$

Proof of (12.2.2a). Observe that there exists an increasing sequence (x_n) of points in X such that

$$x^* = \lim_{n \to \infty} x_n.$$

Since $x^* \in [a, b]$ there exists an open interval I^* in the family $\mathcal{I}$ that contains x^*. Since the sequence (x_n) converges to x^* there exists $k \in \mathbb{N}$ such that $x_k \in I^*$. The interval $[a, x_k]$ is covered by finitely many intervals $I_1, \dots, I_N \in \mathcal{I}$. Clearly $[x_k, x^*] \subset I^*$. Hence the finite collection $I^*, I_1, \dots, I_N \in \mathcal{I}$ covers $[a, x^*]$, i.e., $x^* \in X$. □

Proof of (12.2.2b). We argue by contradiction. Suppose that $x^* \neq b$. Hence $x^* < b$. Since $x^* \in X$, the interval $[a, x^*]$ is covered by finitely many open intervals $I^*, I_1, \dots, I_N \in \mathcal{I}$, where $I^* \ni x^*$. Since I^* is open, there exists $\varepsilon > 0$, such that $\varepsilon < b - x^*$ and $[x^*, x^* + \varepsilon] \subset I^*$. This shows that the interval $[a, x^* + \varepsilon]$ is covered by the finite family $I^*, I_1, \dots, I_N$ and $x^* + \varepsilon \in [a, b]$. Hence $x^* + \varepsilon \in X$. The inequality $x^* + \varepsilon > x^*$ contradicts the fact that $x^* = \sup X$. □

The proof of Theorem 12.2.18 is now complete. □

Proposition 12.2.19. *Let $m, n\in\mathbb{N}$. Suppose that $K\subset\mathbb{R}^m$ and $L\subset\mathbb{R}^n$ satisfy HB. Then the Cartesian product $K\times L\subset\mathbb{R}^{m+n}$ also satisfies HB.*

Proof. Again it suffices to verify only the wHB property. Suppose that $\mathcal{B}$ is a collection of open boxes in $\mathbb{R}^{m\times n}$ that covers $K\times L$. Each box $B\in\mathcal{B}$ is a product $B=B'\times B''$ where B' is an open box in $\mathbb{R}^m$ and B'' is an open box in $\mathbb{R}^n$. To see this note that each open box $B\in\mathcal{B}$ is a product of $m+n$ intervals

$$B=I_1\times\cdots\times I_m\times I_{m+1}\times\cdots\times I_{m+n}.$$

Then

$$B'=I_1\times\cdots\times I_m,\ \ B''=I_{m+1}\times\cdots\times I_{m+n}.$$

If you think of B as a rectangle in the xy-plane, then B' would be its "shadow" on the x axis and B'' would be its "shadow" on the y-axis. For each $\boldsymbol{x}\in K$ we denote by $\mathcal{B}_{\boldsymbol{x}}$ the subfamily of $\mathcal{B}$ consisting of boxes that intersect $\{\boldsymbol{x}\}\times L$; see Figure 12.4.

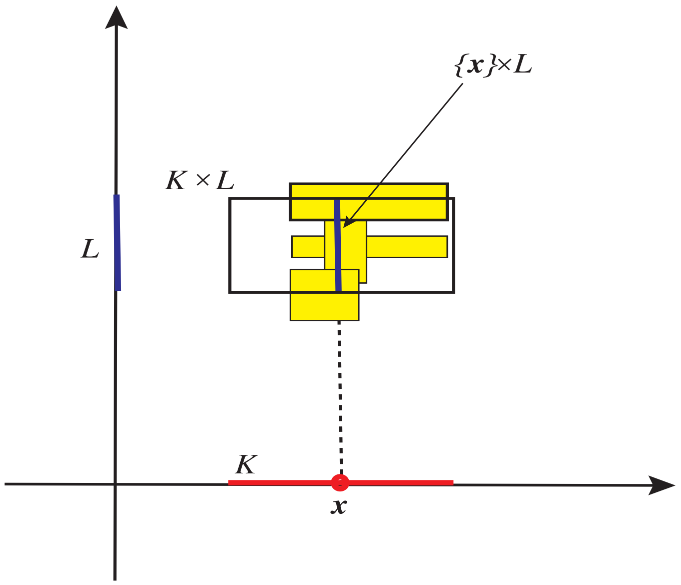

Figure 12.4. *From the collection $\mathcal{B}$ of open boxes covering $K\times L$ we concentrate on the subcollection $\mathcal{B}_{\boldsymbol{x}}$ consisting of boxes that intersect the slice $\{\boldsymbol{x}\}\times L$.*

Lemma 12.2.20. *Fix $\boldsymbol{x}\in K\subset\mathbb{R}^m$. There exists an open box $\tilde{B}_{\boldsymbol{x}}$ in $\mathbb{R}^m$ containing $\boldsymbol{x}$ and a finite subcollection $\mathcal{F}_{\boldsymbol{x}}\subset\mathcal{B}_x$ that covers $\tilde{B}_{\boldsymbol{x}}\times L$.*

Proof of Lemma 12.2.20. The collection $\mathcal{B}_{\boldsymbol{x}}$ covers $\{\boldsymbol{x}\}\times L$ and thus the collection of open n-*dimensional* boxes

$$\left\{\,B'';\ \ B'\times B''\in\mathcal{B}_{\boldsymbol{x}}\,\right\}$$

covers L. Since L satisfies HB, there exists a finite subfamily $\mathcal{F}_{\boldsymbol{x}}\subset\mathcal{B}_x$ such that the collection of n-dimensional boxes

$$\{B'';\ \ B\in\mathcal{F}_{\boldsymbol{x}}\}$$

covers L. For each $B\in\mathscr{F}_{\boldsymbol{x}}$, the m-dimensional box B' contains $\boldsymbol{x}$. The intersection of the family $\{B';\;\; B'\times B\in\mathscr{F}_{\boldsymbol{x}}\}$ is therefore a nonempty m-dimensional box $\tilde{B}_{\boldsymbol{x}}$ that contains $\boldsymbol{x}$. Since

$$\tilde{B}_{\boldsymbol{x}}\times B''\subset B'\times B''=B,\;\;\forall B\in\mathscr{F}_{\boldsymbol{x}},$$

we deduce that

$$\tilde{B}_{\boldsymbol{x}}\times L\subset\bigcup_{B\in\mathscr{F}_{\boldsymbol{x}}}B'_{\boldsymbol{x}}\times B''\subset\bigcup_{B\in\mathscr{F}_{\boldsymbol{x}}}B$$

□

For any $\boldsymbol{x}\in K$ choose an open box $\tilde{B}_{\boldsymbol{x}}\subset\mathbb{R}^m$ as in Lemma 12.2.20. The collection of boxes

$$\left\{\tilde{B}_{\boldsymbol{x}}\right\}_{\boldsymbol{x}\in K}$$

clearly covers K. Since K satisfies HB there exist finitely many points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_\nu\in K$ such that the finite subcollection

$$\left\{\tilde{B}_{\boldsymbol{x}_j}\right\}_{1\leqslant j\leqslant\nu}$$

covers K. Note that each finite subfamily $\mathscr{F}_{\boldsymbol{x}_j}\subset\mathscr{B}$ covers $\tilde{B}_{\boldsymbol{x}_j}\times L$ so the finite family

$$\mathscr{F}=\mathscr{F}_{\boldsymbol{x}_1}\cup\cdots\cup\mathscr{F}_{\boldsymbol{x}_\nu}\subset\mathscr{B}$$

covers $K\times L$.

□

Corollary 12.2.21. *Any closed box in $\mathbb{R}^n$ satisfies HB.* □

We can now state and prove the following very important result.

Theorem 12.2.22. *Let $n\in\mathbb{N}$ and $X\subset\mathbb{R}^n$. Then the following statements are equivalent.*

(i) *The set X is closed and bounded.*

(ii) *The set X satisfies BW.*

(iii) *The set X satisfies HB.*

Proof. We already know that (i) $\Longleftrightarrow$ (ii). Let us prove that (i) $\Rightarrow$ (iii). Thus we want to prove that if X is closed and bounded then X satisfies HB.

Observe first that since X is bounded X is contained in some *closed* cube C. Moreover, since X is closed, the set $U_0=\mathbb{R}^n\backslash X$ is open. Suppose that $\mathscr{U}$ is a family of open sets that covers X. The family $\mathscr{U}_*$ of open sets obtained from $\mathscr{U}$ by adding U_0 to the mix covers C. Indeed, $\mathscr{U}$ covers X and U_0 covers the rest, $C\backslash X$. The closed cube C satisfies HB so there exists a finite subfamily $\mathscr{F}_*$ of $\mathscr{U}_*$ that covers C. If $\mathscr{F}_*$ does not contain the set U_0 then clearly it is a finite subfamily of $\mathscr{U}$ that covers C and, a fortiori, X. If U_0 belongs to $\mathscr{F}_*$, then the family $\mathscr{F}$ obtained from $\mathscr{F}_*$ by removing U_0 will cover X because U_0 does not cover any point on X.

(iii) $\Rightarrow$ (i) To prove that X is bounded choose a family $\mathscr{C}$ of open cubes that covers X. Since X satisfies HB, there exists a finite subfamily $\mathscr{F}\subset\mathscr{C}$ that covers X. The union of the cubes in the *finite* family $\mathcal{F}$ is contained in some large cube, hence X is contained in a large cube and it is therefore bounded.

To prove that X is closed we argue by contradiction. Suppose that $(\boldsymbol{x}_\nu)$ is a sequence of points in X that converges to a point $\boldsymbol{x}_*$ not in X. We set $r_\nu := \operatorname{dist}(\boldsymbol{x}_*, \boldsymbol{x}_\nu)$. Consider the family of open sets

$$U_\nu = \mathbb{R}^n \setminus \overline{B_{r_\nu}(\boldsymbol{x}_*)}, \quad \nu \in \mathbb{N}.$$

Since $r_\nu \to 0$ we have

$$\bigcup_{\nu \geqslant 1} U_\nu = \mathbb{R}^n \setminus \{\boldsymbol{x}_*\} \supset X.$$

However, no finite subfamily of this family covers X. Indeed the union of the open sets in such a finite family is the complement of a closed ball centered at $\boldsymbol{x}_*$ and such a ball contains infinitely many points in the sequence $(\boldsymbol{x}_\nu)$. □

Definition 12.2.23 (Compactness)**.** Let $n \in \mathbb{N}$. A subset $X \subset \mathbb{R}^n$ is called *compact* if it satisfies either one of the equivalent conditions (i), (ii) or (iii) in Theorem 12.2.22. □

Corollary 12.2.24. *Suppose that $S \subset \mathbb{R}$ is a nonempty compact subset of the real axis. Then there exist $s_*, s^* \in S$ such that $s_* \leqslant s \leqslant s^*$, $\forall s \in S$. In other words,*

$$\boxed{\inf S \in S, \quad \sup S \in S}.$$

Proof. We set

$$s_* := \inf S, \quad s^* := \sup S.$$

Since S is compact, it is bounded so that $-\infty < s_* \leqslant s^* < \infty$. We want to prove that $s_*, s^* \in S$.

Now choose a sequence of points (s_ν) in S such that $s_\nu \to s^*$. (The existence of such a sequence is guaranteed by Lemma 6.2.3.)

Since S is compact, it is closed, so the limit of any convergent sequence of points in S is also a point in S. Thus $s^* \in S$. A similar argument shows that $s_* \in S$. □

12.3. Topological properties of continuous maps

The continuous maps enjoy many useful properties not satisfied by many other types of maps. The first property we want to discuss generalizes the intermediate value property of continuous functions of one variable.

Theorem 12.3.1. *Let $m, n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$ is path connected. If $\boldsymbol{F} : X \to \mathbb{R}^m$ is a continuous map, then its image $\boldsymbol{F}(X)$ is path connected.*

Proof. We have to show that for any $\boldsymbol{y}_0, \boldsymbol{y}_1 \in \boldsymbol{F}(X)$ there exists a continuous path in $\boldsymbol{F}(X)$ connecting $\boldsymbol{y}_0$ to $\boldsymbol{y}_1$.

Since $\boldsymbol{y}_0, \boldsymbol{y}_1 \in \boldsymbol{F}(X)$ there exist $\boldsymbol{x}_0, \boldsymbol{x}_1 \in X$ such that $\boldsymbol{F}(\boldsymbol{x}_0) = \boldsymbol{y}_0$ and $\boldsymbol{F}(\boldsymbol{x}_1) = \boldsymbol{y}_1$. Since X is path connected, there exists a continuous path $\gamma : [t_0, t_1] \to X$ such that $\gamma(t_0) = \boldsymbol{x}_0$ and $\gamma(t_1) = \boldsymbol{y}_1$. We obtain a continuous path

$$\boldsymbol{F} \circ \gamma : [t_0, t_1] \to \mathbb{R}^n$$

whose image is contained in the image $\boldsymbol{F}(X)$ of $\boldsymbol{F}$ and satisfying

$$\boldsymbol{F} \circ \gamma(t_i) = \boldsymbol{F}(\gamma(t_i)) = \boldsymbol{F}(\boldsymbol{x}_i) = y_i, \;\; i = 0, 1.$$

Thus the continuous path $\boldsymbol{F} \circ \gamma$ in $\boldsymbol{F}(X)$ connects $\boldsymbol{y}_0$ to $\boldsymbol{y}_1$. □

Corollary 12.3.2. *Let $m, n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$. If $\boldsymbol{F} : X \to \mathbb{R}^m$ is a continuous map, and the image $\boldsymbol{F}(X)$ is not path connected, then X is not path connected.* □

Corollary 12.3.3 (Multi-dimensional intermediate value theorem)**.** *Let $n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$ is a path connected subset. If $f : X \to \mathbb{R}$ is a continuous function, then its image $f(X) \subset \mathbb{R}$ is an interval. In particular, if $\boldsymbol{x}_0, \boldsymbol{x}_1 \in X$ and $c \in \mathbb{R}$ are such that $f(\boldsymbol{x}_0) < c < f(\boldsymbol{x}_1)$, then there exists $\boldsymbol{x} \in X$ such that $f(\boldsymbol{x}) = c$.*

Proof. Theorem 12.3.1 shows that $f(X) \subset \mathbb{R}$ is path connected while Proposition 12.2.4 shows that $f(X)$ must be an interval. In particular, for any points $\boldsymbol{x}_0, \boldsymbol{x}_1 \in X$ such that $f(\boldsymbol{x}_0) < f(\boldsymbol{x}_1)$, the interval $[f(\boldsymbol{x}_0), f(\boldsymbol{x}_1)] \subset \mathbb{R}$ is contained in the range $f(X)$ of f. □

Theorem 12.3.4. *Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. If $\boldsymbol{F} : X \to \mathbb{R}^m$ is continuous and $K \subset X$ is compact, then $F(K)$ is compact.*

Proof. It suffices to prove that the set $\boldsymbol{F}(K)$ satisfies BW. Suppose that $(\boldsymbol{y}_\nu)_{\nu \in \mathbb{N}}$ is a sequence in $\boldsymbol{F}(K)$. We have to show that it admits a subsequence that converges to a point in $\boldsymbol{F}(K)$.

Since $\boldsymbol{y}_\nu \in \boldsymbol{F}(K)$, there exists $\boldsymbol{x}_\nu \in K$ such that $\boldsymbol{y}_\nu = \boldsymbol{F}(\boldsymbol{x}_\nu)$. On the other hand, K satisfies BW so the sequence $(\boldsymbol{x}_\nu)$ admits a subsequence $(\boldsymbol{x}_{\nu_i})$ that converges to a point $\boldsymbol{x}_* \in X$. Since $\boldsymbol{F}$ is continuous we deduce

$$\lim_{i \to \infty} \boldsymbol{y}_{\nu_i} = \lim_{i \to \infty} \boldsymbol{F}(\boldsymbol{x}_{\nu_i}) = \boldsymbol{F}(\boldsymbol{x}_*) \in \boldsymbol{F}(K).$$

□

Corollary 12.3.5 (Weierstrass)**.** *Let $n \in \mathbb{N}$ and suppose that $K \subset \mathbb{R}^n$ is a nonempty compact set. If $f : K \to \mathbb{R}$ is continuous, then there exist $\boldsymbol{x}_*$ and $\boldsymbol{x}^*$ in K such that*

$$f(\boldsymbol{x}_*) \leqslant f(\boldsymbol{x}) \leqslant f(\boldsymbol{x}^*), \;\; \forall \boldsymbol{x} \in K.$$

Proof. According to Theorem 12.3.4 the set $f(K) \subset \mathbb{R}$ is compact. Corollary 12.2.24 implies that there exist $s_*, s^* \in f(K)$ such that $s_* = \inf f(K)$, $s^* = \sup f(K)$. In particular,

$$s_* \leqslant f(\boldsymbol{x}) \leqslant s^*, \ \ \forall \boldsymbol{x} \in K.$$

Since $s_*, s^* \in f(K)$, there exists $\boldsymbol{x}_*, \boldsymbol{x}^* \in K$ such that $s_* = f(\boldsymbol{x}_*)$, $s^* = f(\boldsymbol{x}^*)$. □

Definition 12.3.6. Let $m, n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. A map $\boldsymbol{F} : X \to \mathbb{R}^m$ is called *bounded* if its range $\boldsymbol{F}(X)$ is a bounded subset of $\mathbb{R}^m$. □

Corollary 12.3.7. *Let $m, n \in \mathbb{N}$. Suppose that $K \subset \mathbb{R}^n$ is a compact set and $\boldsymbol{F} : K \to \mathbb{R}^m$ is a continuous map. Then $\boldsymbol{F}$ is a bounded map.*

Proof. The range $\boldsymbol{F}(K)$ is compact, hence bounded. □

Definition 12.3.8. Let $n \in \mathbb{N}$. The *diameter* of a nonempty subset $S \subset \mathbb{R}^n$ is the quantity

$$\operatorname{diam}(S) = \sup_{\boldsymbol{x}, \boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\| \in [0, \infty]. \qquad □$$

We list below a few simple properties of the diameter. Their proofs are left to the reader as an exercise.

Proposition 12.3.9. *Let $n \in \mathbb{N}$. Then the following hold.*

(i) *The set $S \subset \mathbb{R}^n$ is bounded if and only if $\operatorname{diam}(S) < \infty$.*

(ii) *If $S_1 \subset S_2 \subset \mathbb{R}^n$, then $\operatorname{diam}(S_1) \leqslant \operatorname{diam}(S_2)$.*

(iii) *For any $r > 0$*

$$\operatorname{diam}(B_r) = 2r, \ \ \operatorname{diam}(C_r) = 2r\sqrt{n},$$

where $B_r, C_r \subset \mathbb{R}^n$ are the open ball and respectively the open cube of radius r centered at $\mathbf{0} \in \mathbb{R}^n$.

□

Definition 12.3.10. Let $n, m \in \mathbb{N}$, and $X \subset \mathbb{R}^n$. The *oscillation* of a function $\boldsymbol{F} : X \to \mathbb{R}^m$ on a subset S is the quantity

$$\operatorname{osc}(\boldsymbol{F}, S) = \sup_{\boldsymbol{x}, \boldsymbol{y} \in S} \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\|. \qquad □$$

The next result describes alternate characterizations of the oscillation of a *scalar* valued function. Its proof is left to you as an exercise.

Proposition 12.3.11. *Let $n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. For any function $f : X \to \mathbb{R}$ and any subset $S \subset X$ we have the equalities*

$$\operatorname{osc}(f, S) = \sup_{\boldsymbol{x} \in S} f(\boldsymbol{x}) - \inf_{\boldsymbol{y} \in S} f(\boldsymbol{y}) = \sup_{\boldsymbol{x}, \boldsymbol{y} \in S} \big(f(\boldsymbol{x}) - f(\boldsymbol{y}) \big) = \operatorname{diam} f(S). \qquad □$$

Definition 12.3.12. Let $n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. A function $f : X \to \mathbb{R}$ is said to be *uniformly continuous* on the subset $Y \subset X$ if

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 \;\text{ such that }\; \forall S \subset Y : \;\; \operatorname{diam}(S) \leqslant \delta \Rightarrow \operatorname{osc}(f, S) < \varepsilon. \qquad \square$$

Observe that the above uniform continuity condition can be rephrased in the following equivalent way.

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 \;\text{ such that }\; \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in Y : \;\; \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \leqslant \delta \Rightarrow |f(\boldsymbol{y}_1) - f(\boldsymbol{y}_2)| < \varepsilon.$$

Theorem 12.3.13 (Weierstrass)**.** *Let $n \in \mathbb{N}$, $X \subset \mathbb{R}^n$. Suppose that $f : X \to \mathbb{R}$ is continuous. Then f is* uniformly *continuous on any compact set $K \subset X$.*

Proof. Let K be a compact subset of X. We argue by contradiction so we assume that f is not uniformly continuous on K. Hence, there exists $\varepsilon_0 > 0$ such that for any $\nu \in \mathbb{N}$ there exist a subset $S_\nu \subset K$ such that

$$\operatorname{diam}(S_\nu) \leqslant \frac{1}{\nu} \;\text{ and }\; \operatorname{osc}(f, S_\nu) \geqslant \varepsilon_0.$$

Thus, for any $\nu \in \mathbb{N}$, there exist $\boldsymbol{x}_\nu, \boldsymbol{y}_\nu \in S_\nu$ such that

$$\big|\, f(\boldsymbol{x}_\nu) - f(\boldsymbol{y}_\nu) \,\big| \geqslant \frac{\varepsilon_0}{2}. \tag{12.3.1}$$

Note that because $\boldsymbol{x}_\nu, \boldsymbol{y}_\nu \in S_\nu$ and $\operatorname{diam}(S_\nu) < \frac{1}{\nu}$ we have

$$\operatorname{dist}(\boldsymbol{x}_\nu, \boldsymbol{y}_\nu) < \frac{1}{\nu} \to 0 \;\text{ as }\; \nu \to \infty.$$

Since K is compact, the sequence of points $(\boldsymbol{x}_\nu)$ in K has a convergent subsequence $(\boldsymbol{x}_{\nu_j})$

$$\lim_{j \to \infty} \boldsymbol{x}_{\nu_j} = \boldsymbol{x} \in K.$$

Observe that

$$\operatorname{dist}(\boldsymbol{y}_{\nu_j}, \boldsymbol{x}) \leqslant \underbrace{\operatorname{dist}(\boldsymbol{y}_{\nu_j}, \boldsymbol{x}_{\nu_j})}_{< \frac{1}{\nu_j}} + \operatorname{dist}(\boldsymbol{x}_{\nu_j}, \boldsymbol{x}) \to 0 \;\text{ as } j \to \infty.$$

Thus the subsequence $(\boldsymbol{y}_{\nu_j})$ also converges to $\boldsymbol{x}$. Since f is continuous at $\boldsymbol{x}$ we have

$$\lim_{j \to \infty} f(\boldsymbol{x}_{\nu_j}) = \lim_{j \to \infty} f(\boldsymbol{y}_{\nu_j}) = f(\boldsymbol{x})$$

so that

$$\lim_{j \to \infty} \big(f(\boldsymbol{x}_{\nu_j}) - f(\boldsymbol{y}_{\nu_j}) \big) = 0.$$

This contradicts (12.3.1). □

Definition 12.3.14. Let $m, n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^m$, $Y \subset \mathbb{R}^n$.

(i) A map $\boldsymbol{F} : X \to Y$ is called a *homeomorphism* if it is continuous, bijective and the inverse $\boldsymbol{F}^{-1} : Y \to X$ is also continuous.

(ii) The sets X, Y are called *homeomorphic* if there exists a homeomorphism $\boldsymbol{F} : X \to Y$.

□

Corollary 12.3.15. *Let $m, n \in \mathbb{N}$. Suppose that $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ are homeomorphic sets. Then the following hold.*

(i) *The set X is compact if and only if Y is.*

(ii) *The set X is path connected if and only if Y is.*

Proof. Fix a homeomorphism $\boldsymbol{F} : X \to Y$. Then both $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$ are continuous and

$$Y = \boldsymbol{F}(X), \quad X = \boldsymbol{F}^{-1}(Y).$$

The desired conclusions now follow from Theorem 12.3.1 and 12.3.4. □

12.4. Continuous partitions of unity

We conclude this chapter by discussing a technical but very versatile result that will come in handy later. First, we need to discuss a few more topological concepts.

Definition 12.4.1. Let $n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$.

(i) The *closure of* X, denoted by $\boldsymbol{cl}(X)$, is the intersection of all the closed subsets of $\mathbb{R}^n$ that contain X.

(ii) The *interior of* X, denoted by $\boldsymbol{int}(X)$, is the union of all the open sets contained in X.

(iii) The *boundary of* X, denoted ∂X, is the difference $\boldsymbol{cl}(X)\setminus \boldsymbol{int}(X)$.

□

In other words, the closure of a set X is the *smallest* closed subset containing X and its interior is the *largest* open set contained in X. The proof of the following result is left as an exercise.

Proposition 12.4.2. *Let $n \in \mathbb{N}$ and suppose $X \subset \mathbb{R}^n$. Then the following hold.*

(i) *A point $\boldsymbol{x} \in \mathbb{R}^n$ belongs to the closure of X if and only if there exists a sequence of points in X that converges to $\boldsymbol{x}$.*

(ii) *A point $\boldsymbol{x} \in \mathbb{R}^n$ belongs to the interior of X if and only if $\exists r > 0$ such that $B_r(\boldsymbol{x}) \subset X$.*

(iii) *$\partial X = \boldsymbol{cl}(X) \cap \boldsymbol{cl}(\mathbb{R}^n \setminus X)$.*

□

Example 12.4.3. Using the above proposition it is not hard to see that for any $r > 0$, the closure of the open ball $B_r(\mathbf{0}) \subset \mathbb{R}^n$ is the closed ball $\overline{B_r(\mathbf{0})}$. Moreover

$$\partial B_r(\mathbf{0}) = \partial \overline{B_r(\mathbf{0})} := \Sigma_r(\mathbf{0}) = \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| = r \,\big\}. \qquad \square$$

Definition 12.4.4. Let $n \in \mathbb{N}$. The *support* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the subset $\operatorname{supp}(f) \subset \mathbb{R}^n$ defined as the closure of the set of points where f is not zero,

$$\operatorname{supp}(f) := \boldsymbol{cl}\Big(\big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ f(\boldsymbol{x}) \neq 0 \,\big\}\Big).$$

We denote by $C_{\mathrm{cpt}}(\mathbb{R}^n)$ the set of *continuous* functions on $\mathbb{R}^n$ with compact support. $\square$

Clearly, the function identically equal to zero has compact support: its support is empty. The function which is equal to 1 at the origin and zero elsewhere has compact support, but it is not continuous. It turns out that there are plenty of continuous functions with compact support. The next result describes a simple recipe for producing many examples of continuous functions $\mathbb{R}^n \to \mathbb{R}$.

Proposition 12.4.5. *Let $n \in \mathbb{N}$.*

(i) *Suppose that $C, C' \subset \mathbb{R}^n$ are two closed subsets such that $C \cap C' = \varnothing$. Then there exists a continuous function $f : \mathbb{R}^n \to [0,1]$ such that*

$$C = f^{-1}(1),\ \ C' = f^{-1}(0).$$

(ii) *For any positive real numbers $r < R$ and any $\boldsymbol{x}_0 \in \mathbb{R}^n$ there exists a continuous function $f : \mathbb{R}^n \to [0,1]$ such that*

$$f(\boldsymbol{y}) = \begin{cases} 1, & \boldsymbol{y} \in \overline{B_r(\boldsymbol{x}_0)}, \\ \\ 0, & \boldsymbol{y} \in \mathbb{R}^n \setminus B_R(\boldsymbol{x}_0). \end{cases}$$

Proof. (i) We have (see Exercise 12.22)

$$\boldsymbol{x} \in C \Longleftrightarrow \operatorname{dist}(\boldsymbol{x}, C) = 0,\ \ \boldsymbol{x} \in C' \Longleftrightarrow \operatorname{dist}(\boldsymbol{x}, C') = 0.$$

Since C and C' are disjoint we deduce

$$\operatorname{dist}(\boldsymbol{x}, C) + \operatorname{dist}(\boldsymbol{x}, C') > 0,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Now define

$$f : \mathbb{R}^n \to \mathbb{R},\ \ f(\boldsymbol{x}) = \frac{\operatorname{dist}(\boldsymbol{x}, C')}{\operatorname{dist}(\boldsymbol{x}, C) + \operatorname{dist}(\boldsymbol{x}, C')}.$$

The function f is continuous (see Exercise 12.22) and $f(\boldsymbol{x}) \in [0,1]$, $\forall \boldsymbol{x} \in [0,1]$. Note that

$$f(\boldsymbol{x}) = 0 \Longleftrightarrow \operatorname{dist}(\boldsymbol{x}, C') = 0 \Longleftrightarrow \boldsymbol{x} \in C',$$

$$f(\boldsymbol{x}) = 1 \Longleftrightarrow \operatorname{dist}(\boldsymbol{x}, C) = 0 \Longleftrightarrow \boldsymbol{x} \in C.$$

(ii) This is a special case of (i) corresponding to $C = \overline{B_r(\boldsymbol{x}_0)}$ and $C' = \mathbb{R}^n \setminus B_R(\boldsymbol{x}_0)$. $\square$

Definition 12.4.6. Let $n \in \mathbb{N}$, $X \subset \mathbb{R}^n$ and suppose that $\mathcal{U}$ is an open cover of X. A *continuous partition of unity* on X, *subordinated to the open cover* $\mathcal{U}$ is a finite collection of continuous functions $\chi_1, \dots, \chi_\ell : \mathbb{R}^n \to [0,1]$ with the following properties.

(i) For any $i = 1, \dots, \ell$, there exists an open subset U_i in the collection $\mathcal{U}$ such that $\operatorname{supp} \chi_i \subset U_i$.

(ii) $\chi_1(\boldsymbol{x}) + \cdots + \chi_\ell(\boldsymbol{x}) = 1$, $\forall \boldsymbol{x} \in X$.

The partition of unity $\chi_1, \dots, \chi_\ell$ is called *compactly supported* if, additionally,

$$\chi_1, \dots, \chi_\ell \in C_{\mathrm{cpt}}(\mathbb{R}^n). \qquad \square$$

Theorem 12.4.7 (Continuous partitions of unity)**.** *Let $n \in \mathbb{N}$ and suppose that $K \subset \mathbb{R}^n$ is a compact subset. Then, for any open cover $\mathcal{U}$ of K, there exists a compactly supported partition of unity on K subordinated to $\mathcal{U}$.*

Proof. Since the collection $\mathcal{U}$ covers K we deduce that for any $\boldsymbol{x} \in K$ there exists an open set $U_{\boldsymbol{x}}$ in the collection $\mathcal{U}$ such that $\boldsymbol{x} \in U_{\boldsymbol{x}}$. For any $\boldsymbol{x} \in K$ choose $r(\boldsymbol{x}), R(\boldsymbol{x}) > 0$ such that $R(\boldsymbol{x}) > r(\boldsymbol{x})$ and $B_{R(\boldsymbol{x})}(\boldsymbol{x}) \subset U_{\boldsymbol{x}}$.

The family of open balls $\big(B_{r(\boldsymbol{x})}(\boldsymbol{x}) \big)_{\boldsymbol{x} \in K}$ obviously covers K and, since K is compact, we can find finitely many points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_\ell$ such that the collection of open balls

$$B_{r(\boldsymbol{x}_1)}(\boldsymbol{x}_1), \dots, B_{r(\boldsymbol{x}_\ell)}(\boldsymbol{x}_\ell)$$

covers K. Using Proposition 12.4.5(ii) we deduce that, for any $i = 1, \dots, \ell$ there exists a continuous function $f_i : \mathbb{R}^n \to [0,1]$ such that

$$f_i(\boldsymbol{y}) = \begin{cases} 1, & \boldsymbol{y} \in \overline{B_{r(\boldsymbol{x}_i)}(\boldsymbol{x}_i)}, \\ \\ 0, & \boldsymbol{y} \in \mathbb{R}^n \backslash B_{R(\boldsymbol{x}_i)}(\boldsymbol{x}_i). \end{cases}$$

Now define

$$\chi_1 := f_1, \;\; \chi_2 := (1 - f_1) f_2, \;\; \chi_3 := (1 - f_1)(1 - f_2) f_3,$$
$$\chi_j := (1 - f_1) \cdots (1 - f_{j-1}) f_j, \;\; \forall j = 2, \dots, \ell.$$

Note that

$$f_i(\boldsymbol{y}) = 0, \;\; \forall \boldsymbol{y} \in \mathbb{R}^n \backslash B_{R(\boldsymbol{x}_i)}(\boldsymbol{x}_i) \Rightarrow \chi_i(\boldsymbol{y}) = 0, \;\; \forall \boldsymbol{y} \in \mathbb{R}^n \backslash B_{R(\boldsymbol{x}_i)}(\boldsymbol{x}_i).$$

In particular, the function χ_j has compact support contained in $\overline{B_{R(\boldsymbol{x}_j)}(\boldsymbol{x}_j)}$. Since χ_j is the product of functions with values in $[0,1]$, the function χ_j is also valued in $[0,1]$.

Now observe that

$$\chi_1 + \chi_2 = 1 - (1 - f_1) + (1 - f_1) f_2 = 1 - (1 - f_1)(1 - f_2),$$
$$\chi_1 + \chi_2 + \chi_3 = 1 - (1 - f_1)(1 - f_2) + (1 - f_1)(1 - f_2) f_3$$
$$= 1 - \Big((1 - f_1)(1 - f_2) - (1 - f_1)(1 - f_2) f_3 \Big) = 1 - (1 - f_1)(1 - f_2)(1 - f_3).$$

We obtain inductively that

$$\chi_1 + \chi_2 + \cdots + \chi_\ell = 1 - (1 - f_1)(1 - f_2)\cdots(1 - f_\ell).$$

Finally note that

$$\begin{gathered}
\boldsymbol{x} \in \bigcup_{j=1}^{\ell} B_{r(\boldsymbol{x}_j)}(\boldsymbol{x}_j) \Rightarrow \exists i : \;\; \boldsymbol{x} \in B_{r(\boldsymbol{x}_i)}(\boldsymbol{x}_i) \\
\Rightarrow \exists i : \;\; f_i(\boldsymbol{x}) = 1 \Rightarrow \prod_{j=1}^{\ell} \big(1 - f_j(\boldsymbol{x})\big) = 0 \Rightarrow \sum_{j=1}^{\ell} \chi_j(\boldsymbol{x}) = 1.
\end{gathered}$$

□

12.5. Exercises

Exercise 12.1. Consider the function

$$f:\mathbb{R}^2\backslash\{\mathbf{0}\}\to\mathbb{R},\ \ f(x,y)=\frac{xy}{x^2+y^2}.$$

Decide whether the limit

$$\lim_{\boldsymbol{p}\to\mathbf{0}} f(\boldsymbol{p})$$

exists. Justify your answer.

Hint: Analyze the behavior of f along the sequences

$$\boldsymbol{p}_\nu=(1/\nu,1/\nu)\ \text{ and }\ \boldsymbol{q}_\nu=(1/\nu,2/\nu).$$

□

Exercise 12.2. Let $n\in\mathbb{N}$. Prove that the function $f:\mathbb{R}^n\to\mathbb{R}$, $f(\boldsymbol{x})=\|\boldsymbol{x}\|^2$ is not Lipschitz. □

Exercise 12.3. Let $n\in\mathbb{N}$ and suppose that

$$\boldsymbol{\alpha}:[a,b]\to\mathbb{R}^n,\ \ \boldsymbol{\beta}:[b,c]\to\mathbb{R}^n$$

are two continuous paths such that $\boldsymbol{\alpha}(b)=\boldsymbol{\beta}(b)$, i.e., $\boldsymbol{\alpha}$ ends where $\boldsymbol{\beta}$ begins. Define

$$\boldsymbol{\alpha}*\boldsymbol{\beta}:[a,c]\to\mathbb{R}^n,\ \ \boldsymbol{\alpha}*\boldsymbol{\beta}(t)=\begin{cases}\boldsymbol{\alpha}(t), & t\in[a,b],\\ \boldsymbol{\beta}(t), & t\in(b,c].\end{cases}$$

Prove that $\boldsymbol{\alpha}*\boldsymbol{\beta}$ is a *continuous* path. □

Exercise 12.4. (a) Consider a map $\boldsymbol{F}:\mathbb{R}^n\to\mathbb{R}^m$. Show that the following statements are equivalent.

(i) The map $\boldsymbol{F}$ is continuous.

(ii) For any open set $U\subset\mathbb{R}^m$, the preimage $\boldsymbol{F}^{-1}(U)$ is open.

(iii) For any closed set $C\subset\mathbb{R}^m$, the preimage $\boldsymbol{F}^{-1}(C)$ is closed.

(b) Suppose that D is an open subset of $\mathbb{R}^n$ and $\boldsymbol{F}:D\to\mathbb{R}^m$ is a map. Show that the following statements are equivalent.

(i) The map $\boldsymbol{F}$ is continuous.

(ii) For any open set $U\subset\mathbb{R}^m$, the preimage $\boldsymbol{F}^{-1}(U)$ is open.

Hint. (a) You need to understand very well the definition of preimage (1.4.2). □

Exercise 12.5. Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is continuous and $c\in\mathbb{R}$.

(i) Prove that the set $E_1=\big\{\boldsymbol{x}\in\mathbb{R}^n;\ \ f(\boldsymbol{x})<c\big\}$ is open.

(ii) Prove that the set $E_2=\big\{\boldsymbol{x}\in\mathbb{R}^n;\ \ f(\boldsymbol{x})\leqslant c\big\}$ is closed.

(iii) Prove that the set $E_3=\big\{\boldsymbol{x}\in\mathbb{R}^n;\ \ f(\boldsymbol{x})=c\big\}$ is closed.

(iv) Find an example of a function $f : \mathbb{R} \to \mathbb{R}$ that is not continuous yet, for any $c \in \mathbb{R}$, the set $\big\{ x \in \mathbb{R};\;\; f(x) \leqslant c \big\}$ is closed.

Hint. (i)-(iii) Use the previous exercise and Example 11.3.6. □

Exercise 12.6. (a) Suppose that $A \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ and $B \in \mathrm{Mat}_{n\times p}(\mathbb{R})$. Prove that

$$\|A\|_{HS}^2 = \mathrm{tr}(A^\top A) = \mathrm{tr}(AA^\top),$$

and

$$\|A \cdot B\|_{HS} \leqslant \|A\|_{HS} \cdot \|B\|_{HS}, \tag{12.5.1}$$

where $\| - \|_{HS}$ denotes the Hilbert-Schmidt norm defined in Remark 12.1.11 and "tr" denotes the trace of a square matrix defined in Exercise 11.21.

(b) Compute $\|A\|_{HS}$, where A is the 2×2 matrix

$$A = \left[\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}\right].$$

(c) Show that if (A_ν), (B_ν) are two sequences in $\mathrm{Mat}_n(\mathbb{R})$ that converge (see Definition 12.1.12) to the matrices A and respectively B, then $A_\nu B_\nu$ converges to AB.

Hint. (a) Denote by $(A \cdot B)^i_j$ the (i,j)-entry of the product matrix $A \cdot B$. Use (11.1.15) to prove that

$$\big|\,(A \cdot B)^i_j\,\big| \leqslant \|\,(A^i)_\uparrow\,\| \cdot \|B_j\|.$$

(c) Use the same strategy as in the proof of Proposition 11.4.8 . □

Exercise 12.7. Suppose that $(A_\nu)_{\nu\geqslant 1}$ is a sequence of $n \times n$ matrices and $A \in \mathrm{Mat}_n(\mathbb{R})$. Prove that the following statements are equivalent.

(i)

$$\lim_{\nu\to\infty} \|A_\nu - A\|_{HS} = 0.$$

(ii) For any $\boldsymbol{x} \in \mathbb{R}^n$

$$\lim_{\nu\to\infty} A_\nu \boldsymbol{x} = A\boldsymbol{x}.$$

(iii) For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\lim_{\nu\to\infty} \langle A_\nu \boldsymbol{x}, \boldsymbol{y} \rangle = \langle A\boldsymbol{x}, \boldsymbol{y} \rangle.$$

(iv) If the entries of A_ν are $A^i_j(\nu)$, $1 \leqslant i, j \leqslant n$, and the entries of A are A^i_j, then

$$\lim_{\nu\to\infty} A^i_j(\nu) = A^i_j, \;\; \forall 1 \leqslant i, j \leqslant n.$$

Hint. (i) ⇒ (ii) Use (12.5.1). (ii) ⇒ (iii) Use Cauchy-Schwarz. (iii) ⇒ (iv) Use Exercise 11.23. (iv) ⇒ (i) Use the definition of the Frobenius norm. □

Exercise 12.8. To a matrix $R \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ we associate the series of matrices

$$\mathbb{1} + R + R^2 + \cdots$$

with partial sums

$$S_0 = \mathbb{1}, \;\; S_1 = \mathbb{1} + R, \;\; S_2 = \mathbb{1} + R + R^2, \cdots .$$

(i) Show that if $\|R\|_{HS} < 1$, then the sequence (S_N) is convergent to a matrix S satisfying $S(\mathbb{1} - R) = (\mathbb{1} - R)S = \mathbb{1}$, i.e., $\mathbb{1} - R$ is invertible and its inverse is S.

(ii) Prove that the matrix S above satisfies

$$\|S - \mathbb{1}\|_{HS} \leqslant \frac{\|R\|_{HS}}{1 - \|R\|_{HS}}.$$

Hint: (i) +(ii) Use the results in Exercises 11.40 and 12.6. □

Exercise 12.9. Suppose that A is an invertible $n \times n$ matrix. Prove that there exists $\varepsilon > 0$ such that if B is an $n \times n$ matrix satisfying $\|A - B\|_{HS} < \varepsilon$, then B is also invertible.

Hint. Write $C = A - B$ so that $B = A - C = A(\mathbb{1} - A^{-1}C)$. Thus, to prove that B is invertible it suffices to show that $\mathbb{1} - A^{-1}C$ is invertible. Prove that if $\|C\|_{HS} < \frac{1}{\|A^{-1}\|_{HS}}$, then $\|A^{-1}C\|_{HS} < 1$. To conclude invoke Exercise 12.8. □

Exercise 12.10. Let $n \in \mathbb{N}$ and suppose that (A_ν) is a sequence of invertible $n \times n$ matrices that converges with respect to the Hilbert-Schmidt norm to an invertible matrix A. Prove that

$$\lim_{\nu \to \infty} \|A_\nu^{-1} - A^{-1}\|_{HS} = 0.$$

Hint: Write $C_\nu := A - A_\nu$, $R_\nu := A^{-1}C_\nu$. Observe that $C_\nu, R_\nu \to \mathbf{0}$, $A_\nu = A(\mathbb{1} - R_\nu)$ and for ν large

$$\begin{aligned} A_\nu^{-1} - A^{-1} = (\mathbb{1} - R_\nu)^{-1}A^{-1} - A^{-1} &= \Big(\mathbb{1} + R_\nu + R_\nu^2 + \cdots\Big)A^{-1} - A^{-1} \\ &= \Big(R_\nu + R_\nu^2 + \cdots\Big)A^{-1}. \end{aligned}$$

□

Exercise 12.11. Prove Theorem 12.1.20.

Hint. Mimic the proof of Theorem 6.1.10. □

Exercise 12.12. Suppose that X is a nonempty subset of the real axis $\mathbb{R}$. Prove that the following statements are equivalent.

(i) The set X is an interval, i.e., it has the form

(a, b), $[a, b)$, $(a, b]$, $[a, b]$, (a, ∞), $[a, \infty), (-\infty, b)$, or $(-\infty, b]$, or $(-\infty, \infty)$.

(ii) If $x_0, x_1 \in X$ and $x_0 < x_1$, then $[x_0, x_1] \subset X$.

(iii) The set X is convex.

Hint. Clearly (i) $\Rightarrow$ (ii) and (ii) $\Longleftrightarrow$ (iii). The tricky implication is (ii) $\Rightarrow$ (i). Set $m := \inf X$, $M := \sup X$. Show that (ii) $\Rightarrow (m, M) \subset X \subset [m, M]$. □

Exercise 12.13. (i) Prove that the set $\mathbb{R}\backslash\{0\} \subset \mathbb{R}$ is not path connected.

(ii) Prove that if L is a line in $\mathbb{R}^n$ and $\boldsymbol{p} \in L$, then the set $L\backslash\{\boldsymbol{p}\}$ is not path connected.

(iii) Suppose that $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ is a nonzero linear functional. Prove that the set

$$\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \boldsymbol{\xi}(\boldsymbol{x}) \neq 0 \,\}$$

is not path connected.

Hint. For (ii) consider a point $\boldsymbol{q} \in L \setminus \{\boldsymbol{p}\}$ so that $L = \boldsymbol{p}\boldsymbol{q}$. Define $f : \mathbb{R} \to L$, $f(t) = (1-t)\boldsymbol{p} + t\boldsymbol{q}$. Show that f is bijective, Lipschitz and $f^{-1} : L \to \mathbb{R}$ is also Lipschitz. Conclude using Corollary 12.3.15. For (iii) use (i) and Corollary 12.3.3. □

Exercise 12.14. Let $n \in \mathbb{N}$, $n > 1$.

(i) Show that the punctured space $\mathbb{R}^n \setminus \{\boldsymbol{0}\}$ is path connected.

(ii) Show that the *unit Euclidean sphere*

$$\Sigma_1(\boldsymbol{0}) := \{\boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| = 1\}$$

is path connected.

(iii) Show that for any $r > 0$ and any $\boldsymbol{p} \in \mathbb{R}^n$ the *Euclidean sphere of center* $\boldsymbol{p}$ *and radius* r, i.e., the set

$$\Sigma_r(\boldsymbol{p}) := \{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x} - \boldsymbol{p}\| = r \,\},$$

is path connected.

(iv) Prove that for any positive numbers $r < R$ the *annulus*

$$A_{r,R} := \{\boldsymbol{x} \in \mathbb{R}^n;\ \ r < \|\boldsymbol{x}\| < R\}$$

is path connected but not convex.

Hint. (i) Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. If the line $\boldsymbol{p}\boldsymbol{q}$ does not contain $\boldsymbol{0}$ we're done since the segment $[\boldsymbol{p}, \boldsymbol{q}]$ will do the trick. If $\boldsymbol{0} \in \boldsymbol{p}\boldsymbol{q}$, then choose a point $\boldsymbol{r} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ that does not belong to this line. (You need to use the assumption $n > 1$ to prove that such a point exists.) Then $\boldsymbol{0} \notin \boldsymbol{p}\boldsymbol{r}$. Travel from $\boldsymbol{p}$ to $\boldsymbol{r}$ on $[\boldsymbol{p}, \boldsymbol{r}]$ and then from $\boldsymbol{r}$ to $\boldsymbol{q}$ on $[\boldsymbol{r}, \boldsymbol{q}]$. (Need to invoke Remark 12.2.2 and Exercise 12.3.) To prove (ii) use (i). To prove (iii) use (ii). To prove (iv) it helps to first visualize the region $A_{r,R}$ in the special case $n = 2$, $r = 1$, $R = 2$. Use (iii) to prove that this annulus is path connected. □

Exercise 12.15. Let $n \in \mathbb{N}$ and suppose that $S_1, S_2 \subset \mathbb{R}^n$ are two path connected subsets such that $S_1 \cap S_2 \neq \emptyset$. Prove that $S_1 \cup S_2$ is also path connected. □

Exercise 12.16. Prove Proposition 12.3.9. □

Exercise 12.17. Let $n \in \mathbb{N}$. Suppose that $(K_\nu)_{\nu \in \mathbb{N}}$ is a sequence of nonempty compact subsets of $\mathbb{R}^n$ such that

$$K_1 \supset K_2 \supset \cdots \supset K_\nu \supset \cdots .$$

Prove that

$$\bigcap_{\nu \in \mathbb{N}} K_\nu \neq \emptyset,$$

i.e.,

$$\exists \boldsymbol{p} \in \mathbb{R}^n \quad \text{such that} \ \ \boldsymbol{p} \in K_\nu, \quad \forall \nu \in \mathbb{N}.$$

Hint. For any $\nu \in \mathbb{N}$ choose a point $\boldsymbol{p}_\nu \in K_\nu$. Show that a subsequence of $(\boldsymbol{p}_\nu)$ is convergent and then prove that its limit belongs to K_ν for any ν. □

Exercise 12.18. Let $n \in \mathbb{N}$ and suppose that $A, B \subset \mathbb{R}^n$ are nonempty. We regard $A \times B$ as a subset of $\mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$ and we consider the function

$$f : A \times B \to \mathbb{R}, \ \ f(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|.$$

Prove that f is continuous.

Hint. Use Proposition 12.1.4(ii). □

Exercise 12.19. Let $n \in \mathbb{N}$ and suppose that $K \subset \mathbb{R}^n$ is a nonempty compact subset. Recall (see Definition 12.3.8) that

$$\operatorname{diam}(K) := \sup_{\boldsymbol{x}, \boldsymbol{y} \in K} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Prove that there exist $\boldsymbol{x}_*, \boldsymbol{y}_* \in K$ such that

$$\operatorname{diam}(K) = \|\boldsymbol{x}_* - \boldsymbol{y}_*\|.$$

Hint. Use Exercise 12.18, Proposition 12.2.19, and Corollary 12.3.5. □

Exercise 12.20. Prove Proposition 12.3.11. □

Exercise 12.21. Let $X \subset \mathbb{R}^n$, and $f : X \to \mathbb{R}$ a Lipschitz function. Prove that f is uniformly continuous on X. □

Exercise 12.22. Let $n \in \mathbb{N}$. Suppose that $C \subset \mathbb{R}^n$ is a nonempty closed subset. For $\boldsymbol{x} \in \mathbb{R}^n$ we set

$$\operatorname{dist}(\boldsymbol{x}, C) := \inf_{\boldsymbol{p} \in C} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{p}).$$

(i) Prove that for any $\boldsymbol{x} \in \mathbb{R}^n$ there exists $\boldsymbol{y} \in C$ such that

$$\|\boldsymbol{x} - \boldsymbol{y}\| = \operatorname{dist}(\boldsymbol{x}, C).$$

(ii) Prove that the function $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, C)$ is Lipschitz. More precisely

$$\big| f(\boldsymbol{x}) - f(\boldsymbol{y}) \big| \leqslant \big\| \boldsymbol{x} - \boldsymbol{y} \big\|, \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

(iii) Prove that

$$C = f^{-1}(0) = \big\{ \boldsymbol{x} \in \mathbb{R}^n; \ \ \operatorname{dist}(\boldsymbol{x}, C) = 0 \big\}.$$

Hint. (i) Show that there exists a sequence $(\boldsymbol{y}_\nu)$ in C such that $\|\boldsymbol{x} - \boldsymbol{y}_\nu\| \to \operatorname{dist}(\boldsymbol{x}, C)$. Next prove that this sequence is bounded and thus it has a convergent subsequence. (ii) Use the triangle inequality, part (i) and the definition of $\operatorname{dist}(\boldsymbol{x}, C)$ to prove that $L = 1$ is a Lipschitz constant for $f(\boldsymbol{x})$. (iii) Use (i). □

Exercise 12.23. Let $n \in \mathbb{N}$ and suppose that $C \subset \mathbb{R}^n$ is a *closed, convex* subset and $\boldsymbol{x}_0 \in \mathbb{R}^n \backslash C$. Set

$$r := \operatorname{dist}(\boldsymbol{x}_0, C).$$

Prove that the sphere

$$\Sigma_r(\boldsymbol{x}_0) = \big\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; \|\boldsymbol{x} - \boldsymbol{x}_0\| = r \,\big\}$$

intersects the set C in *exactly one point.* This unique point of intersection is called the *projection of* $\boldsymbol{x}_0$ *on* C and it is denoted by $\operatorname{Proj}_C \boldsymbol{x}_0$.

Hint. You need to use Exercise 12.22. □

Exercise 12.24. Let $U \subset \mathbb{R}^n$ be an open set. Prove that the following statements are equivalent.

(i) The set U is path connected.

(ii) Any $\boldsymbol{p}, \boldsymbol{q} \in U$ can be joined by a broken line contained in U. More precisely, this means that for any $\boldsymbol{p}, \boldsymbol{q} \in U$ there exist points $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_N \in U$ such that $\boldsymbol{p} = \boldsymbol{p}_0$, $\boldsymbol{q} = \boldsymbol{p}_N$ and all the line segments

$$[\boldsymbol{p}_0, \boldsymbol{p}_1], \; [\boldsymbol{p}_1, \boldsymbol{p}_2], \dots, [\boldsymbol{p}_{N-1}, \boldsymbol{p}_N]$$

are contained in U.

Hint. (i) ⇒ (ii) Set $C = \mathbb{R}^n \backslash U$ and define $\rho : \mathbb{R}^n \to [0, \infty)$, $\rho(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, C)$. Observe that ρ is Lipschitz and thus continuous. Consider a continuous path $\boldsymbol{\gamma} : [0,1] \to U$ such that $\boldsymbol{\gamma}(0) = \boldsymbol{p}$ and $\boldsymbol{\gamma}(1) = \boldsymbol{q}$. Set $r_0 = \inf_{t \in [0,1]} \rho(\boldsymbol{\gamma}(t))$. Show that $r_0 > 0$ and $B_{r_0}(\boldsymbol{\gamma}(t)) \subset U$, $\forall t \in [0,1]$. Use the uniform continuity of $\boldsymbol{\gamma} : [0,1] \to U$ to show that, for N sufficiently large, we have

$$\|\,\boldsymbol{\gamma}(0) - \boldsymbol{\gamma}(1/N)\,\| < \frac{r_0}{2}, \;\; \dots, \|\,\boldsymbol{\gamma}\big((N-1)/N\big) - \boldsymbol{\gamma}(1)\,\| < \frac{r_0}{2},$$

and conclude that the broken line determined by the points

$$\boldsymbol{p}_0 = \boldsymbol{\gamma}(0), \; \boldsymbol{p}_1 = \boldsymbol{\gamma}(1/N), \;\; \boldsymbol{p}_i = \boldsymbol{\gamma}(i/N), \;\; i = 1, \dots, N,$$

is contained in U. □

Exercise 12.25. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function with the following property: there exist $A, B > 0$ such that

$$f(\boldsymbol{x}) \geqslant A\|\boldsymbol{x}\| - B, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n.$$

(i) Prove that for any $R > 0$ the set

$$\{f \leqslant R\} := \big\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; f(\boldsymbol{x}) \leqslant R \,\big\}$$

is compact.

(ii) Prove that there exists $\boldsymbol{x}_* \in \mathbb{R}^n$ such that $f(\boldsymbol{x}_*) \leqslant f(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.

Hint. (i) Show that the set $\{f \leqslant R\}$ is bounded. (ii) Prove that there exists a sequence $(\boldsymbol{x}_\nu)$ in $\mathbb{R}^n$ such that

$$\lim_{\nu \to \infty} f(\boldsymbol{x}_\nu) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}).$$

Deduce that the sequence $f(\boldsymbol{x}_{\boldsymbol{\nu}})$ is bounded above and then, using (i), prove that the sequence $(\boldsymbol{x}_\nu)$ is bounded and thus it has a convergent subsequence. □

Exercise 12.26. Let $n \in \mathbb{N}$, $d \in \mathbb{R}$. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *positively homogeneous of degree d* if

$$f(t\boldsymbol{x}) = t^d f(\boldsymbol{x}),\ \ \forall t > 0,\ \ \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\}.$$

(i) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a nonconstant, continuous and positively homogeneous function of degree d. Prove that $d > 0$.

(ii) Given $d \in \mathbb{R}$ construct a nonconstant function $f : \mathbb{R}^n \to \mathbb{R}$ that is positively homogeneous of degree d and it is continuous at every point $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$.

Hint. (i). Fix $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and consider the sequence $f(\nu^{-1}\boldsymbol{x})$, $\nu \in \mathbb{N}$. □

Exercise 12.27. Let $n \in \mathbb{N}$ and suppose that $d > 0$ and $f : \mathbb{R}^n \to (0,\infty)$ is continuous, positively homogeneous of degree $d > 0$ and satisfies

$$f(\boldsymbol{x}) > 0,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\}.$$

(i) Prove that there exists $c > 0$ such that

$$f(\boldsymbol{x}) \geqslant c\|\boldsymbol{x}\|^d,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

(ii) Prove that for any $r > 0$ the *sublevel set*

$$\{f \leqslant r\} := \big\{\boldsymbol{x} \in \mathbb{R}^n;\ \ f(\boldsymbol{x}) \leqslant r\,\big\}$$

is compact.

Hint. (i) Consider the unit sphere

$$\Sigma_1 := \big\{\boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| = 1\,\big\}.$$

Use Corollary 12.3.5 to show that the infimum of f on Σ_1 is *strictly positive*. (ii) Use (i). □

Exercise 12.28. For any linear operator $A : \mathbb{R}^n \to \mathbb{R}^m$ we set

$$\|A\| := \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|.$$

(i) Show that if $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, then $\|A\| < \infty$ and

$$\|A\boldsymbol{x}\| \leqslant \|A\| \cdot \|\boldsymbol{x}\|,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

(ii) Show that if $A : \mathbb{R}^n \to \mathbb{R}^m$ and $B : \mathbb{R}^m \to \mathbb{R}^\ell$ are linear operators, then

$$\|B \circ A\| \leqslant \|B\| \cdot \|A\|.$$

(iii) Show that the linear operator $A : \mathbb{R}^n \to \mathbb{R}^m$ is injective if and only if there exists $C > 0$ such that

$$\|A\boldsymbol{x}\| \geqslant C\|\boldsymbol{x}\|,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}.$$

(iv) Prove that if $A, B : \mathbb{R}^n \to \mathbb{R}^m$ are linear operators and $t \in \mathbb{R}$ then

$$\|A + B\| \leqslant \|A\| + \|B\|,\ \ \|tA\| = |t| \cdot \|A\|.$$

Hint. (iii) Use Exercise 11.13 and Exercise 12.27(i) applied to the function $f(\boldsymbol{x}) = \|A\boldsymbol{x}\|$. □

Exercise 12.29. Prove Proposition 12.4.2. □

Exercise 12.30. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$.

(i) Prove that the boundary ∂X is a closed subset of $\mathbb{R}^n$.

(ii) Show that if X is bounded, then ∂X is compact.

□

Exercise 12.31. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$. Prove that the following statements are equivalent.

(i) $\boldsymbol{cl}(X) = \mathbb{R}^n$.

(ii) The set X is dense in $\mathbb{R}^n$.

□

Exercise 12.32. Find the closures, the interiors and the boundaries of the following sets.

(i) $(0,1) \subset \mathbb{R}$.

(ii) $[0,1] \subset \mathbb{R}$

(iii) $(0,1) \times \{0\} \subset \mathbb{R}^2$.

(iv) $\big\{ (x,y) \in \mathbb{R}^2;\ \ 0 \leqslant x, y \leqslant 1 \big\} \subset \mathbb{R}^2$.

□

Exercise 12.33. Suppose that $\mathcal{O} \subset \mathbb{R}^n$ is an open subset and

$$K \subset \mathbb{R} \times \mathcal{O}$$

is a compact subset. For any $t \in \mathbb{R}$ we set

$$K_t := \big\{ \boldsymbol{x} \in \mathbb{R}^n :\ \ (t, \boldsymbol{x}) \in K \big\},\ \ T := \big\{ t \in \mathbb{R};\ \ K_t \neq \varnothing \big\}.$$

(i) Show that T is compact.

(ii) Prove that there exists a compact set $\mathcal{K} \subset \mathcal{O}$ such that

$$K_t \subset \mathcal{K},\ \ \forall t \in \mathbb{R}.$$

□

Exercise 12.34. Let $n \in \mathbb{N}$ and suppose that $K \subset \mathbb{R}^n$ is a nonempty subset. Prove that the following statements are equivalent.

(i) The set K is compact

(ii) Any continuous function $f : K \to \mathbb{R}$ is bounded.

□

Exercise 12.35. Suppose that $U \subset \mathbb{R}^n$ is an open set and $\boldsymbol{p}, \boldsymbol{q} \in U$ are points such that the line segment $[\boldsymbol{p}, \boldsymbol{q}]$ is contained in U. Prove that there exists an open *convex* set V

such that

$$[\boldsymbol{p}, \boldsymbol{q}] \subset V \subset U. \qquad \square$$

12.6. Exercises for extra credit

Exercise* 12.1. Let $n \in \mathbb{N}$ and suppose that $C \subset \mathbb{R}^n$ is a *closed, convex* subset and $\boldsymbol{x}_0 \in \mathbb{R}^n \backslash C$. Prove that there exists a linear functional $\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}$ and a real number c such that

$$\boldsymbol{\xi}(\boldsymbol{x}_0) > c > \boldsymbol{\xi}(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in C.$$

Exercise* 12.2. Let $n \in \mathbb{N}$ and suppose that $f : \mathbb{R}^n \to (0, \infty)$ is continuous and positively homogeneous of degree $d > 0$. Prove that the following statements are equivalent.

(i) The function f is uniformly continuous on $\mathbb{R}^n$.

(ii) $d \leqslant 1$.

□

Exercise* 12.3. Suppose that $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a map satisfying the following conditions.

(i) T is continuous.

(ii) T is injective.

(iii) $T(\mathbf{0}) = \mathbf{0}$, $T(\boldsymbol{i}) = \boldsymbol{i}$, $T(\boldsymbol{j}) = \boldsymbol{j}$.

(iv) For any line $\ell \subset \mathbb{R}^2$, the image $T(\ell)$ is also a line in $\mathbb{R}^2$.

Prove that $T(\boldsymbol{v}) = \boldsymbol{v}$, $\forall \boldsymbol{v} \in \mathbb{R}^2$. □

Exercise* 12.4. Prove that $\mathbb{R}^2$ is *not* homeomorphic to $\mathbb{R}^3$. □

Chapter 13

Multi-variable differential calculus

The concept of differential of a one-variable function extends to functions of several variables. The several-variable situation adds new complexity and subtleties, and the goal of the present chapter is to investigate them.

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at a point $x_0 \in \mathbb{R}$ if and only if it admits a "best" linear approximation near x_0. More geometrically, the graph of f, which is a curve in $\mathbb{R}^2$, can be well approximated in a vicinity of the point $\boldsymbol{p}_0 = (x_0, y_0) \in \mathbb{R}^2$, $y_0 = f(x_0)$, by a straight line, the tangent line to the curve at the point $\boldsymbol{p}_0$. This tangent line is graph of a function of the form $L(x) = A(x - x_0) + y_0$. The slope A of this line is the derivative of f at x_0.

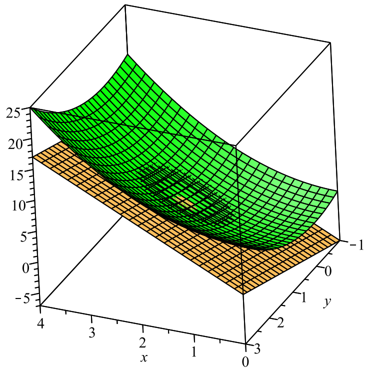

Figure 13.1. *A best linear approximation of the function $f(x, y) = x^2 + y^2$ near the point $(2, 1)$.*

We want to extend this approach to maps $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$. The graph of such map is an m-dimensional "curved" surface in $\mathbb{R}^n \times \mathbb{R}^m$. We seek to find a "best" approximation of this graph near $\boldsymbol{p}_0 = (\boldsymbol{x}_0, \boldsymbol{y}_0)$, $\boldsymbol{y}_0 = \boldsymbol{F}(\boldsymbol{x}_0)$, by a "straight" or "flat" m-dimensional surface; see Figure 13.1 where $n = 2$, $m = 1$. The "straight" surfaces in an Euclidean space are precisely the affine subspaces and we seek to approximate the graph of $\boldsymbol{F}$ near $\boldsymbol{p}_0$ by an affine subspace described as the graph of a map $L : \mathbb{R}^n \to \mathbb{R}^m$ of the form $L(\boldsymbol{x}) = A(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{y}_0$, where $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator. The concept of Fréchet derivative formalizes the above heuristics.

13.1. The differential of a map at a point

Suppose that $m, n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$ is an open subset. Since U is open, we deduce that for any point $\boldsymbol{x}_0 \in U$ there exists $r = r(\boldsymbol{x}_0) > 0$ such that the open ball $B_r(\boldsymbol{x}_0)$ is contained in U. This means that (see Figure 13.2)

$$\boldsymbol{x}_0 + \boldsymbol{h} \in U, \quad \forall \|\boldsymbol{h}\| < r.$$

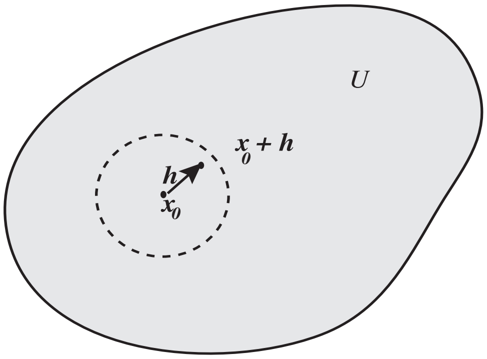

Figure 13.2. *An open set.*

Definition 13.1.1. Suppose that $\boldsymbol{F} : U \to \mathbb{R}^m$ is a map and $\boldsymbol{x}_0 \in U$. We say that F is *Fréchet*[1] *differentiable at* $\boldsymbol{x}_0$ if there exists a linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\boxed{\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{1}{\|\boldsymbol{h}\|} \Big(\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h} \Big) = \boldsymbol{0}}. \tag{13.1.1}$$

□

[1] Named after Maurice René Fréchet (1878-1973), a French mathematician. He made major contributions to point-set topology and introduced the concept of compactness; see Wikipedia.

Remark 13.1.2. Observe that if $\boldsymbol{F}$ is differentiable at $\boldsymbol{x}_0$, then there exists ***exactly one*** linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$ satisfying the condition (13.1.1). More precisely, for any $\boldsymbol{h} \in \mathbb{R}^n \backslash \boldsymbol{0}$ we have

$$\boxed{L\boldsymbol{h} = \lim_{t\to 0} \frac{1}{t}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0)\Big)}. \tag{13.1.2}$$

Indeed, consider a sequence of real numbers $t_\nu \to 0$, $t_\nu \neq 0$. Set $\boldsymbol{h}_\nu = t_\nu \boldsymbol{h}$. Note that

$$\lim_{\nu\to\infty} \boldsymbol{h}_\nu = \boldsymbol{0}$$

so that $\boldsymbol{x}_0 + \boldsymbol{h}_\nu \in U$, for large ν. We have $L\boldsymbol{h}_\nu = t_\nu L\boldsymbol{h}$ and

$$\lim_{\nu\to\infty} \left\| \frac{1}{t_\nu}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t_\nu\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0)\Big) - L\boldsymbol{h} \right\|$$

$$= \|\boldsymbol{h}\| \lim_{\nu\to\infty} \left\| \frac{1}{t_\nu\|\boldsymbol{h}\|}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t_\nu\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - t_\nu L\boldsymbol{h}\Big) \right\|$$

$$= \|\boldsymbol{h}\| \lim_{\nu\to\infty} \frac{1}{|t_\nu| \cdot \|\boldsymbol{h}\|} \left\|\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t_\nu\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - t_\nu L\boldsymbol{h}\Big) \right\|$$

$(|t_\nu| \cdot \|\boldsymbol{h}\| = \|t_\nu \boldsymbol{h}\| = \|\boldsymbol{h}_\nu\|)$

$$= \|\boldsymbol{h}\| \lim_{\nu\to\infty} \frac{1}{\|\boldsymbol{h}_\nu\|} \left\|\Big(\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_\nu) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}_\nu\Big) \right\| \overset{(13.1.1)}{=} 0.$$

□

Definition 13.1.3. The unique linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$ such that the differentiability condition (13.1.1) is satisfied is called the *(Fréchet) differential* of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ and it is denoted by $d\boldsymbol{F}(\boldsymbol{x}_0)$. □

The equality (13.1.2) shows that the Fréchet differential $d\boldsymbol{F}(\boldsymbol{x}_0)$ is determined uniquely by the equality

$$\boxed{d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{h} = \lim_{t\to 0} \frac{1}{t}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0)\Big),\ \ \forall \boldsymbol{h} \in \mathbb{R}^n}. \tag{13.1.3}$$

Remark 13.1.4. (a) Suppose that $\boldsymbol{F} : U \to \mathbb{R}^m$ is differentiable at $\boldsymbol{x}_0$ and $L := d\boldsymbol{F}(\boldsymbol{x}_0)$. The main point of Definition 13.1.1 is that, for small $\boldsymbol{h}$, the variation

$$\Delta_{\boldsymbol{h}}\boldsymbol{F}(\boldsymbol{x}_0) = \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0)$$

is very well approximated by the *linear* quantity $L\boldsymbol{h}$. For $\boldsymbol{h} \in \mathbb{R}^n$ the error of this approximation is

$$R(\boldsymbol{h}) := \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}.$$

The differentiability condition is equivalent to the fact that the error $R(\boldsymbol{h})$ is $o(\boldsymbol{h})$, where $o(\boldsymbol{h})$ stands for “a lot smaller” than $\boldsymbol{h}$ as $\boldsymbol{h} \to \boldsymbol{0}$. More precisely

$$\boxed{R(\boldsymbol{h}) = o(\boldsymbol{h}) \ \text{ as } \boldsymbol{h} \to \boldsymbol{0} \Longleftrightarrow \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{1}{\|\boldsymbol{h}\|} R(\boldsymbol{h}) = \boldsymbol{0}}. \tag{13.1.4}$$

One can prove(see Exercise 13.1) that the condition (13.1.4) is equivalent to the existence of a function

$$\varphi : [0, r) \to [0, \infty)$$

such that

$$\lim_{t \searrow 0} \varphi(t) = 0 \;\text{ and }\; \|R(\boldsymbol{h})\| \leqslant \varphi\big(\,\|\boldsymbol{h}\|\,\big)\|\,\boldsymbol{h}\|, \;\; \forall \|\boldsymbol{h}\| < r. \tag{13.1.5}$$

The equality (13.1.4) can be rewritten as

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) = L(\boldsymbol{h}) + o(\boldsymbol{h}) \;\text{ as } \boldsymbol{h} \to \boldsymbol{0},$$

or, if we set $\boldsymbol{x} := \boldsymbol{x}_0 + \boldsymbol{h}$

$$\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_0) = L(\boldsymbol{x} - \boldsymbol{x}_0) + o(\boldsymbol{x} - \boldsymbol{x}_0) \;\text{ as } \boldsymbol{x} \to \boldsymbol{x}_0. \tag{13.1.6}$$

This last equality can be taken as a definition of the Fréchet differential: *the linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$ is the Fréchet differential of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ iff it satisfies (13.1.6).*

(b) By definition, the differential $d\boldsymbol{F}(\boldsymbol{x}_0)$ is a linear operator $\mathbb{R}^n \to \mathbb{R}^m$ and, as such, it is represented by an $m \times n$ matrix sometimes called the *Jacobian matrix* of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ denoted by

$$\boxed{J_{\boldsymbol{F}}(\boldsymbol{x}_0) \;\text{ or }\; \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}(\boldsymbol{x}_0)}.$$

The n columns of the matrix $J_{\boldsymbol{F}}(\boldsymbol{x}_0)$ consist of the vectors

$$d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_1, \dots, d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_n \in \mathbb{R}^m,$$

where $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ is the natural basis of $\mathbb{R}^n$ and

$$d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_j = \lim_{t \to 0} \frac{1}{t}\big(\,\boldsymbol{F}(\boldsymbol{x}_0 + t\boldsymbol{e}_j) - \boldsymbol{F}(\boldsymbol{x}_0)\,\big), \;\; \forall j = 1, \dots, n. \tag{13.1.7}$$

□

Definition 13.1.5. Let $m, n \in \mathbb{N}$, assume that $U \subset \mathbb{R}^n$ is an open set. If $\boldsymbol{F} : U \to \mathbb{R}^m$ is Fréchet differentiable at $\boldsymbol{x}_0$ and $L : \mathbb{R}^n \to \mathbb{R}^m$ is its Fréchet derivative, then the function $\mathcal{L} = \mathcal{L}_{\boldsymbol{F}, \boldsymbol{x}_0} : \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$\mathcal{L}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}_0) + L(\boldsymbol{x} - \boldsymbol{x}_0) \tag{13.1.8}$$

is called the *linearization* or the *linear approximation* of $\boldsymbol{F}$ at $\boldsymbol{x}_0$. □

Note that the equality (13.1.4) where $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$ (equivalently, $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{h}$), implies that

$$\boldsymbol{F}(\boldsymbol{x}) - \mathcal{L}(\boldsymbol{x}) = o\big(\boldsymbol{x} - \boldsymbol{x}_0\big) \;\text{ as } \boldsymbol{x} \to \boldsymbol{x}_0$$

i.e.,

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{\|\boldsymbol{F}(\boldsymbol{x}) - \mathcal{L}(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0.$$

This shows that, when $\boldsymbol{x} \to \boldsymbol{x}_0$, the difference $\boldsymbol{F}(\boldsymbol{x}) - \mathcal{L}(\boldsymbol{x})$ is a lot smaller than the very small quantity $\|\boldsymbol{x} - \boldsymbol{x}_0\|$.

The equality (13.1.4) implies the following result.

Proposition 13.1.6. *If $U \subset \mathbb{R}^n$ is open, $\boldsymbol{x}_0 \in U$ and the map $\boldsymbol{F} : U \to \mathbb{R}^m$ is Fréchet differentiable at $\boldsymbol{x}_0$, then it is continuous at $\boldsymbol{x}_0$.*

Proof. Using the notation from Remark 13.1.2 we can write

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) = \boldsymbol{F}(\boldsymbol{x}_0) + L\boldsymbol{h} + R(\boldsymbol{h}).$$

Since L is a linear operator, it is a continuous map and thus

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} L\boldsymbol{h} = \boldsymbol{0}.$$

On the other hand, (13.1.4) shows that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} R(\boldsymbol{h}) = \boldsymbol{0}.$$

Hence

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) = \lim_{\boldsymbol{h}\to 0} \big(\boldsymbol{F}(\boldsymbol{x}_0) + L\boldsymbol{h} + R(\boldsymbol{h}) \big) = \boldsymbol{F}(\boldsymbol{x}_0).$$

□

Example 13.1.7. Before we proceed with the general theory let us look at a few special cases

(a) Suppose that $m = n = 1$ and $U \subset \mathbb{R}$ is an interval. In this case $F : U \to \mathbb{R}$ is a function of one real variable, $F = F(x)$. If F is differentiable at x_0, then the differential $dF(x_0)$ is a linear operator $\mathbb{R}^1 \to \mathbb{R}^1$ and, as such, it is described by a 1×1 matrix, i.e., a real number.

We see that F is differentiable at x_0 if and only if there exists a real number m such that

$$\lim_{h\to 0} \frac{1}{|h|}\Big(F(x_0 + h) - F(x_0) - mh \Big) = 0.$$

This happens if and only if F is differentiable at x_0 in the sense of Definition 7.1.2 and m is the derivative of F at x_0, $m = F'(x_0)$.

(b) Suppose that $m > 1$, $n = 1$ and U is an interval so that $\boldsymbol{F} : U \to \mathbb{R}^m$ is a vector valued function depending on a single real variable $x \in U \subset \mathbb{R}$

$$\boldsymbol{F}(x) = \begin{bmatrix} F^1(x) \\ \vdots \\ F^m(x) \end{bmatrix}.$$

If $\boldsymbol{F}$ is differentiable at $x_0 \in U$, then the differential of $d\boldsymbol{F}(x_0)$ is described by an $m \times 1$ matrix, i.e., a matrix consists of one column of height m. We have

$$\boldsymbol{F}(x + h) - \boldsymbol{F}(x) = \begin{bmatrix} F^1(x_0 + h) - F^1(x_0) \\ \vdots \\ F^m(x_0 + h) - F^m(x_0) \end{bmatrix}$$

We deduce that $\boldsymbol{F}$ is differentiable at x_0 if and only if the functions $F^1, \dots, F^m$ are differentiable at x_0 and

$$d\boldsymbol{F}(x_0) = \begin{bmatrix} \frac{dF^1}{dx}(x_0) \\ \vdots \\ \frac{dF^m}{dx}(x_0) \end{bmatrix}.$$

(c) If $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, then L is Fréchet differentiable at any $\boldsymbol{x}_0 \in \mathbb{R}^n$. Moreover, the differential at $\boldsymbol{x}_0$ is the operator L itself. □

Deciding when a function or a map is Fréchet differentiable at a point $\boldsymbol{x}_0$ takes a bit of work. We will describe in the following sections some simple ways of recognizing Fréchet differentiable maps.

13.2. Partial derivatives and Fréchet differentials

Suppose that $U \subset \mathbb{R}^n$ is an open set and $\boldsymbol{F} : U \to \mathbb{R}^m$. The limits in the right-hand side of (13.1.3) play a very important role in differential calculus and for this reason they were given a special name.

Definition 13.2.1. Let $\boldsymbol{x}_0 \in U$ and $\boldsymbol{v} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$. We say that $\boldsymbol{F}$ is *differentiable along the vector $\boldsymbol{v}$ at $\boldsymbol{x}_0$* if the limit

$$\boxed{\partial_{\boldsymbol{v}} \boldsymbol{F}(\boldsymbol{x}_0) = \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial \boldsymbol{v}} := \lim_{t \to 0} \frac{1}{t}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + t\boldsymbol{v}) - \boldsymbol{F}(\boldsymbol{x}_0) \Big)} \tag{13.2.1}$$

exists. This limit is called the *derivative of $\boldsymbol{F}$ along the vector $\boldsymbol{v}$ at the point $\boldsymbol{x}_0$*.

If $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ is the natural basis of $\mathbb{R}^n$, then the derivatives of $\boldsymbol{F}$ along $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ (when they exist) are called the *first-order partial derivatives* of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ and are denoted by

$$\boxed{\partial_{x^1} \boldsymbol{F}(\boldsymbol{x}_0) = \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^1} := \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial \boldsymbol{e}_1}, \dots, \partial_{x^n} \boldsymbol{F}(\boldsymbol{x}_0) = \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^n} := \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial \boldsymbol{e}_n}}.$$

We will refer to $\partial_{x^i} \boldsymbol{F}$ as the *partial derivative of the map $\boldsymbol{F}$ with respect to the variable x^i*. Often we will use the alternate notation

$$\boxed{\boldsymbol{F}'_{x^i} := \frac{\partial \boldsymbol{F}}{\partial x^i}}.$$

□

Remark 13.2.2. Suppose that $F : U \to \mathbb{R}$ is a real valued map depending on n real variables, $F = F(x^1, \dots, x^n)$. You should think of F as measuring some physical quantity at the point $\boldsymbol{x}$ such as temperature or pressure.

In this case the partial derivatives of F at $\boldsymbol{x}_0$ are real numbers. They can be computed as follows. Assume that $\boldsymbol{x}_0 = [x_0^1, \dots, x_0^n]^\top$. Then, for any $t \in \mathbb{R}$ sufficiently small we have

$$\boldsymbol{x}_0 + t\boldsymbol{e}_k = \big[\, x_0^1, \dots, x_0^{k-1}, x_0^k + t, x_0^{k+1}, \dots, x_0^n \,\big]^\top$$

and

$$\frac{F(\boldsymbol{x}_0+t\boldsymbol{e}_k)-F(\boldsymbol{x}_0)}{t}=\frac{F(x_0^1,\dots,x_0^{k-1},x_0^k+t,x_0^{k+1},\dots,\dots x_0^n)-F(x_0^1,\dots,x_0^k,\dots,x_0^n)}{t}.$$

Thus, when computing the partial derivative $\frac{\partial F}{\partial x^k}$ *we treat the variables x^i, $i\neq k$, as constants*, we regard F as a function of a single variable x^k and we derivate as such.

Equivalently, consider the function $g_k(t)=F(\boldsymbol{x}_0+t\boldsymbol{e}_k)$, $|t|$ sufficiently small. Then

$$F'_{x^k}(\boldsymbol{x}_0)=g'_k(0).$$

In other words, if we think of F as measuring say the temperature at a point $\boldsymbol{x}$, then $F'_{x^k}(\boldsymbol{x}_0)$ is the rate of change in the temperature as we travel through the point $\boldsymbol{x}_0$, at unit speed, in the direction of the k-th axis of $\mathbb{R}^n$.

More generally, for any vector $\boldsymbol{v}\neq\boldsymbol{0}$, the image of the path $\boldsymbol{\gamma}:\mathbb{R}\to\mathbb{R}^n$, $\boldsymbol{\gamma}(t)=\boldsymbol{x}_0+t\boldsymbol{v}$, is the line $\ell_{\boldsymbol{x}_0,\boldsymbol{v}}$ through $\boldsymbol{x}_0$ in the direction $\boldsymbol{v}$. Think of $\boldsymbol{\gamma}$ as describing the motion of a particle in $\mathbb{R}^n$ traveling with constant velocity $\boldsymbol{v}$. Next, think of a map $\boldsymbol{F}:U\to\mathbb{R}^m$ as associating m different physical quantities (e.g., pressure, temperature, external forces, etc.) to each point in U. These quantities can be measured by various sensors attached to the moving particle.

The derivative $\partial_{\boldsymbol{v}}\boldsymbol{F}(\boldsymbol{x}_0)$ measures the “infinitesimal rate of change” in the quantities aggregated in $\boldsymbol{F}$ as the moving particle travels through $\boldsymbol{x}_0$. As an object $\partial_{\boldsymbol{v}}\boldsymbol{F}(\boldsymbol{x}_0)$ is an m-dimensional vector. □

Proposition 13.2.3. *If $\boldsymbol{F}:U\to\mathbb{R}^m$ is Fréchet differentiable at $\boldsymbol{x}_0$, then $\boldsymbol{F}$ is differentiable along any direction $\boldsymbol{v}$ and*

$$\boxed{\partial_{\boldsymbol{v}}\boldsymbol{F}(\boldsymbol{x}_0)=d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{v}}. \tag{13.2.2}$$

In particular,

$$\boldsymbol{F}'_{x^j}(\boldsymbol{x}_0)=\frac{\partial\boldsymbol{F}}{\partial x^j}(\boldsymbol{x}_0)=\partial_{\boldsymbol{e}_j}\boldsymbol{F}(\boldsymbol{x}_0)=d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_j,\ \ \forall j=1,\dots,n,$$

and, if $\boldsymbol{v}=[v^1,\dots,v^n]^\top$, then

$$\boxed{\partial_{\boldsymbol{v}}\boldsymbol{F}(\boldsymbol{x}_0)=v^1\frac{\partial\boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^1}+\cdots+v^n\frac{\partial\boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^n}}. \tag{13.2.3}$$

Proof. The equality (13.2.2) is in fact the equality (13.1.3) in disguise. To prove (13.2.3) observe first that the equality $\boldsymbol{v}=[v^1,\dots,v^n]^\top$ signifies that

$$\boldsymbol{v}=v^1\boldsymbol{e}_1+\cdots+v^n\boldsymbol{e}_n.$$

From (13.2.2) and the linearity of $d\boldsymbol{F}(\boldsymbol{x}_0)$ we deduce

$$\begin{aligned}\partial_{\boldsymbol{v}}\boldsymbol{F}(\boldsymbol{x}_0)&=d\boldsymbol{F}(\boldsymbol{x}_0)(v^1\boldsymbol{e}_1+\cdots+v^n\boldsymbol{e}_n)=v^1d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_1+\cdots+v^nd\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_n\\&=v^1\frac{\partial\boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^1}+\cdots+v^n\frac{\partial\boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^n}.\end{aligned}$$

□

Remark 13.2.4. If $F : U \to \mathbb{R}$ is differentiable, then its differential is represented by a $1 \times n$ matrix, i.e., a matrix consisting of a single row of length n. Its entries are the real numbers

$$dF(\boldsymbol{x}_0)\boldsymbol{e}_1 = \frac{\partial F}{\partial x^1}(\boldsymbol{x}_0), \dots, dF(\boldsymbol{x}_0)\boldsymbol{e}_n = \frac{\partial F}{\partial x^n}(\boldsymbol{x}_0).$$

In other words, the differential $dF(\boldsymbol{x}_0)$ is described by the row vector

$$\boxed{dF(\boldsymbol{x}_0) = \left[\frac{\partial F}{\partial x^1}(\boldsymbol{x}_0), \dots, \frac{\partial F}{\partial x^n}(\boldsymbol{x}_0)\right].} \tag{13.2.4}$$

Viewed as a linear form $\mathbb{R}^n \to \mathbb{R}$, the differential $dF(\boldsymbol{x}_0)$ admits the alternate description

$$dF(\boldsymbol{x}_0) = \frac{\partial F}{\partial x^1}(\boldsymbol{x}_0)\boldsymbol{e}^1 + \cdots + \frac{\partial F}{\partial x^n}(\boldsymbol{x}_0)\boldsymbol{e}^n = \sum_{j=1}^{n} \frac{\partial F}{\partial x^j}(\boldsymbol{x}_0)\boldsymbol{e}^j, \tag{13.2.5}$$

where we recall that $\boldsymbol{e}^j$ denotes the linear functional $\mathbb{R}^n \to \mathbb{R}$ given by

$$\boldsymbol{e}^j(\boldsymbol{x}) = x^j.$$

In terms of row vectors we have

$$\boldsymbol{e}^1 = [1, 0, 0, \dots 0], \;\; \boldsymbol{e}^2 = [0, 1, 0, \dots, 0], \dots . \qquad \square$$

Example 13.2.5. For example, if $n = 3$,

$$x := x^1, \;\; y := x^2, \;\; z := x^3, \;\; \boldsymbol{x}_0 = (x_0, y_0, z_0)$$

and

$$F : \mathbb{R}^3 \to \mathbb{R}, \;\; F(x, y, z) = e^{3x+4y+5z},$$

then

$$\frac{\partial F}{\partial x}(\boldsymbol{x}_0) = 3e^{3x_0+4y_0+5z_0}, \;\; \frac{\partial F}{\partial y}(\boldsymbol{x}_0) = 4e^{3x_0+4y_0+5z_0}, \;\; \frac{\partial F}{\partial z}(\boldsymbol{x}_0) = 5e^{3x_0+4y_0+5z_0}.$$

If $\boldsymbol{x}_0 = \boldsymbol{0} = (0, 0, 0)$, then the differential $dF(\boldsymbol{0})$, *if it exists*,[2] *must* be the *single row matrix*

$$dF(\boldsymbol{0}) = [3, 4, 5] = 3\boldsymbol{e}^1 + 4\boldsymbol{e}^2 + 5\boldsymbol{e}^3.$$

In particular, for any vector $\boldsymbol{v} = (v^1, v^2, v^3) \in \mathbb{R}^3 \backslash \{\boldsymbol{0}\}$, we have

$$\partial_{\boldsymbol{v}} F(\boldsymbol{0}) \overset{(13.2.3)}{=} 3v^1 + 4v^2 + 5v^3. \qquad \square$$

We saw that the differentiability of a map at a point $\boldsymbol{x}_0$ guarantees the existence of derivatives at $\boldsymbol{x}_0$ in any direction. We want to investigate the extent to which a converse is true. To do this we first need to clarify a bit the concept of differentiability.

[2] We will see a bit later in Example 13.2.11 that the differential does indeed exist.

Proposition 13.2.6. *Let $m, n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set. Consider a map*

$$\boldsymbol{F} : U \to \mathbb{R}^m, \quad \boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} F^1(\boldsymbol{x}) \\ \vdots \\ F^m(\boldsymbol{x}) \end{bmatrix}, \quad \boldsymbol{x} \in U.$$

Then the following statements are equivalent.

(i) *The map $\boldsymbol{F}$ is Fréchet differentiable at $\boldsymbol{x}_0 \in U$.*

(ii) *Each of the scalar valued functions $F^1, \dots, F^m : U \to \mathbb{R}$ is Fréchet differentiable at $\boldsymbol{x}_0$.*

Proof. (i) ⇒ (ii) Suppose that $\boldsymbol{F}$ is differentiable at $\boldsymbol{x}_0$. We denote by L its differential. We identify L with an $m \times n$ matrix. For $i = 1, \dots, m$ we denote by L^i the i-th *row* of L and we view L^i as a linear map $L^i : \mathbb{R}^n \to \mathbb{R}$. We will show that L^i is the differential of F^i at $\boldsymbol{x}_0$. For $\boldsymbol{h} \in \mathbb{R}^n$ sufficiently small we have

$$\frac{1}{\|\boldsymbol{h}\|}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}\Big) = \frac{1}{\|\boldsymbol{h}\|}\begin{bmatrix} F^1(\boldsymbol{x}_0 + \boldsymbol{h}) - F^1(\boldsymbol{x}_0) - L^1\boldsymbol{h} \\ \vdots \\ F^m(\boldsymbol{x}_0 + \boldsymbol{h}) - F^m(\boldsymbol{x}_0) - L^m\boldsymbol{h} \end{bmatrix}. \tag{13.2.6}$$

We deduce

$$\begin{gathered} \lim_{\boldsymbol{h}\to 0} \frac{1}{\|\boldsymbol{h}\|}\Big(\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}\Big) = \boldsymbol{0} \\ \Longleftrightarrow \lim_{\boldsymbol{h}\to 0} \frac{1}{\|\boldsymbol{h}\|}\Big(F^i(\boldsymbol{x}_0 + \boldsymbol{h}) - F^i(\boldsymbol{x}_0) - L^i\boldsymbol{h}\Big) = \boldsymbol{0}, \quad \forall i = 1, \dots, m. \end{gathered} \tag{13.2.7}$$

The top line of this equivalence states the differentiability of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ and the bottom line of this equivalence amounts to the differentiability at $\boldsymbol{x}_0$ of each of the components $F^1, \dots, F^m$.

(ii) ⇒ (i) Suppose that each of the functions F^i is differentiable at $\boldsymbol{x}_0$. We denote by L^i the differential of F^i at $\boldsymbol{x}_0$. This is a linear map $\mathbb{R}^n \to \mathbb{R}$ which we identify with a row of length n. Denote by L the $m \times n$ matrix with i-th row is L^i, $\forall i = 1, \dots, m$.

The matrix L satisfies the equality (13.2.6) and the equivalence (13.2.7) holds as well. This proves (i).

□

Example 13.2.7. Suppose that $m, n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$ is an open set. If the map is differentiable at $\boldsymbol{x}_0 \in U$, then the differential $d\boldsymbol{F}(\boldsymbol{x}_0)$ is represented by the $m \times n$ Jacobian matrix $J_{\boldsymbol{F}}(\boldsymbol{x}_0)$ with columns

$$d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_1 = \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^1}, \dots, d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{e}_n = \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^n}.$$

Let $F^1, \dots, F^m$ be the components of $\boldsymbol{F}$, so that

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} F^1(\boldsymbol{x}) \\ \vdots \\ F^m(\boldsymbol{x}) \end{bmatrix},$$

Each component F^j, viewed as a function $F^j : U \to \mathbb{R}$ is differentiable at $\boldsymbol{x}_0$. For any $j = 1, \dots, n$ we have

$$\frac{\partial \boldsymbol{F}(\boldsymbol{x}_0)}{\partial x^j} = \lim_{t \to 0} \frac{1}{t}\big(\boldsymbol{F}(\boldsymbol{x}_0 + t\boldsymbol{e}_j) - \boldsymbol{F}(\boldsymbol{x}_0)\big)$$

$$= \lim_{t \to 0} \frac{1}{t} \begin{bmatrix} F^1(\boldsymbol{x}_0 + t\boldsymbol{e}_j) - F^1(\boldsymbol{x}_0) \\ \vdots \\ F^m(\boldsymbol{x}_0 + t\boldsymbol{e}_j) - F^m(\boldsymbol{x}_0) \end{bmatrix} = \begin{bmatrix} \frac{\partial F^1(\boldsymbol{x}_0)}{\partial x^j} \\ \vdots \\ \frac{\partial F^m(\boldsymbol{x}_0)}{\partial x^j} \end{bmatrix}.$$

Hence, the Jacobian matrix of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ is

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}(\boldsymbol{x}_0) = J_{\boldsymbol{F}}(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial F^1(\boldsymbol{x}_0)}{\partial x^1} & \cdots & \frac{\partial F^1(\boldsymbol{x}_0)}{\partial x^n} \\ \vdots & \vdots & \vdots \\ \frac{\partial F^m(\boldsymbol{x}_0)}{\partial x^1} & \cdots & \frac{\partial F^m(\boldsymbol{x}_0)}{\partial x^n} \end{bmatrix}.$$

Using the equality (13.2.4) we see that the first row of $J_{\boldsymbol{F}}(\boldsymbol{x}_0)$ is the differential of F^1 at $\boldsymbol{x}_0$, the second row of $J_{\boldsymbol{F}}(\boldsymbol{x}_0)$ is the differential of F^2 at $\boldsymbol{x}_0$ etc. Thus we can describe the Jacobian $J_{\boldsymbol{F}}$ in the simplified form

$$J_{\boldsymbol{F}} = \begin{bmatrix} dF^1 \\ dF^2 \\ \vdots \\ dF^m \end{bmatrix}. \qquad \square$$

Proposition 13.2.3 shows that the maps $\boldsymbol{F} : U \to \mathbb{R}^m$ that are differentiable at a point $\boldsymbol{x}_0$ have a special property: they admit partial derivatives at $\boldsymbol{x}_0$. However, the existence of partial derivatives at $\boldsymbol{x}_0$ is not enough to guarantee the Fréchet differentiability at $\boldsymbol{x}_0$. The next result describes one very simple and useful condition guaranteeing Fréchet differentiability.

Theorem 13.2.8. *Let $m, n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$ open set. Suppose $F : U \to \mathbb{R}^m$ is a map and $\boldsymbol{x}_0$ is a point in U satisfying the following conditions.*

(i) *There exists $r > 0$ such that $B_r(\boldsymbol{x}_0) \subset U$ and the map F admits partial derivatives at any point $\boldsymbol{x} \in B_r(\boldsymbol{x}_0)$.*

(ii) *For any* $j = 1, \dots, n$

$$\frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} = \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{\partial F(\boldsymbol{x})}{\partial x^j}.$$

Then the map F is Fréchet differentiable at $\boldsymbol{x}_0$.

Proof. According to Proposition 13.2.6 it suffices to consider only the case $m = 1$, i.e., F is a real valued function, $F : U \to \mathbb{R}$. Denote by L the linear map

$$L : \mathbb{R}^n \to \mathbb{R}, \quad L\boldsymbol{h} = \sum_{j=1}^{n} \frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} h^j.$$

We want to prove that L is the Fréchet differential of F at $\boldsymbol{x}_0$, i.e.,

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{1}{\|\boldsymbol{h}\|} \Big| F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) - L\boldsymbol{h} \Big| = 0. \tag{13.2.8}$$

Given $\boldsymbol{h} = h^1\boldsymbol{e}_1 + \cdots + h^n\boldsymbol{e}^n$, $\|\boldsymbol{h}\| < \frac{r}{2}$, we set (see Figure 13.3)

$$\boldsymbol{h}_1 := h^1\boldsymbol{e}_1, \quad \boldsymbol{h}_2 = h^1\boldsymbol{e}_1 + h^2\boldsymbol{e}_2, \quad \boldsymbol{h}_j := h^1\boldsymbol{e}_1 + \cdots + h^j\boldsymbol{e}_j, \dots, j = 1, \dots, n,$$

$$\boldsymbol{x}_j = \boldsymbol{x}_0 + \boldsymbol{h}_j, \quad j = 1, \dots, n.$$

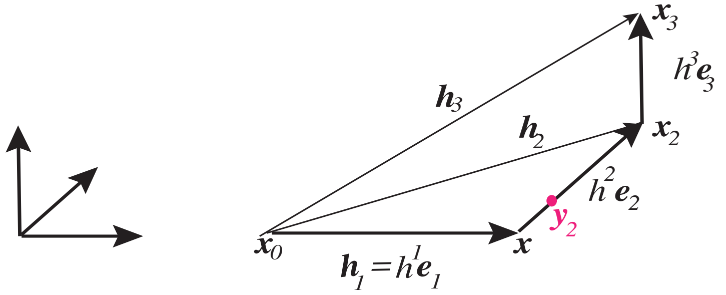

Figure 13.3. *Zig-zagging from $\boldsymbol{x}_0$ to $\boldsymbol{x}_n = \boldsymbol{x} + \boldsymbol{h}$, $n = 3$.*

We have

$$F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) = F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1}) + F(\boldsymbol{x}_{n-1}) - F(\boldsymbol{x}_{n-2}) + \cdots + F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0).$$

For each $j = 1, \dots, n$ define[3]

$$g_j : (-r/2, r/2) \to \mathbb{R}, \quad g_j(t) = F(\boldsymbol{x}_{j-1} + t\boldsymbol{e}_j).$$

Note that,

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + h^j\boldsymbol{e}_j, \quad F(\boldsymbol{x}_{j-1}) = g_j(0), \quad F(\boldsymbol{x}_j) = g_j(h^j)$$

Since F admits partial derivatives at every $\boldsymbol{x} \in B_r(\boldsymbol{x}_0)$ we deduce that the function g_j is differentiable and

$$g_j'(t) = \frac{\partial F(\boldsymbol{x}_{j-1} + t\boldsymbol{e}_j)}{\partial x^j}. \tag{13.2.9}$$

The Lagrange mean value theorem implies that there exists τ_j in the interval $[0, h^j]$ such that

$$F(\boldsymbol{x}_j) - F(\boldsymbol{x}_{j-1}) = g_j(h^j) - g_j(0) = g_j'(\tau_j)h^j.$$

[3] Observe that $\boldsymbol{x}_{j-1} + t\boldsymbol{e}_j \in U$, $\forall |t| < r/2$.

We set $\boldsymbol{y}_j = \boldsymbol{y}_j(\boldsymbol{h}) = \boldsymbol{x}_{j-1} + \tau_j \boldsymbol{e}_k$. Note that $\boldsymbol{y}_j$ is situated on the line segment connecting $\boldsymbol{x}_{j-1}$ to $\boldsymbol{x}_j$. From (13.2.9) we deduce

$$F(\boldsymbol{x}_j) - F(\boldsymbol{x}_{j-1}) = \frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} h^j.$$

Let us observe that

$$\|\boldsymbol{h}_j\| \leqslant \|\boldsymbol{h}\|, \;\; \forall j = 1, \dots, n$$

proving that

$$\operatorname{dist}(\boldsymbol{x}_j, \boldsymbol{x}_0) \leqslant \|\boldsymbol{h}\|, \;\; \forall j = 1, \dots, n.$$

Thus all the points $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n = \boldsymbol{x}_0 + \boldsymbol{h}$ lie in $\overline{B_{\|\boldsymbol{h}\|}(\boldsymbol{x}_0)}$, the closed Euclidean ball of center $\boldsymbol{x}_0$ and radius $\|\boldsymbol{h}\|$. This is a convex subset, and since $\boldsymbol{y}_j$ is situated on the line segment $[\boldsymbol{x}_{j-1}, \boldsymbol{x}_j]$, it is also contained $\overline{B_{\|\boldsymbol{h}\|}(\boldsymbol{x}_0)}$. Hence

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} y_j(\boldsymbol{h}) = \boldsymbol{x}_0, \;\; \forall j = 1, \dots, n. \tag{13.2.10}$$

We can now put together all the facts above. We have

$$F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) = \sum_{j=1}^{n} \frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} h^j$$

$$F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) - L\boldsymbol{h} = \sum_{j=1}^{n} \left(\frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} - \frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} \right) h^j,$$

so that

$$\left| F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) - L\boldsymbol{h} \right| \leqslant \sum_{j=1}^{n} \left| \frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} - \frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} \right| \cdot |h^j|$$

(use the Cauchy-Schwarz inequality)

$$\leqslant \sqrt{\left| \frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} - \frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} \right|^2} \cdot \|\boldsymbol{h}\|.$$

Hence

$$\frac{1}{\|\boldsymbol{h}\|} \left| F(\boldsymbol{x}_0 + \boldsymbol{h}) - F(\boldsymbol{x}_0) - L\boldsymbol{h} \right| \leqslant \sqrt{\left| \frac{\partial F(\boldsymbol{y}_j)}{\partial x^k} - \frac{\partial F(\boldsymbol{x}_0)}{\partial x^j} \right|^2}.$$

If we let $\boldsymbol{h} \to 0$, and take (13.2.10) into account, we obtain the desired conclusion, (13.2.6). □

Definition 13.2.9. Let $m, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ an open set, and $\boldsymbol{F} : U \to \mathbb{R}^m$ a map.

(i) We say that the map $\boldsymbol{F} : U \to \mathbb{R}^m$ is *Fréchet differentiable on* U if it is Fréchet differentiable at every point $\boldsymbol{x} \in U$.

(ii) We say that $\boldsymbol{F}$ is *continuously differentiable*, or C^1, on U, if it admits first order partial derivatives at any $\boldsymbol{x} \in U$ and, for any $j = 1, \dots, n$, the function

$$U \ni \boldsymbol{x} \mapsto \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^j} \in \mathbb{R}^m$$

is continuous.

□

✍ *We will denote by $C^1(U, \mathbb{R}^m)$ the set of C^1-maps $\boldsymbol{F} : U \to \mathbb{R}^m$. For simplicity we will write $C^1(U)$ instead of $C^1(U, \mathbb{R})$.*

From Theorem 13.2.8 we obtain the following very useful result.

Corollary 13.2.10. *Let $m, n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$. If the map $\boldsymbol{F} : U \to \mathbb{R}^m$ is C^1 on U, then it is Fréchet differentiable on U. Moreover, if $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ is the canonical basis of $\mathbb{R}^n$, then*

$$d\boldsymbol{F}(\boldsymbol{x})\boldsymbol{e}_j = \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial x^j}, \;\; \forall \boldsymbol{x} \in U, \;\; \forall j = 1, \dots, n. \qquad \square$$

Example 13.2.11. (a) Consider a linear functional

$$\boldsymbol{\xi} : \mathbb{R}^n \to \mathbb{R}, \;\; \boldsymbol{\xi}(\boldsymbol{x}) = \sum_{j=1}^n \xi_j x^j.$$

We deduce that, for any $\boldsymbol{x} \in \mathbb{R}^n$, and any $j = 1, \dots, n$,

$$\frac{\partial \xi(\boldsymbol{x})}{\partial x^j} = \xi_j.$$

Thus the functions

$$\mathbb{R}^n \ni \boldsymbol{x} \mapsto \frac{\partial \xi(\boldsymbol{x})}{\partial x^j} \in \mathbb{R}$$

are constant and, in particular, continuous. Corollary 13.2.10 implies that the linear function $\boldsymbol{\xi}$ is differentiable on $\mathbb{R}^n$, and its differential at a point $\boldsymbol{x}$ is represented by the row vector

$$[\xi_1, \dots, \xi_n].$$

This is the same row vector that represents $\boldsymbol{\xi}$. Thus we have the equality of linear functions

$$d\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\xi}, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{13.2.11}$$

At this point it is worth mentioning a classical convention that we will use frequently in the sequel.

Note that for any $j = 1, \dots, n$, the linear functional $\boldsymbol{e}^j$ associates to the vector $\boldsymbol{x}$ its j-th coordinate x^j. We can rephrase this by saying that $\boldsymbol{e}^j$ is the function x^j, i.e., $\boldsymbol{e}^j(\boldsymbol{x}) = x^j$. We write this in the less precise fashion $\boldsymbol{e}^j = x^j$. The equality (13.2.11) applied to the linear functional $\boldsymbol{e}^j$ yields the classical convention

$$\boxed{dx^j = d\boldsymbol{e}^j = \boldsymbol{e}^j}. \tag{13.2.12}$$

If now $f : \mathbb{R}^n \to \mathbb{R}$ is a C^1-function, then it is differentiable everywhere and, according to (13.2.5), its differential at $\boldsymbol{x} \in \mathbb{R}^n$ is the linear functional

$$df(\boldsymbol{x}) = \sum_{j=1}^n \frac{\partial f(\boldsymbol{x})}{\partial x^j}\boldsymbol{e}^j = \frac{\partial f(\boldsymbol{x})}{\partial x^1}\boldsymbol{e}^1 + \cdots + \frac{\partial f(\boldsymbol{x})}{\partial x^n}\boldsymbol{e}^n.$$

Using the convention (13.2.12) we obtain another frequently used convention/notation

$$\boxed{df = \sum_{j=1}^n \frac{\partial f}{\partial x^j}dx^j = \frac{\partial f}{\partial x^1}dx^1 + \cdots + \frac{\partial f}{\partial x^n}dx^n.} \tag{13.2.13}$$

The right-hand side of the above equality is classically referred to as the *total differential* of the function f. Moreover, in the above equality we interpret both sides as *functions* on $\mathbb{R}^n$ with values in the space $\operatorname{Hom}(\mathbb{R}^n, \mathbb{R})$ of linear functionals on $\mathbb{R}^n$.

For example, if $n = 1$, so f is a function of a single real variable, then the above equality takes the known form (7.1.5)

$$df = \frac{df}{dx}dx = f'(x)dx.$$

(b) Consider the function $r : \mathbb{R}^2 \backslash \{\mathbf{0}\} \to \mathbb{R}$, $r(x, y) = \sqrt{x^2 + y^2}$. For fixed y the function $x \mapsto \sqrt{x^2 + y^2}$ is differentiable as long as $x^2 + y^2 \neq 0$. Its derivative is

$$\frac{\partial r}{\partial x} = \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r}.$$

A similar argument shows that $\frac{\partial r}{\partial y}$ exists as long as $x^2 + y^2 \neq 0$ and we have

$$\frac{\partial r}{\partial y} = \frac{y}{\sqrt{x^2 + y^2}} = \frac{y}{r}.$$

Thus the function r is differentiable at every point in $\mathbb{R}^2 \backslash \{\mathbf{0}\}$ and

$$dr = \frac{x}{\sqrt{x^2 + y^2}}dx + \frac{y}{\sqrt{x^2 + y^2}}dy.$$

The associated Jacobian matrix is the single row matrix

$$J_r = \left[\begin{array}{cc} \frac{x}{\sqrt{x^2 + y^2}} & \frac{y}{\sqrt{x^2 + y^2}} \end{array} \right].$$

The differential of r at the point$(x_0, y_0) = (3, 4)$ is therefore represented by the row vector

$$\left[\frac{3}{5}, \frac{4}{5} \right].$$

(c) Consider again the function

$$F : \mathbb{R}^3 \to \mathbb{R}, \;\; F(x, y, z) = e^{3x+4y+5z}$$

we discussed in Remark 13.2.4(b). The function F admits partial derivatives

$$\frac{\partial F}{\partial x} = 3e^{3x+4y+5z}, \;\; \frac{\partial F}{\partial y} = 4e^{3x+4y+5z}, \;\; \frac{\partial F}{\partial z} = 5e^{3x+4y+5z}$$

which are continuous functions. Thus $F \in C^1(\mathbb{R}^3)$ and, in particular, it is Fréchet differentiable on $\mathbb{R}^3$. Moreover

$$dF = 3e^{3x+4y+5z}\,dx + 4e^{3x+4y+5z}\,dy + 5e^{3x+4y+5z}\,dz.$$

Again, we interpret both sides of the above equality as functions $\mathbb{R}^3 \to \operatorname{Hom}(\mathbb{R}^3, \mathbb{R})$.

(d) Consider the map $\boldsymbol{F} : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\mathbb{R}^2 \ni \begin{bmatrix} r \\ \theta \end{bmatrix} \overset{\boldsymbol{F}}{\mapsto} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} \in \mathbb{R}^2.$$

You should read the above as follows: the components of the map $\boldsymbol{F}$ are two functions called x and y depending on two variables (r, θ) and

$$x(r, \theta) = r\cos\theta, \;\; y(r, \theta) = r\sin\theta.$$

Clearly the functions x, y are C^1 on their domains and the Jacobian matrix of $\boldsymbol{F}$ is

$$J_{\boldsymbol{F}} = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ & \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

In particular for $(r, \theta) = (1, \pi/2)$ we have

$$J_{\boldsymbol{F}}(1, \pi/2) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

If $\boldsymbol{v} = [3, 4]^\top$, then

$$\partial_{\boldsymbol{v}}\boldsymbol{F}(1, \pi/2) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}.$$

□

Example 13.2.12 (Linearizations)**.** Suppose that $n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is a C^1-function. Then, according to Definition 13.1.5, the linearization (or linear approximation) of f at $\boldsymbol{x}_0$ is the affine function

$$\mathcal{L} : \mathbb{R}^n \to \mathbb{R}, \;\; \mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{x}_0) + df(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0).$$

To see how this looks concretely, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$. We have

$$\partial_x f = 2x, \;\; \partial_y f = 2y.$$

Let us find the linear approximation of this function at the point $(x_0, y_0) = (2, 1)$. We have

$$f(x_0, y_0) = 2^2 + 1^2 = 5, \;\; \partial_x f(x_0, y_0) = 4, \;\; \partial_y f(x_0, y_0) = 2.$$

The differential $df(x_0, y_0)$ is thus described by the row vector $[4, 2]$. The linearization of f at $(2, 1)$ is the affine function

$$\begin{aligned} \mathcal{L}(x, y) &= f(x_0, y_0) + \partial_x f(x_0, y_0)(x - x_0) + \partial_y f(x_0, y_0)(y - y_0) \\ &= 5 + 4(x - 2) + 2(y - 1) = 4x + 2y - 5. \end{aligned}$$

The surface in Figure 13.1 is the graph of f, while the plane in the same figure is the graph of $\mathcal{L}$. □

Definition 13.2.13 (Gradient)**.** Let $n \in \mathbb{N}$. Suppose that $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is a function differentiable at $\boldsymbol{x}_0$. The *gradient* of f at $\boldsymbol{x}_0$ is the *vector* $df(\boldsymbol{x}_0)_\uparrow$ dual to the differential of f at $\boldsymbol{x}_0$. We denote by $\nabla f(\boldsymbol{x}_0)$ the gradient of f at $\boldsymbol{x}_0$. The symbol ∇ is pronounced *nabla*.[4] □

The above definition is rather dense. Let us unpack it. The differential $df(\boldsymbol{x}_0)$ of f at $\boldsymbol{x}_0$ is a linear form $\mathbb{R}^n \to \mathbb{R}$ (or covector) represented by the single *row* matrix

$$\left[\frac{\partial f(\boldsymbol{x}_0)}{\partial x^1}, \ldots, \frac{\partial f(\boldsymbol{x}_0)}{\partial x^n} \right].$$

[4] The name nabla originates from an ancient stringed musical instrument shaped as a harp.

As explained in (11.2.4a) the dual of this covector is the *column* vector

$$\begin{bmatrix} \frac{\partial f(\boldsymbol{x}_0)}{\partial x^1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x}_0)}{\partial x^n} \end{bmatrix} =: \nabla f(\boldsymbol{x}_0).$$

Using (13.2.3) we deduce that, for any $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ we have

$$\partial_{\boldsymbol{v}} f(\boldsymbol{x}_0) = \partial_{x^1} f(\boldsymbol{x}_0) v^1 + \cdots + \partial_{x^n} f(\boldsymbol{x}_0) v^n,$$

i.e.,

$$\boxed{\partial_{\boldsymbol{v}} f(\boldsymbol{x}_0) = df(\boldsymbol{x}_0)\boldsymbol{v} = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{v} \rangle, \quad \forall \boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}}. \tag{13.2.14}$$

The construction of the gradient might appear to the uninitiated as "much ado about nothing" because all we have done was take a row and then transform it into a column. Temporarily it is difficult to justify this algebraic contortion. For now, please take it as an article of faith that there is a method to this "madness."

Example 13.2.14. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$. Then

$$df(x, y) = 2xdx + 2ydy, \quad \nabla f(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}.$$

The correspondence $\mathbb{R}^2 \ni (x, y) \mapsto \nabla f(x, y) \in \mathbb{R}^2$ is often viewed as a vector field on $\mathbb{R}^2$ in that it assigns an "arrow" (or vector) to each point of $\mathbb{R}^2$. □

Example 13.2.15. We define a *direction* in $\mathbb{R}^n$ to be a unit length vector $\boldsymbol{n}$, $\|\boldsymbol{n}\| = 1$. Observe that any nonzero vector $\boldsymbol{v} \in \mathbb{R}^n$ determines a direction

$$\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{v}) = \frac{1}{\|\boldsymbol{v}\|}\boldsymbol{v}.$$

A point $\boldsymbol{x}_0 \in \mathbb{R}^n$ and a direction $\boldsymbol{n}$ canonically determine a path

$$\boldsymbol{\gamma}_{\boldsymbol{x}_0, \boldsymbol{n}} : \mathbb{R} \to \mathbb{R}^n, \quad \boldsymbol{\gamma}_{\boldsymbol{x}_0, \boldsymbol{n}}(t) = \boldsymbol{x}_0 + t\boldsymbol{n}$$

whose image is the line through $\boldsymbol{x}_0$ in the direction $\boldsymbol{n}$.

Given an open set $U \subset \mathbb{R}^n$, a C^1-function $f : U \to \mathbb{R}$, a point $\boldsymbol{x}_0$ and a direction $\boldsymbol{n}$, we define the derivative of f in the direction $\boldsymbol{n}$ at $\boldsymbol{x}_0$ to be the derivative of f along the vector $\boldsymbol{n}$. From (13.2.14) we deduce that

$$\partial_{\boldsymbol{n}} f(\boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{n} \rangle.$$

Suppose that $\nabla f(\boldsymbol{x}_0) \neq \boldsymbol{0}$ and let $\theta \in [0, \pi]$ be the angle between the vectors $\nabla f(\boldsymbol{x}_0)$ and $\boldsymbol{n}$. We have

$$\partial_{\boldsymbol{n}} f(\boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{n} \rangle = \|\nabla f(\boldsymbol{x}_0)\| \cos\theta \leqslant \|\nabla f(\boldsymbol{x}_0)\|.$$

Above, we have equality if and only if $\theta = 0$. Thus $\partial_{\boldsymbol{n}} f(\boldsymbol{x}_0)$ takes its highest possible value if and only if $\boldsymbol{n}$ points in the same direction as $\nabla f(\boldsymbol{x}_0)$ or, equivalently, $\boldsymbol{n}$ is the direction determined by the gradient vector $\nabla f(\boldsymbol{x}_0)$. This shows that *the direction determined by* *the gradient of a function at a point is the direction of fastest growth* of the function at that given point. □

13.3. The chain rule

We can now state and prove a key result in several variable calculus.

Theorem 13.3.1 (Chain rule). *Let $\ell, m, n \in \mathbb{N}$. Suppose that we are given open sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, maps $\boldsymbol{F} : U \to \mathbb{R}^m$, $\boldsymbol{G} : V \to \mathbb{R}^\ell$, and a point $\boldsymbol{u}_0 \in U$ satisfying the following conditions.*

(i) $\boldsymbol{F}(U) \subset V$.

(ii) *$\boldsymbol{F}$ is differentiable at $\boldsymbol{u}_0$ and $\boldsymbol{G}$ is differentiable at $\boldsymbol{v}_0 := \boldsymbol{F}(\boldsymbol{u}_0)$.*

Then the composition $\boldsymbol{G} \circ \boldsymbol{F} : U \to \mathbb{R}^\ell$ is differentiable at $\boldsymbol{u}_0$ and

$$d(\boldsymbol{G} \circ \boldsymbol{F})(\boldsymbol{u}_0) = d\boldsymbol{G}(\boldsymbol{v}_0) \circ d\boldsymbol{F}(\boldsymbol{u}_0). \tag{13.3.1}$$

Idea of proof. Set $A := d\boldsymbol{F}(\boldsymbol{u}_0)$, $B := d\boldsymbol{G}(\boldsymbol{v}_0)$ so that $A \in \operatorname{Hom}(\mathbb{R}^n, \mathbb{R}^m)$, $B \in \operatorname{Hom}(\mathbb{R}^m, \mathbb{R}^\ell)$. From the definition of Fréchet differential we deduce

$$\boldsymbol{G}\big(\,\boldsymbol{F}(\boldsymbol{u})\,\big) - \boldsymbol{G}\big(\,\boldsymbol{F}(\boldsymbol{u}_0)\,\big) \approx B\big(\,\boldsymbol{F}(\boldsymbol{u}) - \boldsymbol{F}(\boldsymbol{u}_0)\,\big),$$
$$\boldsymbol{F}(\boldsymbol{u}) - \boldsymbol{F}(\boldsymbol{u}_0) \approx A(\boldsymbol{u} - \boldsymbol{u}_0).$$

Hence

$$\boldsymbol{G}\big(\,\boldsymbol{F}(\boldsymbol{u})\,\big) - \boldsymbol{G}\big(\,\boldsymbol{F}(\boldsymbol{u}_0)\,\big) \approx B \circ A\big(\,\boldsymbol{u} - \boldsymbol{u}_0\,\big).$$

This shows that $B \circ A$ is the Fréchet differential of $\boldsymbol{G} \circ \boldsymbol{F}$ at $\boldsymbol{u}_0$.

□

The above argument is an almost complete proof capturing the essence of the main idea. We present the missing details below.

Proof. We set $A := d\boldsymbol{F}(\boldsymbol{u}_0)$ and $B := d\boldsymbol{G}(\boldsymbol{v}_0)$. We have to prove that

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{1}{\|\boldsymbol{h}\|} \big\|\, \boldsymbol{G}(\,\boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h})\,) - \boldsymbol{G}(\boldsymbol{v}_0) - B(A\boldsymbol{h})\,\big\| = 0. \tag{13.3.2}$$

We set

$$T\boldsymbol{h} := \boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{u}_0) = \boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h}) - \boldsymbol{v}_0$$

and we deduce

$$\begin{aligned}\boldsymbol{G}(\,\boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h})\,) - \boldsymbol{G}(\boldsymbol{v}_0) - B(A\boldsymbol{h}) &= \boldsymbol{G}(\boldsymbol{v}_0 + T\boldsymbol{h}) - G(\boldsymbol{v}_0) - B(A\boldsymbol{h}) \\ &= \boldsymbol{G}(\boldsymbol{v}_0 + T\boldsymbol{h}) - G(\boldsymbol{v}_0) - B(T\boldsymbol{h}) + B(T\boldsymbol{h} - A\boldsymbol{h}).\end{aligned}$$

Set

$$R_F(\boldsymbol{h}) := \boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{u}_0) - A\boldsymbol{h}, \;\; R_G(\boldsymbol{k}) := \boldsymbol{G}(\boldsymbol{v}_0 + \boldsymbol{k}) - \boldsymbol{G}(\boldsymbol{v}_0) - B\boldsymbol{k}.$$

Since $\boldsymbol{F}$ is differentiable at $\boldsymbol{u}_0$ and $\boldsymbol{G}$ is differentiable at $\boldsymbol{v}_0$ we deduce from (13.1.5) that there exist $r > 0$ and functions $\varphi_F, \varphi_G : [0, r) \to \mathbb{R}$ such that

$$0 = \varphi_F(0) = \lim_{t \searrow 0} \varphi_F(t), \;\; 0 = \varphi_G(0) = \lim_{t \searrow 0} \varphi_G(t), \tag{13.3.3a}$$

$$\|R_F(\boldsymbol{h})\| \leqslant \varphi_F(\|\boldsymbol{h}\|)\|\boldsymbol{h}\|,\ \ \|R_G(\boldsymbol{k})\| \leqslant \varphi_G(\|\boldsymbol{k}\|)\|\boldsymbol{k}\|,\ \ \forall \|\boldsymbol{h}\|,\ \|\boldsymbol{k}\| < r. \tag{13.3.3b}$$

Note that

$$T\boldsymbol{h} - A\boldsymbol{h} = \boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{u}_0) - A\boldsymbol{h} = R_F(\boldsymbol{h}),$$
$$\boldsymbol{G}(\boldsymbol{v}_0 + T\boldsymbol{h}) - G(\boldsymbol{v}_0) - B(T\boldsymbol{h}) = R_G(T\boldsymbol{h}),$$

and

$$\begin{aligned}\boldsymbol{G}(\,\boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h})\,) - \boldsymbol{G}(\boldsymbol{v}_0) - B(A\boldsymbol{h}) &= R_G(T\boldsymbol{h}) + B(\,R_F(\boldsymbol{h})\,)\\ &= R_G(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,) + B(\,R_F(\boldsymbol{h})\,).\end{aligned}$$

Hence

$$\begin{aligned}\|\,\boldsymbol{G}(\,\boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h})\,) - \boldsymbol{G}(\boldsymbol{v}_0) - B(A\boldsymbol{h})\,\| &\leqslant \|\,R_G(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,)\,\| + \|B(\,R_F(\boldsymbol{h})\,)\|\\ &\leqslant \|\,R_G(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,)\,\| + \|\,B\,\|_{HS} \cdot \|\,R_F(\boldsymbol{h})\,)\,\|\\ &\overset{(13.3.3b)}{\leqslant} \varphi_G(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,)\|A\boldsymbol{h} + R_F(\boldsymbol{h})\| + \varphi_F(\|\boldsymbol{h}\|)\|B\|_{HS} \cdot \|\boldsymbol{h}\|\\ &\overset{(13.3.3b)}{\leqslant} \varphi_G(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,)\Big(\|A\|_{HS} + \varphi_F(\|\boldsymbol{h}\|)\,\Big)\|\boldsymbol{h}\| + \varphi_F(\|\boldsymbol{h}\|)\|B\|_{HS} \cdot \|\boldsymbol{h}\|,\end{aligned}$$

and thus

$$\frac{1}{\|\boldsymbol{h}\|}\|\,\boldsymbol{G}(\boldsymbol{F}(\boldsymbol{u}_0 + \boldsymbol{h})) - \boldsymbol{G}(\boldsymbol{v}_0) - B(A\boldsymbol{h})\,\| \leqslant \varphi_G\big(\,A\boldsymbol{h} + R_F(\boldsymbol{h})\,\big)\Big(\|A\|_{HS} + \varphi_F(\|\boldsymbol{h}\|)\Big)$$
$$+\varphi_F(\|\boldsymbol{h}\|)\|B\|_{HS}.$$

The conclusion (13.3.2) is obtained by letting $\boldsymbol{h} \to \boldsymbol{0}$ in the above inequality and invoking (13.3.3a). □

Let us rewrite the chain rule (13.3.1) in a less precise, but more intuitive manner.

We denote by $(u^i)_{1\leqslant i\leqslant n}$ the Euclidean coordinates on $\mathbb{R}^n$, by $(v^j)_{1\leqslant j\leqslant m}$ the Euclidean coordinates on $\mathbb{R}^m$ and by $(x^k)_{1\leqslant k\leqslant \ell}$ the Euclidean coordinates in $\mathbb{R}^\ell$. The map $\boldsymbol{F}$ is described by m functions depending on the variables (u^i)

$$v^j = F^j(u^1, \dots, u^n),\ \ j = 1, \dots, m,$$

while the map $\boldsymbol{G}$ is described by ℓ functions depending on the variables (v^j)

$$x^k = G^k(v^1, \dots, v^m),\ \ k = 1, \dots, \ell.$$

The differential of $\boldsymbol{F}$ at $\boldsymbol{u}_0$ is described by the $m \times n$ Jacobian matrix $J_{\boldsymbol{F}}$ with entries

$$(J_{\boldsymbol{F}})^j_i = \frac{\partial F^j}{\partial u^i} = \frac{\partial v^j}{\partial u^i}.$$

The differential of $\boldsymbol{G}$ at $\boldsymbol{v}_0 = \boldsymbol{F}(\boldsymbol{u}_0)$ is described by the $\ell \times m$ Jacobian matrix $J_{\boldsymbol{G}}$ with entries

$$(J_{\boldsymbol{G}})^k_j = \frac{\partial G^k}{\partial v^j} = \frac{\partial x^k}{\partial v^j}.$$

The composition $\boldsymbol{G} \circ \boldsymbol{F}$ is described by ℓ functions depending on the variables (u^i)

$$x^k = G^k\big(\,F^1(u^1, \dots, u^n), \dots, F^m(u^1, \dots, u^n)\,\big),\ \ k = 1, \dots, \ell.$$

The differential of $\boldsymbol{G} \circ \boldsymbol{F}$ at $\boldsymbol{u}_0$ is described by the $\ell \times n$ matrix $J_{\boldsymbol{G}\circ\boldsymbol{F}}$ with entries

$$(J_{\boldsymbol{G}\circ\boldsymbol{F}})^k_i = \frac{\partial x^k}{\partial u^i}.$$

The chain rule (13.3.1) states that

$$J_{\boldsymbol{G}\circ\boldsymbol{F}}(\boldsymbol{u}_0)=J_{\boldsymbol{G}}(\boldsymbol{v}_0)J_{\boldsymbol{F}}(\boldsymbol{u}_0)=J_{\boldsymbol{G}}\big(\boldsymbol{F}(\boldsymbol{u}_0)\big)J_{\boldsymbol{F}}(\boldsymbol{u}_0), \tag{13.3.4}$$

or, equivalently,

$$\boxed{\frac{\partial x^k}{\partial u^i}=\sum_{j=1}^m\frac{\partial x^k}{\partial v^j}\cdot\frac{\partial v^j}{\partial u^i}=\frac{\partial x^k}{\partial v^1}\cdot\frac{\partial v^1}{\partial u^i}+\cdots+\frac{\partial x^k}{\partial v^m}\cdot\frac{\partial v^m}{\partial u^i}}. \tag{13.3.5}$$

Example 13.3.2. Consider the function

$$f:\mathbb{R}^2\to\mathbb{R},\;\; f(x,y)=(x^2+y^2+1)^{\sin(xy)}.$$

This is the composition of two C^1-maps

$$(x,y)\mapsto(u,v)=\big(\,1+x^2+y^2,\,\sin(xy)\,\big),\;\;(u,v)\mapsto f=u^v.$$

Then

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial x}+\frac{\partial f}{\partial v}\cdot\frac{\partial v}{\partial x}$$

$$=vu^{v-1}\cdot(2x)+u^v\ln u\cdot y\cos(xy)=u^v\cdot\Big(\frac{v}{u}\cdot(2x)+(\ln u)\cdot y\cos(xy)\Big)$$

$$=(x^2+y^2+1)^{\sin(xy)}\left(\frac{2x\sin(xy)}{x^2+y^2+1}+y\cos(xy)\ln(x^2+y^2+1)\right). \qquad \square$$

Example 13.3.3. Suppose that $f:\mathbb{R}^2\to\mathbb{R}$ is a differentiable function depending on two variables $f=f(x,y)$. Suppose additionally that x,y are themselves functions of two variables

$$x=x(r,\theta)=r\cos\theta,\;\; y=y(r,\theta)=r\sin\theta. \tag{13.3.6}$$

We want to compute the partial derivatives $\frac{\partial f}{\partial r}$ and $\frac{\partial f}{\partial\theta}$. First, let us give a geometric interpretation to the functions (13.3.6).

If we fix r, say $r=4$, then we get a path

$$\theta\mapsto\big(\,4\cos\theta,4\sin\theta\,\big)\in\mathbb{R}^2.$$

This describes the motion of a point in the plane with constant angular velocity along the circle of radius 4 centered at the origin; see the thick orange circle in Figure 13.4. If we keep θ fixed, $\theta=\theta_0$, then the resulting path

$$r\mapsto\big(\,r\cos\theta_0,r\sin\theta_0\,\big)$$

describes the motion with speed 1 along a ray emanating at the origin that makes angle θ_0 with the x-axis.

We get two families of curves in the plane: the family of curves obtained by fixing r (circles centered at the origin) and the family of curves obtained by fixing θ (rays). These two families form a *curvilinear grid* in the plane (see Figure 13.4) known as the *polar grid.*

The function f depends on the variables x,y, which themselves depend on the quantities r,θ. $\frac{\partial f}{\partial r}$ measures how fast is f changing when we travel at unit speed along a ray,

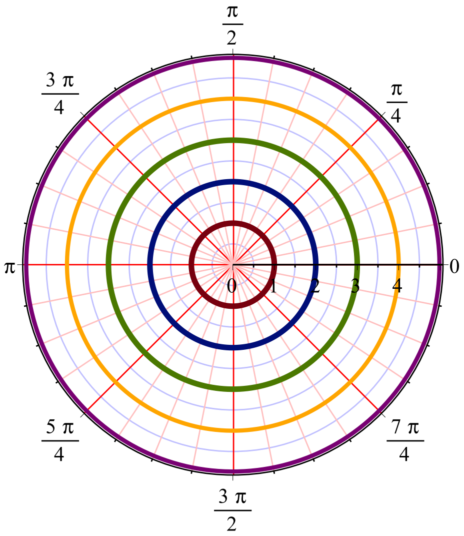

Figure 13.4. *Polar grid.*

while $\frac{\partial f}{\partial \theta}$ measures how fast is f changing when we travel along a circle at constant angular velocity 1rad/sec. The chain rule shows that

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \frac{\partial f}{\partial x}\cos\theta + \frac{\partial f}{\partial y}\sin\theta \tag{13.3.7a}$$

$$\frac{\partial f}{\partial \theta} = -\frac{\partial f}{\partial x} r\sin\theta + \frac{\partial f}{\partial y} r\cos\theta. \tag{13.3.7b}$$

Suppose for example that $f(x,y) = x^2 + y^2$. Note that if x, y depend on r, θ as in (13.3.6), then $x^2 + y^2 = r^2$ and thus

$$\frac{\partial f}{\partial r} = 2r.$$

On the other hand (13.3.7a) implies that

$$\frac{\partial f}{\partial r} = 2x\cos\theta + 2y\sin\theta \overset{(13.3.6)}{=} 2r\cos^2\theta + 2r\sin^2\theta = 2r. \qquad \square$$

Remark 13.3.4 (The naturality of the differential)**.** We want to describe a remarkable "accident" which is extremely important in differential geometry and theoretical physics.

Suppose that f is a differentiable function depending on the n variables $x^1, \dots, x^n$. Using the convention (13.2.13) we have

$$\boxed{df = \frac{\partial f}{\partial x^1}dx^1 + \cdots + \frac{\partial f}{\partial x^n}dx^n}. \tag{13.3.8}$$

Suppose that the quantities $x^1, \dots, x^n$ themselves depend differentiably on a number of variables

$$x^i = x^i(u^1, \dots, u^m), \;\; i = 1, \dots, n. \tag{13.3.9}$$

Through this new dependence we can view the quantity f as a function of the variables $u^1, \dots, u^m$ and, as such, we have

$$df = \frac{\partial f}{\partial u^1} du^1 + \cdots + \frac{\partial f}{\partial u^m} du^m. \tag{13.3.10}$$

? *How do we reconcile (13.3.8) with (13.3.10)?*

The chain rule comes to the rescue. To see that (13.3.8) and (13.3.10) are compatible (noncontradictory) regard the quantities $dx^1, \dots, dx^n$ as the differentials of the functions in (13.3.9), i.e.,

$$dx^i = \frac{\partial x^i}{\partial u^1} du^1 + \cdots + \frac{\partial x^i}{\partial u^m} du^m, i = 1, \dots, n.$$

The equality (13.3.8) becomes

$$df = \frac{\partial f}{\partial x^1}\left(\frac{\partial x^1}{\partial u^1} du^1 + \cdots + \frac{\partial x^1}{\partial u^m} du^m\right) + \cdots + \frac{\partial f}{\partial x^n}\left(\frac{\partial x^n}{\partial u^1} du^1 + \cdots + \frac{\partial x^n}{\partial u^m} du^m\right)$$

$$= \underbrace{\left(\frac{\partial f}{\partial x^1}\frac{\partial x^1}{\partial u^1} + \cdots + \frac{\partial f}{\partial x^n}\frac{\partial x^n}{\partial u^1}\right)}_{=:q_1} du^1 + \cdots + \underbrace{\left(\frac{\partial f}{\partial x^1}\frac{\partial x^1}{\partial u^m} + \cdots + \frac{\partial f}{\partial x^n}\frac{\partial x^n}{\partial u^m}\right)}_{=:q_m} du^m.$$

Hence

$$df = q_1 du^1 + \cdots + q_m du^m. \tag{13.3.11}$$

The chain rule (13.3.5) shows that

$$q_1 = \frac{\partial f}{\partial u^1}, \dots, q_m = \frac{\partial f}{\partial u^m}$$

so the equality (13.3.11) is none other than (13.3.10) in disguise. □

Let us discuss a few simple but useful applications of the chain rule.

Definition 13.3.5 (Differentiable paths)**.** Let $n \in \mathbb{N}$. A *differentiable path* in $\mathbb{R}^n$ is a differentiable map $\boldsymbol{\gamma} : I \to \mathbb{R}^n$, where $I \subset \mathbb{R}$ is an interval. □

A differentiable path $\boldsymbol{\gamma} : (a, b) \to \mathbb{R}^n$ is described by n differentiable functions

$$x^i : (a, b) \to \mathbb{R}, \quad i = 1, \dots, n,$$

such that

$$\boldsymbol{\gamma}(t) = \begin{bmatrix} x^1(t) \\ x^2(t) \\ \vdots \\ x^n(t) \end{bmatrix}.$$

The differential of the map $\boldsymbol{\gamma}$ is an $n \times 1$ matrix, i.e., a matrix consisting of a single column of height n. This matrix is

$$\frac{d}{dt}\boldsymbol{\gamma}(t) = \begin{bmatrix} \frac{dx^1(t)}{dt} \\ \\ \frac{dx^2(t)}{dt} \\ \vdots \\ \frac{dx^n(t)}{dt} \end{bmatrix}$$

We will adopt a convention frequently used by physicists and will denote by an upper dot "$\dot{}$" the *time derivatives*. With this convention we can rewrite the above equality as

$$\dot{\boldsymbol{\gamma}}(t) = \begin{bmatrix} \dot{x}^1(t) \\ \\ \dot{x}^2(t) \\ \vdots \\ \\ \dot{x}^n(t) \end{bmatrix}.$$

If we think of $\boldsymbol{\gamma}$ as describing the motion of a point in $\mathbb{R}^n$, then the vector $\dot{\boldsymbol{\gamma}}(t)$ is the *velocity* of that moving point at the moment of time t.

Proposition 13.3.6 (Derivatives along paths)**.** *Let $n \in \mathbb{N}$. Assume that $U \subset \mathbb{R}^n$ is an open set, $f : U \to \mathbb{R}$ is a Fréchet differentiable function and $\boldsymbol{\gamma} : (a,b) \to U$ a differentiable path. Then*

$$\boxed{\frac{d}{dt}f\big(\boldsymbol{\gamma}(t)\big) = \big\langle \nabla f\big(\boldsymbol{\gamma}(t)\big), \dot{\boldsymbol{\gamma}}(t)\big\rangle, \;\; \forall t \in (a,b)}, \tag{13.3.12}$$

where we recall that $\nabla f(\boldsymbol{x})$ denotes the gradient of f at $\boldsymbol{x}$. The quantity $\langle \nabla f(\boldsymbol{\gamma}), \dot{\boldsymbol{\gamma}}\rangle$ is called the derivative of f along the path $\boldsymbol{\gamma}$.

Proof. As explained above, the path $\boldsymbol{\gamma}$ is described by n differentiable functions

$$\boldsymbol{\gamma}(t) = \big(x^1(t), \dots, x^n(t)\big).$$

We have

$$f\big(\boldsymbol{\gamma}(t)\big) = f\big(x^1(t), \dots, x^n(t)\big).$$

Using the chain rule (13.3.5) we deduce

$$\frac{d}{dt}f\big(\boldsymbol{\gamma}(t)\big) = \frac{\partial f(\boldsymbol{\gamma}(t))}{\partial x^1}\frac{dx^1(t)}{dt} + \cdots + \frac{\partial f(\boldsymbol{\gamma}(t))}{\partial x^n}\frac{dx^n(t)}{dt}$$

$$= \frac{\partial f(\boldsymbol{\gamma})}{\partial x^1}\dot{x}^1 + \cdots + \frac{\partial f(\boldsymbol{\gamma})}{\partial x^n}\dot{x}^n = \langle \nabla f(\boldsymbol{\gamma}), \dot{\boldsymbol{\gamma}}\rangle.$$

□

If we think of the function $f : U \to \mathbb{R}$ as a physical quantity associated to each point in U (say temperature) and of the path $\boldsymbol{\gamma}$ as describing the motion of a point in U, then the derivative of f along the path is the rate of change of f (per unit of time) during the motion.

Example 13.3.7 (Euler's identity)**.** Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is positively homogeneous of degree k, i.e.,

$$f(t\boldsymbol{x}) = t^k f(\boldsymbol{x}), \;\; \forall t > 0, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}.$$

If f is differentiable on $\mathbb{R}^n \backslash \{\boldsymbol{0}\}$, then f satisfies *Euler's identity*

$$\big\langle \boldsymbol{x}, \nabla f(\boldsymbol{x}) \big\rangle = k f(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}. \tag{13.3.13}$$

To prove the above identity, fix $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ and consider the path

$$\boldsymbol{\gamma}_{\boldsymbol{x}} : (0, \infty) \to \mathbb{R}^n, \;\; \boldsymbol{\gamma}_{\boldsymbol{x}}(t) = t\boldsymbol{x}, \;\; \forall t > 0.$$

Observe that

$$f\big(\boldsymbol{\gamma}_{\boldsymbol{x}}(t)\big) = f(t\boldsymbol{x}) = t^k f(\boldsymbol{x}), \;\; \dot{\boldsymbol{\gamma}}_{\boldsymbol{x}}(t) = \boldsymbol{x}, \;\; \forall t > 0.$$

Thus

$$\frac{d}{dt} f\big(\boldsymbol{\gamma}_{\boldsymbol{x}}(t)\big) = k t^{k-1} f(\boldsymbol{x}), \;\; \forall t > 0.$$

On the other hand, the derivative of f along $\boldsymbol{\gamma}_{\boldsymbol{x}}(t)$ is given by (13.3.12)

$$\frac{d}{dt} f\big(\boldsymbol{\gamma}_{\boldsymbol{x}}(t)\big) = \big\langle \dot{\boldsymbol{\gamma}}_{\boldsymbol{x}}(t), \nabla f\big(\boldsymbol{\gamma}_{\boldsymbol{x}}(t)\big) \big\rangle = \big\langle \boldsymbol{x}, \nabla f(t\boldsymbol{x}) \big\rangle.$$

We deduce

$$\big\langle \boldsymbol{x}, \nabla f(t\boldsymbol{x}) \big\rangle = k t^{k-1} f(\boldsymbol{x}), \;\; \forall t > 0.$$

If we set $t = 1$ in the above equality we obtain Euler's identity (13.3.13). □

Theorem 13.3.8 (Lagrange mean value theorem)**.** *Suppose that $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is a differentiable function. Then, for any $\boldsymbol{x}_0, \boldsymbol{x}_1 \in U$ such that $[\boldsymbol{x}_0, \boldsymbol{x}_1] \subset U$, there exists a point $\boldsymbol{p}$ on the line segment $[\boldsymbol{x}_0, \boldsymbol{x}_1]$ such that*

$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}_0) = \langle \nabla f(\boldsymbol{p}), \boldsymbol{x}_1 - \boldsymbol{x}_0 \rangle.$$

Proof. Consider the restriction of f to the line segment $[\boldsymbol{x}_0, \boldsymbol{x}_1]$, i.e., the function $g : [0, 1] \to \mathbb{R}$

$$g(t) = f\big(\boldsymbol{x}_0 + t(\boldsymbol{x}_1 - \boldsymbol{x}_0)\big).$$

According to the 1-dimensional Lagrange mean value theorem there exists $\tau \in (0, 1)$ such that

$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}_0) = g(1) - g(0) = g'(\tau).$$

On the other hand, the derivative of f along the path $t \mapsto \boldsymbol{x}_0 + t(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ is

$$g'(t) = \big\langle \nabla f(\boldsymbol{x}_0 + t(\boldsymbol{x}_1 - \boldsymbol{x}_0)), \boldsymbol{x}_1 - \boldsymbol{x}_0 \big\rangle.$$

This yields the desired conclusion with $\boldsymbol{p} = \boldsymbol{x}_0 + \tau(\boldsymbol{x}_1 - \boldsymbol{x}_0)$. □

Corollary 13.3.9. *Suppose that $U \subset \mathbb{R}^n$ is an open convex set and $f : U \to \mathbb{R}$ is a differentiable function. If there exists $C > 0$ such that $\|\nabla f(\boldsymbol{x})\| \leqslant C$, $\forall \boldsymbol{x} \in U$, then*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant C\|\boldsymbol{x} - \boldsymbol{y}\|, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in U. \tag{13.3.14}$$

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in U$. The mean value theorem shows that there exists a point $\boldsymbol{p}$ on the line segment $[\boldsymbol{x}, \boldsymbol{y}]$ such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| = \big|\langle \nabla f(\boldsymbol{p}), \boldsymbol{x} - \boldsymbol{y}\rangle\big|.$$

The desired conclusion now follows by invoking the Cauchy-Schwarz inequality

$$\big|\langle \nabla f(\boldsymbol{p}), \boldsymbol{x} - \boldsymbol{y}\rangle\big| \leqslant \|\nabla f(\boldsymbol{p})\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \leqslant C\|\boldsymbol{x} - \boldsymbol{y}\|.$$

□

Corollary 13.3.10. *Suppose that $U \subset \mathbb{R}^n$ is an open and path connected set and $f : U \to \mathbb{R}$ is a differentiable function such that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, $\forall \boldsymbol{x} \in U$. Then the function f is constant.*

Proof. Fix $\boldsymbol{p}_0 \in U$. Let $\boldsymbol{q}$ be an arbitrary point in U. Since U is path connected, Exercise 12.24 shows that there exist points $\boldsymbol{p}_1, \dots, \boldsymbol{p}_N$ such that $\boldsymbol{q} = \boldsymbol{p}_N$ and the line segments $[\boldsymbol{p}_{i-1}, \boldsymbol{p}_i]$, $i = 1, \dots, N$, are contained in U. Corollary 13.3.9 then implies

$$f(\boldsymbol{p}_0) = f(\boldsymbol{p}_1) = f(\boldsymbol{p}_2) = \cdots = f(\boldsymbol{p}_{N-1}) = f(\boldsymbol{p}_N) = f(\boldsymbol{q}).$$

We have thus proved that $f(\boldsymbol{q}) = f(\boldsymbol{p}_0)$, $\forall \boldsymbol{q} \in U$, i.e., f is constant. □

Corollary 13.3.11. *Suppose that $U \subset \mathbb{R}^n$ is an open convex set and $\boldsymbol{F} : U \to \mathbb{R}^m$ is a C^1-map. Suppose that there exists a constant $C > 0$ such that $\|J_{\boldsymbol{F}}(\boldsymbol{x})\|_{HS} \leqslant C$, $\forall \boldsymbol{x} \in U$, where $\| - \|_{HS}$ denotes the Hilbert-Schmidt norm of an $m \times n$ matrix; see Remark 12.1.11. Then*

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| \leqslant C\sqrt{m}\|\boldsymbol{x} - \boldsymbol{y}\|, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in U. \tag{13.3.15}$$

Proof. Denote by $F^1, \dots, F^m$ the components of $\boldsymbol{F}$. Note that

$$\|J_{\boldsymbol{F}}(\boldsymbol{x})\|_{HS}^2 = \sum_{i=1}^m \|\nabla F^i(\boldsymbol{x})\|^2, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Hence

$$\|\nabla F^i(\boldsymbol{x})\| \leqslant \|J_{\boldsymbol{F}}(\boldsymbol{x})\|_{HS}, \;\; \forall \boldsymbol{x} \in U, \;\; i = 1, \dots, m.$$

Then, for any $\boldsymbol{x}, \boldsymbol{y} \in U$ we have

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\|^2 = \sum_{i=1}^m \|F^i(\boldsymbol{x}) - F^i(\boldsymbol{y})\|^2 \overset{(13.3.14)}{\leqslant} \sum_{i=1}^m C^2\|\boldsymbol{x} - \boldsymbol{y}\|^2 = C^2 m\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

□

Definition 13.3.12 (Vector fields)**.** Let $n \in \mathbb{N}$.

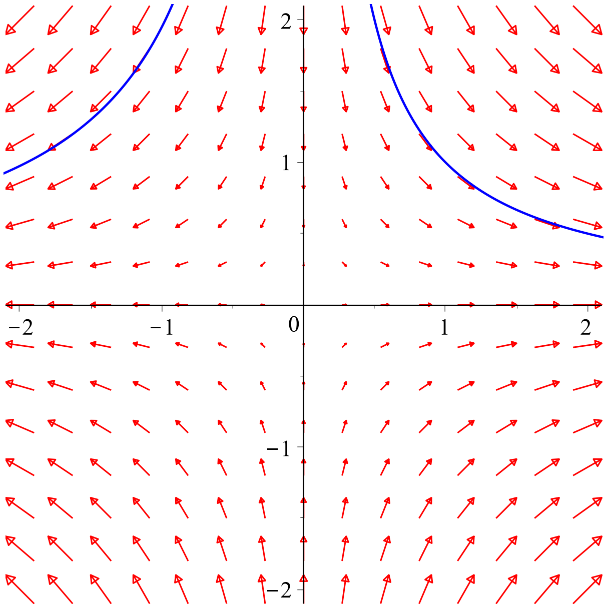

Figure 13.5. *The vector field* $V(x,y) = (2x,-2y)$ *on the square* $S = [-2,2] \times [-2,2] \subset \mathbb{R}^2$ *and two integral curves of this vector field.*

(i) A *vector field* on a set $S \subset \mathbb{R}^n$ is a map

$$\boldsymbol{V} : S \to \mathbb{R}^n, \quad S \ni \boldsymbol{x} \mapsto \boldsymbol{V}(\boldsymbol{x}).$$

(ii) An *integral curve* or *flow line* of a vector field $\boldsymbol{V}$ on a set $S \subset \mathbb{R}^n$ is a differentiable path $\boldsymbol{\gamma} : (a,b) \to \mathbb{R}^n$ such that

$$\boldsymbol{\gamma}(t) \in S, \quad \dot{\boldsymbol{\gamma}}(t) = \boldsymbol{V}\big(\boldsymbol{\gamma}(t)\big), \quad \forall t \in (a,b). \tag{13.3.16}$$

□

Let us emphasize a one aspect in the definition of a vector field that you may overlook. The domain of the vector field, i.e., set S, lives inside the space $\mathbb{R}^n$ and $\boldsymbol{V}$ takes values in <u>*the same*</u> vector space $\mathbb{R}^n$. Intuitively, a vector field $\boldsymbol{V}$ on a set $S \subset \mathbb{R}^n$ associates to each point $\boldsymbol{x} \in S$ a vector $\boldsymbol{V}(\boldsymbol{x}) \in \mathbb{R}^n$ that should be visualized as an arrow $V(\boldsymbol{x})$ originating at $\boldsymbol{x}$. The result is a "hairy" region S, with one "hair" $\boldsymbol{V}(\boldsymbol{x})$ at each location $\boldsymbol{x} \in S$; see Figure 13.5.

An integral curve of the vector field $\boldsymbol{V}$ is then a path $\boldsymbol{\gamma}(t)$ in S whose velocity $\dot{\boldsymbol{\gamma}}(t)$ at each point $\boldsymbol{\gamma}(t)$ is equal to $\boldsymbol{V}\big(\boldsymbol{\gamma}(t)\big)$: this is precisely the arrow the vector field associates to $\boldsymbol{\gamma}(t)$. In particular, this arrow is tangent to the path at this point; see the blue curves in Figure 13.5.

A vector field $\boldsymbol{V}$ on S is determined by n functions on S

$$\boldsymbol{V}(\boldsymbol{x}) = \left[\begin{array}{c} V^1(\boldsymbol{x}) \\ \vdots \\ V^n(\boldsymbol{x}) \end{array}\right], \forall \boldsymbol{x} \in S, \;\; V^i : S \to \mathbb{R}, \;\; i = 1, \dots, n.$$

An integral curve $\boldsymbol{\gamma} : (a,b) \to S \subset \mathbb{R}^n$ of $\boldsymbol{V}$ is then given by n functions $x^1(t), \dots, x^n(t)$, $t \in (a,b)$ satisfying the system of differential equations

$$\left\{\begin{array}{ccc} \dot{x}^1(t) & = & V^1\big(x^1(t), \dots, x^n(t)\big) \\ \vdots & \vdots & \vdots \\ \dot{x}^n(t) & = & V^n\big(x^1(t), \dots, x^n(t)\big) \end{array}\right. . \tag{13.3.17}$$

Example 13.3.13 (Gradient vector fields)**.** Suppose that $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is a smooth function. The gradient of f defines a vector field on U,

$$U \ni \boldsymbol{x} \mapsto \nabla f(\boldsymbol{x}) \in \mathbb{R}^n.$$

This vector field is called the *gradient vector field* of (or associated to) the function f. Such a function f is called a *potential* of the gradient vector field.

The vector field depicted in Figure 13.5 is the gradient vector field of the function $f(x,y) = x^2 - y^2$,

$$\nabla f(x,y) = [2x, -2y]^\top,$$

The integral curves of this vector field are differentiable maps

$$\mathbb{R} \ni t \mapsto \left[\begin{array}{c} x(t) \\ y(t) \end{array}\right] \in \mathbb{R}^2$$

satisfying the system of differential equations

$$\left\{\begin{array}{ccc} \dot{x} & = & 2x \\ \dot{y} & = & -2y \end{array}\right. .$$

Arguing as in Example 8.5.12 we deduce that the solutions of the first equations have the form $x(t) = ae^{2t}$, a constant, while the solutions of the second equation are $y(t) = be^{-2t}$, b constant.

The vector field

$$\mathbb{R}^2 \ni [x,y]^\top \mapsto \boldsymbol{V}(x,y) = \left[\begin{array}{c} -y \\ x \end{array}\right] \in \mathbb{R}^2 \tag{13.3.18}$$

depicted in Figure 13.6 *is not* the gradient of any function. Exercise 13.19 asks you to prove this. □

Definition 13.3.14. Suppose that $\boldsymbol{V}$ is a vector field on the set $S \subset \mathbb{R}^n$. A *prime integral* or *conservation law* of $\boldsymbol{V}$ is a continuous function $f : S \to \mathbb{R}$ that is constant along the flow lines of $\boldsymbol{V}$, i.e., for any integral curve $\boldsymbol{\gamma} : (a,b) \to S$ of $\boldsymbol{V}$, the function

$$(a,b) \ni t \mapsto f(\boldsymbol{\gamma}(t)) \in \mathbb{R}$$

is constant. □

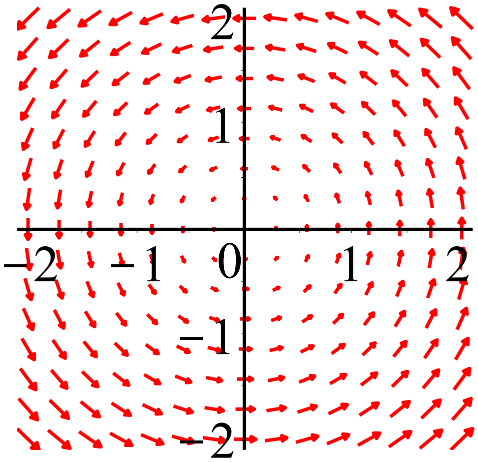

Figure 13.6. *A non-gradient vector field on the square $S = [-2,2] \times [-2,2] \subset \mathbb{R}^2$.*

Proposition 13.3.15. *Suppose that $\boldsymbol{V}$ is a vector field on the open set $U \subset \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is a differentiable function such that*

$$\big\langle\, \nabla f(\boldsymbol{x}), \boldsymbol{V}(\boldsymbol{x})\,\big\rangle = 0, \;\; \forall \boldsymbol{x} \in U.$$

Then f is a prime integral of $\boldsymbol{V}$.

Proof. Let $\boldsymbol{\gamma} : (a,b) \to U$ be an integral curve of $\boldsymbol{V}$. Then

$$\dot{\boldsymbol{\gamma}}(t) = \boldsymbol{V}(\,\boldsymbol{\gamma}(t)\,),$$

and

$$\frac{d}{dt} f(\,\boldsymbol{\gamma}(t)\,) = \langle \nabla f(\,\boldsymbol{\gamma}(t)\,), \dot{\boldsymbol{\gamma}}(t)\,\rangle = \big\langle\, \nabla f\big(\,\boldsymbol{\gamma}(t)\,\big), \boldsymbol{V}\big(\,\boldsymbol{\gamma}(t)\,\big)\,\big\rangle = 0.$$

□

13.4. Higher order partial derivatives

Let $U \subset \mathbb{R}^n$ be an open set and $f : U \to \mathbb{R}$ be a function such that the partial derivatives $\partial_{x^1} f(\boldsymbol{x}), \dots, \partial_{x^n} f(\boldsymbol{x})$ exist at every point $\boldsymbol{x} \in U$. We obtain n new functions

$$\partial_{x^1} f, \dots, \partial_{x^n} f : U \to \mathbb{R}. \tag{13.4.1}$$

We say that f admits *second order* partial derivatives on U if each of the functions (13.4.1) admit partial derivatives on U. We say that f admits *third order* partial derivatives on U if each of the functions (13.4.1) admit second order partial derivatives on U. Inductively, if $k \in \mathbb{N}$, we say that f admits *partial derivatives of order* k on U if each of the functions (13.4.1) admit partial derivatives of order $k-1$ on U.

Recall that f is said to be C^1 on U if the functions (13.4.1) are continuous on U. We say that f is C^2 on U if the functions (13.4.1) are C^1 on U. We say that f is C^3 on U if the functions (13.4.1) are C^2 on U. Inductively, if $k \in \mathbb{N}$, we say that f is C^k on U if the functions (13.4.1) are C^{k-1} on U. We will write $f \in C^k(U)$ to indicate that f is C^k on U. We say that the function f is *smooth* or C^∞ on U, and we denote this $f \in C^\infty(U)$ if

$$f \in C^k(U), \quad \forall k \in \mathbb{N}.$$

Note that $C^k(U)$ stands for the collection of all functions $f : U \to \mathbb{R}$ that are C^k on U. This collection is a vector space.

Suppose that $f : U \to \mathbb{R}$ is a function that admits second order derivatives on U. Thus, each of the first order derivatives $\partial_{x^j} f$, $j = 1, \dots, n$, admits in its turn first order derivatives. We denote

$$\partial^2_{x^k x^j} f \;\text{ or }\; \frac{\partial^2 f}{\partial x^k \partial x^j}$$

the partial derivative of the function $\partial_{x^j} f$ with respect to the variable x^k, i.e.,

$$\boxed{\partial^2_{x^k x^j} f := \partial_{x^k}\big(\,\partial_{x^j} f\,\big)}.$$

More generally, if f is a C^k-function, then for any $i_1, \dots, i_k \in \{1, \dots, n\}$ we define inductively

$$\partial^k_{x^{i_k} \cdots x^{i_1}} f := \partial_{x^{i_k}}\big(\,\partial^{k-1}_{x^{i_{k-1}} \cdots x^{i_1}} f = \partial_{x^{i_k}}\,\big).$$

We have the following important result.

Theorem 13.4.1 (Partial derivatives commute). *Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set. Then for any function $f \in C^2(U)$ we have*

$$\partial^2_{x^k x^j} f(\boldsymbol{x}) = \partial^2_{x^j x^k} f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in U, \quad \forall j, k = 1, \dots, n.$$

Proof. The result is obviously true when $j = k$ so it suffices to consider the case $j \neq k$, say $j < k$. Fix a point $\boldsymbol{a} \in U$. We have to prove that

$$\partial^2_{x^k x^j} f(\boldsymbol{a}) = \partial^2_{x^j x^k} f(\boldsymbol{a}). \tag{13.4.2}$$

We have

$$\partial_{x^j} f(\boldsymbol{x}) = \lim_{s \to 0} \frac{f(\boldsymbol{x} + s\boldsymbol{e}_j) - f(\boldsymbol{x})}{s}, \quad \forall \boldsymbol{x} \in U \tag{13.4.3a}$$

$$\partial_{x^k} f(\boldsymbol{x}) = \lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_k) - f(\boldsymbol{x})}{t}, \quad \forall \boldsymbol{x} \in U \tag{13.4.3b}$$

$$\partial^2_{x^k x^j} f(\boldsymbol{a}) = \lim_{t \to 0} \frac{\partial_{x^j} f(\boldsymbol{a} + t\boldsymbol{e}_k) - \partial_{x^j} f(\boldsymbol{a})}{t}$$

$$\overset{(13.4.3a)}{=} \lim_{t \to 0} \left(\lim_{s \to 0} \frac{1}{t} \cdot \frac{f(\boldsymbol{a} + t\boldsymbol{e}_k + s\boldsymbol{e}_j) - f(\boldsymbol{a} + t\boldsymbol{e}_k) - f(\boldsymbol{a} + s\boldsymbol{e}_j) + f(\boldsymbol{a})}{s} \right).$$

Similarly

$$\partial^2_{x^j x^k} f(\boldsymbol{a}) = \lim_{s \to 0} \frac{\partial_{x^k} f(\boldsymbol{a} + s\boldsymbol{e}_j) - \partial_{x^k} f(\boldsymbol{a})}{s}$$

$$\overset{(13.4.3a)}{=} \lim_{s\to 0}\left(\lim_{t\to 0}\frac{1}{s}\cdot\frac{f(\boldsymbol{a}+t\boldsymbol{e}_k+s\boldsymbol{e}_j)-f(\boldsymbol{a}+s\boldsymbol{e}_j)-f(\boldsymbol{a}+t\boldsymbol{e}_k)+f(\boldsymbol{a})}{t}\right).$$

If we denote by $Q(s,t)$, $s,t\neq 0$, the quantity

$$Q(s,t):=f(\boldsymbol{a}+t\boldsymbol{e}_k+s\boldsymbol{e}_j)-f(\boldsymbol{a}+t\boldsymbol{e}_k)-f(\boldsymbol{a}+s\boldsymbol{e}_j)+f(\boldsymbol{a})$$

then we see that (13.4.2) is equivalent with the equality

$$\lim_{s\to 0}\left(\lim_{t\to 0}\frac{Q(s,t)}{st}\right)=\lim_{t\to 0}\left(\lim_{s\to 0}\frac{Q(s,t)}{st}\right). \tag{13.4.4}$$

The two sides above are examples of *iterated limits*, and they differ only in the order we take the limits. It suggests that Theorem 13.4.1 is at least plausible. However, the complete proof of (13.4.4) is not trivial and requires a bit of sweat.

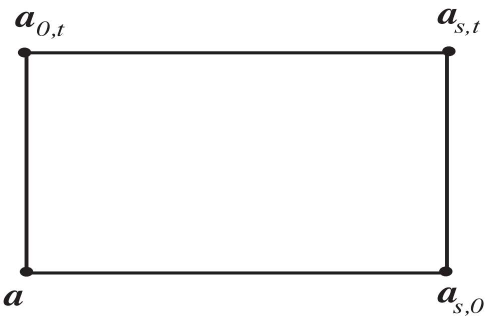

Figure 13.7. *The rectangle $\mathcal{R}_{s,t}$ at $\boldsymbol{a}$ spanned by the vectors $s\boldsymbol{e}_j$ and $t\boldsymbol{e}_k$.*

For simplicity, for $s,t\in\mathbb{R}$ we set (see Figure 13.7)

$$\boldsymbol{a}_{s,t}:=\boldsymbol{a}+s\boldsymbol{e}_j+t\boldsymbol{e}_k.$$

Denote by $\mathcal{R}_{s,t}$ the rectangle with vertices $\boldsymbol{a},\boldsymbol{a}_{s,0},\boldsymbol{a}_{0,t},\boldsymbol{a}_{s,t}$. Note also that

$$Q(s,t)=f(\boldsymbol{a}_{s,t})-f(\boldsymbol{a}_{0,t})-f(\boldsymbol{a}_{s,0})+f(\boldsymbol{a}),$$

and that st is the area of this rectangle.

Applying the Lagrange Mean Value Theorem to the function $\lambda\mapsto g_t(\lambda)=f(\boldsymbol{a}_{\lambda,t})-f(\boldsymbol{a}_{\lambda,0})$ we deduce that, for any s,t small, there exists $\bar{\lambda}=\lambda_{s,t}\in(0,s)$ such that

$$\frac{Q(s,t)}{s}=\frac{g_t(s)-g_t(0)}{s}$$

$$=g_t'(\lambda)=\partial_{x^j}f(\boldsymbol{a}+\lambda\boldsymbol{e}_j+t\boldsymbol{e}_k)-\partial_{x^j}f(\boldsymbol{a}+\lambda\boldsymbol{e}_j)=\partial_{x^j}f(\boldsymbol{a}_{\lambda,t})-\partial_{x^j}f(\boldsymbol{a}_{\lambda,0}).$$

Applying the Lagrange Mean Value Theorem to the function

$$h(\mu)=\partial_{x^j}f(\boldsymbol{a}+\mu\boldsymbol{e}_k+\lambda_{s,t}\boldsymbol{e}_j)$$

we deduce that there exists $\bar{\mu}=\mu_{s,t}$ in the interval $(0,t)$ such that

$$\frac{Q(s,t)}{st}=\frac{\partial_{x^j}f(\boldsymbol{a}+t\boldsymbol{e}_k+\lambda_{s,t}\boldsymbol{e}_j)-\partial_{x^j}f(\boldsymbol{a}+\lambda_{s,t}\boldsymbol{e}_j)}{t}=\frac{h(t)-h(0)}{t}$$

$$=h'(\mu_{s,t})=\partial^2_{x^kx^j}f(\boldsymbol{a}+\mu_{s,t}\boldsymbol{e}_k+\lambda_{s,t}\boldsymbol{e}_j)=\partial^2_{x^kx^j}f(\boldsymbol{a}_{\bar{\lambda},\bar{\mu}}).$$

We denote by $\boldsymbol{p}_{s,t}$ the point $\boldsymbol{a}+\mu_{s,t}\boldsymbol{e}_k+\lambda_{s,t}\boldsymbol{e}_j$. Thus

$$\frac{Q(s,t)}{st}=\partial^2_{x^kx^j}f(\boldsymbol{p}_{s,t}). \tag{13.4.5}$$

Applying the Lagrange Mean Value Theorem to the function $\beta\mapsto u_s(\beta)=Q(s,\beta)$ we deduce that for every s,t sufficiently small there exists $\bar{\beta}=\beta_{s,t}$ in $(0,t)$ such that

$$\frac{Q(s,t)}{t}=\frac{u_s(t)-u_s(0)}{t}=u_s'(\beta)=\partial_{x^k}f(\boldsymbol{a}+s\boldsymbol{e}_j+\bar{\beta}\boldsymbol{e}_k)-\partial_{x^k}f(\boldsymbol{a}+\bar{\beta}\boldsymbol{e}_k).$$

Applying the Lagrange Mean Value Theorem to the function

$$\alpha\mapsto v(\alpha)=\partial_{x^k}f(\boldsymbol{a}+\alpha\boldsymbol{e}_j+\bar{\beta}\boldsymbol{e}_k)$$

we deduce that there exists $\bar{\alpha}=\alpha_{s,t}$ in the interval $(0,s)$ such that

$$\frac{Q(s,t)}{st}=\frac{u_s(t)-u_s(0)}{st}=\frac{\partial_{x^k}f(\boldsymbol{a}+s\boldsymbol{e}_j+\bar{\beta}\boldsymbol{e}_k)-\partial_{x^k}f(\boldsymbol{a}+\bar{\beta}\boldsymbol{e}_k)}{s}$$

$$=\frac{v(s)-v(0)}{s}=v'(\bar{\alpha})=\partial^2_{x^jx^k}f(\boldsymbol{a}+\bar{\alpha}\boldsymbol{e}_j+\bar{\beta}\boldsymbol{e}_k).$$

We denote by $\boldsymbol{q}_{s,t}$ the point $\boldsymbol{a}+\alpha_{s,t}\boldsymbol{e}_j+\beta_{s,t}\boldsymbol{e}_k$. Thus

$$\frac{Q(s,t)}{st}=\partial^2_{x^jx^k}f(\boldsymbol{q}_{s,t}). \tag{13.4.6}$$

In particular, we deduce

$$\partial^2_{x^kx^j}f(\boldsymbol{p}_{s,t})=\frac{Q(s,t)}{st}=\partial^2_{x^jx^k}f(\boldsymbol{q}_{s,t}). \tag{13.4.7}$$

Note that since $\alpha,\lambda\in(0,s)$, $\beta,\mu\in(0,t)$ we have

$$\operatorname{dist}(\boldsymbol{a},\boldsymbol{p}_{s,t})=\sqrt{\bar{\lambda}^2+\bar{\mu}^2}\leqslant\sqrt{s^2+t^2}, \tag{13.4.8a}$$

$$\operatorname{dist}(\boldsymbol{a},\boldsymbol{q}_{s,t})=\sqrt{\bar{\alpha}^2+\bar{\beta}^2}\leqslant\sqrt{s^2+t^2}. \tag{13.4.8b}$$

Fix $r>0$ sufficiently small such that the closed ball $\overline{B_r(\boldsymbol{a})}$ is contained in U. Since the functions

$$\partial^2_{x^kx^j}f,\quad \partial^2_{x^jx^k}f:U\to\mathbb{R}$$

are continuous, they are continuous at $\boldsymbol{a}$. Hence, for any $\varepsilon>0$, there exists $\delta=\delta(\varepsilon)>0$ such that

$$\begin{aligned}\forall\boldsymbol{x}\in U,\ \ \operatorname{dist}(\boldsymbol{a},\boldsymbol{x})<\delta(\varepsilon)\Rightarrow&\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{a})-\partial^2_{x^kx^j}(\boldsymbol{x})\,\bigr|<\frac{\varepsilon}{2},\\ &\bigl|\,\partial^2_{x^jx^k}f(\boldsymbol{a})-\partial^2_{x^jx^k}(\boldsymbol{x})\,\bigr|<\frac{\varepsilon}{2}.\end{aligned} \tag{13.4.9}$$

Fix $\varepsilon>0$. Choose $s,t>0$ small enough such that $\sqrt{s^2+t^2}<\delta(\varepsilon)$ and $\mathcal{R}_{s,t}\subset B_r(\boldsymbol{a})$. The points $\boldsymbol{p}_{s,t}$ and $\boldsymbol{q}_{s,t}$ belong to the rectangle $\mathcal{R}_{s,t}$ and thus, also to U. We deduce

$$\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{a})-\partial^2f_{x^jx^k}(\boldsymbol{a})\,\bigr|$$

$$\leqslant\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{a})-\partial^2_{x^kx^j}f(\boldsymbol{p}_{s,t})\,\bigr|+\boxed{\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{p}_{s,t})-\partial^2_{x^jx^k}f(\boldsymbol{q}_{st})\,\bigr|}+\bigl|\,\partial^2_{x^jx^k}f(\boldsymbol{q}_{st})-\partial^2f_{x^jx^k}(\boldsymbol{a})\,\bigr|$$

$$\overset{(13.4.7)}{=}\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{a})-\partial^2_{x^kx^j}f(\boldsymbol{p}_{s,t})\,\bigr|+\bigl|\,\partial^2_{x^jx^k}f(\boldsymbol{q}_{st})-\partial^2f_{x^jx^k}(\boldsymbol{a})\,\bigr|$$

(use (13.4.8a, 13.4.8b,13.4.9))

$$<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

This shows that

$$\bigl|\,\partial^2_{x^kx^j}f(\boldsymbol{a})-\partial^2f_{x^jx^k}(\boldsymbol{a})\,\bigr|<\varepsilon,\ \ \forall\varepsilon>0.$$

This proves (13.4.2). □

Example 13.4.2 (Gradient vector fields again)**.** Suppose that $U \subset \mathbb{R}^n$ is an open set and $\boldsymbol{V} : U \to \mathbb{R}^n$ is a C^1-vector field

$$\boldsymbol{V}(x^1, \dots, x^n) = \left[\begin{array}{c} V^1(x^1, \dots, x^n) \\ \vdots \\ V^n(x^1, \dots, x^n) \end{array} \right].$$

Let us show that

$$\boxed{\boldsymbol{V} \text{ is a gradient vector field } \Rightarrow \partial_{x^j} V^i = \partial_{x^i} V^j, \ \ \forall \boldsymbol{x} \in U, \ \ i \neq j}. \tag{13.4.10}$$

Indeed, if $\boldsymbol{V}$ is the gradient of some function $f : U \to \mathbb{R}$, then

$$V^i = \partial_{x^i} f, \ \ \forall i.$$

In particular this shows that the function f is C^2 since its partial derivatives are C^1. We deduce

$$\partial_{x^j} V^i = \partial_{x^j}(\partial_{x^i} f) = \partial_{x^i}(\partial_{x^j} f) = \partial_{x^i} V^j.$$

For example, if $\boldsymbol{V}(x, y)$ is a *gradient* vector field on $\mathbb{R}^2$

$$\boldsymbol{V}(x, y) = \left[\begin{array}{c} P(x, y) \\ Q(x, y) \end{array} \right] = P(x, y)\boldsymbol{i} + Q(x, y)\boldsymbol{j},$$

then

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}.$$

Similarly, if $\boldsymbol{V}(x, y, z)$ is a *gradient* vector field on $\mathbb{R}^3$

$$\boldsymbol{V}(x, y, z) = \left[\begin{array}{c} P(x, y, z) \\ Q(x, y, z) \\ R(x, y, z) \end{array} \right] = P(x, y, z)\boldsymbol{i} + Q(x, y, z)\boldsymbol{j} + R(x, y, z)\boldsymbol{k},$$

then

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \ \ \frac{\partial R}{\partial y} = \frac{\partial Q}{\partial z}, \ \ \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}.$$

The converse of (13.4.10) is not true. More precisely, there exist open sets $U \subset \mathbb{R}^n$ and C^1 vector fields $\boldsymbol{V} : U \to \mathbb{R}^n$ satisfying (13.4.10) yet they are not gradient vector fields.

A famous example is the vector field

$$\Theta : \mathbb{R}^2 \backslash \{0\} \to \mathbb{R}^2, \ \ \Theta(x, y) = \left[\begin{array}{c} P(x, y) \\ Q(x, y) \end{array} \right] := \left[\begin{array}{c} -\frac{y}{x^2+y^2} \\ \\ \frac{x}{x^2+y^2} \end{array} \right]$$

Indeed

$$\partial_y P = -\frac{1}{x^2 + y^2} + \frac{2y^2}{(x^2 + y^2)^2} = \frac{-(x^2 + y^2) + 2y^2}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2}.$$

$$\partial_x Q = \frac{1}{x^2 + y^2} - \frac{2x^2}{(x^2 + y^2)^2} = \frac{(x^2 + y^2) - 2x^2}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2}.$$

The reason why Θ *is not* a gradient vector field is rather subtle and can be properly explained once we introduce the concept of integration along paths. What is more surprising,

H. Poincaré proved that if $U \subset \mathbb{R}^n$ is an open *convex* set and $\boldsymbol{V} \to \mathbb{R}^n$ is a C^1 vector field satisfying (13.4.10), then $\boldsymbol{V}$ *is* a gradient vector field. Thus, for any open convex subset $C \subset \mathbb{R}^2 \backslash \{\boldsymbol{0}\}$ there exists a C^2 function $f_C : C \to \mathbb{R}$ such that

$$\Theta(\boldsymbol{x}) = \nabla f_C(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in C.$$

We cannot however find a function $f : \mathbb{R}^2 \backslash \{\boldsymbol{0}\} \to \mathbb{R}$ such that $\Theta = \nabla f$! □

Example 13.4.3. Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is a C^3 function of two variables x, y. Then $\partial_y f$ is a C^2 function and we have

$$\boxed{\partial^3_{xyy} f} = \partial^2_{xy}\big(\partial_y f\big) = \partial^2_{yx}\big(\partial_y f\big) = \boxed{\partial^3_{yxy} f} = \partial_y\big(\partial^2_{xy} f\big) = \partial_y\big(\partial^2_{yx} f\big) = \boxed{\partial^3_{yyx} f}.$$

If additionally f is C^4, then a similar argument shows

$$\partial^4_{xxyy} f = \partial^4_{xyxy} f = \partial^4_{yxxy} f = \partial^4_{yxyx} f = \partial^4_{yyxx} f = \partial^4_{xyyx} f.$$

It is now time to introduce a more convenient notation. Fix $n \in \mathbb{N}$. A *multi-index* of dimension n is an n-tuple

$$\alpha = (\alpha_1, \dots, \alpha_n), \;\; \alpha_1, \dots, \alpha_n \in \mathbb{Z}_{\geqslant 0}.$$

The *size* of the multi-index α is the nonnegative integer

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

Suppose now that $m \in \mathbb{N}$, $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is a C^m-function. Given $k \leqslant m$ we define

$$\partial^k_{x^1} f := \partial^k_{x^1 \cdots x^1} f, \dots, \partial^k_{x^n} f := \partial^k_{x^n \cdots x^n} f.$$

Thus, instead of $\partial^2_{x^1 x^1}$ we will write $\partial^2_{x^1} f$. We define $\partial^0_{x^1} f := f$.

For any multi-index α of dimension n and size $|\alpha| \leqslant m$ we set

$$\partial^\alpha_x f := \partial^{\alpha_1}_{x^1} \partial^{\alpha_2}_{x^2} \cdots \partial^{\alpha_n}_{x^n} f.$$

13.5. Exercises

Exercise 13.1. Let $m, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$, $\boldsymbol{x}_0 \in U$ and $\boldsymbol{F} : U \to \mathbb{R}^m$ a map. Assume U is open. Prove that the following statements are equivalent.

(i) The map $\boldsymbol{F}$ is Fréchet differentiable at $\boldsymbol{x}_0$.

(ii) There exists a linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$ with the following property:

$$\forall \varepsilon > 0,\ \ \exists \delta = \delta(\varepsilon) > 0 :\ \ \forall \boldsymbol{h} \in \mathbb{R}^n,\ \ \|\boldsymbol{h}\| < \delta \Rightarrow \big\|\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}\big\| \leqslant \varepsilon\|\boldsymbol{h}\|.$$

(iii) There exists a linear operator $L : \mathbb{R}^n \to \mathbb{R}^m$, a number $r > 0$ such that $B_r(\boldsymbol{x}_0) \subset U$ and a function $\varphi : [0, r) \to [0, \infty)$ with the following properties

$$\big\|\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}\big\| \leqslant \varphi\big(\,\|\boldsymbol{h}\|\,\big)\|\boldsymbol{h}\|.$$

$$\lim_{t \searrow 0} \varphi(t) = 0 = \varphi(0).$$

Hint. For (ii) ⇒ (iii) use

$$\varphi(t) := \sup_{\|\boldsymbol{h}\|=t} \frac{1}{\|\boldsymbol{h}\|}\big\|\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}_0) - L\boldsymbol{h}\big\|,\ \ t > 0.$$

□

Exercise 13.2. Consider the map $\boldsymbol{F} : \mathbb{R}^3 \to \mathbb{R}^2$ given by

$$\boldsymbol{F}(x, y, z) = \begin{bmatrix} x^3 + y^3 + z^3 \\ xyz \end{bmatrix}.$$

(i) Show that $\boldsymbol{F}$ is differentiable at any point $(x_0, y_0, z_0) \in \mathbb{R}^3$.

(ii) Find the Jacobian matrix of $\boldsymbol{F}$ at the point $(x_0, y_0, z_0) = (1, 1, 1)$.

□

Exercise 13.3. Compute the Jacobian matrices of the maps $\boldsymbol{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\boldsymbol{G}, \boldsymbol{H} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\begin{bmatrix} r \\ \theta \end{bmatrix} \overset{\boldsymbol{F}}{\mapsto} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix},$$

$$\begin{bmatrix} \rho \\ \theta \\ \varphi \end{bmatrix} \overset{\boldsymbol{G}}{\mapsto} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \rho\sin\varphi\cos\theta \\ \rho\sin\varphi\sin\theta \\ \rho\cos\varphi \end{bmatrix},\quad \begin{bmatrix} r \\ \theta \\ z \end{bmatrix} \overset{\boldsymbol{H}}{\mapsto} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ z \end{bmatrix}.$$ □

Exercise 13.4. Show that the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = 2^{xy}$ is C^1 and then find its linear approximation at the point $(x_0, y_0) = (1, 1)$.

Hint. Use Example 13.2.12 as inspiration. □

Exercise 13.5. Let $n \in \mathbb{N}$ and suppose that A is a symmetric $n \times n$ matrix. Define

$$q_A : \mathbb{R}^n \to \mathbb{R},\ \ q_A(\boldsymbol{x}) = \frac{1}{2}\langle A\boldsymbol{x}, \boldsymbol{x}\rangle,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Prove that

$$\nabla q_A(\boldsymbol{x}) = A\boldsymbol{x}, \ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Hint. You need to use the results in Exercise 11.24. □

Exercise 13.6. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *homogeneous of degree* 1, if

$$f(t\boldsymbol{x}) = tf(\boldsymbol{x}), \ \ \forall t \in \mathbb{R}, \ \ \boldsymbol{x} \in \mathbb{R}^n.$$

(i) Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is homogeneous of degree 1, then for any $\boldsymbol{v} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$, the function f is differentiable along $\boldsymbol{v}$ at $\boldsymbol{0}$.

(ii) Show that the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \ \ f(x,y) = \begin{cases} \frac{x^3-y^3}{x^2+y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0). \end{cases}$$

is homogeneous of degree 1, it is continuous at $\boldsymbol{0}$, but it is not Fréchet differentiable at $\boldsymbol{0}$.

(iii) Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is homogeneous of degree 1 and Fréchet differentiable at $\boldsymbol{0}$, then f is linear.

□

Exercise 13.7. Let $m, n \in \mathbb{N}$. Suppose that U is an open subset of $\mathbb{R}^n$, $I \subset \mathbb{R}$ is an open interval,$\gamma : I \to U$ is a C^1-path and $\boldsymbol{F} : U \to \mathbb{R}^m$ is a C^1-map. Let $\boldsymbol{\omega} : I \to \mathbb{R}^m$ denote the C^1-path $\boldsymbol{\omega}(t) = \boldsymbol{F}\big(\gamma(t)\big)$. Prove that

$$\dot{\boldsymbol{\omega}}(t) = d\boldsymbol{F}\big(\gamma(t)\big)\dot{\gamma}(t), \ \ \forall t \in I.$$ □

Exercise 13.8. Let $a, b \in \mathbb{R}$, $a < b$ and suppose that $\boldsymbol{\alpha}, \boldsymbol{\beta} : (a,b) \to \mathbb{R}^3$ are two C^1-paths. Prove that

$$\frac{d}{dt}\Big(\boldsymbol{\alpha}(t) \times \boldsymbol{\beta}(t)\Big) = \dot{\boldsymbol{\alpha}}(t) \times \boldsymbol{\beta}(t) + \boldsymbol{\alpha}(t) \times \dot{\boldsymbol{\beta}}(t).$$

Hint. Use (11.2.6). □

Exercise 13.9. Let $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ and suppose that $\boldsymbol{\alpha}, \boldsymbol{\beta} : (-1,1) \to \mathbb{R}^n$ are two differentiable paths.

(i) Show that

$$\frac{d}{dt}\big\langle \boldsymbol{\alpha}(t), \boldsymbol{\beta}(t) \big\rangle = \big\langle \dot{\boldsymbol{\alpha}}(t), \boldsymbol{\beta}(t) \big\rangle + \big\langle \boldsymbol{\alpha}(t), \dot{\boldsymbol{\beta}}(t) \big\rangle, \ \ \forall t \in (-1,1).$$

(ii) Compute the gradient ∇f. **Hint.** Compare with Exercise 13.5.

(iii) Compute $\frac{d}{dt}\|\boldsymbol{\alpha}(t)\|^2$.

(iv) Show that the function $t \mapsto \|\boldsymbol{\alpha}(t)\|$ is constant if and only if $\boldsymbol{\alpha}(t) \perp \dot{\boldsymbol{\alpha}}(t)$, $\forall t \in (-1,1)$.

(v) What can you say about the motion described by the path $\boldsymbol{\alpha}(t)$ when $\boldsymbol{\alpha}(t) \perp \dot{\boldsymbol{\alpha}}(t)$, $\forall t$?

□

Exercise 13.10. Let $f : \mathbb{R}^n \backslash \{\mathbf{0}\} \to \mathbb{R}$ be a function that is positively homogeneous of degree k, i.e.,

$$f(t\boldsymbol{x}) = t^k f(\boldsymbol{x}), \quad \forall t > 0, \quad \boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}.$$

Show that if f is differentiable, then the partial derivatives $\frac{\partial f}{\partial x^i}$, $i = 1, \dots, n$, are positively homogeneous of degree $k - 1$. □

Exercise 13.11. Suppose that the path

$$\boldsymbol{\gamma} : \mathbb{R} \to \mathbb{R}^2, \quad \boldsymbol{\gamma}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \in \mathbb{R}^2$$

is an integral curve of the vector field $\boldsymbol{V}$ defined in (13.3.18).

(i) Prove that $\|\boldsymbol{\gamma}(t)\| = \|\boldsymbol{\gamma}(0)\|$, $\forall t$.

(ii) Deduce from the above that $\dot{x}(t)^2 + x(t)^2 = \dot{x}(0)^2 + x(0)^2$, $\forall t$.

(iii) Prove that $\ddot{x} = -x$, where $\ddot{f}$ denotes the second order time derivative of a function f.

(iv) Given that $\boldsymbol{\gamma}(0) = (1, 0)$ determine $\boldsymbol{\gamma}(t)$.

Hint. For (i)-(iii) use the differential equations (13.3.17). (iv) Compare with Exercise 7.16. □

Exercise 13.12. Prove that the function $r : \mathbb{R}^n \backslash \{\mathbf{0}\} \to \mathbb{R}$, $r(\boldsymbol{x}) = \|\boldsymbol{x}\|$, is C^1 and then describe its differential. □

Exercise 13.13. Let $n \in \mathbb{N}$. Fix a C^1-function $U : \mathbb{R}^n \backslash \{\mathbf{0}\} \to \mathbb{R}$. Suppose that I is an open interval of the real axis and

$$\boldsymbol{\gamma} : I \to \mathbb{R}^n \backslash \{\mathbf{0}\}, \quad t \mapsto \boldsymbol{\gamma}(t) = [x^1(t), \dots, x^n(t)]^\top,$$

is a C^2-path satisfying Newton's (2nd Law of Dynamics) differential equations

$$\ddot{\boldsymbol{\gamma}}(t) = -\nabla U\big(\boldsymbol{\gamma}(t)\big), \quad \forall t \in I.$$

(i) (*Conservation of energy*) Prove that the function $E : I \to \mathbb{R}$

$$E(t) = \frac{1}{2}\|\dot{\boldsymbol{\gamma}}(t)\|^2 + U\big(\boldsymbol{\gamma}(t)\big),$$

is constant.

(ii) (*Conservation of momenta*) Suppose that there exists a C^1-function $f : (0, \infty) \to \mathbb{R}$ such that

$$U(\boldsymbol{x}) = f(\|\boldsymbol{x}\|), \quad \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}.$$

Prove that for, any $1 \leqslant k < \ell \leqslant n$, the function $P^{k\ell} : I \to \mathbb{R}$

$$P^{k\ell}(t) = \dot{x}^k(t)x^\ell(t) - \dot{x}^\ell(t)x^k(t)$$

is constant.

□

Exercise 13.14. Let $k, n \in \mathbb{N}$.

(i) Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ and $u : \mathbb{R} \to \mathbb{R}$ are C^k-functions, then so is their composition $u \circ f : \mathbb{R}^n \to \mathbb{R}$.

(ii) Let $n \in \mathbb{N}$ and $r > 0$. Prove that for any $0 < r < R$ there exists a *nonzero* function $f \in C^\infty(\mathbb{R}^n)$ such that $f(\boldsymbol{x}) = 1$ if $\|\boldsymbol{x}\| \leqslant r$ and $f(\boldsymbol{x}) = 0$, if $\|\boldsymbol{x}\| \geqslant R$.

(iii) Let $n \in \mathbb{N}$. Suppose that $K \subset \mathbb{R}^n$ is a compact set and $\mathcal{U}$ is an open cover of K. Prove that there exist compactly supported smooth functions

$$\chi_1, \dots, \chi_\ell : \mathbb{R}^n \to [0, \infty)$$

with the following properties.

- For any $i = 1, \dots, \ell$ there exists an open set $U = U_i$ in the family $\mathcal{U}$ such that $\operatorname{supp} \chi_i \subset U_i$.
- $\chi_1(\boldsymbol{x}) + \cdots + \chi_\ell(\boldsymbol{x}) = 1$, $\forall \boldsymbol{x} \in K$.

Hint. (i) Argue by induction on k. (ii) Use the result proved in Exercise 7.8 to construct a smooth function $u : \mathbb{R} \to [0, \infty)$ such that $u(s) = 1$ if $s \leqslant 0$ and $u(s) = 0$ if $s \geqslant 1$. Then, for $a < b$, define

$$u_{a,b} : \mathbb{R} \to [0, \infty), \quad u_{a,b}(t) = u\left(\frac{t-a}{b-a}\right)$$

and show that $u_{a,b}$ is smooth and satisfies $u_{a,b}(t) = 1$ if $t \leqslant a$ and $u_{a,b}(t) = 0$ if $t > b$. Finally, set $f(\boldsymbol{x}) = u_{a,b}(\|\boldsymbol{x}\|^2)$ with $a = r^2$, $b = R^2$ and then show that f will do the trick. (iii) Use (ii) and imitate the proof of Theorem 12.4.7.

□

Exercise 13.15. Let $n \in \mathbb{N}$. For any open set $\mathcal{O} \subset \mathbb{R}^n$ we define the *Laplacian* to be the map

$$\Delta : C^2(\mathcal{O}) \to C(\mathcal{O}), \quad (\Delta f)(\boldsymbol{x}) = \sum_{k=1}^{n} \partial^2_{x^k} f(\boldsymbol{x}). \tag{13.5.1}$$

(i) Show that, $\forall f, g \in C^2(\mathcal{O})$, we have

$$\Delta(f + g) = \Delta f + \Delta g,$$

$$\Delta(fg) = f\Delta g + 2\langle \nabla f, \nabla g\rangle + g\Delta f$$

(ii) Show

$$\Delta\|\boldsymbol{x}\|^p = p(p + n - 2)\|\boldsymbol{x}\|^{p-2}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n\backslash\{\boldsymbol{0}\}, \quad p \in \mathbb{R}.$$

□

Exercise 13.16. Let $n \in \mathbb{N}$, $n \geqslant 2$. Consider the function $U : \mathbb{R}^n\backslash\{\boldsymbol{0}\} \to \mathbb{R}$

$$U(\boldsymbol{x}) = \begin{cases} \log\|\boldsymbol{x}\|, & n = 2, \\ \frac{1}{\|\boldsymbol{x}\|^{n-2}}, & n > 2. \end{cases}$$

Compute $\Delta U(\boldsymbol{x})$, where Δ is the Laplacian defined as in (13.5.1). □

Exercise 13.17. Let $n \in \mathbb{N}$ and consider the function $K : (0, \infty) \times \mathbb{R}^n \to \mathbb{R}$ given by

$$K(t, \boldsymbol{x}) = t^{-n/2} e^{-\frac{\|\boldsymbol{x}\|^2}{4t}}.$$

Compute

$$\partial_t K - \Delta_{\boldsymbol{x}} K = \partial_t K - \Big(\partial^2_{x^1} K + \cdots + \partial^2_{x^n} K \Big).$$

□

Exercise 13.18. Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are C^2 functions. Define

$$w : \mathbb{R}^2 \to \mathbb{R}, \;\; w(t, x) = f(x + t) + g(x - t).$$

Compute

$$\partial_t^2 w - \partial_x^2 w.$$

□

Exercise 13.19. Show that the vector field

$$\boldsymbol{V} : \mathbb{R}^2 \to \mathbb{R}^2, \;\; \boldsymbol{V}(x, y) = \begin{bmatrix} -y \\ x \end{bmatrix}$$

is not a gradient vector field.

Hint. Have a look at Example 13.4.2. □

Exercise 13.20. Let $n \in \mathbb{N}$ and suppose that A is an $n \times n$ matrix.

(i) Prove that for any $\boldsymbol{x} \in \mathbb{R}^n$ the series

$$\sum_{k=0}^{\infty} \frac{1}{k!} A^k \boldsymbol{x} = \boldsymbol{x} + A\boldsymbol{x} + \frac{1}{2!} A^2 \boldsymbol{x} + \frac{1}{3!} A^3 \boldsymbol{x} + \cdots$$

is absolutely convergent.

(ii) Prove that for any $\boldsymbol{x} \in \mathbb{R}^n$ the path

$$\boldsymbol{\gamma} : \mathbb{R} \to \mathbb{R}^n, \;\; \boldsymbol{\gamma}(t) = \sum_{k=0}^{\infty} \frac{1}{k!} (tA)^k \boldsymbol{x}$$

is differentiable and

$$\dot{\gamma}(t) = A\boldsymbol{\gamma}(t), \;\; \forall t \in \mathbb{R}.$$

(iii) Compute $\boldsymbol{\gamma}(t)$ when $n = 2$,

$$\boldsymbol{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} \text{ and } A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

13.6. Exercises for extra credit

Chapter 14

Applications of multi-variable differential calculus

We present below a few of the most frequently encountered applications of multi-dimensional differential calculus.

14.1. Taylor formula

Just like functions of one real variable, the differentiable functions of several variables can be well approximated by certain explicit polynomials. We present in this section two such approximation formulæ that are used frequently in applications. We refer to [13, §2.8] for more general results.

Fix $n \in \mathbb{N}$, an open set $U \subset \mathbb{R}^n$ and a point $\boldsymbol{x}_0 \in U$. Then there exists $r_0 > 0$ such that $\overline{B_{r_0}(\boldsymbol{x}_0)} \subset U$. Suppose that $f : U \to \mathbb{R}$ is a differentiable function.

Theorem 14.1.1 (Multidimensional Taylor formula)**.** *(a) If the function f is C^2, then for any $\boldsymbol{h} = (h^1, \dots, h^n) \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| < r_0$ we have*

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + \sum_{i=1}^{n} \partial_{x^i} f(\boldsymbol{x}_0) h^i + R_1(\boldsymbol{x}_0, \boldsymbol{h}), \tag{14.1.1}$$

where the remainder $R_1(\boldsymbol{x}_0, \boldsymbol{h})$ is described by the integral formula

$$R_1(\boldsymbol{x}_0, \boldsymbol{h}) = \int_0^1 (1-t)\rho_2(\boldsymbol{x}_0, \boldsymbol{h}, t)dt, \;\; \rho_2(\boldsymbol{x}_0, \boldsymbol{h}, t) = \sum_{i,j=1}^{n} \partial^2_{x^i x^j} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i h^j. \tag{14.1.2}$$

Moreover, there exists a constant $C > 0$, $\boxed{\text{independent of } \boldsymbol{h}}$, *such that*

$$\big|\, R_1(\boldsymbol{x}_0, \boldsymbol{h}) \,\big| \leqslant C\|\boldsymbol{h}\|^2, \;\; \forall \|\boldsymbol{h}\| < r_0. \tag{14.1.3}$$

(b) If the function f is C^3, then for any $\boldsymbol{h} = (h^1, \dots, h^n) \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| < r_0$ we have

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + \sum_{i=1}^n \partial_{x^i} f(\boldsymbol{x}_0) h^i + \frac{1}{2} \sum_{i,j=1}^n \partial^2_{x^i x^j} f(\boldsymbol{x}_0) h^i h^j + R_2(\boldsymbol{x}_0, \boldsymbol{h}), \tag{14.1.4}$$

where the remainder $R_2(\boldsymbol{x}_0, \boldsymbol{h})$ is described by the integral formula

$$\begin{gathered} R_2(\boldsymbol{x}_0, \boldsymbol{h}) = \frac{1}{2!} \int_0^1 (1-t)^2 \rho_3(\boldsymbol{x}_0, \boldsymbol{h}, t) dt, \\ \rho_3(\boldsymbol{x}_0, \boldsymbol{h}, t) = \sum_{i,j,k=1}^n \partial^3_{x^i x^j x^k} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i h^j h^k. \end{gathered} \tag{14.1.5}$$

Moreover, there exists a constant $C > 0$, $\boxed{\text{independent of } \boldsymbol{h}}$, *such that*

$$\big|\, R_2(\boldsymbol{x}_0, \boldsymbol{h}) \,\big| \leqslant C\|\boldsymbol{h}\|^3, \;\; \forall \|\boldsymbol{h}\| < r_0. \tag{14.1.6}$$

Proof. We prove only (b). The case (a) is similar and involves simpler computations. Fix $\boldsymbol{h} \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| < r_0$. Consider the C^3-function

$$g : [-1, 1] \to \mathbb{R}, \;\; g(t) = f(\boldsymbol{x}_0 + t\boldsymbol{h}).$$

Using the one dimensional Taylor formula with integral remainder, Proposition 9.6.9, we deduce

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(0) + R_2, \;\; R_2 = \frac{1}{2!} \int_0^1 g^{(3)}(t)(1-t)^2 dt. \tag{14.1.7}$$

Using the chain rule (13.3.12) repeatedly we deduce

$$g'(t) = \sum_{i=1}^n \partial_{x^i} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i, \;\; g''(t) = \sum_{i,j=1}^n \partial^2_{x^i x^j} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i h^j,$$

$$g^{(3)}(t) = \sum_{i,j,k=1}^n \partial^3_{x^i x^j x^k} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i h^j h^k.$$

Using these equalities in (14.1.7) we obtain (14.1.4) and (14.1.5). It remains to prove (14.1.6).

Observe that $|h^i| \leqslant \|\boldsymbol{h}\|$ for any $i = 1, \dots, n$. Hence

$$|\rho_3(\boldsymbol{x}_0, \boldsymbol{h}, t)| \leqslant \sum_{i,j,k=1}^n \big|\, \partial^3_{x^i x^j x^k} f(\boldsymbol{x}_0 + t\boldsymbol{h}) h^i h^j h^k \,\big| \leqslant \|\boldsymbol{h}\|^3 \sum_{i,j,k=1}^n \big|\, \partial^3_{x^i x^j x^k} f(\boldsymbol{x}_0 + t\boldsymbol{h}) \,\big|.$$

For each $i, j, k = 1, \dots, n$ we set

$$M_{i,j,k} = \sup_{\boldsymbol{x} \in B_{r_0}(\boldsymbol{x}_0)} \big|\, \partial^3_{x^i x^j x^k} f(\boldsymbol{x}) \,\big|.$$

The quantity $M_{i,j,k}$ is finite since the function $\partial^3_{x^ix^jx^k} f(\boldsymbol{x})$ is continuous on the *compact* set $\overline{B_{r_0}(\boldsymbol{x}_0)}$. We set

$$M := \max_{i,j,k} M_{i,j,k}.$$

Clearly, the number M is independent of $\boldsymbol{h}$. We deduce that for any $\|\boldsymbol{h}\| < r_0$ we have

$$|\rho_3(\boldsymbol{x}_0, \boldsymbol{h}, t)| \leqslant \|\boldsymbol{h}\|^3 \sum_{i,jk=1}^{n} M_{i,j,k} \leqslant \|\boldsymbol{h}\|^3 \sum_{i,j,k=1}^{n} M = Mn^3\|\boldsymbol{h}\|^3.$$

Hence

$$|R_2(\boldsymbol{x}_0, \boldsymbol{h})| \leqslant \frac{1}{2}\int_0^1 (1-t)^2\, |\rho_3(\boldsymbol{x}_0, \boldsymbol{h}, t)| dt \leqslant \frac{Mn^3}{2}\|\boldsymbol{h}\|^3.$$

□

Definition 14.1.2. The $n \times n$ matrix $\boldsymbol{H}(f, \boldsymbol{x}_0)$ with entries

$$\boldsymbol{H}(f, \boldsymbol{x}_0)_{ij} = \partial^2_{x^ix^j} f(\boldsymbol{x}_0), \;\; 1 \leqslant i, j \leqslant n,$$

is called the *Hessian* of f at $\boldsymbol{x}_0$.[1] □

Since partial derivatives commute, we see that the Hessian is a *symmetric* matrix, i.e.,

$$\partial^2_{x^ix^j} f(\boldsymbol{x}_0) = \partial^2_{x^jx^i} f(\boldsymbol{x}_0).$$

The matrix $\boldsymbol{H}(f, \boldsymbol{x}_0)$ defines a linear operator $\mathbb{R}^n \to \mathbb{R}^n$ given by

$$\boldsymbol{H}(f, \boldsymbol{x}_0)\boldsymbol{h} = \Big(\sum_{j=1}^{n} \boldsymbol{H}(f, \boldsymbol{x}_0)_{1j} h^j\Big)\boldsymbol{e}_1 + \cdots + \Big(\sum_{j=1}^{n} \boldsymbol{H}(f, \boldsymbol{x}_0)_{nj} h^j\Big)\boldsymbol{e}_n$$

$$= \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} \boldsymbol{H}(f, \boldsymbol{x}_0)_{ij} h^j\Big)\boldsymbol{e}_i, \;\; \forall \boldsymbol{h} \in \mathbb{R}^n.$$

We deduce that

$$\sum_{i,j=1}^{n} \partial^2_{x^ix^j} f(\boldsymbol{x}_0) h^i h^j = \big\langle\, \boldsymbol{h}, \boldsymbol{H}(f, \boldsymbol{x}_0)\boldsymbol{h} \,\big\rangle. \tag{14.1.8}$$

We can rewrite the equality (14.1.4) in the more compact form

$$\boxed{f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + \big\langle\, \nabla f(\boldsymbol{x}_0), \boldsymbol{h} \,\big\rangle + \frac{1}{2}\big\langle\, \boldsymbol{h}, \boldsymbol{H}(f, \boldsymbol{x}_0)\boldsymbol{h} \,\big\rangle + R_2(\boldsymbol{x}_0, \boldsymbol{h})}. \tag{14.1.9}$$

[1] Note that when describing the Hessian matrix both indices are *subscripts*. This differs from the way we described the matrix associated to an operator where one index is a superscript, the other is a subscript. This discrepancy is a reflection of the fact that the Hessian of a function is intrinsically a different beast than a linear operator.

Example 14.1.3. Consider the function

$$f:\mathbb{R}^2\to\mathbb{R},\;\; f(x,y)=3x^2+4xy+5y^2.$$

Then

$$\partial_x^2 f=6,\;\; \partial_{xy}^2 f=4,\;\; \partial_y^2 f=10.$$

The Hessian of f at $\mathbf{0}$ is the symmetric 2×2-matrix

$$\boldsymbol{H}(f,\mathbf{0})=\left[\begin{array}{cc}6 & 4\\ 4 & 10\end{array}\right].$$ □

14.2. Extrema of functions of several variables

Fix a natural number n.

Definition 14.2.1. Let $X\subset\mathbb{R}^n$ and $\boldsymbol{x}_0\in X$. Fix a function $f:X\to\mathbb{R}$.

(i) The point $\boldsymbol{x}_0$ is said to be a *local minimum* of the function if there exists $r>0$ with the following property:

$$\forall\boldsymbol{x}\in X,\;\; \operatorname{dist}(\boldsymbol{x},\boldsymbol{x}_0)<r\Rightarrow f(\boldsymbol{x}_0)\leqslant f(\boldsymbol{x}),$$

(ii) The point $\boldsymbol{x}_0$ is said to be a *local maximum* of the function f if there exists $r>0$ with the following property,

$$\forall\boldsymbol{x}\in X\;\; \operatorname{dist}(\boldsymbol{x},\boldsymbol{x}_0)<r\Rightarrow f(\boldsymbol{x}_0)\geqslant f(\boldsymbol{x}).$$

(iii) The point $\boldsymbol{x}_0$ is said to be a *local extremum* of the function f if it is either a local minimum or a local maximum of f.

□

The one-dimensional Fermat Principle[2] (Theorem 7.4.2) has the following multi-dimensional counterpart.

Theorem 14.2.2 (Multidimensional Fermat Principle)**.** *Suppose that $U\subset\mathbb{R}^n$ is an $\underline{\text{open}}$ set and $f:U\to\mathbb{R}$ is a C^1-function. If $\boldsymbol{x}_0\in U$ is a local extremum of f, then $df(\boldsymbol{x}_0)=\mathbf{0}$, i.e.,*

$$\partial_{x^1}f(\boldsymbol{x}_0)=\cdots=\partial_{x^n}f(\boldsymbol{x}_0)=0.$$

Proof. Assume for simplicity that $\boldsymbol{x}_0$ is a local minimum of f. (When $\boldsymbol{x}_0$ is a local maximum of f, then it is a local minimum of $-f$.) Fix $r>0$ sufficiently small with the following properties.

- $B_r(\boldsymbol{x}_0)\subset U$.
- $f(\boldsymbol{x}_0)\leqslant f(\boldsymbol{x})$, $\forall\boldsymbol{x}\in B_r(\boldsymbol{x}_0)$.

[2] See this beautiful lecture by Richard Feynman `http://www.feynmanlectures.caltech.edu/II_19.html`

Fix a vector $\boldsymbol{h} \in \mathbb{R}^n$. For $\varepsilon > 0$ sufficiently small, the line segment $[\boldsymbol{x}_0 - \varepsilon\boldsymbol{h}, \boldsymbol{x}_0 + \varepsilon\boldsymbol{h}]$ is contained in the ball $B_r(\boldsymbol{x}_0)$. Consider now the function

$$g : [-\varepsilon, \varepsilon] \to \mathbb{R}, \;\; g(t) = f(\boldsymbol{x}_0 + t\boldsymbol{h}).$$

We can identify g with the restriction of f to the line segment $[\boldsymbol{x}_0 - \varepsilon\boldsymbol{h}, \boldsymbol{x}_0 + \varepsilon\boldsymbol{h}]$. Note that $g(0) = f(\boldsymbol{x}_0) \leqslant f(\boldsymbol{x}_0 + t\boldsymbol{h}) = g(t)$, $\forall t \in [-\varepsilon, \varepsilon]$. Thus 0 is a minimum point of g and the one-dimensional Fermat principle implies that $g'(0) = 0$. The chain rule (13.3.12) now implies

$$\big\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{h} \big\rangle = g'(0) = 0.$$

We have thus shown that $\big\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{h} \big\rangle = 0$, for any vector $\boldsymbol{h} \in \mathbb{R}^n$. If we choose $\boldsymbol{h} = \nabla f(\boldsymbol{x}_0)$, then we deduce

$$\|\nabla f(\boldsymbol{x}_0)\|^2 = \big\langle \nabla f(\boldsymbol{x}_0), \nabla f(\boldsymbol{x}_0) \big\rangle = 0.$$

□

Definition 14.2.3. Let $U \subset \mathbb{R}^n$ be an open set. A *critical point* of a differentiable function $f : U \to \mathbb{R}$ is a point $\boldsymbol{x}_0 \in U$ such that $df(\boldsymbol{x}_0) = 0$. □

We can rephrase Theorem 14.2.2 as follows.

If $U \subset \mathbb{R}^n$ is an open set, and $\boldsymbol{x}_0 \in U$ is a local extremum of a C^1-function $f : U \to \mathbb{R}$, then $\boldsymbol{x}_0$ must be a critical point of f.

We know now that the local extrema of a C^1-function $f : U \to \mathbb{R}$, if any, are located among the critical points of f. We want to address a sort of converse. Suppose that $\boldsymbol{x}_0 \in U$ is a critical point. Is there any way of deciding whether $\boldsymbol{x}_0$ is a local min, max or neither?

To answer this question we need to introduce some more terminology.

Definition 14.2.4. Suppose that A is a *symmetric* $n \times n$ matrix A. We denote by a_{ij} the entry located on the i-th row and j-th column.

(i) The *quadratic function* associated to A is the function $Q_A : \mathbb{R}^n \to \mathbb{R}$ given by

$$Q_A(\boldsymbol{h}) = \langle \boldsymbol{h}, A\boldsymbol{h} \rangle = \sum_{i,j=1}^{n} a_{ij} h^i h^j.$$

(ii) The matrix A is called *positive definite* if

$$Q_A(\boldsymbol{h}) > 0, \;\; \forall \boldsymbol{h} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}.$$

(iii) The matrix A is called *negative definite* if

$$Q_A(\boldsymbol{h}) < 0, \;\; \forall \boldsymbol{h} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}.$$

(iv) The matrix A is called *indefinite* if there exist $\boldsymbol{h}_0, \boldsymbol{h}_1 \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ such that

$$Q_A(\boldsymbol{h}_0) < 0 < Q_A(\boldsymbol{h}_1).$$

□

Let us observe that the quadratic function associated to a symmetric $n \times n$ matrix A is homogeneous of degree 2, i.e.,

$$Q_A(t\boldsymbol{h}) = t^2 Q_A(\boldsymbol{h}), \ \ \forall t \in \mathbb{R}, \ \ \boldsymbol{h} \in \mathbb{R}^n. \tag{14.2.1}$$

Example 14.2.5. Suppose that $n = 2$,

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \ \ \boldsymbol{h} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Then

$$A\boldsymbol{h} = \begin{bmatrix} ax + by \\ bx + cy \end{bmatrix}, \ \ Q_A(\boldsymbol{h}) = \langle \boldsymbol{h}, A\boldsymbol{h} \rangle = x(ax + by) + y(bx + cy) = ax^2 + 2bxy + cy^2. \ \square$$

Theorem 14.2.6. *Let $U \subset \mathbb{R}$ be an open set and $f : U \to \mathbb{R}$ a C^3-function. Suppose that $\boldsymbol{x}_0$ is a critical point of f. Denote by A the Hessian of f at $\boldsymbol{x}_0$, $A := \boldsymbol{H}(f, \boldsymbol{x}_0)$. Then the following hold.*

(i) *If A is positive definite, then $\boldsymbol{x}_0$ is a local minimum of f.*

(ii) *If A is negative definite, then $\boldsymbol{x}_0$ is a local maximum of f.*

(iii) *If A is indefinite, then $\boldsymbol{x}_0$ is not a local extremum of f.*

Proof. The above claims are immediate consequences of Taylor's formula (14.1.4). Fix $r > 0$ sufficiently small such that $\overline{B_r(\boldsymbol{x}_0)} \subset U$. According to (14.1.4) for any $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| < r$ we have

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + \frac{1}{2} Q_A(\boldsymbol{h}) + R_2(\boldsymbol{x}_0, \boldsymbol{h}). \tag{14.2.2}$$

Moreover, there exists $C > 0$ such that

$$|R_2(\boldsymbol{x}_0, \boldsymbol{h})| \leqslant C\|\boldsymbol{h}\|^3, \ \ \forall \|\boldsymbol{h}\| < r. \tag{14.2.3}$$

To prove (i) we need to use the following very useful technical fact whose proof is outlined in Exercise 14.5.

Lemma 14.2.7. *Suppose that A is a symmetric, positive definite matrix. Then there exists $m > 0$ such that*

$$Q_A(\boldsymbol{h}) \geqslant m\|\boldsymbol{h}\|^2, \ \ \forall \boldsymbol{h} \in \mathbb{R}^n. \qquad \square$$

Suppose now that $A = \boldsymbol{H}(f, \boldsymbol{x}_0)$ is positive definite. Choose a number $m > 0$ as in Lemma 14.2.7. From (14.2.2) and (14.2.3) we deduce

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) \geqslant f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{h}\|^2 - C\|\boldsymbol{h}\|^3 = f(\boldsymbol{x}_0) + \|\boldsymbol{h}\|^2 \Big(\frac{m}{2} - C\|\boldsymbol{h}\| \Big).$$

Choose $\varepsilon > 0$ smaller than both r and $\frac{m}{2C}$. Then, for any $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| < \varepsilon$ we have

$$\boldsymbol{x}_0 + \boldsymbol{h} \in B_\varepsilon(\boldsymbol{x}_0), \ \ \frac{m}{2} - C\|\boldsymbol{h}\| > 0.$$

Thus for any $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| < \varepsilon$ we have $f(\boldsymbol{x}_0 + \boldsymbol{h}) > f(\boldsymbol{x}_0)$. This proves that $\boldsymbol{x}_0$ is a local minimum of f.

The statement (ii) reduces to (i) by observing that the Hessian of $-f$ at $\boldsymbol{x}_0$ is $-A$ and it is positive definite. Thus $\boldsymbol{x}_0$ is a local minimum of $-f$, therefore a local maximum of f.

To prove (iii) choose vectors $\boldsymbol{h}_0, \boldsymbol{h}_1$ such that

$$Q_A(\boldsymbol{h}_0) < 0 < Q_A(\boldsymbol{h}_1).$$

For $t > 0$ sufficiently small we have $\boldsymbol{x}_0 + t\boldsymbol{h}_0, \boldsymbol{x}_0 + t\boldsymbol{h}_1 \in B_r(\boldsymbol{x}_0)$ and

$$f(\boldsymbol{x}_0 + t\boldsymbol{h}_0) = f(\boldsymbol{x}_0) + \frac{1}{2}Q_A(t\boldsymbol{h}_0) + R_2(\boldsymbol{x}_0, t\boldsymbol{h}_0) \overset{(14.2.1)}{=} f(\boldsymbol{x}_0) + \frac{t^2}{2}Q_A(\boldsymbol{h}_0) + R_2(\boldsymbol{x}_0, t\boldsymbol{h}_0)$$

$$\leqslant f(\boldsymbol{x}_0) + \frac{t^2}{2}Q_A(\boldsymbol{h}_0) + Ct^3\|\boldsymbol{h}_0\|^3 = f(\boldsymbol{x}_0) + \frac{t^2}{2}\underbrace{\Big(Q_A(\boldsymbol{h}_0) + 2tC\|\boldsymbol{h}_0\|\Big)}_{=:u(t)}.$$

Observe that

$$\lim_{t\to 0} u(t) = Q_A(\boldsymbol{h}_0) < 0$$

so $u(t)$ is negative for t sufficiently small. Thus, for all t sufficiently small we have

$$f(\boldsymbol{x}_0 + t\boldsymbol{h}_0) < f(\boldsymbol{x}_0),$$

so $\boldsymbol{x}_0$ cannot be a local minimum. Similarly

$$f(\boldsymbol{x}_0 + t\boldsymbol{h}_1) = f(\boldsymbol{x}_0) + \frac{1}{2}Q_A(t\boldsymbol{h}_1) + R_2(\boldsymbol{x}_0, t\boldsymbol{h}_1) = f(\boldsymbol{x}_0) + \frac{t^2}{2}Q_A(\boldsymbol{h}_1) + R_2(\boldsymbol{x}_0, t\boldsymbol{h}_1)$$

$$\geqslant f(\boldsymbol{x}_0) + \frac{t^2}{2}Q_A(\boldsymbol{h}_1) - Ct^3\|\boldsymbol{h}_1\| \overset{(14.2.1)}{=} f(\boldsymbol{x}_0) + \frac{t^2}{2}\underbrace{\Big(Q_A(\boldsymbol{h}_1) - 2Ct\|\boldsymbol{h}_1\|\Big)}_{=:v(t)}.$$

Observe that

$$\lim_{t\to 0} v(t) = Q_A(\boldsymbol{h}_1) > 0$$

so $v(t) > 0$ for all t sufficiently small. Hence, for all t sufficiently small we have

$$f(\boldsymbol{x}_0 + t\boldsymbol{h}_1) > f(\boldsymbol{x}_0)$$

so $\boldsymbol{x}_0$ cannot be a local maximum either.

□

Remark 14.2.8. Deciding when a symmetric matrix A is positive/negative definite or indefinite is a nontrivial task. All the known techniques rely on more linear algebra than we are prepared to assume at this point. It is known that all the eigenvalues of a real symmetric matrix are real. The matrix A is positive/negative definite if all its eigenvalues are positive/negative. The matrix A is indefinite if it admits both positive and negative eigenvalues.

If the dimension of the matrix A is small one can conceive faster ad-hoc methods of deciding if S is positive/negative definite. In Exercise 14.6 we describe a simple method of deciding when a 2×2 symmetric matrix is positive/negative definite. This is a special case of a theorem of J.J. Sylvester[3] [40, Chap. 7]. □

Example 14.2.9. Consider the function

$$f:(0,\infty)\times(0,\infty)\to\mathbb{R},\;\; f(x,y)=x^3y^2(6-x-y).$$

The critical points of f are found solving the system of equations $\partial_x f=\partial_y f=0$, i.e.,

$$\begin{cases} 3x^2y^2(6-x-y)-x^3y^2 &= \;\; 0\\ 2x^3y(6-x-y)-x^3y^2 &= 0. \end{cases} \tag{14.2.4}$$

The first equality in (14.2.4) can be rewritten as

$$x^2y^2\Big(3(6-x-y)-x\Big)=0.$$

Since $x,y>0$ we deduce

$$18-3x-3y-x=0\Rightarrow 4x+3y=18.$$

The second equality in (14.2.4) can be rewritten as

$$x^3y\Big(2(6-x-y)-y\Big)=0$$

and we conclude as above that

$$2x+3y=12.$$

Hence

$$2x=(4x+3y)-(2x+3y)=18-12=6\Rightarrow x=3.$$

Using this information in the equality $2x+3y=12$ we deduce $3y=6$ so $y=2$. Thus, the only critical point of f is $(3,2)$. Let us find the Hessian at this point. We have

$$\partial_x f=x^2y^2\big(3(6-x-y)-x\,\big)=x^2y^2\big(18-4x-3y\,\big),$$

$$\partial_y f=x^3y\Big(2(6-x-y)-y\Big)=x^3y\big(\,12-2x-3y\,\big),$$

$$\partial^2_{xx}f=2xy^2(18-4x-3y)-4x^2y^2,\;\; \partial^2_{xy}f=2x^2y(18-4x-3y)-3x^2y^2,$$

$$\partial^2_{yy}f=3x^2(12-2x-3y)-3x^3y.$$

Hence

$$\partial^2_{xx}f(3,2)=-4\cdot 3^2\cdot 2^2=-144,\;\; \partial^2_{yy}(3,2)=-3\cdot 3^3\cdot 2=-162,$$

$$\partial^2_{xy}f(3,2)=-3\cdot 3^2\cdot 2^2=-108.$$

Hence, the Hessian of f at $(3,2)$ is

$$A:=\begin{bmatrix} -144 & -108\\ -108 & -162\end{bmatrix}.$$

[3] J.J. Sylvester was an English mathematician. He made fundamental contributions to matrix theory, invariant theory, number theory, partition theory, and combinatorics. He played a leadership role in American mathematics in the later half of the 19th century as a professor at Johns Hopkins University and as founder of the American Journal of Mathematics. https://en.wikipedia.org/wiki/James_Joseph_Sylvester

To decide whether the matrix A is positive/negative definite we use the criterion in Exercise 14.6. Note that $-144 < 0$ and

$$\det A = (-144)(-162) - (-108)^2 = 144 \cdot 162 - (108)^2 = 11644 > 0.$$

Hence A is negative definite and thus the stationary point $(3, 2)$ is a local maximum. □

14.3. Diffeomorphisms and the inverse function theorem

We can now discuss a classical theorem that plays a key role in modern differential geometry/topology. The remainder of this chapter assumes familiarity with basic linear algebra concepts such as linear combinations, linear independence, rank and determinant of a matrix.

We begin by introducing a key concept.

Definition 14.3.1 (Diffeomorphisms)**.** Let $n \in \mathbb{N}$, $k \in \mathbb{N} \cup \{\infty\}$ and suppose that $U \subset \mathbb{R}^n$ is an open set. A map $\boldsymbol{F} : U \to \mathbb{R}^n$ is called a C^k-*diffeomorphism* if the following hold.

- The map $\boldsymbol{F}$ is injective and its range $\boldsymbol{F}(U)$ is also an open subset of $\mathbb{R}^n$.
- The inverse map $\boldsymbol{F}^{-1} : \boldsymbol{F}(U) \to U$ is also C^k.

□

Example 14.3.2. (a) Any invertible linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism.

(b) The bijective C^1-map $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ is *not* a diffeomorphism because its inverse is *not* differentiable at 0. □

Example 14.3.3. The map

$$\boldsymbol{F} : (0, \infty) \times (0, 2\pi) \to \mathbb{R}^2, \quad \boldsymbol{F}(r, \theta) = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$$

is a C^1-diffeomorphism. Indeed, it is a C^1 map. To see that it is injective observe that if

$$x = r\cos\theta, \quad y = r\sin\theta$$

then

$$x^2 + y^2 = r^2 \Rightarrow r = \sqrt{x^2 + y^2}.$$

Thus, $r > 0$ is uniquely determined by (x, y). Note that $(x/r, y/r)$ is a point on the unit circle, it is not equal to $(1, 0)$ and uniquely determines the angle θ; recall the trigonometric circle in Section 5.6.

This proves that $\boldsymbol{F}$ is injective and the range is the plane $\mathbb{R}^2$ with the nonnegative x-semiaxis removed. Hence the range is open. One can show directly that $\boldsymbol{F}^{-1}$ is C^1, but this is a rather tedious job. Fortunately there is a faster alternate approach that relies on the main theorem of this section, namely, the inverse function theorem. We will present this approach after we discuss this very important theorem. □

We have the following useful consequence of the chain rule. Its proof is left to the reader as an exercise.

Proposition 14.3.4. *Let $n \in \mathbb{N}$. Suppose that $U \subset \mathbb{R}^n$ is an open set and $\boldsymbol{F} : U \to \mathbb{R}^n$ is a C^1-diffeomorphism. If $\boldsymbol{x}_0 \in U$ and $\boldsymbol{y}_0 = \boldsymbol{F}(\boldsymbol{x}_0)$, then the differential $d\boldsymbol{F}(\boldsymbol{x}_0)$ of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ is invertible and*

$$d\boldsymbol{F}^{-1}(\boldsymbol{y}_0) = d\boldsymbol{F}(\boldsymbol{x}_0)^{-1}. \qquad \square$$

The above result gives a necessary condition for a map to be a diffeomorphism, namely its differential has to be invertible. The next result is a very versatile criterion for recognizing diffeomorphisms. Roughly speaking, it states that maps with invertible differentials are very close to being diffeomorphisms.

Theorem 14.3.5 (Inverse function theorem)**.** *Let $n \in \mathbb{N}$, $k \in \mathbb{N} \cup \{\infty\}$. Suppose that $U \subset \mathbb{R}^n$ is an open set and $\boldsymbol{F} : U \to \mathbb{R}^n$ is a C^k-map. If $\boldsymbol{x}_0 \in U$ is such that the differential $d\boldsymbol{F}(\boldsymbol{x}_0) : \mathbb{R}^n \to \mathbb{R}^n$ is invertible, then there exists an open neighborhood V of $\boldsymbol{x}_0$ with the following properties.*

(i) $V \subset U$.

(ii) *The restriction of $\boldsymbol{F}$ to V defines a C^k-diffeomorphism $\boldsymbol{F} : V \to \mathbb{R}^n$.*

Proof. For simplicity we consider only the case $k = 1$. The case $k > 1$ follows inductively from this special case, [**13**, Prop. 3.2.9]. We follow closely the approach in the proof of [**35**, Thm. 2-11]. Denote by L the differential of $\boldsymbol{F}$ at $\boldsymbol{x}_0$ and set $\boldsymbol{y}_0 := \boldsymbol{F}(\boldsymbol{x}_0)$. We begin with an apparently very special case.

A. *The differential L is the identity operator $\mathbb{R}^n \to \mathbb{R}^n$.* We complete the proof in several steps.

Step 1. *We prove that there exists $r > 0$ such that the closed ball $\overline{B_r(\boldsymbol{x}_0)}$ is contained in U and the restriction of $\boldsymbol{F}$ to this closed ball is injective.*

Using the definition of the differential we can write

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) = \boldsymbol{F}(\boldsymbol{x}_0) + \boldsymbol{h} + R(\boldsymbol{h}) = \boldsymbol{y}_0 + \boldsymbol{h} + R(\boldsymbol{h}), \tag{14.3.1}$$

where

$$\lim_{\|\boldsymbol{h}\| \to 0} \frac{1}{\|\boldsymbol{h}\|} R(h) = 0. \tag{14.3.2}$$

Observe that

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_1) = \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_2) \overset{(14.3.1)}{\iff} R(\boldsymbol{h}_1) - R(\boldsymbol{h}_2) = -(\boldsymbol{h}_1 - \boldsymbol{h}_2).$$

We will show that the last equality above cannot happen if $\boldsymbol{h}_1, \boldsymbol{h}_2$ are sufficiently small and $\boldsymbol{h}_1 \neq \boldsymbol{h}_2$.

The correspondence $\boldsymbol{x} \mapsto J_{\boldsymbol{F}}(\boldsymbol{x})$ is continuous and the Jacobian matrix $J_{\boldsymbol{F}}(\boldsymbol{x}_0)$ is invertible. Thus, for $\boldsymbol{x}$ close to $\boldsymbol{x}_0$ the Jacobian $J_{\boldsymbol{F}}(\boldsymbol{x})$ is also invertible; see Exercise 12.9. Fix a radius $r_0 > 0$ such that $\overline{B_{2r_0}(\boldsymbol{x}_0)} \subset U$ and

$$J_{\boldsymbol{F}}(\boldsymbol{x}) \text{ is invertible } \forall \boldsymbol{x} \in B_{2r_0}(\boldsymbol{x}_0). \tag{14.3.3}$$

Observe that for $\|\boldsymbol{h}\| \leqslant 2r_0$ we have $R(\boldsymbol{h}) = \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{h} - \boldsymbol{y}_0$. This proves that the map $R : B_{2r_0}(\boldsymbol{0}) \to \mathbb{R}^n$ is differentiable and

$$J_R(\boldsymbol{h}) = J_{\boldsymbol{F}}(\boldsymbol{x}_0 + \boldsymbol{h}) - \mathbb{1} = J_{\boldsymbol{F}}(\boldsymbol{x}_0 + \boldsymbol{h}) - J_{\boldsymbol{F}}(\boldsymbol{x}_0).$$

Since the map $\boldsymbol{F}$ is C^1 we have

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \| J_{\boldsymbol{F}}(\boldsymbol{x}_0 + \boldsymbol{h}) - J_{\boldsymbol{F}}(\boldsymbol{x}_0) \|_{HS} = 0,$$

where $\| - \|_{HS}$ denotes the Frobenius norm of a matrix described in Remark 12.1.11. Fix a very small positive constant $\hbar$,

$$\hbar < \frac{1}{10n}. \tag{14.3.4}$$

There exists $r < r_0$ sufficiently small such that

$$\|J_R(\boldsymbol{h})\|_{HS} = \|J_{\boldsymbol{F}}(\boldsymbol{x}_0 + \boldsymbol{h}) - J_{\boldsymbol{F}}(\boldsymbol{x}_0)\|_{HS} < \hbar, \ \ \forall \|\boldsymbol{h}\| < 2r. \tag{14.3.5}$$

Corollary 13.3.11 implies that

$$\begin{gathered}\|\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_1) - \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_2) - (\boldsymbol{h}_1 - \boldsymbol{h}_2)\| = \|R(\boldsymbol{h}_1) - R(\boldsymbol{h}_2)\| \\ \overset{(14.3.5)}{\leqslant} \hbar\sqrt{n}\|\boldsymbol{h}_1 - \boldsymbol{h}_2\| \overset{(14.3.4)}{<} \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|, \ \ \forall \boldsymbol{h}_1, \boldsymbol{h}_2 \in B_{2r}(\boldsymbol{0}), \ \ \boldsymbol{h}_1 \neq \boldsymbol{h}_2.\end{gathered} \tag{14.3.6}$$

This proves that if $\|\boldsymbol{h}_1\|, \|\boldsymbol{h}_2\| < 2r$ and $\boldsymbol{h}_1 \neq \boldsymbol{h}_2$, then

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_1) - \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_2) - (\boldsymbol{h}_1 - \boldsymbol{h}_2) = R(\boldsymbol{h}_1) - R(\boldsymbol{h}_2) \neq -\big(\boldsymbol{h}_1 - \boldsymbol{h}_2\big).$$

Hence

$$\boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_1) - \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}_2) \neq \boldsymbol{0}.$$

In particular, this shows that the restriction of $\boldsymbol{F}$ on $\overline{B_r(\boldsymbol{x}_0)} \subset B_{2r}(\boldsymbol{x}_0) \subset U$ is injective.

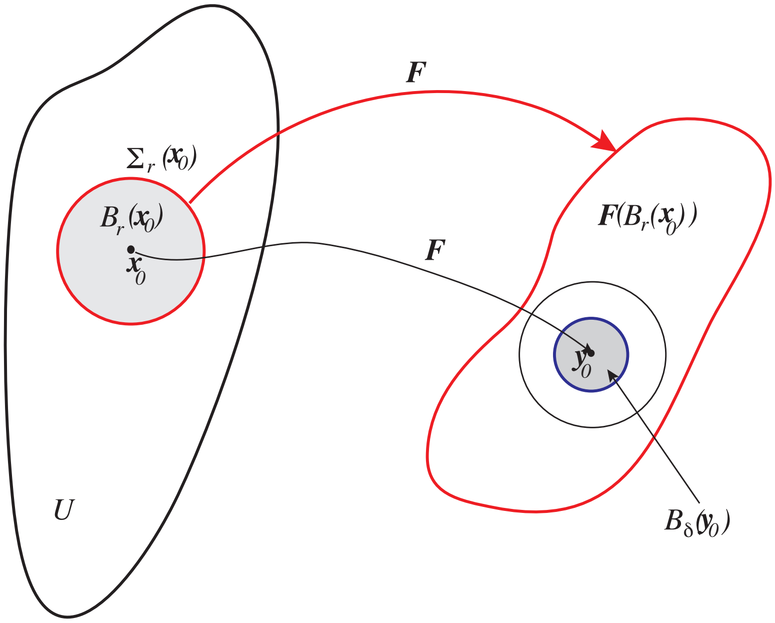

Figure 14.1. *The map $\boldsymbol{F}$ is injective on $B_r(\boldsymbol{x}_0)$, and the image of this ball contains a small ball $B_\delta(\boldsymbol{y}_0)$.*

The sphere

$$\Sigma_r(\boldsymbol{x}_0) = \big\{ \boldsymbol{x} \in \mathbb{R}^n; \ \ \|\boldsymbol{x} - \boldsymbol{x}_0\| = r \big\}$$

is compact, and thus its image $\boldsymbol{F}\big(\Sigma_r(\boldsymbol{x}_0)\big)$ is also compact; see Figure 14.1. Because of the injectivity of $\boldsymbol{F}$ on $\overline{B_r(\boldsymbol{x}_0)}$, the point $\boldsymbol{y}_0 = \boldsymbol{F}(\boldsymbol{x}_0)$ does not belong to the image $\boldsymbol{F}\big(\Sigma_r(\boldsymbol{x}_0)\big)$ of this sphere. Hence,

$$\operatorname{dist}\Big(\boldsymbol{y}_0, \boldsymbol{F}\big(\Sigma_r(\boldsymbol{x}_0)\big)\Big) > 0,$$

so there exists $\delta > 0$ such that

$$\|\boldsymbol{y}_0 - \boldsymbol{F}(\boldsymbol{x})\| > 2\delta, \ \ \forall \boldsymbol{x} \in \Sigma_r(\boldsymbol{x}_0). \tag{14.3.7}$$

Step 2. *We will prove that* $B_\delta(\boldsymbol{y}_0) \subset \boldsymbol{F}\big(\, B_r(\boldsymbol{x}_0)\,\big)$, i.e.,

$$\forall \boldsymbol{y} \in B_\delta(\boldsymbol{y}_0),\ \ \exists \boldsymbol{h} \in \mathbb{R}^n \ \text{ such that } \ \|\boldsymbol{h}\| < r \ \text{ and } \ \boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x}_0 + \boldsymbol{h}). \tag{14.3.8}$$

To do this, let $\boldsymbol{y} \in B_\delta(\boldsymbol{y}_0)$ and consider the function

$$g_{\boldsymbol{y}} : \overline{B_r(\boldsymbol{x}_0)} \to \mathbb{R},\ \ g_{\boldsymbol{y}}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x})\|^2.$$

The function $g_{\boldsymbol{y}}$ is continuous and the closed ball $\overline{B_r(\boldsymbol{x}_0)}$ is compact and thus $g_{\boldsymbol{y}}$ admits a global minimum

$$\boldsymbol{z} \in \overline{B_r(\boldsymbol{x}_0)},\ \ g_{\boldsymbol{y}}(\boldsymbol{z}) \leqslant g_{\boldsymbol{y}}(\boldsymbol{x}),\ \ \forall \boldsymbol{x} \in \overline{B_r(\boldsymbol{x}_0)}.$$

Let us first observe that $\boldsymbol{z} \in B_r(\boldsymbol{x}_0)$. We argue by contradiction. If $\boldsymbol{z} \in \Sigma_r(\boldsymbol{x}_0)$, then

$$\|\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{z})\| \geqslant \|\boldsymbol{y}_0 - \boldsymbol{F}(\boldsymbol{z})\| - \|\boldsymbol{y}_0 - \boldsymbol{y}\| \overset{(14.3.7)}{>} 2\delta - \underbrace{\|\boldsymbol{y}_0 - \boldsymbol{y}\|}_{<\delta} > \delta > \|\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x}_0)\|.$$

Hence

$$g_{\boldsymbol{y}}(\boldsymbol{z}) > g_{\boldsymbol{y}}(\boldsymbol{x}_0),\ \ \forall \boldsymbol{z} \in \Sigma_r(\boldsymbol{x}_0),$$

proving that the absolute minimum $\boldsymbol{z}$ of $g_{\boldsymbol{y}}$ is achieved somewhere inside the *open* ball $B_r(\boldsymbol{x}_0)$. The multidimensional Fermat principle then implies

$$\nabla g_{\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{0} \Longleftrightarrow J_{\boldsymbol{F}}(\boldsymbol{z})\big(\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{z})\,\big) = \boldsymbol{0}.$$

On the other hand, according to (14.3.3), the differential $\boldsymbol{dF}(\boldsymbol{z})$ is invertible. We deduce from the above equality that $\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{z})$ for some $\boldsymbol{z} \in B_r(\boldsymbol{x}_0)$.

Since $\boldsymbol{F} : U \to \mathbb{R}^n$ is continuous the preimage $\boldsymbol{F}^{-1}\big(B_\delta(\boldsymbol{y}_0)\,\big)$ is open (see Exercise 12.4(b)) and so is the set

$$V := \boldsymbol{F}^{-1}\big(B_\delta(\boldsymbol{y}_0)\,\big) \cap B_r(\boldsymbol{x}_0).$$

The above discussion shows that the resulting map $\boldsymbol{F} : V \to B_\delta(\boldsymbol{y}_0)$ is bijective.

Step 3. We prove that the inverse $\boldsymbol{G} := \boldsymbol{F}^{-1} : B_\delta(\boldsymbol{y}_0) \to V$ is Lipschitz continuous.

Let $\boldsymbol{y}_*, \boldsymbol{y} \in B_\delta(\boldsymbol{y}_0)$ We set $\boldsymbol{x} := G(\boldsymbol{y})$, $\boldsymbol{x}_* = \boldsymbol{G}(\boldsymbol{y}_*)$. Then $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{h}$, $\boldsymbol{x}_* = \boldsymbol{x}_0 + \boldsymbol{h}_*$. From (14.3.6) we deduce

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{x}_*\| - \|\boldsymbol{y} - \boldsymbol{y}_*\| &\leqslant \|\boldsymbol{y} - \boldsymbol{y}_* - (\boldsymbol{x} - \boldsymbol{x}_*)\| = \|R(\boldsymbol{h}) - R(\boldsymbol{h}_*)\| \overset{(14.3.6)}{\leqslant} \hbar\sqrt{n}\|\boldsymbol{h} - \boldsymbol{h}_*\| \\ &\overset{(14.3.4)}{\leqslant} \frac{1}{10\sqrt{n}}\|\boldsymbol{h} - \boldsymbol{h}_*\| = \frac{1}{10\sqrt{n}}\|\boldsymbol{x} - \boldsymbol{x}_*\| \leqslant \frac{1}{10}\|\boldsymbol{x} - \boldsymbol{x}_*\|.\end{aligned}$$

We deduce that

$$\|\boldsymbol{G}(\boldsymbol{y}) - \boldsymbol{G}(\boldsymbol{y}_*)\| = \|\boldsymbol{x} - \boldsymbol{x}_*\| \leqslant \frac{10}{9}\|\boldsymbol{y} - \boldsymbol{y}_*\|.$$

Step 4. We prove that the inverse $\boldsymbol{G} := \boldsymbol{F}^{-1} : B_\delta(\boldsymbol{y}_0) \to V$ is differentiable.

Fix $\boldsymbol{y}_* \in B_\delta(\boldsymbol{y}_0)$. There exists $\boldsymbol{x}_* \in V$ such that $\boldsymbol{F}(\boldsymbol{x}_*) = \boldsymbol{y}_*$. Proposition 14.3.4 suggests that the differential of $\boldsymbol{G}$ at $\boldsymbol{y}_*$ should be the inverse of the differential of $\boldsymbol{F}$ at $\boldsymbol{x}_0$. For $\boldsymbol{y} \in B_\delta(\boldsymbol{y}_0)$ we set

$$R(\boldsymbol{y}, \boldsymbol{y}_*) := \Big(\, \boldsymbol{G}(\boldsymbol{y}) - \boldsymbol{G}(\boldsymbol{y}_*) - \boldsymbol{dF}(\boldsymbol{x}_*)^{-1}(\boldsymbol{y} - \boldsymbol{y}_*)\,\Big).$$

We have to prove that

$$\lim_{\boldsymbol{y} \to \boldsymbol{y}_*} \frac{\|R(\boldsymbol{y}, \boldsymbol{y}_*)\|}{\|\boldsymbol{y} - \boldsymbol{y}_*\|} = 0.$$

Observe first that

$$\begin{aligned}\boldsymbol{dF}(\boldsymbol{x}_*)R(\boldsymbol{y}, \boldsymbol{y}_*) &= \boldsymbol{dF}(\boldsymbol{x}_*)\Big(\,\boldsymbol{G}(\boldsymbol{y}) - \boldsymbol{G}(\boldsymbol{y}_*)\,\Big) - (\boldsymbol{y} - \boldsymbol{y}_*) \\ &= \underbrace{\boldsymbol{dF}(\boldsymbol{x}_*)\Big(\,\boldsymbol{G}(\boldsymbol{y}) - \boldsymbol{G}(\boldsymbol{y}_*)\,\Big) - \Big(\,\boldsymbol{F}(\,\boldsymbol{G}(\boldsymbol{y})\,) - \boldsymbol{F}(\,\boldsymbol{G}(\boldsymbol{y}_*)\,)\,\Big)}_{=:Q(\boldsymbol{y}, \boldsymbol{y}_*)}.\end{aligned}$$

Since $\boldsymbol{F}$ is differentiable at $\boldsymbol{x}_*$ we have

$$\lim_{\boldsymbol{y} \to \boldsymbol{y}_*} \frac{\|Q(\boldsymbol{y}, \boldsymbol{y}_*)\|}{\|\boldsymbol{G}(\boldsymbol{y}) - \boldsymbol{G}(\boldsymbol{y}_*)\|} = 0. \tag{14.3.9}$$

On the other hand,

$$\|\boldsymbol{G}(\boldsymbol{y})-\boldsymbol{G}(\boldsymbol{y}_*)\| \leqslant \frac{10}{9}\|\boldsymbol{y}-\boldsymbol{y}_*\|$$

so that

$$\frac{10}{9}\frac{\|Q(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{G}(\boldsymbol{y})-\boldsymbol{G}(\boldsymbol{y}_*)\|} \geqslant \frac{\|Q(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{y}-\boldsymbol{y}_*\|}$$

We deduce that

$$\frac{\|\boldsymbol{dF}(\boldsymbol{x}_*)R(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{y}-\boldsymbol{y}_*\|} \leqslant \frac{10}{9}\frac{\|Q(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{G}(\boldsymbol{y})-\boldsymbol{G}(\boldsymbol{y}_*)\|}.$$

On the other hand, since $\boldsymbol{dF}(\boldsymbol{x}_*)$ is invertible, we deduce from Exercise 12.28(iii) that there exists a constant $C>0$ such that

$$C\|\boldsymbol{h}\| \leqslant \|\boldsymbol{dF}(\boldsymbol{x}_*)\boldsymbol{h}\|,\ \ \forall \boldsymbol{h}\in\mathbb{R}^n.$$

We conclude that

$$C\frac{\|R(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{y}-\boldsymbol{y}_*\|} \leqslant \frac{10}{9}\frac{\|Q(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{G}(\boldsymbol{y})-\boldsymbol{G}(\boldsymbol{y}_*)\|}.$$

Invoking the Squeezing Principle and (14.3.9) we deduce from the above

$$\lim_{\boldsymbol{y}\to\boldsymbol{y}_*}\frac{\|R(\boldsymbol{y},\boldsymbol{y}_*)\|}{\|\boldsymbol{y}-\boldsymbol{y}_*\|}=0.$$

This proves the differentiability of $\boldsymbol{G}$ at $\boldsymbol{y}_*$.

Step 5. We finally prove that map $\boldsymbol{G}$ is C^1. We have to show that the map $\boldsymbol{y}\mapsto J_{\boldsymbol{G}}(\boldsymbol{y})$ is continuous, i.e., the map

$$\boldsymbol{y}\mapsto J_{\boldsymbol{F}}\big(\boldsymbol{G}(\boldsymbol{y})\big)^{-1}$$

is continuous. This follows from Exercise 12.10.

B. We now discuss the general case when we do not assume that $\boldsymbol{dF}(\boldsymbol{x}_0)=\mathbb{1}$. Set $L=\boldsymbol{dF}(\boldsymbol{x}_0)$. Define

$$\Phi: U\to\mathbb{R}^n,\ \ \Phi=L^{-1}\circ\boldsymbol{F}.$$

The chain rule implies that

$$d\Phi=dL^{-1}\circ\boldsymbol{dF}=L^{-1}\circ\boldsymbol{dF}=\mathbb{1}.$$

From Case **A** we deduce that there exists an open neighborhood V of $\boldsymbol{x}_0$ contained in U such that the restriction of Φ to V is a diffeomorphism. From the equality $F=L\circ\Phi$ we deduce that the restriction of F to V is also a diffeomorphism.

□

Remark 14.3.6. (a) The assumption that $\boldsymbol{dF}(\boldsymbol{x}_0)$ is invertible is equivalent with the condition

$$\det J_{\boldsymbol{F}}(\boldsymbol{x}_0)\neq 0.$$

This is easier to verify especially when n is not too large.

(b) If $V\subset\mathbb{R}^n$ is an open neighborhood of $\boldsymbol{x}_0$ satisfying the conditions (i) and (ii) in Theorem 14.3.5, then any smaller open neighborhood $W\subset V$ of $\boldsymbol{x}_0$ satisfies these conditions.

□

We have the following useful consequence of the inverse function theorem. Its proof is left to you as an exercise.

Corollary 14.3.7. *Let $n\in\mathbb{N}$. Suppose that U is an open subset of $\mathbb{R}^n$ and $\boldsymbol{F}: U\to\mathbb{R}^n$ is a C^1-map satisfying the following conditions.*

(i) *The map* $\boldsymbol{F}$ *is injective.*

(ii) *For any* $\boldsymbol{x} \in U$*, the differential* $d\boldsymbol{F}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^n$ *is bijective.*

Then the map $\boldsymbol{F}$ *is a* C^1*-diffeomorphism.* □

Remark 14.3.8. The condition (ii) in the above corollary is equivalent with the condition

$$\det J_{\boldsymbol{F}}(\boldsymbol{x}) \neq 0, \quad \forall \boldsymbol{x} \in U.$$

This is easier to verify especially when n is not too large. □

Example 14.3.9. Consider again the map

$$\boldsymbol{F} : (0,\infty) \times (0,2\pi) \to \mathbb{R}^2, \quad \boldsymbol{F}(r,\theta) = \left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} r\cos\theta \\ r\sin\theta \end{array}\right]$$

in Example 14.3.3. We have seen there that it is injective. According to Corollary 14.3.7, to prove that it is a diffeomorphism it suffices to show that for any $(r,\theta) \in (0,\infty) \times (0,2\pi)$ the Jacobian matrix $J_{\boldsymbol{F}}(r,\theta)$ is invertible. We have

$$J_{\boldsymbol{F}}(r,\theta) = \left[\begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ & \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{array}\right] = \left[\begin{array}{cc} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{array}\right].$$

The determinant of the above matrix is

$$\det J_{\boldsymbol{F}} = (\cos\theta)\cdot(r\cos\theta) - (-r\sin\theta)\cdot(\sin\theta) = r\cos^2\theta + r\sin^2\theta = r > 0.$$

Thus the matrix $J_{\boldsymbol{F}}(r,\theta)$ is invertible for any $(r,\theta) \in (0,\infty) \times (0,2\pi)$. □

Example 14.3.10. The transformation $\boldsymbol{F}$ in Example 14.3.9 is often referred to as the *change to polar coordinates*. A function u depending on the Cartesian coordinates (x,y) can be transformed to a function depending on the coordinates (r,θ),

$$u(x,y) = u(r\cos\theta, r\sin\theta).$$

Often in physics and geometry one is faced with the problem of transforming various quantities expressed in the (x,y)-coordinates to quantities expressed in the polar coordinates (r,θ). We discuss below one such important example.

Suppose $u = u(x,y)$ is a C^2-function. We are deliberately vague about the domain of definition of u since this details is irrelevant to the computations we are about to perform. Its *Laplacian* is the function

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

We want to express the Laplacian in polar coordinates. The chain rule, cleverly deployed, will do the trick.

Note first that for any function (or quantity) q depending on the variables (x, y), $q = q(x, y)$, we have

$$\frac{\partial q}{\partial x} = \frac{\partial q}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial q}{\partial \theta}\frac{\partial \theta}{\partial x}, \tag{14.3.10a}$$

$$\frac{\partial q}{\partial y} = \frac{\partial q}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial q}{\partial \theta}\frac{\partial \theta}{\partial y}. \tag{14.3.10b}$$

Let us concentrate first on the x-derivative. We rewrite (14.3.10a) in the form,

$$\frac{\partial q}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial q}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial q}{\partial \theta}.$$

Since the exact nature of the quantity q is not important in the sequel, we will drop the letter q from our notations. Hence, the above equality becomes

$$\frac{\partial}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta}. \tag{14.3.11}$$

From the equalities $r^2 = x^2 + y^2$, $x = r\cos\theta$ and $y = r\sin\theta$ we deduce

$$\partial_x r = \frac{x}{r} = \cos\theta,\ \ \partial_y r = \frac{y}{r} = \sin\theta. \tag{14.3.12}$$

Derivating the equality $y = r\sin\theta$ with respect to x we deduce

$$0 = \partial_x r(\sin\theta) + r(\cos\theta)\partial_x\theta = \cos\theta\sin\theta + (r\cos\theta)\partial_x\theta = \cos\theta\big(\sin\theta + r\partial_x\theta\big)$$

$$\Rightarrow \partial_x\theta = -\frac{\sin\theta}{r}.$$

Hence (14.3.11) becomes

$$\boxed{\frac{\partial}{\partial x} = \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}}. \tag{14.3.13}$$

Then

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= \frac{\partial}{\partial x}\frac{\partial u}{\partial x} = \Big(\cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\Big)\Big(\cos\theta\frac{\partial u}{\partial r} - \frac{\sin\theta}{r}\frac{\partial u}{\partial\theta}\Big)\\ &= \cos\theta\frac{\partial}{\partial r}\Big(\cos\theta\frac{\partial u}{\partial r} - \frac{\sin\theta}{r}\frac{\partial u}{\partial\theta}\Big) - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\Big(\cos\theta\frac{\partial u}{\partial r} - \frac{\sin\theta}{r}\frac{\partial u}{\partial\theta}\Big)\\ &= \cos\theta\Big(\cos\theta\frac{\partial^2 u}{\partial r^2} + \frac{\sin\theta}{r^2}\frac{\partial u}{\partial\theta} - \frac{\sin\theta}{r}\frac{\partial^2 u}{\partial r\partial\theta}\Big) - \frac{\sin\theta}{r}\Big(-\sin\theta\frac{\partial u}{\partial r} + \cos\theta\frac{\partial^2 u}{\partial\theta\partial r} - \frac{\sin\theta}{r}\frac{\partial^2 u}{\partial\theta^2}\Big)\\ &= \cos^2\theta\partial_r^2 u + \frac{\sin\theta\cos\theta}{r^2}\partial_\theta u - 2\frac{\sin\theta\cos\theta}{r}\partial^2_{r\theta}u + \frac{\sin^2\theta}{r}\partial_r u + \frac{\sin^2\theta}{r^2}\partial^2_\theta u.\end{aligned}$$

Arguing in a similar fashion we have

$$\frac{\partial}{\partial y} = \frac{\partial r}{\partial y}\frac{\partial}{\partial r} + \frac{\partial\theta}{\partial y}\frac{\partial}{\partial\theta}.$$

Derivating with respect to y the equality $x = r\cos\theta$ we deduce in similar fashion that $\partial_y\theta = \frac{\cos\theta}{r}$ so that

$$\boxed{\frac{\partial}{\partial y} = \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial\theta}},$$

$$\begin{aligned}\frac{\partial^2 u}{\partial y^2} &= \Big(\sin\theta\frac{\partial}{\partial r}+\frac{\cos\theta}{r}\frac{\partial}{\partial\theta}\Big)\Big(\sin\theta\frac{\partial u}{\partial r}+\frac{\cos\theta}{r}\frac{\partial u}{\partial\theta}\Big)\\ &= \sin\theta\frac{\partial}{\partial r}\Big(\sin\theta\frac{\partial u}{\partial r}+\frac{\cos\theta}{r}\frac{\partial u}{\partial\theta}\Big)+\frac{\cos\theta}{r}\frac{\partial}{\partial\theta}\Big(\sin\theta\frac{\partial u}{\partial r}+\frac{\cos\theta}{r}\frac{\partial u}{\partial\theta}\Big)\\ &= \sin\theta\Big(\sin\theta\frac{\partial^2 u}{\partial r^2}-\frac{\cos\theta}{r^2}\frac{\partial u}{\partial\theta}+\frac{\cos\theta}{r}\frac{\partial^2 u}{\partial r\partial\theta}\Big)+\frac{\cos\theta}{r}\Big(\cos\theta\frac{\partial u}{\partial r}+\sin\theta\frac{\partial^2 u}{\partial\theta\partial r}-\frac{\sin\theta}{r}\frac{\partial u}{\partial\theta}+\frac{\cos\theta}{r}\frac{\partial^2 u}{\partial\theta^2}\Big)\\ &= \sin^2\theta\partial_r^2 u-\frac{\sin\theta\cos\theta}{r^2}\partial_\theta u+\frac{2\sin\theta\cos\theta}{r}\partial^2_{r\theta}u+\frac{\cos^2\theta}{r}\partial_r u+\frac{\cos^2\theta}{r^2}\partial^2_\theta u.\end{aligned}$$

Putting together all of the above we deduce

$$\boxed{\Delta u=\partial_r^2 u+\frac{1}{r}\partial_r u+\frac{1}{r^2}\partial_\theta^2 u=\frac{1}{r}\frac{\partial}{\partial r}\Big(r\frac{\partial u}{\partial r}\Big)+\frac{1}{r^2}\frac{\partial u}{\partial\theta^2}}. \tag{14.3.14}$$

To see how this works in practice, consider the special case $u=(x^2+y^2)^{\frac{p}{2}}$. Since $x^2+y^2=r^2$ we deduce $u=r^p$ and

$$\Delta u=\partial_r^2\big(r^p\big)+\frac{1}{r}\partial_r\big(r^p\big)=p(p-1)r^{p-2}+pr^{p-2}=p^2r^{p-2}. \qquad \square$$

14.4. The implicit function theorem

To understand the meaning of the implicit function theorem it is useful to start with a simple example.

Example 14.4.1. Consider the function $f:\mathbb{R}^2\to\mathbb{R}$, $f(x,y)=x^2+y^2-1$. The level set

$$f^{-1}(0)=\big\{(x,y)\in\mathbb{R}^2;\;\; f(x,y)=1\,\big\}$$

is the circle C_1 of radius 1 centered at the origin of $\mathbb{R}^2$. This curve cannot be the graph of any function, but portions of it are graphs. For example, the part of C_1 above the x-axis

$$\big\{\,(x,y)\in C_1;\;\; y>0\,\big\},$$

is the graph of a function. To see this, we solve for y the equality $x^2+y^2=1$, and since $y>0$, we obtain the unique solution

$$y=\sqrt{1-x^2}.$$

We say that the function $\sqrt{1-x^2}$ is a function defined *implicitly* by the equality $f(x,y)$.

This is not an isolated phenomenon. The implicit function theorem states that for many equations of the type $f(x,y)=$ *const* the solution set is locally the graph of a function g, although we cannot describe g as explicitly as in the above simple example. □

Theorem 14.4.2 (Implicit function theorem. Version 1)**.** *Let $m,n\in\mathbb{N}$. Suppose that*

$$\mathcal{O}\subset\mathbb{R}^n\times\mathbb{R}^m$$

is an open set, $\boldsymbol{F}=\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v}):\mathcal{O}\to\mathbb{R}^m$ is a C^1 map and $(\boldsymbol{u}_0,\boldsymbol{v}_0)\in\mathcal{O}$ is a point satisfying the following properties.

(i) $\boldsymbol{F}(\boldsymbol{u}_0,\boldsymbol{v}_0)=\boldsymbol{0}$.

(ii) *The restriction of the differential $d\boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{v}_0)$ to the subspace*

$$\mathbf{0} \times \mathbb{R}^m \subset \mathbb{R}^n \times \mathbb{R}^m$$

induces an invertible linear map $\mathbf{0} \times \mathbb{R}^m \to \mathbb{R}^m$.

Then there exists an open neighborhood U of $\boldsymbol{u}_0 \in \mathbb{R}^n$, an open neighborhood V of $\boldsymbol{v}_0 \in \mathbb{R}^m$ and a C^1-map $\boldsymbol{G} : U \to V$ with the following properties

- *$U \times V \subset \mathcal{O}$.*
- *If $(\boldsymbol{u}, \boldsymbol{v}) \in U \times V$, then $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{v}) = \mathbf{0}$ if and only if $\boldsymbol{v} = \boldsymbol{G}(\boldsymbol{u})$.*

In other words, for any $\boldsymbol{u} \in U$, the equation $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{v}) = 0$ has a unique solution $\boldsymbol{v} \in V$. This unique solution is denoted by $\boldsymbol{G}(\boldsymbol{u})$. We say that $\boldsymbol{G}$ is the function implicitly defined by the equation $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{v}) = \mathbf{0}$.

Proof. If we represent $L := d\boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{v}_0)$ as an $m \times (n+m)$ matrix, then it has a block decomposition

$$L = \left[\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}}, \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{v}} \right] = [A\ B],$$

where A is an $m \times n$ matrix and B is a $m \times m$ matrix. The matrix $B = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{v}}$ describes the restriction of L to the subspace $\mathbf{0} \times \mathbb{R}^m$ and assumption (ii) implies that B is invertible.

Consider the new map $\boldsymbol{H} : \mathcal{O} \to \mathbb{R}^n \times \mathbb{R}^m$,

$$\boldsymbol{H}(\boldsymbol{u}, \boldsymbol{v}) = \big(\boldsymbol{u}, \boldsymbol{F}(\boldsymbol{u}, \boldsymbol{v}) \big).$$

The differential of $\boldsymbol{H}$ at $(\boldsymbol{v}_0, \boldsymbol{u}_0)$ is a linear map $T : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ described by an $(m+n) \times (m+n)$-matrix with block decomposition

$$T = \left[\begin{array}{cc} \mathbb{1}_n & \mathbf{0} \\ & \\ \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{u}} & \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{v}} \end{array} \right] = \left[\begin{array}{cc} \mathbb{1}_n & \mathbf{0} \\ A & B \end{array} \right].$$

Since B is invertible, we deduce that T is also invertible since $\det T = \det B \neq 0$. Note that $\boldsymbol{H}(\boldsymbol{u}_0, \boldsymbol{v}_0) = (\boldsymbol{u}_0, \mathbf{0})$.

From the inverse function theorem we deduce that there exists an open neighborhood W of $(\boldsymbol{u}_0, \boldsymbol{v}_0)$ contained in $\mathcal{O}$ such that the restriction of $\boldsymbol{H}$ to W is a diffeomorphism. By making W smaller as in Remark 14.3.6, we can assume that W has the form $W = U \times V$, where $U \subset \mathbb{R}^n$ is an open neighborhood of $\boldsymbol{u}_0$ in $\mathbb{R}^n$ and $V \subset \mathbb{R}^m$ is an open neighborhood of $\boldsymbol{v}_0$ in $\mathbb{R}^m$.

We denote by $\mathcal{W}$ the image of $U \times V$ via $\boldsymbol{H}$, $\mathcal{W} := \boldsymbol{H}(U \times V)$. Let $\Phi : \mathcal{W} \to U \times V$ denote the inverse of $\boldsymbol{H} : U \times V \to \mathcal{W}$. The diffeomorphism Φ has the form

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{u}, \boldsymbol{v}) = \big(\Psi(\boldsymbol{x}, \boldsymbol{y}), \Xi(\boldsymbol{x}, \boldsymbol{y}) \big) \in U \times V \subset \mathbb{R}^n \times \mathbb{R}^m,$$

where

$$\Psi : \mathcal{W} \to \mathbb{R}^n, \ \ \Xi : \mathcal{W} \to \mathbb{R}^m$$

are C^1-maps. Note that if $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{W}$ and,

$$(\boldsymbol{u},\boldsymbol{v})=\Phi(\boldsymbol{x},\boldsymbol{y})=\big(\,\Psi(\boldsymbol{x},\boldsymbol{y}),\Xi(\boldsymbol{x},\boldsymbol{y})\,\big),$$

then

$$\begin{aligned}(\boldsymbol{x},\boldsymbol{y})&=\boldsymbol{H}(\boldsymbol{u},\boldsymbol{v})=\big(\,\boldsymbol{u},\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})\,\big).\\ &=\Big(\,\Psi(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{F}\big(\Phi(\boldsymbol{x},\boldsymbol{y}),\Xi(\boldsymbol{x},\boldsymbol{y})\,\big)\,\Big).\end{aligned}$$

We deduce that $\boldsymbol{u}=\boldsymbol{x}$, i.e., $\Xi(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{v}$. Hence the inverse Φ has the form

$$\Phi(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{u},\boldsymbol{v})=\big(\,\boldsymbol{x},\Xi(\boldsymbol{x},\boldsymbol{y})\,\big),$$

where

$$\boldsymbol{u}=\boldsymbol{x},\ \ \boldsymbol{y}=\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v}),$$

Note that

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})=0&\Longleftrightarrow(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{H}(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u},\boldsymbol{0})\\ &\Longleftrightarrow(\boldsymbol{u},\boldsymbol{v})=\Phi(\boldsymbol{u},\boldsymbol{0})=\big(\,\boldsymbol{u},\Xi(\boldsymbol{u},\boldsymbol{0})\,\big).\end{aligned}$$

The sought out map $\boldsymbol{G}$ is then

$$\boldsymbol{G}(\boldsymbol{u})=\Xi(\boldsymbol{u},\boldsymbol{0}).$$

□

Remark 14.4.3. (a) The above proof shows that there exist

- an open set $\mathcal{W}\subset\mathbb{R}^n\times\mathbb{R}^m$ containing $(\boldsymbol{u}_0,\boldsymbol{0})$,
- an open neighborhood V of $\boldsymbol{v}_0$ in $\mathbb{R}^m$,
- an open neighborhood U of $\boldsymbol{u}_0$ in $\mathbb{R}^n$, and
- a diffeomorphism $\Phi:\mathcal{W}\to\mathbb{R}^n\times\mathbb{R}^m$,

with the following properties.

(i) $\Phi(\boldsymbol{u}_0,\boldsymbol{0})=(\boldsymbol{u}_0,\boldsymbol{v}_0)$, $\Phi(\mathcal{W})=U\times V$.

(ii) The diffeomorphism Φ maps the part of the plane $\mathbb{R}^n\times\boldsymbol{0}$ contained in $\mathcal{W}$ bijectively to the part of the set $\boldsymbol{F}=\boldsymbol{0}$ contained in $U\times V$.

(b) The assumption (ii) in the statement of the Implicit Function Theorem can be rephrased in a more convenient way. In the proof assumption (ii) was used to conclude that the $(n+m)\times m$ matrix $J_{\boldsymbol{F}}$ representing $d\boldsymbol{F}(\boldsymbol{x}_0,\boldsymbol{y}_0)$ has the property that the matrix B determined by the columns and the *last* m rows is invertible. The condition (ii) is then equivalent with the condition

$$\det B\neq 0.$$

Note that if $F^1,\dots,F^m$ are the components of $\boldsymbol{F}$ and $v^1,\dots,v^m$ are the components of $\boldsymbol{v}$, then B is the $m\times m$ matrix with entries

$$B^i_j=\frac{\partial F^i}{\partial v^j},\ \ 1\leqslant i,j\leqslant m.$$

The condition implies that $d\boldsymbol{F}(\boldsymbol{u}_0,\boldsymbol{v}_0)$ is surjective.

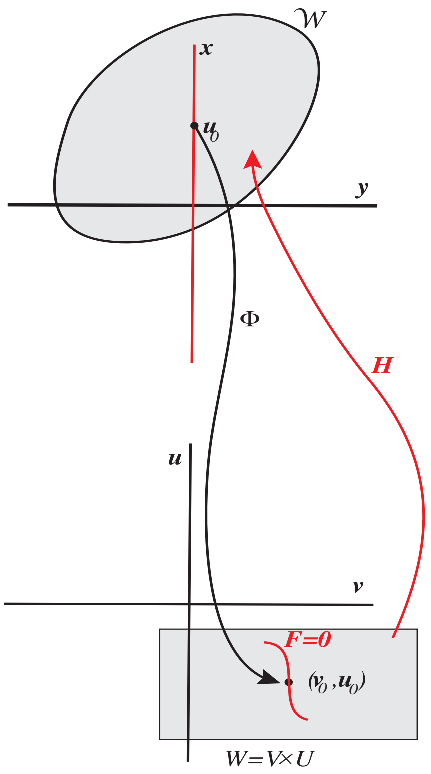

Figure 14.2. *The map* Φ *sends a portion of the subspace* $\mathbf{0}\times\mathbb{R}^n$ *bijectively to a portion of the zero set* $\{\boldsymbol{F}=\mathbf{0}\}$.

If we assume *only* that the differential $d\boldsymbol{F}(\boldsymbol{u}_0,\boldsymbol{v}_0):\mathbb{R}^{n+m}\to\mathbb{R}^m$ is *surjective*, then the $m\times(n+m)$-matrix representing this linear operator has maximal rank m and thus, there exist m columns so that the matrix determined by these columns and all the m rows is invertible; see e.g. [40, Thm. 6.1]. If we reorder the components of a vector in $\mathbb{R}^{n+m}$ we can then assume that these m columns are the last m columns.

(c) Note that the surjectivity of the linear operator $d\boldsymbol{F}(\boldsymbol{u}_0,\boldsymbol{v}_0):\mathbb{R}^{n+m}\to\mathbb{R}^m$ is equivalent with the linear independence of the m rows of $J_{\boldsymbol{F}}(\boldsymbol{u}_0,\boldsymbol{v}_0)$. If $F^1,\dots,F^m$ are the components of F, then the rows of $J_{\boldsymbol{F}}$ describe the differentials $dF^1,\dots,dF^m$ and we see that the rows are linearly independent if and only if the gradients $\nabla F^1,\dots,\nabla F^m$ are linearly independent. □

In view of the last remark, we can give an equivalent but more flexible formulation of Theorem 14.4.2. First, let us introduce some convenient terminology. A *codimension* m *coordinate subspace* of an Euclidean space $\mathbb{R}^n$ is a linear subspace of $\mathbb{R}^n$ described by the vanishing of a given group of m coordinates.

For example, the subspace of $\mathbb{R}^5$ of the form

$$S = \big\{ \, (x^1, 0, x^3, x^4, 0); \;\; x^1, x^3, x^4 \in \mathbb{R} \, \big\}$$

is a codimension 2 coordinate subspace described by the vanishing of the coordinates x^2, x^5. It is naturally isomorphic to $\mathbb{R}^3 = \mathbb{R}^{5-2}$. The codimension 3 subspace

$$\big\{ \, (0, x^2, 0, 0, x^5); \;\; x^2, x^5 \in \mathbb{R} \, \big\}$$

described by the vanishing of the coordinates x^1, x^3, x^4 is none other than $S^\perp$, the orthogonal complement of S. It is naturally isomorphic to $\mathbb{R}^2$. Note that we have a natural decomposition

$$(x^1, x^2, x^3, x^4, x^5) = \underbrace{(x^1, 0, x^3, x^4, 0)}_{\in S} + \underbrace{(0, x^2, 0, 0, x^5)}_{\in S^\perp}.$$

In general, a codimension m coordinate subspace of $\mathbb{R}^N$ is naturally isomorphic to $\mathbb{R}^{N-m}$ and thus has dimension $N - m$. The orthogonal complement $S^\perp$ is another coordinate subspace of codimension $N - m$. Moreover any $\boldsymbol{z} \in \mathbb{R}^N$ admits a *unique decomposition* of the form

$$\boldsymbol{z} = \boldsymbol{u} + \boldsymbol{v}, \;\; \boldsymbol{u} \in S, \;\; \boldsymbol{v} \in S^\perp.$$

The vectors $\boldsymbol{u}, \boldsymbol{v}$ are called the *projections* of $\boldsymbol{z}$ on S and respectively $S^\perp$.

Theorem 14.4.4 (Implicit function theorem. Version 2)**.** *Let $m, n \in \mathbb{N}$ and set $N := n+m$. Suppose that $\mathcal{O} \subset \mathbb{R}^N$ is an open set, $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{x}) : \mathcal{O} \to \mathbb{R}^m$ is a C^1 map and $\boldsymbol{p}_0 \in \mathcal{O}$ is a point satisfying the following properties.*

(i) $\boldsymbol{F}(\boldsymbol{p}_0) = \boldsymbol{0}$.

(ii) *The differential $L = d\boldsymbol{F}(\boldsymbol{p}_0) : \mathbb{R}^N \to \mathbb{R}^m$ is surjective.*

Label the coordinates $(x^i)_{1 \leqslant i \leqslant N}$ of $\boldsymbol{x} \in \mathbb{R}^N$ so that

$$\det \left[\frac{\partial F^i}{\partial x^{n+j}}(\boldsymbol{p}_0) \right]_{1 \leqslant i,j \leqslant m} \neq 0.$$

Denote by S the codimension m coordinate plane defined by the equations

$$x^{n+1} = \cdots = x^{n+m} = 0.$$

Then there exist

- *an open ball $U \subset S$ centered at $\boldsymbol{u}_0$, the projection of $\boldsymbol{p}_0$ on S,*
- *an open ball $V \subset S^\perp$ centered at $\boldsymbol{v}_0$, the projection of $\boldsymbol{p}_0$ on $S^\perp$ and a C^1-map $\boldsymbol{G} : U \to V$*

with the following properties:

- *$V \times U \subset \mathcal{O}$;*
- *If $(\boldsymbol{v}, \boldsymbol{u}) \in U \times V$, then $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{V}) = \boldsymbol{0}$ if and only if $\boldsymbol{v} = \boldsymbol{G}(\boldsymbol{u})$.*

□

Remark 14.4.5. Under the assumptions of the above theorem, we say that we can solve for $x^{n+1}, \dots, x^{n+m}$ in terms of the remaining variables $x^1, \dots, x^n$. The map $\boldsymbol{G}$ is then described by m functions $g^1, \dots, g^m$, depending on the "free" variables $x^1, \dots, x^n$, such that

$$x^{m+1} = g^1\big(x^1, \dots, x^n\big), \dots, x^m = g^m\big(x^1, \dots, x^n\big),$$

if and only if

$$F^1\big(x^1, \dots, x^m, x^{m+1}, \dots, x^{m+n}\big) = \cdots = F^m\big(x^1, \dots, x^m, x^{m+1}, \dots, x^{m+n}\big) = 0,$$

where $F^1(\boldsymbol{x}), \dots, F^m(\boldsymbol{x})$ are the components of $\boldsymbol{F}(\boldsymbol{x}) \in \mathbb{R}$. In the notation of the above theorem we have

$$\boldsymbol{v} = (x^{n+1}, \dots, x^{n+m}), \;\; \boldsymbol{u} = \big(x^1, \dots, x^n\,\big). \qquad \square$$

Example 14.4.6. Consider a function $f : \mathbb{R}^{n+1} \to \mathbb{R}$. We denote by $(x^0, x^1, \dots, x^n)$ the Cartesian coordinates $\mathbb{R}^{n+1}$. Consider the zero set of f,

$$Z_f := \big\{\, (x^0, x^1, \dots, x^n) \in \mathbb{R}^{n+1};\;\; f(x^0, x^1, \dots, x^n) = 0 \,\big\}.$$

Assume for some $\boldsymbol{x}_0 = (x_0^0, x_0^1, \dots, x_0^n)$ we have $\boldsymbol{x}_0 \in Z_f$ and the differential of f at $\boldsymbol{x}_0$ is surjective as a linear map $\mathbb{R}^{n+1} \to \mathbb{R}$.

The differential of f at $\boldsymbol{x}_0$ is the $1 \times (n+1)$ matrix

$$\big[\, \partial_{x^0} f(\boldsymbol{x}_0), \partial_{x^1} f(\boldsymbol{x}_0), \dots, \partial_{x^n} f(\boldsymbol{x}_0) \,\big].$$

The differential $df(\boldsymbol{x}_0)$ is surjective if and only if it is nonzero, i.e., one of the partial derivatives

$$\partial_{x^0} f(\boldsymbol{x}_0),\; \partial_{x^1} f(\boldsymbol{x}_0), \dots, \partial_{x^n} f(\boldsymbol{x}_0)$$

is nonzero. Without loss of generality we can assume that $\partial_{x^0} f(\boldsymbol{x}_0) \neq 0$.

Consider the coordinate subspace S described by $x^0 = 0$. Explicitly,

$$S = \big\{\, (0, x^1, \dots, x^n);\;\; x^1, \dots, x^n \in \mathbb{R} \,\big\}.$$

Loosely speaking, the implicit function theorem says that, in a neighborhood of $\boldsymbol{x}_0$, we can solve the equation

$$f(x^0, x^1, \dots, x^n) = 0$$

uniquely for x^0 in terms of $x^1, \dots, x^n$.

More precisely, the implicit function theorem states that there exists an open (n-dimensional) ball B^n in $S \cong \mathbb{R}^n$ centered at $\boldsymbol{u}_0 = (\boldsymbol{x}_0^1, \dots, \boldsymbol{x}_0^n)$, an open interval $I \subset \mathbb{R}$ containing $v_0 = x_0^0$, and a C^1-function $g : B \to I$ such that

$$(x^0, x^1, \dots, x^n) \in I \times B^n \text{ and } f(x^0, x^1, \dots, x^n) = 0 \Longleftrightarrow x^0 = g(x^1, \dots, x^n).$$

The function g is only locally defined, and it is called the *implicit function* determined by the equation

$$f(x^0, x^1, \dots, x^n) = 0,$$

i.e.,

$$f(x^0,x^1,\dots,x^n)=0\Longleftrightarrow x^0=g(x^1,\dots,x^n)\Longleftrightarrow f\big(\,g(x^1,\dots,x^n),\ x^1,\dots,x^n\,\big)=0. \tag{14.4.1}$$

We often express this by saying that, in an open neighborhood of $\boldsymbol{x}_0$, along the zero set Z_f the coordinate x^0 is a function of the remaining coordinates

$$x^0=x^0(x^1,\dots,x^n)$$

and thus, locally, Z_f is the graph of a C^1-function depending on the n variables $(x^1,\dots,x^n)$.

From the equality $f(x^0,x^1,\dots,x^n)=0$ we can determine the partial derivatives of x^0 at $\boldsymbol{u}_0$, when x^0 is viewed as a function of $(x^1,\dots,x^n)$. Derivating the equality

$$f(x^0,x^1,\dots,x^n)=0$$

with respect to x^i, $i=1,\dots,n$, while keeping in mind that x^0 is really a function of the variables $x^1,\dots,x^n$, we deduce from the chain rule that

$$f'_{x^0}\frac{\partial x^0}{\partial x^i}+f'_{x_i}=0\Rightarrow\frac{\partial x^0}{\partial x^i}=-\frac{f'_{x^i}}{f'_{x^0}}.$$

Hence

$$\boxed{\frac{\partial x^0}{\partial x^i}(\boldsymbol{u}_0)=g'_{x^i}(\boldsymbol{u}_0)=-\frac{f'_{x^i}\big(\,x^0_0,\boldsymbol{u}_0\big)}{f'_{x^0}\big(\,x^0_0,\boldsymbol{u}_0\,\big)}}. \tag{14.4.2}$$

□

Example 14.4.7. Consider the subset

$$Z=\big\{(x,y,z)\in\mathbb{R}^3;\ \ 2^{xyz}=2\,\big\}.$$

A portion of this set is depicted in Figure 14.3.

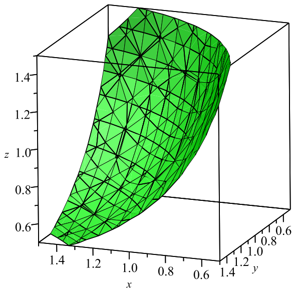

Figure 14.3. *The surface in* $\mathbb{R}^3$ *described by the equation* $2^{xyz}=2$.

Note that $Z\neq\varnothing$ since $(1,1,1)\in Z$. Equivalently, Z is the zero set of the function $f:\mathbb{R}^3\to\mathbb{R}$, $f(x,y,z)=2^{xyz}-2$. Note that

$$\frac{\partial f}{\partial z}=xy2^{xyz}\ln 2,\ \ \frac{\partial f}{\partial z}(1,1,1)=2\ln 2.$$

The implicit function theorem shows that there exists a small open ball B in $\mathbb{R}^2$ centered at $(1,1)$, an open interval $I \subset \mathbb{R}$ centered at 1 and a C^1-function $g : B \to I$ such that

$$(x,y,z) \in Z \cap \Big(B \times I \Big) \Longleftrightarrow z = g(x,y).$$

In other words, in an open neighborhood of $(1,1,1)$, the set Z is the graph of a C^1-function $z = z(x,y)$. Let us compute the partial derivatives of $z(x,y)$ at $(1,1)$.

Derivating with respect to x the equality $2^{xyz} = 2$ in which we treat z as a function of the variables (x,y) we deduce

$$(yz + xy\partial_x z)2^{xyz} \ln 2 = 0 \Rightarrow x\frac{\partial z}{\partial x} + zy = 0 \Rightarrow \frac{\partial z}{\partial x} = -\frac{zy}{x}.$$

When $(x,y) = (1,1)$, we have $z = 1$ and we deduce

$$\frac{\partial z}{\partial x}(1,1) = -1.$$

We can give an alternate verification of this equality. Namely, observe that we can solve for z *explicitly* the equality $2^{xyz} = 2$. More precisely, we have

$$\log_2 \Big(2^{xyz} \Big) = \log_2(2) \Rightarrow xyz = 1 \Rightarrow z = \frac{1}{xy} \Rightarrow \frac{\partial z}{\partial x} = -\frac{1}{x^2 y} \Rightarrow \frac{\partial z}{\partial x}(1,1) = -1. \qquad \square$$

Example 14.4.8. Consider the map

$$\boldsymbol{F} : \mathbb{R}^3 \to \mathbb{R}^2, \;\; \boldsymbol{F}(x,y,z) = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} xyz - 2 \\ x + y + z - 4 \end{bmatrix}.$$

The zero set Z of $\boldsymbol{F}$ consists of the points $(x,y,z) \in \mathbb{R}^3$ satisfying the equations

$$xyz = 2, \;\; x + y + z = 4. \tag{14.4.3}$$

Note that $(1,1,2) \in Z$. The Jacobian of the map $\boldsymbol{F}$ at a point $(x,y,z) \in \mathbb{R}^3$ is the 2×3-matrix

$$J = J(x,y,z) = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\ \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \end{bmatrix} = \begin{bmatrix} yz & xz & xy \\ 1 & 1 & 1 \end{bmatrix}.$$

Consider the minor of the above matrix determined by the y and z columns,

$$\det \begin{bmatrix} xz & xy \\ \\ 1 & 1 \end{bmatrix} = xz - xy = x(y - z).$$

Note that this minor is nonzero at the point $(x,y,z) = (1,1,2)$. The implicit function theorem then implies that, near $(1,1,2)$ we can solve (14.4.3) for y,z in terms of x. In other words, there exists a tiny (open) box B centered at $(1,1,2)$, such that the intersection of Z with B coincides with the graph of a C^1-map

$$I \ni x \mapsto \big(y(x), z(x) \big) \in \mathbb{R}^2,$$

where I is some open interval on the x-axis centered at $x=1$. To find the derivatives of $y(x)$ and $z(x)$ at $x=1$ we derivate (14.4.3) with respect to x keeping in mind that y and z depend on x. We deduce

$$yz + xzy' + xyz' = 0,\ \ 1 + y' + z' = 0.$$

At the point $(x,y,z) = (1,1,2)$ we have $yz = xz = 2$, $xy = 1$, and the above equations become

$$\begin{cases} 2y' + z' &= -2\\ y' + z' &= -1. \end{cases}$$

Note that the matrix of this linear system is the sub-matrix of $J(1,1,2)$ corresponding to the y,z columns. This matrix is nondegenerate so we can solve uniquely the above system. In fact, if we subtract the second equation from the first we deduce $y' = -1$. Using this information in the 2nd equation we deduce $z' = 0$. Hence

$$y'(x)\big|_{x=1} = -1,\ \ z'(x)\big|_{x=1} = 0. \qquad \square$$

14.5. Submanifolds of $\mathbb{R}^n$

The implicit function theorem discussed in the previous section leads to a very important concept that clarifies and generalizes our intuitive concepts of curves and surfaces.

14.5.1. Definition and basic examples. A submanifold of dimension m in the n-dimensional Euclidean space $\mathbb{R}^n$ is a set that locally "feels" like an m-dimensional vector subspace of $\mathbb{R}^n$. This is not very precise and we will address this lack of precision in Definition 14.5.1. Before we do this we want to build some intuition. Let us consider a controversy that plagued the humanity for centuries.

We now know that the surface of the Earth is spherical, but this was not what people initially believed. Anybody that walked in a wide open field could see clearly that the Earth is "obviously flat" as far as the eyes can see. The problem is that "*as far as the eyes can see*" is not far enough when compared to the size of the Earth. Our eyesight can only reach as far as the horizon: this is where the Earth's surface begins "to bend".

This phenomenon is not restricted to spheres. Take a surface in $\mathbb{R}^3$, say the surface in Figure 14.4. Any tiny region on this surface is nearly flat, and it can appear to be so to an inhabitant on this surface.

Another way to express it is to say that *any* tiny region on this surface together with a tiny region around it but outside the surface can be straightened so it now looks like a tiny region of the vector subspace $\mathbb{R}^2$ sitting in $\mathbb{R}^3$.

For example, the origin $(0,0,0)$ lives on the surface depicted in Figure 14.4. In Figure 14.5 we depicted the image of the tiny region $|x|, |y| < 0.2$ of this surface containing the origin magnified by a factor of 10. In fact, if we push the magnification factor to ∞, then this tiny region will approach a two-dimensional vector subspace of $\mathbb{R}^3$ that is intimately related to the surface namely, the plane tangent to the surface at the origin.

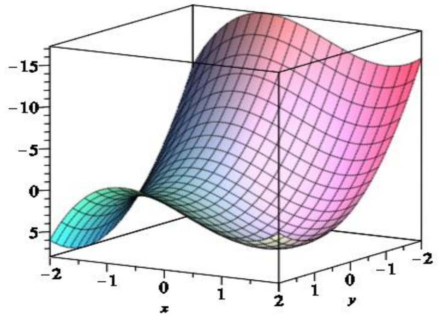

Figure 14.4. *The surface $z = -x^3 + x^2 - 2y^2 + 3x - 4y$, $|x|, |y| < 2$.*

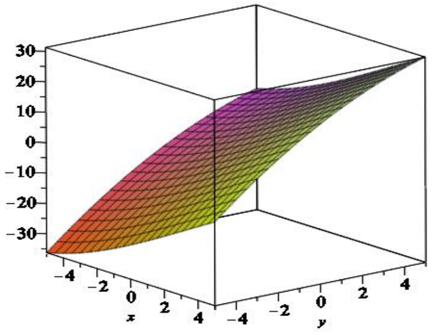

Figure 14.5. *The tiny region of the surface $z = -x^3 + x^2 - 2y^2 + 3x - 4y$ corresponding to $|x|, |y| < 0.2$ could seem flat under magnification.*

The local straightening property is indeed the defining feature of a surface in $\mathbb{R}^3$. The next definition is a mouthful but it describes in precise terms the essential features of a surface and its higher dimensional cousins, the submanifolds of Euclidean spaces. Let $n \in \mathbb{N}$, $m \in \mathbb{N}_0$, $m \leqslant n$ and $k \in \mathbb{N} \cup \{\infty\}$.

Definition 14.5.1 (Submanifolds)**.** An *m-dimensional C^k-submanifold* of $\mathbb{R}^n$ is a subset $X \subset \mathbb{R}^n$ such that, for any $\boldsymbol{p}_0 \in X$, there exists a pair $(\mathcal{U}, \Psi)$ with the following properties.

(i) $\mathcal{U}$ is an open neighborhood of $\boldsymbol{p}_0$ in $\mathbb{R}^n$,

(ii) $\Psi : \mathcal{U} \to \mathbb{R}^n$ is a C^k-diffeomorphism. We set $\boldsymbol{q}_0 := \Psi(\boldsymbol{p}_0)$, $U := \Psi(\mathcal{U})$.

(iii) If $\mathbb{R}^m \times \mathbf{0}$ denotes the coordinate subspace

$$\mathbb{R}^m \times \mathbf{0} = \Big\{ \big(x^1, \dots, x^m, x^{m+1}, \dots, x^n \big) \in \mathbb{R}^n; \;\; x^{m+1} = \cdots = x^n = 0 \Big\},$$

then $\boldsymbol{q}_0 \in \mathbb{R}^m \times \mathbf{0}$ and $\Psi\big(X \cap \mathcal{U} \big) = \big(\mathbb{R}^m \times \mathbf{0} \big) \cap U$.

An open set $\mathcal{U}$ as above is called a *coordinate neighborhood* of p_0 *adapted to* X. The pair $(\mathcal{U}, \Psi)$ is called a *straightening diffeomorphism* near $\boldsymbol{p}_0$. The induced map $\Psi : X \cap \mathcal{U} \to \mathbb{R}^m$ is called a *local coordinate chart* of X at $\boldsymbol{p}_0$. The inverse map $\Psi^{-1} : \big(\mathbb{R}^m \times \mathbf{0}\big) \cap U \to X \cap \mathcal{U}$ is called a *local parametrization* of X near $\boldsymbol{p}_0$; see Figure 14.6. □

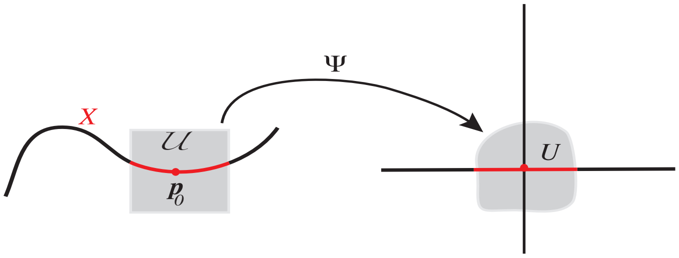

Figure 14.6. *The map Ψ straightens the "curved" portion of X located in $\mathcal{U}$.*

Remark 14.5.2. (a) A local chart maps a piece of the m-dimensional submanifold X bijectively onto an open subset of the "flat" m-dimensional space $\mathbb{R}^m$. A local parametrization of X "deforms" an open subset of the "flat" m-dimenisonal space $\mathbb{R}^m$ bijectively onto a piece of the m-dimensional submanifold.

Intuitively, a 1-dimensional submanifold of $\mathbb{R}^3$ is a curve, while a 2-dimensional submanifold of $\mathbb{R}^3$ is a surface.

(b) If $X \subset \mathbb{R}^n$ is an m-dimensional submanifold, $\boldsymbol{p}_0 \in X$ and $\mathcal{U}$ is a coordinate neighborhood of $\boldsymbol{p}_0$ adapted to X, then any open neighborhood $\mathcal{V}$ of $\boldsymbol{p}_0$ in $\mathbb{R}^n$ such that $\mathcal{V} \subset \mathcal{U}$ is also a coordinate neighborhood of $\boldsymbol{p}_0$ adapted to X. □

Example 14.5.3. (a) A point in $\mathbb{R}^n$ is a 0-dimensional submanifold of $\mathbb{R}^n$. An open subset of $\mathbb{R}^n$ is an n-dimensional submanifold of $\mathbb{R}^n$.[4] □

Our next result is a direct consequence of the inverse function theorem and describes an alternate characterization of submanifolds.

Proposition 14.5.4 (Parametric description of a submanifold)**.** *Let $m, n \in \mathbb{N}$, $m < n$. Suppose that $U \subset \mathbb{R}^m$ is an open set and*

$$\Phi : U \to \mathbb{R}^n,\ \ \Phi(\boldsymbol{u}) = \begin{bmatrix} \Phi^1(\boldsymbol{u}) \\ \vdots \\ \Phi^n(\boldsymbol{u}) \end{bmatrix}$$

*is a C^k-*parametrization*, i.e., a C^k-map satisfying the following properties.*

(i) *The map Φ is injective.*

[4] Can you verify this claim?

(ii) *The map Φ is an immersion, i.e., for any $\boldsymbol{u} \in U$ the $n \times m$ Jacobian matrix*

$$J_\Phi := \Big(\partial_{u^j} \Phi^i(\boldsymbol{u}) \Big)_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}}$$

has maximal rank m, i.e., it is injective when viewed as a linear operator $\mathbb{R}^m \to \mathbb{R}^n$.

(iii) *The inverse $\Phi^{-1} : \Phi(U) \to U$ is continuous.*

Then the following hold.

(A) *The set $X = \Phi(U)$ is an m-dimensional C^k-submanifold of $\mathbb{R}^n$. (The map Φ is referred to as a* parametrization *of X.)*

(B) *If $\ell \in \mathbb{N}$, $V \subset \mathbb{R}^\ell$ is an open set and $\boldsymbol{G} : V \to \mathbb{R}^n$ is a C^k-map such that $\boldsymbol{G}(V) \subset X$, then the map $\Phi^{-1} \circ \boldsymbol{G} : V \to U$ is C^k.*

Proof. Fix $\boldsymbol{u}_0 \in U$ and set $\boldsymbol{x}_0 := \Phi(\boldsymbol{u}_0)$. The Jacobian matrix $J_\Phi(\boldsymbol{u}_0)$ is an $n \times m$ matrix with (maximal) rank m. Thus (see [40, Thm.6.1]) there exist m rows such that the matrix determined by these rows and all the m columns of $J_\Phi(\boldsymbol{x}_0)$ is invertible. Without loss of generality we can assume that these m rows are the first m rows. We denote by $J_\Phi^m(\boldsymbol{u}_0)$ this $m \times m$ matrix. Define

$$\boldsymbol{F} : U \times \mathbb{R}^{n-m} \to \mathbb{R}^n, \quad \boldsymbol{F}(\boldsymbol{u}, \boldsymbol{v}) = \begin{bmatrix} \Phi^1(\boldsymbol{u}) \\ \vdots \\ \Phi^m(\boldsymbol{u}) \\ \Phi^{m+1}(\boldsymbol{u}) + v^1 \\ \vdots \\ \Phi^n(\boldsymbol{u}) + v^n \end{bmatrix}.$$

Note that $\boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{0}) = \Phi(\boldsymbol{u}_0) = \boldsymbol{p}_0$. The Jacobian matrix of $\boldsymbol{F}$ and $(\boldsymbol{u}_0, \boldsymbol{0})$ is the $n \times n$ matrix with the block decomposition

$$J_{\boldsymbol{F}}(\boldsymbol{u}_0, \boldsymbol{0}) = \begin{bmatrix} J_\Phi^m(\boldsymbol{u}_0) & \boldsymbol{0}_{m \times (n-m)} \\ & \\ A_{(n-m) \times m} & \mathbb{1}_{n-m} \end{bmatrix},$$

where $\boldsymbol{0}_{m \times (n-m)}$ denotes the $m \times (n-m)$ matrix with all entries equal to 0 and $A_{(n-m) \times m}$ is an $(n-m) \times m$ matrix whose explicit description is irrelevant for our argument.

Since $\det J_\Phi^m(\boldsymbol{u}_0)$ is invertible, we deduce that $\det J_{\boldsymbol{F}}(\boldsymbol{u}_0, \boldsymbol{0}) \neq 0$, so $J_{\boldsymbol{F}}(\boldsymbol{u}_0, \boldsymbol{0})$ is invertible. We can then apply the Inverse Function Theorem to conclude that there exists $\rho > 0$ sufficiently small such that the restriction of $\boldsymbol{F}$ to the open set $B_\rho^m(\boldsymbol{u}_0) \times B_\rho^{n-m}(\boldsymbol{0}) \subset \mathbb{R}^m \times \mathbb{R}^{n-m}$ is a diffeomorphism. Since $\boldsymbol{F}^{-1}$ is continuous, there an open neighborhood $\mathcal{O}$ of $\boldsymbol{p}_0$ in $\mathbb{R}^n$ such that

$$\mathcal{O} \cap \Phi(U) = \Phi\big(B_\rho^m(\boldsymbol{u}_0) \big)$$

Now choose $r < \rho$ sufficiently small so that, if $W_r = B_r^m(\boldsymbol{u}_0) \times B_r^{n-m}(\boldsymbol{0})$, then $\mathcal{W}_r := \boldsymbol{F}(W_r) \subset \mathcal{O}$.

The inverse $\Psi : \mathcal{W}_r \to W_r \subset \mathbb{R}^m \times \mathbb{R}^{n-m} = \mathbb{R}^n$ is a local straightening of $X = \Phi(U)$ around $\Phi(\boldsymbol{u}_0)$. It sends $\mathcal{W}_r \cap X$ to the m-dimensional ball $B_r^m(\boldsymbol{u}_0) \times \boldsymbol{0}_{n-m} \subset \mathbb{R}^m \times \mathbb{R}^{n-m}$.

Suppose $\boldsymbol{G}$ is as in the statement of the proposition. Fix $\boldsymbol{v}_0 \in \mathbb{V}$. Then there exists $\boldsymbol{u}_0 \in U$ such that $\Phi(\boldsymbol{u}_0) = \boldsymbol{G}(\boldsymbol{v}_0)$. Choose a local straightening Ψ of X around $\Phi(\boldsymbol{u}_0)$. Now observe that the restriction $\Phi^{-1} \circ \boldsymbol{G}$ to the *open* neighborhood $\boldsymbol{G}^{-1}(\mathcal{W})$ of $\boldsymbol{v}_0$ is equal to $\Psi \circ \boldsymbol{G}$. □

Remark 14.5.5. A C^1-map $\Phi : I \to \mathbb{R}^n$, $I \subset \mathbb{R}$ interval, is an immersion if and only if the derivative $\Phi'(t)$ is nonzero for any $t \in I$. □

Example 14.5.6. (a) Let $\boldsymbol{p} \in \mathbb{R}^n$, $\boldsymbol{v} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$. The line $\ell_{\boldsymbol{p},\boldsymbol{v}}$ is a 1-dimensional submanifold. To see this consider the map

$$\boldsymbol{\gamma} : \mathbb{R} \to \mathbb{R}^n, \;\; \boldsymbol{\gamma}(t) = \boldsymbol{p} + t\boldsymbol{v}.$$

This is an immersion since $\dot{\boldsymbol{\gamma}}(t) = \boldsymbol{v} \neq \boldsymbol{0}$, $\forall t \in \mathbb{R}$. It is also an injection since $\boldsymbol{v} \neq \boldsymbol{0}$. According to (11.1.5), the image of $\boldsymbol{\gamma}$ is the line $\ell_{\boldsymbol{p},\boldsymbol{v}}$. The inverse

$$\boldsymbol{\gamma}^{-1} : \ell_{\boldsymbol{p},\boldsymbol{v}} \to \mathbb{R}$$

is given by

$$\boldsymbol{\gamma}^{-1}(q) = \frac{1}{\|\boldsymbol{v}\|}\langle q - p, \boldsymbol{v}\rangle.$$

The above map is clearly continuous.

(b) Consider the map $\Phi : (0, \pi) \to \mathbb{R}^2$,

$$\Phi(\theta) = (\cos\theta, \sin\theta).$$

This map is injective (why?) and it is an immersion since

$$\Phi'(\theta) = (-\sin\theta, \cos\theta), \;\; \|\Phi'(\theta)\|^2 = 1 \neq 0.$$

Its image is the half circle centered at the origin and contained in the upper half space $\{y > 0\}$. Its inverse associates to a point (x, y) on this circle the angle $\theta = \arccos x \in (0, \pi)$. The map $(x, y) \mapsto \arccos x$ is obviously continuous.

(c) A helix H in $\mathbb{R}^3$ is a curve described by the parametrization (see Figure 14.7)

$$\boldsymbol{\alpha} : (0, 1) \to \mathbb{R}^3, \;\; \boldsymbol{\alpha}(t) = \big(\, r\cos(at), r\sin(at), bt\,\big),$$

where r, a, b are fixed *nonzero* real numbers $r > 0$. Note that the above map is an immersion since

$$\|\dot{\boldsymbol{\alpha}}(t)\|^2 = a^2r^2\sin^2 t + a^2r^2\cos^2 t + b^2 = a^2r^2 + b^2 \neq 0.$$

The map $\boldsymbol{\alpha}$ is clearly injective since its third component bt is such. Its inverse associates to a point (x, y, z) on the helix, the real number $t = z/b$. The map $(x, y, z) \mapsto z/b$ is obviously continuous. In Figure 14.7 we have depicted an example of helix with $r = 1$, $a = 4\pi$, $b = 1$.

(d) Consider the map $\Phi : (-\pi, \pi) \times (-\pi, \pi) \to \mathbb{R}^3$ given by

$$\Phi(\theta, \varphi) = \begin{bmatrix} (3 + \cos\varphi)\cos\theta \\ (3 + \cos\varphi)\sin\theta \\ \sin\varphi \end{bmatrix}.$$

This is an injective immersion; see Exercise 14.13. Its image is the torus in Figure 14.8. □

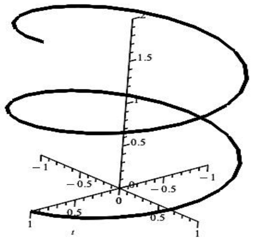

Figure 14.7. *The helix described by the parametrization* $\big(\cos(4\pi t), \sin(4\pi t), 2t\big)$ *is winding up a cylinder of radius* $r = 1$. *During one second, it winds twice around the axis of the cylinder while climbing up* 2 *units of distance.*

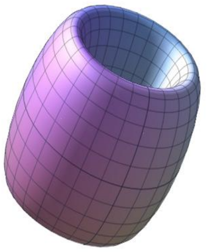

Figure 14.8. *A two-dimensional torus in* $\mathbb{R}^3$.

Corollary 14.5.7 (Graphical description of a submanifold)**.** *Let* $m, k \in \mathbb{N}$. *Suppose that* $U \subset \mathbb{R}^m$ *is an open set and* $\boldsymbol{F} : \mathbb{R}^m \to \mathbb{R}^k$ *is a* C^1*-map. Then the graph of* $\boldsymbol{F}$,

$$\Gamma_{\boldsymbol{F}} := \Big\{ \big(\boldsymbol{x}, \boldsymbol{F}(\boldsymbol{x})\big) \in \mathbb{R}^m \times \mathbb{R}^k;\;\; \boldsymbol{x} \in U \Big\} \subset \mathbb{R}^m \times \mathbb{R}^k \cong \mathbb{R}^{m+k},$$

is an m*-dimensional* C^1*-submanifold of* $\mathbb{R}^m \times \mathbb{R}^k$.

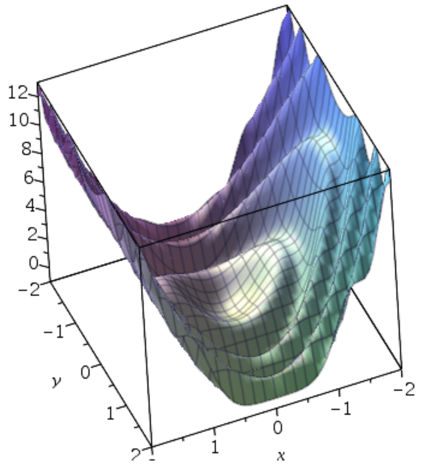

Figure 14.9. *The graph of the map* $f : (-2,2) \times (-2,2) \to \mathbb{R}$, $f(x,y) = 3x^2 + \sin(3x^2 + 3y^2)$ *is a 2-dimensional submanifold of* $\mathbb{R}^3$.

Proof. [5] Observe that the map $\Phi : \mathbb{R}^m \to \mathbb{R}^{m+k}$, $\Phi(\boldsymbol{x}) = \big(\,\boldsymbol{x}, \boldsymbol{F}(\boldsymbol{x})\,\big)$ is a parametrization. The conclusion now follows from Proposition 14.5.4. □

Remark 14.5.8. The condition (iii) in Proposition 14.5.4 is difficult to verify in concrete situations. However, the parametrizations play an important role in integration problems and it would be desirable to have a simple way of recognizing them. We mention below, without proof, one such method.

Suppose that $X \subset \mathbb{R}^n$ is an m-dimensional submanifold, $U \subset \mathbb{R}^m$ is an open set and $\Phi : U \to \mathbb{R}^n$ is an injective immersion such that $\Phi(U) \subset X$. Then Φ is a parametrization, i.e., it satisfies assumption (iii) in Proposition 14.5.4. □

The implicit function theorem coupled with the above corollary imply immediately the following result. We let the reader supply the proof.

Proposition 14.5.9 (Implicit description of a submanifold)**.** *Let* $k, m, n \in \mathbb{N}$, $m < n$. *Suppose that* U *is an open subset of* $\mathbb{R}^n$ *and* $\boldsymbol{F} : U \to \mathbb{R}^m$ *is a* C^k*-map satisfying*

$$\forall \boldsymbol{x} \in U : \quad \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0} \Rightarrow \text{ the differential } d\boldsymbol{F}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m \text{ is surjective.} \tag{14.5.1}$$

Then the set

$$\{\boldsymbol{F} = \boldsymbol{0}\} := \big\{\,\boldsymbol{x} \in U \subset \mathbb{R}^n; \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}\,\big\}$$

is an $(n-m)$*-dimensional* C^k*-submanifold of* $\mathbb{R}^n$. □

Remark 14.5.10. Let us rephrase the above result in, hopefully, more intuitive terms.

[5] Remember the simple trick used in this proof. It will come in handy later.

Recall that an $m \times n$ matrix $m < n$ has maximal rank if and only if its rows are linearly independent. The differential $d\boldsymbol{F}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$ is surjective if and only if it has maximal rank m. Denote by $F^1, \dots, F^m$ the components map $\boldsymbol{F} : U \to \mathbb{R}^m$. Then the rows of $J_{\boldsymbol{F}}(\boldsymbol{x})$ correspond to the gradients of the components F^j.

The zero set $\{\boldsymbol{F} = \boldsymbol{0}\}$ is a subset $U \subset \mathbb{R}^n$ described by m equations in n unknowns

$$F^1(x^1, \dots, x^n) = 0, \cdots, F^m(x^1, \dots, x^n) = 0.$$

The condition (14.5.1) is equivalent with the following *transversality* property.

If $\boldsymbol{x} \in U$ and $F^1(\boldsymbol{x}) = \cdots = F^m(\boldsymbol{x}) = 0$, then the gradients $\nabla F^1(\boldsymbol{x}), \dots, \nabla F^m(\boldsymbol{x})$ are linearly independent.

The above result shows that if the transversality condition is satisfied, then the common zero locus

$$\mathcal{Z}(F^1, \dots, F^m) := \big\{\, \boldsymbol{x} \in U;\ \ F^1(\boldsymbol{x}) = \cdots = F^m(\boldsymbol{x}) = 0 \,\big\}$$

is a C^k-submanifold of $\mathbb{R}^n$ of dimension $n - m$. In this case we say that the submanifold $\mathcal{Z}(F^1, \dots, F^m)$ $\mathcal{Z}$ *is cut out transversally* by the equations $F^j(\boldsymbol{x}) = 0$, $j = 1, \dots, m$.

Note that the transversality is automatically satisfied if $\boldsymbol{F}(\boldsymbol{x})$ is a *submersion*, i.e., for any $\boldsymbol{x} \in U$, the differential $d\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$ is surjective.

As an illustration, consider the situation in Example 14.4.8. There we proved that the equations

$$xyz = 2, \quad x + y + z = 4$$

satisfy the transversality conditions in a small open neighborhood U of the point $(1, 1, 2)$. Thus in this neighborhood these equations describe a submanifold of dimension $3 - 2 = 1$, i.e., a curve. □

From Propositions 14.5.4 and 14.5.9 we deduce the following useful characterization of submanifolds.

Theorem 14.5.11. *Let $m, n \in \mathbb{N}$, $m < n$, and $X \subset \mathbb{R}^n$. The following statements are equivalent.*

(i) *The set X is an m-dimensional submanifold of $\mathbb{R}^n$.*

(ii) *For any $\boldsymbol{x}_0 \in X$ there exists an open neighborhood V of $\boldsymbol{x}_0$ and a submersion $\boldsymbol{F} : V \to \mathbb{R}^{n-m}$ such that*

$$V \cap X = \big\{\, \boldsymbol{x} \in V;\ \ \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0} \,\big\}.$$

(iii) *For any $\boldsymbol{x}_0 \in X$ there exists an open neighborhood V of $\boldsymbol{x}_0$, an open neighborhood U of $\boldsymbol{0}$ in $\mathbb{R}^m$ and a parametrization $\Phi : U \to \mathbb{R}^n$ such that*

$$V \cap X = \Phi\big(U\big).$$

□

Outline of the proof. Proposition 14.5.4 shows that (iii) ⇒ (i) while Proposition 14.5.9 shows that (ii) ⇒ (i). The opposite implications (i) ⇒ (ii) and (i) ⇒ (iii) follow from the definition of a submanifold. □

Example 14.5.12 (The unit circle)**.** The unit circle is the closed subset of $\mathbb{R}^2$ defined by

$$\mathbb{S}^1 := \left\{ (x,y) \in \mathbb{R}^2;\;\; x^2+y^2=1 \right\}.$$

Then $\mathbb{S}^1$ is a curve in $\mathbb{R}^2$, i.e., a 1-dimensional submanifold of $\mathbb{R}^2$. To see this consider the smooth function

$$f:\mathbb{R}^2\to\mathbb{R},\;\; f(x,y)=x^2+y^2-1.$$

Then $\mathbb{S}^1$ can be identified with the level set $\{f=0\}$. Note that $df=2xdx+2ydy$. Hence if $(x_0,y_0)\in\mathbb{S}^1$, then at least one of the coordinates x_0,y_0 is nonzero so that $df(x_0,y_0)\neq\mathbf{0}$ proving that the differential $df(x_0,y_0):\mathbb{R}^2\to\mathbb{R}$ is surjective. The implicit function theorem then implies that $\mathbb{S}^1$ is a 1-dimensional submanifold of $\mathbb{R}^2$. There are several ways of constructing useful local coordinates.

For example, in the region $y>0$ the correspondence

$$\mathbb{S}^1\cap\{y>0\}\to\mathbb{R},\;\; (x,y)\mapsto x$$

is a local coordinate chart. The corresponding parametrization is the map

$$(-1,1)\to\mathbb{S}^1\cap\{y>0\},\;\; x\mapsto\big(x,\sqrt{1-x^2}\,\big).$$

This follows from the fact that the portion $\mathbb{S}^1\cap\{y>0\}$ is the graph of the smooth map

$$F:(-1,1)\to\mathbb{R},\;\; F(x)=\sqrt{1-x^2}.$$

Another very convenient choice is that of *polar coordinates*.

The location of a point $\boldsymbol{p}=(x,y)$ in the Cartesian plane $\mathbb{R}^2$, other than the origin $\mathbf{0}$, is uniquely determined by two parameters: the distance to the origin $r=\|\boldsymbol{p}\|=\sqrt{x^2+y^2}$, and the angle θ the vector $\boldsymbol{p}$ makes with the x-axis, measured *counterclockwisely*; see Figure 14.10.

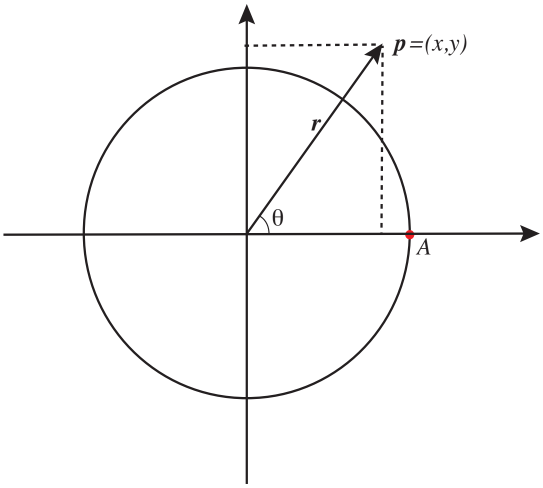

Figure 14.10. *Constructing the polar coordinates.*

The Cartesian coordinates x, y are related to the parameters r, θ via the equalities

$$\begin{cases} x &= & r\cos\theta \\ y &= & r\sin\theta. \end{cases} \tag{14.5.2}$$

Exercise 14.11 shows that the map

$$\Psi : (0, \infty) \times (0, 2\pi) \to \mathbb{R}^2, \ \ (r, \theta) \mapsto (r\cos\theta, r\sin\theta)$$

is a diffeomorphism whose image is the region $\mathbb{R}^2_*$, the plane $\mathbb{R}^2$ with the nonnegative x-semiaxis removed. The parameters (r, θ) are called *polar coordinates*. Denote by $\mathbb{S}^1_*$ the circle $\mathbb{S}^1$ with the point $A = (1, 0)$ removed; see Figure 14.10. The inverse of the diffeomorphism Ψ maps $\mathbb{S}^1$ to a portion line $r = 1$ in the (r, θ) plane. The correspondence that associates to a point $\boldsymbol{p} \in \mathbb{S}^1_*$ the angle θ is a local coordinate chart. □

Example 14.5.13 (The unit sphere). The unit sphere is the closed subset $\mathbb{S}^2$ of $\mathbb{R}^3$ defined by

$$\mathbb{S}^2 := \left\{ (x, y, z) \in \mathbb{R}^3; \ \ x^2 + y^2 + z^2 = 1 \right\}.$$

Arguing exactly as in the case of the unit circle, we can invoke the implicit function theorem to deduce that $\mathbb{S}^2$ is a surface in $\mathbb{R}^3$, i.e., 2-dimensional submanifold of $\mathbb{R}^3$.

Besides the Cartesian coordinates in $\mathbb{R}^3$ there are two other particularly useful choices of coordinates. To describe them pick a point $\boldsymbol{p} = (x, y, z) \in \mathbb{R}^3$ not situated on the z-axis. Note that the location of $\boldsymbol{p}$ is completely known if we know the altitude z of $\boldsymbol{p}$ and the location of the projection of $\boldsymbol{p}$ on the (x, y)-plane. We denote by $\boldsymbol{q}$ this projection so that $\boldsymbol{q} = (x, y) \in \mathbb{R}^2$; see Figure 14.11.

The location of $\boldsymbol{q}$ is completely determined by its polar coordinates (r, θ) so that the location of $\boldsymbol{p}$ is completely determined if we know the parameters r, θ, z. These parameters are called the *cylindrical coordinates* in $\mathbb{R}^3$. The Cartesian coordinates are related to the cylindrical coordinates via the equalities

$$\begin{cases} x &= r\cos\theta \\ y &= r\sin\theta \\ z &= z. \end{cases} \tag{14.5.3}$$

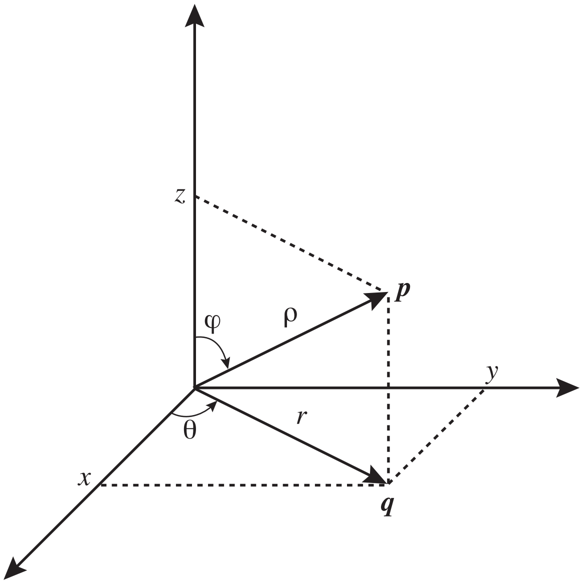

Figure 14.11. *Constructing the cylindrical and spherical coordinates.*

Observe that if we know the distance ρ of $\boldsymbol{p}$ to the origin, $\rho = \|\boldsymbol{p}\| = \sqrt{x^2+y^2+z^2}$, and the angle $\varphi \in (0,\pi)$ the vector $\boldsymbol{p}$ makes with the z-axis, then we can determine the altitude z via the equality $z = \rho\cos\varphi$ and the parameter r via the equality $r = \rho\sin\varphi$; see Figure 14.11. This shows that the parameters r, θ, φ uniquely determine the location of $\boldsymbol{p}$. These parameters are called the *spherical coordinates* in $\mathbb{R}^3$.

The Cartesian coordinates are related to the spherical coordinates via the equalities

$$\begin{cases} x &= \rho\sin\varphi\cos\theta \\ y &= \rho\sin\varphi\sin\theta \\ z &= \rho\cos\varphi. \end{cases} \tag{14.5.4}$$

Exercise 14.11 shows that the equalities (14.5.3) and (14.5.4) describe diffeomorphisms defined on certain open subsets of $\mathbb{R}^3$. Note that in spherical coordinates the unit sphere $\mathbb{S}^2$ is described by the very simple equation $\rho = 1$. The position of a point on $\mathbb{S}^2$ not situated at the poles is completely determined by the two angles φ and θ. Intuitively, φ gives the Latitude of the point, while θ determines the Longitude. □

14.5.2. Tangent spaces.

Definition 14.5.14. Let $m, n \in \mathbb{N}$, $m \leqslant n$, suppose that $X \subset \mathbb{R}^n$ is an m-dimensional submanifold of $\mathbb{R}^n$ and $\boldsymbol{x}_0 \in X$.

(i) A *path in X through $\boldsymbol{x}_0$* is a C^1-path $\boldsymbol{\gamma} : I \to \mathbb{R}^n$, $I \subset \mathbb{R}$ open interval, such that $0 \in I$, $\boldsymbol{\gamma}(0) = \boldsymbol{x}_0$ and $\boldsymbol{\gamma}(I) \subset X$; see Figure 14.12.

(ii) A vector $\boldsymbol{v} \in \mathbb{R}^n$ is said to be *tangent to X at $\boldsymbol{x}_0$* if there exists a path $\boldsymbol{\gamma} : I \to \mathbb{R}^n$ in X through $\boldsymbol{x}_0$ such that $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$. We denote by $T_{\boldsymbol{x}_0}X$ the set of vectors tangent to X at $\boldsymbol{x}_0$. We will refer to $T_{\boldsymbol{x}_0}X$ as the *tangent space to X at $\boldsymbol{x}_0$*.

□

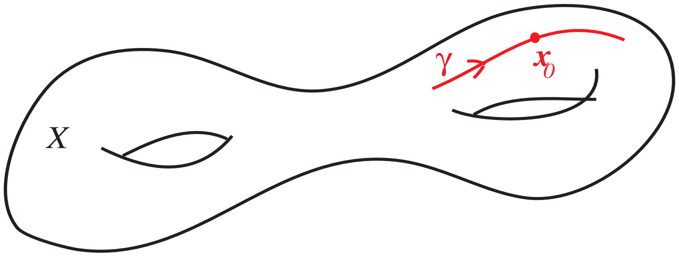

Figure 14.12. *A path $\boldsymbol{\gamma}$ in the surface $X \subset \mathbb{R}^3$ through a point $\boldsymbol{x}_0 \in X$. The velocity of $\boldsymbol{\gamma}$ at $\boldsymbol{x}_0$ is, by definition, a vector tangent to X at $\boldsymbol{x}_0$.*

Example 14.5.15. Let $m, n \in \mathbb{N}$, $m < n$. Denote by $\mathbb{R}^m \times \mathbf{0}$ the subspace of $\mathbb{R}^n$ defined by the equations $x^{m+1} = \cdots = x^n = 0$. Fix an open set $U \subset \mathbb{R}^n$ and denote by Y the intersection $Y := U \cap (\mathbb{R}^m \times \mathbf{0})$. Then Y is an m-dimensional submanifold of $\mathbb{R}^n$. We want to prove that

$$T_{\boldsymbol{y}_0}Y = \mathbb{R}^m \times \mathbf{0}, \quad \forall \boldsymbol{y}_0 \in Y. \tag{14.5.5}$$

In particular $T_{\boldsymbol{y}_0}Y$ is a vector subspace of $\mathbb{R}^n$.

Clearly $T_{\boldsymbol{y}_0}Y \subset \mathbb{R}^m \times \mathbf{0}$. To see this observe that if $\boldsymbol{\gamma} : I \to \mathbb{R}^n$ is a path in Y through $\boldsymbol{y}_0$, then it has the form

$$\boldsymbol{\gamma}(t) = \begin{bmatrix} \gamma^1(t) \\ \vdots \\ \gamma^m(t) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n.$$

In particular, $\dot{\boldsymbol{\gamma}}(0) \in \mathbb{R}^m \times \mathbf{0}$.

To prove that $\mathbb{R}^m \times \mathbf{0} \subset T_{\boldsymbol{y}_0} Y$ consider a vector

$$\boldsymbol{v} = \begin{bmatrix} v^1 \\ \vdots \\ v^m \\ \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^m \times \mathbf{0}.$$

Then, there exists $\varepsilon > 0$ sufficiently small such that $\boldsymbol{x}_0 + t\boldsymbol{v} \in U$, $\forall t \in (-\varepsilon, \varepsilon)$. The path

$$\boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to \mathbb{R}^n, \;\; \boldsymbol{\gamma}(t) = \boldsymbol{x}_0 + t\boldsymbol{v}$$

is in Y through $\boldsymbol{y}_0$ and $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$, i.e., $\boldsymbol{v} \in T_{\boldsymbol{y}_0} Y$. □

The above example is a manifestation of a more general phenomenon.

Proposition 14.5.16. *Let $m, n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$ is an m-dimensional C^1-submanifold. Then for any $\boldsymbol{x}_0 \in X$ the tangent space $T_{\boldsymbol{x}_0} X$ is an m-dimensional vector subspace of $\mathbb{R}^n$.*

Proof. Let $\boldsymbol{x}_0 \in X$. Fix a straightening diffeomorphism $\Psi : \mathcal{U} \to \mathbb{R}^n$ of X at $\boldsymbol{x}_0$ and set $U := \Psi(\mathcal{U})$. Denote by Φ the inverse $\Psi^{-1} : U \to \mathbb{R}^n$ and by L the differential of Φ at $(\boldsymbol{u}_0, \mathbf{0})$. Then $\Psi(\boldsymbol{x}_0) = (\boldsymbol{u}_0, \mathbf{0}) \in \mathbb{R}^m \times \mathbf{0} \subset \mathbb{R}^n$. Note that $\boldsymbol{\omega}$ is a path in X through $\boldsymbol{x}_0$ if and only if $\boldsymbol{\gamma} := \Psi \circ \boldsymbol{\omega}$ is a path in $\mathbb{R}^m \times \mathbf{0}$ through $(\boldsymbol{u}_0, \mathbf{0})$. Moreover $\boldsymbol{\omega} = \Phi \circ \boldsymbol{\gamma}$. Thus any path $\boldsymbol{\omega}$ in X through $\boldsymbol{x}_0$ has the form $\boldsymbol{\omega} = \Phi \circ \boldsymbol{\gamma}$ for some path $\boldsymbol{\gamma}$ in $\mathbb{R}^m \times \mathbf{0}$ through $(\boldsymbol{u}_0, \mathbf{0})$ and (see Exercise 13.7)

$$\dot{\boldsymbol{\omega}}(0) = L\dot{\boldsymbol{\gamma}}(0).$$

This proves that

$$T_{\boldsymbol{x}_0} X = L\big(T_{\boldsymbol{x}_0, \mathbf{0}} \mathbb{R}^m \times \mathbf{0}\big) \overset{(14.5.5)}{=} L\big(\mathbb{R}^m \times \mathbf{0}\big).$$

Thus $T_{\boldsymbol{x}_0} X$ is the image of the m-dimensional vector subspace $\mathbb{R}^m \times \mathbf{0}$ via the linear map L. Since L is injective, the image also has dimension m. □

The above proof leads to the following useful consequence.

Corollary 14.5.17. *Let $m, n \in \mathbb{N}$, $m < n$. Suppose that $U \subset \mathbb{R}^m$ is open and $\Phi : U \to \mathbb{R}^n$ is a parametrization (see Proposition 14.5.4) with image $X = \Phi(U)$. If $\boldsymbol{u}_0 \in U$ and $\boldsymbol{x}_0 = \Phi(\boldsymbol{u}_0)$, then the tangent space $T_{\boldsymbol{x}_0} X$ is equal to the range of the differential $d\Phi(\boldsymbol{u}_0)$, i.e.,*

$$T_{\boldsymbol{x}_0} X = d\Phi(\boldsymbol{u}_0)\mathbb{R}^m.$$

In particular, the vectors

$$\partial_{u^1}\Phi(\boldsymbol{u}_0), \; \partial_{u^2}\Phi(\boldsymbol{u}_0), \dots, \partial_{u^m}\Phi(\boldsymbol{u}_0)$$

form a basis of $T_{\boldsymbol{x}_0} X$. □

Proposition 14.5.18. *Let $m, n \in \mathbb{N}$, $m < n$. Suppose that $V \subset \mathbb{R}^n$ is open and $\boldsymbol{F} : V \to \mathbb{R}^m$ is a C^1-map such that, for any $\boldsymbol{x} \in X = \boldsymbol{F}^{-1}(\boldsymbol{0})$, the differential $d\boldsymbol{F}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$ is onto, so the Jacobian matrix $J_{\boldsymbol{F}}(\boldsymbol{x})$ has rank m. Then X is a smooth submanifold of $\mathbb{R}^n$ of dimension $n - m$ and*

$$\forall \boldsymbol{x}_0 \in X, \;\; T_{\boldsymbol{x}_0} X = \ker d\boldsymbol{F}(\boldsymbol{x}_0) = \big\{ \boldsymbol{v} \in \mathbb{R}^n ; \;\; d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{v} = 0 \big\}.$$

Proof. The fact that X is a smooth submanifold of dimension $n - m$ follows from the implicit function theorem; see Proposition 14.5.9. Let $\boldsymbol{x}_0 \in X$. The range $\boldsymbol{R}\big(d\boldsymbol{F}(\boldsymbol{x}_0) \big)$ of the linear operator $d\boldsymbol{F}(\boldsymbol{x}_0) : \mathbb{R}^n \to \mathbb{R}^m$ has dimension m since this linear operator is surjective. We deduce

$$\dim \ker d\boldsymbol{F}(\boldsymbol{x}_0) = n - \dim \boldsymbol{R}\big(d\boldsymbol{F}(\boldsymbol{x}_0) \big) = n - m = \dim T_{\boldsymbol{x}_0} X.$$

Hence it suffices to show that $T_{\boldsymbol{x}_0} X \subset \ker d\boldsymbol{F}(\boldsymbol{x}_0)$.

Let $\boldsymbol{v} \in T_{\boldsymbol{x}_0} X$. Thus, there exists a C^1-path $\boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $\boldsymbol{\gamma}(t) \in X$, $\forall t \in (-\varepsilon, \varepsilon)$, $\boldsymbol{\gamma}(0) = \boldsymbol{x}_0$, $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$ and consequently

$$\boldsymbol{F}\big(\boldsymbol{\gamma}(t) \big) = \boldsymbol{0}, \;\; \forall t \in (-\varepsilon, \varepsilon).$$

Derivating the last equality at $t = 0$ using the chain rule we deduce

$$\boldsymbol{0} = \frac{d}{dt}\Big|_{t=0} \boldsymbol{F}\big(\boldsymbol{\gamma}(t) \big) = d\boldsymbol{F}\big(\boldsymbol{\gamma}(0) \big) \dot{\boldsymbol{\gamma}}(0) = d\boldsymbol{F}(\boldsymbol{x}_0)\boldsymbol{v} \Rightarrow \boldsymbol{v} \in \ker d\boldsymbol{F}(\boldsymbol{x}_0).$$

□

Remark 14.5.19. The last result has a more geometric equivalent reformulation. The map $\boldsymbol{F}$ in Proposition 14.5.18 has m components,

$$\boldsymbol{F}(\boldsymbol{x}) = \left[\begin{array}{c} F^1(\boldsymbol{x}) \\ \vdots \\ F^m(\boldsymbol{x}) \end{array} \right].$$

The differential of $\boldsymbol{F}$ is represented by the $m \times n$ matrix

$$d\boldsymbol{F}(\boldsymbol{x}) = \left[\begin{array}{c} dF^1(\boldsymbol{x}) \\ \vdots \\ dF^m(\boldsymbol{x}) \end{array} \right],$$

where the i-th row describes the differential of F^i. Note that

$$\boldsymbol{v} \in \ker d\boldsymbol{F}(\boldsymbol{x}) \Longleftrightarrow dF^i(\boldsymbol{x})(\boldsymbol{v}) = 0, \;\; \forall i = 1, \dots, m$$

$$\Longleftrightarrow \langle \nabla F^i(\boldsymbol{x}), \boldsymbol{v} \rangle = 0, \;\; \forall i = 1, \dots, m \Longleftrightarrow \boldsymbol{v} \perp \nabla F^i(\boldsymbol{x}), \;\; \forall i = 1, \dots, m$$

$$\Longleftrightarrow v^1 \partial_{x^1} F^i(\boldsymbol{x}_0) + v^2 \partial_{x^2} F^i(\boldsymbol{x}_0) + \cdots + v^n \partial_{x^n} F^i(\boldsymbol{x}_0) = 0, \;\; \forall i = 1, \dots, m$$

□

To put the above remark in its proper geometric perspective we need to survey a few linear algebra facts. For a given vector subspace $V \subset \mathbb{R}^n$ we denote by $V^\perp$ the set of vectors $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x} \perp \boldsymbol{v}$, $\forall \boldsymbol{v} \in V$. The set $V^\perp$ is called the *orthogonal complement* of V in $\mathbb{R}^n$. Often we will use the notation $\boldsymbol{x} \perp V$ to indicate $\boldsymbol{x} \in V^\perp$. The orthogonal complement enjoys several useful properties.

Proposition 14.5.20. *Let $n \in \mathbb{N}$ and suppose that V is a vector subspace of $\mathbb{R}^n$. Then the following hold.*

(i) *The orthogonal complement $V^\perp$ is also a vector subspace of $\mathbb{R}^n$. Moreover, if the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_m$ span V, then*

$$\boldsymbol{x} \in V^\perp \Longleftrightarrow \boldsymbol{x} \perp \boldsymbol{v}_i, \quad \forall i = 1, \dots, m.$$

(ii) *For any $\boldsymbol{x} \in \mathbb{R}^n$ there exists a unique $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}) \in V$ such that $\boldsymbol{x} - \boldsymbol{v} \in V^\perp$. This vector is called the* orthogonal projection *of $\boldsymbol{x}$ on V.*

(iii) $\dim V + \dim V^\perp = n = \dim \mathbb{R}^n$.

(iv) $(V^\perp)^\perp = V$.

□

For a proof of the above proposition and additional information we refer to [40, Sec. 5.3].

Corollary 14.5.21. *Let $k, n \in \mathbb{N}$, $k < n$. Suppose that $U \subset \mathbb{R}^n$ is an open set and*

$$F^1, \dots, F^k : U \to \mathbb{R}$$

are C^1-functions. Set

$$X := \left\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ F^1(\boldsymbol{x}) = \cdots = F^k(\boldsymbol{x}) = 0 \right\}.$$

Assume that

$$\boxed{\textit{for any } \boldsymbol{x} \in X, \textit{ the vectors } \nabla F^1(\boldsymbol{x}), \dots, \nabla F^k(\boldsymbol{x}) \textit{ are linearly independent}}. \tag{14.5.6}$$

Then the following hold.

(i) *The subset X is a C^1-submanifold of dimension $m = n - k$.*

(ii) *For any $\boldsymbol{x} \in X$ we have*

$$T_{\boldsymbol{x}} X = \operatorname{span}\left\{ \nabla F^1(\boldsymbol{x}), \dots, \nabla F^k(\boldsymbol{x}) \right\}^\perp.$$

(iii) *For any $\boldsymbol{x} \in X$, $\boldsymbol{v} \in \mathbb{R}^n$ we have*

$$\boxed{\boldsymbol{v} \perp T_{\boldsymbol{x}} X \Longleftrightarrow \boldsymbol{v} \in \operatorname{span}\{\nabla F^1(\boldsymbol{x}), \dots, \nabla F^k(\boldsymbol{x})\}}. \tag{14.5.7}$$

Proof. Consider the map $\boldsymbol{F} : U \to \mathbb{R}^k$,

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} F^1(\boldsymbol{x}) \\ \vdots \\ F^k(\boldsymbol{x}) \end{bmatrix}.$$

The Jacobian matrix $J_{\boldsymbol{F}}(\boldsymbol{x})$ that represents $d\boldsymbol{F}(\boldsymbol{x})$ is a $k \times n$ matrix and its rows are described by the differentials $dF^1(\boldsymbol{x}), \dots, dF^k(\boldsymbol{x})$; see Example 13.2.7. The assumption (14.5.6) shows that for $\boldsymbol{x} \in X$ the (*row*) rank of the matrix $J_{\boldsymbol{F}}(\boldsymbol{x})$ is k. This implies that, for any $\boldsymbol{x} \in X$, the operator $J_{\boldsymbol{F}}(\boldsymbol{x})$ is onto; see [**40**, Sec. 2.7]. Proposition 14.5.18 now implies that X is a C^1-submanifold of dimension $m = n - k$.

From Remark 14.5.19 and Proposition 14.5.20(i) we deduce

$$\boxed{T_{\boldsymbol{x}}X = \big(\operatorname{span}\{\,\nabla F^1(\boldsymbol{x}), \dots, \nabla F^k(\boldsymbol{x})\,\}\,\big)^{\perp}}.$$

The equivalence (14.5.7) is now a consequence of Proposition 14.5.20(iv). □

Example 14.5.22 (Hypersurfaces)**.** A *hypersurface* in $\mathbb{R}^n$ is a C^1-submanifold of dimension $n-1$. We can use Corollary 14.5.21 to produce hypersurfaces as follows.

Suppose that $U \subset \mathbb{R}^n$ is an open subset and $f : U \to \mathbb{R}$ is a C^1-function such that

$$\boxed{\forall \boldsymbol{x} \in U,\ \ f(\boldsymbol{x}) = 0 \Rightarrow \nabla f(\boldsymbol{x}) \neq \boldsymbol{0}}.$$

Then the zero set of f,

$$X := \big\{\boldsymbol{u} \in U;\ \ f(\boldsymbol{u}) = 0\,\big\}$$

is a hypersurface in $\mathbb{R}^n$. Moreover, for all $\boldsymbol{x}_0 \in X$, the tangent space of X at $\boldsymbol{x}_0$ is a hyperplane (through $\boldsymbol{0}$) and $\nabla f(\boldsymbol{x}_0)$ is a normal vector of this hyperplane. In particular, it is described by the equation

$$T_{\boldsymbol{x}_0}X = \big\{\boldsymbol{v} \in \mathbb{R}^n;\ \ \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{v}\rangle = 0\,\big\}.$$

As an example, consider a C^1-function $h : \mathbb{R}^2 \to \mathbb{R}$. As we know, its graph

$$\Gamma_h := \big\{\,(x,y,z) \in \mathbb{R}^3;\ \ z = h(x,y)\,\big\}$$

is a hypersurface in $\mathbb{R}^3$. If we define $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = h(x,y) - z$, we see that we can alternatively characterize Γ_h as the zero set of f.

Note that for any $(x_0, y_0, z_0) \in \mathbb{R}^3$ we have

$$\nabla f(x_0, y_0, z_0) = \begin{bmatrix} \partial_x h(x_0, y_0) \\ \partial_y h(x_0, y_0) \\ -1 \end{bmatrix} \neq \boldsymbol{0}.$$

Thus the tangent space to Γ_h at $\boldsymbol{p}_0 = (x_0, y_0, z_0)$, $z_0 = h(x_0, y_0)$ consists of the vectors

$$\dot{\boldsymbol{r}} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \in \mathbb{R}^3$$

such that $\langle \nabla f(\boldsymbol{p}_0), \dot{\boldsymbol{r}}\rangle = 0$, i.e.,

$$\partial_x h(x_0,y_0)\dot{x} + \partial_y h(x_0,y_0)\dot{y} - \dot{z} \Longleftrightarrow \dot{z} = \partial_x h(x_0,y_0)\dot{x} + \partial_y h(x_0,y_0)\dot{y}.$$

We see that $T_{\boldsymbol{p}_0}\Gamma_h$ is the graph of the differential

$$dh(x_0,y_0) : \mathbb{R}^2 \to \mathbb{R},\ \ dh(x_0,y_0)(\dot{x},\dot{y}) = \partial_x h(x_0,y_0)\dot{x} + \partial_y h(x_0,y_0)\dot{y}.$$

As an even more concrete example consider the sphere

$$\Sigma := \big\{ (x,y,z) \in \mathbb{R}^3;\ \ x^2+y^2+z^2 = 3 \big\}.$$

It is the zero set of the function $f(x,y,z) = x^2+y^2+z^2-3$. Note that

$$\nabla f(x,y,z) = \begin{bmatrix} 2x \\ 2y \\ 2z \end{bmatrix} \neq \boldsymbol{0},\ \ \forall (x,y,z) \neq \boldsymbol{0}.$$

This shows that Σ is a hypersurface in $\mathbb{R}^3$. The point $\boldsymbol{p}_0 = (1,1,1)$ lives on this sphere and

$$T_{\boldsymbol{p}_0}\Sigma = \{\dot{\boldsymbol{r}} = (\dot{x},\dot{y},\dot{z}) \in \mathbb{R}^3;\ \ \dot{x}+\dot{y}+\dot{z} = 0\,\}.$$

We treat the equality $\dot{x}+\dot{y}+\dot{z} = 0$ as a homogeneous linear system consisting of one equation in the three unknowns $\dot{x},\dot{y},\dot{z}$. We see that the solutions of this system satisfy $\dot{x} = -\dot{y}-\dot{z}$ so that

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} -\dot{y}-\dot{z} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \dot{y}\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + \dot{z}\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

where $\dot{y},\dot{z}$ are arbitrary. This shows that the vectors

$$\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix},\ \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

form a basis of $T_{\boldsymbol{p}_0}\Sigma$. □

Example 14.5.23. Consider the map

$$\boldsymbol{F} : \mathbb{R}^4 \to \mathbb{R}^2,\ \ \boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} F^1(\boldsymbol{x}) \\ F^2(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \|\boldsymbol{x}\|^2 - 1 \\ \\ x^1+x^2+x^3+x^4-1 \end{bmatrix}$$

and the set

$$S := \big\{ \boldsymbol{x} \in \mathbb{R}^4;\ \ \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0} \big\}.$$

In more concrete terms, S is the locus of points $\boldsymbol{x} \in \mathbb{R}^4$ satisfying the equations

$$\begin{cases} \|\boldsymbol{x}\|^2 &= 1 \\ x^1+x^2+x^3+x^4 &= 1. \end{cases}$$

Let us observe first that $S \neq \varnothing$ since the basic vectors $\boldsymbol{e}_1,\dots,\boldsymbol{e}_4 \in \mathbb{R}^4$ belong to S. We will prove that S is a 2-dimensional submanifold. In view of Corollary 14.5.21 it suffices to verify that for any $\boldsymbol{x} \in S$ the gradients $\nabla F^1(\boldsymbol{x})$ and $\nabla F^2(\boldsymbol{x})$ are linearly independent, i.e., they are not collinear.

Observe that

$$\nabla F^1(\boldsymbol{x}) = 2\boldsymbol{x}, \;\; \nabla F^2(\boldsymbol{x}) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Note that if $\nabla F^1(\boldsymbol{x})$ and $\nabla F^2(\boldsymbol{x})$ were collinear, then

$$x^1 = x^2 = x^3 = x^4 = c.$$

Since $\boldsymbol{x} \in S$ we deduce $4c^2 = 1$ and $4c = 1$. This is obviously impossible. Hence S is a 2-dimensional submanifold. To find the tangent space of S at $\boldsymbol{e}_1 = (1, 0, 0, 0)$ observe first that

$$\nabla F^1(\boldsymbol{e}_1) = 2\boldsymbol{e}_1 = (2, 0, 0, 0), \;\; \nabla F^2(\boldsymbol{e}_1) = (1, 1, 1, 1),$$

and we deduce that $T_{\boldsymbol{e}_1}S$ consists of vectors $(\dot{x}^1, \dot{x}^2, \dot{x}^3, \dot{x}^4)$ satisfying the homogeneous linear system

$$\begin{array}{rcl} \dot{x}^1 & = & 0 \\ \dot{x}^1 + \dot{x}^2 + \dot{x}^3 + \dot{x}^4 & = & 0. \end{array}$$

This system is equivalent with the system in upper echelon form

$$\begin{array}{rcl} 2\dot{x}^1 \qquad\qquad\qquad & = & 0 \\ \dot{x}^2 + \dot{x}^3 + \dot{x}^4 & = & 0. \end{array}$$

The general solution of the last system is

$$\begin{bmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \\ \dot{x}^4 \end{bmatrix} = \dot{x}^3 \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \dot{x}^4 \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}.$$

This shows that the vectors

$$\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \;\; \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

form a basis of the tangent space $T_{\boldsymbol{e}_1}S$. □

14.5.3. Lagrange multipliers. To understand the significance of the Lagrange multiplier theorem we consider simple question that can be addressed by it.

Example 14.5.24. Find the minimum of the *cost function*

$$h : \mathbb{R}^3 \to \mathbb{R}, \;\; h(x, y, z) = x + y + z,$$

subject to the *constraint*

$$x^2 + y^2 + z^2 = 3.$$

Note that the above constraint equation defines a submanifold S in $\mathbb{R}^3$, more precisely, the sphere of radius $\sqrt{3}$ centered at the origin. The question can now be rephrased as asking to find the minimum value of the restriction of h to S.

If you think of h as describing say the temperature in $\mathbb{R}^3$ at a given moment, then the question asks to find the coldest point on the sphere S. The Lagrange multiplier theorem describes a simple criterion for recognizing (local) minima or maxima of functions defined on a submanifold of $\mathbb{R}^n$. □

Theorem 14.5.25. *Let $n \in \mathbb{N}$. Suppose that S is a submanifold of $\mathbb{R}^n$, $\mathcal{O} \subset \mathbb{R}^n$ is an open subset containing S and $h : \mathcal{O} \to \mathbb{R}$ is a C^1 function. If $\boldsymbol{x}_0$ is a local minimum (or maximum) of the restriction of h to S, then*

$$\nabla h(\boldsymbol{x}_0) \perp T_{\boldsymbol{x}_0}S, \;\; \textit{i.e.,} \; \big\langle \nabla h(\boldsymbol{x}_0), \boldsymbol{v} \big\rangle = 0, \;\; \forall \boldsymbol{v} \in T_{\boldsymbol{x}_0}S.$$

Proof. Suppose that $\boldsymbol{x}_0$ is a local minimum of the restriction of h to S. (The local maximum follows from this case applied to the function $-h$.) Let $\boldsymbol{v} \in T_{\boldsymbol{x}_0}S$. We deduce that there exists an open interval $I \subset \mathbb{R}$ containing 0 and a C^1 path $\boldsymbol{\gamma} : I \to \mathbb{R}^n$ such that $\boldsymbol{\gamma}(t) \in S$, $\forall t \in I$ and $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$.

Since $\boldsymbol{x}_0$ is a local minimum of h on S, there exists $r > 0$ such that

$$h(\boldsymbol{x}_0) \leqslant h(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in B_r(\boldsymbol{x}_0) \cap S.$$

On the other hand, since $\boldsymbol{\gamma}$ is continuous and $\boldsymbol{\gamma}(0) = \boldsymbol{x}_0$, there exists $\varepsilon > 0$ sufficiently small such that $(-\varepsilon, \varepsilon) \subset I$ and $\boldsymbol{\gamma}(t) \in B_r(\boldsymbol{x}_0)$, $\forall t \in (-\varepsilon, \varepsilon)$. Hence

$$h(\boldsymbol{\gamma}(0)) = h(\boldsymbol{x}_0) \leqslant h(\boldsymbol{\gamma}(t)), \;\; \forall t \in (-\varepsilon, \varepsilon).$$

In other words, $0 \in I$ is a local minimum of the function $h \circ \boldsymbol{\gamma} : I \to \mathbb{R}$. Fermat's theorem then implies

$$0 = \frac{d}{dt}\Big|_{t=0} h(\boldsymbol{\gamma}(t)) \overset{(13.3.12)}{=} \big\langle \nabla h(\,\boldsymbol{\gamma}(0)\,), \dot{\boldsymbol{\gamma}}(0) \big\rangle = \langle \nabla h(\boldsymbol{x}_0), \boldsymbol{v} \rangle.$$

□

Corollary 14.5.26 (Lagrange Multipliers Theorem)**.** *Let $k, n \in \mathbb{N}$, $k < n$. Suppose that $\mathcal{O} \subset \mathbb{R}^n$ is an open set, and we are given C^1-functions $h, F^1, \dots, F^k : \mathcal{O} \to \mathbb{R}$ with the property*

$$\boxed{F^1(\boldsymbol{x}) = \cdots = F^k(\boldsymbol{x}) = 0 \Rightarrow \textit{the vectors } \nabla F^1(\boldsymbol{x}), \dots, \nabla F^k(\boldsymbol{x}) \textit{ are linearly independent}}. \tag{14.5.8}$$

If $\boldsymbol{x}_0 \in \mathcal{O}$ minimizes h subject to the constraints

$$F^1(\boldsymbol{x}) = \cdots = F^k(\boldsymbol{x}) = 0,$$

then there exist real numbers $\lambda_1, \dots, \lambda_k$ such that

$$\nabla h(\boldsymbol{x}_0) = \lambda_1 \nabla F^1(\boldsymbol{x}_0) + \cdots + \lambda_k \nabla F^k(\boldsymbol{x}_0).$$

The real numbers $\lambda_1, \dots, \lambda_k$ *are called* Lagrange multipliers.[6]

Proof. In view of Corollary 14.5.21, the assumption (14.5.8) implies that the constrained set

$$S := \{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ F^1(\boldsymbol{x}) = \cdots = F^k(\boldsymbol{x}) = 0 \,\}$$

is a submanifold of $\mathbb{R}^n$ of dimension $n-k$.

If $\boldsymbol{x}_0$ is a minimum of the restriction of h on S, then Theorem 14.5.25 shows that

$$\nabla h(\boldsymbol{x}_0) \in (T_{\boldsymbol{x}_0} S)^{\perp}.$$

Corollary 14.5.21 implies that

$$\nabla h(\boldsymbol{x}_0) \in \operatorname{span}\{\, \nabla F^1(\boldsymbol{x}_0), \dots, \nabla F^k(\boldsymbol{x}_0) \,\} = (T_{\boldsymbol{x}_0} S)^{\perp}.$$

This implies the existence of numbers $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ such that

$$\nabla h(\boldsymbol{x}_0) = \lambda_1 \nabla F^1(\boldsymbol{x}_0) + \cdots + \lambda_k \nabla F^k(\boldsymbol{x}_0).$$

□

Example 14.5.27. We want to find the minimum of the function

$$h : \mathbb{R}^3 \to \mathbb{R},\ \ h(x,y,z) = x + y + z,$$

subject to the constraint

$$x^2 + y^2 + z^2 = 1.$$

The set S consisting of the points satisfying this constraint is the unit sphere in $\mathbb{R}^3$ centered at the origin. This is compact and, since h is continuous, its restriction to S has an absolute minimum attained at some point $\boldsymbol{p}_0 = (x_0, y_0, z_0)$.

Set $f(x,y,z) := x^2 + y^2 + z^2 - 1$ so the constraint is described by the equation $f(x,y,z) = 0$. Since

$$\nabla f(x,y,z) = 2 \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

we deduce that if $x^2 + y^2 + z^2 = 1$, then $\nabla f(x,y,z) \neq \mathbf{0}$ so (14.5.8) is satisfied.

Corollary 14.5.26 implies that there exists a Lagrange multiplier $\lambda \in \mathbb{R}$ such that

$$\nabla h(\boldsymbol{p}_0) = \lambda \nabla f(\boldsymbol{p}_0) \Longleftrightarrow (1,1,1) = \lambda(2x_0, 2y_0, 2z_0).$$

We obtain the system of 4 equations

$$\begin{cases} x_0^2 + y_0^2 + z_0^2 &= 1 \\ 1 &= 2\lambda x_0 \\ 1 &= 2\lambda y_0 \\ 1 &= 2\lambda z_0, \end{cases}$$

[6] Observe that the number of Lagrange multipliers is equal to the number of constraints $F^1(\boldsymbol{x}) = 0, \dots, F^k(\boldsymbol{x}) = 0$.

in 4 unknowns, x_0, y_0, z_0, λ. From the last 3 equations we deduce

$$x_0 = y_0 = z_0 = \frac{1}{2\lambda}$$

and

$$3 = (2\lambda)^2(x_0^2 + y_0^2 + z_0^2) \Rightarrow 3 = 4\lambda^2 \Rightarrow \lambda^2 = \frac{3}{4} \Rightarrow \lambda = \pm\frac{\sqrt{3}}{2}.$$

Thus $\boldsymbol{p}_0$ can only be one of the two points

$$\boldsymbol{p}_0^{\pm} = \pm\frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Since $h(\boldsymbol{p}_0^-) < 0 < h(\boldsymbol{p}_0^+)$ we deduce that the minimum of h subject to the constraint $x^2 + y^2 + z^2 = 1$ is $-\sqrt{3}$ and it is attained at the point $\boldsymbol{p}_0^-$. □

14.6. Exercises

Exercise 14.1. Let $n \in \mathbb{N}$, $r > 0$ and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a C^1-function such that $f(\boldsymbol{x}) = 0$, $\forall \boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\| = r$. Show that there exists $\boldsymbol{x}_0 \in \mathbb{R}^n$ such that

$$\|\boldsymbol{x}_0\| < r \text{ and } \nabla f(\boldsymbol{x}_0) = \boldsymbol{0}.$$

Hint. Use the proof of Rolle's Theorem 7.4.5 as inspiration. □

Exercise 14.2. Consider the symmetric 3×3-matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}.$$

Show that the associated quadratic function $Q_A : \mathbb{R}^3 \to \mathbb{R}$ satisfies

$$Q_A(x, y, z) = x^2 + 4y^2 + 6z^2 + 4xy + 6xz + 10yz, \quad \forall x, y, z \in \mathbb{R}.$$ □

Exercise 14.3. Suppose that A is a symmetric $n \times n$ matrix with associated quadratic function Q_A. Prove that

$$\nabla Q_A(\boldsymbol{x}) = 2A\boldsymbol{x}, \quad \boldsymbol{H}(Q_A, \boldsymbol{x}) = 2A, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$ □

Exercise 14.4. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a C^2-function and $\boldsymbol{p}_0 \in \mathbb{R}^n$. Show that the Hessian of f at $\boldsymbol{p}_0$ is equal to the Jacobian of the map $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ at the same point. □

Exercise 14.5. Suppose that A is a symmetric, positive definite $n \times n$ matrix. Prove that there exists $m > 0$ such that

$$Q_A(\boldsymbol{h}) \geqslant m\|\boldsymbol{h}\|^2, \quad \forall \boldsymbol{h} \in \mathbb{R}^n.$$

Hint. Set

$$\Sigma_1 := \{\boldsymbol{h} \in \mathbb{R}^n;\ \|\boldsymbol{h}\| = 1\}, \quad m := \inf_{\boldsymbol{h} \in \Sigma_1} Q_A(\boldsymbol{h}).$$

Show that Σ_1 is compact and deduce that $m > 0$. Next, use (14.2.1) to prove that $Q_A(\boldsymbol{h}) \geqslant m\|\boldsymbol{h}\|^2$, $\forall \boldsymbol{h}$. □

Exercise 14.6. Consider the symmetric 2×2-matrix

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

(i) Prove that A is positive definite if and only if $a > 0$ and $ac - b^2 > 0$.

(ii) Prove that A is negative definite if and only if $a < 0$ and $ac - b^2 > 0$.

(iii) Prove that A is indefinite if and only if $ac - b^2 < 0$.

Hint: (i) Investigate when $Q_A(x, 1) > 0$ for any $x \in \mathbb{R}$. (ii) Investigate when $Q_A(x, 1) < 0$ for any $x \in \mathbb{R}$. □

Exercise 14.7. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open and convex set. A function $f : U \to \mathbb{R}$ is called *convex* if

$$f(t\boldsymbol{p} + (1-t)\boldsymbol{q}) \leqslant tf(\boldsymbol{p}) + (1-t)f(\boldsymbol{q}), \quad \forall t \in [0, 1], \ \ \boldsymbol{p}, \boldsymbol{q} \in U.$$

(i) Prove that if the C^1 function $f : U \to \mathbb{R}$ is convex, then

$$f(\boldsymbol{q}) \geqslant f(\boldsymbol{p}) + \langle \nabla f(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p} \rangle, \quad \forall \boldsymbol{p}, \boldsymbol{q} \in U.$$

(ii) Prove that a C^1 function $f : U \to \mathbb{R}$ is convex if and only if

$$\langle \nabla f(\boldsymbol{p}) - \nabla f(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle \geqslant 0, \quad \forall \boldsymbol{p}, \boldsymbol{q} \in U.$$

(iii) Prove that a C^2 function $f : U \to \mathbb{R}$ is convex if and only if

$$\langle \boldsymbol{H}(f, \boldsymbol{p})\boldsymbol{v}, \boldsymbol{v} \rangle \geqslant 0, \quad \forall \boldsymbol{p} \in U, \quad \boldsymbol{v} \in \mathbb{R}^n,$$

where $\boldsymbol{H}(f, \boldsymbol{p})$ is the Hessian of f at $\boldsymbol{p}$; see Definition 14.1.2.

Hint: Have a look at Section 8.3. and observe that f is convex if and only if for any $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$ the function $g_{\boldsymbol{p},\boldsymbol{q}} : \mathbb{R} \to \mathbb{R}$, $g_{\boldsymbol{p},\boldsymbol{q}}(t) = f(1-t)\boldsymbol{p} + t\boldsymbol{q})$ is convex. □

Exercise 14.8. Consider the smooth function

$$f : (0, \infty) \times (0, \infty) \to \mathbb{R}, \quad f(x, y) = xy + \frac{1}{x} + \frac{1}{y}.$$

(i) Show that the point $\boldsymbol{p}_0 = (1, 1)$ is the only critical point of f.

(ii) Show that the point $\boldsymbol{p}_0 = (1, 1)$ is a local minimum of f.

(iii) Prove that

$$f(x, y) \geqslant 2\sqrt{x + y}, \quad \forall x, y > 0.$$

Hint: Use the inequality $a^2 + b^2 \geqslant 2ab$, $\forall a, b \in \mathbb{R}$.

(iv) Prove that the point $\boldsymbol{p}_0$ in (i) is *the global minimum point of* f, i.e.,

$$f(\boldsymbol{p}_0) < f(\boldsymbol{p}), \quad \forall \boldsymbol{p} \in (0, \infty) \times (0, \infty), \quad \boldsymbol{p} \neq \boldsymbol{p}_0.$$

Hint: Set $\mu := \inf_{x,y>0} f(x, y) \geqslant 0$. Choose a sequence (x_n, y_n) such that

$$x_n, y_n > 0, \quad \mu \leqslant f(x_n, y_n) < \mu + \frac{1}{n}, \quad \forall n \geqslant 1.$$

Prove that (x_n, y_n) is bounded. Conclude using Bolzano-Weierstrass.

□

Exercise 14.9. Let $n \in \mathbb{N}$ and suppose that $U, V \subset \mathbb{R}^n$ are open sets.

(i) Prove that if $\boldsymbol{G} : V \to \mathbb{R}^n$ is a C^1 diffeomorphism and $W \subset V$ is an open set, then $\boldsymbol{G}(W)$ is also an open set.

(ii) Prove that if $\boldsymbol{F} : U \to \mathbb{R}^n$ and $\boldsymbol{G} : V \to \mathbb{R}^n$ are C^1-diffeomorphisms and $\boldsymbol{F}(U) \subset V$, then the composition $\boldsymbol{G} \circ \boldsymbol{F} : U \to \mathbb{R}^n$ is also a C^1-diffeomorphism.

□

Exercise 14.10. Prove Corollary 14.3.7. □

Exercise 14.11. Prove that the following maps are C^1 diffeomorphisms,[7] and then find their ranges and inverses.

$$\begin{gathered}
\boldsymbol{F} : (0,\infty)\times(0,2\pi)\to\mathbb{R}^2,\ \ \boldsymbol{F}(r,\theta) = [r\cos\theta, r\sin\theta]^\top,\\
\boldsymbol{G} : (0,\infty)\times(0,2\pi)\times\mathbb{R}\to\mathbb{R}^3,\ \ \boldsymbol{G}(r,\theta,z) = [r\cos\theta, r\sin\theta, z]^\top,\\
\boldsymbol{H} : (0,\infty)\times(0,2\pi)\times(0,\pi)\to\mathbb{R}^3,\ \ \boldsymbol{H}(\rho,\theta,\varphi) = [\rho\cos\varphi\cos\theta, \rho\cos\varphi\sin\theta, \rho\cos\varphi]^\top.
\end{gathered}$$

Hint. Prove that each of the above maps and then show that Corollary 14.3.7 applies in each of these cases. □

Exercise 14.12. Let $m, n \in \mathbb{N}$, $m < n$. Prove that if S_1, S_2 are two codimension m coordinate subspaces of $\mathbb{R}^n$, then there exists a bijective linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T(S_1) = S_2$. □

Exercise 14.13. Consider the map $\Phi : (-\pi,\pi)\times(-\pi,\pi)\to\mathbb{R}^3$ given by

$$\Phi(\theta,\varphi) = \begin{bmatrix} (2+\cos\varphi)\cos\theta \\ (2+\cos\varphi)\sin\theta \\ \sin\varphi \end{bmatrix}.$$

Prove that Φ is an injective immersion. □

Exercise 14.14. Prove Proposition 14.3.4. □

Exercise 14.15. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = e^{\sin(xy)} - 1$.

(i) Show that $f(1,0) = 0$.

(ii) Show that there exist open intervals I centered at 1 and J centered at 0 and a C^1-function $g : I \to \mathbb{R}$ such that the intersection of the level set $\{f = 0\}$ with the rectangle $I \times J \subset \mathbb{R}^2$ coincides with the graph of g.

(iii) Compute $g'(1)$.

□

Exercise 14.16. Show that the equation

$$xy - z\log y + e^{xz} = 1$$

be solved uniquely in the form $y = g(x,z)$ in an open neighborhood of $(0,1,1)$. □

Exercise 14.17. Show that the system of equations

$$\begin{cases} u^2+v^2-x^2-y &= 0\\ u+v-x^2+y &= 0, \end{cases}$$

can be solved uniquely for (u,v) in terms of (x,y) in an open neighborhood of

$$(u_0,v_0,x_0,y_0) = (1,2,2,1).$$ □

[7] Exercise 13.3 asks you to compute the Jacobian matrices of these maps.

Exercise 14.18. Consider the map

$$\boldsymbol{F} : (0, 2\pi) \times (0, \pi/2) \to \mathbb{R}^3, \quad \boldsymbol{F}(\theta, \varphi) = \begin{bmatrix} \sin\varphi\cos\theta \\ \sin\varphi\sin\theta \\ \cos\varphi \end{bmatrix}.$$

Show $\boldsymbol{F}$ is an injective immersion and then find its image. □

Exercise 14.19. (a) Suppose that $f : \mathbb{R} \to (0, \infty)$ is a C^1-function. Its graph is the curve

$$\Gamma_f := \{ (x, y) \in \mathbb{R}^2; \;\; y = f(x) \}.$$

Denote by S_f the region in the space $\mathbb{R}^3$ swept when rotating Γ_f about the x axis; see Figure 14.13. Show that S_f is 2-dimensional submanifold of $\mathbb{R}^3$.

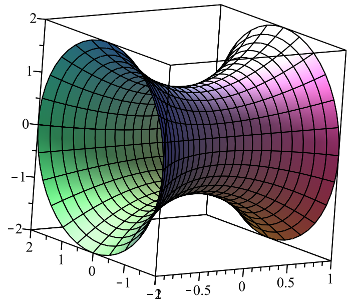

Figure 14.13. *The surface of revolution S_f, $f(x) = x^2 + 1$.*

(b) Suppose that $0 < a < b$ and $h : (a, b) \to \mathbb{R}$ is a C^1-function. Denote by Σ_h the region in the space $\mathbb{R}^3$ swept when rotating Γ_h about the y-axis; see Figure 14.14 in the special case $h(x) = (x - 1)(x - 2)$. Show that Σ_h is a 2-dimensional submanifold of $\mathbb{R}^3$.

Hint. (a) Show that S_f is described by the equation $f(x)^2 = y^2 + z^2$ and then show that this equation satisfies the assumptions of Proposition 14.5.9. (b) Show that Σ_h is the graph of a function depending on x, z i.e., it can be described by an equation of the form $y = F(x, z)$ for some C^1-function F. □

Exercise 14.20. Prove that the map $\boldsymbol{F} : (0, \infty) \times \mathbb{R} \to \mathbb{R}^3$ given by

$$F(r, t) = \begin{bmatrix} r\cos t \\ r\sin t \\ t \end{bmatrix}$$

satisfies all the conditions (i)-(iii) in Proposition 14.5.4. (Its image is a *helicoid* and it is depicted Figure 14.15.) □

Exercise 14.21. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = xy - z\log y + e^{xz} - 1$.

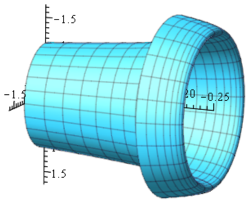

Figure 14.14. *The surface of revolution* Σ_h, $h(x) = (x-1)(x-2)$, $x \in (1, 1.75)$.

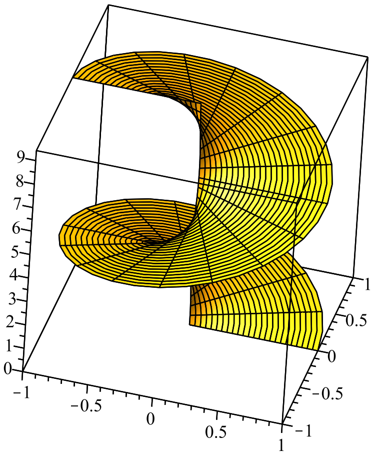

Figure 14.15. *A helicoid.*

(i) Show that there exists an open neighborhood U of $\boldsymbol{p}_0 := (0,1,1)$ such that $\nabla f(\boldsymbol{p}) \neq \boldsymbol{0}\ \forall \boldsymbol{p} \in U$.

(ii) Let U be as above. Show that the set

$$Z = \left\{ \boldsymbol{p} \in U;\ \ f(\boldsymbol{p}) = 0 \right\}$$

is a 2-dimensional submanifold of $\mathbb{R}^3$ containing $\boldsymbol{p}_0$.

(iii) Find a basis of the tangent space $T_{\boldsymbol{p}_0} Z$.

□

Exercise 14.22. Consider the map $\boldsymbol{F} : \mathbb{R}^4 \to \mathbb{R}^2$ given by

$$\boldsymbol{F}(u,v,x,y) = \begin{bmatrix} u^2 + v^2 - x^2 - y \\ u + v - x^2 + y \end{bmatrix}.$$

Set $\boldsymbol{p}_0 := (1,2,2,1) \in \mathbb{R}^4$.

(i) Show that there exists an open neighborhood U of $\boldsymbol{p}_0$ in $\mathbb{R}^4$ such that, $\forall \boldsymbol{p} \in U$, the differential $d\boldsymbol{F}(\boldsymbol{p})$ is surjective as a linear map $\mathbb{R}^4 \to \mathbb{R}^2$, i.e., the Jacobian $J_{\boldsymbol{F}}(\boldsymbol{p})$ has rank 2 for any $\boldsymbol{p} \in U$.

(ii) Let U be as above. Show that the set

$$Z = \left\{ \boldsymbol{p} \in U;\ \ \boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{0} \right\}$$

is a 2-dimensional submanifold of $\mathbb{R}^4$ containing $\boldsymbol{p}_0$.

(iii) Find a basis of the tangent space $T_{\boldsymbol{p}_0} Z$.

Hint. (i) Use the main theorem in Sec.6, Chapter 3 of [40]. □

Exercise 14.23. Show that the set

$$S := \left\{ (x,y,z) \in \mathbb{R}^3;\ \ x^2 + y^2 - z^2 = 1 \right\}$$

is a 2-dimensional submanifold of $\mathbb{R}^3$ and then describe a basis of the tangent space to S at the point $\boldsymbol{p}_0 = (1,1,1)$. □

Exercise 14.24. Let $n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ an open set and $S \subset U$ a C^1-submanifold of dimension k. Prove that if $\boldsymbol{F} : U \to \mathbb{R}^n$ is a C^1-diffeomorphism, then $\boldsymbol{F}(S)$ is also C^1-submanifold of $\mathbb{R}^n$ of dimension k.

Hint. Observe that $\boldsymbol{F}^{-1} : \boldsymbol{F}(U) \to \mathbb{R}^n$ is a diffeomorphism. Use Exercise 14.9 to show that if (V, Ψ) is a straightening diffeomorphism of S at $\boldsymbol{p}_0$, then $(\boldsymbol{F}(V), \Psi \circ \boldsymbol{F}^{-1})$ is a straightening diffeomorphism of $\boldsymbol{F}(S)$ at $\boldsymbol{F}(\boldsymbol{p}_0)$. □

Exercise 14.25. Let $n \in \mathbb{N}$.

(i) Prove that any vector subspace $U \subset \mathbb{R}^n$ is a submanifold of dimension $\dim U$.

(ii) Suppose that $X \subset \mathbb{R}^n$ is a nonempty affine subspace and $\boldsymbol{p}_0 \in X$. Define $T : \mathbb{R}^n \to \mathbb{R}^n$, $T(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{p}_0$. Show that $U = T(X)$ is a vector subspace of $\mathbb{R}^n$.

(iii) Prove that the above map T is a diffeomorphism with image $\mathbb{R}^n$.

(iv) Deduce that X is a submanifold of $\mathbb{R}^n$ of dimension $\dim U$.

Hint. (i) requires linear algebra, namely that a basis of U can be extended to a basis of $\mathbb{R}^n$. (ii),(iii) are easy. For (iv) use (i)-(iii) and Exercise 14.24. □

Exercise 14.26. Let $n \in \mathbb{N}$, $n \geqslant 2$.

(i) Show that a hyperplane H of $\mathbb{R}^n$ is a submanifold of dimension $n-1$.

(ii) Let $f : \mathbb{R}^n \to \mathbb{R}$ is a C^1-function and set

$$Z_f := \left\{ \boldsymbol{p} \in \mathbb{R}^n;\ \ f(\boldsymbol{p}) = 0 \right\}.$$

Suppose that $\nabla f(\boldsymbol{p}) \neq \boldsymbol{0}$, $\forall \boldsymbol{p} \in Z_f$. According to Example 14.5.22 the zero set Z_f is a *hypersurface* of $\mathbb{R}^n$, i.e., a submanifold of dimension $n-1$. Fix a point

$\boldsymbol{p}_0 \in Z_f$ and set $\boldsymbol{v}_0 := \nabla f(\boldsymbol{p}_0)$. Describe explicitly in terms of $\boldsymbol{p}_0$ and $\boldsymbol{v}_0$ the hyperplane of $\mathbb{R}^n$ satisfying

$$\boldsymbol{p}_0 \in H \;\text{ and }\; T_{\boldsymbol{p}_0} H = T_{\boldsymbol{p}_0} Z_f.$$

This hyperplane is called the *affine tangent space* of Z_f at $\boldsymbol{p}_0$.

□

Exercise 14.27. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set containing the origin $\boldsymbol{0}$. Consider a C^1-function $f : U \to \mathbb{R}$. We set

$$c_0 = f(\boldsymbol{0}), \;\; c_i := \partial_{x^i} f(\boldsymbol{0}), \;\; i = 1, \dots, n.$$

The graph of f,

$$\Gamma_f := \big\{ (\boldsymbol{x}, y) \in U \times \mathbb{R};\;\; y = f(\boldsymbol{x}) \,\big\} \subset \mathbb{R}^{n+1},$$

is an n-dimensional submanifold of $\mathbb{R}^{n+1}$. Note that the point $\boldsymbol{p}_0 := (\,\boldsymbol{0}, f(\boldsymbol{0})\,)$ belongs to the graph.

(i) Describe explicitly in terms of the constants $c_1, \dots, c_n$ a basis of the tangent space $T_{\boldsymbol{p}_0}\Gamma_f$.

(ii) Describe explicitly in terms of the constants $c_0, c_1, \dots, c_n$ a hyperplane $H \subset \mathbb{R}^{n+1}$ such that

$$\boldsymbol{p}_0 \in H \;\text{ and }\; T_{\boldsymbol{p}_0} H = T_{\boldsymbol{p}_0}\Gamma_f.$$

Hint. Observe that Γ_f can be described as the hypersurface of $\mathbb{R}^{n+1}$ described in the coordinates $(x^1, \dots, x^n, y)$ by the equation $f(x^1, \dots, x^n) - y = 0$. □

Exercise 14.28. Find the minimum of $h(x, y, z, t) = t$ subject to the constraints

$$x^2 + y^2 + z^2 + t^2 = x + y + z + t = 1.$$

Hint. Apply Corollary 14.5.26. You also need to use the results in Example 14.5.23. □

Exercise 14.29. From among all rectangular parallelepipeds of given volume $V > 0$ find the ones that have the least surface area. □

Exercise 14.30. Determine the outer dimensions of a plastic box, with no lid, with walls of a given thickness δ and a given internal volume V that requires the least amount of plastic to produce. □

Exercise 14.31 (Raleigh-Ritz)**.** Let $n \in \mathbb{N}$ and suppose that A is a symmetric $n \times n$ matrix. Define

$$q_A : \mathbb{R}^n \to \mathbb{R}, \;\; q_A(\boldsymbol{x}) = \langle A\boldsymbol{x}, \boldsymbol{x} \rangle, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Set

$$\mu := \inf_{\|\boldsymbol{x}\|=1} q_A(\boldsymbol{x}).$$

(i) Show that there exists $\boldsymbol{u} \in \mathbb{R}^n$ such that $\|\boldsymbol{u}\| = 1$ and $q_A(\boldsymbol{u}) = \mu$.

(ii) Use Lagrange multipliers to show that any vector $\boldsymbol{u}$ as above is an eigenvector of A corresponding to the eigenvalue μ, i.e., $A\boldsymbol{u} = \mu\boldsymbol{u}$.

Hint. You may want to have a look at Exercise 13.5.qed

14.7. Exercises for extra credit

Exercise* 14.1. Let $n \in \mathbb{N}$ and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex C^1-function. Show that the map

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n, \ \ \Phi(x) = \boldsymbol{x} + \nabla f(\boldsymbol{x})$$

is surjective. □

Chapter 15

Multidimensional Riemann integration

In this chapter we want to extend the concept of Riemann integral to functions of several variables. While there are many similarities, the higher dimensional situation displays new phenomena and difficulties that do not have a 1-dimensional counterpart.

15.1. Riemann integrable functions of several variables

15.1.1. The Riemann integral over a box. In this chapter, for simplicity, we define a *box* in $\mathbb{R}^n$ to be a *closed* box in the sense of Definition 12.2.8, i.e., a closed set B of the form

$$B = [a_1, b_1] \times \cdots \times [a_n, b_n], \tag{15.1.1}$$

where a's and b's are real numbers satisfying $a_1 \leqslant b_1, \dots, a_n \leqslant b_n$. Note that we allow the possibility that $a_i = b_i$ for some i's. In particular, a set consisting of a single point is a very special case of box.

A *vertex* of the box in (15.1.1) is a point $\boldsymbol{x} = (x^1, \dots, x^n)$ such that

$$x^i = a_i \ \text{ or } \ x^i = b_i, \ \ \forall i = 1, \dots, n.$$

A *facet* of B is the set obtained by intersecting B with a coordinate hyperplane of the form $x^i = a_i$ or $x^i = b_i$.[1]

The n*-dimensional volume* of the box B in (15.1.1) is the nonnegative real number

$$\operatorname{vol}(B) = \operatorname{vol}_n(B) := (b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n). \tag{15.1.2}$$

[1] In dimension 2 a box is a rectangle and the facets are the boundary edges of that rectangle.

The box B is called *nondegenerate* if $\operatorname{vol}_n(B) > 0$, i.e., $a_i < b_i$, $\forall i = 1, \dots, n$. Note that when $n = 1$, a box B in $\mathbb{R}$ is a compact interval and $\operatorname{vol}_1(B)$ is precisely the length of the interval B. In this case the facets of B are the endpoints of the interval B

Recall that the diameter of a set $S \subset \mathbb{R}^n$ is (see Definition 12.3.8)

$$\operatorname{diam}(S) = \sup_{\boldsymbol{x}, \boldsymbol{y} \in S} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}).$$

It is not hard to see that if B is the box in (15.1.1), then

$$\operatorname{diam}(B) = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

Note that the intersection of two boxes B, B' is either empty, or another box. For example if

$$B = I_1 \times \cdots \times I_n, \;\; B' = I_1' \times \cdots \times I_n',$$

the $I_j, I_k' \subset \mathbb{R}$ are compact intervals, then $B \cap B'$ is the (possibly empty) box

$$(I_1 \cap I_1') \times \cdots \times (I_n \cap I_n').$$

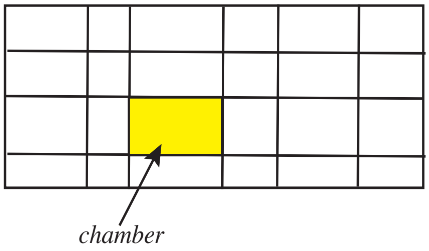

Figure 15.1. *A partition of a 2-dimensional box. The sum of the areas of the chambers is equal to the area of the big box that contains them.*

Definition 15.1.1. Let $n \in \mathbb{N}$ and suppose that

$$B = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$$

is a nondegenerate box.

(a) A *partition* of B is an n-tuple $\boldsymbol{P} = (\boldsymbol{P}_1, \dots, \boldsymbol{P}_n)$, where, for each $j = 1, \dots, n$, $\boldsymbol{P}_j$, is a partition of the interval $[a_j, b_j]$; see Definition 9.2.1.

(b) A *chamber* of $\boldsymbol{P}$ is a box of the form $I_1 \times \cdots \times I_n$, where $I_j \subset [a_j, b_j]$ is an interval of the partition $\boldsymbol{P}_j$. We denote by $\mathscr{C}(\boldsymbol{P})$ the set of chambers of a partition $\boldsymbol{P}$ and by $\mathcal{P}(B)$ the set of partitions of B.

(c) The *mesh size* or *mesh* of a partition $\boldsymbol{P}$ is the positive number

$$\|\boldsymbol{P}\| := \max_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{diam}(C).$$

(d) A partition $\boldsymbol{P}'$ of B is said to be *finer* than another partition $\boldsymbol{P}$, and we denote this by $\boldsymbol{P}' > \boldsymbol{P}$, if any chamber of $\boldsymbol{P}'$ is contained in some chamber of $\boldsymbol{P}$. □

The next result is immediate in dimension 1 and, although it is very intuitive, it takes a bit more work in higher dimensions. We leave its proof to you as an exercise.

Lemma 15.1.2. *Let $n \in \mathbb{N}$. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $\boldsymbol{P}$ is a partition of B. Then (see Figure 15.1)*

$$\operatorname{vol}_n(B) = \sum_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{vol}_n(C). \tag{15.1.3}$$

□

✍ In the sequel, for the sake of readability, we introduce the following notation:

$$m_S(f) := \inf_{x \in S} f(x), \quad M_S(f) := \sup_{x \in S} f(x),$$

for any function $f : X \to \mathbb{R}$, and any $S \subset X$.

Definition 15.1.3. Let $n \in \mathbb{N}$. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a *bounded* function.

(a) For any partition $\boldsymbol{P}$ of B we define the *lower Darboux sum* of f over $\boldsymbol{P}$ to be

$$\boldsymbol{S}_*(f, \boldsymbol{P}) := \sum_{C \in \mathscr{C}(\boldsymbol{P})} m_C(f) \operatorname{vol}_n(C).$$

The *upper Darboux sum* of f over $\boldsymbol{P}$ is

$$\boldsymbol{S}^*(f, \boldsymbol{P}) := \sum_{C \in \mathscr{C}(\boldsymbol{P})} M_C(f) \operatorname{vol}_n(C).$$

(b) The *mean oscillation* of f over a partition $\boldsymbol{P}$ of B is the real number

$$\omega(f, \boldsymbol{P}) := \sum_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{osc}(f, C) \operatorname{vol}_n(C).$$

□

Arguing as in the one-dimensional case (see Proposition 9.3.2) one can show that for any nondegenerate box $B \subset \mathbb{R}^n$, any bounded function $f : B \to \mathbb{R}$ and any partition $\boldsymbol{P}$ of B we have

$$m_B(f) \operatorname{vol}_n(B) \leqslant \boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}) \leqslant M_B(f) \operatorname{vol}_n(B), \tag{15.1.4a}$$

$$\omega(f, \boldsymbol{P}) = \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}). \tag{15.1.4b}$$

Indeed

$$m_B(f) \operatorname{vol}_n(B) \overset{(15.1.3)}{=} \sum_{C \in \mathscr{C}(\boldsymbol{P})} m_B(f) \operatorname{vol}_n(C)$$

$(m_B(f) \leqslant m_C(f),\ \forall C \in \mathscr{C}(\boldsymbol{P}))$

$$\leqslant \sum_{C \in \mathscr{C}(\boldsymbol{P})} m_C(f) \operatorname{vol}_n(C) \leqslant \sum_{C \in \mathscr{C}(\boldsymbol{P})} M_C(f) \operatorname{vol}_n(C)$$

$(M_C(f) \leqslant M_B(f),\ \forall C \in \mathscr{C}(\boldsymbol{P}))$

$$\leqslant \sum_{C\in\mathscr{C}(\boldsymbol{P})} M_B(f)\operatorname{vol}_n(C) \overset{(15.1.3)}{=} M_B(f)\operatorname{vol}_n(B).$$

The next result is the higher dimensional counterpart of Proposition 9.3.6.

Lemma 15.1.4. *Let $n \in \mathbb{N}$. Assume that $B \subset \mathbb{R}^n$ is a nondegenerate box and $\boldsymbol{P}$,$\boldsymbol{P}'$ are partitions of B such that $\boldsymbol{P}'$ is finer than $\boldsymbol{P}$, $\boldsymbol{P}' \succ \boldsymbol{P}$. Then, for any bounded function $f : B \to \mathbb{R}$ we have*

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}), \tag{15.1.5a}$$

$$\omega(f,\boldsymbol{P}') \leqslant \omega(f,\boldsymbol{P}). \tag{15.1.5b}$$

Proof. We already know that $\boldsymbol{S}_*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}')$ so it suffices to prove

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}') \ \text{ and } \ \boldsymbol{S}^*(f,\boldsymbol{P}') \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}).$$

We will prove only the first one. The proof of the second inequality above is entirely similar, and it follows from the first inequality applied to the function $-f$. Suppose that $\boldsymbol{P} = (B_\alpha)_{\alpha\in\mathcal{A}}$.

For every chamber $C \in \mathscr{C}(\boldsymbol{P})$ we denote by $\boldsymbol{P}'_C$ the collection of chambers of the partition $\boldsymbol{P}'$ that are contained in C. Note that the collection $\boldsymbol{P}'_C$ is the collection of chambers of some partition of C. We have

$$\boldsymbol{S}^*(f,\boldsymbol{P}') = \sum_{C'\in\mathscr{C}(\boldsymbol{P}')} M_{C'}(f)\operatorname{vol}_n(C') = \sum_{C\in\mathscr{C}(\boldsymbol{P})}\left(\sum_{C'\in\boldsymbol{P}'_C} M_{C'}(f)\operatorname{vol}_n(C')\right)$$

$(M_{C'}(f) \leqslant M_C(f)$ when $C' \subset C)$

$$\leqslant \sum_{C\in\mathscr{C}(\boldsymbol{P})}\left(\sum_{C'\in\boldsymbol{P}'_C} M_C(f)\operatorname{vol}_n(C')\right) = \sum_{C\in\mathscr{C}(\boldsymbol{P})} M_C(f)\boxed{\left(\sum_{C'\in\boldsymbol{P}'_C}\operatorname{vol}_n(C')\right)}$$

$$\overset{(15.1.3)}{=} \sum_{C\in\mathscr{C}(\boldsymbol{P})} M_C(f)\boxed{\operatorname{vol}_n(C)} = \boldsymbol{S}^*(f,\boldsymbol{P}).$$

This proves (15.1.5a). To prove (15.1.5b) note that

$$\omega(f,\boldsymbol{P}') = \boldsymbol{S}^*(f,\boldsymbol{P}') - \boldsymbol{S}_*(f,\boldsymbol{P}') \overset{(15.1.5a)}{\leqslant} \boldsymbol{S}^*(f,\boldsymbol{P}) - \boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \omega(f,\boldsymbol{P}).$$

□

Consider a nondegenerate box

$$B = [a_1,b_1]\times\cdots\times[a_n,b_n] \subset \mathbb{R}^n$$

and two partitions $\boldsymbol{P}', \boldsymbol{P}''$ of it. The intersection of a chamber $C' \in \mathscr{C}(\boldsymbol{P}')$ with a chamber $C'' \in \mathscr{C}(\boldsymbol{P}'')$ is either empty, a degenerate box or a nondegenerate box. The collection of all the possible nondegenerate boxes formed by such overlaps coincides with the collection of chambers of a new partition of B that we denote by $\boldsymbol{P}' \vee \boldsymbol{P}''$. Equivalently if $\boldsymbol{P}'$ and $\boldsymbol{P}''$ are described by n-tuples of partitions of the intervals $[a_j,b_j]$,

$$\boldsymbol{P}' = (\boldsymbol{P}'_1,\dots,\boldsymbol{P}'_n),\ \ \boldsymbol{P}'' = (\boldsymbol{P}''_1,\dots,\boldsymbol{P}''_n),$$

then

$$\boldsymbol{P}' \vee \boldsymbol{P}'' = (\boldsymbol{P}'_1 \vee \boldsymbol{P}''_1, \dots, \boldsymbol{P}'_n \vee \boldsymbol{P}''_n),$$

where $\boldsymbol{P}'_j \vee \boldsymbol{P}''_j$ is defined at page 256.

By construction, any chamber of $\boldsymbol{P}' \vee \boldsymbol{P}''$ is contained both in a chamber of the partition $\boldsymbol{P}'$ and in a chamber of $\boldsymbol{P}''$. In other words,

$$\boldsymbol{P}' \vee \boldsymbol{P}'' \succ \boldsymbol{P}', \;\; \boldsymbol{P}''.$$

Lemma 15.1.4 implies that, for any bounded function $f : B \to \mathbb{R}$ we have

$$\boldsymbol{S}_*(f, \boldsymbol{P}') \leqslant \boldsymbol{S}_*(f, \boldsymbol{P}' \vee \boldsymbol{P}'') \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}' \vee \boldsymbol{P}'') \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}'').$$

We have thus shown that

$$\boldsymbol{S}_*(f, \boldsymbol{P}') \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}''), \;\; \forall \boldsymbol{P}', \boldsymbol{P}'' \in \mathcal{P}(B).$$

Hence, the collection of lower Darboux sums of f is bounded above by any upper Darboux sum, and the collection of upper Darboux sums of f is bounded below by any lower Darboux sum. We set

$$\underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| := \sup_{\boldsymbol{P} \in \mathcal{P}(B)} \boldsymbol{S}_*(f, \boldsymbol{P}), \;\; \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| := \inf_{\boldsymbol{P} \in \mathcal{P}(B)} \boldsymbol{S}^*(f, \boldsymbol{P})$$

The number $\underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}|$ is called the *lower Darboux integral* of f over B, and the number $\overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}|$ is called the *upper Darboux integral* of f over B. Note that

$$\underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

Definition 15.1.5. Let $n \in \mathbb{N}$. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a bounded function. The function f is called *Riemann integrable* over B if

$$\underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| = \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

The common value of these numbers is called the *Riemann integral* of f over B and it is denoted by

$$\boxed{\int_B f(\boldsymbol{x})|d\boldsymbol{x}| \;\text{ or }\; \int_B f(x^1, \dots, x^n)|dx^1 \cdots dx^n|.}$$

We denote by $\mathcal{R}(B)$ the set of Riemann integrable functions $f : B \to \mathbb{R}$. □

Remark 15.1.6. When $n = 1$, and $B = [a, b]$, $a < b$, then

$$\int_{[a,b]} f(x)|dx| = \int_a^b f(x)dx = -\int_b^a f(x)dx,$$

where $\int_a^b f(x)dx$ is the usual 1-dimensional Riemann integral. For this reason we will set

$$\int_a^b f(x)|dx| = \int_b^a f(x)|dx| := \int_{[a,b]} f(\boldsymbol{x})|d\boldsymbol{x}|.$$ □

Example 15.1.7. Suppose $f : B \to \mathbb{R}$ is a constant function, $f(\boldsymbol{x}) = c_0$, $\forall \boldsymbol{x} \in B$. Then, for any partition $\boldsymbol{P}$ of B, we have

$$\boldsymbol{S}_*(f, \boldsymbol{P}) = \sum_{C \in \mathscr{C}(\boldsymbol{P})} c_0 \operatorname{vol}_n(C) = c_0 \sum_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{vol}_n(C) \overset{(15.1.3)}{=} c_0 \operatorname{vol}_n(B)$$

and, similarly,

$$\boldsymbol{S}^*(f, \boldsymbol{P}) = c_0 \operatorname{vol}_n(B).$$

This proves that f is integrable and

$$\int_B c_0 |d\boldsymbol{x}| = c_0 \operatorname{vol}_n(B). \qquad \square$$

Definition 15.1.8. Let $n \in \mathbb{N}$ and suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a *bounded* function. We define a *sample* of a partition $\boldsymbol{P}$ to be an assignment $\underline{\xi}$ that associates to each chamber C of $\boldsymbol{P}$ a point ξ_C located in the chamber C. The *Riemann sum of* f determined by the partition $\boldsymbol{P}$ and the sample $\underline{\xi}$ is the real number

$$\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) := \sum_{C \in \mathscr{C}(\boldsymbol{P})} f(\xi_C) \operatorname{vol}_n(C). \qquad \square$$

From the definition of Riemann and Darboux sums we deduce immediately that, for any partition $\boldsymbol{P}$ of B and any sample $\underline{\xi}$ of $\boldsymbol{P}$ we have

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}).$$

Note also, that if $f, g : B \to \mathbb{R}$ are two bounded functions, $a, b \in \mathbb{R}$, $\boldsymbol{P}$ is a partition of B and $\underline{\xi}$ is a sample of $\boldsymbol{P}$, then

$$\boldsymbol{S}\big(af + bg, \boldsymbol{P}, \underline{\xi}\big) = a\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) + b\boldsymbol{S}(g, \boldsymbol{P}, \underline{\xi}). \tag{15.1.6}$$

Our next result suggests a method of approximation of Riemann integrals.

Proposition 15.1.9. *Let $n \in \mathbb{N}$. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a Riemann integrable function. Then, for any partition $\boldsymbol{P}$ of B and any sample $\underline{\xi}$ of $\boldsymbol{P}$, we have*

$$\left| \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) - \int_B f(\boldsymbol{x}) |d\boldsymbol{x}| \right| \leqslant \omega(f, \boldsymbol{P}), \tag{15.1.7a}$$

$$\left| \boldsymbol{S}_*(f, \boldsymbol{P}) - \int_B f(\boldsymbol{x})\, |d\boldsymbol{x}| \right|, \ \left| \boldsymbol{S}^*(f, \boldsymbol{P}) - \int_B f(\boldsymbol{x})\, |d\boldsymbol{x}| \right| \leqslant \omega(f, \boldsymbol{P}). \tag{15.1.7b}$$

Proof. We have

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \int_B f(\boldsymbol{x})\, |d\boldsymbol{x}| \leqslant \boldsymbol{S}^*(f, \boldsymbol{P})$$

and

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}).$$

The conclusion follows by observing that the two numbers

$$\boldsymbol{S}(f, \boldsymbol{P}, \underline{\xi}),\ \ \int_B f(\boldsymbol{x})|d\boldsymbol{x}|$$

are both situated in the interval $\big[\, \boldsymbol{S}_*(f, \boldsymbol{P}), \boldsymbol{S}^*(f, \boldsymbol{P})\, \big]$ of length $\omega(f, \boldsymbol{P})$. □

Our next result is a higher dimensional version of the Riemann-Darboux Theorem 9.3.11.

Theorem 15.1.10 (Riemann-Darboux)**.** *Let $n \in \mathbb{N}$. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a bounded function. Then the following statements are equivalent.*

(i) *The function f is Riemann-integrable over B.*

(ii) *For any $\varepsilon > 0$ there exists a partition $\boldsymbol{P}$ of B such that the mean oscillation of f over $\boldsymbol{P}$ is $< \varepsilon$, i.e.,*

$$\omega(f, \boldsymbol{P}) < \varepsilon.$$

Proof. (i) ⇒ (ii). We know that f is Riemann integrable. We set

$$I := \int_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

We have

$$\underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| = \sup_{\boldsymbol{P} \in \mathcal{P}(B)} \boldsymbol{S}_*(f, \boldsymbol{P}) = \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| = \inf_{\boldsymbol{P} \in \mathcal{P}(B)} \boldsymbol{S}^*(f, \boldsymbol{P}) = I$$

Thus, for any $\varepsilon > 0$ there exists partitions $\boldsymbol{P}', \boldsymbol{P}''$ of B such that

$$I - \frac{\varepsilon}{2} < \boldsymbol{S}_*(f, \boldsymbol{P}') \leqslant I \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}'') < I + \frac{\varepsilon}{2}.$$

If we set $\boldsymbol{P} := \boldsymbol{P}' \vee \boldsymbol{P}''$, then we deduce

$$I - \frac{\varepsilon}{2} < \boldsymbol{S}_*(f, \boldsymbol{P}') \leqslant \boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}'') < I + \frac{\varepsilon}{2}.$$

Hence

$$\omega(f, \boldsymbol{P}) = \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}) < I + \frac{\varepsilon}{2} - \Big(I - \frac{\varepsilon}{2}\Big) = \varepsilon.$$

(ii) ⇒ (i) Let $\varepsilon > 0$. There exists a partition $\boldsymbol{P}$ of B such that

$$\omega(f, \boldsymbol{P}) = \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}) < \varepsilon.$$

On the other hand,

$$\boldsymbol{S}_*(f, \boldsymbol{P}) \leqslant \underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \boldsymbol{S}^*(f, \boldsymbol{P})$$

so that

$$\overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| - \underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}) - \boldsymbol{S}_*(f, \boldsymbol{P}) < \varepsilon.$$

In other words

$$0 \leqslant \overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| - \underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \varepsilon, \;\; \forall \varepsilon > 0,$$

i.e.,

$$\overline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| - \underline{\int}_B f(\boldsymbol{x})|d\boldsymbol{x}| = 0$$

and thus the function f is Riemann integrable. □

Corollary 15.1.11. *Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a continuous function. Then f is Riemann integrable over B.*

Proof. The box B is compact and thus f is uniformly continuous. Thus, for any $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that, for any set $S \subset B$ satisfying $\operatorname{diam}(S) < \delta(\varepsilon)$ we have

$$\operatorname{osc}(f, S) < \frac{\varepsilon}{\operatorname{vol}_n(B)}.$$

Choose a partition $\boldsymbol{P}$ of B such that $\|\boldsymbol{P}\| < \delta(\varepsilon)$. In particular, we deduce that

$$\operatorname{osc}(f, C) < \frac{\varepsilon}{\operatorname{vol}_n(B)}, \;\; \forall C \in \mathscr{C}(\boldsymbol{P}).$$

We have

$$\omega(f, \boldsymbol{P}) = \sum_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{osc}(f, C) \operatorname{vol}_n(C) < \frac{\varepsilon}{\operatorname{vol}_n(B)} \sum_{C \in \mathscr{C}(\boldsymbol{P})} \operatorname{vol}_n(C) = \varepsilon.$$

□

Theorem 15.1.12. *Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f_1, \dots, f_N : B \to \mathbb{R}$ are Riemann integrable functions. Fix a positive constant R such that*

$$|f_i(\boldsymbol{x})| \leqslant R, \;\; \forall i = 1, \dots, N, \;\; \forall \boldsymbol{x} \in B.$$

If

$$H : \underbrace{[-R, R] \times \cdots \times [-R, R]}_{N} \to \mathbb{R}$$

is a Lipschitz function, then the function

$$f : B \to \mathbb{R}, \;\; f(\boldsymbol{x}) = H\big(f_1(\boldsymbol{x}), \dots, f_N(\boldsymbol{x})\big)$$

is also Riemann integrable.

Proof. Fix a Lipschitz constant $L > 0$ of H, i.e.,

$$|H(\boldsymbol{y}_1) - H(\boldsymbol{y}_2)| \leqslant L\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|, \;\; \forall \boldsymbol{y} \in \underbrace{[-R, R] \times \cdots [-R, R]}_{N} = \overline{C_R(\mathbf{0})} \subset \mathbb{R}^N.$$

Define $\boldsymbol{F}:\mathbb{R}^n\to\overline{C_R(\boldsymbol{0})}$

$$\boldsymbol{F}(\boldsymbol{x})=\begin{bmatrix} f_1(\boldsymbol{x})\\ \vdots\\ f_N(\boldsymbol{x})\end{bmatrix}.$$

We first prove that, for any subset $S\subset B$, we have

$$\operatorname{osc}(f,S)\leqslant L\sqrt{N}\sum_{i=1}^N\operatorname{osc}(f_i,S). \tag{15.1.8}$$

Indeed, for any $\boldsymbol{x}_1,\boldsymbol{x}_2\in S$, we have

$$\big|\,f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\,\big|=\big|\,H(\,\boldsymbol{F}(\boldsymbol{x}_1)\,)-H(\,\boldsymbol{F}(\boldsymbol{x}_2)\,)\,\big|$$

$$\leqslant L\big\|\,\boldsymbol{F}(\boldsymbol{x}_1)-\boldsymbol{F}(\boldsymbol{x}_2)\,\big\|\leqslant L\sqrt{N}\big\|\,\boldsymbol{F}(\boldsymbol{x}_1)-\boldsymbol{F}(\boldsymbol{x}_2)\,\big\|_\infty$$

$$=L\sqrt{N}\max_{1\leqslant i\leqslant N}\big|\,f_i(\boldsymbol{x}_1)-f_i(\boldsymbol{x}_2)\,\big|\leqslant L\sqrt{N}\sum_{i=1}^N\operatorname{osc}(f_i,S).$$

Since the functions f_i are Riemann integrable, we deduce that, for any $i=1,\dots,N$, and for any $\varepsilon>0$, we can find a partition $\boldsymbol{P}_i$ of B such that

$$\omega(f_i,\boldsymbol{P}_i)<\frac{\varepsilon}{LN\sqrt{N}}. \tag{15.1.9}$$

Choose a partition $\boldsymbol{P}$ that is finer than all the partitions $\boldsymbol{P}_1,\dots,\boldsymbol{P}_N$. E.g., we can choose

$$\boldsymbol{P}=\boldsymbol{P}_1\vee\boldsymbol{P}_2\vee\cdots\vee\boldsymbol{P}_N.$$

We deduce from (15.1.5b) that

$$\omega(f_i,\boldsymbol{P})<\frac{\varepsilon}{LN\sqrt{N}},\quad\forall i=1,\dots,N.$$

We have

$$\omega(f,\boldsymbol{P})=\sum_{C\in\mathscr{C}(\boldsymbol{P})}\operatorname{osc}(f,C)\operatorname{vol}_n(C)\overset{(15.1.8)}{\leqslant}L\sqrt{N}\sum_{C\in\mathscr{C}(\boldsymbol{P})}\sum_{i=1}^N\operatorname{osc}(f_i,C)\operatorname{vol}_n(C)$$

$$=L\sqrt{N}\sum_{i=1}^N\left(\sum_{C\in\mathscr{C}(\boldsymbol{P})}\operatorname{osc}(f_i,C)\operatorname{vol}_n(C)\right)=L\sqrt{N}\sum_{i=1}^N\omega(f_i,\boldsymbol{P})$$

$$\overset{(15.1.9)}{<}L\sqrt{N}\sum_{i=1}^N\frac{\varepsilon}{LN\sqrt{N}}=\varepsilon.$$

This proves that $f(\boldsymbol{x})$ is Riemann integrable. □

Theorem 15.1.13. *Let $n\in\mathbb{N}$ and suppose that $B\subset\mathbb{R}^n$ is a nondegenerate box. Then the following hold.*

(i) *If* $f,g\in\mathcal{R}(B)$ *and* $s,t\in\mathbb{R}$, *then* $sf+tg\in\mathcal{R}(B)$ *and*

$$\int_B\big(sf(\boldsymbol{x})+tg(\boldsymbol{x})\big)|d\boldsymbol{x}|=s\int_B f(\boldsymbol{x})|d\boldsymbol{x}|+t\int_B g(\boldsymbol{x})\,|d\boldsymbol{x}|. \tag{15.1.10}$$

(ii) *If* $f,g\in\mathcal{R}(B)$, *then* $fg\in\mathcal{R}(B)$.

(iii) *If* $f,g\in\mathcal{R}(B)$ *and* $f(\boldsymbol{x})\leqslant g(\boldsymbol{x})$, $\forall\boldsymbol{x}\in B$, *then*

$$\int_B f(\boldsymbol{x})|d\boldsymbol{x}|\leqslant\int_B g(\boldsymbol{x})|d\boldsymbol{x}|.$$

(iv) *If* $f\in\mathcal{R}(B)$, *then* $|f|\in\mathcal{R}(B)$ *and*

$$\left|\int_B f(\boldsymbol{x})|d\boldsymbol{x}|\right|\leqslant\int_B|f(\boldsymbol{x})|\,|d\boldsymbol{x}|.$$

Proof. (i) Let $H:\mathbb{R}^2\to\mathbb{R}$ be the linear function $H(x,y)=sx+ty$. Then H is Lipschitz and

$$sf(\boldsymbol{x})+tg(\boldsymbol{x})=H\big(f(\boldsymbol{x}),g(\boldsymbol{x})\big),\ \ \forall\boldsymbol{x}\in B.$$

Theorem 15.1.12 now implies that $sf(\boldsymbol{x})+tg(\boldsymbol{x})$ is Riemann integrable.

Arguing as in the proof of Theorem 15.1.12, we can find a sequence of partitions $\boldsymbol{P}_\nu$, $\nu\in\mathbb{N}$ of B such that

$$\omega(sf+tg,\boldsymbol{P}_\nu),\ \ \omega(f,\boldsymbol{P}_\nu),\ \ \omega(g,\boldsymbol{P}_\nu)<\frac{1}{\nu},\ \ \forall\nu\in\mathbb{N}.$$

Next, choose a sample $\underline{\xi}_\nu$ of $\boldsymbol{P}_\nu$ for any $\nu\in\mathbb{N}$. Proposition 15.1.9 now implies that

$$\lim_{\nu\to\infty}\boldsymbol{S}(f,\boldsymbol{P}_\nu,\underline{\xi}_\nu)=\int_B f(\boldsymbol{x})|d\boldsymbol{x}|,$$

$$\lim_{\nu\to\infty}\boldsymbol{S}(g,\boldsymbol{P}_\nu,\underline{\xi}_\nu)=\int_B g(\boldsymbol{x})|d\boldsymbol{x}|,$$

$$\lim_{\nu\to\infty}\boldsymbol{S}(\,sf+tg,\boldsymbol{P}_\nu,\underline{\xi}_\nu\,)=\int_B\big(sf(\boldsymbol{x})+tg(\boldsymbol{x})\big)|d\boldsymbol{x}|$$

On the other hand

$$\boldsymbol{S}(\,sf+tg,\boldsymbol{P}_\nu,\underline{\xi}_\nu\,)=s\boldsymbol{S}(f,\boldsymbol{P}_\nu,\underline{\xi}_\nu)+t\boldsymbol{S}(g,\boldsymbol{P}_\nu,\underline{\xi}_\nu).$$

If we let $\nu\to\infty$ in the last equality we obtain (15.1.10).

(ii) We begin by proving that for any $u\in\mathcal{R}(B)$, its square u^2 is also Riemann integrable. To see this, fix $R>0$ such that $|u(\boldsymbol{x})|\leqslant R$, $\forall\boldsymbol{x}\in B$. The function $H:[-R,R]\to\mathbb{R}$, $H(t)=t^2$ is Lipschitz because, for any $s,t\in[-R,R]$, we have

$$|H(s)-H(t)|=|s^2-t^2|=|s+t|\cdot|s-t|\leqslant(|s|+|t|)\cdot|t-s|\leqslant 2R|s-t|.$$

Then $u(\boldsymbol{x})^2=H(\,u(\boldsymbol{x})\,)$ is Riemann integrable.

To deal with the general case, note that, according to (i) $f+g,f-g\in\mathcal{R}(B)$. We deduce that $(f+g)^2,(f-g)^2\in\mathcal{R}(B)$ and thus

$$fg=\frac{1}{4}\Big(\,(f+g)^2-(f-g)^2\,\Big)\in\mathcal{R}(B).$$

(iii) The function $g(\boldsymbol{x}) - f(\boldsymbol{x})$ is Riemann integrable and nonnegative. In particular, we deduce that, for any partition $\boldsymbol{P}$ of B we have

$$0 \leqslant \boldsymbol{S}_*(g-f,\boldsymbol{P}) \leqslant \int_B \big(g(\boldsymbol{x})-f(\boldsymbol{x})\big)|d\boldsymbol{x}| = \int_B g(\boldsymbol{x})|d\boldsymbol{x}| - \int_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

(iv) The function $H:\mathbb{R}\to\mathbb{R}$, $H(x)=|x|$ is Lipschitz and Theorem 15.1.12 implies that $|f|\in\mathcal{R}(B)$ for any $f\in\mathcal{R}(B)$. Observe next that

$$-|f(\boldsymbol{x})| \leqslant f(\boldsymbol{x}) \leqslant |f(\boldsymbol{x})|,\ \ \forall \boldsymbol{x}\in B.$$

Using (i) and (iii) we deduce

$$-\int_B |f(\boldsymbol{x})||d\boldsymbol{x}| \leqslant \int_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \int_B |f(\boldsymbol{x})||d\boldsymbol{x}| \Longleftrightarrow \left|\int_B f(\boldsymbol{x})|d\boldsymbol{x}|\right| \leqslant \int_B |f(\boldsymbol{x})||d\boldsymbol{x}|.$$

□

Proposition 15.1.14. *Fix $n\in\mathbb{N}$ and a nondegenerate box $B\subset\mathbb{R}^n$. If $f_\nu : B\to\mathbb{R}$, $\nu\in\mathbb{N}$, is a sequence of Riemann integrable functions that converges uniformly to the function $f:B\to\mathbb{R}$, then f is also Riemann integrable and*

$$\lim_{\nu\to\infty}\int_B f_\nu(\boldsymbol{x})|d\boldsymbol{x}| = \int_B f(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.1.11}$$

Proof. Let $\hbar>0$. Since f_ν converges uniformly to f, there exists $N=N(\hbar)$ such that

$$\forall \nu \geqslant N(\hbar),\ \ \forall \boldsymbol{x}\in B : f_\nu(\boldsymbol{x})-\hbar < f(\boldsymbol{x}) < f_\nu(\boldsymbol{x})+\hbar. \tag{15.1.12}$$

We deduce from the above inequality that for any box $C\subset B$ and $\nu\geqslant N(\hbar)$ we have

$$m_C(f_\nu)-\hbar \leqslant m_C(f) \leqslant M_C(f) \leqslant M_C(f_\nu)+\hbar.$$

Hence, for any partition $\boldsymbol{P}$ of B we have

$$\boldsymbol{S}_*(f_\nu,\boldsymbol{P}) - \hbar\operatorname{vol}_n(B) \leqslant \boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}^*(f_\nu,\boldsymbol{P}) + \hbar\operatorname{vol}_n(B). \tag{15.1.13}$$

In particular, for any partition $\boldsymbol{P}$ of B, we have

$$\omega(f,\boldsymbol{P}) \leqslant \omega(f_\nu,\boldsymbol{P}) + 2\hbar\operatorname{vol}_n(B),\ \ \forall\nu\geqslant N(\hbar). \tag{15.1.14}$$

Now let $\varepsilon>0$. Choose $\hbar=\hbar(\varepsilon)$ such that

$$2\hbar\operatorname{vol}_n(B) < \frac{\varepsilon}{2}.$$

Now fix a natural number $\nu > N(\hbar(\varepsilon))$, where $N(\hbar)$ is as in (15.1.12). Since f_ν is Riemann integrable we can find a partition $\boldsymbol{P}_\varepsilon$ of B such that

$$\omega(f_\nu,\boldsymbol{P}_\varepsilon) < \frac{\varepsilon}{2}.$$

We deduce from (15.1.14) that

$$\omega(f,\boldsymbol{P}_\varepsilon) < \varepsilon$$

proving that f is Riemann integrable. From the inequalities (15.1.13) we can now conclude that

$$\int_B f_\nu(\boldsymbol{x})|d\boldsymbol{x}| - \hbar\,\mathrm{vol}_n(B) \leqslant \int_B f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \int_B f_\nu(\boldsymbol{x})|d\boldsymbol{x}| + \hbar\,\mathrm{vol}_n(B), \quad \forall \nu \geqslant N(\hbar)$$

i.e.,

$$\left| \int_B f_\nu(\boldsymbol{x})|d\boldsymbol{x}| - \int_B f(\boldsymbol{x})|d\boldsymbol{x}| \right| \leqslant \hbar\,\mathrm{vol}_n(B), \quad \forall \nu \geqslant N(\hbar).$$

This last inequality proves (15.1.11). □

Let us interrupt the flow of arguments to take stock of what we have achieved so far.

- We have defined concepts of Riemann integrability/integral associated to functions of several variables defined on a box B.
- We showed that the set $\mathcal{R}(B)$ of functions that are Riemann integrable on B is quite large: it contains all the continuous functions, and it is closed with respect to the algebraic operations of addition and multiplication of functions.
- The uniform limits of Riemann integrable functions are Riemann integrable.

If we compare the current state of affairs with the one-dimensional situation we realize that we have several glaring gaps in our developing story. First, our supply of Riemann integrable functions is still "meagre" since, unlike the one-dimensional case, we have not yet produced any example of a Riemann integrable function that is not continuous. Second, we have not yet indicated any concrete and practical way of computing Riemann integrals of functions of several variables.

The first issue is resolved by a remarkable result of *Henri Lebesgue*.[2] To state it we need to define the concept of *negligible subset of* $\mathbb{R}^n$.

Definition 15.1.15. A subset $S \subset \mathbb{R}^n$ is called *negligible* if, for any $\varepsilon > 0$, there exists a countable family $(B_\nu)_{\nu\in\mathbb{N}}$ of *closed* boxes in $\mathbb{R}^n$ that covers S and such that

$$\sum_{\nu\in\mathbb{N}} \mathrm{vol}_n(B_\nu) < \varepsilon.$$

□

Example 15.1.16. Suppose that $f : [a,b] \to \mathbb{R}$ is a Riemann integrable function. Then its graph

$$\Gamma_f := \big\{\, (x, f(x)\,) \in \mathbb{R}^2;\;\; x \in [a,b] \,\big\}$$

is a negligible subset of $\mathbb{R}^2$.

To see this fix $\varepsilon > 0$ and choose a partition $\boldsymbol{P}$ of $[a,b]$ such that $\omega(f,\boldsymbol{P}) < \varepsilon$. Suppose that

$$\boldsymbol{P} = a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

[2] Henri Lebesgue (1875-1941) was a French mathematician famous for his theory of integration; see Wikipedia for more details on his life and work.

set

$$m_i := \inf_{x \in [x_{i-1}, x_i]} f(x), \quad M_i := \sup_{x \in [x_{i-1}, x_i]} f(x), \quad i = 1, \dots, n.$$

For $i = 1, \dots, n$ we denote by B_i the box $[x_{i-1}, x_i] \times [m_i, M_i]$. From the definition of m_i and M_i we deduce that

$$\Gamma_f \subset \bigcup_{i=1}^{n} B_i,$$

$$\sum_{i=1}^{n} \operatorname{vol}_2(B_i) = \sum_{i=1}^{n} \big(M_i - m_i \big) \big(x_i - x_{i-1} \big) = \omega(f, \boldsymbol{P}) < \varepsilon.$$

This shows that, for any $\varepsilon > 0$ we can find a fine collection of rectangles that covers the graph and such that the sum of their areas is $< \varepsilon$.

□

In Exercise 15.7 we describe several examples of negligible sets, and some elementary properties of such sets. We have the following result that vastly generalizes Corollary 15.1.11.

Theorem 15.1.17 (Lebesgue). *Let $n \in \mathbb{N}$ and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a bounded function that is identically zero outside some box $B \subset \mathbb{R}^n$. Then the following statements are equivalent.*

(i) *The function f is Riemann integrable.*

(ii) *The set of points of discontinuity of f is negligible.*

□

For a proof we refer to [**45**, §11.1.2].

15.1.2. A conditional Fubini theorem. In this subsection we take a stab at the second problem and we describe a very versatile result showing that, under certain conditions, one can reduce the computation of an integral of a function of n-variables to the computation of Riemann integrals of functions with *fewer than n variables.*

Theorem 15.1.18 (Fubini). *Let $m, n \in \mathbb{N}$. Suppose that*

$$B^m = [a_1, b_1] \times \cdots \times [a_m, b_m]$$

is a nondegenerate box in $\mathbb{R}^m$ and

$$B^n = [a_{m+1}, b_{m+1}] \times \cdots \times [a_{m+n}, b_{m+n}]$$

is a nondegenerate box in $\mathbb{R}^n$. Suppose that

- *the function $f : B^m \times B^n \to \mathbb{R}$ is Riemann integrable on the box*

$$B = B^m \times B^n \subset \mathbb{R}^{m+n}$$

and,

- *for any $\boldsymbol{x} \in B^m$, the function*

$$B^n \ni \boldsymbol{y} \mapsto f_{\boldsymbol{x}}(\boldsymbol{y}) := f(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}$$

  is Riemann integrable.

Then, the marginal *(function)*

$$B^m \ni \boldsymbol{x} \mapsto M_f^1(\boldsymbol{x}) := \int_{B^n} f(\boldsymbol{x}, \boldsymbol{y}) |d\boldsymbol{y}| \in \mathbb{R}$$

is Riemann integrable and

$$\boxed{\int_{B^m \times B^n} f(\boldsymbol{x}, \boldsymbol{y}) |d\boldsymbol{x} d\boldsymbol{y}| = \int_{B^m} M_f^1(\boldsymbol{x}) |d\boldsymbol{x}| = \int_{B^m} \left(\int_{B^n} f(\boldsymbol{x}, \boldsymbol{y}) |d\boldsymbol{y}| \right) |d\boldsymbol{x}|.} \tag{15.1.15}$$

The last term in the above equality is called a repeated *or* iterated integral. *Often, for mnemonic purposes, we use the alternate notation*

$$\int_{B^m} |d\boldsymbol{x}| \left(\int_{B^n} f(\boldsymbol{x}, \boldsymbol{y}) |d\boldsymbol{y}| \right) := \int_{B^m} \left(\int_{B^n} f(\boldsymbol{x}, \boldsymbol{y}) |d\boldsymbol{y}| \right) |d\boldsymbol{x}|.$$

Main idea behind Fubini Before we embark in the proof we want to explain the simple principle behind this result. Suppose that we want to add all the numbers situated at the nodes of a grid such as the one in Figure 15.2.

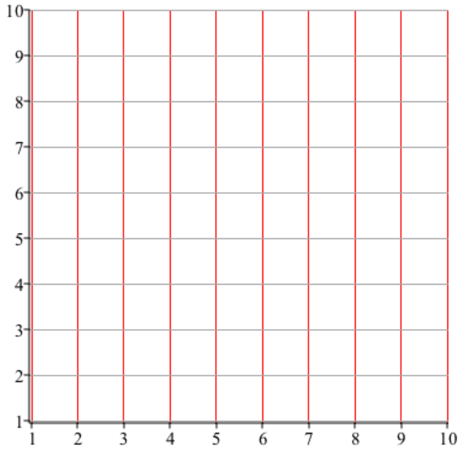

Figure 15.2. *Adding numbers on a grid.*

Fubini says that one could proceed as follows: for each number $k = 1, \ldots, 10$ on the horizontal (or the first) margin there is a tower of grid nodes above it. Add the numbers above it to obtain the value M_k^1 (the 1st marginal at k). Next add all these marginal values $M_1^1 + \ldots + M_{10}^1$ to recover the sum of all the numbers situated at nodes. One can proceed in a similar fashion using the vertical margin, obtaining first a marginal M^2.

Proof. We follow the approach in [35, Thm.3-10]. Consider a partition

$$\boldsymbol{P}_\varepsilon = (\boldsymbol{P}_1, \dots, \boldsymbol{P}_m, \boldsymbol{P}_{m+1}, \dots, \dots, \boldsymbol{P}_{m+n})$$

of B. We denote by $\boldsymbol{P}^m$ the partition

$$(\boldsymbol{P}_1, \dots, \boldsymbol{P}_m)$$

of B^m, and by $\boldsymbol{P}^n$ the partition

$$(\boldsymbol{P}_{m+1}, \dots, \dots, \boldsymbol{P}_{m+n})$$

of B^n. Every chamber C of $\boldsymbol{P}$ is the Cartesian product of a chamber C^m of $\boldsymbol{P}^m$ and a chamber C^n of $\boldsymbol{P}^n$. Moreover

$$\operatorname{vol}_{m+n}(C) = \operatorname{vol}_m(C^m) \operatorname{vol}_n(C^n).$$

For simplicity we set $\mathscr{C}^m := \mathscr{C}(\boldsymbol{P}^m)$ and $\mathscr{C}^n := \mathscr{C}(\boldsymbol{P}^n)$. To simplify the exposition we will continue to use the notation

$$m_U(g) := \inf_{u \in U} g(u), \;\; M_U(g) := \sup_{u \in U} g(u),$$

for any set U and for any bounded real valued function g defined on a set containing U.

We have

$$\begin{aligned}\boldsymbol{S}_*(f, \boldsymbol{P}) &= \sum_{\substack{C^m \in \mathscr{C}^m, \\ C^n \in \mathscr{C}^n}} m_{C^m \times C^n}(f) \cdot \operatorname{vol}_m(C^m) \operatorname{vol}_n(C^n) \\ &= \sum_{C^m \in \mathscr{C}^m} \left(\sum_{C^n \in \mathscr{C}^n} m_{C^m \times C^n}(f) \cdot \operatorname{vol}_n(C^n) \right) \operatorname{vol}_m(C^m).\end{aligned}$$

Now observe that for any $C^m \in \mathscr{C}^m$, any $C^n \in \mathscr{C}^n$, and any $\boldsymbol{x} \in C^m$ we have

$$m_{C^m \times C^n}(f) \leqslant m_{C^n}(f_{\boldsymbol{x}}).$$

Hence, for any $\boldsymbol{x} \in \mathscr{C}^m$ we have

$$\begin{aligned}\sum_{C^n \in \mathscr{C}^n} m_{C^m \times C^n}(f) \cdot \operatorname{vol}_n(C^n) &\leqslant \sum_{C^n \in \mathscr{C}^n} m_{C^n}(f_{\boldsymbol{x}}) \cdot \operatorname{vol}_n(C^n) = \boldsymbol{S}_*(f_{\boldsymbol{x}}, \boldsymbol{P}^n) \\ &\leqslant \int_{B^n} f_{\boldsymbol{x}}(\boldsymbol{y}) |d\boldsymbol{y}| = M_f^1(\boldsymbol{x}).\end{aligned}$$

Thus

$$\sum_{C^n \in \mathscr{C}^n} m_{C^m \times C^n}(f) \cdot \operatorname{vol}_n(C^n) \leqslant \inf_{\boldsymbol{x} \in C^m} M_f^1(\boldsymbol{x})$$

so that

$$\begin{aligned}\boldsymbol{S}_*(f, \boldsymbol{P}) &= \sum_{C^m \in \mathscr{C}^m} \left(\sum_{C^n \in \mathscr{C}^n} m_{C^m \times C^n}(f) \cdot \operatorname{vol}_n(C^n) \right) \operatorname{vol}_m(C^m) \\ &\leqslant \sum_{C^m \in \mathscr{C}^m} \inf_{\boldsymbol{x} \in C^m} M_f^1(\boldsymbol{x}) \cdot \operatorname{vol}_m(C^m) = \boldsymbol{S}_*\big(M_f^1, \boldsymbol{P}^m \big).\end{aligned}$$

A similar argument shows that

$$\boldsymbol{S}^*\big(M_f^1, \boldsymbol{P}^m \big) \leqslant \boldsymbol{S}^*(f, \boldsymbol{P}).$$

Hence, for any partition $\boldsymbol{P}$ of $B^m \times B^n$ we have

$$\boldsymbol{S}_*(f,\boldsymbol{P}) \leqslant \boldsymbol{S}_*\big(\, M_f^1, \boldsymbol{P}^m \,\big) \leqslant \boldsymbol{S}^*\big(\, M_f^1, \boldsymbol{P}^m \,\big) \leqslant \boldsymbol{S}^*(f,\boldsymbol{P}).$$

We deduce from the above that

$$\underline{\int}_{B^m\times B^n} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}||d\boldsymbol{y}| \leqslant \underline{\int}_{B^m} M_f^1(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \overline{\int}_{B^m} M_f^1(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \overline{\int}_{B^m\times B^n} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}||d\boldsymbol{y}|.$$

Since f is Riemann integrable,

$$\underline{\int}_{B^m\times B^n} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}||d\boldsymbol{y}| = \overline{\int}_{B^m\times B^n} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}||d\boldsymbol{y}|$$

so all the above inequalities are in fact equalities. □

Remark 15.1.19. (a) Note that when $f : B^m \times B^n \to \mathbb{R}$ is continuous, all the assumptions in Theorem 15.1.18 are automatically satisfied. In particular, if $f : [a_1,b_1]\times\cdots\times[a_n,b_n] \to \mathbb{R}$ is continuous, then

$$\begin{aligned}
&\int_{[a_1,b_1]\times\cdots\times[a_n,b_n]} f(x^1,\dots,x^n)|dx^1\cdots dx^n| \\
&= \int_{[a_1,b_1]} |dx^1| \int_{[a_2,b_2]\times\cdots\times[a_n,b_n]} f(x^1,x^2,\dots,x^n)|dx^2\cdots dx^n| \\
&= \int_{[a_1,b_1]} |dx^1| \int_{[a_2,b_2]} |dx^2| \int_{[a_3,b_3]\times\cdots\times[a_n,b_n]} f(x^1,x^2,x^3,\dots,x^n)|dx^2\cdots dx^n| \\
&= \int_{[a_1,b_1]} |dx^1| \int_{[a_2,b_2]} |dx^2| \cdots \int_{[a_n,b_n]} f(x^1,x^2,\dots,x^n)|dx^n| \\
&= \int_{a_1}^{b_1} dx^1 \int_{a_2}^{b_2} dx^2 \cdots \int_{a_n}^{b_n} f(x^1,x^2,\dots,x^n)dx^n.
\end{aligned}$$

For example

$$\begin{aligned}
&\int_{[0,\pi/2]\times[0,\pi]} \sin(x+y)|dxdy| = \int_0^{\frac{\pi}{2}} dx \int_0^{\pi} \sin(x+y)\,dy \\
&= \int_0^{\frac{\pi}{2}} \Big(-\cos(x+y)\big|_{y=0}^{y=\pi} \Big) dx = \int_0^{\frac{\pi}{2}} \Big(\cos x - \cos(\pi + x) \Big) dx \\
&= \int_0^{\frac{\pi}{2}} \cos x\,dx - \int_0^{\frac{\pi}{2}} \cos(x+\pi)dx \\
&= \sin x\Big|_{x=0}^{x=\frac{\pi}{2}} - \sin(x+\pi)\Big|_{x=0}^{x=\frac{\pi}{2}} = \sin\frac{\pi}{2} - \sin\frac{3\pi}{2} + \sin\pi = 2.
\end{aligned}$$

(b) A completely similar argument shows that when $f : B^m \times B^n \to \mathbb{R}$ is Riemann integrable and, for any $\boldsymbol{y} \in B^n$, the function

$$f_{\boldsymbol{y}} : B^m \to \mathbb{R},\quad f_{\boldsymbol{y}}(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{y}),$$

is Riemann integrable, then the second *marginal* function

$$M_f^2 : B^n \to \mathbb{R},\ \ M_f^2(\boldsymbol{y}) = \int_{B^n} f_{\boldsymbol{y}}(\boldsymbol{x})|d\boldsymbol{x}|,$$

is Riemann integrable and we have

$$\boxed{\int_{B^m\times B^n} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}|\,|d\boldsymbol{y}| = \int_{B^n} M_f^2(\boldsymbol{y})|d\boldsymbol{y}| = \int_{B^n}\left(\int_{B^m} f(\boldsymbol{x},\boldsymbol{y})|d\boldsymbol{x}|\right)|d\boldsymbol{y}|.} \tag{15.1.16}$$

□

Remark 15.1.20 (Changing the order of integration)**.** Suppose $B = [a,b]\times[c,d] \subset\in \mathbb{R}^2$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a Riemann integrable function such that for any $x \in [a,b]$ the function

$$[c,d] \ni y \mapsto f(x,y)$$

is Riemann integrable and, for any $y \in [c,d]$, the function

$$[a,b] \ni x \mapsto f(x,y)$$

is Riemann integrable. We obtain in this fashion *two* marginals

$$M_f^1 : [a,b] \to \mathbb{R},\ \ M_f^1(x) = \int_c^d f(x,y)dy$$

and,

$$M_f^2 : [c,d] \to \mathbb{R},\ \ M_f^2(y) = \int_a^b f(x,y)dx.$$

Fubini's theorem then implies that

$$\int_a^b M_f^1(x)dx = \int_{[a,b]\times[c,d]} f(x,y)|dxdy| = \int_c^d M_f^2(y)dy.$$

Using the concrete definitions of the marginals we obtain the equality

$$\boxed{\int_a^b dx\int_c^d f(x,y)dy = \int_c^d dy\int_a^b f(x,y)dx}. \tag{15.1.17}$$

This equality often leads to surprising conclusions and it is commonly known as the *changing-the-order-of-integration trick.* □

15.1.3. Functions Riemann integrable over $\mathbb{R}^n$. Let $n \in \mathbb{N}$ and suppose that $X \subset \mathbb{R}^n$ is an arbitrary set. For any function $f : X \to \mathbb{R}$ we denote by f^0 its *extension by zero outside* X, i.e., f^0 is defined on the entire space $\mathbb{R}^n$, and

$$f^0(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & \boldsymbol{x} \in X,\\ 0, & \boldsymbol{x} \in \mathbb{R}^n\backslash X.\end{cases}$$

Proposition 15.1.21. *Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a Riemann integrable function. Then, for any box B' that contains B, the restriction $f^0_{B'}$ of f^0 to B' is Riemann integrable and*

$$\int_{B'} f^0_{B'}(\boldsymbol{x})\,|d\boldsymbol{x}| = \int_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

Proof. It is important to have a heuristic explanation why the above result is plausible. We know that the Riemann integral can be very well approximated by appropriate Riemann sums. Start with a partition $\boldsymbol{P}$ of the inside box B. Extend it in some way to a partition $\boldsymbol{P'}$ of the surrounding box B'. Observe that a "typical" Riemann sum of $f^0_{B'}$ associated to $\boldsymbol{P'}$ is equal to a Riemann sum of f associated to $\boldsymbol{P}$ since the value of f^0 outside B is 0. Thus the Riemann integral of f over B ought to be arbitrarily close to the Riemann integral of f^0 over B'.

The set D_f of points of discontinuity of f is negligible since f is Riemann integrable. The set of points of discontinuity of $f^0_{B'}$ is contained in the union $D_f \cup \partial B$. Since each of the faces of B is contained in a coordinate hyperplane, we deduce from Exercise 15.7 that each facet is negligible. The boundary ∂B is the union of facets and thus it is negligible; see Exercise 15.7. Hence $f^0_{B'}$ is Riemann integrable.

Fix $\varepsilon > 0$. Now choose a partition $\boldsymbol{P}$ of B such that

$$\omega(f, \boldsymbol{P}) < \frac{\varepsilon}{2}.$$

Now, extend $\boldsymbol{P}$ to a partition $\boldsymbol{P'}$ of B'. Since $f^0_{B'}$ is Riemann integrable we can find a partition $\boldsymbol{Q'} \succ \boldsymbol{P'}$ of B' such that

$$\omega\big(\, f^0_{B'}\, \boldsymbol{Q'}\big) < \frac{\varepsilon}{2}. \tag{15.1.18}$$

The partition $\boldsymbol{Q'}$ induces a partition $\boldsymbol{Q}$ of B that is finer than $\boldsymbol{P}$. Thus

$$\omega(f, \boldsymbol{Q}) \leqslant \omega(f, \boldsymbol{P}) < \frac{\varepsilon}{2}. \tag{15.1.19}$$

Now choose a sample $\underline{\xi}$ of $\boldsymbol{Q'}$ such that, for each chamber C of $\boldsymbol{Q'}$ not contained in B, the corresponding sample $\xi(C)$ is contained in the interior of C. In particular, this shows that $f^0\big(\,\xi(C)\,\big) = 0$ for such a chamber and sample point. We deduce

$$\boldsymbol{S}(f^0_{B'}, \boldsymbol{Q'}, \underline{\xi}) = \sum_{C \in \mathscr{C}(\boldsymbol{Q'})} f^0\big(\,\xi(C)\,\big)\operatorname{vol}_n(C) = \sum_{\substack{C \in \mathscr{C}(\boldsymbol{Q'})\\ C \subset B}} f\big(\,\xi(C)\,\big)\operatorname{vol}_n(C)$$

$$= \sum_{C \in \mathscr{C}(\boldsymbol{Q})} f\big(\,\xi(C)\,\big)\operatorname{vol}_n(C) = \boldsymbol{S}(f, \boldsymbol{Q}, \underline{\xi}).$$

On the other hand, Proposition 15.1.9 coupled with (15.1.18) and (15.1.19) imply that

$$\left|\int_{B'} f^0_{B'}(\boldsymbol{x})|d\boldsymbol{x}| - \boldsymbol{S}(f^0_{B'}, \boldsymbol{Q'}, \underline{\xi})\right| < \omega\big(\, f^0_{B'}\, \boldsymbol{Q'}\big) < \frac{\varepsilon}{2},$$

$$\left|\int_B f(\boldsymbol{x})|d\boldsymbol{x}| - \boldsymbol{S}(f, \boldsymbol{Q}, \underline{\xi})\right| < \omega\big(\, f^0_{B'}\, \boldsymbol{Q'}\big) < \frac{\varepsilon}{2}.$$

Hence

$$\left|\int_{B'} f^0_{B'}(\boldsymbol{x})|d\boldsymbol{x}| - \int_B f(\boldsymbol{x})|d\boldsymbol{x}|\right| < \varepsilon, \;\; \forall \varepsilon > 0,$$

and thus,

$$\int_{B'} f^0_{B'}(\boldsymbol{x})|d\boldsymbol{x}| = \int_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

□

Let us introduce an important concept.

Definition 15.1.22. Let $n \in \mathbb{N}$. The *indicator function* of a subset $S \subset \mathbb{R}^n$ is the function

$$I_S : \mathbb{R}^n \to \mathbb{R},\ \ I_S(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in S, \\ 0, & \boldsymbol{x} \in \mathbb{R}^n \setminus S. \end{cases}$$

In other words, I_S is the extension by 0 of the function on S equal to the constant 1. □

If B, B' are nondegenerate boxes such that $B \subset B'$, then Proposition 15.1.21 shows that the restriction of I_B to B' is Riemann integrable and it is discontinuous if $B \neq B'$. Moreover

$$\int_{B'} I_B|_{B'}(\boldsymbol{x})|d\boldsymbol{x}| = \int_B |d\boldsymbol{x}| = \operatorname{vol}_n(B).$$

Definition 15.1.23. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *Riemann integrable* (over $\mathbb{R}^n$) if it satisfies the following conditions.

(i) There exists a nondegenerate box $B \subset \mathbb{R}^n$ such that $f(\boldsymbol{x}) = 0$, $\forall \boldsymbol{x} \in \mathbb{R}^n \setminus B$.

(ii) The restriction of $f|_B$ of f to B is Riemann integrable.

We set

$$\int_{\mathbb{R}^n} f(\boldsymbol{x})|d\boldsymbol{x}| := \int_B f|_B(\boldsymbol{x})|d\boldsymbol{x}|,$$

where B is a box satisfying the conditions (i) and (ii) above. We denote by $\mathcal{R}_n$ the set of Riemann integrable functions $f : \mathbb{R}^n \to \mathbb{R}$. □

Remark 15.1.24. (a) From the definition we deduce that if $f : \mathbb{R}^n \to \mathbb{R}$ is Riemann integrable, then it must have compact support.

(b) We need to verify the consistency of the above definition of the integral. More precisely, we need to verify that, if $f : \mathbb{R}^n \to \mathbb{R}$ is Riemann integrable and B_1, B_2 are two boxes satisfying the conditions (i)+(ii) in the Definition 15.1.23, then

$$\int_{B_1} f|_{B_1}(\boldsymbol{x})|d\boldsymbol{x}| = \int_{B_2} f|_{B_2}(\boldsymbol{x})|d\boldsymbol{x}|.$$

To prove this choose a box B such that $B \supset B_1 \cup B_2$. Observe that the function f can be identified with the extension by 0 of either functions $f|_{B_1}$ and $f|_{B_2}$. From Proposition 15.1.21 we now deduce that $f|_B$ is Riemann integrable (on B) and

$$\int_{B_1} f|_{B_1}(\boldsymbol{x})|d\boldsymbol{x}| = \int_B f|_B(\boldsymbol{x})|d\boldsymbol{x}| = \int_{B_2} f|_{B_2}(\boldsymbol{x})|d\boldsymbol{x}|.$$

We see that the indicator function of a nondegenerate (closed) box $B \subset \mathbb{R}^n$ is Riemann integrable and

$$\int_{\mathbb{R}^n} I_B(\boldsymbol{x})|d\boldsymbol{x}| = \operatorname{vol}_n(B).$$

(c) Theorem 15.1.13 shows that if $f, g \in \mathcal{R}_n$ and $s, t \in \mathbb{R}$, then

$$sf + tg,\ fg \in \mathcal{R}_n$$

and

$$\int_{\mathbb{R}^n} \big(sf(\boldsymbol{x}) + tg(\boldsymbol{x}) \big) d\boldsymbol{s} = s \int_{\mathbb{R}^n} f(\boldsymbol{x}) |d\boldsymbol{x}| + t \int_{\mathbb{R}^n} g(\boldsymbol{x}) |d\boldsymbol{x}|.$$

If additionally $f(\boldsymbol{x}) \leqslant g(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$, then

$$\int_{\mathbb{R}^n} f(\boldsymbol{x}) |d\boldsymbol{x}| \leqslant \int_{\mathbb{R}^n} g(\boldsymbol{x}) |d\boldsymbol{x}|.$$ □

Recall that $C_{\text{cpt}}(\mathbb{R}^n)$ denotes the set of continuous functions $\mathbb{R}^n \to \mathbb{R}$ with compact support.

Corollary 15.1.25. *Let $n \in \mathbb{N}$. Then $C_{\text{cpt}}(\mathbb{R}^n) \subset \mathcal{R}_n$, i.e., any compactly supported function $f : \mathbb{R}^n \to \mathbb{R}$ is Riemann integrable.*

Proof. Since the support of f is compact, there exists a (closed) box $B \supset \operatorname{supp}(f)$. Thus B contains all the points where f is nonzero so that f is identically zero outside B. Moreover since f is continuous, it is integrable on B.

□

The compactly supported continuous functions play an important role in the theory of Riemann integration due to the following approximation result.

Theorem 15.1.26. *Let $n \in \mathbb{N}$ and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a bounded function and U is an open set containing the support of f. Then the following statements are equivalent.*

(i) *The function f is Riemann integrable (on $\mathbb{R}^n$).*

(ii) *For any $\varepsilon > 0$ there exist functions $g, G \in C_{\text{cpt}}(\mathbb{R}^n)$ such that*

$$\operatorname{supp}(g),\ \operatorname{supp}(G) \subset U,$$

$$g(\boldsymbol{x}) \leqslant f(\boldsymbol{x}) \leqslant G(\boldsymbol{x}),\ \ \forall \boldsymbol{x} \in \mathbb{R}^n \ \ \textit{and}\ \ 0 \leqslant \int_{\mathbb{R}^n} \big(G(\boldsymbol{x}) - g(\boldsymbol{x}) \big) |d\boldsymbol{x}| < \varepsilon.$$

□

The proof of this theorem is not extremely demanding but it would distract us from the main “story”. The curious reader can find the details in [14, §6.9].

15.1.4. Volume and Jordan measurability. The concept of Riemann integral can be used to define the notion of n-dimensional volume. Intuitively, the n-dimensional volume of subset $S \subset \mathbb{R}^n$ ought to be a nonnegative number $\operatorname{vol}_n(S)$ that satisfies the Inclusion-Exclusion Principle

$$\operatorname{vol}_n(S_1 \cup S_2) = \operatorname{vol}_n(S_1) + \operatorname{vol}_n(S_2) - \operatorname{vol}_n(S_1 \cap S_2),$$

it "depends continuously" on S and, when S is a box, this notion of volume should coincide with our original definition (15.1.2). Additionally, we would like this volume to stay unchanged when we rigidly move S around $\mathbb{R}^n$. (Typical examples of rigid transformations are translations and rotations about an "axis".)

The famous Banach-Tarski "paradox"[3] shows that we cannot associate a notion of n-dimensional volume with the above properties to *all* subsets of $\mathbb{R}^n$. We can however associate a notion of volume with these properties to many subsets of $\mathbb{R}^n$.

Definition 15.1.27. (a) A bounded set $S \subset \mathbb{R}^n$ is called *Jordan*[4] *measurable* if the indicator function I_S is Riemann integrable. We denote by $\mathscr{J}(\mathbb{R}^n)$ the collection of Jordan measurable subsets of $\mathbb{R}^n$.

(b) The *n-dimensional volume* of a Jordan measurable set $S \subset \mathbb{R}^n$ is the nonnegative number

$$\operatorname{vol}_n(S) := \int_{\mathbb{R}^n} I_S(\boldsymbol{x})|d\boldsymbol{x}|. \qquad \square$$

Remark 15.1.28. If B is a (closed) box, then the volume of B as defined in the above definition, coincides with the volume of B as defined in (15.1.2). This follows from Example 15.1.7 and Proposition 15.1.21. □

We mention several useful consequences of the above result.

Corollary 15.1.29. *A bounded subset $S \subset \mathbb{R}^n$ is Jordan measurable if and only if its boundary ∂S is negligible.*

Proof. Note that S is Jordan measurable if and only if its indicator function

$$I_S : \mathbb{R}^n \to \mathbb{R}, \quad I_S(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in S, \\ 0, & \boldsymbol{x} \in \mathbb{R}^n \backslash S, \end{cases}$$

is Riemann integrable. According to Lebesgue's theorem this happens if and only if the set of points of discontinuity of I_S is negligible. Now observe that the boundary of S is precisely the set of discontinuities of I_S. □

Proposition 15.1.30. *Let $n \in \mathbb{N}$.*

(i) (Inclusion-Exclusion Principle) *If $S_1, S_2 \in \mathscr{J}(\mathbb{R}^n)$, then*

$$S_1 \cup S_2, \quad S_1 \cap S_2 \in \mathscr{J}(\mathbb{R}^n)$$

and

$$\operatorname{vol}_n(S_1 \cup S_2) = \operatorname{vol}_n(S_1) + \operatorname{vol}_n(S_2) - \operatorname{vol}_n(S_1 \cap S_2). \tag{15.1.20}$$

(ii) (Monotonicity) *If $S_1, S_2 \in \mathscr{J}(\mathbb{R}^n)$ and $S_1 \subset S_2$, then*

$$\operatorname{vol}_n(S_1) \leqslant \operatorname{vol}_n(S_2).$$

[3] Search Wikipedia for more details about this famous result.

[4] Named after the French mathematician Camille Jordan (1838-1922). See Wikipedia for more on Jordan.

Proof. (i) Since S_1, S_2 are bounded, there exists a box B that contains both S_1 and S_2. We deduce that the restrictions to B of both functions I_{S_1} and I_{S_2} are Riemann integrable. Thus the restrictions to B of the functions $I_{S_1} + I_{S_2}$ and $I_{S_1} I_{S_2}$ are Riemann integrable. Observing that

$$I_{S_1 \cap S_2} = I_{S_1} I_{S_2} \;\text{ and }\; I_{S_1 \cup S_2} = I_{S_1} + I_{S_2} - I_{S_1 \cap S_2}$$

we deduce that $S_1 \cap S_2$ and $S_1 \cup S_2$ are Jordan measurable and

$$\begin{aligned} \operatorname{vol}_n(S_1 \cup S_2) &= \int_{\mathbb{R}^n} \big(I_{S_1}(\boldsymbol{x}) + I_{S_2}(\boldsymbol{x}) - I_{S_1 \cap S_2}(\boldsymbol{x}) \big) |d\boldsymbol{x}| \\ &= \operatorname{vol}_n(S_1) + \operatorname{vol}_n(S_2) - \operatorname{vol}_n(S_1 \cap S_2). \end{aligned}$$

(i) Note that

$$S_1 \subset S_2 \Rightarrow I_{S_1}(\boldsymbol{x}) \leqslant I_{S_2}(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in \mathbb{R}^n \Rightarrow \int_{\mathbb{R}^n} I_{S_1}(\boldsymbol{x}) |d\boldsymbol{x}| \leqslant \int_{\mathbb{R}^n} I_{S_2}(\boldsymbol{x}) |d\boldsymbol{x}|.$$

□

Example 15.1.31 (Cavalieri's Principle)**.** Let $n \in \mathbb{N}$. Consider a Jordan measurable set $S \subset \mathbb{R} \times \mathbb{R}^n = \mathbb{R}^{1+n}$. We denote by $x^0, \dots, x^n$ the coordinates in $\mathbb{R}^{1+n}$. For any $t \in \mathbb{R}$ we denote by S_t the intersection of S with the hyperplane $\{x^0 = t\}$. We will refer to S_t as the *slice* of S over t; see Figure 15.3.

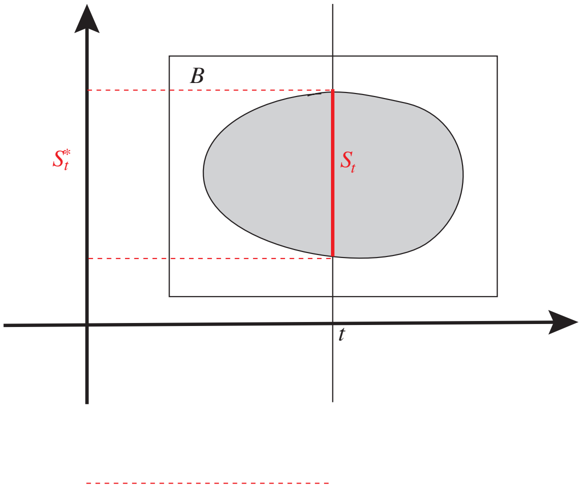

Figure 15.3. *Slicing a 2-dimensional region by vertical lines.*

Denote by S_t^* the projection of S_t on the coordinate subspace $\{0\} \times \mathbb{R}^n$ (pictured as the vertical axis in Figure 15.3). More precisely

$$S_t^* := \big\{ (x^1, \dots, x^n) \in \mathbb{R}^n; \;\; (t, x^1, \dots, x^n) \in S_t \big\}.$$

Cavalieri's principle states that if the all the (projected) slices $S_t^* \subset \mathbb{R}^n$ are Jordan measurable then

$$\operatorname{vol}_{n+1}(S) = \int_{\mathbb{R}} \operatorname{vol}_n(S_t^*)|dt|. \tag{15.1.21}$$

This is an immediate consequence of the Fubini Theorem 15.1.18. Since S is bounded, there exists a box

$$B = [a_0, b_0] \times \underbrace{[a_1, b_1] \times \cdots \times [a_n, b_n]}_{B'}.$$

Let $f : \mathbb{R}^{1+n} \to \mathbb{R}$ be the indicator function of S, $f = I_S$. Note that f is zero outside the box B. Since S is Jordan measurable, the restriction of f to B is Riemann integrable. For any $t \in \mathbb{R}$ the function

$$\mathbb{R} \ni (x^1, \dots, x^n) \mapsto f_t(x^1, \dots, x^n) \in \mathbb{R}$$

is the indicator function of S_t^*,

$$f_t(x^1, \dots, x^n) = I_{S_t^*}(x^1, \dots, x^n),$$

and thus it is Riemann integrable. Note that $f_t = 0$ if $t \notin [a_0, b_0]$. The marginal function is

$$M_f(t) = \int_{B'} f_t(x^1, \dots, x^n) dx^1 \cdots dx^n = \operatorname{vol}_n(S_t^*).$$

Fubini's Theorem now implies

$$\operatorname{vol}_{n+1}(S) = \int_B f(x^0, x^1, \dots, x^n)|dx^0 dx^1 \cdots dx^n| = \int_{a_0}^{b_0} \operatorname{vol}_n(S_t^*)|dt|. \qquad \square$$

15.1.5. The Riemann integral over arbitrary regions.

Definition 15.1.32. Let $S \subset \mathbb{R}^n$. A bounded function $f : S \to \mathbb{R}$ is called *Riemann integrable over* S if f^0, its extension by 0 to $\mathbb{R}^n$, is Riemann integrable over $\mathbb{R}^n$. In this case we define the Riemann integral of f over the set S to be

$$\int_S f(\boldsymbol{x})|d\boldsymbol{x}| := \int_{\mathbb{R}^n} f^0(\boldsymbol{x})|d\boldsymbol{x}|.$$

We denote by $\mathcal{R}_n(S)$ the set of Riemann integrable functions on S.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be Riemann integrable over S if $f\boldsymbol{I}_S$ is integrable over $\mathbb{R}^n$. □

Remark 15.1.33. (a) Note that we can rephrase the above definition in a more concise way,

$$\boxed{f \in \mathcal{R}_n(S) \Longleftrightarrow f^0 I_S \in \mathcal{R}_n}.$$

(b) Proposition 15.1.21 shows that, if B is a box, then the set $\mathcal{R}(B)$ of functions $f : B \to \mathbb{R}$ Riemann integrable in the sense of Definition 15.1.5 coincides with the set $\mathcal{R}_n(B)$ in the above definition.

c) If $f \in \mathcal{R}(S)$, then

$$\int_S f(\boldsymbol{x})\,|d\boldsymbol{x}| = \int_B f^0(\boldsymbol{x})\boldsymbol{I}_S(\boldsymbol{x})\,|d\boldsymbol{x}|,$$

where B is *any* box in $\mathbb{R}^n$ that contains S. □

Proposition 15.1.34. *Let $n \in \mathbb{N}$.*

(i) (Additivity of integrals with respect to domains) *Suppose that $S_1, S_2 \subset \mathbb{R}^n$ are Jordan measurable sets and $f : S_1 \cup S_2 \to \mathbb{R}$ is Riemann integrable. Then the restrictions of f to S_1 and S_2 are Riemann integrable and*

$$\int_{S_1\cup S_2} f(\boldsymbol{x})|d\boldsymbol{x}| = \int_{S_1} f(\boldsymbol{x})|d\boldsymbol{x}| + \int_{S_2} f(\boldsymbol{x})|d\boldsymbol{x}| - \int_{S_1\cap S_2} f(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.1.22}$$

(ii) (Monotonicity) *If S is Jordan measurable and $f, g : S \to \mathbb{R}$ are Riemann integrable functions such that $f(\boldsymbol{x}) \leqslant g(\boldsymbol{x})$, $\forall \boldsymbol{x} \in S$, then*

$$\int_S f(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \int_S g(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.1.23}$$

In particular

$$\left|\int_S f(\boldsymbol{x})|d\boldsymbol{x}|\right| \leqslant \Big(\sup_{\boldsymbol{x}\in S}|f(\boldsymbol{x})|\Big)\operatorname{vol}_n(S). \tag{15.1.24}$$

Proof. (i) Observe that $f^0, I_{S_1}, I_{S_2} \in \mathcal{R}_n$ so that

$$I_{S_1}f^0,\ \ I_{S_2}f^0 \in \mathcal{R}_n,$$

$$f^0 = I_{S_1\cup S_2}f^0 = I_{S_1}f^0 + I_{S_2}f^0 - I_{S_1\cap S_2}f^0.$$

The equality (15.1.22) follows by integrating over $\mathbb{R}^n$ the above equality.

(ii) The equality (15.1.23) follows from Remark 15.1.24(b) by observing that $f^0(\boldsymbol{x}) \leqslant g^0(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. The inequality (15.1.24) follows by integrating over S the inequality

$$-\Big(\sup_{\boldsymbol{x}\in S}|f(\boldsymbol{x})|\Big) \leqslant f(\boldsymbol{x}) \leqslant \Big(\sup_{\boldsymbol{x}\in S}|f(\boldsymbol{x})|\Big),\ \ \forall \boldsymbol{x} \in S.$$

□

Corollary 15.1.35. *Let $n \in \mathbb{N}$. Suppose that $S_1, S_2 \subset \mathbb{R}^n$ are Jordan measurable sets and $f : S_1 \cup S_2 \to \mathbb{R}$ is Riemann integrable. If $\operatorname{vol}_n(S_1 \cap S_2) = 0$, then*

$$\int_{S_1\cup S_2} f(\boldsymbol{x})|d\boldsymbol{x}| = \int_{S_1} f(\boldsymbol{x})|d\boldsymbol{x}| + \int_{S_2} f(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.1.25}$$

Proof. From (15.1.24) we deduce that

$$\int_{S_1\cap S_2} f(\boldsymbol{x})|d\boldsymbol{x}| = 0.$$

We now see that (15.1.25) is a special case of (15.1.22). □

Proposition 15.1.36. *Suppose that $K \subset \mathbb{R}^n$ is a* compact *Jordan measurable set and $f : K \to \mathbb{R}$ is a continuous function. Then f is integrable over K.*

Proof. Fix a box B that contains K. As usual, denote by f^0 the extension by 0 of f. Observe that the set of points of discontinuity of f^0 is contained in the boundary ∂K which is negligible since K is Jordan measurable. Lebesgue's Theorem 15.1.17 then shows that f^0 is integrable. □

The next result is also a consequence of Lebesgue's Theorem.

Corollary 15.1.37. *Suppose that $K \subset \mathbb{R}^n$ is a compact Jordan measurable set, $Z \subset \mathbb{R}^n$ is a negligible closed subset and $f : K \to \mathbb{R}$ is a bounded function. Then the following are equivalent.*

(i) *The function f is Riemann integrable over K.*

(ii) *The function f is Riemann integrable over $K \backslash Z$.*

If either (i) or (ii) holds, then

$$\int_{K\backslash Z} f(\boldsymbol{x})\,|d\boldsymbol{x}| = \int_K f(\boldsymbol{x})\,|d\boldsymbol{x}|.$$ □

15.2. Fubini theorem and iterated integrals

We now have at our disposal all the information we need to prove a version of the Fubini Theorem 15.1.18 that involves easily verifiable assumptions.

15.2.1. An unconditional Fubini theorem. Before we can state the version of the Fubini theorem most frequently used in applications we need to introduce a very versatile concept.

Definition 15.2.1. Let $n \in \mathbb{N}$. A compact set $D \subset \mathbb{R}^{n+1}$ is called a *domain of simple type* if there exist a *compact Jordan measurable* set $K \subset \mathbb{R}^n$ and *continuous* functions

$$\beta, \tau : K \to \mathbb{R}$$

with the following properties

- $\beta(x^1, \dots, x^n) \leqslant \tau(x^1, \dots, x^n)$, $\forall (x^1, \dots, x^n) \in K$.
- The point $(x^1, \dots, x^n, y) \in \mathbb{R}^{n+1}$ belongs to D if and only if
$$(x^1, \dots, x^n) \in K \ \text{ and } \ \beta(x^1, \dots, x^n) \leqslant y \leqslant \tau(x^1, \dots, x^n).$$

We will denote this domain by $D(K, \beta, \tau)$. The region K is called the *cross section* of D, the function β is called the *bottom* of D and the function τ is called the *top* of D; see Figure 15.4. □

Thus $D(K, \beta, \tau)$ is the region between the graphs of β and τ, where β sits at the bottom and τ at the top.

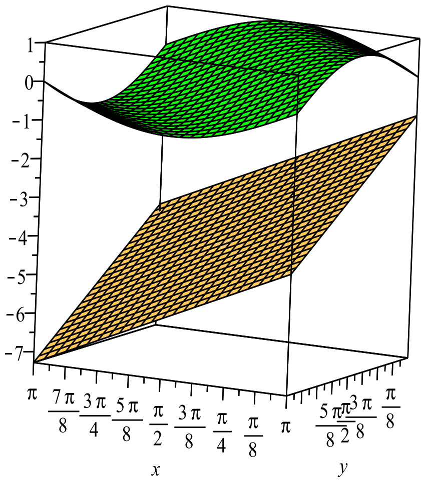

Figure 15.4. *A simple type region in* $\mathbb{R}^3$ *with cross section* $K = [0, \pi] \times [0, \pi]$, *a flat bottom* $\beta(x, y) = -2 - x - y$ *and curved top* $\tau(x, y) = \sin(x + y)$.

Proposition 15.2.2. *Let* $n \in \mathbb{N}$. *Suppose that* $D = D(K, \beta, \tau) \subset \mathbb{R}^{n+1}$ *is a simple type domain with cross section* $K \subset \mathbb{R}^n$ *and top/bottom functions* $\tau, \beta : K \to \mathbb{R}$. *Then* D *is compact and Jordan measurable.*

Proof. As usual we denote by $m_S(f)$ and $M_S(f)$ the infimum and respectively the supremum of a function $f : S \to \mathbb{R}$. Note that

$$D(K, \beta, \tau) \subset K \times [m_\beta, M_\tau]$$

so $D(K, \beta, \tau)$ is bounded. Since K is compact, we deduce that K is closed. The continuity of β, τ shows that $D(K, \beta, \tau)$ is closed and thus compact.

Fix a box $B \subset \mathbb{R}^n$ that contains K. Set

$$\widehat{B} = B \times I, \;\; I := [m_\beta, M_\tau], \;\; L = M_\tau - m_\beta.$$

Fix a partition $\boldsymbol{P} = (\boldsymbol{P}_1, \dots, \boldsymbol{P}_n)$ of B. The set of chambers $\mathscr{C}(\boldsymbol{P})$ decomposes as

$$\mathscr{C}(\boldsymbol{P}) = \mathscr{C}_e(\boldsymbol{P}) \cup \mathscr{C}_b(\boldsymbol{P}) \cup \mathscr{C}_i(\boldsymbol{P}),$$

where $\mathscr{C}_e(\boldsymbol{P})$ consists of chambers that do not intersect K, $\mathscr{C}_i(\boldsymbol{P})$ consists of chambers located in the interior of K and $\mathscr{C}_b(\boldsymbol{P})$ consists of chambers that intersect both K and its complement. For $\nu \in \mathbb{N}$ we denote by $\boldsymbol{U}_\nu$ the uniform partition of I into ν sub-intervals of equal length L/ν. We denote by $I_1, \dots, I_\nu$ sub-intervals of the partition $\boldsymbol{U}_\nu$.

We obtain a partition $\widehat{\boldsymbol{P}}_\nu = (\boldsymbol{P}_1, \dots, \boldsymbol{P}_n, \boldsymbol{U}_\nu$ of $\widehat{B} = B \times I$. The chambers of $\widehat{\boldsymbol{P}}_\nu$ have the form $C \times I_k$, $C \in \mathscr{C}(\boldsymbol{P})$, $k = 1, \dots, \nu$. Note that

$$\operatorname{osc}(I_D, C \times I_k) = 0, \;\; \forall C \in \mathscr{C}_e(\boldsymbol{P}), \; k = 1, \dots, \nu,$$

and

$$\operatorname{osc}(I_D, C \times I_k) \leqslant 1, \;\; \forall C \in \mathscr{C}_b(\boldsymbol{P}), \; k = 1, \dots, \nu$$

In particular, we deduce that,

$$\forall C \in \mathscr{C}_b(\boldsymbol{P}),\quad \sum_{k=1}^{\nu} \operatorname{osc}(I_D, C\times I_k)\operatorname{vol}_{n+1}(C\times I_k) \leqslant \sum_{k=1}^{\nu}\operatorname{vol}_n(C)\frac{L}{\nu} = L\operatorname{vol}_n(C).$$

If $C \in \mathscr{C}_i(\boldsymbol{P})$, and $I_k \subset \big(\,M_C(\beta), m_C(\tau)\,\big)$, then $\operatorname{osc}(I_D, C\times I_k\,) = 0$. We deduce that if $\operatorname{osc}(I_D, C\times I_k\,) = 1$, then

$$I_k \subset \mathcal{I}_\nu(C) := \left[m_C(\beta) - \frac{L}{\nu}, M_C(\beta) + \frac{L}{\nu}\right] \cup \left[m_C(\tau) - \frac{L}{\nu}, M_C(\tau) + \frac{\nu}{\nu}\right].$$

Hence, $\forall C \in \mathscr{C}_i(\boldsymbol{P})$ we have

$$\sum_{k=1}^{\nu} \operatorname{osc}(I_D, C\times I_k)\operatorname{vol}_{n+1}(C\times I_k) \leqslant \operatorname{vol}_n(C)\sum_{k\in\mathcal{I}_\nu(C)}\operatorname{vol}_1(I_k)$$

$$\leqslant \operatorname{vol}_n(C)\operatorname{vol}_1\big(\,\mathcal{I}_\nu(C)\,\big) = \left(\,\operatorname{osc}(\beta, C) + \operatorname{osc}(\tau, C) + \frac{4L}{\nu}\,\right)\operatorname{vol}_n(C).$$

Putting together all of the above we deduce

$$\begin{aligned}\omega(I_D, \widehat{\boldsymbol{P}}_\nu) &= \sum_{C\in\mathscr{C}_b(\boldsymbol{P})}\sum_{k=1}^{\nu}\operatorname{osc}(I_D, C\times I_k)\operatorname{vol}_{n+1}(C\times I_k)\\ &\quad+ \sum_{C\in\mathscr{C}_i(\boldsymbol{P})}\sum_{k=1}^{\nu}\operatorname{osc}(I_D, C\times I_k)\operatorname{vol}_{n+1}(C\times I_k)\\ &\leqslant L\sum_{C\in\mathscr{C}_b(\boldsymbol{P})}\operatorname{vol}_n(C) + \frac{4L}{\nu}\underbrace{\sum_{C\in\mathscr{C}_i(\boldsymbol{P})}\operatorname{vol}_n(C)}_{\leqslant \operatorname{vol}_n(B)} + \sum_{C\in\mathscr{C}_i(\boldsymbol{P})}\big(\,\operatorname{osc}(\beta, C) + \operatorname{osc}(\tau, C)\,\big)\operatorname{vol}_n(C).\end{aligned}$$

Hence

$$\begin{aligned}\omega(I_D, \widehat{\boldsymbol{P}}_\nu) &\leqslant L\sum_{C\in\mathscr{C}_b(\boldsymbol{P})}\operatorname{vol}_n(C) + \frac{4L\operatorname{vol}_n(B)}{\nu}\\ &\quad+ \sum_{C\in\mathscr{C}_i(\boldsymbol{P})}\big(\,\operatorname{osc}(\beta, C) + \operatorname{osc}(\tau, C)\,\big)\operatorname{vol}_n(C).\end{aligned}\tag{15.2.1}$$

Fix $\varepsilon > 0$. Since K is Jordan measurable, we can find a partition $\boldsymbol{Q}_\varepsilon$ of B such that

$$\sum_{C\in\mathscr{C}_b(\boldsymbol{Q}^\varepsilon)}\operatorname{vol}_n(C) = \omega(I_K, \boldsymbol{Q}_\varepsilon) < \frac{\varepsilon}{3}.$$

Choose $\nu = \nu(\varepsilon) > 0$ sufficiently large so that

$$\frac{4L}{\nu}\operatorname{vol}_n(B) < \frac{\varepsilon}{3}.$$

Since $\beta, \tau : K \to \mathbb{R}$ are uniformly continuous, we can find $\delta = \delta(\varepsilon) > 0$ such that, for any set $S \subset K$ with $\operatorname{diam}(S) < \delta$ we have

$$\operatorname{osc}(\beta, S) + \operatorname{osc}(\tau, C) < \frac{\varepsilon}{3\operatorname{vol}_n(B)}.$$

Now choose a partition $\boldsymbol{P}^\varepsilon$ of $\boldsymbol{P}$ such that $\boldsymbol{P}^\varepsilon \succ \boldsymbol{Q}^\varepsilon$ and $\|\boldsymbol{P}^\varepsilon\| < \delta(\varepsilon)$. We deduce from (15.2.1) that for $\nu > \nu(\varepsilon)$ we have

$$\omega\big(\,I_D, \widehat{\boldsymbol{P}^\varepsilon_\nu}\,\big) < \varepsilon.$$

□

Theorem 15.2.3 (Fubini)**.** *Let $n \in \mathbb{N}$ and suppose $D = D(K, \beta, \tau)$ is a simple type domain with cross section K and bottom/top functions $\beta, \tau : K \to \mathbb{R}$. We denote by $(\boldsymbol{x}, y)$ the coordinates in $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$. If $f : D \to \mathbb{R}$ is continuous, then f is integrable over D, the marginal function*

$$M_f : K \to \mathbb{R}, \;\; M_f(\boldsymbol{x}) = \int_{\beta(\boldsymbol{x})}^{\tau(\boldsymbol{x})} f(\boldsymbol{x}, y)|dy|$$

is Riemann integrable and

$$\int_D f(\boldsymbol{x}, y)|d\boldsymbol{x}dy| = \int_K M_f(\boldsymbol{x})|d\boldsymbol{x}| = \int_K \left(\int_{\beta(\boldsymbol{x})}^{\tau(\boldsymbol{x})} f(\boldsymbol{x}, y)\,|dy| \right) |d\boldsymbol{x}|. \tag{15.2.2}$$

Proof. According to Proposition 15.2.2 the region $D \subset \mathbb{R}^{n+1}$ is compact and Jordan measurable. Proposition 15.1.36 now implies that f is Riemann integrable on D.

Fix a box in $B \subset \mathbb{R}^n$ and a compact interval $I = [m, M] \subset \mathbb{R}$ such that

$$\beta(\boldsymbol{x}), \;\; \tau(\boldsymbol{x}) \in I, \;\; \forall \boldsymbol{x} \in K.$$

Then the box $B' = B \times [m, M] \subset \mathbb{R}^{n+1}$ contains D and the extension f^0 is Riemann integrable on B'. Next observe that for any $\boldsymbol{x} \in B$ the function

$$f^0_{\boldsymbol{x}} : I \to \mathbb{R}, \;\; f^0_{\boldsymbol{x}}(y) = f(\boldsymbol{x}, y)$$

is Riemann integrable. Indeed, if $\boldsymbol{x} \in B \backslash K$, then $f^0_{\boldsymbol{x}}$ is identically 0. On the other hand, if $\boldsymbol{x} \in K$, then

$$f^0_{\boldsymbol{x}}(y) = \begin{cases} f(\boldsymbol{x}, y), & y \in [\beta(\boldsymbol{x}), \tau(\boldsymbol{x})], \\ 0, & y \in [m, \beta(\boldsymbol{x})) \cup (\tau(\boldsymbol{x}), M]. \end{cases}$$

The continuity of f coupled with Corollary 9.4.7 imply the Riemann integrability of $f^0_{\boldsymbol{x}}$. We can now apply the conditional Fubini Theorem 15.1.18 to the function $f^0 : B' \to \mathbb{R}$. Note that the marginal function $M^1_{f^0}$ is Riemann integrable on B and coincides with the extension by 0 of the marginal function $M_f : K \to \mathbb{R}$, i.e., $M^1_{f^0} = (M_f)^0$. Thus M_f is Riemann integrable on K. We deduce

$$\int_{B'} f(\boldsymbol{x}, y)|d\boldsymbol{x}dy| = \int_{B \times I} f^0(\boldsymbol{x}, y)|d\boldsymbol{x}dy|$$

$$= \int_B M^1_{f^0}(\boldsymbol{x})|d\boldsymbol{x}| = \int_B (M^1_f)^0(\boldsymbol{x})|d\boldsymbol{x}| = \int_K M^1_f(\boldsymbol{x})|d\boldsymbol{x}|.$$

□

Remark 15.2.4. (a) The last integral in (15.2.2) is an example of *iterated* or *repeated* integral.

(b) In the definition of simple type domains in $\mathbb{R}^{n+1}$ the last coordinate x^{n+1} plays a distinguished role. We did this only to simplify the notation in the various proofs. The results we proved above hold for regions $D \subset \mathbb{R}^{n+1}$ of the type

$$D = \{ (x^1, \dots, x^{n+1}) \in \mathbb{R}^{n+1};\ \ \boldsymbol{x} \in K,\ \ \beta(\boldsymbol{x}) \leqslant x^j \leqslant \tau(\boldsymbol{x}) \}$$

where

$$\boldsymbol{x} := (x^1, \dots, x^{j-1}, x^{j+1}, \dots, x^{n+1}) \in \mathbb{R}^n.$$

We will refer to domains of this type as *domains of simple type with respect to the* x^j*-axis.* We want to point out that a given domain could be of simple type with respect to many axes.

Theorem 15.2.3 has an obvious extension to domains that are simple type with respect to an arbitrary axis x^j: replace y by x^j everywhere in the statement and the proof of this theorem. □

15.2.2. Some applications. Theorem 15.2.3 is best understood by witnessing it at work.

Example 15.2.5. Consider the triangle T in Figure 15.5. Its vertices have coordinates $(0,0), (1,1), (0,2)$. It is limited by the y-axis and the lines $y = x$, $y = 2 - x$. This triangle is a domain of simple type with respect to both x and y-axis.

Viewed as a simple type domain with respect to the y-axis it has description

$$T = \{ (x,y) \in \mathbb{R}^2;\ \ x \in [0,1],\ \ \beta(x) \leqslant y \leqslant \tau(x) \},$$

where $\beta(x) = x$ and $\tau(x) = 2 - x$.

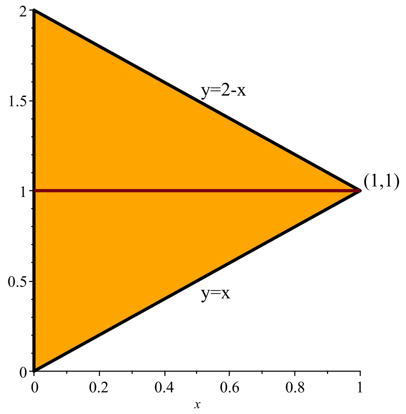

Figure 15.5. *An isosceles triangle in the plane.*

Viewed as a simple type domain with respect to the x-axis it has description

$$T = \{ (x,y) \in \mathbb{R}^2;\ \ y \in [0,2],\ \ \bar{\beta}(y) \leqslant x \leqslant \bar{\tau}(y) \},$$

where $\bar{\beta}(y) = 0$ and

$$\bar{\tau}(y) = \begin{cases} y, & y \in [0,1], \\ 2 - y, & y \in (1,2]. \end{cases}$$

Consider a continuous function $f : T \to \mathbb{R}$. Using Fubini's theorem we deduce

$$\int_0^1 \left(\int_x^{2-x} f(x,y)|dy| \right) |dx| = \int_T f(x,y)|dxdy|$$
$$= \int_0^1 \left(\int_0^y f(x,y)|dx| \right) |dy| + \int_1^2 \left(\int_0^{2-y} f(x,y)\,|dx| \right) |dy|. \qquad \square$$

Our next application is another version of *Cavalieri's Principle.*

Proposition 15.2.6. *Let $n \in \mathbb{N}$. Suppose that $K \subset \mathbb{R}^n$ is a compact Jordan measurable set and $\beta, \tau : K \to \mathbb{R}$ continuous functions such that*

$$\beta(\boldsymbol{x}) \leqslant \tau(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in K.$$

Then the $(n+1)$-dimensional volume of the region $D(K,\beta,\tau) \subset \mathbb{R}^{n+1}$ between the graphs of β and τ is

$$\operatorname{vol}_{n+1}\big(D(K,\beta,\tau) \big) = \int_K \big(\tau(\boldsymbol{x}) - \beta(\boldsymbol{x}) \big) |d\boldsymbol{x}|. \tag{15.2.3}$$

In particular, for any continuous function $h : K \to \mathbb{R}$, the $(n+1)$-dimensional volume of the graph Γ_h of h is 0

$$\operatorname{vol}_{n+1}(\Gamma_h) = 0. \tag{15.2.4}$$

Proof. The equality (15.2.3) is the special case of (15.2.2) corresponding to $f = 1$. The equality (15.2.4) follows from (15.2.3) by choosing $\beta = \tau = h$. □

Example 15.2.7. For every $n \in \mathbb{N}$ we denote by $\boldsymbol{T}_n$ the n-dimensional *simplex*[5] defined by the conditions

$$\boldsymbol{T}_n := \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ x^i \geqslant 0,\ \ \forall i = 1,\dots,n,\ x^1 + \cdots + x^n \leqslant 1 \big\}.$$

The region $\boldsymbol{T}_n$ can be alternatively described as the region

$$\boldsymbol{T}_n := \big\{ (\boldsymbol{x}_*, x^n) \in \mathbb{R}^n;\ \ \boldsymbol{x}_* \in \boldsymbol{T}_{n-1},\ \ 0 \leqslant x^n \leqslant 1 - (x^1 + \cdots + x^{n-1}) \big\},$$

where $\boldsymbol{x}_* := (x^1, \dots, x^{n-1})$.

Observe that $\boldsymbol{T}_1 = [0,1]$, that $\boldsymbol{T}_2$ is a domain of simple type in $\mathbb{R}^2$ with respect to the x^2 axis, with bottom 0 and top $1 - x^1$ and cross section $\boldsymbol{T}_1$ and thus $\boldsymbol{T}_2$ is compact and Jordan measurable. We deduce inductively that $\boldsymbol{T}_n$ is simple type with respect to the x^n-axis, with bottom 0, top $1 - (x^1 + \cdots + x^{n-1})$ and cross section $\boldsymbol{T}_{n-1}$ and thus $\boldsymbol{T}_n$ is compact and Jordan measurable. We want to compute its volume.

To keep the insanity at bay we introduce the simplifying notation

$$s_k = s_k(x^1, \dots, x^k) := x^1 + \cdots + x^k.$$

Note that $s_k = s_{k-1} + x^k$ and

$$\boldsymbol{T}_k = \big\{ (x^1, \dots, x^k) \in \mathbb{R}^k;\ \ (x^1, \dots, x^{k-1}) \in \boldsymbol{T}_{k-1},\ 0 \leqslant x^k \leqslant 1 - s_{k-1} \big\}.$$

[5] The concept of simplex is the higher dimensional generalization of the more familiar concepts of triangles and tetrahedra.

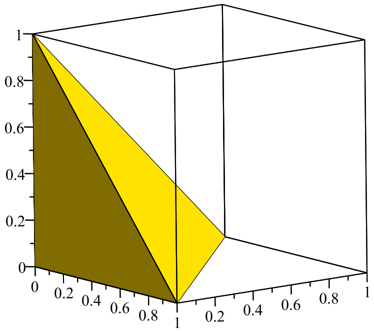

Figure 15.6. *The tetrahedron* $\boldsymbol{T}_3$.

For any $k \in \mathbb{N}$ and $s \in \mathbb{R}$ we set

$$I_k(s) := \int_0^{1-s} (1-s-x)^k dx.$$

Making the change in variables $u := 1-s-x$ we deduce

$$I_k(s) = -\int_{1-s}^0 u^k du = \frac{1}{k}(1-s)^k. \tag{15.2.5}$$

Using Fubini's Theorem (15.2.2) and the equality (15.2.5) we deduce

$$\operatorname{vol}_n(\boldsymbol{T}_n) \overset{(15.2.2)}{=} \int_{\boldsymbol{T}_{n-1}} \big(1-s_{n-1}\big)|dx^1\cdots dx^{n-1}|$$

$$\overset{(15.2.2)}{=} \int_{\boldsymbol{T}_{n-2}} \left(\int_0^{1-s_{n-2}} \big(1-\big(s_{n-2}+x^{n-1}\big)\big)dx^{n-1}\right)|dx^1\cdots dx^{n-2}|$$

$$= \int_{\boldsymbol{T}_{n-2}} \underbrace{\left(\int_0^{1-s_{n-2}} \big((1-\ s_{n-2}) - x^{n-1}\big)dx^{n-1}\right)}_{I_1(s_{n-2})}|dx^1\cdots dx^{n-2}|$$

$$\overset{(15.2.5)}{=} \frac{1}{2}\int_{\boldsymbol{T}_{n-2}} (1-s_{n-2})^2|dx^1\cdots dx^{n-2}|$$

$$\overset{(15.2.2)}{=} \frac{1}{2}\int_{\boldsymbol{T}_{n-3}} \underbrace{\left(\int_0^{1-s_{n-3}} \Big((1-s_{n-3})-x_{n-2}\Big)^2 dx^{n-2}\right)}_{I_2(s_{n-3})}|dx^1\cdots dx^{n-3}|$$

$$\overset{(15.2.5)}{=} \frac{1}{2\cdot 3}\int_{\boldsymbol{T}_{n-3}} \big(1-s_{n-3}\big)^3|dx^1\cdots dx^{n-3}|.$$

Continuing in this fashion we deduce

$$\operatorname{vol}_n(\boldsymbol{T}_n) = \frac{1}{k!}\int_{\boldsymbol{T}_{n-k}} \bigl(1 - s_{n-k}\bigr)^k |dx^1 \cdots dx^{n-k}|, \;\; \forall k = 1, \dots, n.$$

In particular, if we let $k = n - 1$, we deduce

$$\operatorname{vol}_n(\boldsymbol{T}_n) = \frac{1}{(n-1)!}\int_{\boldsymbol{T}_1} \bigl(1 - s_1\bigr)^{n-1} |dx^1| = \frac{1}{(n-1)!}\int_0^1 \bigl(1 - x^1\bigr)^{n-1} dx^1 = \frac{1}{n!}. \qquad \square$$

15.3. Change in variables formula

In the last section of this chapter we discuss a fundamental result in the theory of integration of functions of several variables. The importance of change-in-variables formula goes beyond its applications to the computation of many concrete integrals. It will serve as a guiding principle when defining the integral over "curved" spaces, i.e., submanifolds.

15.3.1. Formulation and some classical examples. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is open and $\Phi : U \to \mathbb{R}^n$ is a C^1-diffeomorphism

$$U \ni \boldsymbol{x} \mapsto \boldsymbol{y} = (y^1, \dots, y^n) = \bigl(\Phi^1(\boldsymbol{x}), \dots, \Phi^n(\boldsymbol{x})\bigr)$$

We denote by $J_\Phi(\boldsymbol{x})$ the Jacobian matrix of Φ at the point $\boldsymbol{x} \in U$. We set $V := \Phi(U)$ so V is an open subset of the target space $\mathbb{R}^n$.

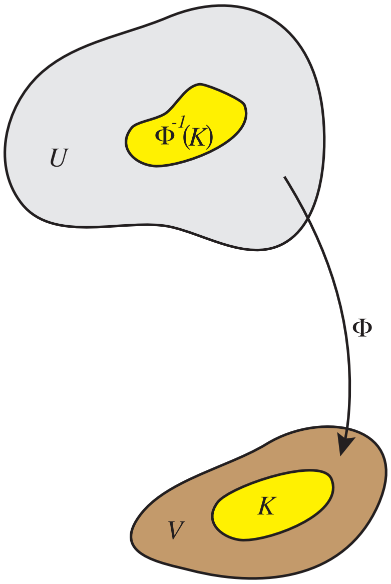

Figure 15.7. *A diffeomorphism $\Phi : U \to V$.*

Theorem 15.3.1. *Suppose that $f : V \to \mathbb{R}$ is a bounded function that vanishes outside a compact subset $K \subset V$. Then the following hold.*

(i) *The function f is integrable if and only if the function*

$$U \ni \boldsymbol{x} \mapsto f\big(\,\Phi(\boldsymbol{x})\,\big)|\det J_\Phi(\boldsymbol{x})| \in \mathbb{R}$$

is Riemann integrable.

(ii) *We have the* change in variables formula *(see 15.7)*

$$\boxed{\int_V f(\boldsymbol{y})|d\boldsymbol{y}| = \int_U f\big(\,\Phi(\boldsymbol{x})\,\big)\,\big|\det J_\Phi(\boldsymbol{x})\big|\,|d\boldsymbol{x}|}. \tag{15.3.1}$$

□

Remark 15.3.2. (a) We say that Φ changes the "*old*" variables (or coordinates) $y^1, \dots, y^n$ on V to the "*new*" variables (or coordinates) $x^1, \dots, x^n$ on U. The change is described by the equations

$$y^k = \Phi^k(x^1, \dots, x^n),\quad k = 1, \dots, n.$$

Thus the "*old*" variables $\boldsymbol{y}$ are expressed as functions of the "*new*" variables $\boldsymbol{x}$. We often use the slightly ambiguous but more suggestive notation

$$\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})\big(\,=\Phi(\boldsymbol{x})\,\big),$$

to express the dependence of the "old" coordinates $\boldsymbol{y}$ on the "new" coordinates $\boldsymbol{x}$. The inverse transformation $\Phi^{-1}(\boldsymbol{y})$ is often replaced by the simpler and more intuitive notation

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})\big(\,=\Phi^{-1}(\boldsymbol{y})\,\big),$$

indicating that $\boldsymbol{x}$ depends on $\boldsymbol{y}$ via the unspecified transformation Φ^{-1}.

Frequently we will use the more intuitive notation

$$\left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right| := |\det J_\Phi|.$$

In concrete applications we are given a compact Jordan measurable subset $K \subset V$ and a Riemann integrable function $f : K \to \mathbb{R}$. The change of variables formula applied to the function $I_K(\boldsymbol{y})f(\boldsymbol{y})$ can then be rewritten in the more intuitive form (Figure 15.3.1)

$$\boxed{\int_K f(\boldsymbol{y})|d\boldsymbol{y}| = \int_{\Phi^{-1}(K)} f\big(\,\boldsymbol{y}(\boldsymbol{x})\,\big)\left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right||d\boldsymbol{x}|}. \tag{15.3.2}$$

Implicit in the above equality is the conclusion that the compact $\Phi^{-1}(K)$ is also Jordan measurable.

(b) Let us observe that if in (15.3.1) we set $g(\boldsymbol{x}) := f\big(\Phi(x)\big)$, then we can rewrite this equality in the form

$$\boxed{\int_V g\big(\,\boldsymbol{x}(\boldsymbol{y})\,\big)|d\boldsymbol{y}| = \int_U g(\boldsymbol{x})\left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right||d\boldsymbol{x}|}. \tag{15.3.3}$$

Clearly (15.3.3) is equivalent to (15.3.1). □

We will present an outline of the proof in the next subsection. The best way of understanding how it works is through concrete examples.

Example 15.3.3 (The volume of a parallelepiped)**.** Let $n \in \mathbb{N}$. Suppose that we are given n *linearly independent* vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^n$.

The *parallelepiped* spanned by $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is the set $P(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \subset \mathbb{R}^n$ consisting of all the vectors $\boldsymbol{y}$ of the form

$$\boldsymbol{y} = x^1\boldsymbol{v}_1 + \cdots + x^n\boldsymbol{v}_n,\ \ x^1, \dots, x^n \in [0,1]. \tag{15.3.4}$$

We want to prove that $P(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is Jordan measurable and then compute its volume. To this end consider the linear map $V : \mathbb{R}^n \to \mathbb{R}^n$ uniquely determined by the requirements

$$V\boldsymbol{e}_j = \boldsymbol{v}_j,\ \ \forall j = 1, \dots, n.$$

In other words, the columns of the matrix representing V are given by the (column) vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$. Since the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ are linearly independent the operator V is invertible and thus defines a diffeomorphism $\mathbb{R}^n \to \mathbb{R}^n$. Being linear, the operator V coincides with its differential at every $\boldsymbol{x} \in \mathbb{R}^n$ so

$$J_V = V\ \ \det J_V = \det V.$$

We denote by C the cube $C = [0,1]^n \subset \mathbb{R}^n$. The equation (15.3.4) can be written in the form

$$\boldsymbol{y} \in P(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \Longleftrightarrow \boldsymbol{y} = x^1V\boldsymbol{e}_1 + \cdots + x^nV\boldsymbol{e}_n,\ \ \boldsymbol{x} = (x^1, \dots, x^n) \in [0,1]^n = C.$$

In other words, $P = V(C)$. In particular, this shows that P is compact. The equality $P = V(C)$ translates into an equality of indicator functions

$$I_P(V\boldsymbol{x}) = I_C(\boldsymbol{x}).$$

Theorem 15.3.1 implies that P is Jordan measurable and

$$\boxed{\operatorname{vol}_n\big(\,P(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)\,\big) = \int_{\mathbb{R}^n} I_P(\boldsymbol{y})|d\boldsymbol{y}| = \int_{\mathbb{R}^n} I_C(\boldsymbol{x})|\det V||d\boldsymbol{x}| = |\det V|}. \tag{15.3.5}$$

When $n = 2$, and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$ are not collinear, then $P(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is the parallelogram spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. For example, if

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix},\ \ \boldsymbol{v}_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

then

$$V = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix},\ \ \det V = 1 \cdot 4 - 2 \cdot 3 = -2,\ \ \operatorname{area}\big(\,P(\boldsymbol{v}_1, \boldsymbol{v}_2)\,\big) = |\det V| = 2. \qquad \square$$

Example 15.3.4 (Polar coordinates)**.** Consider the map

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^2, (r, \theta) \mapsto (x, y) = (r\cos\theta, r\sin\theta).$$

The geometric significance of this map was explained in Example 14.5.12 and can be seen in Figure 15.8.

The location of a point $\boldsymbol{p} \in \mathbb{R}^2 \backslash \{\mathbf{0}\}$ can be indicated using the "old" *Cartesian coordinates*, or the "new" *polar coordinates* (r, θ), where $r = \operatorname{dist}(\boldsymbol{p}, \mathbf{0})$ and θ is the angle between the line segment $[\mathbf{0}, \boldsymbol{p}]$ and the x-axis, measured counterclockwisely.

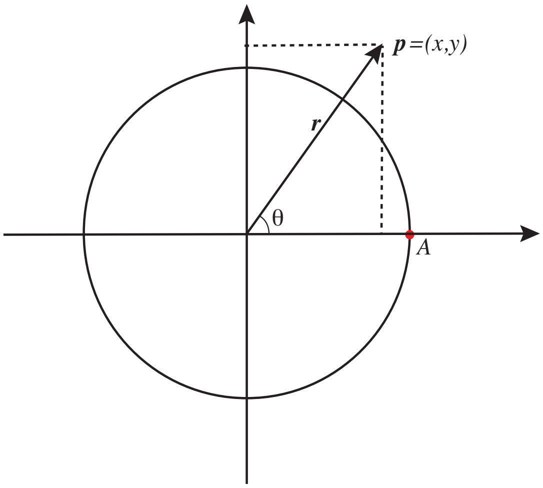

Figure 15.8. *Constructing the polar coordinates.*

The Jacobian of this map is

$$J_\Phi = \left[\begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ & \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{array}\right] = \left[\begin{array}{cc} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{array}\right]$$

so

$$\det J_\Phi = r. \tag{15.3.6}$$

Let us observe that, for any $T \in \mathbb{R}$, the restriction of Φ to the region $(0, \infty) \times (T, T + 2\pi)$ produces a bijection onto $\mathbb{R}^2_* :=$ the plane $\mathbb{R}^2$ with the nonnegative x-semiaxis removed. We will work exclusively with the restriction

$$\Phi_{\text{polar}} := \Phi\big|_{[0,\infty)\times[0,2\pi]}.$$

Note two "problems" with Φ_{polar}.

- The domain $[0, \infty) \times [0, 2\pi]$ is *not* an open subset of $\mathbb{R}^2$.
- The map Φ_{polar} is *not* injective, but its restriction to the open subset $(0, \infty) \times (0, 2\pi)$ is injective.

For every Jordan measurable compact set $K \subset \mathbb{R}^2$ we set

$$K_{\text{polar}} := \Phi^{-1}_{\text{polar}}(K) = \big\{ (r, \theta) \in [0, \infty) \times [0, 2\pi];\ \ (r\cos\theta, r\sin\theta) \in K \big\}. \tag{15.3.7}$$

Observe that K_{polar} is compact. If K *does not* intersect the nonnegative x-semiaxis, then Theorem 15.3.1 applies directly to this situation and shows that if $f = f(x,y) : K \to \mathbb{R}$ is Riemann integrable, then so is the function

$$f\circ\Phi_{\text{polar}} : K_{\text{polar}} \to \mathbb{R},\;\; f\circ\Phi_{\text{polar}}(r,\theta) = f(r\cos\theta, r\sin\theta),$$

and we have

$$\boxed{\int_{K_{\text{polar}}} f(r\cos\theta, r\sin\theta)r|drd\theta| = \int_K f(x,y)|dxdy|}. \tag{15.3.8}$$

When K does intersect this semi-axis Theorem 15.3.1 does not apply directly because of the above two "problems". However the equality (15.3.8) continues to hold even in this case. This requires a separate argument.

Since we will be frequently confronted with such problems, we state below a general result that deals with these situations. We refer the curious reader to [**29**, Thm. XX.4.7] or [**45**, Sec. 11.5.7, Thm.2] for a proof of this result.

Theorem 15.3.5. *Let $n \in \mathbb{N}$. We are given a compact Jordan measurable set $K \subset \mathbb{R}^n$, an open set $U \supset K$ and a C^1 map $\Phi : U \to \mathbb{R}^n$. Suppose that the restriction of Φ to the interior of K is a diffeomorphism. If $f : \Phi(K) \to \mathbb{R}$, is Riemann integrable, then the function*

$$K \ni \boldsymbol{x} \mapsto f\big(\Phi(\boldsymbol{x})\big)|\det J_\Phi(\boldsymbol{x})| \in \mathbb{R}$$

is Riemann integrable and

$$\boxed{\int_{\Phi(K)} f(\boldsymbol{y})|d\boldsymbol{y}| = \int_K f\big(\Phi(\boldsymbol{x})\big)\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{x}|}. \tag{15.3.9}$$

□

Here is an immediate application of the above result. Suppose that $K = K_R$ is the rectangle (see Figure 15.9)

$$K = K_R := \big\{ (r,\theta) \in \mathbb{R}^2;\;\; r \in [0,R],\; \theta \in [0,2\pi] \big\}$$

We have a differentiable map $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$,

$$\Phi(r,\theta) = (r\cos\theta, r\sin\theta).$$

and $\Phi(K) = D = D_R$ is the closed disk (in $\mathbb{R}^2$) of radius R centered at the origin,

$$D_R = \big\{ (x,y) \in \mathbb{R}^2;\;\; x^2+y^2 \leqslant R^2 \big\}.$$

Using the terminology in (15.3.7) we have $K = D_{\text{polar}}$. The boundary S of K is depicted in red on Figure 15.9. The interior of K is $K\backslash S$. The induced map $\Phi : K\backslash S \to \mathbb{R}^2$ is a diffeomorphism with image the complement of the set Z also depicted in red on Figure 15.9. Theorem 15.3.5 implies that for any Riemann integrable function $f : D \to \mathbb{R}$, the function

$$K \ni (r,\theta) \mapsto f(r\cos\theta, r\sin\theta) \in \mathbb{R}$$

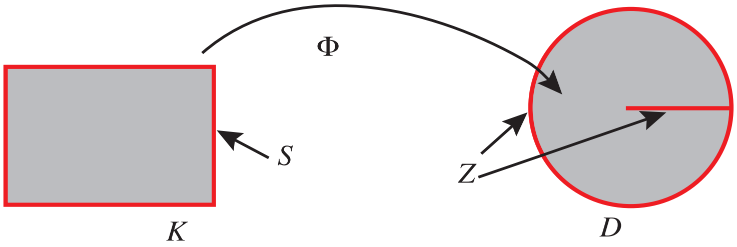

Figure 15.9. *The polar coordinate change transformation sends a rectangle U to a disk V.*

is Riemann integrable and we have

$$\int_D f(x,y)|dxdy| = \int_K f(r\cos\theta, \sin\theta)r|drd\theta|. \tag{15.3.10}$$

More generally, suppose that S is a compact, Jordan measurable subset of the (x,y)-plane. Since S is bounded, it is contained in some closed disk D_R of radius R, centered at the origin. Form the associated closed rectangle K_R in the (r,θ)-plane

$$K_R = \big\{ (r,\theta);\ \ r \in [0,R],\ \ \theta \in [0,2\pi] \big\}.$$

Suppose that $f : S \to \mathbb{R}$ is a Riemann integrable function. As usual, we denote by f^0 the extension by 0 of f and by I_{D_R} the indicator function of D_R. Note that $f^0 = f^0 I_{D_R}$. By definition

$$\int_S f|dxdy| = \int_{\mathbb{R}^2} I_{D_R} f^0|dxdy| = \int_{D_R} f^0|dxdy|.$$

Applying (15.3.10) to the function $f^0 : D_R \to \mathbb{R}$ we deduce

$$\boxed{\int_{S_{\text{polar}}} f(r\cos\theta, r\sin\theta)r|drd\theta| = \int_S f(x,y)|dxdy|}, \tag{15.3.11}$$

where S_{polar} is defined as in (15.3.7).

To see how (15.3.11) works in practice, consider the continuous function

$$f : \mathbb{R}^2\backslash\{\mathbf{0}\} \to \mathbb{R},\ \ f(x,y) = \log(x^2+y^2),$$

where log denotes the natural logarithm. We want to compute the integral of this function over the annulus

$$A := \big\{ \boldsymbol{p} \in \mathbb{R}^2;\ \ 1 \leqslant \operatorname{dist}(\boldsymbol{p},\mathbf{0}) \leqslant 2 \big\}.$$

Then

$$A_{\text{polar}} = \big\{ (r,\theta);\ \ r \in [1,2],\ \ \theta \in [0,2\pi] \big\}$$

and

$$\log(x^2+y^2) = \log(r^2) = 2\log r.$$

We deduce

$$\int_A \log(x^2+y^2)|dxdy| = \int_{\substack{1\leqslant r\leqslant 2,\\ 0\leqslant\theta\leqslant 2\pi}} 2(\log r)r|drd\theta|$$

(use Fubini)

$$= 2\int_1^2\left(\int_0^{2\pi}|d\theta|\right)r\log r|dr| = 2\pi\int_1^2 2r\log r dr.$$

We have

$$\int_1^2 2r\log r dr = \int_1^2 \log r d(r^2) = (r^2\log r)\Big|_{r=1}^{r=2} - \int_1^2 r^2 d(\log r)$$

$$= 4\log 2 - \int_1^2 r dr = 4\log 2 - \frac{1}{2}\left(r^2\Big|_{r=1}^{r=2}\right) = 4\log 2 - \frac{3}{2}.$$

Thus

$$\int_A \log(x^2+y^2)dxdy = \pi\big(8\log 2 - 3\big).$$ □

Example 15.3.6 (Cylindrical and spherical coordinates)**.** Denote by $\mathcal{O}$ the open subset of $\mathbb{R}^3$ obtained by removing the half-plane H contained in the xz-plane defined by

$$H := \big\{(x,y,z)\in\mathbb{R}^3;\ \ y=0,\ \ x\geqslant 0\big\}. \tag{15.3.12}$$

The location of a point $\boldsymbol{p}\in\mathcal{O}$ is determined either by its Cartesian coordinates (x,y,z), or by its altitude and the location of its projection $\boldsymbol{q}$ on the xy-plane. In turn, this projection is determined by its polar coordinates (r,θ); see Figure 15.10.

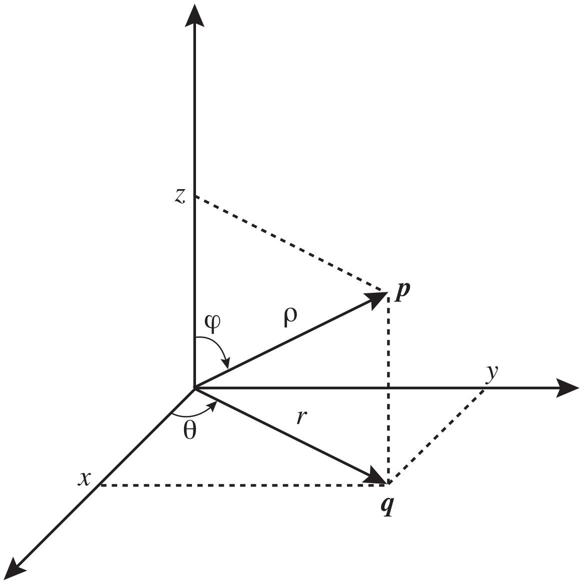

Figure 15.10. *Constructing the cylindrical and spherical coordinates.*

The *cylindrical coordinates* (r,θ,z) are related to the Cartesian coordinates via the equalities

$$\left\{\begin{array}{rcl} x &=& r\cos\theta\\ y &=& r\sin\theta \qquad r>0,\ \ \theta\in(0,2\pi),\ \ z\in\mathbb{R}.\\ z &=& z, \end{array}\right.$$

The above equalities define a transformation

$$\Phi_{\mathrm{Cart}\leftarrow\mathrm{cyl}} : [0,\infty) \times [0,2\pi] \times \mathbb{R} \to \mathbb{R}^3,\ \ (r,\theta,z) \mapsto \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ z \end{bmatrix},$$

whose restriction to the open set $(0,\infty) \times (0,2\pi) \times \mathbb{R}$ is a diffeomorphism with image $\mathcal{O} = \mathbb{R}^3 \backslash H$, where H is the half-plane in (15.3.12). The Jacobian matrix of the transformation $\Phi_{\mathrm{Cart}\leftarrow\mathrm{cyl}}$ is

$$J_{\mathrm{Cart}\leftarrow\mathrm{cyl}} := \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We have

$$\boxed{\det J_{\mathrm{Cart}\leftarrow\mathrm{cyl}} = \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r}. \tag{15.3.13}$$

Alternatively, the location of a point $\boldsymbol{p} \in \mathcal{O}$ is determined if we know the distance ρ to the origin $\rho = \|\boldsymbol{p}\|$, the angle $\varphi \in (0,\pi)$ the line $\boldsymbol{0p}$ makes with the z-axis, and the polar coordinate θ of the projection of $\boldsymbol{p}$ on the xy-plane; see Figure 15.10. The parameters ρ, φ, θ are called the *spherical coordinates* of $\boldsymbol{p}$. The spherical coordinates (ρ, φ, θ) determine the cylindrical coordinates (r, θ, z) via the equalities

$$r = \rho\sin\varphi,\ \ \theta = \theta,\ \ z = \rho\cos\varphi.$$

The Jacobian matrix of the transformation $\Phi_{\mathrm{cyl}\leftarrow\mathrm{sph}}(\rho,\varphi,\theta) = (r,\theta,z)$ is

$$J_{\mathrm{cyl}\leftarrow\mathrm{sph}} = \begin{bmatrix} \frac{\partial r}{\partial\rho} & \frac{\partial r}{\partial\varphi} & \frac{\partial r}{\partial\theta} \\ \\ \frac{\partial \theta}{\partial\rho} & \frac{\partial \theta}{\partial\varphi} & \frac{\partial \theta}{\partial\theta} \\ \\ \frac{\partial z}{\partial\rho} & \frac{\partial z}{\partial\varphi} & \frac{\partial z}{\partial\theta} \end{bmatrix} = \begin{bmatrix} \sin\varphi & \rho\cos\varphi & 0 \\ 0 & 0 & 1 \\ \cos\varphi & -\rho\sin\varphi & 0 \end{bmatrix}.$$

Expanding along the second row we deduce

$$\det J_{cyl\leftarrow sph} = -\det \begin{bmatrix} \sin\varphi & \rho\cos\varphi \\ \cos\varphi & -\rho\sin\varphi \end{bmatrix} = \rho. \tag{15.3.14}$$

We deduce that the Cartesian coordinates (x, y, z) are related to the spherical coordinates (ρ, φ, θ) via the equalities

$$\begin{cases} x &= & \rho\sin\varphi\cos\theta \\ y &= & \rho\sin\varphi\sin\theta \\ z &= & \rho\cos\varphi, \end{cases},\ \ \rho > 0,\ \ \theta \in (0,2\pi),\ \ \varphi \in (0,\pi).$$

The above equations define a transformation

$$\Phi_{\mathrm{Cart}\leftarrow\mathrm{cyl}} : [0,\infty) \times [0,2\pi) \times \mathbb{R} \to \mathbb{R}^3,\ \ (\rho,\varphi,\theta) \mapsto \begin{bmatrix} x \\ y \\ x \end{bmatrix} = \begin{bmatrix} \rho\sin\varphi\cos\theta \\ \rho\sin\varphi\sin\theta \\ \rho\cos\varphi \end{bmatrix}.$$

The restriction of $\Phi_{\text{Cart}\leftarrow\text{sph}}$ to the open set $(0,\infty)\times(0,\pi)\times(0,2\pi)$ is a diffeomorphism with image $\mathcal{O}$. The Jacobian matrix of the above transformation is

$$J_{\text{Cart}\leftarrow\text{sph}} = \begin{bmatrix} \frac{\partial x}{\partial\rho} & \frac{\partial x}{\partial\varphi} & \frac{\partial x}{\partial\theta} \\ \\ \frac{\partial y}{\partial\rho} & \frac{\partial y}{\partial\varphi} & \frac{\partial y}{\partial\theta} \\ \\ \frac{\partial z}{\partial\rho} & \frac{\partial z}{\partial\varphi} & \frac{\partial z}{\partial\theta} \end{bmatrix} = \begin{bmatrix} \sin\varphi\cos\theta & \rho\cos\varphi\cos\theta & -\rho\sin\varphi\sin\theta \\ \\ \sin\varphi\sin\theta & \rho\cos\varphi\sin\theta & \rho\sin\varphi\cos\theta \\ \\ \cos\varphi & -\rho\sin\varphi & 0 \end{bmatrix}.$$

Since $\Phi_{\text{Cart}\leftarrow\text{sph}} = \Phi_{\text{Cart}\leftarrow\text{cyl}}\circ\Phi_{\text{cyl}\leftarrow\text{sph}}$ we deduce from the chain rule that

$$J_{\text{Cart}\leftarrow\text{sph}} = J_{\text{Cart}\leftarrow\text{cyl}}\cdot J_{\text{cyl}\leftarrow\text{sph}}.$$

Hence $\det J_{\text{Cart}\leftarrow\text{sph}} = \det J_{\text{Cart}\leftarrow\text{cyl}}\det J_{\text{cyl}\leftarrow\text{sph}}$. Using (15.3.13) and (15.3.14) we deduce

$$\boxed{\det J_{\text{Cart}\leftarrow\text{sph}} = r\rho = \rho^2\sin\varphi}. \tag{15.3.15}$$

Let us see how we can use these facts in concrete situations. Suppose that $K\subset\mathbb{R}^3$ is a compact measurable set contained in some large ball $\overline{B_R(\mathbf{0})}\subset\mathbb{R}^3$. We set

$$K_{\text{cyl}} := \bigl\{\,(r,\theta,z)\in[0,\infty)\times[0,2\pi]\times\mathbb{R};\;\;\Phi_{\text{Cart}\leftarrow\text{cyl}}(r,\theta,z)\in K\,\bigr\},$$
$$K_{\text{sph}} := \bigl\{\,(r,\varphi,z)\in[0,\pi]\times[0,2\pi]\times\mathbb{R};\;\;\Phi_{\text{Cart}\leftarrow\text{sph}}(r,\varphi,\theta)\in K\,\bigr\}.$$

Suppose that $f:K\to\mathbb{R}$ is a Riemann integrable function. If K does not intersect the "forbidden" half-plane H, then Theorem 15.3.1 applies and yields

$$\boxed{\int_K f(x,y,z)dxdydz = \int_{K_{\text{cyl}}} f(r\cos\theta, r\sin\theta, z)rdrd\theta dz}. \tag{15.3.16a}$$

$$\boxed{\int_K f(x,y,z)\,|dxdydz| = \int_{K_{\text{sph}}} f(\rho\sin\varphi\cos\theta,\rho\sin\varphi\sin\theta,\rho\cos\varphi)\rho^2\sin\varphi\,|d\rho d\varphi d\theta|}. \tag{15.3.16b}$$

If K does intersect the "forbidden" half-plane H, then using Theorem 15.3.5 we deduce as in Example 15.3.4 that (15.3.16a) and (15.3.16b) continue to hold in this case as well. Let us see how the above formulæ work in concrete situations.

Fix a number $\varphi_0\in(0,\pi/2)$. The locus of points in $\mathbb{R}^3$ such that $\varphi=\varphi_0$ describes a circular cone; Fig 15.11. The z-axis is a symmetry axis of this cone.

If we set $m_0 := \tan\varphi_0$, then we observe that, along the surface of this cone the cylindrical coordinates coordinates r,z satisfy

$$m_0 = \tan\varphi_0 = \frac{r}{z}\quad r = m_0 z.$$

The "inside part" of this cone is described by the inequality

$$\frac{r}{z}\leqslant m_0 \Longleftrightarrow r\leqslant m_0 z.$$

Consider two positive numbers $z_0<z_1$ and denote by $R=R(m_0,z_0,z_1)$ the region inside this cone contained between the horizontal planes $z=z_0$ and $z=z_1$; see Figure 15.12.

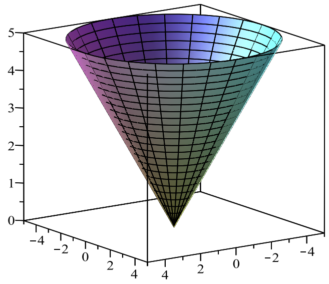

Figure 15.11. *A cone.*

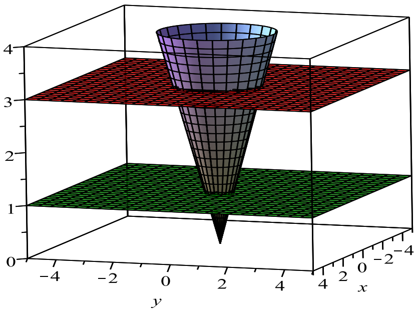

Figure 15.12. *A truncated cone.*

We want to compute the volume of this truncated cone. In cylindrical coordinates it corresponds to the region R_{cyl} described by the inequalities

$$0 \leqslant r \leqslant m_0 z, \quad z_0 \leqslant z \leqslant z_1, \quad 0 \leqslant \theta \leqslant 2\pi.$$

Using the change in variables formula (15.3.16a)

$$\begin{aligned}\mathrm{vol}_3(R) = \int_R |dxdydz| &= \int_{\substack{(\theta,z)\in[0,2\pi]\times[z_0,z_1]\\ 0\leqslant r\leqslant m_0 z}} r\,|drd\theta dz|\\ &= \int_0^{2\pi}\left(\int_{z_0}^{z_1}\left(\int_0^{m_0 z} rdr\right)dz\right)d\theta\end{aligned}$$

$$= \int_0^{2\pi}\left(\int_{z_0}^{z_1}\frac{1}{2}(m_0 z)^2 dz\right)d\theta = \frac{m_0^2}{2}\int_0^{2\pi}\left(\frac{z_1^3 - z_0^3}{3}d\theta\right) = \frac{\pi m_0^2}{3}\big(\, z_1^3 - z_0^3\,\big).$$

To see the spherical coordinates at work, it is useful to relate them to more familiar notions. Note that the surface $\rho = const$ is a sphere. The surface $\varphi = const$ is a cone while the surface $\theta = const$ is a half-plane with edge z-axis. Since $z = \rho\cos\varphi$ we deduce that in spherical coordinates the truncated cone R corresponds to the region R_{sph} described by the inequalities

$$0 \leqslant \varphi \leqslant \varphi_0, \ \ \theta \in [0, 2\pi], \ \ z_0 \leqslant \rho\cos\varphi \leqslant z_1.$$

We deduce

$$\mathrm{vol}_3(R) = \int_{R_{\mathrm{sph}}} \rho^2 \sin\varphi |d\rho d\varphi d\theta|$$

$$= \int_0^{2\pi}\left(\int_0^{\varphi_0}\left(\int_{\frac{z_0}{\cos\varphi}}^{\frac{z_1}{\cos\varphi}} \rho^2 d\rho\right)\sin\varphi d\varphi\right)d\theta = \frac{z_1^3 - z_0^3}{3}\int_0^{2\pi}\left(\int_0^{\varphi_0}\frac{\sin\varphi}{\cos^3\varphi}d\varphi\right)d\theta$$

(make the change in variables $u = \cos\varphi$)

$$= \frac{z_1^3 - z_0^3}{3}\int_0^{2\pi}\left(\int_{\cos\varphi_0}^1 \frac{1}{u^3}du\right)d\theta = \frac{z_1^3 - z_0^3}{6}\int_0^{2\pi}\underbrace{\left(\frac{1}{\cos^2\varphi_0} - 1\right)}_{=\tan^2\varphi_0 = m_0^2}d\theta$$

$$= \frac{\pi m_0^2(z_1^3 - z_0^3)}{3}.$$

This is in perfect agreement with the computation using cylindrical coordinates.

To verify the validity of these computations, we present an alternate computation based on Cavalieri's principle. The intersection of the region R with the horizontal plane $z = t$ is a disk R_t of radius $r_t = m_0 t$. We have

$$\mathrm{vol}_2(R_t) = \pi(m_0 t)^2.$$

Then

$$\mathrm{vol}_3(R) = \int_{z_0}^{z_1}\mathrm{vol}_2(R_t)dt = \pi m_0^2\int_{z_0}^{z_1} t^2 dt = \frac{\pi m_0^2}{3}\big(\, z_1^3 - z_0^3\,\big).$$

Let us rewrite this volume formula in a more familiar form. The bottom of the truncated cone R is a disk of radius $r_0 = m_0 z_0$ and the top is a disk of radius $r_1 = m_0 z_1$. We denote by h the height of this truncated cone $h := z_1 - z_0$. Then

$$\mathrm{vol}_3(R) = \frac{\pi}{3}(z_1 - z_0)\big(\,(m_0 z_1)^2 + (m_0 z_1)(m_0 z_0) + (m_0 z_0)^2\,\big) = \frac{\pi h}{3}\big(\, r_1^2 + r_1 r_0 + r_0^2\big).$$

□

Example 15.3.7 (Spherical coordinates in arbitrary dimensions)**.** Let $n \in \mathbb{N}$, $n \geqslant 2$. We will construct inductively spherical coordinates on $\mathbb{R}^n$.

For $n = 2$, these are versions of the polar coordinates (r, θ). Define (ρ_2, θ_2) via the well known equalities

$$x^1 = \rho_2\cos\theta_2, \ \ x^2 = \rho_2\sin\theta_2, \ \ \rho_2 = \|\boldsymbol{x}\|, \ \ \theta_2 \in [0, 2\pi]. \tag{15.3.17}$$

Observe that the numbers ρ_2 and θ_2 completely determine the location of the point $\boldsymbol{x}$ in $\mathbb{R}^2$.

Suppose now that we have constructed the spherical coordinates $\theta_2, \dots, \theta_n, \rho_n$ on $\mathbb{R}^n$. In particular, the coordinates $x^1, \dots, x^n$ are functions depending on these coordinates,

$$x^1 = f^1(\theta_2, \dots, \theta_n, \rho_n), \dots, x^n = f^n(\theta_2, \dots, \theta_n, \rho_n), \tag{15.3.18a}$$

$$\rho_n(\boldsymbol{x}) = \|\boldsymbol{x}\|, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{15.3.18b}$$

We will construct spherical coordinates $\theta_2, \dots, \theta_n, \theta_{n+1}, \rho_{n+1}$ on $\mathbb{R}^n$.

For $\boldsymbol{x} = (x^1, \dots, x^{n+1}) \in \mathbb{R}^{n+1}$ we denote by $\bar{\boldsymbol{x}}$ is projection onto the coordinate hyperplane $\mathbb{R}^n \times \{0\} = \{x^{n+1} = 0\}$ and we think of $\bar{\boldsymbol{x}}$ as a point in $\mathbb{R}^n$; see Figure 15.13. In other words, we have

$$\boldsymbol{x} = \big(\, \underbrace{x^1, \dots, x^n}_{\bar{\boldsymbol{x}}},\, x^{n+1}\big) = \big(\bar{\boldsymbol{x}}, x^{n+1}\big).$$

We denote by ρ_{n+1} the distance from $\boldsymbol{x}$ to the origin, i.e.,

$$\rho_{n+1} = \|\boldsymbol{x}\| = \sqrt{(x^1)^2 + (x^2)^2 + \cdots + (x^{n+1})^2}.$$

Observe that the location of $\boldsymbol{x}$ is completely determined once we know x^{n+1} and the location of $\bar{\boldsymbol{x}}$ in $\mathbb{R}^n$; see Figure 15.13. According to the induction assumption the location of $\bar{\boldsymbol{x}}$ is completely determined by the previously defined spherical coordinates $(\theta_2, \dots, \theta_n, \rho_n)$, $\rho_n = \|\bar{\boldsymbol{x}}\|$.

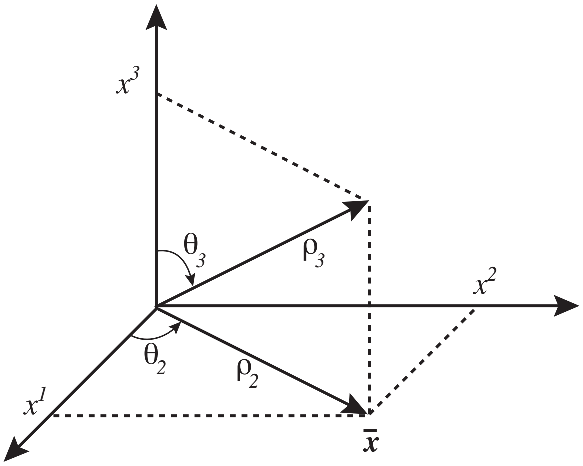

Figure 15.13. *Constructing spherical coordinates in n dimensions.*

For $\boldsymbol{x} \in \mathbb{R}^{n+1} \backslash \{\boldsymbol{0}\}$ denote by $\theta_{n+1} = \theta_{n+1}(\boldsymbol{x}) \in [0, \pi]$ the angle the vector $\boldsymbol{x}$ makes with the x^{n+1}-axis. More precisely, we have (see Figure 15.13)

$$x^{n+1} = \langle \boldsymbol{x}, \boldsymbol{e}_{n+1} \rangle = \|\boldsymbol{x}\| \cdot \|\boldsymbol{e}_{n+1}\| \cos\theta_{n+1} = \rho_{n+1} \cos\theta_{n+1},$$

and

$$\rho_n = \rho_{n+1} \sin\theta_{n+1}.$$

This shows that the quantities $(\rho_{n+1}, \theta_{n+1})$ determine the coordinate x^{n+1} and the spherical coordinate ρ_n of $\bar{\boldsymbol{x}}$. Thus, the quantities

$$\theta_2, \dots, \theta_n, \theta_{n+1}, \rho_{n+1}$$

completely determine the location of $\boldsymbol{x}$ in $\mathbb{R}^{n+1}$. More precisely, using (15.3.18a) we obtain equalities

$$\begin{array}{rcl} x^1 & = & f^1(\theta_2, \dots, \theta_n, \rho_{n+1}\sin\theta_{n+1}) \\ \vdots & \vdots & \vdots \\ x^n & = & f^n(\theta_2, \dots, \theta_n, \rho_{n+1}\sin\theta_{n+1}) \\ x^{n+1} & = & \rho_{n+1}\cos\theta_{n+1}, \end{array} \tag{15.3.19}$$

where

$$\boxed{\rho_{n+1} > 0, \;\; \theta_2 \in [0, 2\pi], \;\; \theta_3, \dots, \theta_{n+1} \in [0, \pi]}.$$

Let us see more explicitly the manner in which the spherical coordinates $\theta_2, \dots, \theta_{n+1}, \rho_{n+1}$ determined the Cartesian coordinates $x^1, \dots, x^{n+1}$. Using (15.3.17) we deduce that, for $n + 1 = 3$ we have

$$x^1 = \rho_3 \sin\theta_3 \cos\theta_2, \;\; x^2 = \rho_3 \sin\theta_3 \sin\theta_2, \;\; x^3 = \rho_3 \cos\theta_3.$$

We recognize here an old "friend", the spherical coordinates in $\mathbb{R}^3$, $\rho = \rho_3$, $\theta = \theta_2$, $\varphi = \theta_3$; see Figure 15.13. Using these freshly obtained equalities and the inductive scheme (15.3.19) we deduce that for $n + 1 = 4$ we have

$$\begin{aligned} x^1 &= \rho_4 \sin\theta_4 \sin\theta_3 \cos\theta_2, \\ x^2 &= \rho_4 \sin\theta_4 \sin\theta_3 \sin\theta_2, \\ x^3 &= \rho_4 \sin\theta_4 \cos\theta_3, \\ x^4 &= \rho_4 \cos\theta_4. \end{aligned}$$

The general pattern should be clear

$$\boxed{\begin{array}{rcl} x^1 & = & \rho_n \sin\theta_n \cdots \sin\theta_4 \sin\theta_3 \cos\theta_2, \\ x^2 & = & \rho_n \sin\theta_n \cdots \sin\theta_4 \sin\theta_3 \sin\theta_2, \\ x^3 & = & \rho_n \sin\theta_n \cdots \sin\theta_4 \cdot \cos\theta_3, \\ \vdots & \vdots & \vdots \\ x^{n-1} & = & \rho_n \sin\theta_n \cos\theta_{n-1}, \\ x^n & = & \rho_n \cos\theta_n. \end{array}} \tag{15.3.20}$$

We interpret the above equalities as defining a map $\Phi_n = \Phi_n(\theta_2, \cdots \theta_n, \rho_n)$ from an open subset of a vector space with coordinates $(\theta_2, \dots, \theta_n, \rho_n)$ to another vector space with coordinates $(x^1, \dots, x^n)$. We want to compute $\delta_n := \det J_{\Phi_n}$.

We will achieve this inductively by observing that we can write Φ_{n+1} as a composition

$$\begin{bmatrix}\theta_2\\ \vdots\\ \theta_{n+1}\\ \rho_{n+1}\end{bmatrix} \overset{\Psi_{n+1}}{\mapsto} \begin{bmatrix}\theta_2\\ \vdots\\ \theta_n\\ \rho_n\\ x^{n+1}\end{bmatrix} = \begin{bmatrix}\theta_2\\ \vdots\\ \theta_n\\ \rho_{n+1}\sin\theta_{n+1}\\ \rho_{n+1}\cos\theta_{n+1}\end{bmatrix},$$

$$\begin{bmatrix}\theta_2\\ \vdots\\ \theta_n\\ \rho_n\\ x^{n+1}\end{bmatrix} \overset{\hat{\Phi}_n}{\mapsto} \begin{bmatrix}\bar{\boldsymbol{x}}\\ x^{n+1}\end{bmatrix} = \begin{bmatrix}\Phi_n(\theta_2,\dots,\theta_n,\rho_n)\\ x^{n+1}\end{bmatrix}$$

From the equality

$$\Phi_{n+1} = \hat{\Phi}_n \circ \Psi_{n+1}$$

and the chain rule we deduce

$$\det J_{\Phi_{n+1}} = \det J_{\hat{\Phi}_n} \cdot \det J_{\Psi_{n+1}}.$$

A simple computation[6] shows that

$$\det J_{\hat{\Phi}_n} = \det J_{\Phi_n},\ \ \det J_{\Psi_{n+1}} = \rho_{n+1}. \tag{15.3.21}$$

Hence

$$\begin{aligned}\delta_{n+1} = \rho_{n+1}\delta_n = \rho_{n+1}\rho_n\delta_{n-1} = \cdots \rho_{n+1}\rho_n\cdots\rho_3\delta_2\\ = \rho_{n+1}\rho_n\cdots\rho_3\rho_2.\end{aligned}$$

From the equalities

$$\rho_n = \rho_{n+1}\sin\theta_{n+1},\ \ \rho_2 = \rho_3\sin\theta_3$$

we deduce

$$\delta_2 = \rho_2,\ \ \delta_3 = \rho_3^2\sin\theta_3,\ \ \delta_4 = \rho_4\rho_3^2\sin\theta_3 = \rho_4^3(\sin\theta_4)^2\sin\theta_3,$$

and, in general,

$$\boxed{\det J_{\Phi_{n+1}} = \delta_{n+1} = \rho_{n+1}^n(\sin\theta_{n+1})^{n-1}(\sin\theta_n)^{n-2}\cdots\sin\theta_3}. \tag{15.3.22}$$

Note that since $\theta_3,\dots,\theta_n \in (0,\pi)$ we deduce that $\det J_{\Phi_{n+1}} > 0$ so $\det J_{\Phi_{n+1}} = |\det J_{\Phi_{n+1}}|$. □

Example 15.3.8 (The volume of the unit n-dimensional ball)**.** Denote by $\boldsymbol{\omega}_n$ the volume of the closed unit n-dimensional ball

$$\overline{B_1^n(\mathbf{0})} := \{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| \leqslant 1\,\}.$$

Consider the box

$$\mathcal{B}_n := \{\, (\theta_2,\dots,\theta_n,\rho_n) \in \mathbb{R}^n,\ \ \theta_2 \in [0,2\pi],\ \theta_3,\dots,\theta_n \in [0,\pi],\ \ \rho_n \in [0,1]\,\}.$$

The transformation Φ_n sends this box to the closed unit n-dimensional ball $\overline{B_1^n(\mathbf{0})}$.

The equality (15.3.22) shows that the determinant of the Jacobian of Φ_n is bounded on $\mathcal{B}_n$. Applying Theorem 15.3.5 we deduce that

$$\boldsymbol{\omega}_n = \operatorname{vol}_n\left(\overline{B_1^n(\mathbf{0})}\right) = \int_{\mathcal{B}_n} \rho_n^{n-1}(\sin\theta_n)^{n-2}(\sin\theta_{n-1})^{n-3}\cdots\sin\theta_3|d\theta_2 d\theta_3\cdots d\theta_n d\rho_n|$$

[6] You need to perform this simple computation.

(set $\rho := \rho_n$ and use Fubini)

$$= \left(\int_0^1 \rho^{n-1} d\rho\right)\left(\int_0^{2\pi} d\theta_2\right)\left(\int_0^{\pi} \sin\theta_3 d\theta_3\right)\cdots\left(\int_0^{\pi} (\sin\theta_n)^{n-2} d\theta_n\right)$$

$$= \frac{2\pi}{n}\left(\int_0^{\pi} \sin\theta d\theta\right)\cdots\left(\int_0^{\pi} (\sin\theta)^{n-2} d\theta\right).$$

If we set

$$J_k := \int_0^{\pi} (\sin\theta)^k d\theta,$$

then we deduce

$$\boldsymbol{\omega}_n = \frac{2\pi}{n} J_1 J_2 \cdots J_{n-2}.$$

Using the equality

$$\sin(\pi - \theta) = \sin\theta, \quad \forall \theta \in \mathbb{R}$$

we deduce

$$\int_0^{\pi} (\sin\theta)^k d\theta = \int_0^{\frac{\pi}{2}} (\sin\theta)^k d\theta + \int_{\frac{\pi}{2}}^{\pi} (\sin\theta)^k d\theta = 2 \underbrace{\int_0^{\frac{\pi}{2}} (\sin\theta)^k d\theta}_{=:I_k}.$$

Hence

$$\boldsymbol{\omega}_n = \frac{2^{n-1}\pi}{n} I_1 I_2 \cdots I_{n-2} \tag{15.3.23}$$

We have compute the integrals I_k earlier in (9.6.15).

$$I_{2j} = \frac{\pi}{2}\frac{(2j-1)!!}{(2j)!!}, \quad I_{2j-1} = \frac{(2j-2)!!}{(2j-1)!!},$$

where the bi-factorial $n!!$ is defined in (9.6.14).

Thus, if $n = 2k$, then

$$I_1 \cdots I_{n-2} = (I_1 I_2)\cdots(I_{2k-3} I_{2k-2}) = \prod_{j=1}^{k-1} I_{2j-1} I_{2j} = \left(\frac{\pi}{2}\right)^{k-1} \prod_{j=1}^{k-1} \frac{(2j-2)!!}{(2j)!!} =$$

$$= \left(\frac{\pi}{2}\right)^{k-1} \frac{1}{(2k-2)!!} = \frac{\pi^{k-1}}{2^{2k-2}(k-1)!}.$$

For $n = 2k+1$ we have

$$I_1 \cdots I_{n-2} = (I_1 I_2)\cdots(I_{2k-3} I_{2k-2}) I_{2k-1} = \left(\frac{\pi}{2}\right)^{k-1} \frac{1}{((2k-2)!!} \cdot \frac{(2k-2)!!}{(2k-1)!!}$$

$$= \left(\frac{\pi}{2}\right)^{k-1} \frac{1}{(2k-1)!!}.$$

Using (15.3.23) we deduce that the volume $\boldsymbol{\omega}_n$ of the unit n dimensional ball is

$$\boxed{\boldsymbol{\omega}_{2k} = \frac{\pi^k}{k!}, \quad \boldsymbol{\omega}_{2k+1} = \frac{2^{k+1}\pi^k}{(2k+1)!!}}. \tag{15.3.24}$$

We list below the values of $\boldsymbol{\omega}_n$ for small n.

n	0	1	2	3	4	5
$\boldsymbol{\omega}_n$	1	2	π	$\frac{4\pi}{3}$	$\frac{\pi^2}{2}$	$\frac{8\pi^2}{15}$

.

Let us mention one simple consequence of the above computations that will come in handy later.

$$n\boldsymbol{\omega}_n = \left(\int_0^{2\pi} d\theta_2\right)\left(\int_0^{\pi} \sin\theta_3\, d\theta_3\right)\cdots\left(\int_0^{\pi} (\sin\theta_n)^{n-2}\, d\theta_n\right). \tag{15.3.25}$$

□

Example 15.3.9 (Integrals of radially symmetric functions)**.** Here is another useful application of the n-dimensional spherical coordinates. For $0 \leqslant r < R$ define

$$A(r,R) = A^n(r,R) = \left\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ r \leqslant \|\boldsymbol{x}\| \leqslant R \,\right\}.$$

Suppose that $f : A(r,R) \to \mathbb{R}$ is a continuous *radially symmetric* function. This means that there exists $u : [r,R] \to \mathbb{R}$ is a continuous function such that

$$f(\boldsymbol{x}) = u\big(\,\|\boldsymbol{x}\|\,\big),\ \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

We want to show that

$$\boxed{\int_{A^n(r,R)} u(\|\boldsymbol{x}\|)|d\boldsymbol{x}| = \int_{r\leqslant\|\boldsymbol{x}\|\leqslant R} u(\|\boldsymbol{x}\|)\,|d\boldsymbol{x}| = n\boldsymbol{\omega}_n \int_r^R u(\rho)\rho^{n-1}d\rho}. \tag{15.3.26}$$

When $n = 2$ this formula reads

$$\int_{r\leqslant\sqrt{x^2+y^2}\leqslant R} u\big(\sqrt{x^2+y^2}\,\big)\,|dxdy| = 2\pi\int_r^R tu(t)dt,$$

and when $n = 3$ it reads

$$\int_{r\leqslant\sqrt{x^2+y^2+z^2}\leqslant R} u\big(\sqrt{x^2+y^2+z^2}\,\big)\,|dxdydz| = 4\pi\int_r^R t^2u(t)dt.$$

To prove (15.3.26) we use the n-dimensional spherical coordinates $(\theta_2,\dots,\theta_n,\rho=\rho_n)$. We deduce

$$\int_{A^n(r,R)} u(\|\boldsymbol{x}\|)|d\boldsymbol{x}| = \int_{\substack{r\leqslant\rho\leqslant R\\ \theta_2\in[0,2\pi]\ \theta_3,\dots,\theta_n\in[0,\pi]}} u(\rho)\rho^{n-1}\left(\prod_{j=2}^n(\sin\theta_j)^{j-2}\right)|d\theta_2 d\theta_3\cdots d\theta_n d\rho|$$

$$= \left(\int_r^R u(\rho)\rho^{n-1}d\rho\right)\left(\int_0^{2\pi} d\theta_2\right)\left(\int_0^{\pi}\sin\theta_3 d\theta_3\right)\cdots\left(\int_0^{\pi}(\sin\theta_n)^{n-2}d\theta_n\right)$$

$$\overset{(15.3.25)}{=} n\boldsymbol{\omega}_n\int_r^R u(\rho)\rho^{n-1}d\rho.$$

□

15.3.2. Proof of the change of variables formula. We will carry the proof of the change in variables formula (15.3.1) or, equivalently, (15.3.3), in several steps that we describe loosely below.

Step 1. *(De)composition.* If the change in variables formula is valid for the diffeomorphisms Φ_0, Φ_1, then it is valid for their composition $\Phi_1 \circ \Phi_0$, whenever this composition makes sense.

Step 2. *Localization.* Suppose that the diffeomorphism $\Phi : U \to \mathbb{R}^n$ has the property that, for any $\boldsymbol{p} \in U$, there exists an open neighborhood $\mathcal{O}_{\boldsymbol{p}}$ of $\boldsymbol{p}$ such that $\mathcal{O}_{\boldsymbol{p}} \subset U$ and the restriction of Φ to $\mathcal{O}_{\boldsymbol{p}}$ is a diffeomorphism satisfying Theorem 15.3.1. We will show that the entire diffeomorphism Φ satisfies this theorem. This step uses the partition of unity trick.

Step 3. *Elementary diffeomorphism.* We describe a class of so called *elementary* diffeomorphisms for which the change in variables holds. In conjunction with Step 1 we deduce that the change in variable formula holds for *quasi-elementary* diffeomorphisms, i.e., diffeomorphisms that are compositions of several elementary ones. This step uses Fubini's theorem coupled with the one-dimensional change in variables formula.

Step 4. *Everything is (quasi)elementary.* We show that for any diffeomorphism $\Phi : U \to \mathbb{R}^n$, ($U$ open subset in $\mathbb{R}^n$) and for any $\boldsymbol{x} \in U$, there exists a tiny open neighborhood $\mathcal{O}$ of $\boldsymbol{x}$ such that $\mathcal{O} \subset U$ and the restriction of Φ to $\mathcal{O}$ is a quasi-elementary diffeomorphism. This step relies on the inverse function theorem.

Clearly Steps 1-4 imply the validity of Theorem 15.3.1. We present the details below. For simplicity we consider only the case when the integrand f in (15.3.1) is a continuous function that vanishes outside a compact set $K \subset V$. The general case follows from this special case by using Theorem 15.1.26.

Step 1. Let $\Phi : U_0 \to \mathbb{R}^n$, $\Psi : U_1 \to \mathbb{R}^n$ be two diffeomorphisms satisfying Theorem 15.3.1 such that $\Phi(U_0) \subset U_1$. We want to prove that $\Psi \circ \Phi$ also satisfies this theorem. Set

$$V_1 := \Phi(U_0), \;\; V_2 := \Psi(V_1).$$

Suppose that $f : V_2 \to \mathbb{R}$ is continuous and vanishes outside a compact set $K_2 \subset V_2$. The function

$$g : V_1 \to \mathbb{R}, \;\; g(\boldsymbol{y}) = f\big(\,\Psi(\boldsymbol{y})\,\big)\big|\det J_{\Psi}(\boldsymbol{y})\,\big|$$

is continuous and vanishes outside the compact set $K_1 := \Psi^{-1}(K_2) \subset V_1$. Since Ψ satisfies Theorem 15.3.1 we deduce that

$$\int_{V_2} f(\boldsymbol{z})|d\boldsymbol{z}| = \int_{V_1} g(\boldsymbol{y})|d\boldsymbol{y}|. \tag{15.3.27}$$

Similarly, the function

$$h : U_0 \to \mathbb{R}, \;\; h(\boldsymbol{x}) = g\big(\,\Phi(\boldsymbol{x})\,\big)\big|\det J_{\Phi}(\boldsymbol{x})\,\big|$$

is continuous and vanishes outside the compact set $K_0 = \Phi^{-1}(K_1) \subset U_0$. Since Φ satisfies Theorem 15.3.1 we conclude that

$$\int_{V_1} g(\boldsymbol{y})|d\boldsymbol{y}| = \int_{U_0} h(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.3.28}$$

Now observe that

$$h(\boldsymbol{x}) = f\big(\,\Psi\circ\Phi(\boldsymbol{x})\,\big)\cdot\big|\,\det J_\Psi(\Phi(\boldsymbol{x}))\,\big|\cdot\big|\det J_\Phi(\boldsymbol{x})\,\big|$$

The chain rule (13.3.4) implies that

$$J_{\Psi\circ\Phi}(\boldsymbol{x}) = J_\Psi(\Phi(\boldsymbol{x}))J_\Phi(\boldsymbol{x})$$

so

$$\big|\,\det J_{\Psi\circ\Phi}(\boldsymbol{x})\,\big| = \big|\,\det J_\Psi(\Phi(\boldsymbol{x}))\,\big|\cdot\big|\det J_\Phi(\boldsymbol{x})\,\big|$$

and

$$h(\boldsymbol{x}) = f\big(\,\Psi\circ\Phi(\boldsymbol{x})\,\big)\cdot\big|\,\det J_{\Psi\circ\Phi}(\boldsymbol{x})\,\big|.$$

We deduce from (15.3.27) and (15.3.28) that

$$\int_{V_2} f(\boldsymbol{z})|d\boldsymbol{z}| = \int_{U_0} f\big(\,\Psi\circ\Phi(\boldsymbol{x})\,\big)\cdot\big|\,\det J_{\Psi\circ\Phi}(\boldsymbol{x})\,\big||d\boldsymbol{x}|.$$

This shows that $\Psi\circ\Phi$ satisfies Theorem 15.3.1

Step 2. Suppose that $\Phi : U \to \mathbb{R}^n$ is a diffeomorphism such that, for any point $\boldsymbol{p} \in U$ there exists an open neighborhood $\mathcal{O}_{\boldsymbol{p}}$ of $\boldsymbol{p}$ with the following properties.

(i) $\mathcal{O}_{\boldsymbol{p}} \subset U$. Set $\hat{\mathcal{O}}_{\boldsymbol{p}} := \Phi(\mathcal{O}_{\boldsymbol{p}})$.

(ii) Any continuous function $g : V \to \mathbb{R}$ that vanishes outside a compact set $C_{\boldsymbol{p}} \subset \hat{\mathcal{O}}_{\boldsymbol{p}}$ satisfies the change in variables formula (15.3.3).

We will show that this condition implies that any continuous function $g : V \to \mathbb{R}$ that vanishes outside a compact set $C \subset V$ satisfies the change in variables formula (15.3.3).

The collection of open sets $\big\{\,\hat{\mathcal{O}}_{\boldsymbol{p}}\,\big\}_{\boldsymbol{p}\in\Phi^{-1}(C)}$ is an open cover of C. Fix a compactly supported partition of unity on K subordinated to this open cover. This consists of finitely many *compactly supported* continuous functions

$$\chi_1, \dots, \chi_\ell : \mathbb{R}^n \to [0,1]$$

satisfying the following properties.

- For any $j = 1, \dots, \ell$ there exists $\boldsymbol{p}_j \in C$ such that $\operatorname{supp}\chi_j \subset \hat{\mathcal{O}}_{\boldsymbol{p}_j}$.
- $\chi_1(\boldsymbol{y}) + \cdots + \chi_\ell(\boldsymbol{y}) = 1$, $\forall \boldsymbol{y} \in C$.

Set $g_j := \chi_j g$. Observe that since $g(\boldsymbol{y}) = 0$, $\forall \boldsymbol{y} \in V\backslash C$, we have

$$g_1(\boldsymbol{y}) + \cdots + g_\ell(\boldsymbol{y}) = \big(\,\chi_1(\boldsymbol{y}) + \cdots + \chi_\ell(\boldsymbol{y})\,\big)g(\boldsymbol{y}) = g(\boldsymbol{y}), \quad \forall \boldsymbol{y} \in V.$$

Clearly g_j is continuous and supported on a compact set contained in $\mathcal{O}_{\boldsymbol{p}_j}$. It satisfies the change-in-variables formula (15.3.3)

$$\int_V g_j(\boldsymbol{x})\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{x}| = \int_{\Phi(\mathcal{O}_{\boldsymbol{p}_j})} g_j(\boldsymbol{y})|d\boldsymbol{x}|$$

$$= \int_{\mathcal{O}_{\boldsymbol{p}_j}} g_j\big(\Phi(\boldsymbol{x})\big)\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{y}| = \int_U g_j\big(\Phi(\boldsymbol{x})\big)\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{x}|.$$

Summing these equalities we deduce

$$\int_U g\big(\Phi(\boldsymbol{x})\big)\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{x}| = \sum_{j=1}^{\ell}\int_U g_j\big(\Phi(\boldsymbol{x})\big)\big|\det J_\Phi(\boldsymbol{x})\big||d\boldsymbol{x}|$$

$$= \sum_{j=1}^{\ell}\int_V g_j(\boldsymbol{y})|d\boldsymbol{y}| = \int_V g(\boldsymbol{y})|d\boldsymbol{y}|.$$

This proves that the diffeomorphism Φ satisfies the change-in-variables formula (15.3.3).

Step 3. Let $j \in \{1, \dots n\}$. We say that a diffeomorphism

$$\Phi : U \to \mathbb{R}^n,\ \ U \ni \boldsymbol{x} \mapsto \boldsymbol{y} = \Phi(\boldsymbol{x}) = \begin{bmatrix} \Phi^1(\boldsymbol{x}) \\ \Phi^2(\boldsymbol{x}) \\ \vdots \\ \Phi^n(\boldsymbol{x}) \end{bmatrix}$$

is j-*elementary* if, for any $i \neq j$, we have

$$\Phi^i(x^1, \dots, x^n) = x^i.$$

Note that the inverse of a j-elementary diffeomorphism is also a j-elementary diffeomorphism. We say that Φ is elementary if it is j-elementary for some index $j = 1, \dots, n$. We will show that if $\Phi : U \to \mathbb{R}^n$ is elementary, then it satisfies Theorem 15.3.1. To complete this we rely on the localization trick discussed in Step 2. Assume for simplicity that Φ is 1-elementary, i.e.,

$$\Phi(\boldsymbol{x}) = \begin{bmatrix} \varphi(x^1, \dots, x^n) \\ x^2 \\ \vdots \\ x^n \end{bmatrix}$$

where $\varphi : U \to \mathbb{R}$ is a C^1 function. The Jacobian matrix of Φ is

$$J_\Phi(\boldsymbol{x}) = \begin{bmatrix} \varphi'_{x^1}(\boldsymbol{x}) & \varphi'_{x^2}(\boldsymbol{x}) & \varphi'_{x^3}(\boldsymbol{x}) & \varphi'_{x^4}(\boldsymbol{x}) & \cdots & \varphi'_{x^n}(\boldsymbol{x}) \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Observe that

$$\det J_\Phi(\boldsymbol{x}) = \varphi'_{x^1}(\boldsymbol{x}). \tag{15.3.29}$$

Since Φ is a diffeomorphism, we deduce $\det J_\Phi(\boldsymbol{x}) \neq 0$, $\forall \boldsymbol{x} \in U$, i.e.,

$$\varphi'_{x^1}(\boldsymbol{x}) \neq 0, \;\; \forall \boldsymbol{x} \in U.$$

Let $\boldsymbol{p} \in U$. We want to show that there exists an open neighborhood $\mathcal{O}_{\boldsymbol{p}}$ of $\boldsymbol{p}$ in U such that the restriction of Φ to that neighborhood satisfies Theorem 15.3.1.

Fix $r > 0$ small enough such that the open cube $C_{2r}(\boldsymbol{p})$ (see Definition 11.3.12) is contained in U. We will prove that the restriction of Φ to $C_r(\boldsymbol{p})$ satisfies the change-in-variables formula.

The cube $C_{2r}(\boldsymbol{p})$ is connected and the continuous function φ'_{x^1} does not vanish in this cube. Hence it must have constant sign. Assume for simplicity that[7]

$$\varphi'_{x^1}(\boldsymbol{x}) > 0, \;\; \forall \boldsymbol{x} \in C_{2r}(\boldsymbol{p}).$$

Note that

$$C_r(\boldsymbol{p}) = \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\;\; |x^i - p^i| < r, \;\; \forall i = 1, \dots, n \,\big\}.$$

For any $\boldsymbol{x} = (x^1, \dots, x^n)$ we set $\bar{\boldsymbol{x}} := (x^2, \dots, x^n)$ and denote by $C'_r(\bar{\boldsymbol{p}}) \subset \mathbb{R}^{n-1}$ the open $(n-1)$-dimensional cube of radius r centered at $\bar{\boldsymbol{p}}$. Using this notation we have

$$C_r(\boldsymbol{p}) = \big\{\, (x^1, \bar{\boldsymbol{x}}) \in \mathbb{R} \times \mathbb{R}^{n-1};\;\; x^1 \in (p^1 - r, p^1 + r), \;\; \bar{\boldsymbol{x}} \in C'_r(\bar{\boldsymbol{p}}) \,\big\}.$$

For each $\bar{\boldsymbol{x}} \in C'_r(\bar{\boldsymbol{p}})$ the function $x^1 \mapsto \varphi(x^1, \bar{\boldsymbol{x}})$ sends the interval $(p^1 - r, p^1 + r)$ to the interval

$$I_{\bar{\boldsymbol{x}}} := \big\{\, y^1 \in \mathbb{R};\;\; \varphi(p^1 - r, \bar{\boldsymbol{x}}) < y^1 < \varphi(p^1 + r, \bar{\boldsymbol{x}}) \,\big\}.$$

We deduce that the image of $C_r(\boldsymbol{p})$ via Φ is the simple-type domain

$$\mathscr{C}(r, \boldsymbol{p}) = \big\{\, (y^1, \bar{\boldsymbol{y}}) \in \mathbb{R} \times \mathbb{R}^{n-1};\;\; \varphi(p^1 - r, \bar{\boldsymbol{y}}) < y^1 < \varphi(p^1 + r, \bar{\boldsymbol{y}}), \;\; \bar{\boldsymbol{y}} \in C'_r(\bar{\boldsymbol{p}}) \,\big\}.$$

[7] The case $\varphi'_{x^1} < 0$ is dealt with in a similar fashion.

Suppose that $f : \mathscr{C}(r, \boldsymbol{p}) \to \mathbb{R}$ is a continuous function that vanishes outside a compact set $K \subset \mathscr{C}(r, \boldsymbol{p})$. Using Fubini's theorem we deduce

$$\int_{\mathscr{C}(r,\boldsymbol{p})} f(\boldsymbol{y})d\boldsymbol{y} = \int_{C_r'(\bar{\boldsymbol{p}})} \left(\int_{\varphi(p^1-r,\bar{\boldsymbol{x}})}^{\varphi(p^1+r,\bar{\boldsymbol{x}})} f(y^1,\bar{\boldsymbol{y}})dy^1 \right) d\bar{\boldsymbol{y}}.$$

Fix $\bar{\boldsymbol{x}}$ and thus $\bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}$. Using the one-dimensional change-in-variables formula (9.6.20) we deduce

$$\int_{\varphi(p^1-r,\bar{\boldsymbol{x}})}^{\varphi(p^1+r,\bar{\boldsymbol{x}})} f(y^1,\bar{\boldsymbol{y}})dy^1 = \int_{p^1-r}^{p^1+r} f\bigl(\varphi(x^1,\bar{\boldsymbol{x}}),\bar{\boldsymbol{y}})\varphi'_{x^1}(x^1,\bar{\boldsymbol{y}})dx^1.$$

Hence

$$\int_{\mathscr{C}(r,\boldsymbol{p})} f(\boldsymbol{y})d\boldsymbol{y} = \int_{C_r'(\bar{\boldsymbol{p}}))} \left(\int_{p^1-r}^{p^1+r} f\bigl(\varphi(x^1,\bar{\boldsymbol{x}}),\bar{\boldsymbol{y}})\varphi'_{x^1}(x^1,\bar{\boldsymbol{y}})dx^1 \right) d\bar{\boldsymbol{y}}.$$

(rename by $\bar{\boldsymbol{x}}$ the variables $\bar{\boldsymbol{y}}$)

$$= \int_{C_r'(\bar{\boldsymbol{p}}))} \left(\int_{p^1-r}^{p^1+r} f\bigl(\varphi(x^1,\bar{\boldsymbol{x}}),\bar{\boldsymbol{x}}\bigr)\varphi'_{x^1}(x^1,\bar{\boldsymbol{x}})dx^1 \right) d\bar{\boldsymbol{x}}$$

(use Fubini again)

$$= \int_{C_r(\boldsymbol{p})} f\bigl(\varphi(x^1),\bar{\boldsymbol{x}}\bigr)\varphi'_{x^1}(x^1,\bar{\boldsymbol{x}})|d\boldsymbol{x}| \overset{(15.3.29)}{=} \int_{C_r(\boldsymbol{p})} f\bigl(\Phi(\boldsymbol{x})\bigr) \cdot \bigl| \det J_\Phi(\boldsymbol{x}) \bigr| |d\boldsymbol{x}|.$$

Step 4. To give you a taste of the main idea we consider first the special case $n = 2$. Then, $\Phi(\boldsymbol{x})$ has the form

$$\boldsymbol{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} \mapsto \Phi(\boldsymbol{x}) = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} \phi^1(x^1,x^2) \\ \phi^2(x^1,x^2) \end{bmatrix}.$$

The Jacobian matrix of Φ is

$$J_\Phi = \begin{bmatrix} \frac{\partial\phi^1}{\partial x^1} & \frac{\partial\phi^1}{\partial x^2} \\ & \\ \frac{\partial\phi^2}{\partial x^1} & \frac{\partial\phi^2}{\partial x^2} \end{bmatrix}.$$

Since Φ is a diffeomorphism, $\det J_\Phi(\boldsymbol{p}) \neq 0$ so at least one of the entries $\frac{\partial\phi^i}{\partial x^j}$ must be nonzero at $\boldsymbol{p}$. After a possible relabeling of the variables $\boldsymbol{y}$ and/or $\boldsymbol{x}$ we can assume that $\frac{\partial\phi^1}{\partial x^1} \neq 0$. The implicit function theorem shows that in the equation

$$y^1 - \phi^1(x^1,x^2) = 0$$

we can locally solve for x^1 in terms of y^1 and x^2, i.e., we can regard x^1 as an implicitly defined C^1-function depending on the variables y^1, x^2,

$$x^1 = \psi^1(y^1,x^2) \Longleftrightarrow y^1 = \phi^1\bigl(\psi^1(y^1,x^2),x^2\bigr). \tag{15.3.30}$$

This shows that the map

$$\boldsymbol{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} \mapsto \Phi_1(\boldsymbol{x}) = \begin{bmatrix} y^1 \\ x^2 \end{bmatrix} = \begin{bmatrix} \phi^1(x^1,x^2) \\ x^2 \end{bmatrix}$$

is a 1-elementary diffeomorphism defined on an open neighborhood U_1 and its inverse, denoted by Υ_1, is the 1-elementary diffeomorphism described explicitly by

$$\begin{bmatrix} y^1 \\ x^2 \end{bmatrix} \overset{\Upsilon_1}{\mapsto} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} = \begin{bmatrix} \psi^1(y^1, x^2) \\ x^2 \end{bmatrix}.$$

Now consider the composition $\Psi_2 = \Phi \circ \Upsilon_1$. More precisely,

$$\begin{bmatrix} y^1 \\ x^2 \end{bmatrix} \overset{\Upsilon_1}{\mapsto} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} \overset{\Phi}{\mapsto} \begin{bmatrix} y^1 \\ y^2 \end{bmatrix} \overset{(15.3.30)}{=} \begin{bmatrix} y^1 \\ \phi^2\big(\,\psi^1(y^1, x^2), x^2\,\big) \end{bmatrix}.$$

This shows that Ψ_2 is 2-elementary. The equality $\Psi_2 = \Phi \circ \Upsilon_1 = \Phi \circ \Phi_1^{-1}$ implies that

$$\Phi = \Psi_2 \circ \Phi_1,$$

i.e., locally, Φ is the composition of elementary diffeomorphisms.

We outline now how the above approach extends to arbitrary n, referring for details to [**44**, Sec. 8.6.4]. Given a diffeomorphism $\Phi : U \to \mathbb{R}^n$ and a point $\boldsymbol{p} \in U$ one shows, using the implicit function theorem, that there exist elementary diffeomorphisms $\Psi_n, \dots, \Psi_1$ such that Ψ_1 is defined on an open neighborhood $\mathcal{O}$ of $\boldsymbol{p}$ and

$$\Phi(\boldsymbol{x}) = \Psi_n \circ \Psi_{n-1} \circ \cdots \circ \Psi_1(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in \mathcal{O}$$

This process proceeds gradually. First, using the inverse function theorem one constructs an open neighborhood U_{n-1} of $\boldsymbol{p}$ in U and a diffeomorphism $\Phi_{n-1} : U_{n-1} \to \mathbb{R}^n$ with the following two properties.

(A) The n-th component of Φ_{n-1} has the special form $\Phi^n_{n-1}(\boldsymbol{x}) = x^n$, $\forall \boldsymbol{x} \in U_{n-1}$.

(B) The diffeomorphism $\Psi_n := \Phi \circ \Phi^{-1}_{n-1}$ is n-elementary.

Note that $\Phi = \Psi_n \circ \Phi_{n-1}$. Proceeding inductively, again relying on the implicit function theorem, one constructs open sets

$$U_n \supset U_{n-1} \supset U_{n-2} \supset \cdots \supset U_1 \ni \boldsymbol{p}$$

and, for any $k = 2, \dots, n$, diffeomorphisms $\Phi_k : U_k \to \mathbb{R}^n$, with the following properties.

- $\Phi_n = \Phi$.
- If Φ^j_k is the j-th component of Ψ_k, then

$$\Phi^j_k(\boldsymbol{x}) = x^j, \;\; \forall j > k, \;\; \boldsymbol{x} \in U_k. \tag{15.3.31}$$

- The diffeomorphism $\Psi_k := \Phi_k \circ \Phi^{-1}_{k-1}$ is k-elementary.

We deduce

$$\begin{aligned} \Phi(\boldsymbol{x}) = \Phi_n &= \big(\Phi_n \circ \Phi^{-1}_{n-1}\big) \circ \big(\Phi_{n-1} \circ \Phi^{-1}_{n-2}\big) \circ \cdots \circ \big(\Phi_2 \circ \Phi^{-1}_1\big) \circ \Phi_1 \\ &= \Psi_n \circ \cdots \circ \Psi_2 \circ \Phi_1(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in U_1. \end{aligned}$$

By construction, the diffeomorphism Ψ_k is k-elementary, $\forall k \geqslant 2$, while (15.3.31) with $k = 1$, shows that Φ_1 is 1-elementary.

This completes our outline of the proof of the change-in-variables formula (15.3.1). For a different, more intuitive but more laborious approach we refer to [**29**, §XX.4]

15.4. Improper integrals

Concrete problems arising in mathematics and natural science force us to integrate functions that are not covered by the theory developed so far. For example, we might want to integrate an unbounded function, or we might want to integrate a function over an unbounded region. The goal of this section is to explain how to handle such issues. Let use first introduce a notation.

> ✍ *For any set $A \subset \mathbb{R}^n$ we denote by $\mathfrak{J}(A)$ the collection of Jordan measurable subsets of A and by $\mathfrak{J}_c(A)$ the collection of compact Jordan measurable subsets of A.*

15.4.1. Locally integrable functions. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set.

Definition 15.4.1. A function $f : U \to \mathbb{R}$ is called *locally integrable* if, for any $\boldsymbol{x} \in U$, there exists a closed box $B = B_{\boldsymbol{x}}$ such that $B_{\boldsymbol{x}} \subset U$, $\boldsymbol{x} \in \boldsymbol{int}(B_{\boldsymbol{x}})$ and the restriction of f to $B_{\boldsymbol{x}}$ is Riemann integrable. □

Example 15.4.2. (a) Any continuous function $f : U \to \mathbb{R}$ is locally integrable.

(b) If the open set $U \subset \mathbb{R}^n$ is Jordan measurable, then any Riemann integrable function $f : U \to \mathbb{R}$ is also locally integrable. □

Proposition 15.4.3. *For any function $f : U \to \mathbb{R}$ the following statements are equivalent.*

(i) *The function f is locally integrable.*

(ii) *For any compact Jordan measurable set $K \subset U$, the restriction of f to K is Riemann integrable on K.*

Proof. Clearly (ii) ⇒ (i) since any closed box is compact and Jordan measurable. Let us prove (i) ⇒ (ii). Suppose that $K \subset U$ is compact and Jordan measurable. For any $\boldsymbol{x} \in K$ choose a closed box $B_{\boldsymbol{x}}$ satisfying the conditions in Definition 15.4.1. Consider a partition of unity on K subordinated to the open cover

$$\Big\{\, \boldsymbol{int}(B_{\boldsymbol{x}});\;\; \boldsymbol{x} \in K \,\Big\}_{\boldsymbol{x} \in K}.$$

Recall (see Definition 12.4.6) that this consists of a finite collection of compactly supported continuous functions

$$\chi_1, \dots, \chi_\ell : \mathbb{R}^n \to \mathbb{R}$$

with the following properties;

$$\chi_1(\boldsymbol{x}) + \cdots + \chi_\ell(\boldsymbol{x}) = 1, \;\; \forall \boldsymbol{x} \in K, \tag{15.4.1a}$$

$$\forall i = 1, \dots, \ell \;\; \exists \boldsymbol{x}_i \in K \;\;\; \operatorname{supp} \chi_i \subset \boldsymbol{int}(B_{\boldsymbol{x}_i}). \tag{15.4.1b}$$

Denote by f^0 the extension of f by 0. The functions $I_{B_{\boldsymbol{x}_i}} f^0$ are Riemann integrable and so are the functions

$$\chi_i I_{B_{\boldsymbol{x}_i}} f^0 \overset{(15.4.1a)}{=} \chi_i f^0.$$

Hence $\chi_1 f^0 + \cdots + \chi_\ell f^0$ is Riemann integrable. Since K is Jordan measurable, we deduce that the function

$$I_K\big(\chi_1 + \cdots + \chi_\ell\big) f^0 \overset{(15.4.1b)}{=} I_K f^0$$

is also Riemann integrable. □

Definition 15.4.4. A *compact exhaustion* of U is a sequence of compact sets $(K_\nu)_{\nu\in\mathbb{N}}$ with the following properties.

(i) $K_\nu \subset \boldsymbol{int}(K_{\nu+1})$, $\forall \nu \in \mathbb{N}$.

(ii)

$$U = \bigcup_{\nu\in\mathbb{N}} K_\nu.$$

The compact exhaustion $(K_\nu)_{\nu\in\mathbb{N}}$ is called *Jordan measurable* if all the compact sets K_ν are Jordan measurable. □

Observe that if $(K_\nu)_{\nu\in\mathbb{N}}$ is a compact exhaustion of the open set U, then the collection of interiors $(\boldsymbol{int}\, K_\nu)_{\nu\in\mathbb{N}}$ is *increasing* and covers U, i.e.,

$$\boldsymbol{int}\, K_1 \subset \boldsymbol{int}\, K_2 \subset \cdots \subset \boldsymbol{int}\, K_\nu \subset \boldsymbol{int}\, K_{\nu+1} \subset \cdots, \quad \bigcup_{\nu\geqslant 1} \boldsymbol{int}\, K_\nu = U.$$

Example 15.4.5. The collection

$$K_\nu = \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| \leqslant \nu \big\},\ \ \nu \in \mathbb{N},$$

is a Jordan measurable compact exhaustion of $\mathbb{R}^n$. The collection

$$K_\nu = \Big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ \frac{1}{\nu} \leqslant \|\boldsymbol{x}\| \leqslant 1 - \frac{1}{\nu} \Big\},\ \ \nu \in \mathbb{N},$$

is a Jordan measurable compact exhaustion of $B_1(\mathbf{0})\setminus\{\mathbf{0}\}$. □

Proposition 15.4.6. *Any open $U \subset \mathbb{R}^n$ set admits Jordan measurable compact exhaustions.*

Proof. Denote by C the complement of U in $\mathbb{R}^n$, $C := \mathbb{R}^n\setminus U$. By definition, C is a closed set. For $\nu \in \mathbb{N}$ we set

$$K_\nu := \overline{B_\nu(\mathbf{0})} \cap \Big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ \operatorname{dist}(\boldsymbol{x}, C) \geqslant \frac{1}{\nu} \Big\}.$$

Clearly K_ν is closed as the intersection of two closed sets. It is bounded since it is contained in the closed ball of radius ν. Hence K_ν is compact. Note that K_ν is contained in U since $\boldsymbol{x} \in U = \mathbb{R}^n\setminus C$ if and only if $\operatorname{dist}(\boldsymbol{x}, C) > 0$. Obviously

$$U = \bigcup_{\nu\in\mathbb{N}} K_\nu.$$

Note that

$$K_\nu \subset B_{\nu+1}(\mathbf{0}) \cap \left\{ \boldsymbol{x} \in \mathbb{R}^n;\ \operatorname{dist}(\boldsymbol{x}, C) > \frac{1}{\nu+1} \right\} \subset \boldsymbol{int}(K_{\nu+1})$$

Hence the collection $(K_\nu)_{\nu\in\mathbb{N}}$ is a compact exhaustion of U. However, it may not be Jordan measurable. We can modify it to a Jordan measurable one as follows.

Since K_ν is compact there exist finitely many closed boxes contained in $\boldsymbol{int}(K_{\nu+1})$ such that their interiors cover K_ν. Denote by $\tilde{K}_\nu$ the union of these finitely many closed boxes. Clearly $\tilde{K}_\nu$ is compact and Jordan measurable, and

$$K_\nu \subset \tilde{K}_\nu \subset K_{\nu+1} \subset \tilde{K}_{\nu+1}.$$

The collection $(\tilde{K}_\nu)_{\nu\in\mathbb{N}}$ is a Jordan measurable compact exhaustion of U. □

Proposition 15.4.7. *Suppose that $U \subset \mathbb{R}^n$ is a Jordan measurable open set and $f : U \to \mathbb{R}$ is a Riemann integrable function. Then, for any Jordan measurable compact exhaustion $(K_\nu)_{\nu\in\mathbb{N}}$ of U we have*

$$\int_U f(\boldsymbol{x})|d\boldsymbol{x}| = \lim_{\nu\to\infty} \int_{K_\nu} f(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.4.2}$$

Proof. Fix a Jordan measurable compact exhaustion $(K_\nu)_{\nu\in\mathbb{N}}$ and set

$$M := \sup_{\boldsymbol{x}\in U} |f(\boldsymbol{x})|.$$

Since f is Riemann integrable, hence bounded, and therefore $M < \infty$. Fix $\varepsilon > 0$.

Since U is Jordan measurable, its boundary is negligible. We can thus cover ∂U with finitely many open boxes such that their union Δ_ε has volume $< \frac{\varepsilon}{M}$. The set $S_\varepsilon := U \backslash \Delta_\varepsilon$. Observe that $S_\varepsilon \subset U$ is closed[8] and bounded so it is compact. The increasing collection of open sets $\boldsymbol{int}(K_\nu)$, $\nu \in \mathbb{N}$, is an open cover of the *compact* set S_ε and thus there exists $N = N(\varepsilon)$ such that $S_\varepsilon \subset K_\nu$ for all $\nu \geqslant N(\varepsilon)$. Note that this implies

$$U\backslash K_\nu \subset U\backslash S_\varepsilon \subset \Delta(\varepsilon)$$

so that

$$\operatorname{vol}_n(U\backslash K_\nu) \leqslant \operatorname{vol}_n(\Delta_\varepsilon) < \frac{\varepsilon}{M},\ \ \forall \nu \geqslant N(\varepsilon).$$

For $\nu \geqslant N(\varepsilon)$ we have

$$\left| \int_U f(\boldsymbol{x})|d\boldsymbol{x}| - \int_{K_\nu} f(\boldsymbol{x})|d\boldsymbol{x}| \right| = \left| \int_{U\backslash K_\nu} f(\boldsymbol{x})|d\boldsymbol{x}| \right| \leqslant \int_{U\backslash K_\nu} |f(\boldsymbol{x})||d\boldsymbol{x}|$$

$$\leqslant \int_{U\backslash K_\nu} M|d\boldsymbol{x}| = M \operatorname{vol}_n(U\backslash K_\nu) < \varepsilon.$$

This proves (15.4.2). □

[8] Why?

15.4.2. Absolutely integrable functions.

Let $U \subset \mathbb{R}^n$ be an open set. Recall that for any $A \subset \mathbb{R}^n$ we denoted by $\mathscr{J}_c(A)$ the collection of compact, Jordan measurable subsets of A.

Definition 15.4.8. A locally integrable function $f : U \to \mathbb{R}$ is called *absolutely integrable* if

$$\sup_{K \in \mathscr{J}_c(U)} \int_K |f(\boldsymbol{x})| \, |d\boldsymbol{x}| < \infty.$$

We will denote by $\mathcal{R}_a(U)$ the collection of absolutely integrable functions $f : U \to \mathbb{R}$. □

Proposition 15.4.9. *Let $f : U \to \mathbb{R}$ be a locally integrable function. Then the following statements are equivalent.*

(i) *The function f is absolutely integrable.*

(ii) *For any Jordan measurable compact exhaustion $(K_\nu)_{\nu \in \mathbb{N}}$ of U the sequence*

$$\int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}|$$

is bounded

(iii) *There exists a Jordan measurable compact exhaustion $(K_\nu)_{\nu \in \mathbb{N}}$ of U such that the sequence*

$$\int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}|$$

is bounded.

Moreover, if any of the above conditions is satisfied, then

$$\lim_{\nu \to \infty} \int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}| = \sup_{K \in \mathscr{J}_c(U)} \int_K |f(\boldsymbol{x})| \, |d\boldsymbol{x}|.$$

Proof. Clearly (i) ⇒ (ii) ⇒ (iii). We only have to prove (iii) ⇒ (i). Suppose that $(K_\nu)_{\nu \in \mathbb{N}}$ is a Jordan measurable compact exhaustion of U such that the sequence

$$\int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}|$$

is bounded. We denote by L its supremum. We have to prove that,

$$\sup_{K \in \mathscr{J}_c(U)} \int_K |f(\boldsymbol{x})| \, |d\boldsymbol{x}| = L.$$

Since for every ν we have

$$\int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}| \leqslant \sup_{K \in \mathscr{J}_c(U)} \int_K |f(\boldsymbol{x})| \, |d\boldsymbol{x}$$

we deduce

$$L = \sup_\nu \int_{K_\nu} |f(\boldsymbol{x})| \, |d\boldsymbol{x}| \leqslant \sup_{K \in \mathscr{J}_c(U)} \int_K |f(\boldsymbol{x})| \, |d\boldsymbol{x}| =: L^*.$$

To prove that $L^* \leqslant L$ it suffices to show that, for any $K \in \mathcal{J}_c(U)$ we can find $\nu \in \mathbb{N}$ such that $K \subset \boldsymbol{int}(K_\nu)$. Indeed, if this were the case we would have

$$\int_K |f(\boldsymbol{x})|\,|d\boldsymbol{x}| \leqslant \int_{K_\nu} |f(\boldsymbol{x})|\,|d\boldsymbol{x}| \leqslant L.$$

To prove the claim note that the *increasing* collection of open sets $\boldsymbol{int}(K_\nu)$, $\nu \in \mathbb{N}$, is an open cover of U and thus also of the *compact* set K. Thus there exist natural numbers $\nu_1 < \cdots < \nu_\ell$ such that the finite collection $\boldsymbol{int}(K_{\nu_1}), \dots, \boldsymbol{int}(K_{\nu_\ell})$ covers K. Since

$$\boldsymbol{int}(K_{\nu_1}) \subset \cdots \subset \boldsymbol{int}(K_{\nu_\ell})$$

we deduce $K \subset \boldsymbol{int}(K_{\nu_\ell})$.

The last conclusion of the proposition follows by observing that the sequence

$$\int_{K_\nu} |f(\boldsymbol{x})|\,|d\boldsymbol{x}|,\ \ \nu \in \mathbb{N},$$

is nondecreasing so

$$\lim_{\nu\to\infty} \int_{K_\nu} |f(\boldsymbol{x})|\,|d\boldsymbol{x}| = \sup_{\nu\in\mathbb{N}} \int_{K_\nu} |f(\boldsymbol{x})|\,|d\boldsymbol{x}| = L = L^*.$$

□

Definition 15.4.10. Suppose that $f : U \to [0,\infty)$ is a *nonnegative* locally integrable function. We set

$$\int_U^* f(\boldsymbol{x})\,|d\boldsymbol{x}| := \sup_{K\in\mathcal{J}_c(U)} \int_K f(\boldsymbol{x})\,|d\boldsymbol{x}|.$$

□

Remark 15.4.11. Note that if U is Jordan measurable and $f : U \to [0,\infty)$ is integrable, then it is absolutely integrable. Moreover, Propositions 15.4.7 and 15.4.9 show that

$$\int_U^* f(\boldsymbol{x})|d\boldsymbol{x}| = \int_U f(\boldsymbol{x})dx.$$

□

To proceed we need to introduce a useful trick. For any $x \in \mathbb{R}$ we set

$$x_+ := \max(x,0),\ \ x_- := \max(-x,0) = (-x)_+.$$

We will refer to $x_\pm$ as the *positive/negative part* of x. For example,

$$(2)_+ = 2,\ \ (2)_- = 0,\ \ (-3)_+ = 0,\ \ (-3)_- = 3.$$

Note that

$$x = x_+ - x_-,\ \ |x| = x_+ + x_-,\ \ \forall x \in \mathbb{R}.$$

The functions $x \mapsto x_\pm$ can be given the alternate descriptions

$$\boxed{x_+ = \frac{|x|+x}{2},\ \ x_- = \frac{|x|-x}{2},\ \ \forall x \in \mathbb{R}}.$$

This shows that the functions $x \mapsto x_\pm$ are Lipschitz since they are linear combinations of the Lipschitz functions $x \mapsto x$ and $x \mapsto |x|$.

Suppose now that U is an open set and $f : U \to \mathbb{R}$ is *absolutely integrable*. Theorem 15.1.12 implies that the functions $f_{\pm} : U \to \mathbb{R}$, $f_{\pm}(\boldsymbol{x}) = f(\boldsymbol{x})_{\pm}$, are locally integrable. From the equality

$$|f(\boldsymbol{x})| = f_{+}(\boldsymbol{x}) + f_{-}(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in U,$$

we deduce that the functions $f_{\pm}$ are also absolutely integrable. The equality $f = f_{+} - f_{-}$ suggests the following concept.

Definition 15.4.12. Suppose that the function $f : U \to \mathbb{R}$ is absolutely integrable. We set

$$\int_U^{**} f(\boldsymbol{x})\,|d\boldsymbol{x}| := \int_U^{*} f_{+}(\boldsymbol{x})\,|d\boldsymbol{x}| \; - \; \int_U^{*} f_{-}(\boldsymbol{x})\,|d\boldsymbol{x}|.$$

We say that $\int_U^{**} f(\boldsymbol{x})|d\boldsymbol{x}|$ is the *improper integral* over U of the absolutely integrable function $f : U \to \mathbb{R}$. □

Remark 15.4.13. (a) If $f : U \to \mathbb{R}$ is absolutely integrable, then, for any Jordan measurable compact exhaustion $(K_\nu)_{\nu \in \mathbb{N}}$ of U we have

$$\int_U^{**} f(\boldsymbol{x})|d\boldsymbol{x}| = \lim_{\nu\to\infty} \int_{K_\nu} f_{+}(\boldsymbol{x})\,|d\boldsymbol{x}| - \lim_{\nu\to\infty} \int_{K_\nu} f_{-}(\boldsymbol{x})\,|d\boldsymbol{x}| = \lim_{\nu\to\infty} \int_{K_\nu} f(\boldsymbol{x})|d\boldsymbol{x}|. \tag{15.4.3}$$

(b) If $f : U \to [0, \infty)$ is absolutely integrable, then we deduce from the equality $f = f_{+}$ that

$$\int_U^{**} f(\boldsymbol{x})\,|d\boldsymbol{x}| := \int_U^{*} f_{+}(\boldsymbol{x})\,|d\boldsymbol{x}| = \int_U^{*} f(\boldsymbol{x})\,|d\boldsymbol{x}|.$$

(c) If U is Jordan measurable, and $f : U \to \mathbb{R}$ is Riemann integrable, then we deduce from Remark 15.4.11 that

$$\int_U^{**} f(\boldsymbol{x})\,|d\boldsymbol{x}| = \int_U f(\boldsymbol{x})\,|d\boldsymbol{x}|.$$

□

✍ *In view of the above remarks, and in order to control the proliferation of notations for closely related concepts, we will continue to use the notation $\int_U f(\boldsymbol{x})\,|d\boldsymbol{x}|$ when referring to the improper integral of f over U. Moreover, we will say that the integral $\int_U f(\boldsymbol{x})\,|d\boldsymbol{x}|$ is absolutely convergent to indicate that the function $f : U \to \mathbb{R}$ is absolutely integrable and the integral is defined by the procedure described above.*

Theorem 15.4.14 (Comparison principle)**.** *Suppose that $f, F : U \to \mathbb{R}$ are locally integrable functions such that*

$$|f(x)| \leqslant |F(\boldsymbol{x})|, \;\; \forall \boldsymbol{x} \in U,$$

(i) *If F is absolutely integrable, then f is also absolutely integrable.*

(ii) *If f is not absolutely integrable, then neither is F.*

Proof. Clearly (i) $\iff$ (ii) so it suffices to prove (i). We have

$$\int_K \big|\, f(\boldsymbol{x}) \,\big|\, |d\boldsymbol{x}| \leqslant \int_K \big|\, F(\boldsymbol{x}) \,\big|\, |d\boldsymbol{x}|, \;\; \forall K \in \mathfrak{J}_c(U)$$

so

$$\sup_{K \in \mathfrak{J}_c(U)} \int_K \big|\, f(\boldsymbol{x}) \,\big|\, |d\boldsymbol{x}| \leqslant \sup_{K \in \mathfrak{J}_c(U)} \int_K \big|\, F(\boldsymbol{x}) \,\big|\, |d\boldsymbol{x}| < \infty.$$

□

15.4.3. Examples. We want to discuss a few simple but important examples.

Example 15.4.15. Fix $n \in \mathbb{N}$. For $\alpha > 0$ define

$$p_\alpha : \mathbb{R}^n \backslash \{\boldsymbol{0}\} \to \mathbb{R}, \;\; p_\alpha(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-\alpha}.$$

Suppose that U is the punctured unit ball in $\mathbb{R}^n$,

$$U = \big\{\, \boldsymbol{x} \in \mathbb{R}^n; \;\; 0 < \|\boldsymbol{x}\| < 1 \,\big\}.$$

The collection of annuli

$$K_\nu := \left\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; \frac{1}{\nu} \leqslant \|\boldsymbol{x}\| \leqslant 1 - \frac{1}{\nu} \right\}, \;\; \nu \in \mathbb{N}$$

is a Jordan measurable compact exhaustion of U. Using (15.3.26) we deduce that

$$\int_{K_\nu} p_\alpha(\boldsymbol{x}) |d\boldsymbol{x}| = n\boldsymbol{\omega}_n \int_{\frac{1}{\nu}}^{1-\frac{1}{\nu}} \rho^{n-1-\alpha} d\rho = n\boldsymbol{\omega}_n \times \begin{cases} \frac{1}{n-\alpha} \rho^{n-\alpha} \Big|_{1/\nu}^{1-1/\nu}, & \alpha \neq n, \\ \\ \log \rho \Big|_{1/\nu}^{1-1/\nu}, & \alpha = n. \end{cases}$$

The last quantity has a finite limit as $\nu \to \infty$ if and only if $\alpha < n$. We deduce that the function $p_\alpha(\boldsymbol{x})$ is absolutely convergent on U if and only if $\alpha < n$.

Fix $\beta > 0$. Consider now the complement in $\mathbb{R}^n$ of the closed unit ball,

$$V := \big\{\, \boldsymbol{x} \in \mathbb{R}^n; \;\; \|\boldsymbol{x}\| > 1 \,\big\}.$$

The collection of annuli

$$C_\nu := \left\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; 1 + \frac{1}{\nu} \leqslant \|\boldsymbol{x}\| \leqslant \nu \right\}, \;\; \nu \in \mathbb{N}$$

is a Jordan measurable compact exhaustion of V. We deduce as above that

$$\int_{C_\nu} p_\beta(\boldsymbol{x}) |d\boldsymbol{x}| = n\boldsymbol{\omega}_n \times \begin{cases} \frac{1}{n-\beta} \rho^{n-\beta} \Big|_{1+1/\nu}^{\nu}, & \beta \neq n, \\ \\ \log \rho \Big|_{1+1/\nu}^{\nu}, & \beta = n. \end{cases}$$

The last quantity has a finite limit as $\nu \to \infty$ if and only if $\beta > n$, so the function $p_\beta(\boldsymbol{x})$ is absolutely convergent on V if and only if $\beta > n$. □

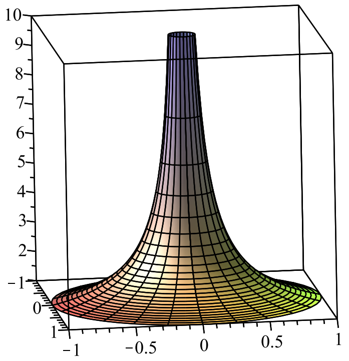

Figure 15.14. *The graph of* $p_1(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-1}$, $\boldsymbol{x} \in \mathbb{R}^2 \backslash \{\mathbf{0}\}$.

Example 15.4.16 (Gaussian integrals)**.** Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = e^{-x^2-y^2}.$$

It is locally integrable since it is continuous. To investigate if it is absolutely integrable consider the disks

$$D_\nu := \left\{ (x,y) \in \mathbb{R}^2; \ \sqrt{x^2+y^2} \leqslant \nu \right\}, \quad \nu \in \mathbb{N}.$$

Using polar coordinates $x = r\cos\theta$, $y = r\sin\theta$ we deduce

$$\int_{D_\nu} f(x,y)|dxdy| = \int_0^{2\pi} \left(\int_0^\nu e^{-r^2} rdr \right) d\theta$$

$$= 2\pi \int_0^\nu e^{-r^2} rdr = \pi \int_0^\nu e^{-r^2} d(r^2) \overset{u=r^2}{=} \pi \int_0^{\nu^2} e^{-u} du$$

$$= \pi\big(1 - e^{-\nu^2} \big).$$

Since $(D_\nu)_{\nu\in\mathbb{N}}$ is a Jordan measurable compact exhaustion of $\mathbb{R}^2$ we deduce that

$$\int_{\mathbb{R}^2} e^{-x^2-y^2} dxdy = \lim_{\nu\to\infty} \int_{D_\nu} e^{-x^2-y^2} dxdy = \pi.$$

On the other hand, if we set $S_\nu = [-\nu,\nu] \times [-\nu,\nu]$ we deduce

$$\pi = \int_{\mathbb{R}^2} e^{-x^2-y^2} |dxdy| = \lim_{\nu\to\infty} \int_{S_\nu} e^{-x^2-y^2} |dxdy|$$

$$= \lim_{\nu\to\infty} \int_{-\nu}^{\nu} \left(\int_{-\nu}^{\nu} e^{-x^2} e^{-y^2} dx \right) dy = \lim_{\nu\to\infty} \left(\int_{-\nu}^{\nu} e^{-x^2} dx \right) \left(\int_{-\nu}^{\nu} e^{-y^2} dy \right)$$

$$= \lim_{\nu\to\infty} \left(\int_{-\nu}^{\nu} e^{-x^2} dx \right)^2 = \left(\int_{\mathbb{R}} e^{-x^2} dx \right)^2.$$

We have thus obtained the following famous result

$$\boxed{\int_{\mathbb{R}} e^{-x^2}dx = \sqrt{\pi}} \tag{15.4.4}$$

In particular

$$\int_{\mathbb{R}} e^{-\frac{x^2}{2r}}dx \overset{x=\sqrt{2r}y}{=\!=} \sqrt{2r}\int_{\mathbb{R}} e^{-y^2}dy = \sqrt{2\pi r}.$$

Hence

$$\boxed{\frac{1}{\sqrt{2\pi r}}\int_{\mathbb{R}} e^{-\frac{x^2}{2r}}dx = 1, \;\; \forall r > 0}. \tag{15.4.5}$$

The last equality plays an important role in probability. The equalities (15.4.4) and (15.4.5) have a captivating history involving many illustrious names. We refer to [**32**] for a very lively presentation of its history. □

Example 15.4.17 (The volume of the unit n-dimensional ball)**.** We want to have another look at $\boldsymbol{\omega}_n$, the volume of the unit ball in $\mathbb{R}^n$. We will obtain a new description for $\boldsymbol{\omega}_n$ using an elegant trick of H. Weyl that is based on computing the integral

$$\boldsymbol{I}_n := \int_{\mathbb{R}^n} e^{-\|\boldsymbol{x}\|^2}\,|d\boldsymbol{x}|$$

in two different ways. Note first that

$$\boldsymbol{I}_n = \int_{\mathbb{R}^n} e^{-x_1^2-\cdots-x_n^2}|dx_1\cdots dx_n| = \int_{\mathbb{R}^n} e^{-x_1^2}\cdots e^{-x_n^2}|dx_1\cdots dx_n|$$

(use Fubini)

$$= \left(\int_{\mathbb{R}} e^{-x_1^2}dx_1\right)\cdots\left(\int_{\mathbb{R}} e^{-x_n^2}dx_n\right) = \left(\int_{\mathbb{R}} e^{-x^2}dx\right)^n \overset{(15.4.4)}{=} \pi^{\frac{n}{2}}.$$

On the other hand, the function $e^{-\|\boldsymbol{x}\|^2}$ is radially symmetric and we deduce from (15.3.26) that

$$\boldsymbol{I}_n = n\boldsymbol{\omega}_n\int_0^\infty e^{-\rho^2}\rho^{n-1}d\rho \overset{\rho=\sqrt{t}}{=} \frac{n\boldsymbol{\omega}_n}{2}\int_0^\infty e^{-t}t^{\frac{n}{2}-1}dt.$$

At this point we want to recall the definition of the Gamma function (9.7.7)

$$\Gamma(x) := \int_0^\infty e^{-t}t^{x-1}dt, \;\; x > 0.$$

Thus

$$\pi^{\frac{n}{2}} = \boldsymbol{I}_n = \frac{n\boldsymbol{\omega}_n}{2}\Gamma\left(\frac{n}{2}\right)$$

We deduce

$$\boldsymbol{\omega}_n = \frac{\pi^{\frac{n}{2}}}{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)}. \tag{15.4.6}$$

This can be simplified a bit by using the identity (9.7.9)

$$\Gamma(x+1) = x\Gamma(x), \;\; \forall x > 0.$$

We deduce

$$\frac{n}{2}\Gamma\big(\frac{n}{2}\big)=\Gamma\big(\frac{n}{2}+1\big),$$

and thus the volume $\boldsymbol{\omega}_n$ of the unit n-dimensional ball is

$$\boldsymbol{\omega}_n=\frac{\pi^{\frac{n}{2}}}{\Gamma\big(\frac{n}{2}+1\big)}. \tag{15.4.7}$$

To see how this relates to (15.3.24) we use again the identity (9.7.9) and we deduce

$$\Gamma(m)=m!\ \ \forall m\in\mathbb{N}$$

$$\Gamma\Big(\frac{2n+1}{2}\Big)=\frac{2n-1}{2}\Gamma\Big(\frac{2n-1}{2}\Big)=\cdots=\frac{(2n-1)!!}{2^{n-1}}\Gamma(1/2).$$

On the other hand

$$\Gamma(1/2)=\int_0^\infty e^{-t}t^{-1/2}dt\overset{t=x^2}{=}2\int_0^\infty e^{-x^2}dx\overset{(15.4.4)}{=}\sqrt{\pi}.$$

□

Example 15.4.18 (Euler's Beta function)**.** For $x,y>0$ we set

$$\boxed{B(x,y):=\int_0^1 t^{x-1}(1-t)^{y-1}dt.} \tag{15.4.8}$$

This integral is convergent since $x-1,y-1>-1$. The resulting function

$$(0,\infty)\times(0,\infty)\ni(x,y)\mapsto B(x,y)\in(0,\infty),$$

is known as *Euler's Beta function.*

If we make the change in variables in the integral (15.4.8)

$$u=\frac{t}{1-t},$$

then we observe that $u=0$ when $t=0$ and $u\to\infty$ as $t\nearrow 1$. Moreover, we have

$$(1-t)u=t\Rightarrow u=t(1+u)\Rightarrow t=\frac{u}{1+u}=1-\frac{1}{1+u}$$

$$\Rightarrow 1-t=\frac{1}{1+u},\quad dt=\frac{1}{(1+u)^2}du,$$

$$t^{x-1}(1-t)^{y-1}dt=\left(\frac{u}{1+u}\right)^{x-1}\left(\frac{1}{1+u}\right)^{y-1}\frac{1}{(1+u)^2}du=\frac{u^{x-1}}{(1+u)^{x+y}}du,$$

so that

$$\boxed{B(x,y)=\int_0^\infty\frac{u^{x-1}}{(1+u)^{x+y}}du,\quad\forall x,y>0.} \tag{15.4.9}$$

Using (9.7.11) we deduce

$$\frac{1}{(1+u)^{x+y}}=\frac{1}{\Gamma(x+y)}\int_0^\infty s^{x+y-1}e^{-(1+u)s}ds.$$

Using this in (15.4.9) we deduce

$$\begin{aligned} B(x,y) &= \frac{1}{\Gamma(x+y)}\int_0^\infty u^{x-1}\left(\int_0^\infty s^{x+y-1}e^{-(1+u)s}ds\right)du \\ &= \frac{1}{\Gamma(x+y)}\underbrace{\int_0^\infty\left(\int_0^\infty u^{x-1}s^{x+y-1}e^{-(1+u)s}ds\right)du}_{=:\boldsymbol{I}}. \end{aligned} \tag{15.4.10}$$

At this point we want to invoke Fubini's theorem which allows us to conclude that we can interchange the order of integration so that

$$\begin{aligned} \boldsymbol{I} &:= \int_0^\infty\left(\int_0^\infty u^{x-1}s^{x+y-1}e^{-(1+u)s}ds\right)du = \int_0^\infty\left(\int_0^\infty s^{x+y-1}u^{x-1}e^{-(1+u)s}du\right)ds \\ &= \int_0^\infty s^{x+y-1}\left(\int_0^\infty u^{x-1}e^{-(1+u)s}du\right)ds = \int_0^\infty s^{x+y-1}\left(\int_0^\infty u^{x-1}e^{-s}e^{-su}du\right)ds \\ &= \int_0^\infty e^{-s}s^{x+y-1}\underbrace{\left(\int_0^\infty u^{x-1}e^{-su}du\right)}_{\overset{(9.7.11)}{=}\frac{\Gamma(x)}{s^x}}ds \\ &= \int_0^\infty e^{-s}s^{x+y-1}\frac{\Gamma(x)}{s^x}ds = \Gamma(x)\int_0^\infty s^{y-1}e^{-s}ds = \Gamma(x)\Gamma(y). \end{aligned}$$

Thus

$$\boldsymbol{I} = \Gamma(x)\Gamma(y),$$

and

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \qquad \square$$

Remark 15.4.19. The Gamma and Beta functions are ubiquitous in mathematics. They belong to the category of so called *special functions*. The facts we have presented barely scratch the surface of the beautiful theory of these functions. There are many sources from which to learn about the Gamma function but, as entry point in this theory no source comes close to the little gem [**2**] by Emil Artin.[9] First published 1931 in German, it remains to the day an example of beautiful mathematical writing. □

[9] Emil Artin (1898-1962) was one of the most influential mathematicians of the 20th century. He was Professor at the University of Hamburg Germany until 1937 when he was forced to emigrate to US due to the political tensions in Germany at that time. His first US position was at the University of Notre Dame where, coincidently, the present book was also born.

15.5. Exercises

Exercise 15.1. Prove Lemma 15.1.2.

Hint. The proof is not hard, but it takes some thinking to write down a short and precise exposition. Start with the case $n = 2$ and then argue by induction on n. □

Exercise 15.2. Consider the triangle

$$T := \left\{ (x, y) \in \mathbb{R}^2;\ \ x, y \geqslant 0,\ \ x + y \leqslant 1 \right\}$$

and the function

$$f : [0,1] \times [0,1] \to \mathbb{R},\ \ f(x,y) = \begin{cases} 1, & (x,y) \in T, \\ 0, & (x,y) \notin T. \end{cases}$$

For each $n \in \mathbb{N}$ denote by $\boldsymbol{P}_n$ the partition of $[0,1] \times [0,1]$,

$$\boldsymbol{P}_n = \left(\boldsymbol{U}_n^x, \boldsymbol{U}_n^y\right),$$

where $\boldsymbol{U}_n^x$ is the uniform partition of order n of the interval $[0,1]$ on the x-axis (see Example 9.2.2) and $\boldsymbol{U}_n^y$ is the uniform partition of order n of the interval $[0,1]$ on the y-axis.

Compute $\boldsymbol{\omega}(f, \boldsymbol{P}_n)$ and then show that

$$\lim_{n \to \infty} \boldsymbol{\omega}(f, \boldsymbol{P}_n) = 0.$$

Hint. To understand what is going on investigate first the partition $\boldsymbol{P}_4$ depicted in Figure 15.15. □

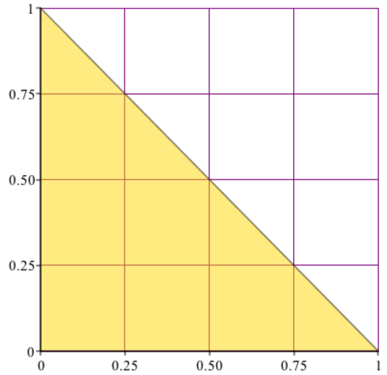

Figure 15.15. *The partition $\boldsymbol{P}_4$ of the square $[0,1] \times [0,1]$. The triangle T is described by the shaded area.*

Exercise 15.3. Let $n \in \mathbb{N}$ and suppose that $B := [a_1, b_1] \times \cdots\cdots \times [a_n, b_n] \subset \mathbb{R}^n$ is a closed nondegenerate box. For any Riemann integrable function $f : B \to \mathbb{R}$ we define its mean on B to be the real number

$$\mathrm{Mean}(f) := \frac{1}{\mathrm{vol}_n(B)} \int_B f(\boldsymbol{x}) |d\boldsymbol{x}|.$$

(i) Prove that for any Riemann integrable function $f : B \to \mathbb{R}$ we have

$$\inf_{\boldsymbol{x}\in B} f(\boldsymbol{x}) \leqslant \operatorname{Mean}(f) \leqslant \sup_{\boldsymbol{x}\in B} f(\boldsymbol{x}).$$

(ii) Prove that for any continuous function $f : B \to \mathbb{R}$ there exists $\bar{\boldsymbol{x}} \in B$ such that

$$f(\bar{\boldsymbol{x}}) = \operatorname{Mean}(f).$$

Hint. (ii) Use (i) and Corollary 12.3.3. □

Exercise 15.4. Denote by $\overline{C_r^n}$ the closed cube in $\mathbb{R}^n$ of radius r centered at $\mathbf{0}$,

$$\overline{C_r^n} = \left\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ |x^i| \leqslant r,\ \ \forall i = 1, \dots, n \right\}.$$

Suppose that $f : \overline{C_1^n} \to \mathbb{R}$ is a continuous function. Prove that

$$\lim_{r\searrow 0} \frac{1}{\operatorname{vol}_n\left(\overline{C_r^n}\right)} \int_{\overline{C_r^n}} f(\boldsymbol{x})|d\boldsymbol{x}| = f(\mathbf{0}).$$

Hint. Use Exercise 15.3(ii). □

Exercise 15.5. Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate closed box and $f, g : B \to \mathbb{R}$ are Riemann integrable functions. Fix $p, q \in (1, \infty)$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

(i) Prove that there exists $C > 0$ such that, for any partition $\boldsymbol{P}$ of B, we have

$$\omega(|f|^p, \boldsymbol{P}) < C\omega(f, \boldsymbol{P}),\ \ \omega(|g|^q, \boldsymbol{P}) < C\omega(g, \boldsymbol{P})$$

(ii) Prove that there exists a sequence of partitions $(\mathbb{P}_\nu)_{\nu\in\mathbb{N}}$ of B such that

$$\lim_{\nu\to\infty} \omega(f, \mathbb{P}_\nu) = \lim_{\nu\to\infty} \omega(g, \mathbb{P}_\nu) = 0.$$

(iii) Prove that

$$\left|\int_B f(\boldsymbol{x})g(\boldsymbol{x})|d\boldsymbol{x}|\right| \leqslant \left(\int_B |f(\boldsymbol{x})|^p|d\boldsymbol{x}|\right)^{\frac{1}{p}} \left(\int_B |g(\boldsymbol{x})|^q|d\boldsymbol{x}|\right)^{\frac{1}{q}}. \tag{15.5.1}$$

Hint. (i) Have a look at the proof of (15.1.8). (iii) Use Proposition 15.1.9 coupled with (i) and (ii) to reduce (15.5.1) to (8.3.15). □

Exercise 15.6. Fix $n \in \mathbb{N}$.

(i) Show that a subset $S \subset \mathbb{R}^n$ is negligible (see Definition 15.1.15) if and only if, for any $\varepsilon > 0$, there exists a sequence $(B_\nu)_{\nu\geqslant 1}$ of *closed* boxes such that

$$S \subset \bigcup_{\nu\geqslant 1} \boldsymbol{int}(B_\nu)\ \text{ and }\ \sum_{\nu\geqslant 1} \operatorname{vol}_n(B_\nu) < \varepsilon.$$

(ii) Show that if the *compact* subset $S \subset \mathbb{R}^n$ is negligible, then for any $\varepsilon > 0$ there exists *finitely many* closed boxes $B_1, \dots, B_N$ such that

$$S \subset \bigcup_{\nu=1}^{N} \boldsymbol{int}(B_\nu) \;\text{ and }\; \sum_{\nu=1}^{N} \operatorname{vol}_n(B_\nu) < \varepsilon.$$

Hint. (i) Observe that for any closed box B and any $\hbar > 0$ there exists a closed box B' such that $B \subset \boldsymbol{int}(B')$ and $\operatorname{vol}_n(B') - \operatorname{vol}_n(B) < \hbar$. □

Exercise 15.7. Prove that the following sets are negligible.

(i) A subset of a negligible subset.

(ii) The union of a sequence $(\mathcal{N}_\nu)_{\nu \geqslant 1}$ of negligible subsets of $\mathbb{R}^n$.

(iii) The coordinate hyperplane

$$H_t^i := \big\{ (x^1, x^2, \dots, x^n) \in \mathbb{R}^n \;\; x^i = t \big\} \subset \mathbb{R}^n,$$

where $i \in \{1, \dots, n\}$ and t is a fixed real number.

Hint. (ii) For $\nu = 1, 2, \dots$ cover $\mathcal{N}_\nu$ by a countable family of boxes $B_{\nu,1}, B_{\nu,2}, \dots$, such that

$$\sum_{k=1}^{\infty} \operatorname{vol}_n(B_{\nu,k}) < \frac{\varepsilon}{2^\nu}.$$

Then use the fact that the set $\mathbb{N} \times \mathbb{N}$ is countable; see Example 3.1.17. (iii) Use (ii). □

Exercise 15.8. Suppose that $f : [a,b] \to [0,\infty)$ and $g : [c,d] \to [0,\infty)$ are two nonnegative Riemann integrable functions. Define

$$h : [a,b] \times [c,d] \to \mathbb{R}, \;\; h(x,y) = f(x)g(y).$$

Prove that h is Riemann integrable and

$$\int_{[a,b]\times[c,d]} h(x,y)|dxdy| = \left(\int_a^b f(x)dx \right) \left(\int_c^d g(y)dy \right).$$

Hint. Choose a partition $\boldsymbol{P} = (\boldsymbol{P}_x, \boldsymbol{P}_y)$ of the box $[a,b] \times [c,d]$ and express the Darboux sum $\boldsymbol{S}_*(h, \boldsymbol{P})$ in terms of $\boldsymbol{S}_*(f, \boldsymbol{P}_x)$ and $\boldsymbol{S}_*(g, \boldsymbol{P}_y)$. Do the same for the upper Darboux sums. □

Exercise 15.9. Let $B = [0,1] \times [1,2] \subset \mathbb{R}^2$. Using Theorem 15.1.18 compute

$$\int_B \frac{1}{(x^1 + x^2)^2} \, |dx^1 dx^2|.$$

□

Exercise 15.10. Consider a nondegenerate box $B \subset \mathbb{R}^n$, a continuous function $f : B \to \mathbb{R}$ and a continuous convex function $\Phi : \mathbb{R} \to \mathbb{R}$. Prove that

$$\Phi\left(\frac{1}{\operatorname{vol}_n(B)} \int_B f(\boldsymbol{x}) \, |d\boldsymbol{x}| \right) \leqslant \frac{1}{\operatorname{vol}_n(B)} \int_B \Phi\big(f(\boldsymbol{x}) \big) |d\boldsymbol{x}|.$$

Hint. Use Riemann sums and Jensen's inequality (8.3.8). □

Exercise 15.11. (a) Let $a, b \in \mathbb{R}$, $a < b$. For any $\varepsilon > 0$ construct *explicitly* a continuous function $g_\varepsilon : [a,b] \to \mathbb{R}$ satisfying the following properties.

(i) $g_\varepsilon(a) = g_\varepsilon(b) = 0$.

(ii) $0 \leqslant g_\varepsilon(x) \leqslant 1$, $\forall x \in [a, b]$.

(iii)
$$\int_a^b dx - \varepsilon \leqslant \int_a^b g_\varepsilon(x)dx \leqslant \int_a^b dx.$$

(b) Let $n \in \mathbb{N}$ and suppose that $B = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$ is a closed nondegenerate box. Prove that, for any $\varepsilon > 0$ there exists a continuous function $h_\varepsilon : B \to \mathbb{R}$ satisfying the following conditions.

(i) $h_\varepsilon(\boldsymbol{x}) = 0$ for any point $\boldsymbol{x}$ the boundary of B, i.e., a point $\boldsymbol{x} = (x^1, \dots, x^n)$ such that $x^i = a_i$ or $x^i = b_i$ for some $i = 1, \dots, n$.

(ii) $0 \leqslant h_\varepsilon(\boldsymbol{x}) \leqslant 1$, $\forall \boldsymbol{x} \in B$.

(iii)
$$\int_B |d\boldsymbol{x}| - \varepsilon \leqslant \int_B h_\varepsilon(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \int_B |d\boldsymbol{x}|.$$

Hint. (a) Think of a function whose graph looks like a trapezoid. (b) Seek h_ε of the form
$$h_\varepsilon(x^1, \dots, x^n) = g_\varepsilon^1(x^1) \cdots g_\varepsilon^n(x^n),$$
where $g_\varepsilon : [a_i, b_i] \to \mathbb{R}$ are chosen as in (a). Use Fubini to reach the desired conclusion. □

Exercise 15.12. Let $n \in \mathbb{N}$ and suppose that $B = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$ is a closed nondegenerate box. Suppose that $f : B \to [0, \infty)$ is a *nonnegative* Riemann integrable function. Prove that, for any $\varepsilon > 0$ there exists a continuous function $h_\varepsilon : B \to \mathbb{R}$ satisfying the following conditions.

(i) $h_\varepsilon(\boldsymbol{x}) = 0$ for any point $\boldsymbol{x}$ on the boundary of B, i.e., a point $\boldsymbol{x} = (x^1, \dots, x^n)$ such that $x^i = a_i$ or $x_i = b_i$ for some $i = 1, \dots, n$.

(ii) $0 \leqslant h_\varepsilon(\boldsymbol{x}) \leqslant f(\boldsymbol{x})$, $\forall \boldsymbol{x} \in B$.

(iii)
$$\int_B f(\boldsymbol{x})|d\boldsymbol{x}| - \varepsilon \leqslant \int_B h_\varepsilon(\boldsymbol{x})|d\boldsymbol{x}| \leqslant \int_B f(\boldsymbol{x})|d\boldsymbol{x}|.$$

Hint. Choose a partition $\boldsymbol{P}$ of B such that the mean oscillation $\omega(f, \boldsymbol{P})$ is very small. Next use Exercise 15.11 (b) on each chamber of the partition $\boldsymbol{P}$. □

Exercise 15.13. Let $n, N \in \mathbb{N}$ and suppose that $S_1, \dots, S_N \subset \mathbb{R}^n$ are Jordan measurable sets. Prove that their union is also Jordan measurable and
$$\operatorname{vol}_n\left(\bigcup_{k=1}^N S_k\right) \leqslant \sum_{k=1}^N \operatorname{vol}_n(S_k).$$

Hint. Argue by induction using Proposition 15.1.30. □

Exercise 15.14. (a) Suppose that $K \subset \mathbb{R}^n$ is compact. Prove that K negligible if and only if it is Jordan measurable and $\operatorname{vol}_n(K) = 0$.

(b) Suppose that $B \subset \mathbb{R}^n$ is a *closed* nondegenerate box. Prove that $\boldsymbol{int}(B)$ is Jordan measurable and

$$\operatorname{vol}_n\big(\, \boldsymbol{int}(B)\,\big) = \operatorname{vol}_n(B).$$

Hint. (a) If K is negligible use Corollary 15.1.29 to prove that it is Jordan measurable. To show that $\operatorname{vol}_n(K) = 0$ use Exercises 15.6 and 15.13. Conversely, suppose that K is Jordan measurable and $\operatorname{vol}_n(K) = 0$. Choose a box B containing K. Then $\int_B I_K(\boldsymbol{x})|d\boldsymbol{x}| = 0$. Thus for $\varepsilon > 0$ one can choose a partition $\boldsymbol{P}_\varepsilon$ of B such that $S^*(I_K, \boldsymbol{P}_\varepsilon) < \varepsilon$. Use this partition to show that you can cover K by finitely many boxes whose volumes add up to less than ε. (b) Invoke Proposition 15.1.21 and Lebesgue's Theorem to conclude that ∂B is negligible and then use (a). □

Exercise 15.15. Suppose that $f : [a, b] \times [a, b] \to \mathbb{R}$ is a continuous function. Prove that

$$\int_a^b dy \int_a^y f(x, y)\, dx = \int_a^b dx \int_x^b f(x, y)\, dy.$$

Hint. Compute the double integral $\int_C f(x, y)dxdy$ in two different ways for a suitable Jordan measurable set $C \subset [a, b] \times [a, b]$. □

Exercise 15.16. Denote by T the triangle in the plane determined by the lines

$$y = 2x, \quad y = \frac{x}{2}, \quad x + y = 6.$$

(i) Find the coordinates of the vertices of this triangle.

(ii) Draw a picture of this triangle.

(iii) Compute the integral

$$\int_T xy\,|dxdy|.$$

□

Exercise 15.17. Find the area of the region $R \subset \mathbb{R}^2$ defined as the intersection of the disks of radius 1 centered at the vertices of the square $[0, 1] \times [0, 1]$; see Figure 15.16.

Exercise 15.18. Consider the box $B = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ and suppose that $f : B \to \mathbb{R}$ is continuous function. Define

$$F : B \to \mathbb{R}, \quad F\big(x, y\big) = \int_{[a_1, x] \times [a_2, y]} f(s, t)|dsdt|.$$

Show that the restriction of F to the interior of B is C^1 and then compute the partial derivatives $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$. □

Exercise 15.19. Suppose that $B \subset \mathbb{R}^n$ is a closed nondegenerate box and $f : B \to \mathbb{R}$ is a continuous function such that $f(\boldsymbol{x}) \geqslant 0$, $\forall \boldsymbol{x} \in B$. Prove that the following statements are equivalent.

(i) $f(\boldsymbol{x}) = 0$, $\forall \boldsymbol{x} \in B$.

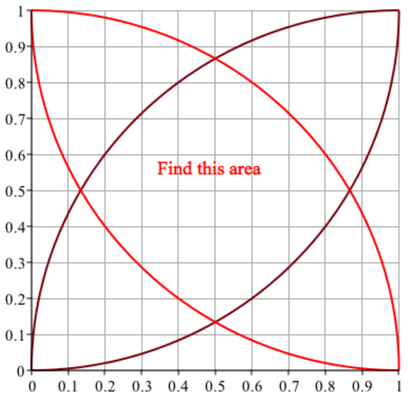

Figure 15.16. *The overlap of 4 disks centered at the vertices of a square.*

(ii)
$$\int_B f(\boldsymbol{x})\,|d\boldsymbol{x}| = 0.$$

□

Exercise 15.20. Let $n \in \mathbb{N}$ and $r > 0$.

(i) Prove that the n-dimensional closed ball
$$\overline{B_r^n(\mathbf{0})} := \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| \leqslant r \,\big\},$$
is Jordan measurable.

(ii) Prove that the sphere
$$\Sigma_r(\mathbf{0}) := \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| = r \,\big\},$$
is Jordan measurable and $\operatorname{vol}_n(\Sigma_r(\mathbf{0})) = 0$.

(iii) Prove that the n-dimensional open ball
$$B_r^n(\mathbf{0}) := \big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x}\| < r \,\big\},$$
is Jordan measurable and
$$\operatorname{vol}_n\big(\, B_r^n(\mathbf{0})\,\big) = \operatorname{vol}_n\big(\,\overline{B_r^n(\mathbf{0})}\,\big)$$

Hint. (i) Use induction on n and Proposition 15.2.2. (ii) Use Corollary 15.1.29. Conclude using Exercise 15.14 and Proposition 15.1.30(i). (iii) Follows from (i) and (ii). □

Exercise 15.21. For any $n \in \mathbb{N}$ and $r > 0$ denote by $\boldsymbol{\omega}_n(r)$ the n-dimensional volume of the n-dimensional ball $B_r^n(\mathbf{0}) \subset \mathbb{R}^n$. For simplicity, set $\boldsymbol{\omega}_n := \boldsymbol{\omega}_n(1)$, so that $\boldsymbol{\omega}_n$ is the volume of the unit ball in $\mathbb{R}^n$.

(i) Show that $\boldsymbol{\omega}_n(r) = \boldsymbol{\omega}_n r^n$.

(ii) Use Cavalieri's Principle to show that

$$\boldsymbol{\omega}_n := 2\boldsymbol{\omega}_{n-1}\int_0^1 \big(1-x^2\big)^{\frac{n-1}{2}}dx.$$

(iii) Prove that $\boldsymbol{\omega}_n$ satisfies the equalities (15.3.24).

Hint. (i) Use the change-in-variables formula for the diffeomorphism $\Phi:\mathbb{R}^n\to\mathbb{R}^n$, $\Phi(\boldsymbol{x})=r\boldsymbol{x}$ that sends $B_1(\mathbf{0})$ onto $B_r(\mathbf{0})$. (iii) Relate the integral in (ii) to the integrals (9.6.12) and (9.6.13). Then proceed by induction on n.□

Exercise 15.22. Prove that Lebesgue's Theorem 15.1.17 implies Corollary 15.1.37. □

Exercise 15.23. Fix a continuous function $f:\mathbb{R}\to\mathbb{R}$ and define recursively the sequence of functions $F_n:\mathbb{R}\to\mathbb{R}$, $n=0,1,2\dots,$

$$F_0(x)=f(x),\;\; F_n(x)=\frac{1}{(n-1)!}\int_0^x(x-y)^{n-1}f(y)dy,\;\; n\geqslant 1.$$

(i) Show that, for all $n\in\mathbb{N}$, $F_n\in C^1(\mathbb{R})$, $F_n(0)=0$ and $F_n'(x)=F_{n-1}(x)$, $\forall x\in\mathbb{R}$.

(ii) Show that if $(G_n)_{n\geqslant 1}$ is a sequence of C^1-functions on $\mathbb{R}$ such that

$$G_n(0)=0,\;\; \forall n\geqslant 1,\;\; G_1'(x)=f(x),$$
$$G_{n+1}'(x)=G_n(x),\;\; \forall n\geqslant 1,\;\; x\in\mathbb{R},$$

then $G_n(x)=F_n(x)$, $\forall n\geqslant 1$, $x\in\mathbb{R}$.

(iii) Show that

$$F_n(x)=\int_0^x dx^1\int_0^{x^1}dx^2\cdots\int_0^{x^{n-1}}f(x^n)dx^n.$$

□

Exercise 15.24. Suppose that $K\subset\mathbb{R}^2$ is a compact Jordan measurable set that is symmetric with respect to the y-axis, i.e.,

$$(x,y)\in K\Longleftrightarrow(-x,y)\in K.$$

Let $f:\mathbb{R}^2\to\mathbb{R}$ be a continuous function that is odd in the x-variable, i.e.,

$$f(-x,y)=-f(x,y),\;\; \forall(x,y)\in K.$$

(i) Prove that

$$\int_K f(x,y)|dxdy|=0.$$

(ii) Let

$$D:=\big\{(x,y)\in\mathbb{R}^2;\;\; x^2+y^2\leqslant 1\big\}.$$

Compute

$$\int_D x^{23}y^{24}|dxdy|\;\;\text{and}\;\;\int_D(xy)^{24}|dxdy|.$$

Hint. (i) Use the change-in-variables formula. For (ii) you need to use the equalities (9.6.15) in Example 9.6.8. □

Exercise 15.25. Let $n \in \mathbb{N}$ and suppose that $K \subset \mathbb{R}^n$ is a compact, Jordan measurable set.

(i) Suppose that $L : \mathbb{R}^n \to \mathbb{R}^n$ is a linear isomorphism. Prove that

$$\operatorname{vol}_n\big(\,L(K)\,\big) = |\det L|\operatorname{vol}_n(K).$$

(ii) Suppose that $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear isometry, i.e., A is a linear operator and

$$\langle A\boldsymbol{x}, A\boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle,\ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Show that

$$\operatorname{vol}_n\big(\,A(K)\,\big) = \operatorname{vol}_n(K).$$

(iii) Let $\boldsymbol{v} \in \mathbb{R}^n$ and define $T_{\boldsymbol{v}} : \mathbb{R}^n \to \mathbb{R}^n$, $T_{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{v} + \boldsymbol{x}$. Show that

$$\operatorname{vol}_n\big(\,T_{\boldsymbol{v}}(K)\,\big) = \operatorname{vol}_n(K).$$

Hint. (ii) Use Exercise 11.25 and (i). □

Exercise 15.26. Let $n \in \mathbb{N}$ and suppose that $a_1, a_2, \dots, , a_n > 0$. Consider the ellipsoid

$$\Sigma(a_1, \dots, a_n) := \Big\{\, \boldsymbol{x} \in \mathbb{R}^n;\ \sum_{j=1}^n \frac{(x^j)^2}{a_j^2} \leqslant 1 \,\Big\}$$

(i) Prove that

$$\operatorname{vol}_n\big(\,\Sigma(a_1, \dots, a_n)\,\big) = \boldsymbol{\omega}_n a_1 \cdots a_n,$$

where $\boldsymbol{\omega}_n$ is the volume of the unit ball in $\mathbb{R}^n$.

(ii) Show that

$$\int_{\Sigma(a_1,\dots,a_n)} x^1 x^2 \cdots x^n \,|dx^1 \cdots dx^n| = 0.$$

Hint. (i) Make the change in variables $x^i = a_i y^i$. (ii) Make the change in variables $x^1 = -y^1, x^2 = y^2, \dots, x^n = y^n$. □

Exercise 15.27. Let $m, n \in \mathbb{R}$ and suppose that $\rho : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a continuous function. We denote by $\boldsymbol{t} = (t^1, \dots, t^m)$ the Euclidean coordinates in $\mathbb{R}^m$ and by $\boldsymbol{x} = (x^1, \dots, x^n)$ the Euclidean coordinates on $\mathbb{R}^n$. Fix a Riemann integrable function $f : \mathbb{R}^n \to \mathbb{R}$. (Note in particular that f must have compact support.) Define

$$\hat{f} : \mathbb{R}^m \to \mathbb{R},\ \ \hat{f}(\boldsymbol{t}) = \int_{\mathbb{R}^n} \rho(\boldsymbol{t}, \boldsymbol{x}) f(\boldsymbol{x}) |d\boldsymbol{x}|.$$

(i) Show that $\hat{f}$ is continuous.

(ii) Suppose additionally that ρ is C^1. Show that $\hat{f}$ is also C^1 and

$$\frac{\partial \hat{f}}{\partial t^k}(\boldsymbol{t}) = \int_{\mathbb{R}^n} \frac{\partial \rho}{\partial t^k}(\boldsymbol{t}, \boldsymbol{x}) f(\boldsymbol{x}) |d\boldsymbol{x}|.$$

Hint. Fix (closed) box B such that $\operatorname{supp} f \subset B$. (i) Prove that if $\boldsymbol{t}_\nu \to \boldsymbol{t}$ as $\nu \to \infty$, then the functions $\boldsymbol{x} \mapsto \rho(\boldsymbol{t}_\nu, \boldsymbol{x}) f(\boldsymbol{x})$ converge uniformly on B to the function $\boldsymbol{x} \mapsto \rho(\boldsymbol{t}, \boldsymbol{x}) f(\boldsymbol{x})$. Conclude using Proposition 15.1.14. (ii) Use Lagrange's Mean Value Theorem 13.3.8 and the fact that the partial derivatives of ρ are bounded on the compact subsets of $\mathbb{R}^m \times \mathbb{R}^n$. □

Exercise 15.28. Suppose that $\rho : \mathbb{R}^n \to \mathbb{R}$ is a *nonnegative, compactly supported* continuous function such that

$$\int_{\mathbb{R}^n} \rho(\boldsymbol{x})\, |d\boldsymbol{x}| = 1. \tag{15.5.2}$$

Given a Riemann integrable function f and $\varepsilon > 0$ we define $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$,

$$f_\varepsilon(\boldsymbol{x}) = \varepsilon^{-n} \int_{\mathbb{R}^n} \rho\big(\varepsilon^{-1}(\boldsymbol{x} - \boldsymbol{y})\big) f(\boldsymbol{y})\, |d\boldsymbol{y}|.$$

(i) Prove that

$$f_\varepsilon(\boldsymbol{x}) = \varepsilon^{-n} \int_{\mathbb{R}^n} \rho\big(\varepsilon^{-1}\boldsymbol{z}\big) f(\boldsymbol{x} - \boldsymbol{z})\, |d\boldsymbol{z}| = \int_{\mathbb{R}^n} \rho(\boldsymbol{z}) f(\boldsymbol{x} - \varepsilon\boldsymbol{z})\, |d\boldsymbol{z}|.$$

(ii) Prove that the function f_ε is continuous and compactly supported.

(iii) Show that

$$\int_{\mathbb{R}^n} f_\varepsilon(\boldsymbol{x})\, |d\boldsymbol{x}| = \int_{\mathbb{R}^n} f(\boldsymbol{x})\, |d\boldsymbol{x}|.$$

(iv) Show that if, additionally, f is continuous, then

$$\lim_{\varepsilon \to 0} f_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}.$$

Hint. (i) Use the change in variables formula. (ii) Use Exercise 15.27(i). (iii) Use (i), (15.5.2) and Fubini. (iv) Use (15.5.2) and Proposition 15.1.14. □

Exercise 15.29. Show that the improper integral

$$\int_{0<x<y} (y^2 - x^2) e^{-y} dx dy$$

is absolutely convergent and then compute its value.

Hint. First draw the region $R := \{0 < x < y\}$. Note that in the region R the function $f(x,y) = (y^2 - x^2)e^{-y}$ is positive. Consider the compact exhaustion

$$K_\nu := \big\{ (x,y) \in \mathbb{R}^2;\ \ 1/\nu \leqslant x,\ y - x \geqslant 1/\nu,\ \ y \leqslant \nu \big\},\ \ \nu \in \mathbb{N},$$

compute the integrals

$$I_\nu = \int_{K_\nu} (y^2 - x^2) e^{-y} |dxdy|,$$

and study the limit of I_ν as $\nu \to \infty$. □

Exercise 15.30. Fix $n \in \mathbb{N}$ and a continuous function $f : \mathbb{R} \to \mathbb{R}$. For $c > 0$ we denote by T_n^c the simplex

$$T_n^c := \big\{ \boldsymbol{x} \in \mathbb{R}^n;\ \ x^1, \dots, x^n \geqslant 0,\ \ x^1 + \cdots + x^n \leqslant c \big\}$$

(i) Prove that

$$\operatorname{vol}_n(T_n^c) = \frac{c^n}{n!}$$

(ii) Show that

$$\int_{T_n^c} f(x^1 + \cdots + x^n)\, |d\boldsymbol{x}| = \frac{1}{(n-1)!} \int_0^c f(t) t^{n-1} dt.$$

(iii) Show that the integral

$$\int_{x^1,\dots,x^n \geqslant 0} \sin\big(\pi(x^1 + \cdots + x^n)\big)\, |d\boldsymbol{x}|$$

is not absolutely convergent.

Hint. (i) Reduce to Example 15.2.7 via the change in variables $x^i = cy^i$, $i = 1, \dots, n$. (ii) Make the change in variables $u^1 = x^1, \dots, u^{n-1} = x^{n-1}, u^n = x^1 + \cdots + x^n$ and then use Fubini coupled with (i). For (iii) use (ii). □

Exercise 15.31. Let $n \in \mathbb{N}$, $n > 1$. Show that, for any $R > 0$ the n-dimensional improper integral

$$\int_{0 < \|\boldsymbol{x}\| \leqslant R} \ln \|\boldsymbol{x}\|\, |d\boldsymbol{x}|$$

is absolutely convergent and then compute its value. □

Exercise 15.32. Prove the equalities (15.3.21). □

15.6. Exercises for extra credit

Chapter 16

Integration over submanifolds

We begin here the study of integration over "curved" regions of $\mathbb{R}^n$. This is a rather elaborate theory belonging properly to the area of mathematics called *differential geometry.* Time constraints prevent us from covering it in all its details and generality. Think of this as a first and low dimensional encounter with the subject.

16.1. Integration along curves

16.1.1. Integration of functions along curves. We start with a simpler situation. Fix $n \in \mathbb{N}$ and suppose that we are given a *convenient* C^1-curve $C \subset \mathbb{R}^n$, i.e., a curve (1-dimensional submanifold) that is the image of an *injective, immersion* $\boldsymbol{\alpha} : (a,b) \to \mathbb{R}^n$. Recall that $\boldsymbol{\alpha}$ is an immersion if $\boldsymbol{\alpha}'(t) \neq \mathbf{0}$, $\forall t \in (a,b)$. We think of $\boldsymbol{\alpha}(t)$ as describing the position at time t of a moving particle. The fact that $\boldsymbol{\alpha}$ is an immersion signifies that the particle does not double back. We will refer to a map $\boldsymbol{\alpha}$ with the above properties as a *parametrization* of the convenient curve.

During a tiny time interval $[t_0, t_0 + dt]$ the particle travels from $\boldsymbol{\alpha}(t_0)$ to $\boldsymbol{\alpha}(t_0 + dt)$. The arc of the curve C from $\boldsymbol{\alpha}(t_0)$ to $\boldsymbol{\alpha}(t_0 + dt)$ "does not bend too much" during the infinitesimal period of time dt so the distance $ds(t_0)$ covered by the particle along C can be approximated by the length of the line segment joining $\boldsymbol{\alpha}(t_0)$ to $\boldsymbol{\alpha}(t_0 + dt)$ i.e.,

$$ds(t_0) \approx \|\boldsymbol{\alpha}(t_0 + dt) - \boldsymbol{\alpha}(t_0)\| \approx \|\boldsymbol{\alpha}'(t_0)\| \, |dt|.$$

Thus, the length of C, viewed as the total distance covered by the particle ought to be

$$\operatorname{length}(C) \overset{?}{=} \int_a^b \|\boldsymbol{\alpha}'(t)\| \, |dt|.$$

Why the question mark? Intuition tells us that the length of a curve should be independent of the way a particle travels along it without backtracking. The right-hand side seems to depend on such travel as expressed by the parametrization $\boldsymbol{\alpha}$.

To deal with this issue we need to answer the following concrete question. Suppose that $\boldsymbol{\beta} : (c,d) \to \mathbb{R}^n$ is another parametrization of the curve C; see Figure 16.1. Can we conclude that

$$\int_a^b \|\boldsymbol{\alpha}'(t)\|\,|dt| = \int_c^d \|\boldsymbol{\beta}'(\tau)\|\,|d\tau|? \tag{16.1.1}$$

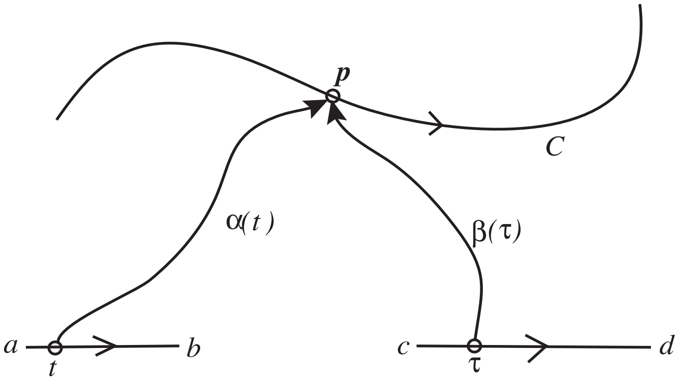

Figure 16.1. *Different parametrizations of the same convenient curve C.*

A point $\boldsymbol{p} \in C$ corresponds uniquely via $\boldsymbol{\alpha}$ to a point $t \in (a,b)$. Via $\boldsymbol{\beta}$ the point $\boldsymbol{p}$ corresponds uniquely to a point $\tau \in (c,d)$. More precisely we have

$$\boldsymbol{\alpha}(t) = \boldsymbol{p} = \boldsymbol{\beta}(\tau).$$

The correspondence

$$t \mapsto \boldsymbol{p} \mapsto \tau = \tau(t),$$

produces a bijection $(a,b) \to (c,d)$ that can be described formally by the equality $\tau = \beta^{-1}\big(\alpha(t)\big)$. Proposition 14.5.4(B) implies that the function $(a,b) \ni t \mapsto \tau(t) \in (c,d)$ is C^1.

The change in variables formula for the Riemann integral implies

$$\int_c^d \|\boldsymbol{\beta}'(\tau)\| d\tau = \int_a^b \|\boldsymbol{\beta}'(\,\tau(t)\,)\| \cdot \left|\frac{d\tau}{dt}\right| dt.$$

Derivating with respect to t the equality $\boldsymbol{\beta}\big(\,\tau(t)\,\big) = \boldsymbol{\alpha}(t)$ we deduce

$$\boldsymbol{\beta}'(\,\tau(t)\,)\frac{d\tau}{dt} = \boldsymbol{\alpha}'(t). \tag{16.1.2}$$

Using the equalities

$$\boxed{ds(\tau) = \|\boldsymbol{\beta}'(\tau)\||d\tau|, \;\; ds(t) = \|\boldsymbol{\alpha}'(t)\||dt|, \;\; |d\tau| = \left|\frac{d\tau}{dt}\right| |dt|},$$

we deduce from (16.1.2) that $ds(\tau) = ds(t)$. For this reason we will rewrite the equality

$$\operatorname{length}(C) = \int_a^b \|\boldsymbol{\alpha}'(t)\|\,|dt| = \int_c^d \|\boldsymbol{\beta}'(\tau)\|\,|d\tau| \tag{16.1.3}$$

in the simpler form

$$\boxed{\operatorname{length}(C) = \int_C ds}.$$

More generally, the same argument as above shows that, if $f : C \to \mathbb{R}$ is a continuous function, then we have the equality

$$f\big(\boldsymbol{\alpha}(t)\big)\|\boldsymbol{\alpha}'(t)\|dt = f\big(\boldsymbol{\beta}(\tau)\big)\|\boldsymbol{\beta}'(\tau)\|d\tau$$

and we set

$$\boxed{\int_C f(\boldsymbol{p})ds := \int_a^b f\big(\boldsymbol{\alpha}(t)\big)\|\boldsymbol{\alpha}'(t)\|\,|dt| = \int_c^d f\big(\boldsymbol{\beta}(\tau)\big)\|\boldsymbol{\beta}'(\tau)\|\,|d\tau|}. \tag{16.1.4}$$

The integral in the left-hand side of (16.1.4) is called the *integral of f along the curve C*. This type of integral is traditionally known as *line integral of the first kind.*

☞ *The integrals in the right-hand side of (16.1.4) could be improper integrals and, as such, they may or may not be convergent. We say that the integral over the convenient curve C is well defined if the integrals in the right-hand side of (16.1.4) are convergent.*

Example 16.1.1. Consider the arc of helix H in $\mathbb{R}^3$ described by the parametrization (see Figure 16.2)

$$\boldsymbol{\alpha} : (0,1) \to \mathbb{R}^3,\quad \boldsymbol{\alpha}(t) = \big(\cos(4\pi t), \sin(4\pi t), 2t\big).$$

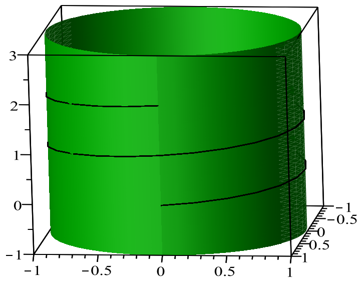

Figure 16.2. *The helix described by the parametrization* $\big(\cos(4\pi t), \sin(4\pi t), 2t\big)$ *is winding up a cylinder of radius* 1.

It is not hard to see that $\boldsymbol{\alpha}$ is an injective immersion. Moreover

$$\dot{\boldsymbol{\alpha}}(t) = \big(-4\pi\sin(4\pi t), 4\pi\sin(4\pi t), 2\big),$$

$$\|\dot{\boldsymbol{\alpha}}(t)\| = \sqrt{(4\pi\sin 4\pi t)^2 + (4\pi\cos 4\pi t)^2 + 2^2} = \sqrt{16\pi^2+4} = 2\sqrt{4\pi^2+1}.$$

We deduce that

$$\operatorname{length}(H) = \int_H ds = \int_0^1 \|\dot{\boldsymbol{\alpha}}(t)\|dt = 2\sqrt{4\pi^2+1}. \qquad \square$$

Example 16.1.2. Suppose that $f : [a,b] \to \mathbb{R}$ is a C^1 function. Its graph with endpoints removed is the set

$$\Gamma_f = \big\{ (x,y) \in \mathbb{R}^2;\ \ x \in (a,b),\ \ y = f(x) \big\}.$$

This is a curve that admits the parametrization

$$\boldsymbol{\alpha} : (a,b) \to \mathbb{R}^2,\ \ \boldsymbol{\alpha}(x) = \big(x,\ f(x)\big),\ \ x \in (a,b).$$

Then

$$\boldsymbol{\alpha}'(x) = \big(1, f'(x)\big),\ \ \|\boldsymbol{\alpha}'(x)\| = \sqrt{1+f'(x)^2}.$$

We deduce that the length of Γ_f is given by

$$\operatorname{length}(\Gamma_f) = \int_a^b \sqrt{1+f'(x)^2}dx.$$

This is in perfect agreement with our earlier definition (9.8.1). □

If the curve C is the union of pairwise disjoint convenient curves $C_1,\dots,C_\ell$, then, for any continuous function $f : C \to \mathbb{R}$ we set

$$\boxed{\int_C f(\boldsymbol{p})ds = \int_{\bigcup_{j=1}^{\ell} C_j} f(\boldsymbol{p})ds = \sum_{j=1}^{\ell}\int_{C_j} f(\boldsymbol{p})ds}.$$

Remark 16.1.3 (A few points don't matter)**.** Suppose that $C \subset \mathbb{R}^n$ is a convenient curve and $\boldsymbol{p}_0$ is a point on C. If we remove the point $\boldsymbol{p}_0$ we obtain a new curve C' consisting of two arcs C_0, C_1 of the original curve; see Figure 16.3.

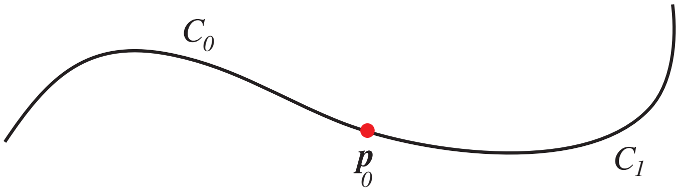

Figure 16.3. *Cutting a convenient curve C in two parts by removing a point $\boldsymbol{p}_0$.*

We want to show that the removal of one point does not affect the computation of an integral, i.e., if $f : C \to \mathbb{R}$ is a continuous function, then

$$\int_C f(\boldsymbol{p})ds = \int_{C'} f(\boldsymbol{p})ds := \int_{C_0} f(\boldsymbol{p})ds + \int_{C_1} f(\boldsymbol{p})ds.$$

Indeed, fix a parametrization $\boldsymbol{\alpha} : (a,b) \to \mathbb{R}^n$ of C. Then, there exists $t_0 \in (a,b)$ such that $\boldsymbol{\alpha}(t_0) = \boldsymbol{p}_0$. Assume that C_0 is the curve swept by the moving point $\boldsymbol{\alpha}(t)$ as t runs from a to t_0 and C_1 is swept when t runs from t_0 to b. Then

$$\int_C f(\boldsymbol{p})ds = \int_a^b f\big(\boldsymbol{\alpha}(t)\big)\|\dot{\boldsymbol{\alpha}}(t)\|dt$$

$$= \int_a^{t_0} f\big(\boldsymbol{\alpha}(t)\big)\|\dot{\boldsymbol{\alpha}}(t)\|dt + \int_{t_0}^b f\big(\boldsymbol{\alpha}(t)\big)\|\dot{\boldsymbol{\alpha}}(t)\|dt$$

$$= \int_{C_0} f(\boldsymbol{p})ds + \int_{C_1} f(\boldsymbol{p})ds =: \int_{C'} f(\boldsymbol{p})ds. \qquad \square$$

Definition 16.1.4. A curve C is called *quasi-convenient* if it admits a *convenient cut*, i.e., a finite subset $Z = \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_\ell\} \subset C$ whose complement $C \backslash \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_\ell\}$ is a union of finitely many pairwise disjoint convenient curves. □

Example 16.1.5. The unit circle in $\mathbb{R}^2$,

$$C := \big\{(x,y) \in \mathbb{R}^2;\ \ x^2 + y^2 = 1\big\}$$

is quasi-convenient. Indeed, the set consisting of the point $(1,0)$ is a convenient cut because its complement admits the parametrization

$$\boldsymbol{\alpha} : (0, 2\pi) \to \mathbb{R}^2,\ \ \boldsymbol{\alpha}(t) = \big(\cos t, \sin t\big). \qquad \square$$

Suppose now that C is a quasi-convenient curve and $f : C \to \mathbb{R}$ is a continuous function. The integral of f along C, denoted by

$$\int_C f(\boldsymbol{p})ds$$

is defined as follows. Choose a convenient cut Z. Then the complement $C_Z = C \backslash Z$ is a union of finitely many pairwise disjoint convenient curves and then we set

$$\boxed{\int_C f(\boldsymbol{p})ds := \int_{C_Z} f(\boldsymbol{p})ds}.$$

To show that this definition is independent of the convenient cut Z, suppose that Z_0, Z_1 are two convenient cuts. Then $Z := Z_0 \cup Z_1$ is also a convenient cut and $Z_0, Z_1 \subset Z$. Note that C_Z is obtained from either C_{Z_0} or C_{Z_1} by removing a few points. From Remark 16.1.3 we deduce

$$\int_{C_{Z_0}} f(\boldsymbol{p})ds = \int_{C_Z} f(\boldsymbol{p})ds = \int_{C_{Z_1}} f(\boldsymbol{p})ds.$$

Example 16.1.6. Suppose that C is the unit circle in $\mathbb{R}^2$

$$C := \big\{(x,y) \in \mathbb{R}^2;\ \ x^2 + y^2 = 1\big\}.$$

Denote by Z the convenient cut $Z = \{(1,0)\}$. Then

$$\operatorname{length}(C) = \operatorname{length}(C_Z).$$

The complement C_Z admits the parametrization

$$\boldsymbol{\alpha} : (0, 2\pi) \to \mathbb{R}^2, \quad \boldsymbol{\alpha}(t) = \big(\cos t, \sin t \big).$$

We deduce

$$\dot{\boldsymbol{\alpha}}(t) = \big(-\sin t, \cos t \big),$$

$$\operatorname{length}(C_Z) = \int_0^{2\pi} \|\dot{\boldsymbol{\alpha}}(t)\| dt = \int_0^{2\pi} \sqrt{\sin^2 t + \cos^2 t} dt = 2\pi. \qquad \square$$

Definition 16.1.7 (Curves with boundary)**.** Let $k, n \in \mathbb{N}$. A C^k*-curve with boundary* in $\mathbb{R}^n$ is a <u>compact</u> subset $C \subset \mathbb{R}^n$ such that, for any point $\boldsymbol{p}_0 \in C$, there exists an open neighborhood $\mathcal{U}$ of $\boldsymbol{p}_0$ in $\mathbb{R}^n$ and a C^k-diffeomorphism $\Psi : \mathcal{U} \to \mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}$ with the following property: either the image of $\mathcal{U} \cap C$ is an interval (a, b) on the x^1-axis, $a < 0 < b$,

$$\Psi\big(\mathcal{U} \cap C \big) = (a, b) \times \mathbf{0}_{n-1} \in \mathbb{R} \times \mathbb{R}^{n-1}, \quad \Psi(\boldsymbol{p}_0) = (0, \mathbf{0}_{n-1}) \in \mathbb{R} \times \mathbb{R}^{n-1}, \tag{I}$$

or, the image of $\mathcal{U} \cap C$ is an interval $(a, 0]$ on the x^1-axis, $a < 0$,

$$\Psi\big(\mathcal{U} \cap C \big) = (a, 0] \times \mathbf{0}_{n-1} \in \mathbb{R} \times \mathbb{R}^{n-1}, \quad \Psi(\boldsymbol{p}_0) = (0, \mathbf{0}_{n-1}) \in \mathbb{R} \times \mathbb{R}^{n-1}. \tag{B}$$

The pair $(\mathcal{U}, \Psi)$ is called a (local) *straightening diffeomorphism* (or s.d. for brevity) at $\boldsymbol{p}_0$.

If the alternative (B) occurs for some choice of straightening diffeomorphism, then we say that $\boldsymbol{p}_0 \in C$ is a *boundary point.* The *boundary* of a C^k-curve with boundary C, denoted by ∂C, is the (possibly empty) subset of C consisting of the boundary points. The curve C is called *closed* if $\partial C = \emptyset$.

A point $\boldsymbol{p}_0 \in C$ is called an *interior point* if it is not a boundary point, i.e., the alternative (I) holds for any choice of straightening diffeomorphism. The *interior* of C, denoted by C°, is the collection of interior points,

$$C^0 = C \backslash \partial C. \qquad \square$$

Example 16.1.8. (a) Suppose that $f : [a, b] \to \mathbb{R}$ is a C^1 function. Then its graph

$$\Gamma_f := \big\{ (x, y) \in \mathbb{R}^2; \;\; x \in [a, b], \;\; y = f(x) \big\}$$

is a curve with boundary. Its boundary consists of the endpoints of the graph

$$\partial \Gamma_f = \big\{ (a, f(a)), \; (b, f(b)) \big\}.$$

(b) More generally, suppose that

$$\boldsymbol{\alpha} : [a, b] \to \mathbb{R}^n, \quad \boldsymbol{\alpha}(t) = \begin{bmatrix} \boldsymbol{\alpha}^1(t) \\ \boldsymbol{\alpha}^2(t) \\ \vdots \\ \boldsymbol{\alpha}^n(t) \end{bmatrix}$$

is an *injective* map such that each of the components $\boldsymbol{\alpha}^i$ is a C^1-function and, $\forall t \in [a, b]$

$$\dot{\boldsymbol{\alpha}}(t) \neq \mathbf{0}.$$

Then the image of $\boldsymbol{\alpha}$ is a curve C with boundary consisting of two points

$$\partial C = \big\{ \boldsymbol{\alpha}(a),\ \boldsymbol{\alpha}(b) \big\}.$$

The proof of this fact is a variation of the proof of Proposition 14.5.4 and we will skip it. A curve with boundary obtained in this fashion is called *convenient*. A map $\boldsymbol{\alpha}$ as above is called a *parametrization* of the convenient curve (with boundary).

(c) The unit circle in $\mathbb{R}^2$ is a closed curve. □

We have the following result whose proof we omit.

Theorem 16.1.9 (Classification of curves with boundary)**.** *(a) Any curve with boundary $C \subset \mathbb{R}^n$ is the union of finitely many pairwise disjoint path connected curves with boundary called the* connected components *of C*

(b) If $C \subset \mathbb{R}^n$ is a path connected *C^k-curve with boundary, then it is either a convenient curve with boundary if $\partial C \neq \emptyset$, or, if C is closed, there exist $T > 0$ and a C^k-immersion $\boldsymbol{\alpha} : \mathbb{R} \to \mathbb{R}^n$ with the following properties*

- $\boldsymbol{\alpha}(\mathbb{R}) = C$,
- $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(t + T)$, $\forall t \in \mathbb{R}$.
- *The restriction to $[0, T)$ is injective.*

*A map with the above properties is called a T-*periodic parametrization *of the closed connected curve C.* □

Example 16.1.10. The closed curve winding around the gold torus in Figure 16.4 admits the 2π-periodic parametrization

$$\boldsymbol{\alpha} : \mathbb{R} \to \mathbb{R}^3,\ \ \boldsymbol{\alpha}(t) = \big((3 + \cos(3t)\,)\cos(2t), (3 + \cos(3t)\,)\sin(2t), \sin(3t)\,\big). \qquad □$$

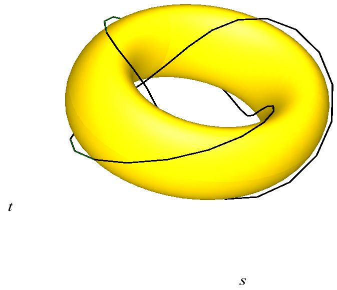

Figure 16.4. *A torus knot*

Remark 16.1.11. Suppose that $C \subset \mathbb{R}^n$ is a closed C^1-curve and $\boldsymbol{\alpha} : \mathbb{R} \to \mathbb{R}^n$ is a T-periodic parametrization of C. Set $\boldsymbol{p}_0 := \boldsymbol{\alpha}(0)$. Then $C' := C \setminus \{\boldsymbol{p}_0\}$. Then C' is a convenient curve and the restriction of $\boldsymbol{\alpha}$ to the open interval $(0, T)$ is a parametrization of C'. □

From the above classification theorem we deduce that if C is a curve with boundary, then its interior C° is a quasi-convenient curve. If $f : C \to \mathbb{R}$ is a continuous function, then we define

$$\boxed{\int_C f(\boldsymbol{p})ds := \int_{C^\circ} f(\boldsymbol{p})ds}.$$

Remark 16.1.12. We can think of a connected curve with boundary in $\mathbb{R}^n$ as a "bent wire". A function $f : C \to (0, \infty)$ can be thought of as a linear density: the quantity $f(\boldsymbol{p})ds$ would be the mass of an infinitesimal arc C of length ds starting at $\boldsymbol{p}$. The integral

$$\int_C f(\boldsymbol{p})ds$$

would then represent the mass of that "bent wire". □

Example 16.1.13. Consider the curve $C \subset \mathbb{R}^3$ obtained by intersecting the unit sphere $\{x^2 + y^2 + z^2 = 1\}$ with the cone spanned by the rays at the origin that make an angle of $\frac{\pi}{4}$ with the positive z-semiaxis: see Figure 16.5. We want to compute the integral

$$I = \int_C xyds. \tag{16.1.5}$$

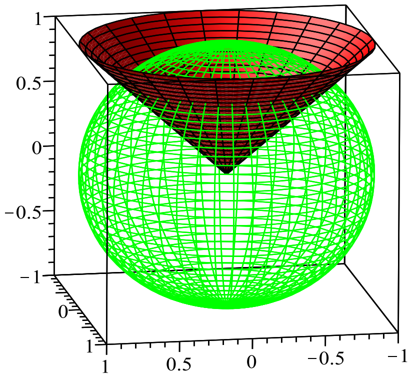

Figure 16.5. *A cone intersecting a sphere.*

The resulting curve is a circle. To find a periodic parametrization for this circle we use spherical coordinates, (ρ, θ, φ),

$$x = \rho \sin\varphi \cos\theta,\ \ y = \rho \sin\varphi \sin\theta,\ \ z = \rho\cos\varphi,\ \ \rho > 0,\ \ \theta \in [0, 2\pi],\ \varphi \in [0, \pi] \tag{16.1.6}$$

In these coordinates the unit sphere is described by the equation $\rho = 1$ and the cone is described by the equation $\varphi = \frac{\pi}{4}$. Taking into account that

$$\cos\frac{\pi}{4} = \sin\frac{\pi}{4} = \frac{\sqrt{2}}{2},$$

we deduce from (16.1.6) that a parametrization for C is described by

$$x = \frac{\sqrt{2}}{2}\sin\theta,\ \ y = \frac{\sqrt{2}}{2}\cos\theta,\ \ z = \frac{\sqrt{2}}{2},\ \ \theta \in [0, 2\pi]. \tag{16.1.7}$$

The arclength element ds on C is then

$$ds = \sqrt{x'(\theta)^2 + y'(\theta)^2 + z'(\theta)^2}\, d\theta = \frac{\sqrt{2}}{2}\, d\theta.$$

To compute the integral (16.1.5) we use the parametrization (16.1.7) and we deduce

$$\int_C xyds = \int_0^{2\pi} \left(\frac{\sqrt{2}}{2}\right)^3 \cos\theta \sin\theta d\theta = \left(\frac{\sqrt{2}}{2}\right)^3 \int_0^{2\pi} \frac{1}{2}\sin 2\theta\, d\theta = 0.$$

□

The next result shows that integral along a curve with boundary has several features in common with the Riemann integrals.

Proposition 16.1.14. *Suppose that $\Gamma \subset \mathbb{R}^n$ is a curve with boundary. (We recall that, by our definition, the curves with boundary are* compact*.) We denote by $C^0(\Gamma)$ the vector space of continuous functions $\Gamma \to \mathbb{R}$. Then the first kind integral along C defines a* linear map

$$C^0(\Gamma) \ni f \mapsto \int_\Gamma f ds \in \mathbb{R}$$

satisfying the monotonicity property:

$$\int_\Gamma f ds \leqslant \int_\Gamma g ds$$

if $f(\boldsymbol{p}) \leqslant g(\boldsymbol{p})$, $\forall \boldsymbol{p} \in \Gamma$. □

We omit the simple proof of the above result.

16.1.2. Integration of differential 1-forms over paths. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set. A *differential form of degree* 1 or a *differential* 1*-form* on U is an expression ω of the form

$$\omega = \omega_1 dx^1 + \cdots + \omega_n dx^n,$$

where $\omega_1, \dots, \omega_n : U \to \mathbb{R}$ are continuous functions. The precise meaning of a 1-form is a bit more complicated to explain at this point but, for the goals we have in mind, it is irrelevant. We will denote by $\Omega^1(U)$ the collection of 1-forms on U.

Example 16.1.15. (a) If $f \in C^1(U)$, then its total differential as described in (13.2.13) is a 1-form

$$df = \partial_{x^1} f dx^1 + \cdots + \partial_{x^n} f dx^n.$$

A 1-form of this type is called *exact*.

(b) Suppose $\boldsymbol{F} : U \to \mathbb{R}^n$ is a continuous vector field on U,

$$\boldsymbol{F}(\boldsymbol{p}) = \begin{bmatrix} F^1(\boldsymbol{p}) \\ F^2(\boldsymbol{p}) \\ \vdots \\ F^n(\boldsymbol{p}) \end{bmatrix},$$

then the *infinitesimal work* is the 1-form

$$W_{\boldsymbol{F}} := F^1 dx^1 + \cdots + F^n dx^n.$$

Traditionally, in classical mechanics the infinitesimal work is denoted by $\boldsymbol{F} \cdot d\boldsymbol{p}$ or $\langle \boldsymbol{F}, d\boldsymbol{p} \rangle$, where "$\cdot$" is short-hand for inner product and $d\boldsymbol{p}$ denotes the "infinitesimal displacement"

$$d\boldsymbol{p} = \begin{bmatrix} dx^1 \\ dx^2 \\ \vdots \\ dx^n \end{bmatrix}.$$

Note that if $f \in C^1(U)$, then $df = W_{\nabla f}$.

(c) The *angular form* on $\mathbb{R}^2 \backslash \{\boldsymbol{0}\}$ is the infinitesimal work associated to the angular vector field (see Figure 16.6)

$$\Theta : \mathbb{R}^2 \backslash \{0\} \to \mathbb{R}^2, \;\; \Theta(x, y) := \begin{bmatrix} -\frac{y}{x^2+y^2} \\ \\ \frac{x}{x^2+y^2} \end{bmatrix}.$$

More explicitly

$$W_\Theta = -\frac{y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy. \qquad \square$$

Definition 16.1.16. Let $n \in \mathbb{N}$ and suppose that $U \subset \mathbb{R}^n$ is an open set. For any differential 1-form

$$\omega = \omega_1 dx^1 + \cdots + \omega_n dx^n \in \Omega^1(U),$$

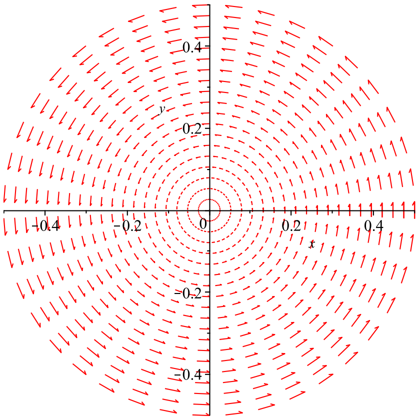

Figure 16.6. *The angular vector field in the punctured plane.*

and any C^1-path

$$\boldsymbol{\gamma} : [a,b] \to U, \;\; \boldsymbol{\gamma}(t) = \big(\gamma^1(t), \dots, \gamma^n(t)\big), \;\; \forall a \leqslant t \leqslant b,$$

we define the *integral of* ω *along the path* $\boldsymbol{\gamma}$ to be the real number

$$\int_{\boldsymbol{\gamma}} \omega := \int_a^b \Big(\omega_1\big(\boldsymbol{\gamma}(t)\big)\dot{\gamma}^1(t) + \cdots + \omega_n\big(\boldsymbol{\gamma}(t)\big)\dot{\gamma}^n(t)\Big)dt.$$

The integral $\int_{\boldsymbol{\gamma}} \omega$ is traditionally known as the *line integral of the second kind* □

When ω is the infinitesimal work of a vector field $\boldsymbol{F}$, $\omega = W_{\boldsymbol{F}}$, then following the physicists' tradition, one uses the notation

$$\int_{\boldsymbol{\gamma}} \boldsymbol{F} \cdot d\boldsymbol{p} := \int_{\boldsymbol{\gamma}} W_{\boldsymbol{F}}.$$

Example 16.1.17. Fix a natural number N, a real number $R > 0$ and consider the path

$$\boldsymbol{\gamma}_N : [0, 2\pi N] \to \mathbb{R}^2 \backslash \{\boldsymbol{0}\}, \;\; \boldsymbol{\gamma}_N(t) = \big(x(t), y(t)\big) := \big(R\cos t, R\sin t\big).$$

Let $W_\Theta \in \Omega^1\big(\mathbb{R}^2\backslash\{0\}\big)$ be the angular form defined in Example 16.1.15(c), i.e.,

$$W_\Theta = \frac{-y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy.$$

Then

$$\int_{\boldsymbol{\gamma}_N} W_\Theta = \int_{\boldsymbol{\gamma}_N} \left(\frac{-y}{x^2+y^2}\,dx + \frac{x}{x^2+y^2}\,dy\right)$$

($dx = \dot{x}dt$, $dy = \dot{y}dt$)

$$= \int_0^{2\pi N} \frac{-y}{x^2+y^2}\,\dot{x}dt + \int_0^{2\pi N} \frac{x}{x^2+y^2}\,\dot{y}dt$$

Observing that along $\boldsymbol{\gamma}_N$ we have $x^2+y^2=R^2$, $\dot{x}=-R\sin t$, $\dot{y}=R\cos t$, we deduce

$$\begin{aligned}\int_{\gamma_N} W_\Theta &= \int_0^{2\pi N}\left(\frac{-R\sin t}{R^2}(-R\sin t)+\frac{R\cos t}{R^2}R\cos t\right)dt\\ &= \int_0^{2\pi N}\big(\sin^2 t+\cos^2 t\big)dt = 2\pi N.\end{aligned}$$

□

Theorem 16.1.18 (1-dimensional Stokes' formula)**.** *Let $n\in\mathbb{N}$ and suppose that $\mathcal{O}\subset\mathbb{R}^n$ is an open subset. Then, for any $f\in C^1(\mathcal{O})$, and any C^1-path $\boldsymbol{\gamma}:[a,b]\to\mathcal{O}$ we have*

$$\boxed{\int_{\boldsymbol{\gamma}} df = f\big(\boldsymbol{\gamma}(b)\big)-f\big(\boldsymbol{\gamma}(a)\big)}. \tag{16.1.8}$$

Proof. Denote by $\big(x^1(t),\dots,x^n(t)\big)$ the coordinates of the point $\boldsymbol{\gamma}(t)$. Then

$$\int_{\boldsymbol{\gamma}} df = \int_{\boldsymbol{\gamma}}\Big(\partial_{x^1}f\,dx^1+\cdots+\partial_{x^n}f\,dx^n\Big)$$

$$=\int_a^b\Big(\partial_{x^1}f\big(x^1(t),\dots,x^n(t)\big)\,\dot{x}^1+\cdots+\partial_{x^n}f\big(x^1(t),\dots,x^n(t)\big)\,\dot{x}^n\Big)dt$$

(use the chain rule)

$$=\int_a^b\frac{d}{dt}f\big(\boldsymbol{\gamma}(t)\big)dt = f\big(\boldsymbol{\gamma}(b)\big)-f\big(\boldsymbol{\gamma}(a)\big).$$

□

Remark 16.1.19. (i) Although very simple, the above result has very important consequences. First, let us observe that df is the infinitesimal work of the vector field ∇f

$$df = W_{\nabla f}.$$

In classical mechanics the function $U=-f$ is often called the *potential* of the gradient vector field $\boldsymbol{F}=\nabla f$.

Think of $\boldsymbol{\gamma}(t)$ as describing the motion of a particle interacting with the force field ∇f. For example, ∇f can be the gravitational field and the particle is a "heavy" particle, i.e., a particle with positive mass. The integral

$$\int_{\boldsymbol{\gamma}} df = \int_{\boldsymbol{\gamma}} W_{\nabla f}$$

is interpreted in classical mechanics as the total energy required to generate the travel of the particle described by the path $\boldsymbol{\gamma}$. Stokes' formula (16.1.8) shows that, when the force field is a gradient vector field, then this total energy depends only on the endpoints of the travel and not on what happened in between. In particular, if the path $\boldsymbol{\gamma}$ is closed, $\boldsymbol{\gamma}(b)=\boldsymbol{\gamma}(a)$, so the particle ends at the same point where it started, this total energy is trivial!

(ii) Let $U \subset \mathbb{R}^n$ be an open set. As we have mentioned earlier a 1-form $\omega \in \Omega^1(U)$ is *exact* if there exists $f \in C^1(U)$ such that $\omega = df$. A function f such that $df = \omega$ is called an antiderivative of ω.

If

$$\omega = \sum_{i=1}^{n} \omega_d x^i,$$

then ω is exact if there exists $f \in C^1(U)$ such that

$$\omega_i = \frac{\partial f}{\partial x^i}, \quad \forall i = 1, \dots, n.$$

Since $\partial_{x^i}\partial_{x^j} f = \partial_{x^j}\partial_{x^i} f$, $\forall i, j$, we deduce that, if ω is exact then

$$\boxed{\frac{\partial \omega_i}{\partial x^j} = \frac{\partial \omega_j}{\partial x^i}, \quad \forall i, j.} \tag{16.1.9}$$

A 1-form satisfying (16.1.9) is called *closed*. In other words,

$$\omega \text{ exact } \Rightarrow \omega \text{ closed.}$$

A famous result, known by the name of *Poincaré Lemma* [**35**, Thm. 4-11], states that the converse is true if U is *convex*,

$$U \text{ convex, } \omega \text{ closed } \Rightarrow \omega \text{ exact.}$$

The result is not true without the convexity assumption. Consider for example the 1-form

$$W_\Theta = \underbrace{\frac{-y}{x^2+y^2}}_{} \, dx + \underbrace{\frac{x}{x^2+y^2}}_{} \, dy \in \Omega^1\big(\mathbb{R}^2 \backslash \{0\}\big).$$

As we will see soon in Example 16.1.35 we have

$$P'_y = Q'_y,$$

so the form W_Θ is closed.

On the other hand, it is not exact because, as shown in Example 16.1.17, its integral over the counterclockwise oriented unit circle centered at the origin is 2π.

This curious phenomenon is the beginning of a rather deep story called *cohomology*.

□

Definition 16.1.20. Let $n \in \mathbb{N}$. A *piecewise C^1-path* in $\mathbb{R}^n$ is a continuous map $\boldsymbol{\gamma} : [a, b] \to \mathbb{R}^n$ such that, there exists a partition

$$a = t_0 < t_1 < \cdots < t_\ell = b$$

of $[a, b]$ with the property that, for any $i = 1, \dots, \ell$, the restriction of $\boldsymbol{\gamma}$ to the subinterval $[t_{i-1}, t_i]$ is C^1. A partition of $[a, b]$ with the above properties is said to be *adapted* to $\boldsymbol{\gamma}$.

□

Example 16.1.21. Consider the continuous path $\boldsymbol{\gamma} : [0,4] \to \mathbb{R}^2$ defined by

$$\boldsymbol{\gamma}(t) = \begin{cases} \big(t, 0\big), & t \in [0,1],\\ \big(1, t-1\big), & t \in (1,2],\\ \big(3-t, 1\big), & t \in (2,3],\\ \big(0, 4-t\big), & t \in (3,4]. \end{cases}$$

The image of this path is the boundary of the unit square $[0,1] \times [0,1] \subset \mathbb{R}^2$; see Figure 16.7. □

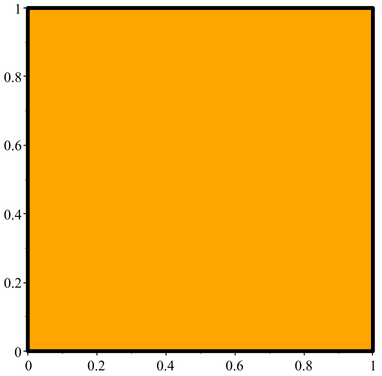

Figure 16.7. *The unit square* $[0,1]^2$ *and its boundary.*

One can integrate differential forms over piecewise C^1-paths. Suppose that $U \subset \mathbb{R}^n$ is an open set,

$$\omega = \omega_1 dx^1 + \cdots + \omega_n dx^n \in \Omega^1(U)$$

and $\boldsymbol{\gamma} : [a,b] \to U$ is a C^1-path. If $a = t_0 < t_1 < \cdots < t_\ell = b$ is *any* partition of $[a,b]$ adapted to $\boldsymbol{\gamma}$ then we set

$$\boxed{\int_{\boldsymbol{\gamma}} \omega = \sum_{i=1}^{\ell} \int_{t_{i-1}}^{t_i} \Big(\omega_1\big(\boldsymbol{\gamma}(t)\big)\dot{\gamma}^1 + \cdots + \omega_n\big(\boldsymbol{\gamma}(t)\big)\dot{\gamma}^n\Big) dt}.$$

One can show that the right-hand side is independent of the choice of the partition adapted to $\boldsymbol{\gamma}$.

Example 16.1.22. Let $\boldsymbol{\gamma}$ denote the piecewise path described in Example 16.1.21. Suppose that

$$\omega = -ydx + xdy.$$

If we denote by $(x(t), y(t))$ the moving point $\boldsymbol{\gamma}(t)$ then we deduce

$$\int_{\boldsymbol{\gamma}} (-ydx + xdy) = \int_0^1 \big(-y\dot{x} + x\dot{y}\big) dt + \int_1^2 \big(-y\dot{x} + x\dot{y}\big) dt$$

$$+\int_2^3\big(-y\dot{x}+x\dot{y}\big)dt+\int_3^4\big(-y\dot{x}+x\dot{y}\big)dt.$$

Observing that on the intervals $[0,1]$ and $[2,3]$ we have $\dot{y}=0$ while on the others we have $\dot{x}=0$ we deduce

$$\int_\gamma(-ydx+xdy)=-\underbrace{\int_0^1 y\dot{x}dt}_{y=0,}+\underbrace{\int_1^2 x\dot{y}dt}_{x=1,\,\dot{y}=1}-\underbrace{\int_2^3 y\dot{x}dt}_{y=1,\,\dot{x}=-1}+\underbrace{\int_3^4 x\dot{y}dt}_{x=0}$$

$$=\int_1^2 dt+\int_2^3 dt=2.$$ □

16.1.3. Integration of 1-forms over oriented curves. There is a simple way of producing a path given a connected curve, namely to assign an orientation to that curve. Loosely speaking, an orientation describes a direction of motion along the curve without specifying the speed of that motion; see Figure 16.8

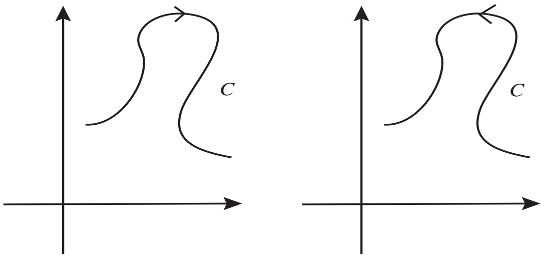

Figure 16.8. *Two orientations on the same planar curve C.*

The direction of motion at a point on the curve C would be given by a unit vector tangent to the curve at that point. At each point $\boldsymbol{p}\in C$ there are exactly *two* unit vectors tangent to C and an orientation would correspond to a choosing one such vector at each point, and the choices would vary continuously from one point to another. Here is a precise definition.

Definition 16.1.23. Let $n\in\mathbb{N}$.

(i) An *orientation* on a C^1-curve $C\subset\mathbb{R}^n$ is a continuous map $\boldsymbol{T}:C\to\mathbb{R}^n$ such that, $\forall\boldsymbol{p}\in C$, the vector $\boldsymbol{T}(\boldsymbol{p})$ has length 1 and it is tangent to C at $\boldsymbol{p}\in C$.

(ii) An *oriented curve* is a pair $(C,\boldsymbol{T})$, where C is a curve and $\boldsymbol{T}$ is an orientation on C.

(iii) An *orientation* on a (compact) C^1-curve with boundary C is an orientation on its interior C°.

(iv) Suppose that C is a convenient curve and $\boldsymbol{T}$ is an orientation of C. A parametrization $\boldsymbol{\gamma}:(a,b)\to\mathbb{R}^n$ is said to be *compatible with the orientation* if, $\forall t\in(a,b)$, the tangent vectors

$$\dot{\boldsymbol{\gamma}}(t),\;\;\boldsymbol{T}\big(\boldsymbol{\gamma}(t)\big)\in T_{\boldsymbol{\gamma}(t)}C$$

point in the same direction, i.e., $\big\langle\, \dot{\boldsymbol{\gamma}}(t), \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big)\,\big\rangle > 0$.

□

Let us observe that if $C \subset \mathbb{R}^n$ is a convenient curve, then any parametrization $\boldsymbol{\gamma}(a,b) \to \mathbb{R}^n$ of C defines an orientation on $\boldsymbol{T} : C \to \mathbb{R}^n$ on C according to the rule

$$\boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big) = \frac{1}{\|\dot{\boldsymbol{\gamma}}(t)\|}\dot{\boldsymbol{\gamma}}(t).\ \ \forall t \in (a,b).$$

This is called the *orientation induced by the parametrization*. Clearly the parametrization is compatible with the orientation it induces since

$$\big\langle\, \dot{\boldsymbol{\gamma}}(t), \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big)\,\big\rangle = \|\dot{\boldsymbol{\gamma}}(t)\| > 0.$$

It turns out that any orientation on a convenient curve is induced by a parametrization.

Lemma 16.1.24. *Suppose that C is a convenient curve in $\mathbb{R}^n$. For any parametrization $\boldsymbol{\gamma} : (a,b) \to \mathbb{R}^n$ of C we denote by $\boldsymbol{\gamma}_-$ the parametrization*

$$\boldsymbol{\gamma}_- : (-b,-a) \to \mathbb{R}^n,\ \ \boldsymbol{\gamma}_-(t) := \boldsymbol{\gamma}(-t).$$

If $\boldsymbol{T}$ is an orientation on C, then exactly one of the parametrizations $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_-$ is compatible with the orientation.

Proof. Observe that for any $t \in (a,b)$, the *nonzero* vectors $\boldsymbol{T}\big(\boldsymbol{\gamma}(t)\big)$ and $\dot{\boldsymbol{\gamma}}(t)$ belong to the 1-dimensional space $T_{\boldsymbol{\gamma}(t)}C$. They are therefore collinear and thus

$$\big\langle\, \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big), \dot{\boldsymbol{\gamma}}(t)\,\big\rangle \neq 0,\ \ \forall t \in (a,b)$$

Since the function

$$(a,b) \ni t \mapsto \big\langle\, \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big), \dot{\boldsymbol{\gamma}}(t)\,\big\rangle \in \mathbb{R}$$

and is nowhere zero, we deduce that either

$$\big\langle\, \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big), \dot{\boldsymbol{\gamma}}(t)\,\big\rangle > 0,\ \ \forall t \in (a,b),$$

or,

$$\big\langle\, \boldsymbol{T}\big(\,\boldsymbol{\gamma}(t)\,\big), \dot{\boldsymbol{\gamma}}(t)\,\big\rangle < 0,\ \ \forall t \in (a,b).$$

In the first case we deduce that $\boldsymbol{\gamma}$ is compatible with $\boldsymbol{T}$, while in the second case we deduce that $\boldsymbol{\gamma}_-$ is compatible with $\boldsymbol{T}$ □

Suppose now that $U \subset \mathbb{R}^n$ is an open set,

$$\omega = \omega_1 dx^1 + \cdots + \omega_n dx^n \in \Omega^1(U),$$

$C \subset \mathbb{R}^n$ is a convenient curve and $\boldsymbol{T}$ is an orientation on C.

To define the integral of ω on the oriented convenient curve $(C,\boldsymbol{T})$ we need to first choose a parametrization $\boldsymbol{\alpha} : (a,b) \to \mathbb{R}^n$ of C compatible with the orientation. We then set

$$\int_C \omega = \int_{(C,\boldsymbol{T})} \omega := \int_{\boldsymbol{\alpha}} \omega = \int_a^b \Big(\,\omega_1\big(\,\boldsymbol{\alpha}(t)\,\big)\dot{\boldsymbol{\alpha}}^1(t) + \cdots + \omega_n\big(\,\boldsymbol{\alpha}(t)\,\big)\dot{\boldsymbol{\alpha}}^n(t)\,\Big)dt.$$

For the above definition to be consistent, the right-hand side has to be independent of parametrization compatible with the given orientation.

Lemma 16.1.25. *The above definition is independent of the choice of the parametrization of C compatible with the orientation.*

Proof. Indeed, if $\boldsymbol{\beta}:(c,d)\to\mathbb{R}^n$ is another such parametrization, then, as argued in Subsection 16.1.1 (see Figure 16.1) there exists a continuous C^1-bijection $(a,b)\ni t\mapsto\tau(t)\in(c,d)$, such that

$$\boldsymbol{\beta}\big(\tau(t)\big)=\boldsymbol{\alpha}\big(t\big),\;\;\frac{d\boldsymbol{\beta}}{d\tau}\big(\tau(t)\big)\frac{d\tau}{dt}=\dot{\boldsymbol{\alpha}}(t),\;\;\forall t\in(a,b).$$

Since the tangent vectors

$$\frac{d\boldsymbol{\beta}}{d\tau}\big(\tau(t)\big),\;\;\frac{d\boldsymbol{\alpha}}{dt}(t)$$

point in the same directions we deduce

$$\frac{d\tau}{dt}>0,\;\;\forall t\in(a,b).$$

The 1-dimensional change-in-variables formula implies that

$$\int_c^d\omega_i\big(\beta(\tau)\big)\frac{d\beta^i}{d\tau}d\tau=\int_a^b\omega_i\big(\beta(\tau(t))\big)\frac{d\beta^i}{d\tau}\frac{d\tau}{dt}dt=\int_a^b\omega_i\big(\boldsymbol{\alpha}(t)\big)\frac{d\alpha^i}{dt}dt,\;\;\forall i=1,\dots,n.$$

Hence

$$\int_{\boldsymbol{\alpha}}\omega=\sum_{i=1}^n\int_a^b\omega_i\big(\boldsymbol{\alpha}(t)\big)\frac{d\alpha^i}{dt}dt=\sum_{i=1}^n\int_c^d\omega_i\big(\beta(\tau)\big)\frac{d\beta^i}{d\tau}d\tau=\int_{\boldsymbol{\beta}}\omega.$$

□

Proposition 16.1.26. *Let $U\subset\mathbb{R}^n$ be an open set, and $(C,\boldsymbol{T})$ an oriented convenient curve inside U. Then, for any 1-form $\omega\in\Omega^1(U)$ we have*

$$\int_{(C,-\boldsymbol{T})}\omega=-\int_{(C,\boldsymbol{T})}\omega.$$

Proof. Let

$$\omega=\omega_1dx^1+\cdots+\omega_ndx^n$$

Fix a parametrization

$$\boldsymbol{\gamma}:(a,b)\to\mathbb{R}^n,\;\;\boldsymbol{\gamma}(t)=\begin{bmatrix}x^1(t)\\x^2(t)\\\vdots\\x^n(t)\end{bmatrix},$$

compatible with the orientation $\boldsymbol{T}$. Then $\boldsymbol{\gamma}_-:(-b,-a)\to\mathbb{R}^n$ is a parametrization compatible with $-\boldsymbol{T}$. We have

$$\int_{(C,-\boldsymbol{T})}\omega=\int_{\boldsymbol{\gamma}_-}\omega=\int_{-b}^{-a}\Big(-\omega_1\big(\boldsymbol{\gamma}(-t)\big)\dot{x}^1(-t)-\cdots-\omega_n\big(\boldsymbol{\gamma}(-t)\big)\dot{x}^n(-t)\Big)dt$$

$(t := -\tau)$

$$= \int_b^{a} \Big(\omega_1\big(\boldsymbol{\gamma}(\tau)\big)\dot{x}^1(\tau) + \cdots + \omega_n\big(\boldsymbol{\gamma}(\tau)\big)\dot{x}^1(\tau)\Big)d\tau$$

$$= -\int_{a}^b \Big(\omega_1\big(\boldsymbol{\gamma}(\tau)\big)\dot{x}^1(\tau) + \cdots + \omega_n\big(\boldsymbol{\gamma}(\tau)\big)\dot{x}^1(\tau)\Big)d\tau$$

$$= -\int_{\boldsymbol{\gamma}} \omega = -\int_{(C,\boldsymbol{T})} \omega.$$

□

If C is an oriented quasi-convenient curve, choose a convenient cut $\{\boldsymbol{p}_1, \dots, \boldsymbol{p}_\ell\}$ so that C becomes a disjoint union of convenient curves $C_1, \dots, C_N$ equipped with orientations. Then define

$$\boxed{\int_C \omega := \sum_{i=1}^N \int_{C_i} \omega}.$$

One can show that the right-hand side of the above equality is independent of the choice of the convenient cut.

Let us finally observe that the interior of an oriented (compact) curve with boundary C is oriented and quasi-convenient and we define

$$\boxed{\int_C \omega := \int_{C^\circ} \omega}.$$

Example 16.1.27. Suppose that C is the quarter of the unit circle contained in the first quadrant and equipped with the counter-clockwise orientation; see Figure 16.9.

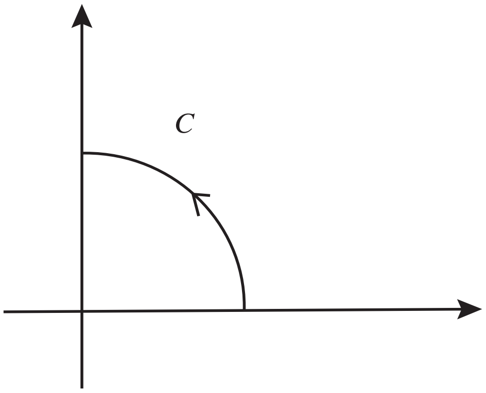

Figure 16.9. *An arc of the unit circle equipped with the counterclockwise orientation.*

Its interior is a convenient curve and the map

$$(0, \pi/2) \ni t \mapsto (\cos t, \sin t) \in \mathbb{R}^2$$

is a parametrization compatible with the counterclockwise orientation. We have

$$\int_C W_\Theta = \int_C \left(\frac{-y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy \right)$$

(along C we have $dx = -\sin t dt$, $dy = \cos t dt$, $x^2+y^2=1$)

$$= \int_0^{\frac{\pi}{2}} \Big((-\sin t)(-\sin t)dt + (\cos t)(\cos t)dt \Big) = \int_0^{\frac{\pi}{2}} dt = \frac{\pi}{2}.$$ □

Definition 16.1.28. A 1-*dimensional chain* in $\mathbb{R}^n$ is a collection

$$\mathscr{C} := \{ (C_1, m_1), \dots, (C_\nu, m_\nu) \}$$

where $C_1, \dots, C_n$ are compact, *oriented*, curves (with or without boundary) and $m_1, \dots, m_\nu$ are integers called the *local multiplicities* of the chain.

The integral of a 1-form ω over the above chain $\mathscr{C}$ is the real number

$$\int_{\mathscr{C}} \omega := \sum_{k=1}^{\nu} m_k \int_{C_k} \omega.$$ □

16.1.4. The 2-dimensional Stokes' formula: a baby case. The integrals of 1-forms over closed curves can often be computed as certain double integrals over appropriate regions. This is roughly speaking the content of Stokes' formula.

Definition 16.1.29. Let $k, n \in \mathbb{N}$. A *domain* of $\mathbb{R}^n$ is an open, path connected subset of $\mathbb{R}^n$. A domain $D \subset \mathbb{R}^n$ is called C^k if its boundary is an $(n-1)$-dimensional C^k-submanifold of $\mathbb{R}^n$. □

Example 16.1.30. In Figure 16.10 we have depicted two C^k-domains in $\mathbb{R}^2$. The domain D_1 is a closed disk and its boundary is a circle. The boundary of D_2 consists of three closed curves. □

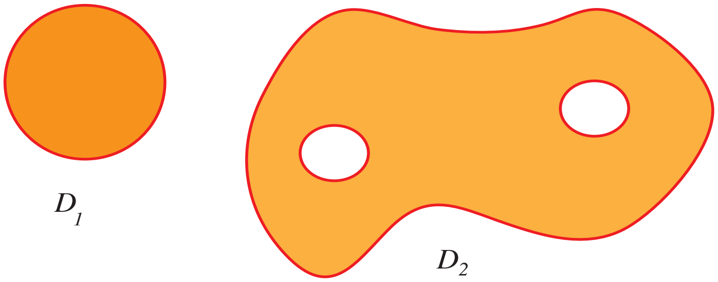

Figure 16.10. *Two domains in* $\mathbb{R}^2$.

Suppose that $D \subset \mathbb{R}^2$ is a bounded C^k domain. As one can see from Figure 16.10, its boundary is a disjoint union of compact C^k curves in $\mathbb{R}^2$. Each of these components is equipped with a natural orientation called the *induced orientation*. It is determined by the right hand rule; see Figure 16.11 and 16.12.

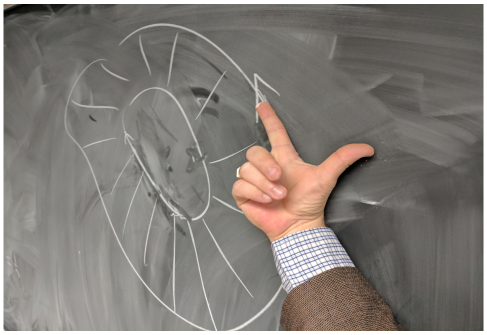

Figure 16.11. *Right-hand rule.*

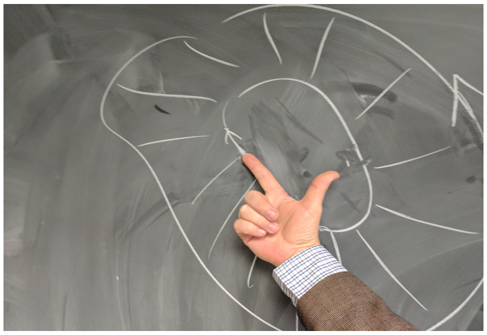

Figure 16.12. *Right-hand rule.*

Here is the explanation for the above figures: place your right hand on the boundary component, palm-up, so that the thumb points to the exterior of your domain. The index finger will then point in the direction given by the orientation of that boundary component. Another way of visualizing is as follows: if we walk along ∂D in the direction prescribed by the induced orientation, then the domain D will be to our left; see Figure 16.13.

Here is a more rigorous explanation. Given a point $\boldsymbol{p} \in \partial D$, there are exactly two unit vectors that are perpendicular to the tangent line $T_{\boldsymbol{p}}\partial D$. These vectors are called the *normal vectors* to ∂C at $\boldsymbol{p}$. Denote by $\boldsymbol{\nu}$ or $\boldsymbol{\nu}(\boldsymbol{p})$ the *outer normal vector at* $\boldsymbol{p} \in \partial D$: this vector is perpendicular to $T_{\boldsymbol{p}}\partial D$ and, when placed at $\boldsymbol{p}$, it points towards the exterior of D. The induced orientation at $\boldsymbol{p}$ is obtained from $\boldsymbol{\nu}(\boldsymbol{p})$ after a 90 degree counterclockwise rotation.

Thus, the boundary of a bounded C^1-domain $D \subset \mathbb{R}^2$ is a disjoint union of closed curves carrying orientations. It is therefore a 1-dimensional chain in the sense of Definition 16.1.28. Thus, we can integrate 1-forms on such boundaries.

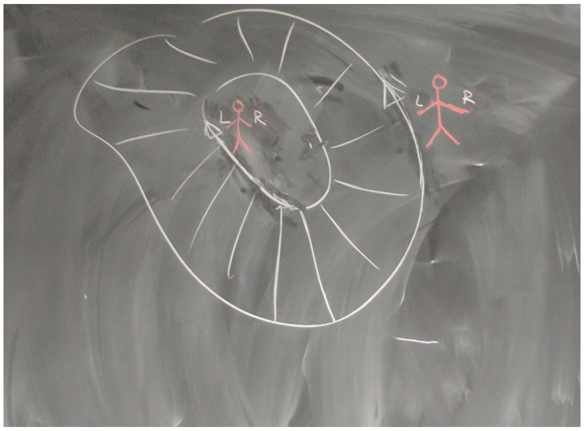

Figure 16.13. *Walking around the boundary.*

Theorem 16.1.31 (Planar Stokes' theorem)**.** *Let $U \subset \mathbb{R}^2$ be a bounded C^1-domain. Suppose that $\boldsymbol{F}$ is a C^1-vector field defined on an open set $\mathcal{O}$ that contains $\mathbf{cl}(U)$,*

$$\boldsymbol{F} : \mathcal{O} \to \mathbb{R}^2, \;\; \boldsymbol{F}(x,y) = \big(\, P(x,y), Q(x,y) \,\big)$$

Let $\boldsymbol{\nu} : \partial U \to \mathbb{R}^2$ be the <u>outer</u> normal vector field. We denoted by $\partial_+ U$ the boundary of U equipped with the induced orientation. Then the following hold

$$\int_{\partial_+ U} (Pdx + Qdy) = \int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) |dxdy|. \tag{16.1.10a}$$

$$\int_{\partial_+ U} \langle \boldsymbol{F}, \boldsymbol{\nu} \rangle ds = \int_U \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) |dxdy| \tag{16.1.10b}$$

□

Remark 16.1.32. It turns out that the equality (16.1.10b) follows rather easily from (16.1.10a). The hard part is proving (16.1.10a). □

Definition 16.1.33. Let $U \subset \mathbb{R}^n$ be an open set and

$$\boldsymbol{F} : U \to \mathbb{R}^n, \;\; \boldsymbol{F}(\boldsymbol{p}) = \begin{bmatrix} F^1(\boldsymbol{p}) \\ F^2(\boldsymbol{p}) \\ \vdots \\ F^n(\boldsymbol{p}) \end{bmatrix}.$$

a C^1 vector field. The *divergence* of $\boldsymbol{F}$ is the function

$$\operatorname{div} : U \to \mathbb{R}, \;\; \operatorname{div} \boldsymbol{F}(\boldsymbol{p}) = \frac{\partial F^1}{\partial x^1}(\boldsymbol{p}) + \cdots + \frac{\partial F^n}{\partial x^n}(\boldsymbol{p}).$$

□

Using the concept of divergence we can rephrase (16.1.10b) in the traditional form

$$\boxed{\int_{\partial D}\langle \boldsymbol{F},\boldsymbol{\nu}\rangle ds = \int_D \operatorname{div}\boldsymbol{F}\,|dxdy|}. \tag{16.1.11}$$

The integral in the left-hand side is usually called the *outer flux of* $\boldsymbol{F}$ *through* ∂D. The last equality is a special case of the *flux-divergence formula*

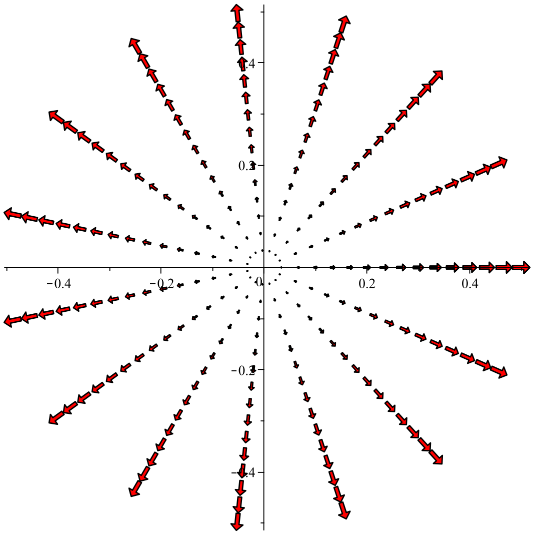

Figure 16.14. *The radial vector field in the plane.*

Example 16.1.34. The *radial* vector field $\mathcal{R}:\mathbb{R}^2\to\mathbb{R}^2$ is given by

$$\mathcal{R}(x,y)=\begin{bmatrix} x\\ y\end{bmatrix}.$$

Thus the vector field $\mathcal{R}$ associates to the point $\boldsymbol{p}\in(x,y)$, the point $\boldsymbol{p}$ itself but viewed as a vector in $\mathbb{R}^2$; see Figure 16.14. In other words $\mathcal{R}$, is the identity map under another guise. Note that

$$\operatorname{div}\mathcal{R}=\partial_x x+\partial_y y=2.$$

If D is a bounded domain with C^1-boundary, then the flux-divergence formula shows that the outer flux of $\mathcal{R}$ through the boundary ∂D is equal to twice the area of D. Thus we can compute the area of D by performing computations involving only boundary data and no interior data. □

Example 16.1.35. Suppose that $U\subset\mathbb{R}^2$ is a bounded C^1-domain whose boundary $C=\partial U$ is a closed connected curve. We assume that the origin $\mathbf{0}$ is not on the boundary of U. The induced orientation on ∂U is the counterclockwise orientation; see Figure 16.15. We denote by $\partial_+ U$ the boundary ∂U equipped with this orientation. We want to compute

$$\int_{\partial_+ U} W_\Theta = \int_{\partial_+ U} \left(\underbrace{-\frac{y}{x^2+y^2}}_{P} \, dx \; + \; \underbrace{\frac{x}{x^2+y^2}}_{Q} \, dy \right). \tag{16.1.12}$$

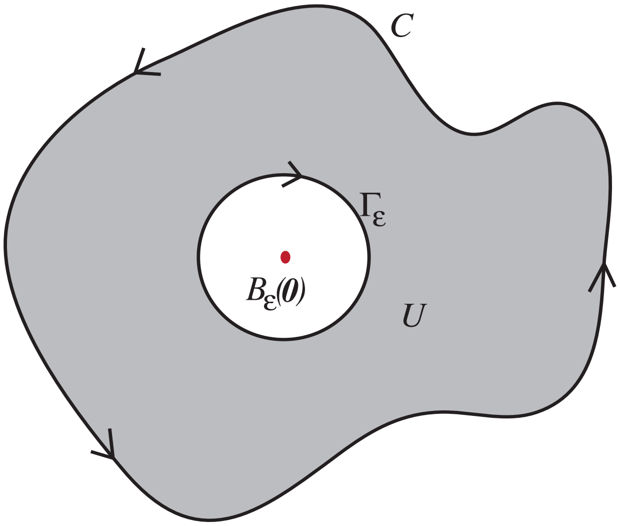

Figure 16.15. *Integrating the angular form.*

We distinguish two cases.

1. $\mathbf{0} \notin U$. To compute (16.1.12) we use (16.1.10a). To find $Q'_x - P'_y$ we set $r := \sqrt{x^2+y^2}$, so

$$P = -\frac{y}{r^2}, \;\; Q = \frac{x}{r^2}.$$

We have

$$r'_x = \frac{x}{r}, \;\; r'_y = \frac{y}{r}, \;\; Q'_x = \frac{1}{r^2} - \frac{2x}{r^3} \cdot \frac{x}{r} = \frac{1}{r^2} - \frac{2x^2}{r^4},$$

$$P'_y = -\frac{1}{r^2} + \frac{2y}{r^3} \cdot \frac{y}{r} = -\frac{1}{r^2} + \frac{2y^2}{r^4}$$

Thus

$$Q'_x - P'_y = \frac{2}{r^2} - \frac{2(x^2+y^2)}{r^4} = 0, \;\; \forall (x,y) \neq (0,0).$$

Using (16.1.10a) we deduce

$$\int_{\partial_+ U} W_\Theta = \int_U \big(Q'_x - P'_y \big) \, |dxdy| = 0.$$

2. $\mathbf{0} \in U$. The above approach does not work. Theorem 16.1.31 requires that the vector field Θ be defined on an open set containing U. This is not the case. The vector field Θ

is not defined at the origin $\mathbf{0}$, and in fact the components P and Q do not have a limit as $(x,y) \to (0,0)$ so they cannot be the restrictions of some continuous functions on U. Although this issue may seem trivial, it has enormous consequences.

To deal with this issue we will tread lightly around the singular point $\mathbf{0}$. Observe first that there exists $\varepsilon > 0$ such that the closed ball $\overline{B_\varepsilon(\mathbf{0})}$ is contained in U. Denote by U_ε the domain obtained by removing this ball from U,

$$U_\varepsilon := U \setminus \overline{B_\varepsilon(\mathbf{0})}.$$

The boundary of the domain U_ε has two components: the boundary ∂U and the boundary Γ_ε of $B_\varepsilon(\mathbf{0})$. Each of these components is equipped with an induced orientation: on ∂U this is the counterclockwise orientation, while on Γ_ε this is the clockwise orientation; see Figure 16.15. Observe that the clockwise orientation on Γ_ε is the opposite of the orientation induced as boundary of $B_\varepsilon(\mathbf{0})$. We denote by $\partial_+ B_\varepsilon(\mathbf{0})$ the curve Γ_ε equipped with the *counter*clockwise orientation. We have

$$\int_{\partial_+ U_\varepsilon} W_\theta = \int_{\partial_+ U} W_\Theta - \int_{\partial_+ B_\varepsilon(\mathbf{0})} W_\Theta.$$

Now observe that Θ is defined and C^1 on the open set $\mathbb{R}^2 \setminus \mathbf{0}$ that contains U_ε. We can now safely invoke Theorem 16.1.31 to conclude that

$$0 = \int_{U_\varepsilon} \big(Q'_x - P'_y \big) = \int_{\partial_+ U_\varepsilon} W_\Theta = \int_{\partial_+ U} W_\Theta - \int_{\partial_+ B_\varepsilon(\mathbf{0})} W_\Theta.$$

Hence

$$\int_{\partial_+ U} W_\Theta = \int_{\partial_+ B_\varepsilon(\mathbf{0})} W_\Theta.$$

To compute the right-hand side of the above equality observe that a parametrization of Γ_ε compatible with the counterclockwise orientation is

$$\boldsymbol{\gamma}(t) = \big(\varepsilon \cos t, \varepsilon \sin t \big), \quad t \in [0, 2\pi].$$

Arguing exactly as in Example 16.1.27 we deduce

$$\int_{\partial_+ B_\varepsilon(\mathbf{0})} W_\Theta = \int_{\boldsymbol{\gamma}} \left(\frac{-y}{x^2+y^2} dx + \frac{x}{x^2+y^2} dy \right) =$$

$$= \int_0^{2\pi} \Big((-\sin t)(-\sin t) dt + (\cos t)(\cos t) dt \Big) = \int_0^{2\pi} dt = 2\pi.$$

This proves that

$$\int_{\partial_+ U} W_\Theta = 2\pi.$$

To put things in perspective let us mention that a famous result of Jordan[1] states that if C is a connected compact C^1-curve, then it is the boundary of a bounded C^1-domain U.

[1] I recommend T. Hales' very lively discussion of this result in [20].

We equip it with the orientation as boundary of U. If $\mathbf{0}\notin\partial U$, then $\int_C$ is well defined and the above computation shows

$$\frac{1}{2\pi}\int_{\partial_+ U} W_\Theta = I_U(\mathbf{0}) = \int_C W_\Theta = \begin{cases} 1, & \mathbf{0}\in U,\\ 0, & \mathbf{0}\notin U.\end{cases}$$

This "quantization" result has profound consequences in mathematics. □

The Planar Stokes' Theorem 16.1.31 extends to more general domains.

Definition 16.1.36. A domain $D\subset\mathbb{R}^2$ is called *piecewise* C^k if there exists a finite subset F of the boundary ∂D such that each component of $\partial D\backslash F$ is the interior of a C^k-curve with boundary. □

Example 16.1.37. Suppose that $\beta,\tau:[a,b]$ are C^1-functions such that

$$\beta(x)<\tau(x),\ \ \forall x\in(a,b).$$

Then the simple type domain

$$D(\beta,\tau)=\big\{\,(x,y)\in\mathbb{R}^2;\ \ x\in(a,b),\ \ \beta(x)<y<\tau(x)\,\big\} \tag{16.1.13}$$

is a piecewise C^1-domain. In particular an open rectangle $(a,b)\times(c,d)$ is a piecewise C^k domain. □

Suppose that $D\subset\mathbb{R}^2$ is a bounded piecewise C^1 domain. Remove a finite subset F of the boundary to obtain a union of convenient C^1-curves. In fact, each of these components is the interior of a (compact) curve with boundary. We denote by $\partial^* U$ the complement of this finite set. Using the right-hand rule we obtain orientations on each component of ∂^*U. We denote by $\partial^*_+ U$ the chain obtained this way, where the multiplicity of each oriented component of $\partial^* U$ is equal to 1. We have the following generalization of Theorem 16.1.31.

Theorem 16.1.38 (Planar Stokes' theorem)**.** *Let $U\subset\mathbb{R}^2$ be a bounded piecewise C^1-domain. Suppose that $\boldsymbol{F}$ is a C^1-vector field defined on an open set $\mathcal{O}$ that contains $\mathbf{cl}(U)$,*

$$\boldsymbol{F}:\mathcal{O}\to\mathbb{R}^2,\ \ \boldsymbol{F}(x,y)=\big(\,P(x,y),Q(x,y)\,\big)$$

Then

$$\int_{\partial^*_+ U}(Pdx+Qdy)=\int_U\left(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}\right)|dxdy|. \tag{16.1.14}$$

□

16.2. Integration over surfaces

The various concepts of integrals over curves have higher dimensional counterparts, called integrals overs submanifolds of a Euclidean space. In this section we will explain this concept only in the special case of 2-dimensional submanifolds. The proper presentation of the more general case requires a more complicated formalism that might bury the geometry of the construction. The restriction to the 2-dimensional situation will afford us more geometric transparency and a lighter algebraic burden. The extension to higher dimensions involves few new geometric ideas, but requires more algebraic travails.

The most basic example of an integral over a surface is the concept of area. To define it we consider first the simplest of situations, when the surface in question is contained in a 2-dimensional vector subspace.

16.2.1. The area of a parallelogram. Suppose that we are given two linearly independent, i.e., non-collinear, vectors

$$\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2, \;\; \boldsymbol{v}_1 = \begin{bmatrix} v_1^1 \\ \\ v_1^2 \end{bmatrix}, \;\; \boldsymbol{v}_1 = \begin{bmatrix} v_2^1 \\ \\ v_2^2 \end{bmatrix}.$$

We denote by V the 2×2 matrix with columns $\boldsymbol{v}_1, \boldsymbol{v}_2$, i.e.,

$$V = \begin{bmatrix} v_1^1 & v_2^1 \\ & \\ v_1^2 & v_2^2 \end{bmatrix}.$$

The parallelogram spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2$ coincides with the parallelepiped spanned by these vectors defined in (15.3.4). More precisely, it is the set

$$P(\boldsymbol{v}_1, \boldsymbol{v}_2) = \big\{\, x^1\boldsymbol{v}_1 + x^2\boldsymbol{v}_2;\;\; x^1, x^2 \in [0,1] \,\big\} \subset \mathbb{R}^2. \tag{16.2.1}$$

According to (15.3.5), the area of this parallelogram is equal to the absolute value of the determinant of V,

$$\operatorname{area}\big(\, P(\boldsymbol{v}_1, \boldsymbol{v}_2) \,\big) = \operatorname{vol}_2\big(\, P(\boldsymbol{v}_1, \boldsymbol{v}_2) \,\big) = \big|\, \det V \,\big|.$$

This equality has one "flaw": we need to know the coordinates of the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$. For the applications we have in mind we would like a formula that does involve this knowledge. To achieve this we consider the product of matrices

$$G = G(\boldsymbol{v}_1, \boldsymbol{v}_2) := V^T \cdot V = \begin{bmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle \\ \langle \boldsymbol{v}_2, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_2, \boldsymbol{v}_2 \rangle \end{bmatrix}. \tag{16.2.2}$$

Note two things.

(i) The matrix $G(\boldsymbol{v}_1, \boldsymbol{v}_2)$ called the *Gramian* of $\boldsymbol{v}_1, \boldsymbol{v}_2$, is symmetric, and it is determined only by the scalar products $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$.

(ii) We have

$$\det G = \det V^\top \det V = \big(\, \det V \,\big)^2.$$

We deduce the following very important formula

$$\operatorname{area}\big(\,P(\boldsymbol{v}_1,\boldsymbol{v}_2)\,\big)=\sqrt{\det G(\boldsymbol{v}_1,\boldsymbol{v}_2)}. \tag{16.2.3}$$

Now suppose that $n\in\mathbb{N}$, $n\geqslant 2$, and $\boldsymbol{u}_1,\boldsymbol{u}_2\in\mathbb{R}^n$ are two linearly independent vectors. They define an $n\times 2$ matrix

$$U=[\boldsymbol{u}_1\ \boldsymbol{u}_2]$$

whose columns are the vectors $\boldsymbol{u}_1,\boldsymbol{u}_2$ We define their Gramian $G(\boldsymbol{u}_1,\boldsymbol{u}_2)$ according to formula (16.2.2)

$$G(\boldsymbol{u}_1,\boldsymbol{u}_2):=U^\top U=\left[\begin{array}{cc}\langle\boldsymbol{u}_1,\boldsymbol{u}_1\rangle & \langle\boldsymbol{u}_1,\boldsymbol{u}_2\rangle\\ \langle\boldsymbol{u}_2,\boldsymbol{u}_1\rangle & \langle\boldsymbol{u}_2,\boldsymbol{u}_2\rangle\end{array}\right].$$

The parallelogram spanned by $\boldsymbol{u}_1,\boldsymbol{u}_2$ is defined as in (16.2.1),

$$P(\boldsymbol{u}_1,\boldsymbol{u}_2):=\big\{x^1\boldsymbol{u}_1+x^2\boldsymbol{u}_2;\ \ x^1,x^2\in[0,1]\,\big\}\subset\mathbb{R}^n.$$

We define the area of $P(\boldsymbol{u}_1,\boldsymbol{u}_2)$ by the formula

$$\boxed{\operatorname{area}\big(\,P(\boldsymbol{u}_1,\boldsymbol{u}_2)\,\big):=\sqrt{\det G(\boldsymbol{u}_1,\boldsymbol{u}_2)}}. \tag{16.2.4}$$

For example, if

$$\boldsymbol{u}_1=(1,1,1)\in\mathbb{R}^3,\ \ \boldsymbol{u}_2=(1,0,-1)\in\mathbb{R}^3,$$

Then $\langle\boldsymbol{u}_1,\boldsymbol{u}_1\rangle=3$, $\langle\boldsymbol{u}_1,\boldsymbol{u}_2\rangle=0$, $\langle\boldsymbol{u}_2,\boldsymbol{u}_2\rangle=2$,

$$G(\boldsymbol{u}_1,\boldsymbol{u}_2)=\left[\begin{array}{cc}3&0\\0&2\end{array}\right],\ \ \det G(\boldsymbol{u}_1,\boldsymbol{u}_2)=6,\ \ \operatorname{area}\big(\,P(\boldsymbol{u}_1,\boldsymbol{u}_2)\,\big)=\sqrt{6}.$$

Remark 16.2.1. Let us point one other feature of (16.2.4) that adds extra plausibility to our definition by diktat of the area of a parallelogram in a higher dimensional space.

Recall (see Exercise 11.25) that an orthogonal operator $S:\mathbb{R}^n\to\mathbb{R}^n$ is a linear map S such that

$$\langle S\boldsymbol{u},S\boldsymbol{v}\rangle=\langle\boldsymbol{u},\boldsymbol{v}\rangle,\ \ \forall\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n.$$

Orthogonal operators preserve distances between points. In particular, an orthogonal operator is bijective and it is natural to expect that it will map a parallelogram to another parallelogram with the same area. The area as defined in (16.2.1) displays this orthogonal invariance.

To see this, note that the definition of an orthogonal operator implies that, for any orthogonal operator S, we have

$$G(\boldsymbol{u}_1,\boldsymbol{u}_2)=G(S\boldsymbol{u}_1,S\boldsymbol{u}_2),$$

Note that the image via S of the parallelogram spanned by $\boldsymbol{u}_1,\boldsymbol{u}_2$ is the parallelogram spanned by $S\boldsymbol{u}_1,S\boldsymbol{u}_2$, i.e.,

$$S\big(\,P(\boldsymbol{u}_1,\boldsymbol{u}_2)\,\big)=P(S\boldsymbol{u}_1,S\boldsymbol{u}_2).$$

In particular, we deduce that

$$\operatorname{area}\big(\,SP(\boldsymbol{u}_1,\boldsymbol{u}_2)\,\big)=\operatorname{area}\big(\,P(S\boldsymbol{u}_1,S\boldsymbol{u}_2)\,\big)=\operatorname{area}\big(\,P(\boldsymbol{u}_1,\boldsymbol{u}_2)\,\big).$$

Let us also observe that if $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^n$ are two linearly independent vectors then there exists an orthogonal operator $S : \mathbb{R}^n \to \mathbb{R}^n$ such that the vectors $\boldsymbol{v}_1 := S\boldsymbol{u}_1$ and $\boldsymbol{v}_2 := S\boldsymbol{u}_2$ belong to the subspace $\mathbb{R}^2 \times \mathbf{0} \subset \mathbb{R}^n$.[2] As we have explained at the beginning of this subsection the area of the parallelogram spanned by the vector $\boldsymbol{v}_1, \boldsymbol{v}_2 \subset \mathbb{R}^2$, defined in terms of Riemann integrals *must be given by* (16.2.4). Hence, due to the orthogonal invariance of the Gramian, the area of the parallelogram spanned by $\boldsymbol{u}_1, \boldsymbol{u}_2$ must also be given by (16.2.4). □

16.2.2. Compact surfaces (with boundary). The concept of curve with boundary has a 2-dimensional counterpart.

Definition 16.2.2. Let $k, n \in \mathbb{N}$, $n \geqslant 2$. A C^k*-surface with boundary* in $\mathbb{R}^n$ is a *compact* subset $\Sigma \subset \mathbb{R}^n$ such that, for any point $\boldsymbol{p}_0 \in \Sigma$, there exists an open neighborhood $\mathcal{U}$ of $\boldsymbol{p}_0$ in $\mathbb{R}^n$ and a C^k-diffeomorphism $\Psi : \mathcal{U} \to \mathbb{R}^n$ such that image $\overline{U} := \Psi(\mathcal{U} \cap \Sigma)$ is contained in the subspace $\mathbb{R}^2 \times \mathbf{0} \subset \mathbb{R}^n$ and it is either

(I) an open disk in $\mathbb{R}^2$ centered at $\Psi(\boldsymbol{p}_0)$ or

(B) the point $\Psi(\boldsymbol{p}_0)$ lies on the y-axis and $\overline{U}$ is the intersection of a disk as above with the half-plane

$$H_- := \left\{ (x,y) \in \mathbb{R}^2;\ \ x \leqslant 0 \right\}.$$

The pair $(\mathcal{U}, \Psi)$ is called a *straightening diffeomorphism* at $\boldsymbol{p}_0$. The pair $\left(\mathcal{U} \cap \Sigma, \Psi|_{\mathcal{U}\cap\Sigma} \right)$ is called a *local coordinate chart* of X at $\boldsymbol{p}_0$.

In the case (B), the point $\boldsymbol{p}_0 \in X$ is called a *boundary point* of X. Otherwise $\boldsymbol{p}_0$ is called an *interior* point.

The set of boundary points of X is called the *boundary* of X and it is denoted by ∂X. The set of interior points of X is called the *interior* of X and it is denoted by X°. The surface with boundary is called *closed* if its boundary is empty, $\partial X = \varnothing$. □

Remark 16.2.3. Before we present several examples of surfaces with boundary we want to mention without proofs a few technical facts.

- If $X \subset \mathbb{R}^n$ is a surface with boundary and $\boldsymbol{p}_0$ is a boundary point, then, *for any* straightening diffeomorphism $(\mathcal{U}, \Psi)$ at $\boldsymbol{p}_0$, the image $\Psi(U \cap X)$ is a half-disk B_r^-.
- The boundary of a surface with boundary is a *closed curve.*

□

Example 16.2.4 (Bounded C^k-domains in the plane)**.** Suppose that $D \subset \mathbb{R}^2$ is a bounded C^k domain. For $n \geqslant 2$ we regard $\mathbb{R}^2$ as a subspace of $\mathbb{R}^n$. One can show that

[2] This follows from a simple application of the Gram-Schmidt procedure.

- the closure $\boldsymbol{cl}(D)$ of a bounded C^k domain in $\mathbb{R}^2$ is a C^k-surface with boundary in $\mathbb{R}^n$
- as a subset of $\mathbb{R}^2$, the closure $\boldsymbol{cl}(D)$ is Jordan measurable.

We will not present the proofs of the above claims. □

Example 16.2.5 (Graphs)**.** Suppose that $D \subset \mathbb{R}^2$ is a bounded C^k domain and f is a C^k function defined on some open set U containing the closure of D, $f : U \to \mathbb{R}$. Then the graph of f over D

$$\Gamma_f(D) := \big\{ (x, y, z) \in \mathbb{R}^3;\ \ (x, y) \in D,\ \ z = f(x, y) \big\}$$

is a surface with boundary. Figure 16.16 depicts such a graph.

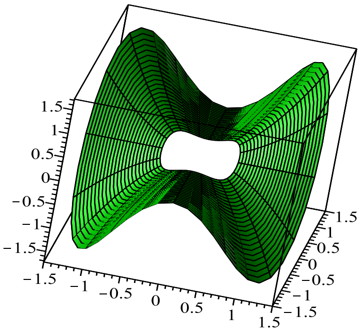

Figure 16.16. *The graph of* $f(x, y) = 2xy^2$ *over the annular domain* $0.5 \leqslant \sqrt{x^2 + y^2} \leqslant 1.5$ *in* $\mathbb{R}^2$.

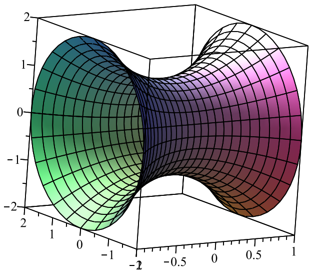

Figure 16.17. *The surface of revolution generated by the graph of the function* $f : [-2, 1] \to (0, \infty)$, $f(x) = x^2 + 1$.

Example 16.2.6 (Surfaces of revolution)**.** Given a C^1-function $f : [a, b] \to (0, \infty)$, then by rotating its graph about the x-axis we obtain a surface with boundary S_f whose boundary consists of two circles; see Figure 16.17. □

Example 16.2.7 (Cutting surfaces transversally by a hypersurface)**.** Suppose that $S \subset \mathbb{R}^n$ is a closed C^k-surface and $f : \mathbb{R}^n \to \mathbb{R}$ is a C^k-function. Suppose that

$$\forall \boldsymbol{x} \in \mathbb{R}^n \;\; f(\boldsymbol{x}) = 0 \Rightarrow \nabla f(\boldsymbol{x}) \neq 0.$$

The zero set

$$Z = \left\{ \boldsymbol{x} \in \mathbb{R}^n; \;\; f(\boldsymbol{z}) = 0 \right\}$$

is a hypersurface of $\mathbb{R}^n$. One expects that the intersection of the surface S with the hypersurface Z is a curve; think e.g. of a plane Z in $\mathbb{R}^3$ intersecting a surface S in $\mathbb{R}^3$.

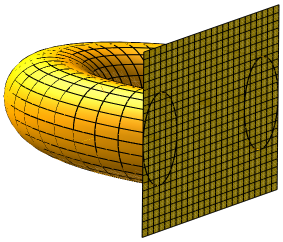

Figure 16.18. *A half-torus in* $\mathbb{R}^3$.

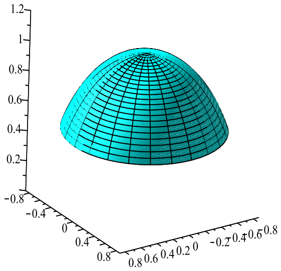

Figure 16.19. *A half-sphere in* $\mathbb{R}^3$.

This intuition is indeed true if this intersection is *transversal*. This means that, for any $\boldsymbol{p} \in S \cap Z$, the gradient vector $\nabla f(\boldsymbol{p})$ is *not perpendicular* to the tangent plane $T_{\boldsymbol{p}}S$. This claim follows from the implicit function theorem, but we will omit the details.

If this transversality condition is satisfied then

$$S_+ := \left\{ \boldsymbol{p} \in S; \;\; f(\boldsymbol{p}) \geqslant 0 \right\}$$

is a surface with boundary $\partial S_+ = S \cap Z$. To get a better hold of this fact, think that S is the surface of a floating iceberg and $f(\boldsymbol{p})$ denotes the altitude of a point. Then, S_+

consists of the points on the iceberg with altitude $\geqslant 0$, i.e., the points on the surface above the water level.

For example we cut the torus in Figure 14.8 with the vertical plane $y = 0$ we obtain the half-torus depicted in Figure 16.18

Also, if we cut the sphere $S = \{x^2 + y^2 + z^2 = 1\}$ with the plane $z = \frac{1}{2}$, then the part above the level $z = \frac{1}{2}$ is the polar cap depicted in Figure 16.19. □

16.2.3. Integrals over surfaces. Let $n \in \mathbb{N}$, $n \geqslant 2$, and suppose that $\Sigma \subset \mathbb{R}^n$ is a C^1-surface. Fix a point $\boldsymbol{p}_0$ and a straightening diffeomorphism $(\mathcal{U}, \Psi)$ near $\boldsymbol{p}_0$; see Definition 14.5.1.

The image of Ψ is an open subset $U \subset \mathbb{R}^n$ and the restriction of Ψ to $\mathcal{U} \cap \Sigma$ is a continuous bijection onto an open subset

$$\overline{U} \subset \mathbb{R}^2 = U \cap \mathbb{R}^2 \times \mathbf{0} \subset \mathbb{R}^n.$$

Its inverse $\Psi^{-1} : U \to \mathcal{U}$ is a C^1-map and it induces a C^1-map $\Phi : \overline{U} \to \mathbb{R}^n$ such that $\Phi(\overline{U}) = \mathcal{U} \cap \Sigma$. The map Φ is called the *local parametrization* of the surface Σ associated to the straightening diffeomorphism (U, Ψ).

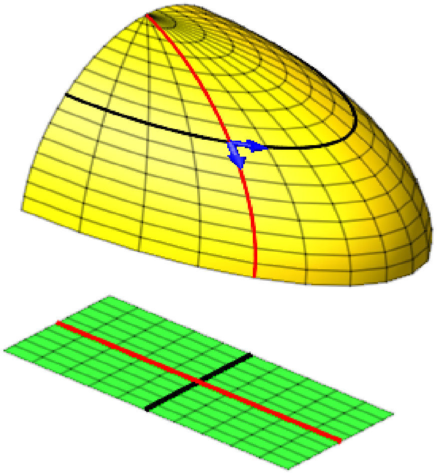

Figure 16.20. *A local parametrization deforms a (flat) planar region to a patch of (curved) surface.*

The local parametrization Φ deforms the flat planar region $\overline{U}$ to a patch $\Phi(\overline{U})$ of the curved surface Σ; see Figure 16.20. The region $\overline{U}$ lies in the two-dimensional vector space $\mathbb{R}^2$, and we denote by (s, t) the coordinates in this space. Moreover, it may help to think of the plane $\mathbb{R}^2$ as made of horizontal and vertical lines woven together. These lines form a grid dividing the plane into infinitesimally small rectangular patches of size $ds \times dt$; see Figure 16.21.

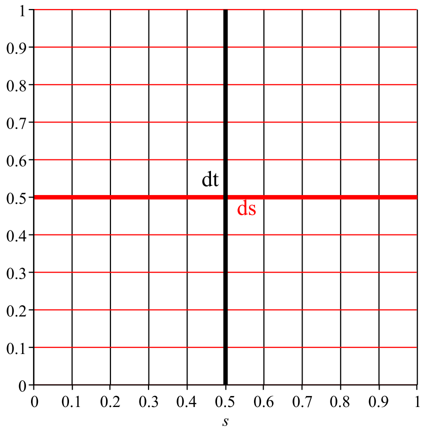

Figure 16.21. *A planar infinitesimal grid.*

The parametrization Φ takes the rectangular (s,t)-grid to a curvilinear grid on the surface Σ; see Figure 16.20. For any point $\boldsymbol{p}$ in the patch $\Phi(\overline{U}) \subset \Sigma$ there exists a *unique* point $(s,t) \in \overline{U}$ such that $\boldsymbol{p} = \Phi(s,t)$. We will write this in a simplified form as

$$\boldsymbol{p} = \boldsymbol{p}(s,t).$$

If we keep t fixed and vary s we get a (red) horizontal line in the (s,t)-plane; see Figure 16.20 and 16.21. This (red) line is mapped by Φ to a (red) curve on Σ. Similarly, if we keep s fixed and vary t we get a vertical line in the (s,t)-plane mapped to a curve on Σ.

Now take a tiny rectangular patch of size $ds \times dt$ with lower left-hand corner situated at some point (s,t). The map Φ sends it to a tiny (infinitesimal) patch of the surface. Because this is so small we can assume it is almost flat and we can approximate it with the parallelogram spanned by the vectors (see Figure 16.20).

$$\boldsymbol{p}'_s ds = \Phi'_s ds \;\text{ and }\; \boldsymbol{p}'_t dt = \Phi'_t dt.$$

The area of this infinitesimal tangent parallelogram is usually referred to as the *area element*. It is denoted by $|dA|$ and we have

$$|dA| = \sqrt{\det G(\boldsymbol{p}'_s ds, \boldsymbol{p}'_t dt)} = \sqrt{\det G(\boldsymbol{p}'_s, \boldsymbol{p}'_t)}\ |ds dt|.$$

More intuitively, the parametrization Φ identifies a point $\boldsymbol{p} \in \mathfrak{U} \cap \Sigma$ with a point $(s,t) \in \overline{U} \subset \mathbb{R}^2$. We write this $\boldsymbol{p} = \boldsymbol{p}(s,t)$ and we rewrite the above equality

$$|dA| = \sqrt{\det G(\boldsymbol{p}'_s, \boldsymbol{p}'_t)}\ |ds dt|.$$

The quantity $\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\;|dsdt|$ describes the area element in the s,t coordinates. We provisorily define the area of $\mathcal{U}\cap\Sigma$ to be

$$\boxed{\operatorname{area}(\mathcal{U}\cap\Sigma)=\int_{\mathcal{U}\cap\Sigma}|dA|:=\int_{\overline{U}}\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\;|dsdt|}. \tag{16.2.5}$$

Suppose that $(\mathcal{V},\hat{\Psi})$ is another straightening diffeomorphism of Σ near $\boldsymbol{p}_0$ such that $\mathcal{U}\cap\Sigma=\mathcal{V}\cap\Sigma$. We set

$$S:=\mathcal{U}\cap\Sigma=\mathcal{V}\cap\Sigma.$$

The restriction of $\hat{\Psi}$ to $\mathcal{V}\cap\Sigma$ is a continuous bijection onto an open subset

$$\overline{V}\subset\mathbb{R}^2=\mathbb{R}^2\times\boldsymbol{0}\subset\mathbb{R}^n.$$

Its inverse is given by a C^1-map $\hat{\Phi}:\overline{V}\to\mathbb{R}^n$ such that $\hat{\Phi}(\overline{V})=\mathcal{U}\cap\Sigma$. We denote by (u,v) the Euclidean coordinates in the plane $\mathbb{R}^2$ that contains $\overline{V}$. A point $\boldsymbol{p}\in S$ can now be identified either with a point $(s,t)\in\overline{U}$ or with a point $(u,v)\in\overline{V}$. We write this $\boldsymbol{p}=\boldsymbol{p}(s,t)$ and respectively $\boldsymbol{p}=\boldsymbol{p}(u,v)$.

We obtain in this fashion a map $\overline{U}\mapsto\overline{V}$ that associates to the point $(s,t)\in\overline{U}$ the unique point $(u,v)\in\overline{V}$ such that $\boldsymbol{p}(s,t)=\boldsymbol{p}(u,v)$. Formally, this is the compostion of C^1-maps $\hat{\Psi}\circ\Phi$. In particular, this is a C^1-map.

We will indicate this correspondence $(s,t)\mapsto(u,v)$ by writing $u=u(s,t)$ and $v=v(s,t)$. To recap

$$u=u(s,t),\;\;v=v(s,t)\Longleftrightarrow\boldsymbol{p}(s,t)=\boldsymbol{p}(u,v).$$

We now have two possible definitions of area(S)

$$\operatorname{area}(S)=\int_{\overline{U}}\sqrt{G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\;|dsdt|\;\;\text{or}\;\;\operatorname{area}(S)=\int_{\overline{V}}\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\;|dudv|.$$

We want to show that they both yield the same result.

Lemma 16.2.8.

$$\int_{\overline{U}}\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\;|dsdt|=\int_{\overline{V}}\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\;|dudv|.$$

Proof. We argue as in the proof of the equality (16.1.1). The key fact is the equality

$$\boldsymbol{p}(s,t)=\boldsymbol{p}(u,v).$$

Differentiating with respect to s,t we deduce

$$\boldsymbol{p}'_s=\boldsymbol{p}'_u u'_s+\boldsymbol{p}'_v v'_s,\;\;\boldsymbol{p}'_t=\boldsymbol{p}'_u u'_t+\boldsymbol{p}'_v v'_t \tag{16.2.6}$$

Let us introduce the $n\times 2$ matrices

$$P_{s,t}:=[\boldsymbol{p}'_s\;\boldsymbol{p}'_t],\;\;P_{u,v}:=[\boldsymbol{p}'_u\;\boldsymbol{p}'_v].$$

Thus the columns of $P_{s,t}$ consist of the vectors $\boldsymbol{p}'_s,\boldsymbol{p}'_t\in\mathbb{R}^n$, and the columns of $P_{u,v}$ consist of the vectors $\boldsymbol{p}'_u,\boldsymbol{p}'_v\in\mathbb{R}^n$ We can rewrite (16.2.6)

$$P_{s,t}=P_{u,v}\cdot\underbrace{\begin{bmatrix}u'_s & u'_t\\ v'_s & v'_t\end{bmatrix}}_{J}.$$

We note that J is the Jacobian matrix of the transformation $(s,t)\mapsto(u,v)$

$$J=\frac{\partial(u,v)}{\partial(s,t)}.$$

Then

$$G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)=P^T_{s,t}P_{s,t}=J^TP^T_{u,v}P_{u,v}J=J^TG(\boldsymbol{p}'_u,\boldsymbol{p}'_v)J$$

so

$$\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)=(\det J^T)\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)(\det J)=\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)(\det J)^2,$$

and thus

$$\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}=\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\cdot|\det J|.$$

The change in variables formula then implies

$$\int_V\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\,|dudv|=\int_U\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\left|\det\frac{\partial(u,v)}{\partial(s,t)}\right||dsdt|$$
$$=\int_U\sqrt{\det G(\boldsymbol{p}'_u,\boldsymbol{p}'_v)}\cdot|\det J|\,|dsdt|=\int_U\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\,|dsdt|.$$

□

Definition 16.2.9 (Convenient surfaces with boundary)**.** A *parametrized surface with boundary* is a triplet (Σ,D,Φ) with the following properties.

- $D\subset\mathbb{R}^2$ is a bounded domain in $\mathbb{R}^2$ with C^1-boundary;
- Φ is an injective immersion $\Phi:U\to\mathbb{R}^n$, where U is an open subset of $\mathbb{R}^2$ containing the closure $\boldsymbol{cl}(D)$ of D.
- $\Sigma=\Phi\big(\,\boldsymbol{cl}(D)\,\big)$.

The induced map $\Phi:\boldsymbol{cl}(D)\to\Sigma$ is called a *parametrization* of Σ. A surface with boundary is called *convenient* if it can be parametrized as above. □

For example, the surfaces depicted in Figure 16.16, 16.18, 16.19 and 16.17 are convenient.

Suppose now that $\Sigma\subset\mathbb{R}^n$ is a convenient surface with boundary with a parametrization $\Phi:\boldsymbol{cl}(D)\to\Sigma$. Here $D\subset\mathbb{R}^2$ is a bounded domain with C^1-boundary. If we denote by s,t the Euclidean coordinates in the plane $\mathbb{R}^2$ containing D, then we can describe the parametrization Φ as describing point $\boldsymbol{p}$ on Σ depending on the coordinates,

$$\boldsymbol{p}=\boldsymbol{p}(s,t).$$

If $f:\Sigma\to\mathbb{R}$ is a function on Σ, then we say that it is integrable over Σ if the function

$$D\ni(s,t)\mapsto f\big(\,\boldsymbol{p}(s,t)\,\big)\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}$$

is Riemann integrable. If this is the case, then we define *the integral of f over Σ* to be the number

$$\boxed{\int_\Sigma f(\boldsymbol{p})|dA(\boldsymbol{p})|:=\int_D f\big(\,\boldsymbol{p}(s,t)\,\big)\sqrt{\det G(\boldsymbol{p}'_s,\boldsymbol{p}'_t)}\,|dsdt|}.$$

The left-hand side of the above equality makes no reference of any parametrization while the right-hand side is obviously described in terms of a concrete parametrization. We want to show, that if we change the parametrization, then the resulting right-hand side does not change its value, although it might look dramatically different.

If $\overline{\Phi}:\overline{D}\to\Sigma$ is another parametrization of Σ,

$$\overline{D}\in(u,v)\mapsto\boldsymbol{p}(u,v)=\overline{\Phi}(u,v),$$

then Lemma 16.2.8 shows that

$$\int_D f\big(\boldsymbol{p}(s,t)\big)\sqrt{\det G(\boldsymbol{p}_s',\boldsymbol{p}_t')}\,|dsdt|=\int_D f\big(\boldsymbol{p}(u,v)\big)\sqrt{\det G(\boldsymbol{p}_u',\boldsymbol{p}_v')}\,|dudv|.$$

Example 16.2.10 (Integration along graphs)**.** Suppose that $D\subset\mathbb{R}^2$ is a bounded domain with C^1-boundary, and $h:\boldsymbol{cl}(D)\to\mathbb{R}$ is a C^1-function. Its graph

$$\Gamma_f:=\big\{(x,y,z);\ \ (x,y)\in\boldsymbol{cl}(D),\ z=h(x,y)\,\big\}$$

is a convenient surface with boundary. The map

$$(x,y)\mapsto\boldsymbol{p}(x,y):=\big(\,x,y,h(x,y)\,\big)\in\mathbb{R}^3$$

is a parametrization of Γ_h. We have

$$\boldsymbol{p}_x'=\big(\,1,0,h_x'(x,y)\,\big),\ \ \boldsymbol{p}_y'=\big(\,0,1,h_y'(x,y)\,\big),$$

$$\langle\boldsymbol{p}_x',\boldsymbol{p}_x'\rangle=1+\big|h_x'\big|^2,\ \ \langle\boldsymbol{p}_y',\boldsymbol{p}_y'\rangle=1+\big|h_y'\big|^2,\ \ \langle\boldsymbol{p}_x',\boldsymbol{p}_y'\rangle=h_x'h_y',$$

$$\begin{aligned}\det G(\boldsymbol{p}_x',\boldsymbol{p}_y')=\langle\boldsymbol{p}_x',\boldsymbol{p}_x'\rangle\cdot\langle\boldsymbol{p}_y',\boldsymbol{p}_y'\rangle-\langle\boldsymbol{p}_x',\boldsymbol{p}_y'\rangle^2&=\big(1+\big|h_x'\big|^2\big)\big(1+\big|h_y'\big|^2\big)-\big(\,h_x'h_y'\big)^2\\&=1+\big|h_x'\big|^2+\big|h_y'\big|^2=1+\big\|\nabla h\big\|^2.\end{aligned}$$

Hence, the area element on the graph of h is

$$\boxed{|dA|=\sqrt{1+\big\|\nabla h\big\|^2}\,|dxdy|}.$$

In particular, the area of Γ_h is

$$\operatorname{area}(\Gamma_h)=\int_D\sqrt{1+\big\|\nabla h\big\|^2}\,|dxdy|.\qquad\square$$

Example 16.2.11 (Revolving graphs about the z-axis)**.** Suppose that $0<a<b$ and $f:[a,b]\to\mathbb{R}$ is a C^1-function. We denote by S_f the surface in $\mathbb{R}^3$ obtained by revolving the graph of f in the (x,z) plane,

$$\Gamma_f=\big\{(x,z);\ \ x\in[a,b],\ \ z=f(x)\,\big\}$$

about the z-axis. Denote by D the annulus in the (x,y)-plane described by the condition

$$a\leqslant r\leqslant b,\ \ r=\sqrt{x^2+y^2}.$$

Then S_f is the graph of the function

$$\hat{f}:D\to\mathbb{R},\ \ \hat{f}(x,y)=f(r)=f\big(\sqrt{x^2+y^2}\,\big).$$

Note that

$$\hat{f}'_x = f'(r)\frac{x}{r},\ \ \hat{f}'_y = f'(r)\frac{y}{r},\ \ 1 + \|\nabla\hat{f}\|^2 = 1 + |f'(r)|^2.$$

Thus

$$\boxed{|dA| = \sqrt{1 + |f'(r)|^2}}\,|dxdy|.$$

□

Suppose in general that Σ is a compact surface with, or without boundary, and

$$f : \Sigma \to \mathbb{R}$$

is a continuous function. We want to define the integral of f over Σ. We distinguish two cases.

Case 1. Let $\boldsymbol{p}_0 \in \Sigma$ and suppose that (U, Ψ) is a straightening diffeomorphism at $\boldsymbol{p}_0$ (see Definition 16.2.2). This induces a homeomorphism

$$\Psi : U \cap \Sigma \to D = \Psi(U \cap \Sigma) \subset \mathbb{R}^2.$$

where D is either an open disk or an open half-disk. We denote by (s, t) the Euclidean coordinates on $\mathbb{R}^2$. If

$$\boxed{\operatorname{supp} f \subset U}, \tag{16.2.7}$$

then we define

$$\int_\Sigma f\,|dA| = \int_D f\big(\boldsymbol{p}(s,t)\big)\sqrt{\det G(\boldsymbol{p}'_s, \boldsymbol{p}'_t)}\,|dsdt|.$$

Case 2. Let us explain how to deal with the general case, when the support of f is not necessarily contained in the domain of a straightening diffeomorphism as in (16.2.7).

Fix an *atlas* of Σ, i.e., a collection

$$\Big\{ (U_\alpha, \Psi_\alpha) \Big\}_{\alpha\in\mathcal{A}}$$

where each (U_α, Ψ_α) is a straightening diffeomorphism of Σ at a point $\boldsymbol{p}_\alpha \in \Sigma$ and

$$\Sigma \subset \bigcup_{\alpha\in\mathcal{A}} U_\alpha.$$

Now choose a continuous partition of unity $\chi_1, \dots, \chi_N$ subordinated to the open cover $(U_\alpha)_{\alpha\in\mathcal{A}}$ of Σ. Thus, each χ_i is a compactly supported continuous function $\mathbb{R}^n \to \mathbb{R}$ and there exists $\alpha(i) \in \mathcal{A}$ such that

$$\operatorname{supp}\chi_i \subset U_{\alpha(i)}. \tag{16.2.8}$$

Moreover

$$\chi_1(\boldsymbol{p}) + \cdots + \chi_N(\boldsymbol{p}) = 1,\ \ \forall \boldsymbol{p} \in \Sigma.$$

In particular

$$f(\boldsymbol{p}) = \chi_1(\boldsymbol{p})f(\boldsymbol{p}) + \cdots + \chi_N(\boldsymbol{p})f(\boldsymbol{p}),\ \ \forall \boldsymbol{p} \in \Sigma.$$

Due to the inclusions (16.2.8), each of the functions $\chi_i f$ satisfies the support condition (16.2.7). We define

$$\boxed{\int_\Sigma f|dA| := \sum_{i=1}^N \int_\Sigma \chi_i f|dA|},$$

where each of the integrals in the right-hand side are defined as in **Case 1**.

Clearly, the definition used in Case 2 depends on several choices.

- A choice of atlas, i.e., a collection $\big\{ (U_\alpha, \Psi_\alpha); \;\; \alpha \in \mathcal{A} \big\}$ of local charts covering Σ.
- A choice of a partition of unity subordinated to the open cover $(U_\alpha)_{\alpha\in\mathcal{A}}$ of Σ.

One can show that the end result is independent of these choices. We omit the proof of this fact. This type of integral over a surface is traditionally known as *surface integral of the first kind*.

The above definition of the integral of a continuous function over a compact surface is not very useful for concrete computations. In concrete situations surfaces are *quasi-parametrized*.

Definition 16.2.12. Suppose that $\Sigma \subset \mathbb{R}^n$ is a compact C^1-surface with or without boundary. A *quasi-parametrization* of Σ is an injective C^1-immersion $\Phi : D \to \mathbb{R}^n$, where $D \subset \mathbb{R}^2$ is a bounded domain, such that $\Phi(D)$ *almost covers* Σ, i.e., $\Phi(D) \subset \Sigma$ and $\Sigma \backslash \Phi(D)$ is a finite union of compact C^1-curves with or without boundary. □

The next result extends Theorem 15.3.5 to "curved" situations.

Theorem 16.2.13. *Let $n \in \mathbb{N}$ and suppose that $\Sigma \subset \mathbb{R}^n$ is a compact surface, with or without boundary. Suppose that*

$$\Phi : D \to \mathbb{R}^n, \;\; (s,t) \mapsto \boldsymbol{p}(s,t) := \Phi(s,t) \in \mathbb{R}^n$$

is quasi-parametrization of Σ. Then, for any continuous function $f : \Sigma \to \mathbb{R}$ we have

$$\int_\Sigma f(\boldsymbol{p})\,|dA(\boldsymbol{p})| = \int_D f\big(\boldsymbol{p}(s,t)\big)\sqrt{\det G(\boldsymbol{p}'_s, \boldsymbol{p}'_t)}\,|dsdt|.$$ □

Example 16.2.14. Consider the unit sphere

$$S := \big\{ (x,y,z) \in \mathbb{R}^3; \;\; x^2 + y^2 + z^2 = 1 \big\}.$$

In spherical coordinates (ρ, φ, θ) this sphere is described by the equation $\rho = 1$.

The spherical coordinates define an injective immersion

$$(0,\pi) \times (0, 2\pi) \ni (\varphi, \theta) \mapsto \boldsymbol{p}(\varphi,\theta) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi)$$

that almost covers S: its image is the complement in the sphere of the "meridian" obtained by intersecting the sphere with the half-plane

$$y = 0, x \geqslant 0.$$

Thus this map almost covers S. Note that

$$\boldsymbol{p}'_\varphi = (\cos\varphi\cos\theta, \cos\varphi\sin\theta, -\sin\varphi),\ \ \boldsymbol{p}'_\theta = (-\sin\varphi\sin\theta, \sin\varphi\cos\theta, 0).$$

Then

$$\langle \boldsymbol{p}'_\varphi, \boldsymbol{p}'_\varphi \rangle = (\cos\varphi\cos\theta)^2 + (\cos\varphi\sin\theta)^2 + \sin^2\varphi = 1,$$

$$\langle \boldsymbol{p}'_\varphi, \boldsymbol{p}'_\theta \rangle = 0,\ \ \langle \boldsymbol{p}'_\theta, \boldsymbol{p}'_\theta \rangle = (\sin\varphi\sin\theta)^2 + (\sin\varphi\cos\theta)^2 = \sin^2\varphi,$$

so that

$$\sqrt{\det G(\boldsymbol{p}'_\varphi, \boldsymbol{p}'_\theta)} = \sin\varphi.$$

This shows that, *in spherical coordinates, the area element of the sphere is*

$$\boxed{|dA(\boldsymbol{p})| = \sin\varphi|d\varphi d\theta|}.$$

If $f : S \to \mathbb{R}$ is a continuous function, then

$$\int_S f(\boldsymbol{p})\,|dA(\boldsymbol{p})| = \int_{\substack{0\leqslant\theta\leqslant,2\pi,\\ 0\leqslant\varphi\leqslant\pi}} f(\varphi,\theta)\sin\varphi\,|d\varphi d\theta|.$$

In particular

$$\operatorname{area}(S) = \int_{\substack{0\leqslant\theta\leqslant2\pi,\\ 0\leqslant\varphi\leqslant\pi}} \sin\varphi\,|d\varphi d\theta| = \underbrace{\left(\int_0^{2\pi} d\theta\right)}_{=2\pi}\underbrace{\left(\int_0^{\pi}\sin\varphi d\varphi\right)}_{=2} = 4\pi. \qquad \square$$

Example 16.2.15. Consider again the unit sphere

$$S := \{\, (x,y,z) \in \mathbb{R}^3;\ \ x^2 + y^2 + z^2 = 1 \,\}.$$

Fix a real number $c \in (0,1)$ and consider the polar cap

$$S_c := \{\, (x,y,z) \in S;\ \ z \geqslant c \,\}.$$

We want to compute the area of this surface with boundary. This time we will use cylindrical coordinates (r,θ,z). In cylindrical coordinates the sphere is described by the equation

$$r^2 + z^2 = 1.$$

In the northern hemisphere $z > 0$ we have

$$z = \sqrt{1-r^2} = \sqrt{1-x^2-y^2}.$$

Along the polar cap we have $z \geqslant c$ so

$$\sqrt{1-r^2} \geqslant c \Rightarrow 1 - r^2 \geqslant c^2 \Rightarrow r^2 \leqslant 1 - c^2 \Rightarrow s \leqslant \sqrt{1-c^2}.$$

Thus the polar cap admits the quasi-parametrization

$$\boldsymbol{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ \sqrt{1-r^2} \end{bmatrix}, \;\; 0 \leqslant r \leqslant \sqrt{1-c^2}, \;\; \theta \in [0,2\pi].$$

Then

$$\boldsymbol{p}_r' = \begin{bmatrix} \cos\theta \\ \sin\theta \\ -\frac{r}{\sqrt{1-r^2}} \end{bmatrix}, \;\; \boldsymbol{p}_\theta' = \begin{bmatrix} -r\sin\theta \\ r\cos\theta \\ 0 \end{bmatrix},$$

$$\langle \boldsymbol{p}_r', \boldsymbol{p}_r' \rangle = 1 + \frac{r^2}{1-r^2} = \frac{1}{1-r^2}, \;\; \langle \boldsymbol{p}_\theta', \boldsymbol{p}_\theta' \rangle = r^2, \;\; \langle \boldsymbol{p}_r', \boldsymbol{p}_\theta' \rangle = 0$$

$$\det G(\boldsymbol{p}_r', \boldsymbol{p}_\theta') = \langle \boldsymbol{p}_r', \boldsymbol{p}_r' \rangle \cdot \langle \boldsymbol{p}_\theta', \boldsymbol{p}_\theta' \rangle = \frac{r^2}{1-r^2},$$

Then, in polar coordinates (r,θ) the area on the unit sphere is given by

$$|dA| = \frac{r}{\sqrt{1-r^2}}|drd\theta|,$$

and we deduce

$$\operatorname{area}(S_c) = \int_{\substack{0 \leqslant r \leqslant \sqrt{1-c^2} \\ \theta \in [0,2\pi]}} \frac{r}{\sqrt{1-r^2}} drd\theta$$

$$= 2\pi \int_0^{\sqrt{1-c^2}} \frac{r}{\sqrt{1-r^2}} dr = -2\pi \int_0^{\sqrt{1-c^2}} \frac{d(1-r^2)}{2\sqrt{1-r^2}}$$

$$= -2\pi \left(\sqrt{1-r^2} \Big|_{r=0}^{r=\sqrt{1-c^2}} \right) = 2\pi(1-c). \qquad \square$$

Example 16.2.16. Suppose that $g : [a,b] \to (0,\infty)$ is a C^1-function. We denote by $S_g \subset \mathbb{R}^3$ the surface obtained by rotating the graph of g about the x-axis. Using polar coordinates in the (y,z)-plane

$$y = r\cos\theta, \;\; z = r\sin\theta$$

we can describe S_g via the equation $r = g(x)$. The injective immersion

$$(0,2\pi) \times (a,b) \ni (\theta, x) \mapsto \boldsymbol{p}(\theta, x) = \big(x, g(x)\cos\theta, g(x)\sin\theta \big) \in \mathbb{R}^3$$

almost covers S_g. We have

$$\boldsymbol{p}_\theta' = \big(0, -g(x)\sin\theta, g(x)\cos\theta \big), \;\; \boldsymbol{p}_x' = \big(1, g'(x)\cos\theta, g'(x)\sin\theta \big)$$

$$\langle \boldsymbol{p}_\theta', \boldsymbol{p}_\theta' \rangle = g(x)^2, \;\; \langle \boldsymbol{p}_\theta', \boldsymbol{p}_x' \rangle = 0,$$

$$\langle \boldsymbol{p}_x', \boldsymbol{p}_x' \rangle = 1 + |g'(x)|^2,$$

$$\det G(\boldsymbol{p}_\theta', \boldsymbol{p}_x') = g(x)^2 \big(1 + |g'(x)|^2 \big)$$

so

$$|dA(\boldsymbol{p})| = g(x)\sqrt{1 + |g'(x)|^2}\, |dxd\theta|.$$

In particular

$$\operatorname{area}(S_g)=\int_{\substack{0\leqslant\theta\leqslant 2\pi,\\ a\leqslant x\leqslant b}} g(x)\sqrt{1+|g'(x)|^2}\,|dxd\theta=\int_0^{2\pi}d\theta\int_a^b g(x)\sqrt{1+|g'(x)|^2}dx$$

$$=2\pi\int_a^b g(x)\sqrt{1+|g'(x)|^2}dx.$$

This is in perfect agreement with the equality (9.8.4) obtained by alternate means.

For example, if $g(x)=R>0$, $\forall x\in[a,b]$, then S_g is the cylinder

$$C_R:=\big\{(x,y,z)\in\mathbb{R}^3;\;\; y^2+z^2=R^2,\;\; x\in[a,b]\,\big\}.$$

Thus

$$\int_{C_R} f(x,y,z)|dA|)=\int_{\substack{a\leqslant x\leqslant b\\ 0\leqslant\theta\leqslant 2\pi}} f\big(\,x,R\cos\theta,R\sin\theta\,\big)\,|dxd\theta|.$$ □

Example 16.2.17. Consider the hypersurface in $\mathbb{R}^3$ given by the equation

$$x^2+y^2=2x.$$

Note that we can rewrite this as

$$x^2-2x+1+y^2=1\Longleftrightarrow(x-1)^2+y^2=1$$

showing that this is a cylinder with radius 1 and vertical axis passing through the point $(1,0,0)$. We want to compute the area of the portion Σ of this cylinder that lies inside the sphere

$$x^2+y^2+z^2=4.$$

We observe first that Σ is symmetric with respect to the plane (x,y). We set

$$\Sigma_+=\big\{\,(x,y,z)\in\Sigma;\;\; z\geqslant 0\,\big\}\;\;\Sigma_-=\big\{\,(x,y,z)\in\Sigma;\;\; z\leqslant 0\,\big\}.$$

Due to the above symmetry of Σ we have

$$\operatorname{area}(\Sigma_+)=\operatorname{area}(\Sigma_-)=\frac{1}{2}\operatorname{area}(\Sigma).$$

Using polar coordinates about $(1,0)$ we obtain the quasiparametrization of Σ_+

$$x=1+\cos\theta,\;\; y=\sin\theta, z=z,$$

where

$$\theta\in(0,2\pi),\;\; 0<z<\sqrt{4-x^2-y^2}=\sqrt{4-2x}=\sqrt{2-2\cos\theta}=2|\sin\theta|.$$

We have

$$\boldsymbol{p}_\theta'=(-\sin\theta,\cos\theta,0),\;\;\boldsymbol{p}_z'=(0,0,1),$$

,

$$\det G(\boldsymbol{p}_\theta',\boldsymbol{p}_z')=0,\;\;\operatorname{area}(\Sigma_+)=\int_0^{2\pi}|\sin\theta|\,d\theta=4.$$

Hence $\operatorname{area}(\Sigma)=8$. □

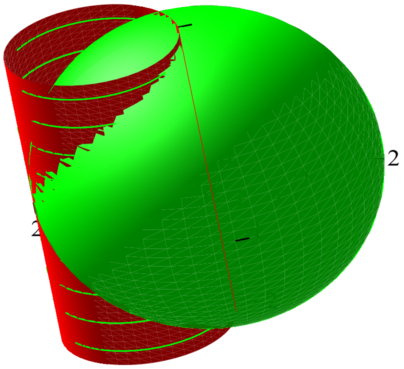

Figure 16.22. *Intersecting a cylinder with a sphere.*

16.2.4. Orientable surfaces in $\mathbb{R}^3$. Suppose that $\Sigma \subset \mathbb{R}^3$ is a surface in $\mathbb{R}^3$. Roughly speaking a surface is orientable if it has two sides. For example, the xy-plane in $\mathbb{R}^3$ is orientable: it has one side facing the positive part of the z-axis, and one side facing the negative part of the z-axis. *Not all surfaces are two-sided.* The most famous and arguably the most important example is the *Möbius strip* (or band) depicted in Figure 16.23. It can be described by the parametrization

$$\begin{aligned} x &= \big(3 + r\cos(t/2)\,\big)\cos(t)\,, \\ y &= \big(3 + r\cos(t/2)\,\big)\sin(t)\,, \quad -1 < r < 1,\;\; 0 \leqslant t \leqslant 2\pi. \\ z &= r\sin(t/2)\,, \end{aligned}$$

Definition 16.2.18. Let $\Sigma \subset \mathbb{R}^3$ be a surface in $\mathbb{R}^3$, with or without boundary. An *orientation on* Σ is a choice of a continuous, unit-normal vector field along Σ, i.e., a continuous map $\boldsymbol{\nu} : \Sigma \to \mathbb{R}^3$ satisfying the following conditions

(i) $\|\boldsymbol{\nu}(\boldsymbol{p})\| = 1$, $\forall \boldsymbol{p} \in \Sigma$.

(ii) $\boldsymbol{\nu}(\boldsymbol{p}) \perp T_{\boldsymbol{p}}\Sigma$, $\forall \boldsymbol{p} \in \Sigma^{\circ}$.[3]

The surface Σ is called *orientable* if it admits an orientation. An *oriented surface* is a pair $(\Sigma, \boldsymbol{\nu})$ consisting of a surface $\Sigma \subset \mathbb{R}^3$ and an orientation $\boldsymbol{\nu}$ on Σ. □

Intuitively, the unit-normal vector field defining an orientation points towards one side of the surface.

[3] We recall that Σ^0 denotes the interior of Σ.

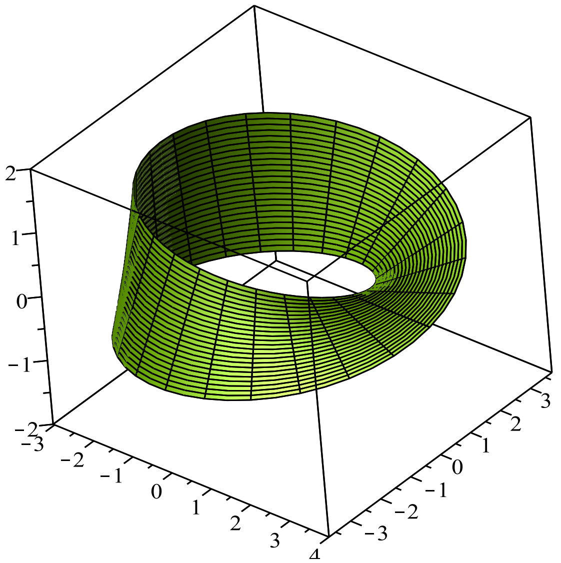

Figure 16.23. *The Möbius strip (band) is the prototypical example of non-orientable surface.*

Example 16.2.19. Suppose that $f : \mathbb{R}^3 \to \mathbb{R}$ is a C^1-function. We denote by Z its zero set,

$$Z = \left\{ \boldsymbol{p} \in \mathbb{R}^3; \;\; f(\boldsymbol{p}) = 0 \right\}.$$

Suppose that

$$\nabla f(\boldsymbol{p}) \neq \boldsymbol{0}, \;\; \forall \boldsymbol{p} \in Z.$$

The implicit function theorem shows that Z is a surface in $\mathbb{R}^3$. It is orientable because the vector field

$$\boldsymbol{\nu} : Z \to \mathbb{R}^3, \;\; \boldsymbol{\nu}(\boldsymbol{p}) = \frac{1}{\|\nabla f(\boldsymbol{p})\|} \nabla f(\boldsymbol{p}),$$

is an orientation on Z. □

Example 16.2.20. Suppose that $U \subset \mathbb{R}^3$ is a bounded domain with C^1-boundary. Then its boundary is orientable. The *induced orientation* of the boundary is that defined by the unit normal vector field that points towards *the exterior* of U. We will use the notation $\partial_+ U$ when referring to the boundary of U equipped with this induced orientation. □

Important orientation convention. Suppose that $\Sigma \subset \mathbb{R}^3$ is a compact C^1-surface with nonempty boundary $\partial\Sigma$. Fix an orientation on Σ described by the normal unit vector field $\boldsymbol{\nu} : \Sigma \to \mathbb{R}^3$. Then we can equip the boundary with an orientation as follows: a person traveling on $\partial\Sigma$ according to this orientation while the toe-to-head direction is given by $\boldsymbol{\nu}$ will notice the surface Σ to her left-hand side; see Figure 16.24. This orientation of $\partial\Sigma$ is called the *orientation induced by the orientation of* Σ.

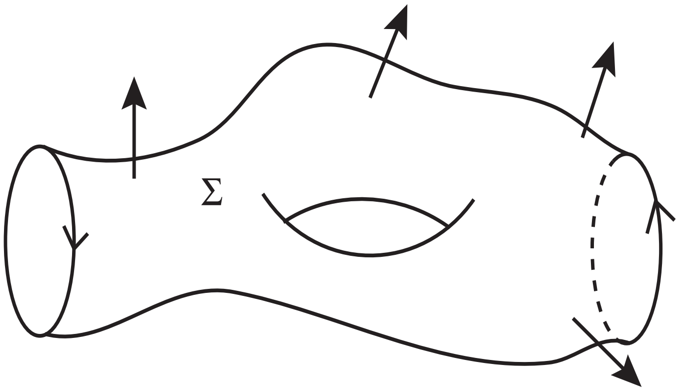

Figure 16.24. *An orientation on a surface induces in a natural way an orientation on its boundary.*

Remark 16.2.21. Observe that if $D \subset \mathbb{R}^2$ is a bounded C^1-domain, then it is also a surface with boundary. The constant unit vector field $\boldsymbol{k}$ along D defines an orientation on D which in turn induces an orientation on the boundary ∂D. This orientation coincides with the induced orientation as described at page 604. □

16.2.5. The flux of a vector field through an oriented surface in $\mathbb{R}^3$.

Definition 16.2.22 (Flux a vector field)**.** Suppose that $\Sigma \subset \mathbb{R}^3$ is compact surface (with or without boundary) and $\boldsymbol{\nu}$ is an orientation on Σ. Suppose that $\boldsymbol{F} : \Sigma \to \mathbb{R}^3$ is a continuous vector field along Σ. The *flux* of $\boldsymbol{F}$ in the direction defined by the orientation is the scalar

$$\operatorname{Flux}(\boldsymbol{F}, \Sigma, \boldsymbol{\nu}) := \int_\Sigma \langle \boldsymbol{F}(\boldsymbol{p}), \boldsymbol{\nu}(\boldsymbol{p}) \rangle |dA(\boldsymbol{p})|.$$ □

Remark 16.2.23 (How one computes the flux of a vector field trough an oriented surface)**.** Suppose $\Sigma \subset \mathbb{R}^3$ is a compact surface (with or without boundary), $\boldsymbol{\nu}$ is an orientation on Σ and $\boldsymbol{F}$ is a continuous vector field along Σ.

$$\Sigma \ni \boldsymbol{p} = (x, y, z) \mapsto \boldsymbol{F}(\boldsymbol{p}) = P(x, y, z)\boldsymbol{i} + Q(x, y, z)\boldsymbol{j} + R(x, y, z)\boldsymbol{k} = \begin{bmatrix} P(x, y, z) \\ Q(x, y, z) \\ R(x, y, z) \end{bmatrix}. \tag{16.2.9}$$

To compute the flux $\operatorname{Flux}(\boldsymbol{F}, \Sigma, \boldsymbol{\nu})$ one typically proceeds as follows.

Step 1. Parametrizing. Fix a quasi-parametrization of Σ, $\Phi : D \to \mathbb{R}^3$, D bounded open domain of $\mathbb{R}^2$,

$$D \ni (s, t) \mapsto \Phi(s, t) = \boldsymbol{p}(s, t) = \begin{bmatrix} x(s, t) \\ y(s, t) \\ z(s, t) \end{bmatrix}. \tag{16.2.10}$$

Step 2. Understanding the orientation. The vectors $\boldsymbol{p}'_s$, $\boldsymbol{p}'_t$ are linearly independent and tangent to Σ. Thus $\boldsymbol{p}'_s \times \boldsymbol{p}'_t$ is perpendicular to the tangent space of Σ at $\boldsymbol{p}(s,t)$. In particular, exactly one of the vectors $\boldsymbol{p}'_s \times \boldsymbol{p}'_t$ or $\boldsymbol{p}'_t \times \boldsymbol{p}'_s = -\boldsymbol{p}'_s \times \boldsymbol{p}'_t$ points in the same direction as the normal $\boldsymbol{\nu}\big(\boldsymbol{p}(s,t)\big)$ defining the orientation. Find $\epsilon(s,t) \in \{\pm 1\}$ such that $\epsilon(s,t)(\boldsymbol{p}'_s \times \boldsymbol{p}'_t)$ and $\boldsymbol{\nu}$ point in the same direction, i.e.,

$$\epsilon(s,t) = \begin{cases} 1, & \big\langle \boldsymbol{p}'_s \times \boldsymbol{p}'_t, \boldsymbol{\nu} \big\rangle > 0, \\ -1, & \big\langle \boldsymbol{p}'_s \times \boldsymbol{p}'_t, \boldsymbol{\nu} \big\rangle < 0. \end{cases}$$

Note that

$$\boxed{\epsilon(s,t) = -\epsilon(t,s)}.$$

Step 3. Integrating. We have

$$\boldsymbol{\nu} = \frac{\epsilon(s,t)}{\|\boldsymbol{p}'_s \times \boldsymbol{p}'_t\|}\boldsymbol{p}'_s \times \boldsymbol{p}'_t.$$

On the other hand (see Exercise 16.8)

$$|dA| = \|\boldsymbol{p}'_s \times \boldsymbol{p}'_t\|\,|dsdt|,$$

$$\boxed{\langle \boldsymbol{F}, \boldsymbol{\nu} \rangle\,|dA| = \epsilon(s,t)\langle \boldsymbol{F}, \boldsymbol{p}'_s \times \boldsymbol{p}'_t \rangle\,|dsdt|}.$$

Observe that

$$\langle \boldsymbol{F}, \boldsymbol{p}'_s \times \boldsymbol{p}'_t \rangle = \det \begin{bmatrix} P & x'_s & x'_t \\ Q & y'_s & y'_t \\ R & z'_s & z'_t \end{bmatrix}.$$

Then

$$\boxed{\operatorname{Flux}(\boldsymbol{F}, \Sigma, \boldsymbol{\nu}) = \int_D \epsilon(s,t) \det \begin{bmatrix} P & x'_s & x'_t \\ Q & y'_s & y'_t \\ R & z'_s & z'_t \end{bmatrix} |dsdt|} \tag{16.2.11}$$

or, equivalently

$$\boxed{\operatorname{Flux}(\boldsymbol{F}, \Sigma, \boldsymbol{\nu}) = \int_D \epsilon(t,s) \det \begin{bmatrix} P & x'_t & x'_s \\ Q & y'_t & y'_s \\ R & z'_t & z'_s \end{bmatrix} |dsdt|}. \tag{16.2.12}$$

Expanding along the first column of the determinant in (16.2.11) we deduce

$$\det \begin{bmatrix} P & x'_s & x'_t \\ Q & y'_s & y'_t \\ R & z'_s & z'_t \end{bmatrix} = P \det \begin{bmatrix} y'_s & y'_t \\ z'_s & z'_t \end{bmatrix} + Q \det \begin{bmatrix} z'_s & z'_t \\ x'_s & x'_t \end{bmatrix} + R \det \begin{bmatrix} x'_s & x'_t \\ y'_s & y'_t \end{bmatrix}. \tag{16.2.13}$$

The above steps are best remembered using the language of *differential forms of degree* 2 or 2-*forms*.

To the vector field $\boldsymbol{F} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k}$ we associate the degree 2 differential form

$$\boxed{\Phi_{\boldsymbol{F}} := Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy}. \tag{16.2.14}$$

Above, the symbol "$\wedge$" is called the *exterior product* or the *wedge*. Don't worry about its meaning yet. For now you only need to know that it differs from a usual product in that it is *anti-commutative*, i.e., *for any differential* 1*-forms* ω *and* η,

$$\boxed{\omega\wedge\eta=-\eta\wedge\omega,\quad \omega\wedge\omega=0}$$

The product $\omega\wedge\eta$ is a differential form of degree 2.

In (16.2.13) we think of x, y, z as functions depending on the variables s, t as in (16.2.10). Then dx, dy, dz are the (total) differentials of these functions as defined in (13.2.13). We have

$$\begin{aligned} dy\wedge dz &= (y'_s ds+y'_t dt)\wedge(z'_s ds+z'_t dt)\\ &=\underbrace{y'_s z'_s ds\wedge ds}_{=0}+y'_s z'_t\, ds\wedge dt+y'_t z'_s\,\underbrace{dt\wedge ds}_{=-ds\wedge dt}+\underbrace{y'_t z'_t dt\wedge dt}_{=0}\\ &=\big(y'_s z'_t-z'_s y'_t\big)ds\wedge dt=\det\begin{bmatrix} y'_s & y'_t\\ z'_s & z'_t\end{bmatrix}ds\wedge dt. \end{aligned}$$

We can write this

$$\det\begin{bmatrix} y'_s & y'_t\\ z'_s & z'_t\end{bmatrix}=\frac{dy\wedge dz}{ds\wedge dt}. \tag{16.2.15}$$

Arguing in a similar fashion we deduce from (16.2.13)

$$\det\begin{bmatrix} P & x'_s & x'_t\\ Q & y'_s & y'_t\\ R & z'_s & z'_t\end{bmatrix}=P\frac{dy\wedge dz}{ds\wedge dt}+Q\frac{dz\wedge dx}{ds\wedge dt}+R\frac{dx\wedge dy}{ds\wedge dt}$$

so

$$Pdy\wedge dz+Qdz\wedge dz+Rdz\wedge dy=\det\begin{bmatrix} P & x'_s & x'_t\\ Q & y'_s & y'_t\\ R & z'_s & z'_t\end{bmatrix}ds\wedge dt.$$

Now observe that

$$\big\langle\,\boldsymbol{F},\boldsymbol{\nu}\,\big\rangle|dA|=\epsilon(s,t)\big\langle\,\boldsymbol{F},\boldsymbol{p}'_s,\boldsymbol{p}'_t\,\big\rangle|dsdt|=\det\begin{bmatrix} P & x'_s & x'_t\\ Q & y'_s & y'_t\\ R & z'_s & z'_t\end{bmatrix}\epsilon(s,t)\,|dsdt|.$$

It is now time to give an idea of what $ds\wedge dt$ is. We "define"

$$\boxed{ds\wedge dt:=\epsilon(s,t)\,|dsdt|}. \tag{16.2.16}$$

This is not an entirely satisfying definition because it is not clear what is the nature of $|dsdt|$. Intuitively, it is the area of an "infinitesimal curvilinear parallelogram" on Σ, but the concept of "infinitesimal parallelogram" is a rather nebulous one. This will have to do for a while. Note that (16.2.15) implies

$$dy\wedge dz=\det\begin{bmatrix} y'_s & y'_t\\ z'_s & z'_t\end{bmatrix}ds\wedge dt=\epsilon(s,t)\det\begin{bmatrix} y'_s & y'_t\\ z'_s & z'_t\end{bmatrix}|dsdt|.$$

The issue of the nature of $\wedge$ aside, we deduce that

$$\big\langle\,\boldsymbol{F},\boldsymbol{\nu}\,\big\rangle|dA|=Pdy\wedge dz+Qdz\wedge dx+Rdx\wedge dy,$$

For this reason we set

$$\boxed{\int_{\Sigma,\boldsymbol{\nu}}\Phi_{\boldsymbol{F}} := \operatorname{Flux}(\boldsymbol{F},\Sigma,\boldsymbol{\nu})},$$

where we recall that

$$\Phi_{\boldsymbol{F}} = Pdy\wedge dz + Qdz\wedge dx + Rdx\wedge dy.$$

The integral

$$\int_{\Sigma,\boldsymbol{\nu}} Pdy\wedge dz + Qdz\wedge dx + Rdx\wedge dy$$

is traditionally called a *surface integral of the second kind*. It depends on a choice of orientation specified by the unit normal vector field $\boldsymbol{\nu}$. The differential form $\Phi_{\boldsymbol{F}}$ is sometimes referred to as the *infinitesimal flux of* $\boldsymbol{F}$. □

Example 16.2.24. Let us see how the above strategy works in a special case. Suppose that Σ is unit sphere

$$\Sigma = \big\{\,(x,y,z)\in\mathbb{R}^3;\;\; x^2+y^2+z^2=1\,\big\}.$$

We fix on Σ the orientation defined by the unit normal vector field pointing towards the exterior of the unit ball bounded by this sphere. Let $\boldsymbol{F}$ be the vector field $\boldsymbol{F}=\boldsymbol{i}+\boldsymbol{j}$.

The spherical coordinates provide a quasi-parametrization of this sphere

$$(\theta,\varphi)\mapsto \boldsymbol{p}(\theta,\varphi)=\left[\begin{array}{ccc} x & = & \sin\varphi\cos\theta\\ y & = & \sin\varphi\sin\theta\\ z & = & \cos\varphi, \end{array}\right],\;\; \theta\in(0,2\pi),\;\; \varphi\in(0,\pi).$$

If we keep φ fixed and we let θ vary increasingly, the moving point $\theta\mapsto\boldsymbol{p}(\theta,\varphi)$ runs West-to East along a parallel; see Figure 16.25. The vector $\boldsymbol{p}'_\theta$ is tangent to this parallel and points East. If we keep θ fixed and we let φ vary increasingly, the moving point $\theta\mapsto\boldsymbol{p}(\theta,\varphi)$ runs North-to-South along a meridian; see Figure 16.25. The vector $\boldsymbol{p}'_\varphi$ is tangent to this meridian and points South.

The right-hand-rule for computing cross products (see page 366) shows that $\boldsymbol{p}'_\varphi\times\boldsymbol{p}'_\theta$ points towards the exterior of the sphere, i.e., in the same direction as the normal $\boldsymbol{\nu}$ defining the chosen orientation on the sphere. Thus, in this case

$$\epsilon(\varphi,\theta)=1=-\epsilon(\theta,\varphi).$$

We have

$$\big\langle\,\boldsymbol{F},\boldsymbol{p}'_\varphi\times\boldsymbol{p}'_\theta\,\big\rangle=\det\left[\begin{array}{ccc} P & x'_\varphi & x'_\theta\\ Q & y'_\varphi & y'_\theta\\ R & z'_\varphi & z'_\theta\end{array}\right]=\det\left[\begin{array}{ccc} 1 & \cos\varphi\cos\theta & -\sin\varphi\sin\theta\\ 1 & \cos\varphi\sin\theta & \sin\varphi\cos\theta\\ 0 & -\sin\varphi & 0\end{array}\right]$$

$$=\det\left[\begin{array}{cc}\cos\varphi\sin\theta & \sin\varphi\cos\theta\\ -\sin\varphi & 0\end{array}\right]-\det\left[\begin{array}{cc}\cos\varphi\cos\theta & -\sin\varphi\sin\theta\\ -\sin\varphi & 0\end{array}\right]$$

$$=\sin^2\varphi\big(\,\cos\theta+\sin\theta\,\big).$$

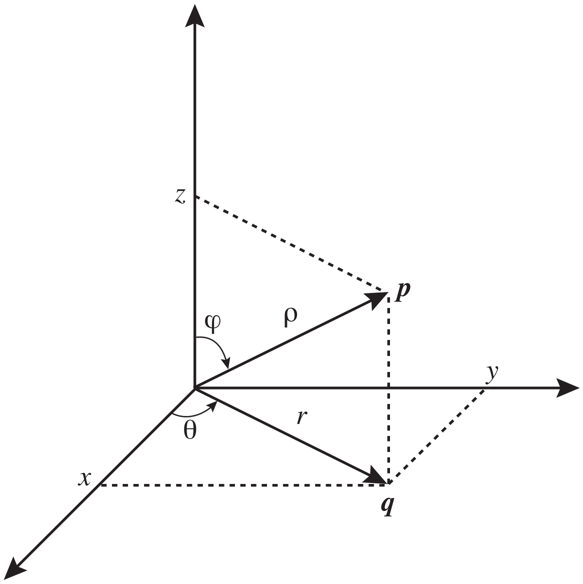

Figure 16.25. *If we vary θ keeping φ fixed the point $\boldsymbol{p}$ runs along a parallel, while if we vary φ keeping theta fixed the point $\boldsymbol{p}$ runs along a meridian.*

We have

$$\int_\Sigma \big\langle\, \boldsymbol{F}, \boldsymbol{\nu} \,\big\rangle \, |dA| = \int_{\substack{0\leqslant\theta\leqslant 2\pi,\\ 0\leqslant\varphi\leqslant\pi}} \sin^2\varphi\big(\,\cos\theta+\sin\theta\,\big)\,|d\varphi d\theta|$$

$$= \left(\int_0^\pi \sin\varphi\, d\varphi\right)\cdot \underbrace{\left(\int_0^{2\pi} (\,\cos\theta+\sin\theta\,)\, d\theta\right)}_{=0} = 0.$$

16.2.6. Stokes' Formulæ. To state these formulæ we need to introduce new concepts. Suppose that $U \subset \mathbb{R}^3$ is an open set and $\boldsymbol{F} : U \to \mathbb{R}^3$ is a C^1 vector field

$$\boldsymbol{F} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k}.$$

The *curl* of $\boldsymbol{F}$ is the continuous vector field $\operatorname{curl} F : U \to \mathbb{R}^3$

$$\boxed{\operatorname{curl}\boldsymbol{F} := \big(\partial_y R - \partial_z Q\big)\boldsymbol{i} + \big(\partial_z P - \partial_x R\big)\boldsymbol{j} + \big(\partial_x Q - \partial_y P\big)\boldsymbol{k}}. \tag{16.2.17}$$

The right-hand side of the above equality looks intimidating, but there are cleverer ways of describing it.

Consider the formal[4] vector field

$$\nabla := \partial_x\boldsymbol{i} + \partial_y\boldsymbol{j} + \partial_z\boldsymbol{k}.$$

We then have

$$\boxed{\operatorname{curl}\boldsymbol{F} = \nabla\times\boldsymbol{F}}, \tag{16.2.18}$$

[4] In mathematics the term *formal* usually refers to objects that exist on paper and whose natures are not important for a particular argument. In this particular case ∇ can be given a precise rigorous meaning.

where "×" denotes the cross product of two vectors in $\mathbb{R}^3$. Recall that it is uniquely determined by the anti-commutativity equalities

$$\boldsymbol{i}\times\boldsymbol{j}=-\boldsymbol{j}\times\boldsymbol{i}=\boldsymbol{k},\;\;\boldsymbol{j}\times\boldsymbol{k}=-\boldsymbol{k}\times\boldsymbol{j}=\boldsymbol{i},\;\;\boldsymbol{k}\times\boldsymbol{i}=-\boldsymbol{i}\times\boldsymbol{k}=\boldsymbol{j},$$

$$\boldsymbol{i}\times\boldsymbol{i}=\boldsymbol{j}\times\boldsymbol{j}=\boldsymbol{k}\times\boldsymbol{k}=\boldsymbol{0}.$$

Let us point out that if we use the classical interpretation of the inner product on $\mathbb{R}^3$ as a "dot product"

$$\boldsymbol{x}\cdot\boldsymbol{y}:=\langle\boldsymbol{x},\boldsymbol{y}\rangle,\;\;\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^3,$$

then the divergence of $\boldsymbol{F}$ can be given the more compact description

$$\boxed{\operatorname{div}\boldsymbol{F}=\nabla\cdot\boldsymbol{F}=\langle\nabla,\boldsymbol{F}\rangle}. \tag{16.2.19}$$

Theorem 16.2.25 (2D Stokes)**.** *Suppose that $\Sigma\subset\mathbb{R}^3$ is a compact C^1-surface with boundary oriented by a unit normal vector field $\boldsymbol{\nu}:\Sigma\to\mathbb{R}^3$. Denote by $\partial_+\Sigma$ the boundary of Σ equipped with the orientation induced by the orientation of Σ as described at page 626. If*

$$\boldsymbol{F}:=P\boldsymbol{i}+Q\boldsymbol{j}+R\boldsymbol{k}$$

is a C^1 vector field defined on an open set $\mathcal{O}$ containing Σ, then

$$\int_{\partial_+\Sigma}W_{\boldsymbol{F}}=\int_\Sigma(\operatorname{curl}\boldsymbol{F})\cdot\boldsymbol{\nu}\,dA=\operatorname{Flux}(\nabla\times\boldsymbol{F},\Sigma,\boldsymbol{\nu})=\int_\Sigma\Phi_{\operatorname{curl}\boldsymbol{F}}. \tag{16.2.20}$$

□

The equality (16.2.20) is often referred to as *Green's formula*

Theorem 16.2.26 (3D Stokes)**.** *Suppose that $U\subset\mathbb{R}^3$ is a bounded C^1-domain. Let $\boldsymbol{\nu}:\partial U\to\mathbb{R}^3$ denote the <u>outer</u> unit normal vector field along ∂U; see Example 16.2.20. If*

$$\boldsymbol{F}:=P\boldsymbol{i}+Q\boldsymbol{j}+R\boldsymbol{k}$$

is a C^1 vector field defined on an open set $\mathcal{O}$ containing $\mathbf{cl}\,U$, then

$$\int_{\partial_+U}\Phi_{\boldsymbol{F}}=\operatorname{Flux}(\boldsymbol{F},\partial U,\boldsymbol{\nu})=\int_U\operatorname{div}\boldsymbol{F}\;|dxdydz|. \tag{16.2.21}$$

□

Often (16.2.21) is referred to as *divergence formula.*

Example 16.2.27. Suppose that $D\subset\mathbb{R}^3$ with a bounded C^1-domain and $\boldsymbol{R}=x\boldsymbol{i}+y\boldsymbol{j}+z\boldsymbol{k}$. Denote by $\boldsymbol{\nu}_{out}$ the outer unit normal vector field. Note that

$$\operatorname{div}\boldsymbol{R}=3.$$

The divergence formula then implies

$$\operatorname{Flux}(\boldsymbol{R},\partial D,\boldsymbol{\nu}_{out})=\int_D 3\,|dxdydz|=3\operatorname{vol}(D).$$

□

Example 16.2.28. Consider the vector field $\boldsymbol{V} : \mathbb{R}^3 \backslash \boldsymbol{0} \to \mathbb{R}^3$

$$\boldsymbol{V}(x,y,z) = \frac{x}{\rho^3}\boldsymbol{i} + \frac{y}{\rho^3}\boldsymbol{j} + \frac{z}{\rho^3}\boldsymbol{k},\;\; \rho = \sqrt{x^2+y^2+z^2}.$$

Let $B_r(0)$ the open ball of radius r centered at $0 \in \mathbb{R}^3$. Its boundary is the sphere $\Sigma_r(0)$ of radius r centered at 0. The outer unit vector field along $\partial B_r(0)$ is

$$\boldsymbol{\nu}_{out}(x,y,z) = \frac{x}{r}\boldsymbol{i} + \frac{y}{r}\boldsymbol{j} + \frac{z}{r}\boldsymbol{k}.$$

Thus, along $\partial B_r(0)$ we have $\rho = r$ and

$$\big\langle\, \boldsymbol{V}(x,y,z), \boldsymbol{\nu}_{out}(x,y,z) \,\big\rangle = \frac{x^2+y^2+z^2}{r^4} = \frac{r^2}{r^4} = \frac{1}{r^2}.$$

Thus

$$\operatorname{Flux}(\boldsymbol{V}, \partial B_r(0), \boldsymbol{\nu}_{out}\rangle = \int_{\Sigma_r} \frac{1}{r^2} dA = \frac{1}{2}\operatorname{area}(\Sigma_r) = 4\pi.$$

Let us compute the divergence of $\boldsymbol{V}$. From the equality $\rho^2 = x^2+y^2+z^2$ we deduce

$$\rho'_x = \frac{x}{\rho},\;\; \rho'_y = \frac{y}{\rho},\;\; \rho'_z = \frac{z}{\rho}$$

$$\partial_x\left(\frac{x}{\rho^3}\right) = \frac{1}{\rho^3} - 3\frac{x^2}{\rho^5},\;\; \partial_y\left(\frac{y}{\rho^3}\right) = \frac{1}{\rho^3} - 3\frac{y^2}{\rho^5},\;\; \partial_z\left(\frac{z}{\rho^3}\right) = \frac{1}{\rho^3} - 3\frac{z^2}{\rho^5}$$

Thus

$$\operatorname{div}\boldsymbol{V} = \frac{3}{\rho^3} - 3\frac{x^2+y^2+z^2}{\rho^5} = 0$$

so that

$$\int_{B_r(0)} \operatorname{div}\boldsymbol{V}\,|dxdydz| = 0.$$

This seems to contradicts the divergence theorem. The problem with this is that the vector field $\boldsymbol{V}$ has a singularity at the origin and it is not defined there. The divergence formula requires that the vector field be defined *everywhere* in the domain.

Suppose now that D is a bounded C^1 domain such that $0 \in D$. We want to compute $\operatorname{Flux}(\boldsymbol{V}, \partial D, \boldsymbol{\nu}_{out})$.

There exists $r_0 > 0$ such that $B_{r_0}(0) \subset D$. For any $0 < \varepsilon < r_0$ we set

$$D_\varepsilon := D \backslash \boldsymbol{cl}\,\big(\, B_\varepsilon(0) \,\big).$$

The vector field is well defined everywhere on D_ε and the divergence formula implies

$$\operatorname{Flux}(\boldsymbol{V}, \partial D_\varepsilon, \boldsymbol{\nu}_{out}) = \int_{D_\varepsilon} \operatorname{div}\boldsymbol{V}\,|dxdydz| = 0.$$

Now observe that

$$\partial D_\varepsilon = \partial D \cup \partial B_\varepsilon(0).$$

The normal vector field along $\partial B_\varepsilon(0)$ that points towards the exterior of D_ε is the normal vector field that points towards the *interior* of $B_\varepsilon(0)$. We denote it with $\boldsymbol{\nu}_{in}(B_\varepsilon)$. Thus

$$0 = \operatorname{Flux}(\boldsymbol{V}, \partial D_\varepsilon, \boldsymbol{\nu}_{out}) = \operatorname{Flux}(\boldsymbol{V}, \partial D, \boldsymbol{\nu}_{out}) + \operatorname{Flux}(\boldsymbol{V}, \partial B_\varepsilon, \boldsymbol{\nu}_{in}(B_\varepsilon))$$

$$= \operatorname{Flux}(\boldsymbol{V}, \partial D, \boldsymbol{\nu}_{out}) - \operatorname{Flux}(\boldsymbol{V}, \partial B_\varepsilon, \boldsymbol{\nu}_{out}(B_\varepsilon)) = \operatorname{Flux}(\boldsymbol{V}, \partial D, \boldsymbol{\nu}_{out}) - 4\pi$$

so that

$$\operatorname{Flux}(\boldsymbol{V}, \partial D, \boldsymbol{\nu}_{out}) = 4\pi. \qquad \square$$

Remark 16.2.29. Suppose that $D \subset \mathbb{R}^2$ is a bounded C^1-domain, and

$$\boldsymbol{F} : U \to \mathbb{R}^2, \quad \boldsymbol{F} = P(x,y)\boldsymbol{i} + Q(x,y)\boldsymbol{j}$$

is a C^1-vector field defined on an open set $U \subset \mathbb{R}^2$ that contains $\boldsymbol{cl}(D)$. Then D can be viewed as a surface with boundary in $\mathbb{R}^3$. If we write

$$\boldsymbol{F} = P(x,y)\boldsymbol{i} + Q(x,y)\boldsymbol{j} + 0\boldsymbol{k}$$

we see that we can view $\boldsymbol{F}$ as a 3-dimensional C^1-vector field defined on the open set $\mathcal{O} = U \times \mathbb{R} \subset \mathbb{R}^3$. The constant vector field $\boldsymbol{k}$ defines an orientation on D, and the induced orientation on ∂D defined by this unit normal vector field coincides with the orientation of ∂D as boundary of a planar domain. Note that

$$\operatorname{curl} \boldsymbol{F} = \big(Q'_x - P'_y\big)\boldsymbol{k}$$

so we deduce from (16.2.20)

$$\int_{\partial_+ D} \boldsymbol{F} \cdot \boldsymbol{p} = \operatorname{Flux}(\operatorname{curl} \boldsymbol{F}, D, \boldsymbol{k}) = \int_D \big(Q'_x - P'_y\big)\, |dxdy|.$$

This shows that (16.1.10a) is a special case of (16.2.20). $\square$

16.3. Differential forms and their calculus

16.3.1. Differential forms on Euclidean spaces. So far we have encountered differential forms of degree 1

$$Pdx + Qdy + Rdz,$$

differential forms of degree 2,

$$Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy,$$

where we recall that the operation "$\wedge$" satisfies the unusual anti-commutativity conditions

$$dx \wedge dy = -dy \wedge dx, \quad dy \wedge dz = -dz \wedge dy, \quad dz \wedge dx = -dx \wedge dz,$$

$$dx \wedge dx = dy \wedge dy = dz \wedge dz = 0.$$

A *differential form of degree* 3 or 3-*form* on an open set $\mathcal{O} \in \mathbb{R}^3$ is an expression of the form

$$\rho\, dx \wedge dy \wedge dz,$$

where $\rho : \mathcal{O} \to \mathbb{R}$ is a continuous function. Again, we avoid explaining the meaning of the quantity $dx \wedge dy \wedge dz$, but we want to point out a few oddities.

$$dx \wedge dy \wedge dz = dx \wedge \big(\, dy \wedge dz\,\big) = dx \wedge \big(\, -dz \wedge dy\,\big)$$

$$= -\big(\, dx \wedge dz\,\big) \wedge dy = \big(\, dz \wedge dx\,\big) \wedge dy = dz \wedge dx \wedge dy$$

This is a bit surprising! The anti-commutative operation "$\wedge$" becomes commutative when 2-forms are involved,

$$dx \wedge dy \wedge dz = dz \wedge dx \wedge dy.$$

We still have not explained the meaning of the quantities $dx \wedge dy$, $dx \wedge dy \wedge dz$ etc. and we will not do so for a while.

Definition 16.3.1. Given an open set $\mathcal{O} \subset \mathbb{R}^n$ and $k = 1, \dots, n$ we denote by $\Omega^k(\mathcal{O})$ the space of *differential forms of degree* k on $\mathcal{O}$, i.e., expressions of the form

$$\omega = \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \omega_{i_1,\dots,i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k},$$

where the coefficients $\omega_{i_1,\dots,i_k}$ are continuous functions on $\mathcal{O}$. We set

$$\Omega^0(\mathcal{O}) := C^0(\mathcal{O}).$$

A differential form is called C^m if its coefficients are C^m-functions. We denote by $\Omega^k(\mathcal{O})_{C^m}$ the space of C^m-forms of degree k. □

To simplify the exposition we introduce the following conventions.

- We set
$$[m] := \{1, \dots, m\}.$$
- We denote by $[n]^{[m]}$ the set of maps $[m] \to [n]$ and by $\mathrm{Inj}(m,n)$ the set of *injections* $[m] \to [n]$, $k \mapsto i_k$. We describe such an injection I using the notation $I = (i_1, \dots, i_m)$. We denote by $\{I\}$ the range of I, $\{I\} := \{i_1, \dots, i_m\}$. We will refer to such injections as multi-indices.
- We denote by $\mathfrak{S}_n$ the group of permutations of n objects, $\mathfrak{S}_n = \mathrm{Inj}(n,n)$.
- We denote by $\mathrm{Inj}^+(m,n)$ the subset of increasing multi-indices $I : [m] \to [n]$.
- if $\sigma \in \mathfrak{S}_m$ i and $I = (i_1, \dots, i_m) \in \mathrm{Inj}_(m,n)$ we set
$$I_\sigma = (i_{\sigma(1)}, \dots, i_{\sigma(m)}) \in \mathrm{Inj}(m,n)$$
- For $I \in \mathrm{Inj}(m,n)$ we set
$$d\boldsymbol{x}^{\wedge I} := dx^{i_1} \wedge \cdots \wedge dx^{i_m}.$$

  The terms $d\boldsymbol{x}^{\wedge I}$, $I \in \mathrm{Inj}(m,n)$ are called *exterior monomials*

Thus, any $\omega \in \Omega^k(\mathcal{O})$ has the form

$$\omega = \sum_{I \in \mathrm{Inj}^+(k,n)} \omega_I d\boldsymbol{x}^{\wedge I}, \;\; \omega_I \in C^0(\mathcal{O}). \tag{16.3.1}$$

The space $\Omega^k(\mathcal{O})$ is a vector space. The addition is defined by

$$\left(\sum_I \alpha_I d\boldsymbol{x}^{\wedge I}\right) + \left(\sum_I \beta_I d\boldsymbol{x}^{\wedge I}\right) = \sum_I (\alpha_I + \beta_I) d\boldsymbol{x}^{\wedge I},$$

and the scalar multiplication is defined in a similar fashion. The elementary monomials dx^i satisfy the anti-commutativity rules

$$dx^i \wedge dx^j = \begin{cases} -dx^j \wedge dx^i, & i \neq j, \\ 0, & i = j. \end{cases} \tag{16.3.2}$$

This implies that if $I \in \operatorname{Inj}(m,n)$ and $\sigma \in \mathfrak{S}_m$, then

$$d\boldsymbol{x}^{\wedge I_\sigma} = \epsilon(\sigma) d\boldsymbol{x}^{\wedge I}, \tag{16.3.3}$$

where $\epsilon(\sigma) \in \{-1, 1\}$ is the signature or parity of the permutation σ.

We can define $d\boldsymbol{x}^{\wedge I}$ in the obvious way for any $I \in [n]^{[m]}$ with the understanding that $d\boldsymbol{x}^{\wedge I} = 0$ if I is not an injection. For example if $I = (2,3,2)$ then

$$d\boldsymbol{x}^{\wedge I} = dx^2 \wedge dx^3 \wedge dx^2 = -dx^2 \wedge dx^2 \wedge dx^3 = 0.$$

For $I \in [n]^{[k]}$ and $J \in [n]^{[\ell]}$ we define

$$I * J := (i_1, \dots, i_k, j_1, \dots, j_\ell). \in [n]^{[k+\ell]}.$$

Then

$$d\boldsymbol{x}^{\wedge I} \wedge d\boldsymbol{x}^{\wedge J} = d\boldsymbol{x}^{\wedge I*J}.$$

This allows us to define a product

$$\wedge : \Omega^k(\mathcal{O}) \times \Omega^\ell(\mathcal{O}) \to \Omega^{k+\ell}(\mathcal{O}), \;\; k + \ell \leqslant n,$$

$$\left(\sum_{I \in \operatorname{Inj}^+(k,n)} \alpha_I d\boldsymbol{x}^{\wedge I} \right) \wedge \left(\sum_{J \in \operatorname{Inj}^+(\ell,n)} \beta_J d\boldsymbol{x}^{\wedge J} \right)$$

$$:= \sum_{\substack{I \in \operatorname{Inj}^+(k,n), \\ J \in \operatorname{Inj}^+(\ell,n)}} \alpha_I \beta_J d\boldsymbol{x}^{\wedge I} \wedge d\boldsymbol{x}^{\wedge J} = \sum_{\substack{I \in \operatorname{Inj}^+(k,n), \\ J \in \operatorname{Inj}^+(\ell,n)}} \alpha_I \beta_J d\boldsymbol{x}^{\wedge I*J}.$$

As we know, the 1-forms can be integrated over *oriented* curves, and the 2-forms can be integrated over *oriented* surfaces. A 3-form $\rho dx \wedge dy \wedge dz \in \Omega(\mathcal{O})$ can also be integrated and we set

$$\int_{\mathcal{O}} \rho \, dx \wedge dy \wedge dz := \int_{\mathcal{O}} \rho \, |dxdydz|,$$

whenever the integral on the right-hand side is absolutely convergent.

Let us point out a curious but important fact: since $dx \wedge dy \wedge dz = -dx \wedge dz \wedge dy$, we have

$$\int_{\mathcal{O}} \rho \, dx \wedge dz \wedge dy = -\int_{\mathcal{O}} \rho \, dx \wedge dy \wedge dz.$$

There is also a notion of derivative of a differential form called *exterior derivative.*

Definition 16.3.2 (Exterior derivative)**.** Let $\mathcal{O} \subset \mathbb{R}^n$ be an open subset. The *exterior derivative* is the linear operator

$$d : \Omega^k(\mathcal{O})_{C^1} \to \Omega^{k+1}(\mathcal{O}), \;\; k = 0, 1, \dots, n-1,$$

defined as follows.

- If $k = 0$ so that $\Omega^0(\mathcal{O})_{C^1} = C^1(\mathcal{O})$, then

$$df = \sum_{i=1}^{n} \partial_{x^i} f dx^i \in \Omega^1(\mathcal{O}), \ \ \forall f \in C^1(\mathcal{O}).$$

- If $k > 0$, then for any $\alpha \in \Omega^k(\mathcal{O})_{C^1}$ we set

$$d\alpha := d\left(\sum_{I \in \mathrm{Inj}^+(k,n)} \alpha_I d\boldsymbol{x}^{\wedge I} \right) = \sum_{I \in \mathrm{Inj}^+(k,n)} d\alpha_I \wedge d\boldsymbol{x}^{\wedge I}.$$

□

Example 16.3.3. The exterior derivative of a 0-form is a 1-form, the exterior derivative of a 1-form is 2-form, the exterior derivative of a 2-form is a 3-form, and the exterior derivative of a 3-form is identically zero. Here is how one computes these exterior derivatives.

The differential of a C^1 form of degree zero, i.e., a C^1-function $f : \mathcal{O} \to \mathbb{R}$ is its total differential

$$df = f'_x dx + f'_y dy + f'_z dz = W_{\nabla f}.$$

If ω is a C^1 differential form of degree 1 on $\mathcal{O}$,

$$\omega = P dx + Q dy + R dz = W_{\boldsymbol{F}}, \ \ \boldsymbol{F} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k},$$

then

$$\begin{aligned} d\omega = dW_{\boldsymbol{F}} &= dP \wedge dx + dQ \wedge dy + dR \wedge dz \\ &= (P'_x dz + P'_y dy + P'_z dz) \wedge dx \\ &+ (Q'_x dx + Q'_y dy + Q'_z dz) \wedge dy \\ &+ (R'_x dx + R'_y dy + R'_z dz) \wedge dz \end{aligned}$$

$$= P'_y \, dy \wedge dx + P'_z \, dz \wedge dx + Q'_x \, dx \wedge dy + Q'_z \, dz \wedge dy + R'_x \, dx \wedge dz + R'_y \, dy \wedge dz$$

$$= \big(R'_y - Q'_z \big) \, dy \wedge dz + \big(P'_z - R'_z \big) \, dz \wedge dx + \big(Q'_x - P'_y \big) \, dx \wedge dy$$

(use (16.2.14) and (16.2.17))

$$= \Phi_{\mathrm{curl}\, \boldsymbol{F}}.$$

Thus

$$\boxed{dW_{\boldsymbol{F}} = \Phi_{\mathrm{curl}\, \boldsymbol{F}}}. \tag{16.3.4}$$

Suppose finally that η is a C^1 form of degree 2 on $\mathcal{O}$

$$\eta = \Phi_{\boldsymbol{F}} = P dy \wedge dz + Q \, dz \wedge dx + R \, dx \wedge dy.$$

Then

$$\begin{aligned} d\eta &= dP \wedge dy \wedge dz + dQ \wedge dz \wedge dx + dR \wedge dx \wedge dy \\ &= \big(P'_x dx + P'_y dy + P'_z dz \big) \, dy \wedge dz \\ &+ \big(Q'_x dx + Q'_y dy + Q'_z dz \big) dz \wedge dz \\ &+ \big(R'_x dx + R'_y dy + R'_z dz \big) \, dx \wedge dy \end{aligned}$$

$$= P'_x \, dx \wedge dy \wedge dz + Q'_y \, dy \wedge dz \wedge dx + R'_z \, dz \wedge dx \wedge dy$$

$$= \big(P'_x + Q'_y + R'_z \big)\, dx \wedge dy \wedge dz = \big(\operatorname{div} \boldsymbol{F} \big)\, dx \wedge dy \wedge dz.$$

Thus

$$\boxed{d\Phi_{\boldsymbol{F}} = \big(\operatorname{div} \boldsymbol{F} \big)\, dx \wedge dy \wedge dz}. \tag{16.3.5}$$

□

In view of the computations in Example 16.3.3 we can give the following equivalent reformulations of Theorem 16.2.25 and Theorem 16.2.26.

Theorem 16.3.4. *Suppose that $\boldsymbol{F}$ is a C^1-vector field defined on the open set $\mathcal{O} \subset \mathbb{R}^3$.*

(i) *If $\Sigma \subset \mathcal{O}$ is an* oriented *compact surface with boundary, then*

$$\int_{\partial_+\Sigma} W_{\boldsymbol{F}} = \int_{\Sigma} dW_{\boldsymbol{F}}.$$

(ii) *If $U \subset \mathbb{R}^3$ is a bounded C^1 domain such that $\mathbf{cl}\, U \subset \mathcal{O}$, then*

$$\int_{\partial_+ U} \Phi_{\boldsymbol{F}} = \int_{U} d\Phi_{\boldsymbol{F}}.$$

□

The above result is not a low dimensional accident. In the remainder of this subsection we hint on how this works in higher dimensions. The key to this process is a more subtle operation on differential forms.

Suppose that $\mathcal{O}_0 \subset \mathbb{R}^{n_0}$ and $\mathcal{O}_1 \subset \mathbb{R}^{n_1}$ are open sets. We denote by $\boldsymbol{x} = (x^i)$ the Cartesian coordinates in $\mathbb{R}^{n_0}$ and by $\boldsymbol{y} = (y^j)$ the Cartesian coordinates in $\mathbb{R}^{n_1}$. Let

$$\Phi : \mathcal{O}_0 \to \mathcal{O}_1$$

be a C^1, map described in the above coordinates by the functions

$$y^j = \Phi^j\big(x^1, \dots, x^{n_0}\big), \quad j = 1, \dots, n_1.$$

For each $k \leqslant \min(n_0, n_1)$, the *pullback* via Φ of a k-form on $\mathcal{O}_1$ (to a k-form on $\mathcal{O}_0$) is the linear operator

$$\Phi^* : \Omega^k(\mathcal{O}_1) \to \Omega^k(\mathcal{O}_0),$$

$$\Phi^*\eta = \Phi^*\left(\sum_{I \in \mathrm{Inj}^+(k, n_1)} \eta_I d\boldsymbol{y}^{\wedge I} \right).$$

$$:= \sum_{I \in \mathrm{Inj}^+(k, n_1)} \eta_I\big(\Phi(\boldsymbol{x})\big) d\Phi^{i_1}(\boldsymbol{x}) \wedge \cdots \wedge d\Phi^{i_k}(\boldsymbol{x}), \quad \forall \eta \in \Omega^k(\mathcal{O}_1).$$

Example 16.3.5. Consider the map $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$, $(r, \theta) \mapsto (x, y) = (r\cos\theta, r\sin\theta)$. Then

$$\Phi^*(dx \wedge dy) = d(r\cos\theta) \wedge d(r\sin\theta)$$

$$= (\cos\theta dr - r\sin\theta d\theta) \wedge (\sin\theta dr + r\cos\theta d\theta)$$

$$= r\cos^2\theta dr \wedge d\theta - r\sin^2\theta \underbrace{d\theta \wedge dr}_{=-dr\wedge d\theta} = (r\cos^2\theta + r\sin^2\theta) dr \wedge d\theta$$
$$= rdr \wedge d\theta = \det J_\Phi dr \wedge d\theta.$$

More generally, given open sets $U, V \subset \mathbb{R}^n$ and $\Phi : U \to V$ a C^1-map, we have

$$\Phi^*\big(dv^1 \wedge \cdots \wedge dv^n\big) = (\det J_\Phi) du^1 \wedge \ldots \wedge du^n, \tag{16.3.6}$$

where (v^i) are the Cartesian coordinates on V and (u^j) are the Cartesian coordinates on U.

(b) Consider the map

$$\Phi : (0,\infty) \times \mathbb{R} \to \mathbb{R}^2\backslash\{\mathbf{0}\},\ \ (r,\theta) \mapsto (x,y) = (r\cos\theta, r\sin\theta).$$

Let

$$\omega = -\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy.$$

Then

$$\Phi^*\omega = \frac{-r\sin\theta\, d(r\cos\theta) + r\cos\theta\, d(r\sin\theta)}{r^2} = d\theta.$$

□

Example 16.3.6. Suppose that $U \subset \mathbb{R}^n$ is an open set that intersects nontrivially the subspace

$$\mathbb{R}^m \times \mathbf{0} = \big\{ (u^1,\ldots,u^n) \in \mathbb{R}^n :\ \ u^i = 0,\ \ \forall i > m \big\}.$$

We set $\overline{U} = U \cap \mathbb{R}^m \times \mathbf{0}$ and we denote by i the inclusion map

$$\overline{U} \ni \bar{\boldsymbol{u}} \mapsto (\bar{\boldsymbol{u}}, \mathbf{0}) \in U.$$

For any $k \leqslant m$ and any $\alpha \in \Omega^k(U)$ we set

$$\alpha|_{\overline{U}} := i^*\alpha. \tag{16.3.7}$$

For example if $k = m$ and

$$\alpha = \sum_{I \in \mathrm{Inj}(m,n)} \alpha_I(u) d\boldsymbol{u}^{\wedge I},$$

then

$$\alpha|_{\overline{U}} = \alpha_{1,2,\ldots,m}(\bar{\boldsymbol{u}}, \mathbf{0}) du^1 \wedge \cdots \wedge u^m \in \Omega^1(\overline{U}).$$

□

The top degree forms on $\mathbb{R}^n$ can be integrated. Let $U \subset \mathbb{R}^n$ be an open set. Denote by $\Omega^k_{\mathrm{cpt}}(U)$ the space of degree k-forms on U with compact support, i.e., forms η such that there exists a compact set $K \subset U$ such that all the coefficients η_I are zero outside K.

Observe first that a degree n form ω defined on open set U in $\mathbb{R}^n$ has the form

$$\omega = \rho_\omega dV_n := \rho_\omega du^1 \wedge \cdots \wedge du^n$$

where ρ_ω is a continuous function on U called the *density* of ω, and $u^1,\ldots,u^n$ are the canonical Cartesian coordinates on $\mathbb{R}^n$. The top degree form dV_n is called the *canonical volume form* on $\mathbb{R}^n$.

Example 16.3.7. Suppose that U, V are open subset of $\mathbb{R}^n$ and $\Phi : U \to V$ is a C^1. Then the density of the top degree form $\omega = \Phi^* dV_n \in \Omega^n(U)$ is

$$\rho_\omega(\boldsymbol{u}) = \det J_\Phi(\boldsymbol{u}).$$

□

Definition 16.3.8. Let $U \subset \mathbb{R}^n$ be an open set. An *orientation* on U is a choice of a nowhere vanishing form η of maximum degree n. Two orientations defined by the top degree forms η_0 or η_1 are two be considered *equivalent* if there exists a continuous function $\rho : U \to (0, \infty)$ such that $\eta_1 = \rho\eta_0$. □

Remark 16.3.9. Observe that if U is an open subset if $\mathbb{R}^n$, then any nowhere vanishing degree n form ω can be described explicitly as a product

$$\omega = \rho_\omega dV_n,$$

where density ρ_ω is a nowhere vanishing continuous function on U. We have a well defined continuous function

$$\epsilon = \epsilon_\omega : U \to \{-1, 1\}, \;\; \epsilon_\omega(\boldsymbol{u}) = \operatorname{sign} \rho_\omega(\boldsymbol{u}) = \frac{\rho_\omega(\boldsymbol{u})}{|\rho_\omega(\boldsymbol{u})|}.$$

The orientation defined by ϵ is therefore equivalent with the orientation defined by $\epsilon_\omega dV_n$. If U is a *path connected* open subset of $\mathbb{R}^n$ then there are precisely two nonequivalent orientations, one defined by the form

$$dV_n := dx^1 \wedge \cdots \wedge dx^n,$$

called the *canonical orientation*, and one defined by $-dV_n$.

If U has k connected components, then there 2^k nonequivalent choices of orientation, each determined by a continuous function[5]

$$\epsilon : U \to \{-1, 1\}.$$

For this reason we will identify the set of possible orientations on U with the set $\mathbb{O}(U)$ of continuous functions $\epsilon : U \to \{-1, 1\}$. The orientation defined my ϵ is, by definition, the orientation defined by the nowhere vanishing top degree form $\epsilon(\boldsymbol{u})dV_n$. □

Definition 16.3.10. An *oriented* open subset of $\mathbb{R}^n$ is a pair (U, ϵ), where U is an open set and $\epsilon \in \mathbb{O}(U)$ is an orientation on U. □

Any oriented open set (U, ϵ) in $\mathbb{R}^n$ defines a linear map

$$\int_{U,\epsilon} : \Omega^n_{\mathrm{cpt}}(U) \to \mathbb{R},$$

$$\int_{U,\epsilon} \omega = \int_{U,\epsilon} \rho_\omega du^1 \wedge \cdots \wedge du^n := \int_U \epsilon(\boldsymbol{u})\rho_\omega(\boldsymbol{u})\, |du^1 \cdots du^n|. \tag{16.3.8}$$

When ϵ is identically equal to 1 we omit it from the notion.

[5] Such a function is constant on the connected components of U.

Let us point out a simple but confusing fact. For any $\sigma \in \mathfrak{S}_n$ we have

$$\epsilon(\sigma)\rho_\omega du^{\sigma(1)} \wedge \cdots \wedge du^{\sigma(n)} = \omega,$$

$$\int_U \rho_\omega du^{\sigma(1)} \wedge \cdots \wedge du^{\sigma(n)} = \epsilon(\sigma) \int_U \rho_\omega du^1 \wedge \cdots \wedge du^n.$$

Because of this it is important to keep in mind the following naive but important advice.

> ☛ ***When integrating differential forms, the order in which we write the coordinates matters!***

Definition 16.3.11. Let (U_i, ϵ_i), $i = 0, 1$ be oriented open sets in $\mathbb{R}^n$, $i = 0, 1$ and $\Phi : U_0 \to U_1$ a C^1-diffeomorphism. We say that Φ is *orientation preserving* if

$$\epsilon_0(\boldsymbol{u})\epsilon_1\big(\Phi(\boldsymbol{u})\big) \det J_\Phi(\boldsymbol{u}) > 0, \;\; \forall \boldsymbol{u} \in U_0. \qquad \square$$

The proof of the next result is left to you as a simple but very instructive Exercise 16.19.

Proposition 16.3.12. *Suppose that (U_i, ϵ_i), $i = 0, 1$ are oriented open sets in $\mathbb{R}^n$, $i = 0, 1$ and $\Phi : U_0 \to U_1$ a C^1-map.*

(i) *For any $k = 0, 1, \dots, n-1$ and any $\alpha \in \Omega^k(U)_{C^1}$ we have*

$$d\big(\Phi^*\alpha\big) = \Phi^*\big(d\alpha\big).$$

(ii) *If $\Phi : (U_0, \epsilon_0) \to (U_1, \epsilon_1)$ is an* orientation preserving *diffeomorphism such that $U_1 = \Phi(U_0)$, then for any $\eta \in \Omega^n_{\mathrm{cpt}}(V)$ we have*

$$\int_{U_1, \epsilon_1} \eta = \int_{U_0, \epsilon_0} \Phi^*\eta. \tag{16.3.9}$$

$\square$

We close this subsection with a technical result which will play a key role in extending to higher dimensions the concept of integration of a differential form.

Proposition 16.3.13. *Suppose that $V_i \subset \mathbb{R}^n$, $i = 0, 1$, are open sets such that*

$$\overline{V}_i := V_i \cap \mathbb{R}^m \times \boldsymbol{0} \neq \varnothing, \;\; i = 0, 1.$$

Let $\Phi : V_0 \to \mathbb{R}^n$ be a C^1-diffeomorphism such that (see Figure 16.26)

$$\Phi(V_0) = V_1, \;\; \Phi(\overline{V}_0) = \overline{V}_1.$$

Then the following hold.

(i) *The induced map $\overline{\Phi} : \overline{V}_0 \to \overline{V}_1 \subset \mathbb{R}^m$ is a C^1-diffeomorphism*

(ii) *If $\omega_1 \in \Omega^m(V_1)$ and $\omega_0 := \Phi^*\omega_1$, then (see (16.3.7))*

$$\omega_0\big|_{\overline{V}_0} = \overline{\Phi}^*\omega_1\big|_{\overline{V}_1}.$$

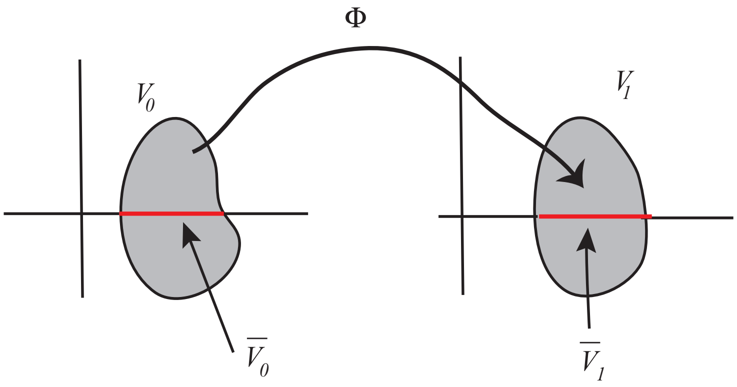

Figure 16.26. *A transition map.*

Proof. Denote by $\boldsymbol{y} = (y^1, \dots, y^n)$ the Cartesian coordinates on V_1 and by $\boldsymbol{x} = (x^1, \dots, x^n)$ the Cartesian coordinates on V_0. We set

$$\bar{\boldsymbol{x}} = (x^1, \dots, x^m),\;\; \bar{\boldsymbol{y}} = (y^1, \dots, y^m),$$
$$\boldsymbol{x}_\perp = (x^{m+1}, \dots, x^n),\;\; \boldsymbol{y}_\perp = (y^{m+1}, \dots, y^n).$$

For simplicity we set

$$\bar{\omega}_i := \omega_i|_{\overline{V}_i},\;\; i = 0, 1.$$

The diffeomorphism Φ is described by a collection of functions

$$y^i = y^i(\boldsymbol{x}),\;\; i = 1, \dots, n.$$

Since $\Phi(\overline{V}_0) = \overline{V}_1$ we deduce

$$y^j(\bar{\boldsymbol{x}}, \mathbf{0}) = 0,\;\; \forall \bar{\boldsymbol{x}} \in \overline{V}_0.,\;\; j > m.$$

We write this

$$\frac{\partial \boldsymbol{y}_\perp}{\partial \bar{\boldsymbol{x}}}(\bar{\boldsymbol{x}}, \mathbf{0}) = 0 \tag{16.3.10}$$

(i) The map $\overline{\Phi}$ is described by the functions

$$y^j = y^j(\bar{\boldsymbol{x}}, \mathbf{0}),\;\; j = 1, \dots, k.$$

We write this succinctly

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{y}}(\bar{\boldsymbol{x}}, \mathbf{0}).$$

The map $\overline{\Phi} : \overline{V}_0 \to \overline{V}_1$ is a homeomorphism since $\Phi : V_0 \to V_1$ is such. We have to show that if $(\bar{\boldsymbol{x}}, \mathbf{0}) \in \overline{V}_0$, then

$$\det J_{\overline{\Phi}}(\bar{\boldsymbol{x}}, \mathbf{0}) \neq 0$$

The Jacobian $J_{\overline{\Phi}}(\bar{\boldsymbol{x}}, \mathbf{0})$ is given by the $m \times m$ matrix

$$\frac{\partial \bar{\boldsymbol{y}}}{\partial \bar{\boldsymbol{x}}}(\bar{\boldsymbol{x}}, \mathbf{0}).$$

We know that $\det J_\Phi(\bar{\boldsymbol{x}}, 0) \neq 0$ since Φ is a diffeomorphism. Now observe that $J_\Phi(\bar{\boldsymbol{x}}, 0)$ has the block decomposition

$$J_\Phi(\bar{\boldsymbol{x}}, 0) = \begin{bmatrix} \partial \bar{\boldsymbol{y}}/\partial \bar{\boldsymbol{x}} & \partial \boldsymbol{y}_\perp/\partial \bar{\boldsymbol{x}} \\ \partial \boldsymbol{y}_\perp/\partial \bar{\boldsymbol{x}} & \partial \boldsymbol{y}_\perp/\partial \boldsymbol{x}_\perp \end{bmatrix}_{(\bar{\boldsymbol{x}}, \mathbf{0})} \overset{(16.3.10)}{=} \begin{bmatrix} \partial \bar{\boldsymbol{y}}/\partial \bar{\boldsymbol{x}} & 0 \\ \partial \boldsymbol{y}_\perp/\partial \bar{\boldsymbol{x}} & \partial \boldsymbol{y}_\perp/\partial \boldsymbol{x}_\perp \end{bmatrix}_{(\bar{\boldsymbol{x}}, \mathbf{0})}.$$

Hence

$$0 \neq \det J_\Phi(\bar{\boldsymbol{x}}, 0) = \det \frac{\partial \bar{\boldsymbol{y}}}{\partial \bar{\boldsymbol{x}}} \cdot \det \frac{\partial \boldsymbol{y}_\perp}{\partial \boldsymbol{x}_\perp} \Rightarrow \det J_{\overline{\Phi}}(\bar{\boldsymbol{x}}, \mathbf{0}) = \det \frac{\partial \bar{\boldsymbol{y}}}{\partial \bar{\boldsymbol{x}}} \neq 0.$$

(ii). Let

$$\omega_1 = \sum_{I\in \mathrm{Inj}^+(m,n)} \omega_I(\boldsymbol{y})d\boldsymbol{y}^{\wedge I}.$$

Then

$$\omega_0 = \sum_{I\in \mathrm{Inj}^+(m,n)} \omega_I\big(\,\boldsymbol{y}(\boldsymbol{x})\,\big)dy^{i_1}(\boldsymbol{x})\wedge\cdots\wedge dy^{i_m}(\boldsymbol{x}),$$

$$\bar{\omega}_1 = \omega_{1,\dots,m}(\bar{\boldsymbol{y}},\boldsymbol{0})dy^1\wedge\cdots\wedge dy^m,$$

$$\bar{\omega}_0 = \sum_{I\in \mathrm{Inj}^+(m,n)} \omega_I\big(\,\boldsymbol{y}(\bar{\boldsymbol{x}},\boldsymbol{0})\,\big)dy^{i_1}(\bar{\boldsymbol{x}},\boldsymbol{0})\wedge\cdots\wedge dy^{i_m}(\bar{\boldsymbol{x}},\boldsymbol{0}).$$

From (16.3.10) we deduce that for $j > m$ we have

$$dy^j(\bar{\boldsymbol{x}},\boldsymbol{0}) = \sum_{i=1}^m \frac{\partial y^j}{\partial x^i}(\bar{\boldsymbol{x}},\boldsymbol{0})dx^i = 0.$$

Hence

$$\bar{\omega}_0 = \omega_{1,2,\dots,m}\big(\,\boldsymbol{y}(\bar{\boldsymbol{x}},\boldsymbol{0})\,\big)dy^1(\bar{\boldsymbol{x}},\boldsymbol{0})\wedge\cdots\wedge dy^m(\bar{\boldsymbol{x}},\boldsymbol{0}) = \bar{\Phi}^*\bar{\omega}_1.$$

□

16.3.2. Orientable submanifolds. . Suppose that X is an m-dimensional C^1-submanifold of $\mathbb{R}^n$, $0 < m < n$. Every point $\boldsymbol{p}\in X$ admits (at least) a *straightening diffeomorphism* $(\mathcal{U},\Psi)$. For brevity we will use the acronym *s.d.* when referring to straightening diffeomorphisms. We recall what this entails (see Definition 14.5.1)

- $\mathcal{U}$ is an open neighborhood of $\boldsymbol{p}\in\mathbb{R}^n$.
- $\Psi:\mathcal{U}\to\mathbb{R}^n$ is a C^1-diffeomorphism with image $U=\Psi(\mathcal{U})$.
- $\Psi\big(\mathcal{U}\cap X\big) = \bar{U} := U\cap\mathbb{R}^m\times\boldsymbol{0}$.
- We denote by $\bar{\Phi}$ the induced map $\Psi^{-1}:\bar{U}\to\mathcal{U}$

An *orientation* for the s.d. $(\mathcal{U},\Psi)$ is a choice of orientation $\epsilon\in\mathbb{O}(\bar{U})$. An *oriented* s.d. is a triplet $(\mathcal{U},\Psi,\epsilon)$, where $(\mathcal{U},\Psi)$ is a s.d. and ϵ is an orientation of that s.d..

If $(\mathcal{U}_0,\Psi_0)$ and $(\mathcal{U}_1,\Psi_1)$ are two straightening diffeomorphism near $\boldsymbol{p}\in X$, we set $\mathcal{V} := \mathcal{U}_0\cap\mathcal{U}_1$, and we get open sets (see Figure 16.27)

$$U_i = \Psi_i(\mathcal{U}_i)\subset\mathbb{R}^n,\;\; V_i = \Psi_i(\mathcal{V})\subset U_i,\;\; i=0,1,$$

$$\bar{U}_i := U_i\cap\mathbb{R}^m\times\boldsymbol{0}\subset\mathbb{R}^m,\;\; \bar{V}_i = V_i\cap\mathbb{R}^m\subset\bar{U}_i,\;\; i=0,1,$$

and C^1-maps

$$\bar{\Phi}_i:\bar{V}_i\to\mathcal{V}.$$

The composition

$$\Phi_{10}:\Psi_1\circ\Psi_0^{-1}:V_0\to V_1$$

is a diffeomorphism. It induces a homeomorphism

$$\bar{\Phi}_{10}:\bar{V}_0\to\bar{V}_1.$$

Note that

$$\bar{\Phi}_{10} = \Psi_1\circ\bar{\Phi}_0.$$

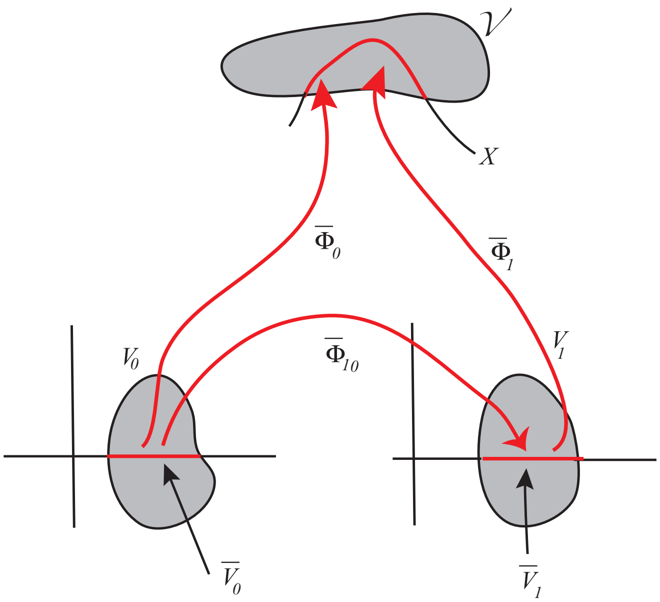

Figure 16.27. *The transition map determined by two overlapping straightening diffeomorphisms.*

According to Proposition 16.3.13(i) the induced map $\overline{\Phi}_{10}$ is a diffeomorphism with image $\overline{V}_1$. We will refer to $\overline{\Phi}_{10}$ as the *transition diffeomorphism* associated to the pair of overlapping s.d.-s $(\mathcal{U}_i, \Psi_i)$, $i = 0, 1$.

Definition 16.3.14. Let $X \subset \mathbb{R}^n$, $1 \leqslant m \leqslant n$, be an m-dimensional C^1-submanifold.

(i) An *atlas* of X is a collection of s.d.-s $\big\{ (\mathcal{U}_i, \Psi_i) \big\}_{i \in I}$ such that the collection $(\mathcal{U}_i)_{i \in I}$ is an open cover of X. For $i, j \in I$ we set $\mathcal{U}_{ij} = \mathcal{U}_i \cap \mathcal{U}_j$.

(ii) An *orientation* of an atlas $\big\{ (\mathcal{U}_i, \Psi_i) \big\}_{i \in I}$ of X is a choice of orientations $\epsilon_i \in \mathbb{O}(\overline{U}_i)$, $\overline{U}_i = \Psi_i(\mathcal{U}_i \cap X) \subset \mathbb{R}^m$, $i \in I$. The orientation is called *coherent* if, for any $i, j \in I$ such that $\mathcal{U}_{ij} \neq \varnothing$, the associated transition map

$$\overline{\Phi}_{ji} : (\overline{V}_i, \epsilon_i) \to (\overline{V}_j, \epsilon_j),$$

is orientation preserving. Above $\overline{V}_i = \Psi_i\big(\mathcal{U}_{ij}\big)$, $\overline{V}_j = \Psi_j\big(\mathcal{U}_{ij}\big)$.

(iii) The submanifold X is called *orientable* if it admits a *coherently* oriented atlas.

(iv) An *orientation* on X is a choice of a coherently oriented atlas.

(v) Two orientations on X given by the coherently oriented atlases

$$\mathcal{A} := \big\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \big\}_{i \in I}, \;\; \mathcal{B} := \big\{ (\mathcal{V}_j, \Psi_j, \epsilon_j) \big\}_{j \in J}$$

are to be considered equivalent if their union is also a *coherently* oriented atlas.

□

Example 16.3.15. Suppose that X is described by $n-m$ equations

$$X := \bigl\{\, \boldsymbol{x} \in \mathbb{R}^n : \;\; F^1(\boldsymbol{x}) = \cdots = F^{n-m}(\boldsymbol{x}) = 0, \;\; F^1, \dots, F^{n-m} \in C^1(\mathbb{R}^n)$$

such that, for

$$\forall \boldsymbol{x} \in X, \text{ the vectors } \nabla F^1(\boldsymbol{x}), \dots, \nabla F^{n-m}(\boldsymbol{x}) \text{ are linearly independent.}$$

Suppose that $\mathcal{A} := (\mathcal{U}_i, \Psi_i)$ is an atlas for X. Set as usual

$$\overline{U}_i := \Psi_i(\mathcal{U}_i \cap X) \subset \mathbb{R}^m, \;\; \overline{\Phi}_i := \Psi_i^{-1}\big|_{\overline{U}_i}.$$

For simplicity we set $\boldsymbol{x}(\boldsymbol{\nu}) := \overline{\Phi}_i(\boldsymbol{u})$. Denote by $u^1, \dots, u^m$ the canonical Cartesian coordinates on $\overline{U}_i$ and we set

$$T_k(\boldsymbol{u}) = \frac{\partial \overline{\Phi}_i}{\partial u^i}(\boldsymbol{u}), \;\; k = 1, \dots, m.$$

The collection $\{T_1(\boldsymbol{u}), \dots, T_m(\boldsymbol{u})\}$ is a basis of the tangent space $T_{\boldsymbol{x}(\boldsymbol{u})}X$. The vectors $\nabla F^j(\boldsymbol{x}(\boldsymbol{u}))$ are perpendicular to this space. It follows that the $n \times n$ matrix $B_i(\boldsymbol{u})$ with columns

$$\nabla F^1(\boldsymbol{x}(\boldsymbol{u})), \dots, \nabla F^{n-m}(\boldsymbol{x}(\boldsymbol{u})), T_1(\boldsymbol{u}), \dots, T_m(\boldsymbol{u})$$

is nonsingular. We obtain an orientation ϵ_i on $\overline{U}_i$ given by

$$\epsilon_i(\boldsymbol{u}) = \operatorname{sign} \det B_i(\boldsymbol{u}).$$

One can show that the collection $\bigl\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \bigr\}$ is a coherently oriented atlas and thus defines an orientation on X.

The natural proof of this fact is based on a bit more differential geometry than I can safely assume you, the reader, may know at this point in time. There exist proofs of this claim that use essentially only linear algebra, but the geometric meaning will be lost in the heap computations. For this reason I have decided not to include a proof of this claim. Instead, I encourage you to supply a proof in the special case $m = 2$, $n = 3$ and compare this with the arguments in Remark 16.2.23. □

16.3.3. Integration along oriented submanifolds. Suppose that $X \subset \mathbb{R}^n$ is an orientable m-dimensional C^1-submanifold. We denote by $\Omega^m_{\text{cpt}}(X)$ the subspace of $\Omega^m_{\text{cpt}}(\mathbb{R}^n)$ consisting of compactly supported degree m forms ω such that $X \cap \operatorname{supp} \omega$ is a compact subset of X. We want to associate to any orientation $\boldsymbol{or}_X$ on X an integration map

$$\int_{X, \boldsymbol{or}_X} : \Omega^m_{\text{cpt}}(X) \to \mathbb{R}.$$

We will build this integral in stages.

Fix an orientation $\vec{\epsilon}$ on X defined by a coherently oriented atlas

$$\mathcal{A} = \bigl\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \bigr\}_{i \in I}$$

We denote by $\Omega^m_{\text{cpt}}(X, \mathcal{A})$ the subspace of $\Omega^m_{\text{cpt}}(\mathbb{R}^n)$ consisting of compactly supported degree m forms ω such that

- The support of ω is contained in the union
$$\mathfrak{U}_{\mathcal{A}} := \bigcup_{i\in I} \mathfrak{U}_i.$$
- $X \cap \operatorname{supp}\omega$ is a compact subset of X.

We will define a canonical linear map
$$\int_X = \int_{X,\mathcal{A}} : \Omega^m_{\mathrm{cpt}}(X,\mathcal{A}) \to \mathbb{R},$$
$$\Omega^m_{\mathrm{cpt}}(X,\mathcal{A}) \ni \omega \mapsto \int_{X,\mathcal{A}} \omega.$$
We achieve this in several steps.

Step 1. The form ω has small support, i.e., $\exists i \in I$ such that $\operatorname{supp}\omega \cap X \subset \mathfrak{U}_i \cap X$. Consider the map
$$\bar{U}_i \ni \boldsymbol{u} \mapsto \boldsymbol{x}(\boldsymbol{u}) = \bar{\Phi}_i(\boldsymbol{u}) \in \mathfrak{U}.$$
Set
$$\bar{\omega}_i := \bar{\Phi}_i^*\omega \in \Omega^m(\bar{U}_i).$$
In this case we set
$$\int_{X,\mathcal{A}} \omega := \int_{\bar{U}_{i,\epsilon_i}} \bar{\omega}_i,$$
where $\int_{U,\epsilon}$ is defined in (16.3.8). Suppose $\operatorname{supp}\omega \cap X \subset \mathfrak{U}_i \cap X$.

Suppose that we also have $\operatorname{supp}\omega \cap X \subset \mathfrak{U}_j \cap X$ for a different $j \in I$. Then, we can propose new definition of $\int_X \omega$
$$\int_{X,\mathcal{A}} \omega := \int_{\bar{U}_{j,\epsilon_j}} \bar{\omega}_j.$$
Set
$$\bar{V}_i = \Psi_i(\mathfrak{U}_i \cap \mathfrak{U}_j \cap X),\ \ \bar{V}_j = \Psi_j(\mathfrak{U}_i \cap \mathfrak{U}_j \cap X).$$
Thus
$$\operatorname{supp}\bar{\omega}_i \subset \bar{V}_i,\ \ \operatorname{supp}\bar{\omega}_j \subset \bar{V}_j.$$
From Proposition 16.3.13 we deduce that
$$\bar{\omega}_i = \bar{\Phi}_{ji}^*\bar{\omega}_j.$$
Since
$$\bar{\Phi}_{ji} : (\bar{V}_i, \epsilon_i) \to (\bar{V}_j, \epsilon_j)$$
is orientation preserving we have
$$\int_{\bar{U}_{i,\epsilon_i}} \bar{\omega}_i = \int_{\bar{V}_{i,\epsilon_i}} \bar{\omega}_i \overset{(16.3.9)}{=} \int_{\bar{V}_{j,\epsilon_j}} \bar{\omega}_j = \int_{\bar{U}_{j,\epsilon_j}} \bar{\omega}_j.$$
Set $X_i := X \cap \mathfrak{U}_i$. We have thus defined an integration map
$$\int_{X_i} : \Omega^m_{\mathrm{cpt}}(X_i) \to \mathbb{R}.$$

This map is linear and it is independent of any other choice of local coordinates we could choose on X_i.

Step 2. Extension to $\Omega^m_{\mathrm{cpt}}(X,\mathcal{A})$. Let $\omega \in \Omega^m_{\mathrm{cpt}}(X,\mathcal{A})$. Choose a continuous partition of unity on $\operatorname{supp}\omega$

$$\psi_1,\dots,\psi_k : \mathbb{R}^n \to \mathbb{R}$$

subordinated to the open cover $(\mathcal{U}_i)_{i\in I}$. For each $a = 1,\dots,k$ choose $i(a) \in I$ such that $\operatorname{supp}\psi_a \subset \mathcal{U}_{i(a)}$. We have

$$\omega = \sum_{a=1}^{l} \psi_a\omega.$$

Note that $\omega_a = \psi_a\omega \in \Omega^m_{\mathrm{cpt}}\big(\,X_{i(a)}\,\big)$. We define

$$\int_{X,\mathcal{A}} \omega = \sum_a \int_{X_{i(a)}} \psi_a\omega$$

A priori, this definition depends on the choice of the partition of unity. Let us show that this is not the case.

Choose another partition of unity on $\operatorname{supp}\omega$, $\phi_1,\dots,\phi_\ell$, subordinated to the cover $(\mathcal{U}_i)_{i\in I}$. For each $b = 1,\dots,\ell$ choose $j(b) \in I$ such that $\operatorname{supp}\phi_b \subset \mathcal{U}_{j(b)}$. Note that

$$\omega_a = \sum_b \underbrace{\phi_b\omega_a}_{\omega_{ab}}$$

Since $\operatorname{supp}\omega_{ab} \subset \mathcal{U}_{(i(a)} \cap \mathcal{U}_{j(b)}$ we have

$$\int_{X_{i(a)}} \omega_a = \sum_b \int_{X_{i(a)}} \omega_{ab} = \sum_b \int_{X_{j(b)}} \omega_{ab},$$

so

$$\sum_a \int_{X_{i(a)}} \psi_a\omega = \sum_a \int_{X_{i(a)}} \omega_a = \sum_b \int_{X_{j(b)}} \underbrace{\sum_a \omega_{ab}}_{=\phi_b\omega} = \sum_b \int_{X_{j(b)}} \phi_b\omega.$$

This proves that the definition of $\int_{X,\mathcal{A}}$ is independent of the choices of partions of unity.

Let us observe that the above proof shows that if $\omega_1,\omega_2 \in \Omega_{\mathrm{cpt}}(X,\mathcal{A})$ coincide in an open neighborhood of X, then

$$\int_{X,\mathcal{A}} \omega_1 = \int_{X,\mathcal{A}} \omega_2.$$

Step 3. The argument in the previous step shows that if $\mathcal{A}$ and $\mathcal{B}$ are two coherently oriented atlases such that $\mathcal{A} \subset \mathcal{B}$, then

$$\Omega^m_{\mathrm{cpt}}(X,\mathcal{A}) \subset \Omega^m_{\mathrm{cpt}}(X,\mathcal{B}),$$

and, for any $\omega \in \Omega^m_{\mathrm{cpt}}(X,\mathcal{A})$, we have

$$\int_{X,\mathcal{A}} \omega = \int_{X,\mathcal{B}} \omega.$$

In particular this shows that if the coherently oriented atlases $\mathcal{A}$ and $\mathcal{B}$ define equivalent orientation, then for any $\omega \in \Omega^m_{\text{cpt}}(X,\mathcal{A}) \cap \Omega^m_{\text{cpt}}(X,\mathcal{B})$ we have

$$\int_{X,\mathcal{A}} \omega = \int_{X,\mathcal{A}\cup B} \omega = \int_{X,\mathcal{B}} \omega.$$

Step 4. Using partitions of unity one can show (but we will skip the details) that for any $\omega \in \Omega^m_{\text{cpt}}$ and for any coherently oriented atlas $\mathcal{A}$ defining an orientation $\boldsymbol{or}_X$ on X there exists a form $\tilde{\omega} \in \Omega^m_{\text{cpt}}(X,\mathcal{A})$ such that $\omega = \tilde{\omega}$ in an open neighborhood of X. We then set

$$\int_{X,\boldsymbol{or}_X} \omega := \int_{X,\mathcal{A}} \tilde{\omega}.$$

Clearly the right-hand side does not depend on any particular choices of $\tilde{\omega}$.

16.3.4. The general Stokes' formula. We first need to introduce the concept of manifolds with boundary. This is a simple generalization of the concept of surface with boundary introduced in Definition 16.2.2. We will skip many technical details.

Definition 16.3.16. Let $k,m,n \in \mathbb{N}$, $n \geqslant m \geqslant 1$. An m-dimensional C^k-*submanifold with boundary* in $\mathbb{R}^n$ is a *compact* subset $X \subset \mathbb{R}^n$ such that, for any point $\boldsymbol{p}_0$, there exists an open neighborhood $\mathcal{U}$ of $\boldsymbol{p}_0$ in $\mathbb{R}^n$ and a C^k-diffeomorphism $\Psi : \mathcal{U} \to \mathbb{R}^n$ such that the image $\overline{U} = \Psi(\mathcal{U} \cap X)$ is contained in the subspace $\mathbb{R}^m \times \boldsymbol{0} \subset \mathbb{R}^n$ and it is either

(I) an open ball in $\mathbb{R}^m$ centered at $\Psi(\boldsymbol{p}_0)$ or

(B) the point $\Psi(\boldsymbol{p}_0)$ lies in plane $\{x^1 = 0\} \subset \mathbb{R}^m$ and $\overline{U}$ it is the intersection of an open ball $B_r(\boldsymbol{p}_0)$ with the half-plane

$$\boldsymbol{H}^m_- := \big\{ (x^1, x^2, \dots, x^m) \in \mathbb{R}^2;\;\; x^1 \leqslant 0 \big\}.$$

The pair $(\mathcal{U},\Psi)$ is called a *straightening diffeomorphism* (abbreviated s.d.) at $\boldsymbol{p}_0$. The pair $\big(\mathcal{U} \cap X, \Psi|_{\mathcal{U}\cap X} \big)$ is called a *local coordinate chart* of X at $\boldsymbol{p}_0$.

In the case (B), the point $\boldsymbol{p}_0 \in X$ is called a *boundary point* of X. Otherwise $\boldsymbol{p}_0$ is called an *interior* point.

The set of boundary points of X is called the *boundary* of X and it is denoted by ∂X. The set of interior points of X is called the *interior* of X and it is denoted by X°. The submanifold with boundary is called *closed* if its boundary is empty, $\partial X = \emptyset$. □

An atlas of a manifold with boundary X is a collection of s.d.-s $\big\{ (\mathcal{U}_i, \Psi_i) \big\}_{i\in I}$ such that

$$X \subset \bigcup_{i\in I} \mathcal{U}_i.$$

An *orientation* of a s.d. $(\mathcal{U},\Psi)$ is an orientation on the interior of $\Psi(X \cap \mathcal{U})$. The transition maps are defined in a similar fashion which leads as in the boundary-less case to the concept of orientation of a manifold with boundary. Equivalently, an orientation on a manifold with boundary is equivalent to a choice of orientation on its interior.

There is a new phenomenon. Namely, an orientation on X induces in a natural fashion an orientation on its boundary ∂X.

Suppose that $\mathcal{A} := \big\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \big\}_{i \in I}$ is a coherently oriented atlas of X. The s.d.-s $(\mathcal{U}_i, \Psi_i)$ are of two types.

- *interior*, i.e., $\mathcal{U} \cap \partial X = \varnothing$ and
- *boundary*, i.e., $\mathcal{U} \cap \partial X \neq \varnothing$.

Consider the subcollection $\mathcal{A}_\partial := \big\{ (\mathcal{U}_a, \Psi_a) \big\}_{a \in A \subset I}$ consisting of all the boundary type s.d.-s in $\mathcal{A}$. The collection $\mathcal{A}_\partial$ is also an atlas for the manifold ∂X. For any $a \in A$ we set

$$\overline{V}_a := \Psi_a(\mathcal{U}_A \cap \partial X) \subset \partial \boldsymbol{H}^m_-.$$

Note that for $a \in A$ we have

$$\overline{U}_a := \Psi_a(\mathcal{U}_a \cap X) = B_{r(a)}(\boldsymbol{p}_a) \cap \boldsymbol{H}^m_-, \;\; \boldsymbol{p}_a \in \partial \boldsymbol{H}^m_-.$$

Denote by $u^1, \dots, u^m$ the Cartesian coordinates on the space $\mathbb{R}^m$ where the half-space $\boldsymbol{H}^m_-$ lives. An orientation ϵ_a on the interior $\boldsymbol{int}\overline{U}_a$ is given by the top degree form

$$\omega_a := \epsilon_a du^1 \wedge du^2 \wedge \cdots \wedge du^m.$$

The boundary $\partial \boldsymbol{H}^m_-$ is the subspace $\mathbb{R}^{m-1}$ with Cartesian coordinates $u^2, \dots, u^m$. The induced orientation on $\overline{V}_a$ is, denoted by $\partial \epsilon_a$ is described by the top degree form

$$\omega_a^\partial := \epsilon_a du^2 \wedge \cdots \wedge du^m.$$

There is a simple mnemonic device to help you remember this construction. It is called the *outer conormal first* convention. Let us explain.

Note that traveling in $\boldsymbol{H}^m_-$ in the direction of increasing u^1 one eventually exits $\boldsymbol{H}^m_-$. Equivalently, observe that along $\partial \boldsymbol{H}^m_-$ the vector field $\boldsymbol{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^m$ is an outer pointing normal vector field. Note that

$$\omega_a = du^1 \wedge \omega_a^\partial,$$

or,

$$\text{orientation interior} = \boxed{\text{outer conormal}} \wedge \text{orientation boundary},$$

whence the terminology outer conormal *first*.

One can show that if

$$\mathcal{A} := \big\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \big\}_{i \in I}$$

is a coherently oriented atlas of the manifold with boundary X, then the collection

$$\mathcal{A}_\partial = \big\{ (\mathcal{U}_a, \Psi_a, \partial\epsilon_a) \big\}_{a \in A}, \;\; A = \big\{ a \in I; \;\; U_a \cap \partial X \neq \varnothing \big\},$$

is a coherently oriented atlas of ∂X. While we will not present all the tedious details, we want to explain the simple fact behind this. Its proof is left to you as an exercise.

Lemma 16.3.17. *Let $(\bar{U}_i, \epsilon_i)$, $i = 0, 1$, be two oriented open subsets of $\mathbb{R}^m$ such that $\bar{U}_i \cap \partial \boldsymbol{H}^m_- \neq \varnothing$ for all $i = 0, 1$, If $\Phi : (\bar{U}_0, \epsilon_0) \to (\bar{U}_1, \epsilon_1)$ is an orientation preserving diffeomorphism such that*

$$\Phi\big(\bar{U}_0 \cap \partial \boldsymbol{H}^m_-\big) \subset \partial \boldsymbol{H}^m_-,$$

then the induced map

$$\Phi_\partial : (\bar{U}_0 \cap \partial \boldsymbol{H}^m_-, \partial\epsilon_0) \to (\bar{U}_1 \cap \partial \boldsymbol{H}^m_-, \partial\epsilon_1),$$

is also orientation preserving. □

Just like in the case of submanifolds without boundary, an orientation ϵ on an m-dimensional manifold with boundary defines an integration map

$$\int_{X,\epsilon} : \Omega^m(\mathbb{R}^n) \to \mathbb{R}.$$

The submanifolds with boundary are, by our definition, compact so we no longer need to work with compactly supported forms.

The construction follows the same four steps as in the bondary-less case so we can safely omit de details.

At the same time, we have another oriented submanifold $(\partial X, \partial\epsilon)$. In particular, we also have an integration map

$$\int_{\partial X, \partial\epsilon} : \Omega^{m-1}(\partial X) \to \mathbb{R}.$$

Theorem 16.3.18 (General Stokes' formula)**.** *Suppose that (X, ϵ) is an m-dimensional oriented C^1-submanifold with boundary of $\mathbb{R}^n$. Then for any $\omega \in \Omega^{m-1}(\mathbb{R}^n)_{C^1}$ we have*

$$\int_{\partial X, \partial\epsilon} \omega = \int_{X,\epsilon} d\omega, \tag{16.3.11}$$

where $d\omega \in \Omega^m(\mathbb{R}^n)$ is the exterior derivative of ω.

Proof. Fix a coherently oriented atlas

$$\mathcal{A} := \big\{ (\mathcal{U}_i, \Psi_i, \epsilon_i) \big\}_{i \in I}$$

that defines the orientation ϵ. Set

$$\mathcal{U}_{\mathcal{A}} := \bigcup_{i \in I} U_i$$

Fix a compact set K such that[6]

$$X \subset \boldsymbol{int}\, K, \quad K \subset \mathcal{U}_{\mathcal{A}} \tag{16.3.12}$$

Using the results in Exercise 13.14 we can find a C^1-partition of unity along K and subordinated to the open cover $(U_i)_{i \in I}$. Recall that this is a *finite* collection $(\chi_s)_{s \in S}$ of compactly supported C^1-functions $\chi_s : \mathbb{R}^n \to \mathbb{R}$ satisfying the following properties.

[6] Can you see why a compact set K satisfying (16.3.12) exists?

- For all $s \in S$ there exists $i = i(s) \in I$ such that $\operatorname{supp} \chi_s \subset \mathcal{U}_{i(s)}$.
- $$\sum_{s \in S} \chi_s(\boldsymbol{x}) = 1, \;\; \forall \boldsymbol{x} \in K.$$

Let $\omega \in \Omega^{m-1}(\mathbb{R}^n)_{C^1}$. For $s \in S$ set

$$\eta_s := \chi_s \omega \in \Omega^{m-1}_{\mathrm{cpt}}(\mathbb{R}^n),$$

and define

$$\eta := \sum_s \eta_s.$$

Note that on $\boldsymbol{int}\, K \supset X$ we have

$$\omega = \eta, \;\; d\omega = d\eta = \sum_s d\eta_s.$$

so it suffices to prove (16.3.11) for η. On the other hand

$$\int_{\partial X, \partial \epsilon} \eta = \sum_s \int_{\partial X, \partial \epsilon} \eta_s,$$

$$\int_{X, \epsilon} d\eta = \sum_s \int_{X, \epsilon} d\eta_s$$

so it suffices to prove (16.3.11) for each of the individual η_s. Thus we have to prove that (16.3.11) holds form $\eta \in \Omega^{m-1}_{\mathrm{cpt}}(\mathbb{R}^n)_{C^1}$ satisfying the additional propperty that there exists an oriented s.d. $(\mathcal{U}, \Psi, \epsilon)$ such that $\operatorname{supp} \eta \subset \mathcal{U}$. We distingush two cases.

Interior case, i.e., $\mathcal{U} \cap \partial X = \emptyset$. In this case η is identically zero in a neighborhood of ∂X so

$$\int_{\partial X, \partial \epsilon} \eta = 0.$$

We have to prove that

$$\int_{X, \epsilon} d\eta = 0.$$

We set $\overline{U} = \Psi(\mathcal{U} \cap X)$ so that $\overline{U}$ is an open subset of $\mathbb{R}^m$. Denote by Φ the inverse of Ψ and by $\overline{\Phi}$ the restriction of Φ to $\overline{U}$. Then, according to Proposition 16.3.12(ii), we have

$$\int_{X, \epsilon} d\eta = \int_{\overline{U}, \epsilon} \overline{\Phi}^* d\eta.$$

We set

$$\bar{\eta} := \overline{\Phi}^* \eta.$$

From Proposition 16.3.12(i) we deduce that

$$\overline{\Phi}^* d\eta = d\overline{\Phi}^* \eta d\bar{\eta}.$$

Thus we have to show that

$$\int_{\overline{U}, \epsilon} d\bar{\eta} = 0,$$

for any $\bar{\eta} \in \Omega^{m-1}_{\mathrm{cpt}}(\overline{U})_{C^1}$.

Let $\bar{\eta}$ be such a degree $(m-1)$ form. Fix a positive number R such that

$$\overline{U} \subset C_R^m := [-R, R]^m \subset \mathbb{R}^m.$$

We have

$$\begin{aligned}\bar{\eta} = \eta_1 du^2 \wedge du^3 \wedge \cdots \wedge du^m + \eta_2 du^1 \wedge du^3 \wedge \cdots \wedge du^m \\ + \cdots + \eta_m du^1 \wedge du^2 \wedge \cdots \wedge du^{m-1}.\end{aligned}$$

This can be written in a more compact form as

$$\bar{\eta} = \sum_{k=1}^m \eta_k du^1 \wedge \cdots \wedge \widehat{du^k} \wedge \cdots \wedge du^m, \tag{16.3.13}$$

where a hat $\widehat{\ }$ indicates a missing entry. We deduce

$$\begin{aligned} d\bar{\eta} &= \left(\frac{\partial \eta_1}{\partial u^1} - \frac{\partial \eta_2}{\partial u^2} + \cdots + (-1)^{m-1} \frac{\partial \eta_m}{\partial u^m} \right) du^1 \wedge du^2 \wedge \cdots \wedge du^m \\ &= \left(\sum_{k=1}^m (-1)^{k-1} \frac{\partial \eta_k}{\partial u^k} \right) du^1 \wedge du^2 \wedge \cdots \wedge du^m. \end{aligned} \tag{16.3.14}$$

Then

$$\int_{\overline{U}, \epsilon} d\bar{\eta} = \sum_{k=1}^m (-1)^{k-1} \epsilon \int_{\overline{U}} \frac{\partial \eta_k}{\partial u^k} \, |du^1 \cdots du^m|.$$

We will prove that

$$\int_{\overline{U}} \frac{\partial \eta_k}{\partial u^k} \, |du^1 \cdots du^m| = 0, \quad \forall k = 1, \dots, m.$$

For simplicity we discuss only the case $k = 1$. The other cases are completely similar. We have

$$\int_{\overline{U}} \frac{\partial \eta_1}{\partial u^1} \, |du^1 \cdots du^m| = \int_{C_R^m} \frac{\partial \eta_1}{\partial u^1} \, |du^1 \cdots du^m|$$

(use Fubini)

$$= \int_{C_R^{m-1}} \left(\int_{-R}^R \frac{\partial \eta_1}{\partial x^1} \, |dx^1| \right) |du^2 \cdots du^m|$$

$$= \int_{C_R^{m-1}} \Big(\underbrace{\eta_1(R, u^2, \dots, u^m)}_{=0} - \underbrace{\eta_1(-R, u^2, \dots, u^m)}_{=0} \Big) = 0.$$

Boundary case, i.e., $\mathcal{U} \cap \partial X = \varnothing$. In this case $\overline{U} := \Psi(\mathcal{U} \cap X)$ is a half-ball (see Figure 16.28)

$$\overline{U} = \boldsymbol{H}_-^m \cap B_r(\boldsymbol{p}_0), \quad \boldsymbol{p}_0 \in \partial \boldsymbol{H}_-^m$$

We can find $L > 0$ such that $\overline{U}$ is contained in the closed box

$$B^- = [-L, L]^m \cap \boldsymbol{H}_-^m \subset \mathbb{R}^m.$$

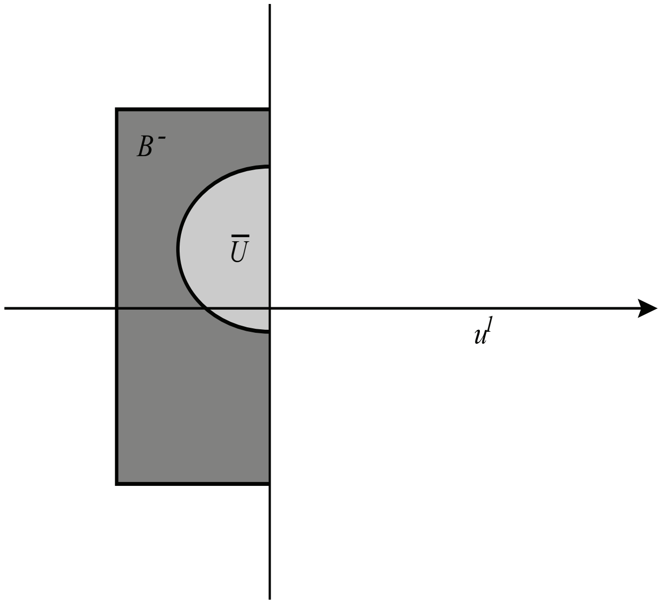

Figure 16.28. *Integrating over a half-ball.*

We define

$$B^0 := [-L,L]^m \cap \partial \boldsymbol{H}^m_- = \{0\} \times [-L,L]^{m-1}.$$

. Arguing as in the previous case we deduce that

$$\int_{X,\epsilon} d\eta = \int_{\overline{U},\epsilon} d\bar{\eta} = \int_{B^-,\epsilon} d\bar{\eta}, \;\; \int_{\partial X, \partial\epsilon} \eta = \int_{B^0,\partial\epsilon} \bar{\eta}.$$

If

$$\bar{\eta} = \sum_{k=1}^m \eta_k du^1 \wedge \cdots \wedge \widehat{du^k} \wedge \cdots \wedge du^m,$$

then

$$\int_{B^-,\epsilon} d\bar{\eta} = \sum_{k=1}^m (-1)^{k-1} \epsilon \int_{B^-} \frac{\partial \eta_k}{\partial u^k} \, |du^1 \cdots du^m|,$$

and

$$\int_{B^0,\partial\epsilon} = \epsilon \int_{B^0} \eta_1(0,u^2,\dots,u^m) \, |du^2 \cdots du^m|.$$

We will show that

$$\int_{B^-} \frac{\partial \eta_1}{\partial u^1} \, |du^1 \cdots du^m| = \int_{B^0} \eta_1(0,u^2,\dots,u^m) \, |du^2 \cdots du^m|, \tag{16.3.15a}$$

$$\int_{B^-} \frac{\partial \eta_k}{\partial u^k} \, |du^1 \cdots du^m| = 0, \;\; \forall k = 2,\dots,m. \tag{16.3.15b}$$

To prove (16.3.15a) we use Fubini's theorem and we deduce

$$\int_{B^-}\frac{\partial\eta_1}{\partial u^1}\,|du^1\cdots du^m| = \int_{|u^k|\leqslant L,\;2\leqslant k\leqslant m}\left(\int_{-L}^0\frac{\partial\eta_1}{\partial u^1}du^1\right)|du^2\cdots du^m|$$

(use the Fundamental Theorem of Calculus)

$$= \int_{|u^k|\leqslant L,\;2\leqslant k\leqslant m}\left(\eta_1(0,u^2,\dots,u^m)-\underbrace{\eta_1(-L,u^2,\dots,u^m)}_{=0}\right)|du^2\cdots du^m|$$

$$\int_{B^0}\eta_1(0,u^2,\dots,u^m)\,|du^2\cdots du^m|$$

The equality (16.3.15b) also follows from Fubini's theorem. We prove only the case $k=m$ which involves simpler notations.

$$\int_{B^-}\frac{\partial\eta_m}{\partial u^m}\,|du^1\cdots du^m|$$

$$= \int_{[-L,0]\times[-L,L]^{m-2}}\left(\int_{-L}^{L}\frac{\partial\eta_m}{\partial u^m}(u^1,\dots,u^{m-1},u^m)\,du^m\right)|du^1du^2\cdots du^{m-1}|$$

(use the Fundamental Theorem of Calculus)

$$= \int_{[-L,0]\times[-L,L]^{m-2}}\underbrace{\left(\eta_m(u^1,\dots,u^{m-1},u^m)\Big|_{u^m=-L}^{u^m=L}\right)}_{=0}|du^1du^2\cdots du^{m-1}| = 0.$$

□

16.3.5. What are these differential forms anyway. In lieu of epilogue to this chapter, I will try to crack open the door to another world to which the considerations in this last section properly belong.

We've developed a theory of integration of objects whose nature was left nebulous. What are these differential forms?

Suppose that U is an open subset of $\mathbb{R}^n$, $n\geqslant 2$. The equality (13.2.12) of Example 13.2.11 explained that the terms dx^i should be viewed as *linear forms*. A linear form, as you know, is a "beast" that, when fed a vector it spits out a number. If you feed a vector $\boldsymbol{v}$ to the "beast" dx^i, it will return the number v^i, the i-th coordinate of the vector $\boldsymbol{v}$.

The exterior monomials $dx^i\wedge dx^j$, $dx^i\wedge dx^j\wedge dx^k$ etc. are more sophisticated "beasts": they are multilinear maps with certain additional properties. To explain their nature we consider a slightly more complicated situation.

Let $m\in\mathbb{N}$ and consider m linear functionals

$$\alpha^1,\dots,\alpha^m\to\mathbb{R}.$$

Their exterior product is m-linear map

$$\alpha^1\wedge\cdots\wedge\alpha^m:\underbrace{\mathbb{R}^n\times\cdots\times\mathbb{R}^n}_{m}\to\mathbb{R},$$

$$\alpha^1\wedge\cdots\wedge\alpha^m(\boldsymbol{v}_1,\dots,\boldsymbol{v}_m):=\det\begin{bmatrix}\alpha^1(\boldsymbol{v}_1) & \alpha^1(\boldsymbol{v}_2) & \cdots & \alpha^1(\boldsymbol{v}_m)\\ \alpha^2(\boldsymbol{v}_1) & \alpha^2(\boldsymbol{v}_2) & \cdots & \alpha^2(\boldsymbol{v}_m)\\ \vdots & \vdots & \ddots & \vdots\\ \alpha^m(\boldsymbol{v}_1) & \alpha^m(\boldsymbol{v}_2) & \cdots & \alpha^m(\boldsymbol{v}_m)\end{bmatrix}. \tag{16.3.16}$$

Note that the m-linear form $\alpha^1\wedge\cdots\wedge\alpha^m$ satisfies the skew-symmetry conditions

$$\alpha^{\sigma(1)}\wedge\cdots\wedge\alpha^{\sigma(m)}=\epsilon(\sigma)\alpha^1\wedge\cdots\wedge\alpha^m,$$

$$\alpha^1\wedge\cdots\wedge\alpha^m(\boldsymbol{v}_{\sigma(1)},\dots,\boldsymbol{v}_{\sigma(m)})=\epsilon(\sigma)\alpha^1\wedge\cdots\wedge\alpha^m(\boldsymbol{v}_1,\dots,\boldsymbol{v}_m),$$

for any permutation $\sigma\in\mathfrak{S}_m$. Let us observe that if $m>n$, then any collection of m vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\in\mathbb{R}^n$ is linearly dependent and we deduce from (16.3.16) that

$$\alpha^1\wedge\cdots\wedge\alpha^m=0,$$

for any linear forms $\alpha_1,\dots,\alpha_m:\mathbb{R}^n\to\mathbb{R}$.

When $m=n$ and $\alpha^i=\dot{x}^i$, then

$$dx^1\wedge\cdots\wedge dx^n(\boldsymbol{v}_1,\dots,\boldsymbol{v}_n)=\det\big[\boldsymbol{v}^i_j\big]_{1\leqslant i,j\leqslant n}\overset{(15.3.5)}{=}\pm\operatorname{vol}\big(\boldsymbol{P}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_n)\big),$$

where we recall that $\boldsymbol{P}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_n)$ denotes the parallelepiped spanned by $\boldsymbol{v}_1,\dots,\boldsymbol{v}_n$.

More generally, if $m<n$, then for any vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m$, the number

$$dx^1\wedge\cdots\wedge dx^m(\boldsymbol{v}^1,\dots,\boldsymbol{v}_m)$$

is equal, up to a sign, with the volume of the parallelepiped spanned by the orthogonal projections of the vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m$ onto the m-dimensional subspace of $\mathbb{R}^n$ with coordinates $x^1,\dots,x^m$.

In general, and exterior form of degree m is an m-linear map

$$\omega:\underbrace{\mathbb{R}^n\times\cdots\times\mathbb{R}^n}_{m}\to\mathbb{R},$$

satisfying the skew-symmetry condition

$$\omega(\boldsymbol{v}_{\sigma(1)},\dots,\boldsymbol{v}_{\sigma(m)})=\epsilon(\sigma)\omega(\boldsymbol{v}_1,\dots,\boldsymbol{v}_m),\ \ \forall\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\in\mathbb{R}^n,\ \ \sigma\in\mathfrak{S}_m. \tag{16.3.17}$$

Such a form can be thought of as "gauging m-dimensional" parallelepipeds. This gauging is of a special kind: its output depends on the order in which we "feed" the vectors spanning the parallelepiped according to (16.3.17) .

Take for the example the case $m=2$. Think of a parallelogram as having two faces: a white face and a black face. When a 2-form gauges a white-face-up parallelogram it outputs a number, but when it gauges the same parallelogram but with its black face up, it outputs the opposite number.

We can use the canonical basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ to express an m-form ω as a linear combination (compare with (16.3.1))

$$\omega = \sum_{I \in \mathrm{Inj}^+(m,n)} \omega_I dx^{\wedge I},$$

where, for any $I = (i_1, \dots, i_m) \in \mathrm{Inj}^+(m,n)$, we have

$$dx^{\wedge I} = dx^{i_1} \wedge \cdots \wedge dx^{i_m}, \;\; \omega_I := \omega(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_m}) \in \mathbb{R}.$$

A differential form of degree m on an open set $U \subset \mathbb{R}^m$ is a continuous assignment of an m-form $\omega_{\boldsymbol{x}}$ to each point $\boldsymbol{x} \in U$. More precisely this means that

$$\omega_{\boldsymbol{x}} = \sum_{I \in \mathrm{Inj}^+(m,n)} \omega_I(\boldsymbol{x}) dx^{\wedge I},$$

where $\omega_I(\boldsymbol{x})$ depends continuously on $\boldsymbol{x}$.

Intuitively we can think that we have a continuous family of “gauges” $\omega_{\boldsymbol{x}}$, where $\omega_{\boldsymbol{x}}$ is to be used to gauge parallelepipeds originating at $\boldsymbol{x}$.

In the case $m = 2$ such a differential form gauges parallelograms. In particular, given a surface $S \in \mathbb{R}^n$, such a form gauges infinitesimal parallelograms on S, i.e., parallelograms spanned by a pair of vectors tangent to the same point $\boldsymbol{x} \in S$. We use the form $\omega_{\boldsymbol{x}}$ to gauge such a parallelogram. An orientation on S is essentially a rule we use to determine the order in which we feed infinitesimal parallelograms to the differential form because we know that the output is order sensitive.

The above intuitive interpretation of differential forms gives a pretty accurate idea on the nature of differential forms. Unfortunately, it is essentially useless if we want to perform meaningful mathematical computations with them.

At this point a deeper look at the concept of differential form is needed and this requires substantial algebraic and analytic considerations. However at this point you have all knowledge you need to digest the classic booklet [**35**] of M. Spivak on this topic. Although it is more than half a century old as I write these lines, it remains very actual and a gem of mathematical writing. However don’t let the tiny size of [**35**] fool you: it has a high density of subtle ideas per square inch of page.

The good news is that you should be very familiar with the first half of [**35**]. The second half, on integration of differential forms and various Stokes’ formulæ, is a rather steep, but very rewarding intellectual climb.

16.4. Exercises

Exercise 16.1. Denote by C the line segment in the plane $\mathbb{R}^2$ that connects the points $\boldsymbol{p}_0 := (3,0)$ and $\boldsymbol{p}_1 := (0,4)$.

(i) Compute the integrals

$$\int_C ds, \quad \int_C f(\boldsymbol{p})ds, \quad f(x,y) = x^2 + y^2.$$

(ii) Compute the integral of the angular form

$$W_\Theta = \frac{-y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy$$

along the segment C equipped with the orientation corresponding to the travel from $\boldsymbol{p}_1$ to $\boldsymbol{p}_0$.

□

Exercise 16.2. Denote by S the open square $(-1,1) \times (-1,1)$ in $\mathbb{R}^2$. Suppose that $P, Q : S \to \mathbb{R}$ are continuous functions. We set

$$\omega = Pdx + Qdy.$$

Prove that the following statements are equivalent.

(i) There exists $f \in C^1(S)$ such that $df = \omega$, i.e.,

$$P = \frac{\partial f}{\partial x}, \quad Q = \frac{\partial f}{\partial y}.$$

(ii) For any piecewise C^1 path $\boldsymbol{\gamma} : [a,b] \to S$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$ we have

$$\int_{\boldsymbol{\gamma}} \omega = 0.$$

(iii) For any piecewise C^1 paths $\boldsymbol{\gamma}_i : [a_i, b_i] \to S$, $i = 1,2$, such that

$$\boldsymbol{\gamma}_1(a_1) = \boldsymbol{\gamma}_2(a_2), \quad \boldsymbol{\gamma}_1(b_1) = \boldsymbol{\gamma}_2(b_2)$$

we have

$$\int_{\boldsymbol{\gamma}_1} \omega = \int_{\boldsymbol{\gamma}_2} \omega.$$

Hint. (iii) ⇒ (i) Define

$$f(x,y) = \int_0^x P(s,0)ds + \int_0^y Q(x,t)dt$$

and use (iii) prove that

$$f(x+h,y) = f(x,y) + \int_x^{x+h} P(s,y)ds, \quad f(x,y+k) = f(x,y) + \int_y^{y+k} Q(x,t)dt,$$

and $\partial_x f = P$, $\partial_y f = Q$.

□

Exercise 16.3. Let $U \subset \mathbb{R}^n$ be an open set and suppose that $f, g : U \to \mathbb{R}$ are C^2-functions. Prove that

$$\Delta g = \operatorname{div} \nabla g, \;\; \operatorname{div}(f\nabla g) = \langle \nabla f, \nabla g\rangle + f\Delta g,$$

where Δg is the Laplacian of g,

$$\Delta g := \sum_{k=1}^{n} \partial^2_{x^k} g. \qquad \square$$

Exercise 16.4. Let $D \subset \mathbb{R}^2$ be a bounded domain with C^1-boundary, U an open set containing $\boldsymbol{cl}\, D$ and $f, g : U \to \mathbb{R}$ are C^2 functions. Denote by $\boldsymbol{\nu}$ the outer normal vector field along ∂D. Prove that

$$\int_D f\Delta g\, |dxdy| = \int_{\partial D} f(\boldsymbol{p}) \frac{\partial g}{\partial \boldsymbol{\nu}}(\boldsymbol{p})|ds| - \int_D \langle \nabla f, \nabla g\rangle\, |dxdy|,$$

$$\int_{\partial D} \frac{\partial g}{\partial \boldsymbol{\nu}}\, |ds| = \int_D \Delta g\, |dxdy|,$$

$$\int_D f\Delta g|dxdy| = \int_{\partial D} \left(f(\boldsymbol{p})\frac{\partial g}{\partial \boldsymbol{\nu}}(\boldsymbol{p}) - g(\boldsymbol{p})\frac{\partial f}{\partial \boldsymbol{\nu}}(\boldsymbol{p}) \right) |ds| + \int_D g\Delta f\, |dxdy|,$$

where we recall that, for $\boldsymbol{p} \in \partial D$, we have

$$\frac{\partial g}{\partial \boldsymbol{\nu}}(\boldsymbol{p}) := \langle \nabla g(\boldsymbol{p}), \boldsymbol{\nu}(\boldsymbol{p})\rangle.$$

Hint. Use Exercise 16.3 and the flux-divergence formula (16.1.11). $\square$

Exercise 16.5. Consider the function $K : \mathbb{R}^2 \to \mathbb{R}$,

$$K(x,y) = \begin{cases} \frac{1}{2\pi}\ln r, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0), \;\; r = \sqrt{x^2+y^2}. \end{cases}$$

Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is a C^2-function with *compact support*.

(i) Show that $\Delta K = 0$ on $\mathbb{R}^2\backslash\{\mathbf{0}\}$.

(ii) Show that the integral $K\Delta f$ is absolutely integrable on $\mathbb{R}^2\backslash\{0\}$

(iii) Show that

$$\int_{\mathbb{R}^2\backslash\{0\}} K\Delta f\, |dxdy| = f(0).$$

Hint. (ii) Have a look back at Exercise 15.31. (iii) Fix $R > 0$ sufficiently large so that the support of f is contained in the disk $D_R = \{r < R\}$. For $\varepsilon > 0$ small we consider the disk $D_\varepsilon = \{r < \varepsilon\}$ and the annulus $A_{\varepsilon,R} := \{\varepsilon < r < R\}$. Use Exercise 16.4 to show that

$$\int_{A_{\varepsilon,R}} K\Delta f\, |dxdy| = \int_{\partial D_\varepsilon} \left(f\frac{\partial K}{\partial \boldsymbol{\nu}} - K\frac{\partial f}{\partial \boldsymbol{\nu}} \right) |ds|,$$

and then prove that

$$\lim_{\varepsilon \searrow 0} \int_{\partial D_\varepsilon} K\frac{\partial f}{\partial \boldsymbol{\nu}}\, ds = 0, \;\; \lim_{\varepsilon \searrow 0} \int_{\partial D_\varepsilon} f\frac{\partial K}{\partial \boldsymbol{\nu}}\, |ds| = f(0).$$

$\square$

Exercise 16.6. Suppose that $\beta, \tau : [a,b]$ are C^1-functions such that

$$\beta(x) < \tau(x), \quad \forall x \in (a,b).$$

Consider the simple type domain

$$D(\beta,\tau) = \big\{ (x,y) \in \mathbb{R}^2; \;\; x \in (a,b), \;\; \beta(x) < y < \tau(x) \big\} \tag{16.4.1}$$

Prove Stokes' formula (16.1.14) when the piecewise C^1 domain is $U = D(\beta,\tau)$.

Hint. To compute $\int_D P'_y |dxdy|$ use Fubini's Theorem 15.2.3. To compute $\int_D Q'_x |dxdy|$ use the change of variables

$$x = u, \;\; y = \beta(u) + (\tau(u) - \beta(u))v, \;\; u \in [a,b], \;\; v \in [0,1],$$

and the chain rule

$$\frac{\partial}{\partial x} = \frac{\partial u}{\partial x}\frac{\partial}{\partial u} + \frac{\partial v}{\partial x}\frac{\partial}{\partial v} = \partial_u - \frac{\beta'(u) + \big(\tau'(u) - \beta'(u)\big)v}{\tau(u) - \beta(u)}\partial_v.$$

(You have to justify the second equality above.) We set

$$f(u,v) := Q\big(x(u,v), y(u,v) \big).$$

Show that

$$\int_D \frac{\partial Q}{\partial x} |dxdy| = \int_{\substack{a \leqslant u \leqslant b \\ 0 \leqslant v \leqslant 1}} \underbrace{\left(\partial_u f - \frac{\beta'(u) + \big(\tau'(u) - \beta'(u)\big)v}{\tau(u) - \beta(u)} \partial_v f \right) \big(\tau(u) - \beta(u) \big)}_{=:g(u,v)} |dudv|.$$

On the other hand, if we write Qdy in (u,v) coordinates we get

$$Qdy = f(u,v)y'_u du + f(u,v)y'_v dv = \underbrace{f(u,v)\big(\beta'(u) + v(\tau'(u) - \beta'(u))\big)}_{=:A(u,v)} du + \underbrace{f(u,v)\tau(u)}_{=:B(u,v)} dv.$$

Show that

$$g(u,v) = \frac{\partial B}{\partial u} - \frac{\partial A}{\partial v}$$

and then compute

$$\int_{\substack{a \leqslant u \leqslant b \\ 0 \leqslant v \leqslant 1}} \left(\frac{\partial B}{\partial u} - \frac{\partial A}{\partial v} \right) |dudv|$$

using Fubini. □

Exercise 16.7. Suppose that $D_1, D_2 \subset \mathbb{R}^2$ are two bounded piecewise C^1 domains that intersect only along portions of their boundaries and $\partial D_1 \cap \partial D_2$ is a piecewise C^1 connected curve. Set $D := D_1 \cup D_2$.

(i) Show that D is also piecewise C^1.

(ii) Assume that $\mathcal{O} \subset \mathbb{R}^2$ is an open set containing the closure of D and $\boldsymbol{F} : \mathcal{O} \to \mathbb{R}^2$ is a continuous vector field, $\boldsymbol{F}(x,y) = \big(P(x,y), Q(x,y) \big)$. Show that

$$\int_{\partial_+^* D} Pdx + Qdy = \int_{\partial_+^* D_1} Pdx + Qdy + \int_{\partial_+^* D_2} Pdx + Qdy.$$

(iii) Conclude that if Stokes' formula (16.1.14) holds for D_1 and D_2, then it also holds for $D = D_1 \cup D_2$.

□

Exercise 16.8. Suppose that $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ are two linearly independent vectors. Prove that

$$\|\boldsymbol{u} \times \boldsymbol{v}\| = \operatorname{area}\big(\, P(\boldsymbol{u}, \boldsymbol{v})\,\big),$$

where the cross product "×" is defined by (11.2.6) and $\operatorname{area}\big(\, P(\boldsymbol{u}, \boldsymbol{v})\,\big)$ is defined by (16.2.4). □

Exercise 16.9. Fix $a, b, r, R \in \mathbb{R}$, $a, R > r > 0$. Consider the map $\Phi : \mathbb{R}^2 \backslash \{\boldsymbol{0}\} \to \mathbb{R}^3$

$$\Phi(x, y) = \begin{bmatrix} \frac{R}{r}x \\ \\ \frac{R}{r}y \\ \\ ar + b \end{bmatrix}, \;\; \text{where} \;\; r = r(x, y) = \sqrt{x^2 + y^2}.$$

Denote by D the annulus

$$D = \big\{\, (x, y) \in \mathbb{R}^2; \;\; 1 < \sqrt{x^2 + y^2} < 2 \,\big\}.$$

(i) Show that Φ satisfies all the conditions (i)-(iii) in Proposition 14.5.4.

(ii) Show that the image S of Φ is a cylinder of radius R with the z-axis as symmetry axis.

(iii) Describe the area element on S in terms of the coordinates x, y induced by Φ.

(iv) Set $\Sigma := \Phi\big(\, \boldsymbol{cl}(D)\,\big)$. Show that Σ is a convenient surface with boundary and Φ defines a parametrization of Σ.

(v) Denote by f the restriction to Σ of the function $f(x, y, z) = z$. Compute

$$\int_\Sigma f(\boldsymbol{p})\, |dA(\boldsymbol{p})|.$$

□

Exercise 16.10. Suppose that $f, g : (0, 1) \to \mathbb{R}$ are C^1 functions such that $f(x) < g(x)$, $\forall x \in (0, 1)$. Let $U \subset \mathbb{R}^2$ be the region $(0, 1) \times [0, 1]$. Construct a diffeomorphism

$$\Phi : (0, 1) \times \mathbb{R} \to \mathbb{R}^2$$

such that $\Phi(U)$ is the region.

$$D = \big\{\, (x, y) \in \mathbb{R}^2; \;\; 0 < x < 1, \;\; f(x) \leqslant y \leqslant g(x) \,\big\}.$$

Conclude that the region D is a surface with boundary in $\mathbb{R}^2$.

Hint. Think of vertically shearing U onto D. □

Exercise 16.11. Suppose that $S \subset \mathbb{R}^n$ is a *compact* surface, with or without boundary.

(i) Show $\operatorname{area}(S) < \infty$.

(ii) If $f : S \to \mathbb{R}$ is continuous and $f(\boldsymbol{p}) \geqslant 0$, $\forall \boldsymbol{p} \in S$, then

$$\int_S f(\boldsymbol{p})\, |dA(\boldsymbol{p})| \geqslant 0.$$

(iii) Prove that if $L : \mathbb{R}^n \to \mathbb{R}^n$ is an orthogonal linear operator, i.e., $L^\top L = \mathbb{1}_n$, then

$$\operatorname{area}\big(\, L(S)\,\big) = \operatorname{area}(S).$$

(iv) If $f, g : S \to \mathbb{R}$ are continuous and $f(\boldsymbol{p}) \geqslant g(\boldsymbol{p})$, $\forall \boldsymbol{p} \in S$, then

$$\int_S f(\boldsymbol{p})\,|dA(\boldsymbol{p})| \geqslant \int_S g(\boldsymbol{p})\,|dA(\boldsymbol{p})|.$$

(v) If $f : S \to \mathbb{R}$ is continuous, $C > 0$ and $|f(\boldsymbol{p})| \leqslant C$, $\forall \boldsymbol{p} \in S$, then

$$\left| \int_S f(\boldsymbol{p})\,|dA(\boldsymbol{p})| \right| \leqslant C \operatorname{area}(S).$$

□

Exercise 16.12. For each $r > 0$ we denote by S_r the sphere of radius r in $\mathbb{R}^3$ centered at the origin. i.e.,

$$S_r := \big\{\, (x, y, z) \in \mathbb{R}^3;\;\; x^2 + y^2 + z^2 = r^2 \,\big\}.$$

(i) Show that $\operatorname{area}(S_r) = 4\pi r^2$.

(ii) Prove that if $f : \mathbb{R}^3 \to \mathbb{R}$ is a continuous function, then

$$\lim_{r \to 0} \frac{1}{4\pi r^2} \int_{S_r} f(\boldsymbol{p})\,|dA(\boldsymbol{p})| = f(\boldsymbol{0}).$$

□

Exercise 16.13. Let S denote the unit sphere in $\mathbb{R}^3$

$$S = \big\{\, (x, y, z) \in \mathbb{R}^3;\;\; x^2 + y^2 + z^2 = 1 \,\big\}.$$

Denote by $\boldsymbol{N}$ the North Pole, i.e., the point on S with coordinates $(0, 0, 1)$. The *stereographic projection* is the map

$$\boldsymbol{F} : S \backslash \{\boldsymbol{N}\} \to \mathbb{R}^2 \times \boldsymbol{0} = \{(x, y, z) \in \mathbb{R}^3 : \;\; z = 0\,\}$$

$$\boldsymbol{F}(\boldsymbol{p}) = \text{intersection of the line } \boldsymbol{Np} \text{ with the plane } \mathbb{R}^2 \times \boldsymbol{0}.$$

(i) For $\boldsymbol{p} \in S\backslash\{\boldsymbol{N}\}$ compute the coordinates (u, v) of $\boldsymbol{F}(\boldsymbol{p})$ in terms of the coordinates (x, y, z) of $\boldsymbol{p}$ and conversely, compute the coordinates (x, y, z) of $\boldsymbol{p}$ in terms of the coordinates (u, v) of $\boldsymbol{F}(\boldsymbol{p})$.

(ii) Prove that $\boldsymbol{F}$ is a homeomorphism and the inverse map $\Phi = \boldsymbol{F}^{-1} : \mathbb{R}^2 \to S\backslash\{\boldsymbol{N}\}$ is an immersion so Φ is a parametrization of $S\backslash\{\boldsymbol{N}\}$.

(iii) Describe the area element dA on $S\backslash\{\boldsymbol{N}\}$ in terms of the coordinates (u, v) defined by Φ.

□

Exercise 16.14. Suppose that $\mathcal{O} \subset \mathbb{R}^3$ is an open set, $f : \mathcal{O} \to \mathbb{R}$ is a C^2-function and $\boldsymbol{F} : \mathcal{O} \to \mathbb{R}^3$ is a C^2-vector field. Compute

$$\operatorname{curl}\big(\nabla f\big), \;\; \operatorname{div}\big(\nabla f\,\big),$$

$$\operatorname{curl}\big(\operatorname{curl}\boldsymbol{F}\big),\ \ \operatorname{div}\big(\operatorname{curl}\boldsymbol{F}\big),\ \ \nabla\big(\operatorname{div}\boldsymbol{F}\big).$$ □

Exercise 16.15. Let $D\subset\mathbb{R}^3$ be a bounded domain with C^1-boundary, U an open set containing $\boldsymbol{cl}\,D$ and $f,g:U\to\mathbb{R}$ are C^2 function. Denote by $\boldsymbol{\nu}$ the outer normal vector field along ∂D. Prove that

$$\int_D f\Delta g\,|dxdydz| = \int_{\partial D} f(\boldsymbol{p})\frac{\partial g}{\partial\boldsymbol{\nu}}(\boldsymbol{p})|dA| - \int_D\langle\nabla f,\nabla g\rangle\,|dxdydz|,$$

$$\int_{\partial D}\frac{\partial g}{\partial\boldsymbol{\nu}}\,|dA| = \int_D\Delta g\,|dxdydz|,$$

$$\int_D f\Delta g|dxdydz| = \int_{\partial D}\left(f(\boldsymbol{p})\frac{\partial g}{\partial\boldsymbol{\nu}}(\boldsymbol{p}) - g(\boldsymbol{p})\frac{\partial f}{\partial\boldsymbol{\nu}}(\boldsymbol{p})\right)|dA| + \int_D g\Delta f\,|dxdydz|,$$

where we recall that, for $\boldsymbol{p}\in\partial D$, we have

$$\frac{\partial g}{\partial\boldsymbol{\nu}}(\boldsymbol{p}) := \langle\nabla g(\boldsymbol{p}),\boldsymbol{\nu}(\boldsymbol{p})\rangle.$$

Hint. Use Exercise 16.3 and the flux-divergence formula (16.2.21). □

Exercise 16.16. Consider the function $K:\mathbb{R}^3\to\mathbb{R}$,

$$K(x,y,z) = \begin{cases}\frac{1}{4\pi\rho} & (x,y,z)\neq(0,0,0),\\ 0, & (x,y,z)=(0,0,0),\quad \rho=\sqrt{x^2+y^2+z^2}.\end{cases}$$

Suppose that $f:\mathbb{R}^3\to\mathbb{R}$ is a C^2-function with *compact support*.

(i) Show that $\Delta K=0$ on $\mathbb{R}^3\backslash\{\boldsymbol{0}\}$.

(ii) Show that the integral $K\Delta f$ is absolutely integrable on $\mathbb{R}^3\backslash\{0\}$.

(iii) Show that

$$\int_{\mathbb{R}^3\backslash\{0\}} K\Delta f\,|dxdydz| = -f(0).$$

Hint. (ii) Have a look back at Example 15.4.15. (iii) Fix $R>0$ sufficiently large so that the support of f is contained in the ball $B_R=\{\rho<R\}$. For $\varepsilon>0$ small we consider the ball $B_\varepsilon=\{\rho<\varepsilon\}$ and the annulus $A_{\varepsilon,R}:=\{\varepsilon<\rho<R\}$. Use Exercise 16.15 to show that

$$\int_{A_{\varepsilon,R}} K\Delta f\,|dxdydz| = -\int_{\partial B_\varepsilon}\left(f\frac{\partial K}{\partial\boldsymbol{\nu}} - K\frac{\partial f}{\partial\boldsymbol{\nu}}\right)|dA|,$$

and then prove that

$$\lim_{\varepsilon\searrow 0}\int_{\partial B_\varepsilon} K\frac{\partial f}{\partial\boldsymbol{\nu}}\,|dA| = 0,\ \ \lim_{\varepsilon\searrow 0}\int_{\partial B_\varepsilon} f\frac{\partial K}{\partial\boldsymbol{\nu}}\,|dA| = f(0).$$

□

Exercise 16.17. Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is a C^1-function with compact support. Denote by H^- the half-space

$$H^- := \big\{(x^1,\dots,x^n)\in\mathbb{R}^n;\ \ x^1\leqslant 0\big\}.$$

Prove that

$$\int_{H^-}\frac{\partial f}{\partial x^1}(\boldsymbol{x})|dx^1\cdots dx^n| = \int_{\mathbb{R}^{n-1}} f(0,x^2,\dots x^n)\,|dx^2\cdots dx^n|,$$

and

$$\int_{H^-} \frac{\partial f}{\partial x^k}(\boldsymbol{x})|dx^1 \cdots dx^n| = 0, \quad \forall k \geqslant 2.$$

Hint. Use Fubini. □

Exercise 16.18. Let $m, n \in \mathbb{N}$, $m \leqslant n$. Consider

$$\omega_1, \dots, \omega_m \in \Omega^1(\mathbb{R}^n),$$

$$\omega_i = \sum_{j=1}^{n} \omega_{ij} dx^j, \quad i = 1, \dots, n.$$

Prove that

$$\omega_1 \wedge \cdots \wedge \omega_m = \sum_{J \in \mathrm{Inj}^+(m,n)} \det \underline{(}\omega_{ij_k}\big)_{1 \leqslant j,k \leqslant m} d\boldsymbol{x}^{\wedge J}.$$

In particular, if $m = n$

$$\omega_1 \wedge \cdots \wedge \omega_n = \Big(\det \big(\omega_{ij}\big)_{1 \leqslant i,j \leqslant n} \Big)\, dx^1 \wedge \cdots \wedge dx^n.$$

□

Exercise 16.19. Prove (16.3.6). □

Exercise 16.20. Prove Proposition 16.3.12.

Hint. For part (i) consider first the case when α is a monomial $\alpha = \alpha_I d\boldsymbol{v}^{\wedge I}$, $\alpha_I \in C^1(V)$, where $I \in \mathrm{Inj}(k,n)$. Start with the case $I = (1, 2, \dots, k)$. □

Chapter 17

Analysis on metric spaces

At the dawn of the twentieth century, as more and more examples appeared on the mathematical scene, mathematicians realized that many of the results of analysis on $\mathbb{R}^n$ extend to more general situations with remarkable consequences. One important difference was the apparently unavoidable need to deal with infinite dimensional vector spaces such as the space of continuous functions on a nontrivial interval. When dealing with infinite dimensions we need to pay attention to foundational issues more carefully than we have done to date.

17.1. Metric spaces

A key concept that allowed the transition to infinite dimensions is the concept of *metric space* introduced by Maurice Fréchet in 1906.

17.1.1. Definition and examples. Loosely speaking, a metric space is a set in which there is a way of measuring how far apart are pairs of points.

Definition 17.1.1. A *metric space* is a pair (X, d), where X is a set and d is a *metric* or *distance function* on X, i.e., a function

$$d : X \times X \to \mathbb{R}$$

satisfying the following conditions.

(i) $\forall x_0, x_1 \in X$, $d(x_0, x_1) \geqslant 0$.

(ii) $\forall x_0, x_1 \in X$, $d(x_0, x_1) = 0$ if and only if $x_0 = x_1$.

(iii) $\forall x_0, x_1 \in X$, $d(x_0, x_1) = d(x_1, x_0)$.

(iv) $\forall x_0, x_1, x_2 \in X$

$$d(x_0,x_2) \leqslant d(x_0,x_1) + d(x_1,x_2).$$

The last inequality above is commonly referred to as the *triangle inequality*. □

Example 17.1.2. (a) Proposition 11.3.4 shows that the Euclidean distance

$$\operatorname{dist} : \mathbb{R}^n \times \mathbb{R}^n, \;\; \operatorname{dist}(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{(x^1-y^1)^2+\cdots+(x^n-y^n)^2}$$

is a metric on $\mathbb{R}^n$. In other words, $(\mathbb{R}^n, \operatorname{dist})$ is a metric space. It usually referred to as the n-dimensional *Euclidean metric space*.

(b) Suppose that (X,d) is a metric space and $S \subset X$ is a nonempty subset. Then d_S, the restriction of d to $S \times S$, is a metric and the metric space (S,d_S) is said to be a *metric subspace* of X.

(c) Suppose that (X,d_X) and (Y,d_Y). Then the function

$$d_{X\times Y} : (X\times Y)\times(X\times Y) \to \mathbb{R},$$

$$d_{X\times Y}\big(\,(x_0,y_0),(x_1,y_1)\,\big) = d_X(x_0,x_1) + d_Y(y_0,y_1)$$

is a metric on the Cartesian product $X \times Y$. The resulting metric space $(X \times Y, d_{X\times Y})$ is called the *product* of the metric spaces (X,d_X), (Y,d_Y). Sometimes we will denote by $d_X \times d_Y$ this product metric. This construction extends in an obvious way to a Cartesian product of finitely many metric spaces.

(d) Suppose that (X,d) is a metric space and $c > 0$ is a positive constant. Define

$$\bar{d}_c : X\times X \to \mathbb{R}, \;\; \bar{d}_c(x_0,x_1) = \min\big(\,d(x_0,x_1), c\,\big).$$

Then $\bar{d}_c$ is also a metric on X.

(e) Suppose that X is an abstract set. Then the function

$$\delta : X\times X \to \mathbb{R}, \;\; \delta(x_0,x_1) = \begin{cases} 1, & x_0 \neq x_1, \\ 0, & x_0 = x_1. \end{cases}$$

defines a metric on X called the *discrete metric*.

(f) Suppose that S is a set and $n \in \mathbb{N}$. Define

$$d_H : S^n \times S^n \to [0,\infty), \;\; d_H\big(\,(s_1,\dots,s_n),(t_1,\dots,t_n)\,\big) = \#\big\{\,k;\;\; s_k \neq t_k\,\big\}.$$

Thus, the distance between the n-tuples $\underline{s} = (s_1,\dots,s_n)$ and $\underline{t} = (t_1,\dots,t_n)$ is equal to the number of distinct entries in identical positions. Equivalently, if δ_S denotes the discrete metric on S, then

$$d_H = \delta_{S^n} = \underbrace{\delta_S \times \cdots \times \delta_S}_{n}.$$

The metric d_H is called the *Hamming metric* on S^n. □

Definition 17.1.3. A *real normed space* is a pair $(X, \|-\|)$ where X is a real vector space and $\|-\|$ is a *norm* on X, i.e., a function

$$\|-\| : X \to \mathbb{R}$$

satisfying the following conditions.

(i) $\forall x \in X$, $\|x\| \geqslant 0$.

(ii) $\forall x \in X$, $\|x\| = 0$ if and only if $x = 0$.

(iii) $\forall x \in X$, $\forall t \in \mathbb{R}$, $\|tx\| = |t| \cdot \|x\|$.

(iv) $\forall x, y \in X$, $\|x + y\| \leqslant \|x\| + \|y\|$.

The last inequality is usually referred to as the *triangle inequality.*

A *complex normed space* is a pair $(X, \|-\|)$ where X is a complex vector space and $\|-\|$ is a *norm* on X, i.e., a function $\|-\| : X \to \mathbb{R}$ satisfying the conditions (i),(ii),(iv) above and

(iii)$_c$ $\forall x \in X$, $\forall t \in \mathbb{C}$, $\|tx\| = |t| \cdot \|x\|$.

□

The next result explains the close connection between normed spaces and metric spaces. Its proof is identical to the proof of Proposition 11.3.4 so we omit it.

Proposition 17.1.4. *Suppose that $(X, \|-\|)$ is a real or complex normed space. Define*

$$d = d_{\|-\|} : X \times X \to \mathbb{R}, \;\; d(x_0, x_1) = \|x_0 - x_1\|.$$

Then d is a metric on X called the metric induced by norm $\|-\|$. □

Example 17.1.5. (a) Suppose that S is a set. Denote by $\mathbb{B}(S)$ the vector space of bounded functions $f : S \to \mathbb{R}$, i.e., functions such that

$$\exists M > 0, \;\; \forall s \in S : |f(s)| < M.$$

Define

$$\|-\|_\infty \to \mathbb{R}, \;\; \|f\|_\infty = \sup_{s \in S} |\, f(s) \,|.$$

Then $\|-\|_\infty$ is a norm on $\mathbb{B}(S)$ called the *sup-norm.*

(b) Suppose that $K \subset \mathbb{R}^n$ is a compact set. As usual, we denote by $C(K)$ the vector space of continuous functions $K \to \mathbb{R}$. Any continuous function on K is bounded so $C(K) \subset \mathbb{B}(K)$. The sup-norm on $\mathbb{B}(K)$ induces a norm on $C(K)$ also called *sup-norm* and also denoted by $\|-\|_\infty$.

(c) For $p \in [1, \infty)$ and $n \in \mathbb{N}$ define

$$\|-\|_p : \mathbb{R}^n \to \mathbb{R}, \;\; \|\boldsymbol{x}\|_p = \Big(|x^1|^p + \cdots + |x^n|^p \Big)^{1/p}.$$

Clearly, $\|-\|_p$ satisfies the properties (i)-(iii) in the definition of norm. The triangle inequality is Minkowski's inequality (8.3.17).

(d) Denote by $\boldsymbol{\ell}_2$ the space of sequences $\underline{x} : \mathbb{N} \to \mathbb{R}$, $x_n := \underline{x}(n)$ such that

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

It becomes a normed space when equipped with the norm

$$\|-\| : \boldsymbol{\ell}_2 \to \mathbb{R},\;\; \|\underline{x}\| := \Big(\sum_{n\in\mathbb{N}} x_n^2\Big)^{1/2}.$$

(e) Suppose that B is a nondegenerate closed box in $\mathbb{R}^n$. For every $p \in [1,\infty)$ we define

$$\|-\|_p : C(B) \to \mathbb{R},\;\; \|f\|_p = \left(\int_B |f(\boldsymbol{x})|^p\,|d\boldsymbol{x}|\right)^{1/p}. \tag{17.1.1}$$

Also, we set

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in B} |f(\boldsymbol{x})|,\;\; \forall f \in C(B).$$

Clearly $\|-\|_\infty$ is a norm. Note that

$$f \in C(B) \;\text{ and }\; \|f\|_p = 0 \Rightarrow f(\boldsymbol{x}) = 0,\;\; \forall \boldsymbol{x} \in B.$$

Indeed, if $f(\boldsymbol{x}_0) \neq 0$ for some $\boldsymbol{x}_0 \in B$, then there exists a tiny closed cube C centered at $\boldsymbol{x}_0$ such that

$$|f(\boldsymbol{x})| > \frac{1}{2}|f(\boldsymbol{x}_0)|,\;\; \forall \boldsymbol{x} \in C \cap B.$$

Then

$$0 = \int_B |f(\boldsymbol{x})|^p\,|d\boldsymbol{x}| \geqslant \int_{C\cap B} |f(\boldsymbol{x})|^p\,|d\boldsymbol{x}| \geqslant \frac{|f(\boldsymbol{x}_0)|^p}{2^p}\operatorname{vol}(C\cap B) > 0.$$

Clearly $\|-\|_1$ satisfies the triangle inequality. We want to prove that the same is true for $\|-\|_p$, $p > 1$.

Fix $p \in (1,\infty)$ and set $q := \frac{p}{p-1}$ so that $\frac{1}{p} + \frac{1}{q} = 1$. We recall Hölder's inequality (15.5.1) in Exercise 15.5 which states that for any $u, v \in \mathcal{R}(B)$ we have

$$\int_B |u(\boldsymbol{x})v(\boldsymbol{x})|\,|d\boldsymbol{x}| \leqslant \|u\|_p \cdot \|v\|_q.$$

Let $f, g \in \mathcal{R}(B)$ and set

$$X := \|f\|_p,\;\; Y := \|g\|_p,\;\; Z := \|f+g\|_p.$$

We want to show that

$$Z \leqslant X + Y.$$

Set $h := |f+g|^{p-1}$. We have

$$Z^p = \int_B |f(\boldsymbol{x}) + g(\boldsymbol{x})|^p\,|d\boldsymbol{x}| = \int_B |f(\boldsymbol{x}) + g(\boldsymbol{x})| \cdot \underbrace{|f(\boldsymbol{x}) + g(\boldsymbol{x})|^{p-1}}_{h(\boldsymbol{x})}\,|d\boldsymbol{x}|$$

$$\leqslant \int_B |f(\boldsymbol{x})| \cdot |h(\boldsymbol{x})|\,|d\boldsymbol{x}| + \int_B |g(\boldsymbol{x})| \cdot |h(\boldsymbol{x})|\,|d\boldsymbol{x}|$$

(use Hölder's inequality (15.5.1))

$$\leqslant \|f\|_p \|h\|_q + \|g\|_p \cdot \|h\|_q.$$

Thus

$$Z^p \leqslant (X+Y)\|h\|_q.$$

Now observe that

$$\|h\|_q = \left(\int_B |h|^{\frac{p}{p-1}}\,|d\boldsymbol{x}|\right)^{\frac{p-1}{p}} = \left(\int_B |f(\boldsymbol{x}) + g(\boldsymbol{x})|^p\,|d\boldsymbol{x}|\right)^{\frac{p-1}{p}} = Z^{p-1}$$

so that

$$Z^p \leqslant (X+Y)Z^{p-1}.$$

Hence, $\forall p \in [1,\infty]$, $\|-\|_p$ is a norm on $\mathcal{R}(B)$. □

Definition 17.1.6. Suppose that (X, d_X) and (Y, d_Y) are metric spaces.

(i) A map $T : X \to Y$ is called an *isometry* if

$$d_Y(Tx_0, Tx_1) = d_X(x_0, x_1), \quad \forall x_0, x_1 \in X.$$

The metric spaces (X, d_X) and (Y, d_Y) are said to be *isometric* if there exists a bijective isometry $T : X \to Y$.

□

Note that if S is a subset of the metric space (X, d), then the natural inclusion is an isometry $(S, d_S) \to (X, d)$.

17.1.2. Basic geometric and topological concepts. A large part of the topological concepts for Euclidean spaces introduced in Sections 11.3 and 11.4 have a counterpart in the more general context of metric spaces.

Definition 17.1.7. Let (X, d) be a metric space.

(i) For $r > 0$ and $x_0 \in X$ we define the *open ball of center* x_0 *and radius* r to be the set

$$B_r(x_0) = B_r^X(x_0) := \left\{\, x \in X;\;\; d(x, x_0) < r \,\right\}.$$

(ii) A subset $U \subset X$ is called *open* if, for any $p \in U$, there exists $r > 0$ such that $B_r(p) \subset U$.

(iii) A *neighborhood* of $x_0 \in X$ is a set V that contains an open ball centered at x_0.

□

Propositions 11.3.7 and 11.3.8 have a metric space counterpart. The proofs are identical and are left to the reader.

Proposition 17.1.8. *Let (X,d) be a metric space. Then the following hold.*

(i) *For any $x_0 \in X$ and any $r > 0$ the open ball $B_r(x_0)$ is an open set.*

(ii) *The whole space X and the empty set are open.*

(iii) *The intersection of two open sets is an open set.*

(iv) *The union of any family of open sets is an open set.*

□

Definition 17.1.9. Let (X,d) be a metric space. A subset $C \subset X$ is called *closed* if its complement $X \backslash C$ is an open set. □

As in the Euclidean case we have the following consequence of Proposition 17.1.8.

Proposition 17.1.10. *Let (X,d) be a metric space. Then the following hold.*

(i) *The whole space X and the empty set are closed sets.*

(ii) *The union of two closed sets is a closed set.*

(iii) *The intersection of any family of closed sets is a closed set.*

□

Remark 17.1.11 (A word of caution)**.** Suppose that (X,d) is metric space and $S \subset X$ is a subset that *is not* open. The metric d induces by restriction a metric d_S on S, Example 17.1.2(b). However, the set S *is an open subset* of the metric space (S, d_S).

For example, the compact interval $[0,1]$ is not an open subset of $(\mathbb{R}, |-|)$ but it is an open set of the metric subspace $[0,1] \subset \mathbb{R}$. □

The proof of the following result is left to the reader as an exercise.

Proposition 17.1.12. *Suppose (X,d) is a metric space and $Y \subset X$. Let $S \subset Y$. The following statements are equivalent.*

(i) *The set S is open (respectively closed) in the metric subspace (Y, d_Y); see Example 17.1.2(b).*

(ii) *There exists an open (respectively closed) subset $U \subset X$ such that $S = U \cap Y$.*

□

Definition 17.1.13. Suppose that (X,d) is a metric space and $S \subset X$.

(i) The *closure of* S in X, denoted by $\boldsymbol{cl}(S)$ or $\boldsymbol{cl}_X(S)$, is the intersection of all the closed subsets of X that contain S.

(ii) The *interior of* S in X, denoted by $\boldsymbol{int}(S)$ or $\boldsymbol{int}_X(S)$, is the union of all the open subset of X contained in S.

(iii) The *boundary of* S, denoted ∂S, is the difference $\boldsymbol{cl}(S)\setminus\boldsymbol{int}(S)$.

□

The metric space setup affords a very useful notion of convergence.

Definition 17.1.14. Let (X,d) be a metric space. We say that a sequence $(x_n)_{n\in\mathbb{N}}$ of points in X is *convergent* if there exists a point $x_*\in X$ such that

$$\lim_{n\to\infty} d(x_n,x_*)=0.$$

In this case we say that $(x_n)_{n\in\mathbb{N}}$ converges to x_*. □

As in the real case, if (x_n) converges to x_* and to x_{**}, then $x_*=x_{**}$. Thus any convergent sequence converges to a single point called *the limit* of the convergent sequence and denoted by

$$\lim_{n\to\infty} x_n \;\text{ or }\; \lim_n x_n.$$

Thus

$$\begin{aligned}\lim_n x_n=x_* &\Longleftrightarrow \forall\varepsilon>0,\ \exists N=N(\varepsilon)>0\ \ \forall n\geqslant N(\varepsilon):\ \ d(x_n,x_*)<\varepsilon\\ &\Longleftrightarrow \forall\varepsilon>0,\ \exists N=N(\varepsilon)>0\ \ \forall n\geqslant N(\varepsilon):\ \ x_n\in B_\varepsilon(x_*).\end{aligned}$$

Example 17.1.15. Let S be a set and consider the normed space $\big(\mathbb{B}(S),\|-\|_\infty\big)$ of bounded functions $S\to\mathbb{R}$. Note that

$$\lim_{n\to\infty}\|f_n-f\|_\infty=0$$

if and only if

$$\forall\varepsilon>0\ \ \exists N=N(\varepsilon)>0\ \ \forall n>N(\varepsilon):\ \ \sup_{s\in S}\big|f_n(s)-f(s)\big|<\varepsilon$$

This means that the sequence of functions $f_n:S\to\mathbb{R}$ converges *uniformly* to the function f (Definition 6.1.9(b)) if and only if $\|f_n-f\|_\infty\to 0$ as $n\to\infty$. □

The following result is an immediate generalization of Proposition 11.4.11, with identical proof.

Proposition 17.1.16. *Let (X,d) be a metric space and $C\subset X$. Then the following are equivalent.*

(i) *The set C is closed.*

(ii) *For any convergent sequence of points in C, its limit is also a point in C.*

□

The above result leads to a very useful characterization of the closure of a subset of a metric space.

Proposition 17.1.17. *Let (X, d) be a metric space, $S \subset X$ and $x \in X$. Then the following statements are equivalent.*

(i) *There exists a sequence $(s_n)_{n \in \mathbb{N}}$ of points in S such that*

$$\lim_{n \to \infty} s_n = x.$$

(ii) $x \in \boldsymbol{cl}(S)$.

Proof. (i) ⇒ (ii) Since $S \subset \boldsymbol{cl}(S)$ we deduce that (s_n) is also a sequence of points in the closed set $\boldsymbol{cl}(S)$. Proposition 17.1.16 implies that the limit x is also a point in $\boldsymbol{cl}(S)$.

(ii) ⇒ (i) Given $x \in \boldsymbol{cl}(S)$ we have to construct a sequence (s_n) of points in S such that

$$\lim_n s_n = x.$$

If $x \in S$, then the constant sequence $s_n = x$, $\forall n$, converges to X. Suppose that $x \in \boldsymbol{cl}(S) \setminus S$. We claim that for any $n \in \mathbb{N}$ the ball of radius $1/n$ centered at x intersects S. Indeed, if that was not the case then, for some n, $B_{1/n}(x) \cap S = \emptyset$ so that

$$S \subset X \setminus B_{1/n}(x).$$

The set $X \setminus B_{1/n}(x)$ is closed since $B_{1/n}(x)$ is open. We have reached a contradiction because x belongs to any closed set containing S and, in particular $x \in X \setminus B_{1/n}(x)$.

The above claim implies that for any $n \in \mathbb{N}$ there exists $s_n \in S$ such that $d(s_n, x) < \frac{1}{n}$ so that

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

□

Definition 17.1.18. Let (X, d) be a metric space. A subset $S \subset X$ is called *dense* (in X) if $\boldsymbol{cl}(S) = X$. □

Corollary 17.1.19. *Let (X, d) be a metric space and $S \subset X$. The following are equivalent.*

(i) *The set S is dense in X.*

(ii) *For any $x \in X$ there exists a sequence of points in S converging to x.*

(iii) *For any nonempty open set $U \subset X$ the set S intersects U nontrivially, $S \cap U \neq \emptyset$.*

Proof. (i) $\iff$ (ii) follows from Proposition 17.1.17.

(i) ⇒ (iii) We argue by contradiction. Suppose U is a nonempty set that does not intersect S. In other words, S is contained in the complement of U, $S \subset U^c$. Since U^c is closed we deduce

$$X = \boldsymbol{cl}(S) \subset U^c \Rightarrow U = \emptyset.$$

(iii) ⇒ (i) We argue again by contradiction. Suppose $\boldsymbol{cl}(S) \neq X$. Thus the open set $U = X \setminus \boldsymbol{cl}(S)$ is nonempty and disjoint from S. This contradicts (iii). □

Definition 17.1.20. A metric space is called *separable* if it admits a dense, *countable* subset. □

For example, the space $\mathbb{R}$ with the Euclidean metric is separable because the set of rational numbers is dense in $\mathbb{R}$.

Definition 17.1.21. A *topology* on a set X is a collection $\mathcal{T}$ of subsets of X satisfying the following properties.

(i) $\varnothing, X \in \mathcal{T}$.

(ii) $\forall U, V\colon\ U, V \in \mathcal{T} \implies U \cap V \in \mathcal{T}$.

(iii) The union of *any* family of subsets in $\mathcal{T}$ is also a subset in $\mathcal{T}$.

The sets in $\mathcal{T}$ are called the *open subsets* of the given topology. A *topological space* is a pair $(X, \mathcal{T})$ where X is a set and $\mathcal{T}$ is a topology on X. □

We have seen that a metric d on a set defines a topology on X called the *metric topology*. A topological space is called *metrizable* if its topology is defined by some metric. Let us point out that different metrics can induce the same topology. For example if d is a metric on X, then the new metric $\bar{d}$ defined by $\bar{d}(x_0, x_1) = \min\big(\, d(x_0, x_1), 1\,\big)$ induces the same topology.

A norm on a vector space defines a metric and in turn, this metric determines a topology called *topology induced by the norm.* The topology defined by the Euclidean norm on $\mathbb{R}^n$ is called *the Euclidean topology* of $\mathbb{R}^n$.

An object or a concept is called *topological* if it can described only in terms of the open sets of a given topology. For example, the concept of closed set or closure of a set are topological concepts. Indeed a closed set is a set whose complement is open. Exercise 17.6 asks you to prove that the concept of convergence of a sequence in a metric space is in fact a topological concept.

Example 17.1.22. (a) For any set X the collection 2^X of all its subsets is a topology on X. It is called the *discrete topology.* It coincides with the topology induced by the discrete metric.

(b) If $(\mathcal{T}_i)_{i \in I}$ is a collection of topologies on a set X, then their intersection

$$\bigcap_{i \in I} \mathcal{T}_i \subset 2^X$$

is another topology on X.

(c) Suppose that $\mathcal{S} = (S_i)_{i \in I}$ is a collection of subsets of set X. Note that the discrete topology 2^X contains the collection $\mathcal{S}$. The intersection of all the topologies that contain $\mathcal{S}$ is a topology on X denoted by $\mathcal{T}[S]$ and called the topology generated by the collection $\mathcal{S}$. □

We will not be using topological spaces in the sequel so we will not delve too long into this topic. However, the curious reader can consult [27, Chap. X] for a very efficient introduction to *point set topology*, as this subject came to be known.

17.1.3. Continuity. The notion of continuity of maps also has an obvious metric space counterpart.

Definition 17.1.23. Suppose that (X, d_X) and (Y, d_Y) are metric spaces and $F : X \to Y$ is a map.

(i) The map F is said to be continuous at the point $x_0 \in X$ if

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 : \;\; \forall x \in X \;\; d_X(x, x_0) < \delta \Rightarrow d_Y\big(\, F(x), F(x_0)\,\big) < \varepsilon. \tag{17.1.2}$$

(ii) The map F is said to be continuous if it is continuous at every point $x \in X$.

(iii) We denote by $C(X, Y)$ the space of continuous maps $X \to Y$. When Y is the real axis $\mathbb{R}$ with the Euclidean metric $\operatorname{dist}(x, y) = |x - y|$ we use the simpler notation $C(X) := C(X, \mathbb{R})$.

□

Set $y_0 = F(x_0)$. Note that F is continuous at x_0 if

$$\forall \varepsilon > 0 \;\; \exists \delta = \delta(\varepsilon) > 0 : \;\; F\big(\, B^X_\delta(x_0)\,\big) \subset B^Y_\varepsilon(y_0). \tag{17.1.3}$$

Proposition 17.1.24. *Suppose that (X, d_X) and (Y, d_Y) are metric spaces and $F : X \to Y$ is a map. Let $x_0 \in X$. Then the following statements are equivalent.*

(i) *The map F is continuous at x_0.*

(ii) *For any sequence $(x_n)_{n\in\mathbb{N}}$ in X that converges to x_0, the sequence $\big(\, F(x_n)\,\big)_{n\in\mathbb{N}}$ converges to $F(x_0)$.*

Proof. (i) ⇒ (ii) Suppose that $(x_n)_{n\in\mathbb{N}}$ is a sequence in X that converges to x_0. For $\varepsilon > 0$ choose $\delta(\varepsilon) > 0$ as in (17.1.2). Then there exists $N = N(\delta(\varepsilon)) = N(\varepsilon) > 0$ such that

$$\forall n > N(\varepsilon), \;\; d_X(x_n, x_0) < \delta(\varepsilon).$$

From (17.1.2) we conclude that

$$\forall n > N(\varepsilon), \;\; d_Y\big(\, F(x_n), F(x_0)\,\big) < \varepsilon$$

so the sequence $F(x_n)$ converges to $F(x_0)$.

(ii) ⇒ (i) We argue by contradiction. Thus

$$\exists \varepsilon_0 > 0, \;\; \forall \delta > 0 \;\; \exists x(\delta) \in X \text{ such that } \; d_X(x(\delta), x_0) < \delta \text{ and } d_Y\big(\, F(x(\delta)), F(x_0)\,\big) > \varepsilon_0.$$

Set $x_n := x(1/n)$. Then $x_n \to x_0$ and $d_Y\big(\, F(x_n), F(x_0)\,\big) > \varepsilon_0$, $\forall n$. Thus the sequence $\big(\, F(x_0)\,\big)_{n\in\mathbb{N}}$ does not converge to $F(x_0)$ contradicting the assumption (ii). □

Definition 17.1.25. Suppose that (X, d_X) and (Y, d_Y) are metric spaces and $F : X \to Y$ a map. Given $K > 0$ we say that F is K-*Lipschitz* if

$$\forall x_0, x_1 \in X \;\; d_Y\big(\, F(x_0), F(x_1)\,\big) \leqslant K d_X(x_0, x_1). \tag{17.1.4}$$

The map is called *Lipschitz* if it is K-Lipschitz for some $K > 0$. We denote by $\mathrm{Lip}(X, Y)$ the space of Lipschitz continuous maps $X \to Y$. □

Corollary 17.1.26. *Suppose that (X, d_X) and (Y, d_Y) are metric spaces. Then any Lipschitz map $X \to Y$ is continuous, i.e.,*

$$\mathrm{Lip}(X, Y) \subset C(X, Y).$$

Proof. Let F be K-Lipschitz map, satisfying (17.1.4). If (x_n) is a sequence in X converging to x. From the inequality

$$d_Y\big(\, F(x_n), F(x)\,\big) \leqslant K d_X(x_n, x), \;\; \forall n.$$

Hence

$$\lim_{n\to\infty} d_Y\big(\, F(x_n), F(x)\,\big) = 0.$$

□

Proposition 17.1.27. *Suppose that (X, d) is a metric space and S is a nonempty subset of X. For every $x \in X$ we set*

$$\mathrm{dist}(x, S) := \inf_{s\in S} d(x, s).$$

(i) *The function $f : X \to \mathbb{R}$, $f(x) = \mathrm{dist}(x, S)$ is 1-Lipschitz, i.e., for any $x, y \in$ we have*

$$|f(x) - f(y)| \leqslant d(x, y).$$

(ii) $f^{-1}(0) = \boldsymbol{cl}\, S$. *In particular, if S is closed, then f is a nonnegative continuous function vanishing precisely on S.*

Proof. (i) Using the triangle inequality we deduce

$$f(x) = \mathrm{dist}(x, S) \leqslant d(x, s) \leqslant d(x, y) + d(x, s).$$

Choosing a sequence $(s_n)_{n\in\mathbb{N}}$ in S such that

$$\lim_{n\to\infty} d(x, s_n) = \mathrm{dist}(x, S)$$

we deduce

$$f(x) \leqslant d(x, y) + d(y, s_n) \;\Rightarrow\; f(x) \leqslant d(x, y) + \lim_{n\to\infty} d(y, s_n) = d(x, y) + f(y).$$

Hence, for any $x, y \in X$ we have

$$f(x) - f(y) \leqslant d(x, y).$$

Reversing the roles of x, y we deduce

$$-\big(\, f(x) - f(y)\,\big) = f(y) - f(x) \leqslant d(y, x) = d(x, y).$$

Hence

$$\bigl|\, f(x) - f(y) \,\bigr| \leqslant d(x, y).$$

(ii) We have to show that $f^{-1}(0) \subset \boldsymbol{cl}\, S$ and $\boldsymbol{cl}\, S \subset f^{-1}(0)$.

Suppose that $x \in f^{-1}(0)$, i.e., $\operatorname{dist}(x, S) = 0$. Thus, there exists a sequence (s_n) in S such that

$$\lim_{n\to\infty} d(x, s_n) = \operatorname{dist}(x, S) = 0.$$

Hence

$$\lim_{n\to\infty} s_n = x,$$

and we deduce from Proposition 17.1.17 that $x \in \boldsymbol{cl}\, S$.

Conversely, suppose that $x \in \boldsymbol{cl}\, S$. Proposition 17.1.17 implies that there exists a sequence (s_n) in S such that $s_n \to x$. Clearly $f(s_n) = \operatorname{dist}(s_n, S) = 0$, $\forall n$. Using Proposition 17.1.24 we deduce that

$$f(x) = \lim_{n\to\infty} f(s_n) = 0,$$

i.e., $x \in f^{-1}(0)$. □

Corollary 17.1.28. *Suppose that (X, d) is a metric space and $x_* \in X$, then the function*

$$f : X \to \mathbb{R}, \;\; f(x) = d(x, x_*)$$

is 1-Lipschitz and thus, continuous. In particular, if $(X, \|-\|)$ is a normed space then the norm function

$$\|-\| : X \to \mathbb{R}, \;\; x \mapsto \|x\| = d_{\|-\|}(x, 0)$$

is 1-Lipschitz. □

Corollary 17.1.29. *Suppose that (X, d) is a metric space and $S \subset X$. Denote by d_S the induced metric on S. Then the natural inclusion*

$$i_S : S \to X, \;\; i_S(s) = s,$$

is continuous.

Proof. Note that

$$d\bigl(\, i_S(s_1), i_S(s_2)\,\bigr) = d(s_1, s_2) = d_S(s_1, s_2)$$

so i_S is Lipschitz. □

Corollary 17.1.30. *Suppose that (X, d) is a metric space. Denote by $\hat{d}$ the product metric on $X \times X$,*

$$\hat{d}\bigl(\, (x_0, y_0), (x_1, y_1)\,\bigr) = d(x_0, x_1) + d(y_0, y_1), \;\; \forall (x_0, y_0), (x_1, y_1) \in X \times X.$$

Then the metric map $d : X \times X \to \mathbb{R}$ is 1-Lipschitz with respect to $\hat{d}$. In particular, it is continuous.

Proof. Let $(x_0, y_0), (x_1, y_1) \in X \times X$. Then

$$d(x_0, y_0) \leqslant d(x_0, x_1) + d(x_1, y_0), \;\; d(x_1, y_1) \geqslant d(x_1, y_0) - d(y_0, y_1)$$

so that

$$d(x_0, y_0) - d(x_1, y_1) \leqslant d(x_0, x_1) + d(x_1, y_0) - d(x_1, y_0) + d(y_0, y_1) = \hat{d}\big(\,(x_0, y_0), (x_1, y_1)\,\big).$$

Reversing the roles of (x_0, y_0) and (x_1, y_1) in the above argument we deduce

$$d(x_1, y_1) - d(x_0, y_0) \leqslant \hat{d}\big(\,(x_0, y_0), (x_1, y_1)\,\big).$$

Hence,

$$\big|\, d(x_0, y_0) - d(x_1, y_1) \,\big| \leqslant \hat{d}\big(\,(x_0, y_0), (x_1, y_1)\,\big). \tag{17.1.5}$$

□

Proposition 17.1.31. *If (X, d) is a metric space, then the space $C(X)$ of continuous real valued functions on X is an $\mathbb{R}$-algebra of functions, i.e.,*

(i) *$C(X)$ is a real vector space.*

(ii) *For any $f, g \in C(X)$ their product $f \cdot g$ is also a continuous function.*

Proof. Let $f, g \in C(X)$ and $t \in \mathbb{R}$. Suppose that (x_n) is a sequence in X converging to x Since f, g are continuous we have

$$\lim_n f(x_n) = f(x), \;\; \lim_n g(x_n) = g(x)$$

so that

$$\lim_n \big(\, f(x_n) + g(x_n)\,\big) = f(x) + g(x), \;\; \lim_n \big(\, tf(x_n)\,\big) = tf(x),$$
$$\lim_n \big(\, f(x_n) g(x_n)\,\big) = f(x)g(x).$$

This proves that the functions $f + g$, tf and $f \cdot g$ are also continuous. □

Definition 17.1.32. Let X be a set. A family $\mathcal{F}$ of functions $f : X \to \mathbb{R}$ is said to be *ample* or to *separate points* if, for any $x_0, x_1 \in X$, $x_0 \neq x_1$, there exists a function f in the family $\mathcal{F}$ such that $f(x_0) \neq f(x_1)$. □

Corollary 17.1.33. *Let (X, d) be a metric space. Then the collection $C(X)$ of continuous functions on X is an algebra that separates points.*

Proof. Let $x_0, x_1 \in X$ such that $x_0 \neq x_1$ Define

$$f : X \to \mathbb{R}, \;\; f(x) = \frac{1}{d(x_1, x_0)} d(x, x_0).$$

Then f is continuous, $f(x_0) = 0$, $f(x_1) = 1$. □

Proposition 17.1.34. *Suppose that (X, d_X) and (Y, d_Y) are metric spaces and $F : X \to Y$ is a map. Then the following statements are equivalent.*

(i) *The map F is continuous.*

(ii) *For any open set $U \subset Y$ the preimage $F^{-1}(U)$ is open (in X).*

(iii) *For any closed set $C \subset Y$ the preimage $F^{-1}(C)$ is closed (in X).*

Proof. (i) ⇒ (iii) Let $C \subset Y$ be closed. To prove that $F^{-1}(C)$ is closed we use Proposition 17.1.16 . Suppose that (x_n) is a sequence of points in $F^{-1}(C)$ converging to $x \in X$. We will prove that $x \in F^{-1}(C)$.

Indeed, since F is continuous, the sequence $F(x_n)$ of points in C converges to $F(x)$. Since C is closed, $F(x) \in C$ so that $x \in F^{-1}(C)$.

(iii) ⇒ (ii) Let $U \subset Y$ be open. Then $C = Y\backslash U$ is closed so $F^{-1}(C)$ is closed. Now observe that

$$F^{-1}(C) = F^{-1}(Y\backslash U) = X\backslash F^{-1}(U).$$

Hence $F^{-1}(U) = X\backslash F^{-1}(C)$ is open.

(ii) ⇒ (i) Fix $x_0 \in X$ and set $y_0 = F(x_0)$. We want to show that F is continuous at x_0. We will prove that F satisfies (17.1.3).

Let $\varepsilon > 0$. The ball $B^Y_\varepsilon(y_0)$ is open and thus its preimage $F^{-1}\big(B^Y_\varepsilon(y_0)\big)$ is open in X. Since $x_0 \in F^{-1}\big(B^Y_\varepsilon(y_0)\big)$ we deduce that there exists $\delta > 0$ such that

$$B^X_\delta(x_0) \subset F^{-1}\big(B^Y_\varepsilon(y_0)\big),$$

i.e., $F\big(B^X_\delta(x_0)\big) \subset B^Y_\varepsilon(y_0)$. □

Corollary 17.1.35. *Let (X, d) be a metric space and $f, g : X \to \mathbb{R}$ two continuous functions. Then the set*

$$E := \big\{x \in X;\ \ f(x) = g(x)\big\}$$

is closed. In particular, if f and g coincide on a dense subset of X, then they are identical.

Proof. The difference $h = f - g$ is a continuous function. Observe that

$$E = h^{-1}\big(\{0\}\big),$$

and since $\{0\}$ is a closed subset of $\mathbb{R}$, its preimage via the continuous function h is also closed.

Suppose now that $Y \subset X$ is a dense subset of X and $f(y) = g(y)$, $\forall y \in Y$. Thus, the set Y is contained in the closed E so the closure of Y is also contained. Since Y is dense $\boldsymbol{cl}\, Y = X$ so $X \subset E \subset X$. □

Corollary 17.1.36. *Suppose that (X, d_X), (Y, d_Y) , (Z, d_Z) are metric spaces and*

$$F : X \to Y \ \ \text{and} \ \ G : Y \to Z$$

are continuous maps. Then their composition $G \circ F : X \to Z$ is also continuous.

Proof. We will show that for any open set $U \subset Z$ the preimage $(G \circ F)^{-1}(U)$ is also open. We have

$$(G \circ F)^{-1}(U) = F^{-1}\big(G^{-1}(U)\big).$$

Since G is continuous the preimage $G^{-1}(U)$ is open, and since F is open, the preimage $F^{-1}\big(G^{-1}(U)\big)$ is also open. □

The above result shows that the concept of continuity is a topological concept.

Definition 17.1.37. Suppose that $(X_0, \mathcal{T}_0)$ and $(X_1, \mathcal{T}_1)$ are topological spaces. A map $F : X_0 \to X_1$ is called *continuous* (with respect to the above topologies) if for any open set $U_1 \in \mathcal{T}_1$ the preimage $F^{-1}(U_1)$ is an open subset $F^{-1}(U_1) \in \mathcal{T}_0$. □

If $(X, \mathcal{T})$ is a topological space, then a function $f : X \to \mathbb{R}$ is called continuous if it is continuous as a map $X \to \mathbb{R}$ where $\mathbb{R}$ is equipped with the Euclidean topology. Equivalently, this means that for any *open* interval $I \subset \mathbb{R}$ the preimage $f^{-1}(I)$ is an open subset of X, i.e.,

$$f^{-1}(I) \in \mathcal{T}.$$

We will denote by $C(X)$ or $C(X, \mathcal{T})$ the space of continuous functions $X \to \mathbb{R}$. As in the case of metric spaces $C(X)$ is an $\mathbb{R}$-algebra. Any constant function is continuous. *However proving that there exist nonconstant functions on an arbitrary topological space is a more challenging task!*

Definition 17.1.38. Suppose that $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ are topological space spaces and $F : X \to Y$ is a map. We say that F is a *homeomorphism* if it satisfies the following conditions.

(i) The map F is continuous.

(ii) The map F is bijective.

(iii) The inverse map $F^{-1} : Y \to X$ is also continuous.

□

17.1.4. Connectedness. The notion of connectedness discussed earlier has a (more subtle) counterpart in the more general case of metric space. Let (X, d) be a metric space. A *clopen* subset of X is a subset that is simultanuously open and closed. Clearly both X and $\varnothing$ are clopen. In some cases, the metric space (X, d) can contain nontrivial clopen subsets.

Suppose $Y = [0, 1] \cup [2, 3]$ and we regard Y as a metric subspace of $\mathbb{R}$. Let $S_0 = [0, 1]$ and $S_1 = [2, 3]$. Since

$$S_0 = (-1, 1.5) \cap Y \text{ and } S_1 = (1.5, 4) \cap Y$$

we deduce from Proposition 17.1.12 that both S_0 and S_1 are open in Y. On the other hand, $S_1 = Y \backslash S_0$ so that S_1 is also closed as the complement of an open subset. Thus S_1 is clopen.

Definition 17.1.39. Let (X, d) be a metric space.

(i) The metric space (X, d) is said to be *connected* if $X, \varnothing$ are the only clopen subsets of X.

(ii) The metric space is called *disconnected* if it is not connected.

(iii) A subset $Y \subset X$ is said to be *connected* if the metric subspace (Y, d_Y) is connected.

□

You should think of a separation as coloring each point of X with one of two colors, black or white so that the resulting black region is nonempty and open and so is the resulting white region.

From Proposition 17.1.12 we deduce that a subset Y of a metric space is disconnected if and only if it admits a *separation*, i.e., a pair of open sets $U_0, U_1 \subset X$ such that

$$U_0 \cap Y,\ U_1 \cap Y \neq \varnothing \ \text{ and } \ U_0 \cap U_1 \cap Y = \varnothing,\ \ Y \subset U_0 \cup U_1.$$

Indeed, the above condition shows that $U_0 \cap Y$ and $U_1 \cap Y$ are nontrivial open subsets of Y that complement each other. Observe that the connectedness property is a topological property.

Proposition 17.1.40. *Suppose that $(Y_i)_{i \in I}$ is a family of connected subsets of the metric space (X, d) that have at least one point in common, i.e.,*

$$\bigcap_{i \in I} Y_i \neq \varnothing.$$

Then their union

$$Y = \bigcup_{i \in I} Y_i$$

is also connected.

Proof. We argue by contradiction. Assume that Y is disconnected. Then there exist open sets $U_0, U_1 \subset X$ such that

$$U_0 \cap Y,\ U_1 \cap Y \neq \varnothing \ \text{ and } \ U_0 \cap U_1 \cap Y = \varnothing,\ \ Y \subset U_0 \cup U_1. \tag{17.1.6}$$

Fix

$$y \in \bigcap_{i \in I} Y_i.$$

Then either $y \in U_0$, or $y \in U_1$. Note that $y \notin U_0 \cap U_1$ since $Y \cap U_0 \cap U_1 = \varnothing$. Suppose that $y \in U_0$. For $i \in I$ the set Y_i is connected and

$$U_0 \cup U_1 \supset Y_i,\ \ U_0 \cap U_1 \cap Y_i = \varnothing,\ \ U_0 \cap Y_i \neq \varnothing.$$

We deduce that $U_1 \cap Y_i = \varnothing\ \forall i \in I$ so

$$U_1 \cap Y = U_1 \cap \left(\bigcup_{i \in I} Y_i \right) = \varnothing.$$

This contradicts (17.1.6). □

Suppose now that (X, d) is a metric space. We define a binary relation " $\sim$ " on X by declaring $x_0 \sim x_1$ if there exists a connected subset Y of X that contains both x_0 and x_1. This relation is reflexive since the singletons $\{x\}$, $x \in X$ are connected subsets. The relation is by definition symmetric. Let us show that it is also transitive.

Suppose $x_0 \sim x_1$ and $x_1 \sim x_2$. Then there exist connected sets $Y_0, Y_1 \subset X$ such that

$$x_0, x_1 \in Y_0, \;\; x_1, x_2 \in Y_1.$$

The overlap $Y_0 \cap Y_1$ is nonempty because it contains the point x_1. Proposition 17.1.40 shows that the union $Y = Y_0 \cup Y_1$ is connected, and since $x_0, x_2 \in Y$, we deduce $x_0 \sim x_2$.

Hence, the binary relation $\sim$ is an equivalence relation on X. Its equivalence classes are called the *connected components* of X. For example, if

$$X = [0,1] \cup \{2\} \cup (3, \infty)$$

then, the three sets that appear in the above union are the connected components of X. The next result is fundamental, very believable, yet its proof is rather subtle.

Theorem 17.1.41. *The interval* $[0,1]$ *is a connected subset of* $\mathbb{R}$.

Proof. Suppose that $U_0, U_1 \subset \mathbb{R}$ are two open sets such that

$$[0,1] \subset U_0 \cup U_1 \;\text{ and }\; U_0 \cap U_1 \cap [0,1] = \varnothing.$$

We want to prove that $[0,1]$ is contained in one of the sets U_0 or U_1. Think of the points in $U_0 \cap [0,1]$ as being colored in white and the ones in $U_1 \cap [0,1]$ as colored in black. Then the above conditions mean that any point in $[0,1]$ is either white, or black but not both and, if a point in $[0,1]$ has a one of these colors, then all the nearby points have the same color.

Since $0 \in U_0 \cup U_1$ we can assume without loss of generality that $0 \in U_0$. We will show that $[0,1] \subset U_0$ by relying on an argument very similar to the one used in the proof of Theorem 12.2.18. Define

$$S := \big\{ s \in [0,1]; \;\; [0,s] \subset U_0, \big\}.$$

Thus S consists of all the white points s such that all the points in $[0,1]$ behind s are also white. Note that $S \neq \varnothing$ since $0 \in S$. We set

$$s^* = \sup S \leqslant 1.$$

Step 1. $s^* > 0$, i.e., s^* is a white point. Indeed, since $0 \in U_0$ and U_0 is an open subset of $\mathbb{R}$, there exists $\varepsilon > 0$ such that $[0, \varepsilon) \subset U_0 \cap [0,1]$.[1] Thus $[0, \varepsilon) \subset S$ so $s^* \geqslant \varepsilon$.

Step 2. $s^* \in U_0 \cap [0,1]$. We argue by contradiction. Suppose that $s^* \notin U_0 \cap [0,1]$. Then $s^* \in U_1 \cap [0,1]$. Since U_1 is open and $s^* > 0$, there exists $\varepsilon > 0$ such that $(s^* - \varepsilon, s^*] \subset U_1 \cap [0,1]$. On the other hand, since s^* is the *least* upper bound of S, there exists $s_\varepsilon \in S \cap (s^* - \varepsilon, s^*]$. Thus $s_\varepsilon \in U_0 \cap U_1 \cap [0,1] = \varnothing$.

[1] The origin 0 is white and thus all the nearby points in $[0,1]$ are also white.

Step 3. $s^* = 1$. Suppose $s^* < 1$. Since U_0 is open, there exists $\varepsilon > 0$ such that

$$[s^*, s^* + \varepsilon) \subset U_0 \cap [0,1].$$

On the other hand, there exists a sequence s_n in S such that $s_n \nearrow s^*$. Thus

$$[0, s_n] \subset U_0 \cap [0,1], \;\; \forall n,$$

so that

$$[0, s^*) = \bigcup_n [0, s_n] \subset U_0 \cap [0,1].$$

This shows that $[0, s^* + \varepsilon) \subset U_0 \cap [0,1]$ so

$$[0, s^* + \varepsilon) \subset S.$$

This contradicts the fact that $s^* = \sup S$. Hence $1 \in S$ so $[0,1] \subset U_0 \cap [0,1]$. □

To produce many examples of connected spaces we need the following simple yet very powerful result.

Theorem 17.1.42. *Suppose that (X_0, d_0) and (X_1, d_1) are metric spaces and $F : X_0 \to X_1$ is a continuous map. If X_0 is connected, then the image of F is a connected subset of X_1.*

Proof. We argue by contradiction. Suppose that $Y = F(X_0)$ is disconnected. Then there exist open sets $U_0, U_1 \in X_1$ such that

$$U_0 \cap Y, \; U_1 \cap Y \neq \varnothing, \;\; Y \subset U_0 \cup U_1, \;\; U_0 \cap U_1 \cap Y = \varnothing$$

Set $V_i = F^{-1}(U_i) \subset X_0$, $i = 0, 1$. Since F is continuous the sets V_i are open in X_0. We have $F^{-1}(Y) = X_0$ and we deduce

$$V_0, \; V_1 \neq \varnothing, \;\; X_0 = V_0 \cup V_1, \;\; V_0 \cap V_1 = \varnothing.$$

This shows that V_0, V_1 are proper, nonempty clopen subsets of X_0. This is impossible since X_0 is connected.

□

Corollary 17.1.43. *Suppose that $F : (X_0, d_0) \to (X_1, d_1)$ is a homeomorphism between metric spaces. Then X_0 is connected if and only if X_1 is connected.*

Proof. Note that both F and its inverse F^{-1} are continuous and

$$X_1 = F(X_0), \;\; X_0 = F^{-1}(X_1).$$

□

Definition 17.1.44. Let (X, d) be a metric space.

(i) A *continuous path* in X is a continuous map $\gamma : [0,1] \to X$.

(ii) The space X is called *path connected* if for any points $x, x' \in X$ there exists a continuous path in X connecting them, i.e., a continuous path $\gamma : [0,1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = x'$.

□

Corollary 17.1.45. *A path connected metric space (X, d) is connected.*

Proof. Fix $x_0 \in X$ and, for any $x \in X$ fix a continuous path $\gamma_x : [0,1] \to X$ connecting x_0 to x and denote by Y_x the image of γ_x, $Y_x = \gamma_x\big([0,1]\big)$. The interval $[0,1]$ is connected and, according to Theorem 17.1.42, the space Y_x is also connected. Clearly $x \in Y_x$, $\forall x \in X$ so

$$\bigcup_{x \in X} Y_x = X.$$

On the other hand

$$x_0 \in \bigcap_{x \in X} Y_x$$

Proposition 17.1.40 now implies that X is connected. □

Corollary 17.1.46. *A subset $S \subset \mathbb{R}$ is connected if and only if it is an interval.*

Proof. Clearly, if S is an interval it is connected because it is path connected. We will show conversely, that if S is connected then S is path connected and thus, according to Proposition 12.2.4 , it is an interval.

Let $s_0, s_1 \in S$. We claim that $[s_0, s_1] \subset S$. Indeed, if there exists $s_* \in (s_0, s_1)$ that is not in S, then note that

$$(-\infty, s_*) \cap S = (-\infty, s_*] \cap S$$

is a nonempty clopen set (it contains s_0) and it is strictly contained in S, since it does not contain s_1. This proves that S is path connected. □

Corollary 17.1.47 (Intermediate Value Theorem)**.** *Suppose that (X, d) is a* connected *metric space and $f : X \to \mathbb{R}$ is a continuous function. If there exist $x_\pm \in X$ such that*

$$f(x_-) < 0 \;\; \textit{and} \;\; f(x_+) > 0,$$

then there exists $x_0 \in X$ such that $f(x_0) = 0$.

Proof. From Theorem 17.1.42 we deduce that $S = f(X)$ is a connected subset of $\mathbb{R}$ and thus it is an interval. Note that $f(x_\pm) \in S$ so $0 \in [f(x_-), f(x_+)] \subset S$. Thus 0 is in the range of f. □

17.1.5. Continuous linear maps. Suppose that $(X, \| - \|_X)$ and $(Y, \| - \|_Y)$ are two normed spaces, either both real or both complex.

Theorem 17.1.48. *Let $T : X \to Y$ be a linear map. Then the following statements are equivalent.*

(i) *The map T is continuous.*

(ii) *The map T is continuous at $0 \in X$.*

(iii) *The map T is* bounded, *i.e., there exists $C > 0$ such that*

$$\|Tx\|_Y \leqslant C, \ \ \forall x \in X \ \ \textit{such that} \ \|x\|_X \leqslant 1.$$

(iv) *There exists $C > 0$ such that*

$$\forall x \in X, \ \ \|Tx\|_Y \leqslant C\|x\|_X. \tag{17.1.7}$$

(v) *The map T is Lipschitz.*

Proof. Clearly (v) ⇒ (i) ⇒ (ii). Note that (iv) ⇒ (v). Indeed if $x_0, x_1 \in X$, then

$$\|Tx_0 - Tx_1\|_Y = \|T(x_0 - x_1)\|_Y \overset{(17.1.7)}{\leqslant} C\|x_0 - x_1\|_X$$

so T is Lipschitz. Thus all we have left to prove is that (ii) ⇒ (iii) ⇒ (iv).

(ii) ⇒ (iii) Since T is continuous at 0 there exists $\delta > 0$ such that

$$\|Tx\|_Y \leqslant 1, \ \ \forall \|x\|_X < \delta.$$

Let $x \in X$, $\|x\|_X \leqslant 1$. Then $\|\delta x\|_X \leqslant \delta$ so

$$\delta\|Tx\|_Y \leqslant 1$$

so that

$$\|Tx\|_Y \leqslant \frac{1}{\delta}, \ \ \forall \|x\|_X \leqslant 1.$$

(iii) ⇒ (iv) Let $x \in X\backslash\{0\}$ and define

$$\bar{x} := \frac{1}{\|x\|_X} x.$$

Since $\|\bar{x}\| = 1$, we deduce that from (iii)

$$\frac{1}{\|x\|_X}\|Tx\|_Y \leqslant C,$$

i.e.,

$$\|Tx\|_Y \leqslant C\|x\|_X, \ \ \forall x \in X\backslash\{0\}.$$

Clearly the above inequality holds trivially for $x = 0$. □

We ought to justify the usage of the term "*bounded*" in property (iii) above. A set in a normed space is called bounded if it is contained in some ball centered at the origin. Property (iii) states that the image of the unit ball in X via T is a bounded subset of Y. Equivalently, this means that T maps bounded subsets of X to bounded subsets of Y.

Definition 17.1.49. For any pair of normed spaces $(X, \|-\|_X)$, $(Y, \|-\|_Y)$ we denote by $\boldsymbol{B}(X, Y)$ the space of bounded or, equivalently, continuous linear maps $X \to Y$. We set $\boldsymbol{B}(X) := \boldsymbol{B}(X, X)$. □

The space $\boldsymbol{B}(X, Y)$ is a vector space itself. For any operator $T \in \boldsymbol{B}(X, Y)$ we set

$$\|T\|_{\mathrm{op}} := \sup_{x\in X\setminus\{0\}} \frac{\|Tx\|_Y}{\|x\|_X} = \inf\Big\{ C>0;\;\; \|Tx\|_Y \leqslant C\|x\|_X,\;\; \forall x\in X \Big\}.$$

Note that

$$\|Tx\|_Y \leqslant \|T\|_{\mathrm{op}}\|x\|_X,\;\; \forall x\in X. \tag{17.1.8}$$

Proposition 17.1.50. *For any normed spaces $(X, \|-\|_X)$, $(Y, \|-\|_Y)$ the function*

$$\|-\|_{\mathrm{op}} : \boldsymbol{B}(X,Y) \to \mathbb{R}$$

is a norm. We will refer to it as the operator norm. *Moreover if $(Z, \|-\|_Z)$ is another normed space, $S : X \to Y$ and $T : Y \to Z$ are continuous linear maps, then*

$$\|T\circ S\|_{\mathrm{op}} \leqslant \|T\|_{\mathrm{op}} \cdot \|S\|_{\mathrm{op}}. \tag{17.1.9}$$

□

The proof is left as an exercise.

Definition 17.1.51. For any real normed space $(X, \|-\|)$ we denote by X^* the space of *continuous* linear maps $X \to \mathbb{R}$. We denote by $\|-\|_*$ the operator norm on the vector space $X^* = \boldsymbol{B}(X,\mathbb{R})$. The resulting normed space $(X^*, \|-\|_*)$ is called the *topological dual* of X. □

Remark 17.1.52. Suppose that $(X, \|-\|)$ is a real normed space. Note that a linear functional $\boldsymbol{\alpha} : X \to \mathbb{E}$ is continuous if and only if

$$\exists C>0,\;\; \forall x\in X:\;\; |\boldsymbol{\alpha}(x)| \leqslant C\|x\|.$$

The norm of α is then

$$\|\alpha\|_* = \sup_{x\neq 0} \frac{\big|\boldsymbol{\alpha}(x)\big|}{\|x\|} = \inf\big\{ C>0;\;\; \big|\alpha(x)\big| \leqslant C\|x\| \big\}. \qquad □$$

Observe that on any normed space X there exists at least one continuous linear functional $\alpha : X \to \mathbb{R}$ namely, the trivial one, identically zero. The next result shows that in fact there are plenty of continuous linear functionals.

Theorem 17.1.53. *Let $(X, \|-\|)$ be a real normed space and $Y \subsetneq X$ a closed subspace. Then, for any $x_0 \in X\setminus Y$ there exists a continuous linear functional $\alpha : X \to \mathbb{R}$ such that $\alpha(x_0) = 1$ and $\alpha(y) = 0$, $\forall y \in Y$.*

Proof. The proof is nonconstructive since it based on Zorn's lemma. Fix $x_0 \in X\setminus Y$ and set $U_0 := \operatorname{span}\{x_0\} + Y$. Since Y is closed and $x_0 \notin Y$ we deduce from Proposition 17.1.27 that $d_0 := \operatorname{dist}(x_0, Y) > 0$. Define

$$\alpha_0 : U_0 \to \mathbb{R},\;\; \alpha_0(tx_0 + y) = t.$$

Note that for $t \neq 0$

$$\|tx_0 + y\| = |t|\|x_0 + t^{-1}y\| \geqslant |t| \inf_{y' \in Y} \|x_0 - y'\| = |t| \operatorname{dist}(x_0, Y) = d_0|t|.$$

We deduce that

$$\big|\,\alpha(t_0 + y)\,\big| = |t| \leqslant \frac{1}{d_0}\|tx_0 + y\|, \;\; \forall t \in \mathbb{R}, \;\; y \in Y,$$

so that α_0 is continuous. Moreover

$$\|\alpha_0\|_* = \frac{1}{d_0} = \frac{1}{\operatorname{dist}(x_0, Y)}.$$

We denote by $\mathcal{U}$ the collection of pairs (U, α), where

- $U \subset X$ is a subspace containing U_0,
- $\alpha : U \to \mathbb{R}$ is a continuous linear functional such that $\alpha\big|_{U_0} = \alpha_0$,
- $\|\alpha\|_* \leqslant \|\alpha_0\|_*$.

Let observe that $\mathcal{U}$ is not empty since $(U_0, \alpha_0) \in \mathcal{U}$.

We have a partial order on $\mathcal{U}$,

$$(U, \alpha) \prec (V, \beta) \Longleftrightarrow U \subset V, \;\; \beta\big|_U = \alpha.$$

Let us observe that any chain $\big\{\,(U_i, \alpha_i)_{i \in I}\,\big\}$ in $\mathcal{U}$ admits an upper bound. Indeed, set

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

and define $\alpha : U \to \mathbb{R}$, $\alpha(u) = \alpha_i(u)$ if $u \in U_i$. Note that if $u \in U_i \cap U_j$, then either $U_i \subset U_j$, or $U_j \subset U_i$. In the first case $\alpha_j(u) = \alpha_i(u)$ since $\alpha_j\big|_{U_i} = \alpha_i$. In the second case $\alpha_j(u) = \alpha_i(u)$ since $\alpha_i\big|_{U_j} = \alpha_j$. Clearly the pair (U, α) belongs to the family $\mathcal{U}$ since for any $u \in U$, exists i such that $u \in U_i$ and

$$|\alpha(u)| = |\alpha_i(u)| \leqslant \|\alpha_i\|_* \|u\| \leqslant \|\alpha_0\|_* \|u\|.$$

Zorn's lemma (Theorem A.1.5) implies that $\mathcal{U}$ admits a maximal element (U_*, α_*). We will show that $U_* = X$. Then α_* is a continuous linear functional with the postulated properties.

To prove that $U_* = X$ we argue by contradiction. Suppose that there exists $x_1 \in X \backslash U_*$. We set

$$U_1 = U_* + \operatorname{span}(x_1).$$

Any $u \in U_1$ admits a unique decomposition of the form $u_1 = u_* + tx_1$, $u_* \in U_*$, $t \in \mathbb{R}$. For $c \in \mathbb{R}$ define

$$\alpha_c : U_1 \to \mathbb{R}, \;\; \alpha_c(u_* + tx_1) = \alpha_*(u_*) + ct.$$

We claim that there exists $c \in \mathbb{R}$ such that $(U_1, \alpha_c) \in \mathcal{U}$. Then

$$(U_*, \alpha_*) \prec (U_1, \alpha_c)$$

contradicting the maximality of (U_*, α_*).

Note that $(U_1, \alpha_c) \in \mathfrak{U}$ iff

(i) $\alpha_c\big|_{U_0} = \alpha_0$, and

(ii) $\exists 0 < K \leqslant \|\alpha_0\|_*$ such that $\big|\, \alpha_c(u_1) \,\big| \leqslant K\|u_1\|$, $\forall u_1 \in U_1$.

Condition (i) is satisfied for any $c \in \mathbb{R}$. We will prove that there exists $c \in \mathbb{R}$ such that α_c satisfies (ii) as well.

Set K_* denote the norm of α_*, $K_* := \|\alpha_*\|_* \leqslant \|\alpha_0\|_*$. We will show that there exists $c \in \mathbb{R}$ such that

$$\alpha_*(u_*) + c \leqslant K_*\|u_* + x_1\| \text{ and } \alpha_*(u_*) - c \leqslant K_*\|u_* - x_1\|, \ \ \forall u_* \in U_*. \tag{17.1.10}$$

Note that if c satisfies (17.1.10) then for any $t > 0$ we have

$$\alpha_c(u_* + tx_1) = t\big(\,\alpha_*\big(\,t^{-1}u_*\,\big) + c\,\big) \leqslant tK_*\big\|\,t^{-1}u_* + x_1\,\big\| = K_*\|u_* + tx_1\|$$

$$-\alpha_c(u_* + tx_1) = -t\big(\,\alpha_*\big(\,-t^{-1}u_*\,\big) - c\,\big) \geqslant -tK_*\big\|\,-t^{-1}u_* - x_1\,\big\| = -K_*\|x_* + tx_1\|$$

In other words

$$\big|\,\alpha_c(u_* + tx_1)\,\big| \leqslant K_*\|x_* + tx_1\|, \ \ \forall t > 0.$$

A similar argument shows that (17.1.10) implies that

$$\big|\,\alpha_c(u_* + tx_1)\,\big| \leqslant K_*\|x_* + tx_1\|, \ \ \forall t < 0.$$

Hence (17.1.10) implies that α_c satisfies the condition (ii) above. In other words, the proof is completed if we show that there exists c satisfying (17.1.10). Note that c satisfies (17.1.10) iff

$$\underbrace{\sup_{u_* \in U_*}\big(\,\alpha_*(u_*) - K_*\|u_* - x_1\|\,\big)}_{=:\lambda} \leqslant c \leqslant \underbrace{\inf_{v_* \in U_*}\big(\,K_*\|v_* + x_1\| - \alpha_*(v_*)\,\big)}_{=:\rho}$$

Since $\|\alpha_*\| = K_*$, for any $u_*, v_* \in U_*$ we have

$$\alpha(u_*) + \alpha(v_*) = \alpha(u_* + v_*) \leqslant K_*\|u_* + v_*\| \leqslant K_*\|u_* - x_1\| + K_*\|v_* + x_1\|$$

i.e.,

$$\alpha_*(u_*) - K_*\|u_* - x_1\| \leqslant K_*\|v_* + x_1\| - \alpha_*(v_*), \ \ \forall u_*, v_* \in U_*.$$

Hence $-\infty < \lambda \leqslant \rho < \infty$ so that if $c \in [\lambda, \rho]$ the condition (17.1.10) is satisfied. Note that

$$\|\alpha_*\|_* = \|\alpha_0\|_* = \frac{1}{d_0}.$$

□

Remark 17.1.54. Theorem 17.1.53 is a very special case of a fundamental result in functional analysis called the *Hahn-Banach Theorem.*

Let us observe that Theorem 17.1.53 implies that the collection of continuous linear functions on a normed space is ample in the sense of Definition 17.1.32.

Recall that this means that for any $x_1, x_2 \in X$, $x_1 \neq x_2$ there exists a continuous linear functional $\alpha \in X^*$ such that $\alpha(x_1) \neq \alpha(x_2)$.

Indeed, if we set $x_0 = x_2 - x_1 \neq 0$, then Theorem 17.1.53 with $Y = \{0\}$ implies that there exists $\alpha \in X^*$ such that

$$\alpha(x_2) - \alpha(x_1) = \alpha(x_2 - x_1) = 1 \neq 0.$$

□

Corollary 17.1.55. *Let $(X, \|-\|)$ be a normed space and $Z \subset X$ a subspace. Then the following are equivalent.*

(i) *The subspace Z is not dense in X.*

(ii) *There exists a nontrivial continuous linear functional $\alpha : X \to \mathbb{R}$ such that $\alpha(z) = 0$, $\forall z \in Z$.*

Proof. Set $Y = \boldsymbol{cl}(Z)$. The implication (ii) ⇒ (i) is obvious. To prove (i) ⇒ (ii) note that $Y \subsetneq X$. The conclusion follows from Theorem 17.1.53.

□

Remark 17.1.56. We should point out the rather paradoxical nature of the above result. We set

$$Z^{\perp} = \big\{ \alpha \in X^*;\;\; \alpha(z) = 0,\; \forall z \in Z \big\}.$$

Observe that $Z^{\perp}$ is a vector subspace of X^*. Corollary 17.1.55 shows that if $Z^{\perp} = 0$, so that $Z^{\perp}$ is as small as possible, then Z has to be very large. How large? Any puny open ball in X must contain at least one point in Z. □

Example 17.1.57. As we have seen, on a given vector space X there are many choices of norms and a linear functional $\alpha : X \to \mathbb{R}$ could be continuous with respect to one choice of norm, but discontinuous with respect to another.

Consider for example the space $X = C([0,1])$ and the linear functional

$$\delta : C([0,1]) \to \mathbb{R},\;\; \delta(f) = f(1).$$

This linear functional is continuous with respect to the sup-norm since

$$\big|\,\delta(f)\,\big| = \big|\,f(1)\,\big| \leqslant \sup_{x \in [0,1]} |f(x) = \|f\|_{\infty}.$$

On the other hand, it is discontinuous with respect to the norm $\|-\|_1$. To see this consider sequence of functions

$$f_n : [0,1] \to \mathbb{R},\;\; f_n(x) = x^n,\;\; n \in \mathbb{N}.$$

Note that f_n is continuous and nonnegative so

$$\|f_n\|_1 = \int_0^1 x^n dx = \frac{1}{n+1}.$$

Thus, the sequence f_n converges to 0 with respect to the norm $\|-\|_1$. On the other hand $\delta(f_n) = 1$, $\forall n$, so $\delta(f_n)$ does not converge to $\delta(0) = 0$. □

Proposition 17.1.58. *Consider two normed spaces $(X, \|-\|_X)$, $(Y, \|-\|_Y)$ and a linear isomorphism $T : X \to Y$. Then the following are equivalent.*

(i) *The map T is a homeomorphism.*

(ii) *There exist constants $0 < c < C$ such that*

$$\forall x \in X,\ \ c\|x\|_X \leqslant \|Tx\|_Y \leqslant C\|x\|_X. \tag{17.1.11}$$

Proof. (i) $\Rightarrow$ (ii) The map T is homeomorphism if and only if both T and T^{-1} are continuous, i.e., if and only if, there exist positive constants C_1, C_2 such that

$$\|Tx\|_Y \leqslant C_1\|x\|_X,\ \ \|T^{-1}y\|_X \leqslant C_2\|y\|_Y,\ \ \forall x \in X,\ y \in Y. \tag{17.1.12}$$

If we let $y = Tx$ in the second inequality we deduce

$$\|x\|_X \leqslant C_2\|Tx\|_Y,\ \ \forall x \in X,$$

so (17.1.11) holds with $C = C_1$ and $c = \frac{1}{C_2}$. The same argument shows that (17.1.11) implies (17.1.12) with $C_2 = \frac{1}{c}$, i.e., (ii) $\Rightarrow$ (i). □

Proposition 17.1.59. *Suppose that $(X, \|-\|_X)$ is a real normed space and $T : \mathbb{R}^n \to X$ is a linear injective map.* We do not assume that T is continuous. *Then there exist constants $0 < c < C$ such that*

$$c\|\boldsymbol{v}\|_2 \leqslant \|T\boldsymbol{v}\|_X \leqslant C\|\boldsymbol{v}\|_2,\ \ \forall \boldsymbol{v} \in \mathbb{R}^n, \tag{17.1.13}$$

where $\|-\|_2$ denotes the Euclidean norm. In particular, if T is bijective, then it also is a homeomorphism.

Proof. Let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be the canonical basis of $\mathbb{R}^n$. Set $x_k = T\boldsymbol{e}_k \in X$, $c_k = \|x_k\|_X$. For any $\boldsymbol{v} = (v^1, \dots, v^n) \in \mathbb{R}^n$ we have

$$T\boldsymbol{v} = T\big(v^1\boldsymbol{e}_1 + \cdots + v^n\boldsymbol{e}_n\big) = v^1T\boldsymbol{e}_1 + \cdots + v^nT\boldsymbol{e}_n = \sum_{k=1}^n v^kx_k,$$

so that

$$\|T\boldsymbol{v}\|_X = \Big\|\sum_{k=1}^n v^kx_k\Big\|_X \leqslant \sum_{k=1}^n |v^k|\cdot\|x_k\|_X = \sum_{k=1}^n |v^k|c_k$$

(use the Cauchy-Schwarz inequality)

$$\leqslant \sqrt{|v^1|^2 + \cdots + |v^n|^2}\cdot\underbrace{\sqrt{c_1^2 + \cdots + c_n^2}}_{=:C} = C\|\boldsymbol{v}\|_2.$$

This proves the second inequality in (17.1.13). In particular, it shows that T *is continuous*.

To prove the first inequality consider the unit sphere

$$\boldsymbol{\Sigma} = \big\{\,\boldsymbol{v} \in \mathbb{R}^n;\ \ \|\boldsymbol{v}\|_2 = 1\,\big\},$$

and the function

$$f : \boldsymbol{\Sigma} \to \mathbb{R},\ \ f(\boldsymbol{v}) = \|T\boldsymbol{v}\|_X.$$

This function is continuous as the composition of three continuous maps

$$\boldsymbol{\Sigma} \xrightarrow{i_{\boldsymbol{\Sigma}}} \mathbb{R}^n \xrightarrow{T} X \xrightarrow{\|-\|} \mathbb{R}.$$

Since $\boldsymbol{\Sigma}$ is compact, there exists $\boldsymbol{v}_0 \in \boldsymbol{\Sigma}$ such that

$$c := f(\boldsymbol{v}_0) \leqslant f(\boldsymbol{v}), \;\; \forall \boldsymbol{v} \in \boldsymbol{\Sigma}.$$

On the other hand, $T\boldsymbol{v}_0 \neq 0$ since T is injective, and thus $c > 0$. Hence

$$\|T\boldsymbol{v}\|_X \geqslant c > 0, \;\; \forall \|\boldsymbol{v}\|_2 = 1.$$

If $\boldsymbol{v} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ and $\bar{\boldsymbol{v}} = \frac{1}{\|\boldsymbol{v}\|_2}\boldsymbol{v}$, then $\|\bar{\boldsymbol{v}}\|_2 = 1$ so that

$$\|T\bar{\boldsymbol{v}}\|_X \geqslant c \Rightarrow \|T\boldsymbol{v}\|_X \geqslant c\|\boldsymbol{v}\|_2.$$

□

Corollary 17.1.60. *Let $n \in \mathbb{N}$. If $(X, \|-\|)$ is an n-dimensional real normed space, then any linear isomorphism $\mathbb{R}^n \to X$ is a homeomorphism $(\mathbb{R}^n, \|-\|_2) \to (X, \|-\|)$.* □

Definition 17.1.61. Let X be a real vector space. Two norms $\|-\|_0$ and $\|-\|_1$ on X are called equivalent if there exist positive constants $C_2 > C_1 > 0$ such that

$$C_1\|x\|_0 \leqslant \|x\|_1 \leqslant C_2\|x\|_0, \;\; \forall x \in X.$$

□

Remark 17.1.62. We see that two norms $\|-\|_i$, $i = 0, 1$, on X are equivalent if and only if the identity map $\mathbb{1}_X$ induces a linear homeomorphism $(X, \|-\|_0) \to (X, \|-\|_1)$. In other words, two norms are equivalent if and only if the topologies they define are identical.

Note also that a sequence $(x_n)_{n\in\mathbb{N}}$ converges with respect to one norm if and only if it converges with respect to the other norm. Moreover a function $f : X \to \mathbb{R}$ is continuous with respect to a norm iff it is continuous with respect to the other. □

Corollary 17.1.63. *On a finite dimensional vector space X any two norms are equivalent.*

Proof. Suppose $n = \dim X$ and a linear isomorphism $T : \mathbb{R}^n \to X$. Given two norms $\|-\|_i$, $i = 0, 1$, we obtain two homeomorphisms

$$T : (\mathbb{R}^n, \|-\|_2) \to (X, \|-\|_i), \;\; i = 0, 1.$$

Now observe that $\mathbb{1}_X$ is the composition of two homeomorphisms

$$(X, \|-\|_0) \xrightarrow{T^{-1}} (\mathbb{R}^n, \|-\|_2) \xrightarrow{T} (X, \|-\|_1).$$

□

Proposition 17.1.64. *Suppose that $(X, \|-\|)$ is a real normed space. Then the following statements are equivalent.*

(i) $\dim X < \infty$.

(ii) *Any linear functional $X \to \mathbb{R}$ is continuous.*

Proof. (i) ⇒ (ii). Let $n = \dim X$. Fix a linear isomorphism $T : \mathbb{R}^n \to X$. According to Proposition 17.1.59, this induces a homeomorphism

$$T : (\mathbb{R}^n, \| - \|_2) \to (X, \| - \|).$$

Suppose that $\alpha : X \to \mathbb{R}$ is a linear map. We obtain a linear map $\beta = \alpha \circ T : \mathbb{R}^n \to \mathbb{R}$ which, according to Proposition 12.1.10, is continuous with respect to the Euclidean norm. We deduce that $\alpha = \beta \circ T^{-1}$ is also continuous.

(ii) ⇒ (i) We argue by contradiction. We will show that if $\dim X = \infty$, then there exists a discontinuous linear map $\alpha : X \to \mathbb{R}$. Assume that $\dim X = \infty$. According to Theorem A.1.6, the vector space X admits a basis $(e_i)_{i \in I}$. Since $\dim X = \infty$ the set I is infinite so there exists a surjection $c : I \to \mathbb{N}$.

Consider the linear map $\alpha : X \to \mathbb{R}$ uniquely determined by the conditions

$$\alpha(e_i) = c(i)\|e_i\|, \quad \forall i \in I.$$

Since the function c is not bounded we deduce that the map α is not bounded, thus not continuous. □

The above result shows that many things that we take for granted in finite dimensions may not necessarily hold in infinite dimensions.

17.2. Completeness

We know that a sequence of real numbers converges if and only if it is Cauchy. Regarding the set $\mathbb{Q}$ of rational numbers as a metric subspace of the real axis we notice that the Cauchy sequences in $\mathbb{Q}$ do not necessarily converge to a point in $\mathbb{Q}$, but we can assign to this sequence a limit that lives in a bigger space $\mathbb{R}$, in which $\mathbb{Q}$ its as a dense subset. This is a reflection of a general paradigm detailed in this section.

17.2.1. Cauchy sequences.

Let (X, d) be a metric space.

Definition 17.2.1. A sequence $(x_n)_{n \in \mathbb{N}}$ of points in X is called *Cauchy* (with respect to the metric d) if

$$\forall \varepsilon > 0, \;\; \exists N = N(\varepsilon) > 0, \;\; \forall m, n > N : \;\; d(x_m, x_n) < \varepsilon. \tag{17.2.1}$$

□

Proposition 17.2.2. *If $(x_n)_{n \in \mathbb{N}}$ is a convergent sequence in X, then it is also Cauchy.*

Proof. Denote by x_* the limit of the sequence (x_n). Then, for any $\varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that,

$$\forall n > N(\varepsilon) \;\; d(x_n, x_*) < \frac{\varepsilon}{2}$$

Then, for any $m, n > N$ we have $d(x_m, x_n) \leqslant d(x_m, x_*) + d(x_*, x_n) < \varepsilon$. □

Example 17.2.3. The converse is not true. Consider the normed space $\big(C([0,1]), \|-\|_1\big)$ and the sequence of continuous functions (see Figure 17.1)

$$f_n : [0,1] \to \mathbb{R}, \quad f_n(x) = \begin{cases} 1, & 0 \leqslant x \leqslant \frac{1}{2}, \\ \text{linear}, & \frac{1}{2} < x \leqslant \frac{1}{2} + \frac{1}{n}, \\ 0, & \frac{1}{2} + \frac{1}{n} < x \leqslant 1. \end{cases}$$

Then, for any $n > m$ we have $f_m(x) \geqslant f_n(x)$, $\forall x \in [0,1]$ and

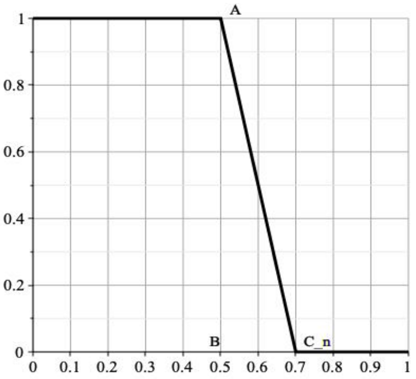

Figure 17.1. *The graph of* $f_5(x)$.

$$\|f_n - f_m\|_1 = \int_0^1 \big(f_m(x) - f_n(x) \big) dx = \text{area}(ABC_m) - \text{area}(ABC_n) = \frac{1}{2m} - \frac{1}{2n},$$

proving that the sequence (f_n) is Cauchy.

Intuitively, the sequence f_n seems to converge in some sense to a function that is equal to 1 on the open interval $(0, 1/2)$ and equal to 0 on $(1/2, 1)$. Such function cannot be continuous.

We will prove that indeed it does not converge in the norm $\| - \|_1$ to any continuous function. We argue by contradiction.

Suppose that f_n converges in the norm $\|-\|_1$ to some continuous function $f : [0,1] \to \mathbb{R}$. Note that for any compact interval $I \subset [0,1]$ we have

$$0 \leqslant \int_I \big| f_n(x) - f(x) \big| dx \leqslant \int_0^1 \big| f_n(x) - f(x) \big| dx = \|f_n - f\|_1 \to 0$$

so that,

$$\lim_{n\to\infty} \int_a^b |f_n(x) - f(x)| dx = 0, \quad \forall 0 \leqslant a < b \leqslant 1. \tag{17.2.2}$$

In particular

$$0=\lim_{n\to\infty}\int_0^{1/2}|f_n(x)-f(x)|dx=\int_0^{1/2}\left|1-f(x)\right|dx.$$

Since $f(x)$ is continuous we deduce (see Exercise 9.9) that $f(x)=1$, $\forall x\in[0,1/2]$.

Similarly, for any $a\in(1/2,1)$ we deduce $|f_n(x)-f(x)|=|f(x)|$, $\forall x\in(a,1]$ if n is sufficiently large. Using (17.2.2) we conclude as before

$$\int_a^1\left|f(x)\right|dx=0,\;\;\forall a\in(1/2,1]$$

so that $f(x)=0$, for $x\in(1/2,1]$. The function f is not continuous at $1/2$ since

$$\lim_{x\nearrow 1/2}f(x)=1\neq 0=\lim_{x\searrow 1/2}f(x).$$

□

Definition 17.2.4. A metric space (X,d) is said to be *complete* if any Cauchy sequence is convergent. A *Banach space* is a normed space such that the associated metric space is complete. □

Example 17.2.5. Theorem 11.4.10 shows that the Euclidean space $(\mathbb{R}^n,\|-\|_2)$ is complete. □

Proposition 17.2.6. *Suppose that (X,d) is a* complete *metric space. Then the following are equivalent.*

(i) *The subset $C\subset X$ is closed.*

(ii) *The metric subspace (C,d) is complete.*

Proof. (i) $\Rightarrow$ (ii) Suppose that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in C. It converges in X since (X,d) is complete. Its limit must be a point in C since C is closed.

(ii) $\Rightarrow$ (i) Suppose that $(x_n)_{n\in\mathbb{N}}$ is a convergent sequence in C. The sequence (x_n) is Cauchy since it is convergent. Because the metric subspace (C,d) is complete, the limit of this convergent sequence is a point in C. □

Example 17.2.7. Example 17.2.3 shows that the normed space $(C([0,1]),\|-\|_1)$ is not complete. On the other hand, the normed space $(C([0,1]),\|-\|_\infty)$ is complete.

Indeed, suppose that $f_n:[0,1]\to\mathbb{R}$, $n\in\mathbb{N}$ is a sequence of continuous functions that is Cauchy with respect to the sup-norm. Hence, for any $\varepsilon>0$ there exists $N=N(\varepsilon)>0$ such that,

$$\forall n,m>N(\varepsilon):\;\;\sup_{x\in[0,1]}\big|f_n(x)-f_m(x)\big|<\varepsilon.$$

Thus,

$$\forall x\in[0,1],\;\;\forall n,m>N(\varepsilon):\;\;\big|f_n(x)-f_m(x)\big|<\varepsilon. \tag{17.2.3}$$

Thus, for each $x \in [0,1]$, the sequence of real numbers $\big(f_n(x) \big)_{n\in\mathbb{N}}$ is Cauchy, hence convergent. Denote by $f(x)$ its limit. Letting $m \to \infty$ in (17.2.3) we deduce

$$\forall \varepsilon > 0, \forall x \in [0,1], \ \ \forall n > N(\varepsilon): \ \ \big| f_n(x) - f(x) \big| \leqslant \varepsilon,$$

i.e.,

$$\forall \varepsilon > 0, \ \ \forall n > N(\varepsilon): \ \ \sup_{x\in[0,1]} \big| f_n(x) - f(x) \big| \leqslant \varepsilon. \tag{17.2.4}$$

In other words, the sequence $(f_n(x))$ converges *uniformly* to $f(x)$ so, according to Theorem 6.1.10, the function $f(x)$ is continuous. Finally observe that (17.2.4) can be rephrased as

$$\lim_{n\to\infty} \|f_n - f\|_\infty = 0. \qquad \square$$

17.2.2. Completions. Let (X, d) be a metric space. We want to show that we can add “virtual” elements to the set X with the property that each new element can be viewed, in a suitable way, as the limit of a Cauchy sequence in X that need not converge in (X, d).

Definition 17.2.8. Let (X, d) be a metric space. A *completion* of (X, d) is a triplet $\big(\overline{X}, \bar{d}, \mathcal{I}\big)$ with the following properties.

(i) $\big(\overline{X}, \bar{d}\big)$ is a *complete* metric space.

(ii) $\mathcal{I}$ is an isometry $\mathcal{I} : (X, d) \to \big(\overline{X}, \bar{d}\big)$.

(iii) The image $\mathcal{I}(X)$ of $\mathcal{I}$ is a dense subset of $\overline{X}$.

□

Note that $\mathbb{R}$ is a completion of $\mathbb{Q}$. If (X, d) is complete, then $(X, d, \mathbb{1}_X)$ is a completion of (X, d).

Let us first observe that, if they exist, the completions are essentially unique. More precisely, the completions have the following *universality property*.

Theorem 17.2.9 (The universality property of completions)**.** *Suppose that* $\big(\overline{X}, \bar{d}, \mathcal{I}\big)$ *is a completion of the metric space* (X, d)*. Then, for any complete metric space* (Y, δ) *and any isometry* $T : (X, d) \to (Y, \delta)$*, there* $\underline{\textit{exists}}$ *a* $\underline{\textit{unique}}$ *isometry* $\overline{T} : \big(\overline{X}, \bar{d}\big) \to (Y, \delta)$ *such that the diagram below is commutative.*

$$\begin{tikzcd}
X \arrow[r, hook, "\mathcal{I}"] \arrow[dr, hook, "T"'] & \overline{X} \arrow[d, dashed, "\overline{T},"] \\
 & Y
\end{tikzcd}$$

i.e., $\overline{T} \circ \mathcal{I} = T$*.*

Proof. Since $\mathcal{I}$ is isometry, it is injective, and we can identify (X, d) with a metric subspace of $(\overline{X}, \bar{d})$. Note that if $F, G : \big(\overline{X}, \bar{d}\big) \to (Y, \delta)$ are two continuous maps such that $F(x) = G(x)$ for any $x \in X \subset \overline{X}$, then $F(\bar{x}) = G(\bar{x})$, for any $\bar{x} \in \overline{X}$.

Indeed, let $\bar{x} \in \overline{X}$. Since X is dense in $\overline{X}$, there exists a sequence (x_n) in X that converges to $\bar{x}$. From the continuity of F and G we deduce

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

This proves the uniqueness part of the theorem.

To prove the existence, consider $\bar{x} \in \overline{X}$. Since X is dense in $\overline{X}$, there exists a sequence (x_n) in X that converges to $\bar{x}$. The sequence (x_n) is Cauchy and, since T is an isometry, so is the sequence (Tx_n) in Y. On the other hand, the metric space (Y, δ) is complete so the sequence Tx_n is convergent. We claim that its limit depends only on $\bar{x}$ and not on the sequence (x_n) we used to approximate.

Indeed, if (x_n) and (x_n') are two sequences in X such that

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} x_n' = \bar{x},$$

Then

$$0 = \lim_{n\to\infty} d(x_n, x_n') = \lim_{n\to\infty} \delta(Tx_n, Tx_n').$$

From Corollary 17.1.30 we deduce that the metric map $\delta : Y \times Y \to \mathbb{R}$ is continuous so

$$0 = \lim_{n\to\infty} \delta(Tx_n, Tx_n') = \delta\big(\lim_n Tx_n, \lim_n Tx_n'\big).$$

Hence we set

$$\overline{T}\bar{x} := \lim_n Tx_n$$

wherever (x_n) is a sequence in X converging to $\bar{x}$. We obtain in this fashion a map $\overline{T} : \overline{X} \to Y$.

Note that if $\bar{x}, \bar{x}' \in \overline{X}$, then for sequences (x_n) and (x_n') converging to $\bar{x}$ and respectively $\bar{x}'$, we have

$$\delta\big(\overline{T}\bar{x}, \overline{T}\bar{x}'\big) = \lim_{n\to\infty} \delta(Tx_n, Tx_n') = \lim_{n\to\infty} \bar{d}(x_n, x_n') = \bar{d}(\bar{x}, \bar{x}').$$

This proves that $\overline{T}$ is an isometry. □

Corollary 17.2.10. *Any two completions of a metric space (X, d) are isometric.*

Proof. Suppose $\big(\overline{X}, \bar{d}, \mathcal{I}\big)$ and $\big(\overline{X}', \bar{d}', \mathcal{I}'\big)$. From the universality property of completions we deduce that there exist unique isometries

$$\bar{\mathcal{I}}' : \overline{X} \to \overline{X}' \;\text{ and }\; \bar{\mathcal{I}} : \overline{X}' \to \overline{X}$$

such that

$$\bar{\mathcal{I}}' \circ \mathcal{I} = \mathcal{I}',\;\; \bar{\mathcal{I}} \circ \mathcal{I}' = \mathcal{I}.$$

Now observe that

$$(\bar{\mathcal{I}} \circ \bar{\mathcal{I}}') \circ \mathcal{I} = \bar{\mathcal{I}} \circ (\bar{\mathcal{I}}' \circ \mathcal{I}) = \bar{\mathcal{I}} \circ \mathcal{I}' = \mathcal{I}.$$

Thus the isometry $T = \bar{\mathfrak{I}} \circ \bar{\mathfrak{I}}'$ makes the diagram below commutative.

$$\begin{tikzcd}
X \arrow[r, "\mathfrak{I}"] \arrow[dr, "\mathfrak{I}"'] & \overline{X} \arrow[d, dashed, "T"] \\
 & \overline{X}
\end{tikzcd}$$

On the other hand, the isometry $\mathbb{1}_{\overline{X}}$ also makes this diagram commutative. There exists exactly one such isometry we deduce

$$\mathbb{1}_{\overline{X}} = T = \bar{\mathfrak{I}} \circ \bar{\mathfrak{I}}'.$$

Arguing in a similar fasion we deduce

$$\mathbb{1}_{\overline{X}'} = T = \bar{\mathfrak{I}}' \circ \bar{\mathfrak{I}}$$

so that $\bar{\mathfrak{I}}'$ is the inverse of $\bar{\mathfrak{I}}$. Hence, $\bar{\mathfrak{I}}'$ is a bijective isometry $\overline{X} \to \overline{X}'$. □

Denote by $\boldsymbol{CS}(X)$ or $\boldsymbol{CS}(X, d)$ the set of Cauchy sequences in (X, d). We will use the notation $\underline{x}$ when referring to a sequence $(x_n)_{n\in\mathbb{N}}$ in X.

☞ *For each $x \in X$ we denote by $\langle x \rangle$ the constant sequence $x_n = x$, $\forall n \in \mathbb{N}$.*

Proposition 17.2.11. *For any Cauchy sequences $\underline{x}$ and $\underline{y}$, the sequence of real numbers $\big(d(x_n, y_n)\big)_{n\in\mathbb{N}}$ is Cauchy. We will denote by $\bar{d}(\underline{x}, \underline{y})$ its limit.*

Proof. Set $d_n := d(x_n, y_n)$. Corollary 17.1.30 implies that, $\forall m, n \in \mathbb{N}$, we have

$$|d_n - d_m| \leqslant d(x_n, x_m) + d(y_n, y_m).$$

Since the sequences $\underline{x}$ and $\underline{y}$ are Cauchy we deduce that,

$$\forall \varepsilon > 0 \;\; \exists N = N(\varepsilon) > 0, \;\; \forall m, n > N(\varepsilon) \;\; d(x_n, x_m) + d(y_n, y_m) < \varepsilon.$$

This proves that the sequence of real numbers (d_n) is Cauchy, hence convergent. □

Remark 17.2.12. Let us observe a few simple things about the map $\bar{d}$.

- If the sequences $\underline{x}$ and $\underline{y}$ converge in (X, d) to x_* and respectively y_*, then

  $$\bar{d}(\underline{x}, \underline{y}) = d(x_*, y_*).$$

  In particular, for any $x, y \in X$, $d(x, y) = \bar{d}\big(\langle x \rangle, \langle y \rangle\big)$.

- For any $\underline{x}, \underline{y} \in \boldsymbol{CS}(X)$

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

- For any $\underline{x}, \underline{y}, \underline{z} \in \boldsymbol{CS}(X)$, we have

  $$\bar{d}(\underline{x}, \underline{z}) \leqslant \bar{d}(\underline{x}, \underline{y}) + \bar{d}(\underline{y}, \underline{z}). \tag{17.2.5}$$

Indeed, this follows by letting $n \to \infty$ in the triangle inequality

$$d(x_n, z_n) \leqslant d(x_n, y_n) + d(y_n, z_n).$$

These facts show that the function $\bar{d} : \boldsymbol{CS}(X) \times \boldsymbol{CS}(X) \to \mathbb{R}$ behaves almost like a metric. We say "almost" because it violates the condition $\bar{d}(\underline{x}, \underline{y}) = 0 \Rightarrow \underline{x} = \underline{y}$. □

We define a binary relation $\sim$ on $\boldsymbol{CS}(X)$ by declaring $\underline{x} \sim \underline{y}$ if and only of $\bar{d}(\underline{x}, \underline{y}) = 0$. Clearly this binary relation is reflexive and symmetric. The inequality (17.2.5) shows that it is also transitive so that '$\sim$' is an equivalence relation.

Denote by $\overline{X}$ the set of equivalence classes of '$\sim$', i.e.

$$\overline{X} = \boldsymbol{CS}(X)/\sim .$$

For any $\underline{x} \in \boldsymbol{CS}(X)$ we denote by $C_{\underline{x}}$ its equivalence class. Note that if $\underline{x} \sim \underline{x}'$,

$$\bar{d}(\underline{x}, \underline{y}) \leqslant \bar{d}(\underline{x}, \underline{x}') + \bar{d}(\underline{x}', \underline{y}) = \bar{d}(\underline{x}', \underline{y})$$

and

$$\bar{d}(\underline{x}', \underline{y}) \leqslant \bar{d}(\underline{x}', \underline{x}) + \bar{d}(\underline{x}, \underline{y}) = \bar{d}(\underline{x}, \underline{y})$$

so that

$$\bar{d}(\underline{x}, \underline{y}) = \bar{d}(\underline{x}', \underline{y}).$$

Thus $\bar{d}$ induces a well defined function

$$\bar{d} : \overline{X} \times \overline{X} \to \mathbb{R} \quad \bar{d}\big(C_{\underline{x}}, C_{\underline{y}}\big) = \bar{d}\big(\underline{x}, \underline{y}\big).$$

The discussion above shows that $\bar{d}$ is a metric. Note that *a Cauchy sequence is convergent if and only if it is equivalent to a constant sequence. In particular, a Cauchy sequence equivalent to a convergent one is also convergent.*

The map

$$\mathcal{I} : X \to \overline{X}, \quad x \mapsto C_{\langle x \rangle}$$

is an isometry. Its image can be identified with the collection of equivalence classes of convergent sequences. We can now state and prove the main result of this subsection.

Theorem 17.2.13 (Existence of completion)**.** *The above triplet $(\overline{X}, \bar{d}, \mathcal{I})$ is a completion of (X, d).*

Proof. Let us first prove that $\mathcal{I}(X)$ is dense in $\overline{X}$. Let $\underline{x} \in \boldsymbol{CS}(X)$, $\underline{x} = (x_n)_{n \in \mathbb{N}}$. For each $m \in \mathbb{N}$ we denote by $\underline{x}^m = (x^m_n)_{n \in \mathbb{N}}$ the constant sequence $\underline{x}^m = \langle x_m \rangle$, i.e.,

$$x^m_n = x_m \quad \forall n \in \mathbb{N}.$$

We will show that

$$\lim_{m \to \infty} \bar{d}\big(\, \underline{x}^m, \underline{x}\big) = 0.$$

Indeed, let $\varepsilon > 0$. Since $\underline{x}$ is a Cauchy sequence we deduce that there exists $N = N(\varepsilon) > 0$ such that

$$\forall m, n > N(\varepsilon) : \quad d(x^m_n, x_n) = d(x_m, x_n) < \varepsilon.$$

Letting $n \to \infty$ we deduce

$$\bar{d}\big(\underline{x}^m, \underline{x}\big) \leqslant \varepsilon \ \ \forall m > N(\varepsilon).$$

To prove that $\big(\overline{X}, \bar{d}\,\big)$ is complete we need to digress a bit.

We say that a sequence of points $(y_n)_{n\in\mathbb{N}}$ in a metric space (Y, δ) is *convenient* if

$$\delta(y_n, y_{n+1}) < \frac{1}{2^{n+5}}, \quad \forall n \in \mathbb{N}.$$

Note that if (y_n) is convenient, then for any $N > n$ we have

$$\delta(y_n, y_N) \leqslant \delta(y_n, y_{n+1}) + \cdots + \delta\big(\, y_{N-1}, y_N\,\big) \leqslant \sum_{k\geqslant n} \frac{1}{2^{k+5}} = \frac{1}{2^{n+4}}. \tag{17.2.6}$$

This proves that any convenient sequence is Cauchy. Clearly, any Cauchy sequence admits a convenient subsequence.

Lemma 17.2.14. *A Cauchy sequence is equivalent with any of its subsequences. In particular, a Cauchy sequence is convergent if and only if it has a convergent subsequence.*

Proof. Suppose that $\underline{x} = (x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and (x_{k_n}) is a subsequence. Then $k_n \geqslant n$ and, since $\underline{x}$ is Cauchy, we deduce that $\forall \varepsilon > 0$ there exists $N = N(\varepsilon) > 0$ such that

$$d(x_{k_n}, x_n) < \varepsilon, \quad \forall n \geqslant N(\varepsilon).$$

In other words,

$$\lim_{n\to\infty} d(x_{k_n}, x_n) = 0$$

so $\underline{x}$ is equivalent to the subsequence (x_{k_n}). The final conclusion follows from the fact that a Cauchy sequence equivalent to a convergent sequence is also convergent. □

Consider a Cauchy sequence $(C_m)_{m\in\mathbb{N}}$ in $\overline{X}$. For each m, pick a convenient Cauchy sequence $\underline{x}^m = (x^m_n)_{n\in\mathbb{N}}$ in X representing C_m. To show that C_m is convergent we will show that a subsequence of C_m is convergent. We will detect this subsequence of a sequence of sequences using a variation of *Cantor's diagonal trick*.[2]

By passing to subsequences we can assume that the Cauchy sequence (C_m) in $\overline{X}$ is also convenient. Since

$$\bar{d}(C_m, C_{m+1}) < \frac{1}{2^{m+5}}, \quad \forall m \in \mathbb{N}$$

we can find an increasing sequence $N_1 < N_2 < \cdots$ of natural numbers such that

$$\boxed{d\big(\, x^m_n, x^{m+1}_n\,\big) < \frac{1}{2^{m+4}}, \quad \forall n \geqslant N_m.} \tag{17.2.7}$$

Consider the sequence in X,

$$\underline{x}^* = (x^*_m), \ \ x^*_m := x^m_{N_m}.$$

[2] https://en.wikipedia.org/wiki/Cantor's_diagonal_argument

We will prove that

$$\underline{x}^* \in \boldsymbol{CS}(X), \tag{17.2.8a}$$

$$\lim_{m\to\infty} \bar{d}\big(\,C_m, C_{\underline{x}^*}\,\big) = 0. \tag{17.2.8b}$$

Observe that

$$d\big(\,x^*_m, x^*_{m+1}\,\big) = d\big(\,x^m_{N_m}, x^{m+1}_{N_{m+1}}\,\big) \leqslant d\big(\,x^m_{N_m}, x^m_{N_{m+1}}\,\big) + d\big(\,x^m_{N_{m+1}}, x^{m+1}_{N_{m+1}}\,\big)$$

(use (17.2.6) and (17.2.7))

$$\leqslant \frac{1}{2^{m+4}} + \frac{1}{2^{m+4}} = \frac{1}{2^{m+3}}.$$

In particular for $m < n$ we deduce

$$d(x^*_m, x^*_n) < \frac{1}{2^{m+2}}.$$

This proves (17.2.8a).

To prove (17.2.8b) we observe that the subsequence $(x^m_{N_k})_{k\in\mathbb{N}}$ of $\underline{x}^m$ is equivalent to $\underline{x}^m$ so

$$\bar{d}\big(\,C_m, C_{\underline{x}^*}\,\big) = \lim_{k\to\infty} d\big(\,x^m_{N_k}, x^*_k\,\big) = \lim_{k\to\infty} d\big(\,x^m_{N_k}, x^k_{N_k}\,\big).$$

Suppose that $m < k$. Then for $m \leqslant j < k$, since $N_k > N_j$, we deduce from (17.2.7) that

$$d\big(\,x^j_{N_k}, x^{j+1}_{N_k}\,\big) \leqslant \frac{1}{2^{j+4}}.$$

We deduce that for $k > m$

$$d\big(\,x^m_{N_k}, x^k_{N_k}\,\big) \leqslant \sum_{j=m}^{k-1} d\big(\,x^j_{N_k}, x^{j+1}_{N_k}\,\big) \leqslant \frac{1}{2^{m+3}}.$$

Hence

$$\bar{d}\big(\,C_m, C_{\underline{x}^*}\,\big) = \lim_{k\to\infty} d\big(\,x^m_{N_k}, x^*_k\,\big) \leqslant \frac{1}{2^{m+3}}.$$

This proves (17.2.8b) and completes the proof of Theorem 17.2.13. □

Proposition 17.2.15. *Suppose that $(X, \|-\|)$ is a $\mathbb{K}$-normed vector space, $\mathbb{K} = \mathbb{R}, \mathbb{C}$, and $\big(\overline{X}, \bar{d}, \mathfrak{I}\big)$ is a completion of X with respect to the metric defined by the norm. We identify as usual X with the subset $\mathfrak{I}(X)$ of $\overline{X}$ so that $\mathfrak{I}(\underline{x}) = x$, $\forall x \in X$. Then $\overline{X}$ has a unique vector space structure such that the map $\mathfrak{I} : X \to \overline{X}$ is linear and the map*

$$\overline{X} \ni \bar{x} \mapsto \|\bar{x}\|_* := \bar{d}\big(\,\bar{x}, 0\,\big) \in [0,\infty)$$

is a norm on $\overline{X}$.

Proof. Let $\bar{x}, \bar{y} \in \overline{X}$. Then there exist sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ in X that converge in $\bar{X}$ to $\bar{x}$ and respectively $\bar{y}$. Observing that

$$\bar{d}\big(\,(x_n + y_n), (x_m + y_m)\,\big) = \|(x_n + y_n) - (x_m + y_m)\| \leqslant \|x_n - x_m\| + \|y_n - y_m\|$$

we deduce that the sequence $\big(\,x_n + y_n\,\big)_{n\in\mathbb{N}}$ is Cauchy and thus it converges in $\overline{X}$.

Note that if $(x_n')_{n\in\mathbb{N}}$ and $(y_n')_{n\in\mathbb{N}}$ are other sequences in X that converge in the metric $\bar{d}$ to $\bar{x}$, then $(x_n'+y_n')_{n\in\mathbb{N}}$ is also convergent and

$$\lim_{n\to\infty}\bar{d}\big(\,(x_n+y_n),(x_n'+y_n')\,\big)=\lim_{n\to\infty}\|(x+y_n)-(x_n'+y_n')\|=0.$$

We denote by $\bar{x}\bar{+}\bar{y}$ the common limit of the sequences $(x+y_n)$ and $(x_n'+y_n')$.

Similarly, for $t\in\mathbb{K}$, we denote by $t\bar{x}$ the common limit of the sequences (tx_n) and (tx_n').

Note that if $\bar{x},\bar{y}\in X$, then choosing (x_n) and (y_n) to be constant sequences $X_n=x$, $y_n=y$, $\forall n$, we deduce that $\bar{x}\bar{+}\bar{y}=\overline{(x+y)}$, where the second "+" denotes the usual addition operation on X. This proves that $\overline{X}$ can be equipped with a structure of vector space such that the map $\mathcal{I}$ is linear.

Let us show that $\|x\|_*$ is a norm. The equality $\|t\bar{x}\|_*=|t|\|\bar{x}\|_*$ follows from the equality

$$d(tx_n,0)=\|tx_n\|=|t|d(x_n,0)$$

by letting $n\to\infty$. To verify the triangle inequality observe have

$$\|\bar{x}+\bar{y}\|_*=\lim_{n\to\infty}\bar{d}\big(\,(x_n+y_n),0\,\big)=\lim_{n\to\infty}\|x_n+y_n\|$$

$$\leqslant\lim_{n\to\infty}\big(\,\|x_n\|+\|y_n\|\,\big)=\bar{d}\big(\,\bar{x},0\,\big)+\bar{d}\big(\,\bar{y},0\,\big)=\|\bar{x}\|_*+\|\bar{y}\|_*.$$

If $\hat{+}$ is another addition operation on $\overline{X}$ such that $\mathcal{I}$ is continuous and $\|-\|_*$ is a norm then

$$\|\bar{x}\hat{+}\bar{y}-(\bar{x}\bar{+}\bar{y})\|_*=\lim_{n\to\infty}\|\bar{x}\hat{+}\bar{y}-(x_n+y_n)\|_*$$

$$=\lim_{n\to\infty}\|\bar{x}\hat{+}\bar{y}-(x_n\hat{+}y_n)\|_*\leqslant\lim_{n\to\infty}\big(\,(\|\bar{x}-x_n\|_*+\|\bar{y}-y_n\|_*\,\big)=0.$$

□

Definition 17.2.16. The Banach space $\big(\overline{X},\|-\|_*\big)$ constructred in Proposition 17.2.15 is called the *completion* of the normed space $(X,\|-\|)$. □

17.2.3. Applications. The completeness of a space is essentially an existence statement: a sequence that looks like it ought to have a limit does indeed have a limit. The applications we have in mind are more special existence statements.

Definition 17.2.17. Suppose that X is a set and $T:X\to X$ is a self-map of X.

(i) A *fixed point* of T is a point $x_*\in X$ such that $Tx_*=x_*$.

(ii) For any $n\in\mathbb{N}$ we define $T^n:X\to X$

$$T^n:=\underbrace{T\circ\cdots\circ T}_{n}.$$

(iii) We say that T is a *contraction* with respect to a metric d on X if there exists $c\in(0,1)$ such that T is c-Lipschitz, i.e.,

$$d\big(\,Tx_0,Tx_1\,\big)\leqslant cd(x_0,x_1),\quad\forall x_0,x_1\in X.$$

□

Theorem 17.2.18 (Banach's fixed point)**.** *Suppose that $T : X \to X$ is a contraction on the* complete *metric space (X, d). Then the following hold.*

(i) *The map T has a unique fixed point x_*.*

(ii) *For any $x \in X$,*

$$\lim_{n\to\infty} T^n x = x_*.$$

Proof. (i) Fix $c \in (0, 1)$ such that T is c-Lipschitz. If x_*, x'_* are fixed points of T then

$$d(x_*, x'_*) = d(Tx_*, Tx'_*) \leqslant cd(x_*, x'_*).$$

Since $c \in (0, 1)$ we deduce $d(x_*, x'_*) = 0$.

(ii) Observe first that T^n is c^n-Lipschitz. Indeed, this is true for $n = 1$ and the general case follows inductively from

$$d\big(T^{n+1}x_0, T^{n+1}x_1\big) \leqslant cd\big(T^n x_0, T^n x_1\big),\ \ \forall x_0, x_1 \in X.$$

Next, observe that for any $x \in X$ and any $k \in \mathbb{N}$ we have

$$d\big(x, T^k x\big) \leqslant d(x, Tx) + d(Tx, T^2 x) + \cdots + d\big(T^{k-1}x, T^k x\big)$$

$$\leqslant d(x, Tx) + cd(x, Tx) + \cdots + c^{k-1}d(x, Tx) < d(x, Tx)\sum_{n=0}^{\infty} c^n = \frac{1}{1-c}d(x, Tx).$$

Now observe that for any $m, n \in \mathbb{N}$, $m < n$, and any $x \in X$ we have

$$d\big(T^m x, T^n x\big) \leqslant c^m d\big(x, T^{n-m}x\big) \leqslant \frac{c^m}{1-c}d\big(x, Tx\big).$$

This proves that the sequence $x_n := T^n x$, $n \in \mathbb{N}$, is Cauchy and thus converges since (X, d) is complete. We denote by x_* its limit. Observe that

$$x_{n+1} = Tx_n,\ \ \forall n \in \mathbb{N}. \tag{17.2.9}$$

Letting $n \to \infty$ in the above equality and taking into account that T is continuous we deduce

$$x_* = Tx_*,$$

i.e., x_* is a fixed point of T, *the unique* fixed point. □

Theorem 17.2.19 (Baire)**.** *Suppose that (X, d) is a* complete *metric space and $(U_n)_{n\in\mathbb{N}}$ is a sequence of nonempty dense open sets. Then their intersection is also dense. More precisely, for every $x \in X$ and $r > 0$ we set*

$$\overline{B}_r(x) := \big\{\, x' \in X;\ \ d(x, x') \leqslant r \,\big\}.$$

Then, $\forall x_0 \in X$ and $r_0 > 0$

$$\overline{B}_{r_0}(x_0) \cap \left(\bigcap_{n\in\mathbb{N}} U_n\right) \neq \varnothing.$$

Proof. We construct inductively a sequence $(x_n)_{n\in\mathbb{N}}$ in X as follows. Choose

$$x_1 \in U_1 \cap B_{r_0}(x_0)$$

and a radius $r_1 < \frac{1}{2}$ such that

$$\overline{B}_{r_1}(x_1) \subset U_1 \cap B_{r_0}(x_0).$$

Since U_2 is dense, the open set $B_{r_1}(x_1) \cap U_2$ is nonempty. Choose $x_2 \in B_{r_1}(x_1) \cap U_2$ and $r_2 < \frac{1}{4}$ such that

$$\overline{B}_{r_2}(x_2) \subset B_{r_1}(x_1) \cap U_2.$$

We proceed inductively and we construct a sequence of points $(x_n)_{n\in\mathbb{N}}$ and radii $r_n < \frac{1}{2^n}$ such that

$$\overline{B}_{r_n}(x_n) \subset B_{r_{n-1}}(x_{n-1}) \cap U_n, \ \ \forall n \geqslant 2.$$

Observe that

$$d(x_{n-1}, x_n) < r_{n-1}.$$

This proves that for any $m < n$ we have

$$d(x_m, x_n) \leqslant r_m + \cdots + r_{n-1} < \frac{1}{2^m} + \cdots + \frac{1}{2^{n-1}} < \frac{1}{2^{m-1}}.$$

This proves that the sequence (x_n) is Cauchy, and thus convergent. We denote by x_* its limit. Note that for any $n > m \geqslant 0$ we have $x_n \in \overline{B}_{r_m}(x_m)$. Since $\overline{B}_{r_m}(x_m)$ is closed we deduce $x_* \in \overline{B}_{r_m}(x_m) \subset U_m \cap B_{r_0}(x_0)$, $\forall m \in \mathbb{N}$. □

Definition 17.2.20 (Baire category)**.** A metric space X is said to be *of the first category* if it is the union of a countable collection of closed sets with empty interiors. Otherwise it is called *of the second category.* □

A set in a metric space is called *nowhere dense* if its closure has empty interior. A set is called *meagre* if it is the union of countably many nowhere dense sets. Thus the sets of first category are meagre. Clearly, any subset of a meagre set is also meagre. Using the above terminology we can rephrase Baire's theorem as follows.

Corollary 17.2.21 (Baire)**.** *A complete metric space* (X, d) *is of second category, i.e., non-meagre.*

Proof. We argue by contradiction. Suppose that X is the union of countably many nowhere dense sets X_n. Then X is also the union of the closed sets $C_n = \boldsymbol{cl}(X_n)$ with empty interiors. The complements $U_n = X\backslash C_n$ are open and *dense*. Indeed if a set U_n were not dense, then it will be disjoint from a small open ball and thus, that ball would be contained in C_n. This is impossible since C_n has empty interior.

Since the union of C_n's is X we deduce that the sets U_n have empty intersection. This contradicts Theorem 17.2.19. □

Intuitively, *meagre* sets are “very thin”. The next result clarifies this intuition: the complement of a meagre set is dense so it has to be quite large.

Proposition 17.2.22. *Suppose that (X, d) is a complete metric space and $M \subset X$ is a meagre subset. Then the complement $X \backslash M$ is dense in X.*

Proof. Let S be a meagre set. Thus

$$S = \bigcup_{n \in \mathbb{N}} S_n,$$

where $C_n := \boldsymbol{cl}(S_n)$ has empty interior $\forall n$. Let M denote the union of closed sets $(C_n)_{n \in \mathbb{N}}$. Then

$$X \backslash S \supset X \backslash M = \bigcap_{n \in \mathbb{N}} U_n, \;\; U_n = X \backslash C_n.$$

Since C_n has empty interior its complement U_n is open and dense, Theorem 17.2.19 shows that $\bigcap_n U_n$ intersects any open ball in X and thus it is dense. □

Observe that U is an open and dense subset of a metric space if and only if its complement is closed and has empty interior. Baire’s theorem can be equivalently reformulated as follows.

Corollary 17.2.23. *If (X, d) is a complete metric space and $(C_n)_{n \in \mathbb{N}}$ is a sequence of closed sets such that*

$$X = \bigcup_{n \in \mathbb{N}} C_n,$$

then at least one of the closed sets C_n has nonempty interior. □

We will present several fundamental applications of Baire’s theorem when we discuss functional analysis. Until then we discuss several surprising applications to one-variable calculus. The next example was first discussed in [**3**].

Example 17.2.24. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a smooth function such that

$$\forall x \in \mathbb{R}, \;\; \exists n \in \mathbb{N} : \;\; f^{(n)}(x) = 0. \tag{17.2.10}$$

Clearly polynomial functions have this property. We want to show that *only* the polynomial functions have this property, i.e., if f satisfies (17.2.10)m then f is a polynomial, i.e.,

$$\exists n \in \mathbb{N}, \;\; \forall x \in \mathbb{R} : \;\; f^{(n)}(x) = 0. \tag{17.2.11}$$

We should pause to compare the differences between (17.2.10) and (17.2.11): they differ only in the order of quantifiers. However, proving that this switch in order leads to an equivalent statement requires quite a bit of imagination.

Denote by $\mathcal{I}$ the collection of all the open intervals (a, b) such that, the restriction of f to (a, b) is a polynomial. Observe that

$$\forall I, J \in \mathcal{Y}, \;\; I \cap J \neq \varnothing \Rightarrow I \cup J \in \mathcal{I}.$$

Indeed, if $f|_I$ is a polynomial P_I and $f|_J$ is a polynomial P_J, then $P_I = P_J = f$ on $I \cap J$. Since two polynomials coinciding on an open interval coincide everywhere we deduce that $P_I = P_J$ and $f|_{I \cup J}$ is also a polynomial.

Hence, the union of an increasing family of intervals in $\mathcal{I}$ is also on interval in $\mathcal{I}$. Let $\Omega \subset \mathbb{R}$ be the union of all the intervals in $\mathcal{I}$. Clearly Ω is open. The above discussion shows that Ω is a union of pairwise disjoint intervals in $\mathcal{I}$

$$\Omega = \bigcup_{n=1}^{N} I_n, \quad I_n = (a_n, b_n) \in \mathcal{I}, \quad 1 \leqslant N \leqslant \infty. \tag{17.2.12}$$

We want to show that $\Omega = \mathbb{R}$. We begin by first proving that it is dense. More precisely, we will show that $\Omega \cap [a, b] \neq \varnothing$, $\forall a < b$.

Indeed, set

$$E_n := \big\{ x \in [a, b]; \;\; f^{(n)}(x) = 0 \,\big\}.$$

Clearly E_n is closed and (17.2.10) shows that

$$[a, b] = \bigcup_n E_n.$$

Baire's theorem applied to the complete metric space $[a, b]$ implies that at least one of the sets E_n has nonempty interior. Thus, there exists an open interval $I \subset E_n$, meaning $f^{(n)}(x) = 0$, $\forall x \in I$, so that $f|I$ is a polynomial of degree $< n$. This proves that Ω is dense in $\mathbb{R}$. Set $X := \mathbb{R}\backslash\Omega$. Thus X is closed, with empty interior. We want to show that X is actually empty.

Let us first observe that X does not contain isolated points. Indeed, suppose that $x_0 \in X$ were an isolated point of X. Then there exists an open interval (a, b) such that $(a, b) \cap X = \{x_0\}$. Then

$$(a, x_0), \;\; (x_0, b) \subset \Omega,$$

so each of these intervals is contained in one of the intervals I_n of the decomposition (17.2.12). Thus for some $m, n \in \mathbb{N}$

$$f^{(m)}(x) = 0, \;\; \forall x \in (a, x_0), \;\; f^{(n)}(x), \;\; \forall x \in (x_0, b).$$

If $p = \max(m, n)$, then

$$f^{(p)}(x) = 0, \;\; \forall x \in (a, b)\backslash\{x_0\}.$$

By continuity we conclude $f^{(p)}(x) = 0$, $\forall x \in (a, b)$ and therefore $(a, b) \subset \Omega$ so $x_0 \notin X$.

We argue by contradiction. Suppose that $X \neq \varnothing$. For $m \in \mathbb{N}$ we set

$$F_m := \big\{ \, x \in X; \;\; f^{(m)}(x) = 0 \,\big\}.$$

Clearly

$$X = \bigcup_{m \in \mathbb{N}} F_m$$

Applying Baire's theorem to X equipped with the induced metric we deduce that there exists $m \in \mathbb{N}$ and an interval (a, b) such that

$$\varnothing \neq (a, b) \cap X \subset F_m. \tag{17.2.13}$$

Note that given $x \in (a, b) \cap X$ there exists a sequence $(x_k) \in (a, b) \cap X$ such that $x_k \neq x$, $\forall k$ and

$$\lim_k x_k = x.$$

Thus

$$f^{(m+1)}(x) = \lim_k \frac{f^{(m)}(x_k) - f^{(m)}(x)}{x_k - x} = 0$$

and we deduce inductively that

$$f^{(n)}(x) = 0, \;\; \forall x \in (a, b) \cap X, \;\; \forall n \geqslant m. \tag{17.2.14}$$

This implies that $(a, b) \cap X \neq (a, b)$ because, if it did, the function f would be a polynomial of degree $< m$ on (a, b) and thus $(a, b) \subset \Omega = \mathbb{R} \backslash X$.

We want to prove that

$$f^{(n)}(x) = 0, \;\; \forall x \in (a, b) \cap \Omega, \;\; \forall n \geqslant m. \tag{17.2.15}$$

We have

$$(a, b) \cap \Omega = \bigcup_{n \in \mathbb{N}} I_n \cap (a, b).$$

Let n such that $J = I_n \cap (a, b) \neq \varnothing$. Note that (a, b) is not contained in I_n because $X \cap (a, b) \neq \varnothing$. Thus J is an interval of the form (c, d) and at least one of the endpoints belongs to $X \cap (a, b)$. For simplicity, we assume $c \in X \cap (a, b)$. On the interval (c, d) the function f is a polynomial of degree k. Thus, on this interval the k-th derivative of f is a nonzero constant. By continuity $f^{(k)}(c) \neq 0$. Since $c \in X$, we deduce from (17.2.14) that $k < m$. Thus, on (c, d) the function f is a polynomial of degree m so that in satisfies (17.2.15).

We conclude that $f^{(m)}(x) = 0$ on (a, b) so f is a polynomial on this interval. In other words this interval is contained in Ω and thus is disjoint from X. This contradicts (17.2.13) so f is indeed a polynomial function on $\mathbb{R}$. □

We conclude this subsection with a famous application of Baire's theorem. It gives a rather surprising answer to a famous question: do that there exist continuous functions that are nowhere differentiable? As mentioned in Remark 7.1.8, Weierstrass constructed the first example of such function. Later on many more examples were constructed. In 1931 Banach and Mazurkiewicz independently offered a surprising answer to this question: yes there are, a lot, so many so that any continuous function can approximated arbitrarily well by such functions. More precisely they proved that the set of continuous nowhere differentiable functions is dense in the Banach space $\big(C([0, 1]), \| - \|_\infty \big)$. Surprisingly, their proof does not produce any concrete examples of such pathological functions. They must exist by virtue of deeper principles. Moreover, their approach requires working in infinite dimensional spaces.

Theorem 17.2.25 (Banach-Mazurkiewicz). *The collection $\mathcal{W}$ of continuous nowhere differentiable functions $f:[0,1]\to\mathbb{R}$ is dense in the Banach space $\big(C([0,1]),\|-\|_\infty\big)$.*

Proof. For simplicity, we set $\mathscr{C}:=C([0,1])$. Here is the strategy. We will construct a meagre subset $\mathcal{A}\subset\mathscr{C}$ such that

$$\mathscr{C}\backslash\mathcal{A}\subset\mathcal{W}. \tag{17.2.16}$$

By Proposition 17.2.22 the set $\mathscr{C}\backslash\mathcal{A}$ is dense in $\mathscr{C}$ and, a fortiori, so is $\mathcal{W}$. Set

$$\mathcal{D}:=\mathscr{C}\backslash\mathcal{W}.$$

Thus, $\mathcal{D}$ consists of continuous functions $[0,1]\to\mathbb{R}$ that are differentiable at some point in $[0,1]$. The condition (17.2.16) is equivalent to the existence of a meagre set $\mathcal{A}$ such that

$$\mathcal{D}\subset\mathcal{A}.$$

First some notation. For each $f\in\mathscr{C}$ and $t\in[0,1]$ we set

$$D_tf:=\sup_{h\neq 0}\left|\frac{f(t+h)-f(t)}{h}\right|\in[0,\infty].$$

For each $n\in\mathbb{N}$ we set

$$\mathcal{A}_n:=\big\{\,f\in\mathscr{C};\;\;\exists t\in[0,1]:\;\;D_tf\leqslant n\,\big\}.$$

Finally, define

$$\mathcal{A}:=\bigcup_{n\in\mathbb{N}}\mathcal{A}_n.$$

The proof of the theorem will be completed in three steps.

Step 1. $\mathcal{D}\subset\mathcal{A}$.

Step 2. For each $n\in\mathbb{N}$ the set $\mathcal{A}_n$ is closed in $(\mathscr{C},\|-\|_\infty)$.

Step 3. For each $n\in\mathbb{N}$, the interior of the set $\mathcal{A}_n$ in $(\mathscr{C},\|-\|_\infty)$ is empty.

Baire's theorem shows that $\mathcal{A}\neq\mathscr{C}$ and, moreover, $\mathscr{C}\backslash\mathcal{A}$ is dense. Any function in $\mathscr{C}\backslash\mathcal{A}$ is nowhere differentiable.

Proof of Step 1. We have to show that for any $f\in\mathcal{D}$ there exists $n\in\mathbb{N}$ and $t\in[0,1]$ such that $D_tf\leqslant n$. To see this, suppose that $t\in[0,1]$ is a point where f is differentiable. Consider the compact interval $J=[-t,1-t]$ and define $q:J\to\mathbb{R}$

$$q(h)=\begin{cases}\frac{f(t+h)-f(t)}{h}, & h\neq 0,\\ f'(t), & h=0.\end{cases}$$

The function q is clearly continuous so it is bounded. Hence

$$D_tf=\sup_{h\in J}|q(h)|<\infty.$$

Hence $f\in\mathcal{A}_n$, $\forall n>D_tf$.

Proof of Step 2. Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{A}_N$ that converges uniformly to a function $f \in \mathscr{C}$. We want to prove that $f \in \mathcal{A}_N$. For each n choose a point $t_n \in [0,1]$ such that $D_{t_n} f_n \leqslant N$. A subsequence of (t_n) is convergent so, after restricting to this subsequence, we can assume that

$$\lim_{n\to\infty} t_n = t_* \in [0,1].$$

For $h \neq 0$ we have

$$\begin{aligned} |f(t_*+h) - f(t_*)| &\leqslant |f(t_*+h) - f_n(t_*+h)| + |f_n(t_*+h) - f_n(t_n)| \\ &\quad + |f_n(t_n) - f_n(t_*)| + |f_n(t_*) - f(t_*)| \\ &\leqslant \|f - f_n\|_\infty + D_{t_n} f_n |t_* + h - t_n| + D_{t_n} f_n |t_n - t_*| + \|f_n - f\|_\infty \\ &= 2\|f - f_n\|_\infty + D_{t_n} f_n \big(|t_* + h - t_n| + |t_n - t_*| \big) \end{aligned}$$

Hence, for any $h \neq 0$ and any $n \in \mathbb{N}$

$$\begin{aligned} \frac{|f(t_*+h) - f(t_*)|}{|h|} &\leqslant \frac{2\|f - f_n\|_\infty}{|h|} + D_{t_n} f_n \cdot \frac{|t_* + h - t_n| + |t_n - t_*|}{|h|} \\ &\leqslant \frac{2\|f - f_n\|_\infty}{|h|} + N \cdot \frac{|t_* + h - t_n| + |t_n - t_*|}{|h|}. \end{aligned}$$

Note, that for fixed $h \neq 0$ we have

$$\lim_{n\to\infty} \frac{|t_* + h - t_n| + |t_n - t_*|}{|h|} = 1.$$

Hence, for any $\varepsilon > 0$ and $h \neq 0$ we can find $n = n(\varepsilon, h)$ sufficiently large such that

$$\frac{2\|f - f_n\|_\infty}{|h|} < \frac{\varepsilon}{2}, \quad \frac{|t_* + h - t_n| + |t_n - t_*|}{|h|} < 1 + \frac{\varepsilon}{2N}.$$

Hence, for any $h \neq 0$ and any $\varepsilon > 0$ we have

$$\frac{|f(t_*+h) - f(t_*)|}{|h|} < N + \varepsilon,$$

so that $D_{t*} f \leqslant N$, i.e., $f \in \mathcal{A}_N$. This proves that $\mathcal{A}_N$ is closed.

Step 3. $\boldsymbol{int}\,\mathcal{A}_N = \varnothing$. Let $f \in \mathcal{A}_N$. We will show that for any $N > 0$ and any $\varepsilon > 0$ there exists a continuous function $f_\varepsilon : [0,1] \to \mathbb{R}$ such that

$$\|f - f_\varepsilon\|_\infty < \varepsilon, \quad f_\varepsilon \notin \mathcal{A}_N, \quad \forall n. \tag{17.2.17}$$

This follows from the following elementary lemma.

Lemma 17.2.26. *Suppose that $g : [a,b] \to \mathbb{R}$ is a continuous function. Then for any $C > 0$ and any $\varepsilon > 0$ there exists a continuous piecewise linear function*

$$\bar{g} = \bar{g}^{\varepsilon, C} : [a,b] \to \mathbb{R}$$

satisfying the following properties.

$$\bar{g}(a) = g(a), \quad \bar{g}(b) = g(b). \tag{17.2.18a}$$

$$\|g-\bar{g}\|_\infty < \operatorname{osc}\big(f,[a,b]\big)+\frac{\varepsilon}{2}. \tag{17.2.18b}$$

$$D_t\bar{g} \geqslant C,\ \ \forall t\in[a,b]. \tag{17.2.18c}$$

Proof of Lemma 17.2.26. Set

$$m := \inf_{t\in[a,b]} f(t),\ \ M := \sup_{t\in[a,b]} f(t),$$

so that

$$\operatorname{osc}(f,[a,b]) = M-m.$$

Fix n sufficiently large so that

$$\frac{n\varepsilon}{(b-a)} > C.$$

Subdivide the interval $[a,b]$ into $2n$ intervals of equal size and set

$$t_k = a+\frac{k(b-a)}{2n},\ \ k=0,1,\dots,2n.$$

$$y_0=f(a), y_1=M+\frac{\varepsilon}{4}, y_2=m-\frac{\varepsilon}{2},\ \ y_{2n-2}=m-\frac{\varepsilon}{2}, y_{2n-1}=M+\frac{\varepsilon}{2}, y_{2n}=f(b).$$

Observe that

$$\frac{|y_k-y_{k-1}|}{t_k-t_{k-1}} > \frac{\varepsilon/2}{(b-a)/2n} = \frac{n\varepsilon}{(b-a)} > C.$$

Denote by $\bar{g}$ the continuous piecewise linear function $[a,b]\to\mathbb{R}$ uniquely determined by the following requirements; see Figure 17.2

- $g(t_k)=y_k$, $\forall k=0,1,2,\dots,2n$.
- $\bar{g}$ is linear on each of the intervals $[t_{k-1},t_k]$, i.e.,

$$g(y) = y_{k-1}+\frac{y_k-y_{k-1}}{t_k-t_{k-1}}(t-t_k),\ \ \forall t\in[t_{k-1},t_k].$$

By construction

$$m-\frac{\varepsilon}{2} \leqslant \bar{g}(t) \leqslant M+\frac{\varepsilon}{2},$$

so that

$$\big|\,g(t)-\bar{g}(t)\,\big| \leqslant (M-m)+\frac{\varepsilon}{2} = \operatorname{osc}\big(f,[a,b]\big)+\frac{\varepsilon}{2},\ \ \forall t\in[a,b].$$

Clearly

$$D_t\bar{g} \geqslant \frac{n\varepsilon}{(b-a)} > K,\ \ \forall t\in[a,b].$$

□

We can now prove (17.2.17). Fix $N\in\mathbb{N}$ and $\varepsilon>0$. Since f is uniformly continuous, there exists $n\in\mathbb{N}$ sufficiently large such that

$$\operatorname{osc}\big(f,[(k-1)/n,k/n]\big) < \frac{\varepsilon}{2},\ \ \forall k=1,\dots,n.$$

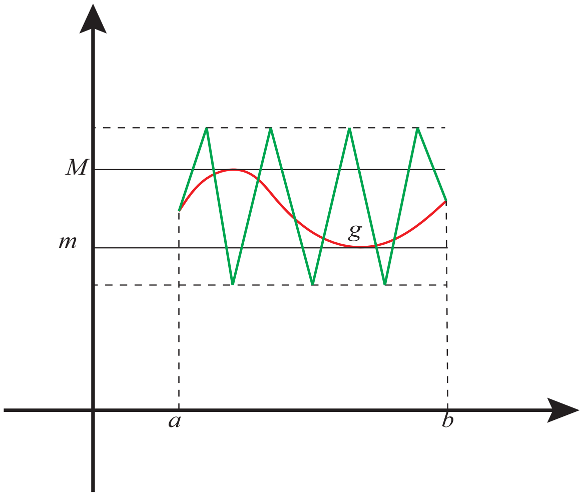

Figure 17.2. *The graph of $\bar{g}(x)$ is a highly oscillating zig-zag.*

For each $k = 1, \dots, n$ we denote by g_k the restriction of f to the interval $I_k = [(k-1)/n, k/n]$. Denote by f_k the function $\bar{g}_k^{\varepsilon,N}$ as in Lemma 17.2.26. Hence

$$\sup_{t \in I_k} \bigl| f(t) - f_k(t) \bigr| < \frac{\varepsilon}{2} + \operatorname{osc}(f, I_k) < \varepsilon, \;\; D_t f_k > N, \;\; \forall t \in I_k.$$

Now define $f_\varepsilon : [0,1] \to \mathbb{R}$ by setting

$$f_\varepsilon(t) = f_k(t), \;\; \forall t \in I_k.$$

□

17.3. Compactness

The concept of compactness in $\mathbb{R}^n$ has a correspondent in the more general case of metric spaces. In this general context compactness is a desired, but less frequent and harder to detect occurrence.

17.3.1. Compact metric spaces. As in the case of $\mathbb{R}^n$, an *open cover* of a set $S \subset X$ of a metric space is a collection $\mathscr{C}$ of open subsets of X whose union contains the subset S. A *subcover* of an open cover $\mathscr{C}$ of S is a subcollection $\mathscr{C}'$ of $\mathscr{C}$ that is itself an open cover of S.

Definition 17.3.1. Fix a metric space (X, d).

(i) The space is called *compact* if it satisfies the *Heine-Borel* (or *HB*) *property*, i.e., *any* open cover of X admits a *finite* subcover.

(ii) The space is called *totally bounded* if, for any $\varepsilon > 0$, the space X can be covered by finitely many open balls of radius ε.

(iii) The space $X$is called *sequentially compact* if it satisfies the *Bolzano-Weierstrass property*, i.e., if any sequence in X admits a convergent subsequence.

□

Remark 17.3.2. Observe that the Heine-Borel property is equivalent to the following dual condition

$\boxed{HB^*}$. *Any collection of closed subsets of X with empty intersection contains a $\underline{\text{finite}}$ subcollection with empty intersection.*

Indeed, suppose that X satisfies the HB property. If $(C_i)_{i\in I}$ is a collection of closed sets such that

$$\bigcap_{i\in I} C_i = \varnothing,$$

then the collection of open subsets

$$\big\{\, U_i = X\backslash C_i;\ \ i \in I\,\big\}$$

is an open cover of X so there exists a finite subset $J \subset I$ such that

$$\bigcup_{j\in J} U_j = X.$$

Clearly the finite subfamily $(C_j)_{j\in J}$ has trivial intersection. This proves $HB \Rightarrow HB^*$. To prove the reverse implication run the above argument in reverse. □

Definition 17.3.3. Let (X, d) be a metric space and $\varepsilon > 0$. An ε-net in X is a subset S such that

$$\forall x \in X,\ \ \exists s \in S:\ \ d(x, s) < \varepsilon.$$ □

We see that a metric space is totally bounded iff, for any $\varepsilon > 0$, the space X contains a *finite* ε-net.

Theorem 17.3.4 (Characterization of compactness)**.** *Let (X, d) be a metric space. The following statements are equivalent.*

(i) *The space (X, d) is compact.*

(ii) *The space (X, d) is sequentialy compact.*

(iii) *The space (X, d) is complete and totally bounded.*

Proof. We follow closely the approach in [12, Sec. 3.16].

(i) ⇒ (ii) Suppose that (x_n) is a sequence in X. Denote by C_n the closure of the set

$$\big\{\, x_n, x_{n+1}, x_{n+2}, \dots\,\big\}.$$

Observe that

$$\bigcap_{n\in\mathbb{N}} C_n \neq \varnothing.$$

Indeed, if that were not the case, then the Heine-Borel property would imply (see Remark 17.3.2) that $C_1 \cap C_2 \cap \cdots \cap C_N = \varnothing$ for some N. This is impossible since

$$\varnothing \neq C_N = C_1 \cap \cdots \cap C_N.$$

Let

$$x_* \in \bigcap_{n\in\mathbb{N}} C_n.$$

Thus, $x_* \in \boldsymbol{cl}\{x_n, x_{n+1}, \dots\}$ for any $n > 0$. Thus, for any $n > 0$ exists $m_n > n$ such that $d(x_*, x_{m_n}) < \frac{1}{n}$. Now define inductively

$$n_1 = m_1,\;\; n_2 = m_{n_1},\;\; n_{k+1} = m_{n_k}.$$

The sequence (n_k) is strictly increasing and

$$d(x_{n_{k+1}}, x_*) < \frac{1}{m_{n_k}} < \frac{1}{n_k}$$

(ii) ⇒ (iii) Suppose that (x_n) is a Cauchy sequence. The Bolzano-Weierstrass property implies that it has a convergent subsequence and, according to Lemma 17.2.14, it must be convergent. This proves that X is complete.

To prove that X is totally bounded we argue by contradiction. Suppose X is not totally bounded. Thus, there exists $r > 0$ such that X cannot be covered by finitely many balls of radius r. We construct inductively a sequence (x_n) in X as follows. Choose x_1 arbitrarily. Then choose

$$x_2 \in X\backslash B_r(x_1),\;\; x_3 \in X\backslash\Big(B_r(x_1) \cup B_r(x_2)\Big), \dots, x_{n+1} \in X\backslash \bigcup_{k=1}^{n} B_r(x_k).$$

The resulting sequence has the property that $d(x_m, x_n) \geqslant r$, $\forall m \neq n$. In particular, none of its subsequences is Cauchy, hence none of its subsequences is convergent. This violates the Bolzano-Weierstrass property.

(iii) ⇒ (i) We will need the following simple fact.

Lemma 17.3.5. *A totally bounded metric space is bounded, i.e.,* $\exists C > 0$ *such that*

$$d(x, x') < C,\;\; \forall x, x' \in X.$$

Proof of Lemma 17.3.5. Cover X by finitely many open balls of radius 1

$$X = B_1(x_1) \cup B_1(x_2) \cup \cdots \cup B_1(x_n).$$

Set

$$\delta := \max_{1\leqslant i,j\leqslant n} d(x_i, x_j)$$

Let $x, x' \in X$. Then there exist i, j such that $x \in B_1(x_i)$, $x' \in B_1(x_j)$ so that

$$d(x, x') \leqslant d(x, x_i) + d(x_i, x_j) + d(x_j, x') < \delta + 2.$$

□

We argue by contradiction. Suppose that X is complete and totally bounded, yet it does not satisfy the Heine-Borel property. Thus, there exists an open cover $(U_i)_{i\in I}$ of X that contains no finite subcover. Fix $x_0 \in X$. Since X is bounded, there exists $r > 0$ and $x_0 \in X$ such that $X = B_r(x_0)$. Set $B_0 := B_r(x_0)$.

Next, fix a finite cover by balls of radius $r/2$. One of these balls cannot be covered by finitely many of the open sets U_i. We denote it by $B_1 = B_{r/2}(x_1)$. The ball B_1 can be covered by finitely many balls of radius $r/4$ and one of these balls $B_2 = B_{r/4}(x_2)$ intersects B_1 and cannot be covered by finitely many of the U_i's. Note that

$$d(x_1, x_2) < \frac{r}{2} + \frac{r}{4}.$$

Inductively, we find a sequence of open balls $B_k = B_{r/2^k}(x_k)$ such that, none of them can be covered by finitely many of the U_i's and $B_k \cap B_{k-1} \neq \varnothing$, $\forall k \geqslant 2$. We deduce inductively that

$$d(x_{k-1}, x_k) < \frac{r}{2^{k-1}} + \frac{r}{2^k} = \frac{3r}{2^k}.$$

Observe that the sequence (x_k) is Cauchy because $\forall n > m$ we have

$$d(x_n, x_m) \leqslant d(x_m, x_{m+1}) + \cdots + d(x_{n-1}, x_n) < 3r\big(2^{-m-1} + \cdots + 2^{-n-1}\big) < 3r2^{-m}.$$

Hence, the sequence (x_k) converges to a point $x_* \in X$. Thus there exists $i_0 \in I$ such that $x_* \in U_{i_0}$. In particular, there exists $\varepsilon > 0$ such that $B_\varepsilon(x_*) \subset U_{i_0}$ now choose k sufficiently large so that

$$r2^{-k} + d(x_k, x_*) < \frac{\varepsilon}{2}.$$

Then $B_k = B_{r/2^k}(x_k) \subset B_\varepsilon(x_*) \subset U_{i_0}$ contradicting the fact that B_k cannot be covered by finitely many U_i's. □

Corollary 17.3.6. *A compact metric space (X, d) is separable, i.e., it admits a dense countable subset.*

Proof. Since X is compact, it is totally bounded and thus, for any $n \in \mathbb{N}$, there exists a finite $\frac{1}{n}$-net S_n. The set

$$S = \bigcup_{n\in\mathbb{N}} S_n$$

is at most countable. It is also dense because, for any $x \in X$ and any $n \in \mathbb{N}$ there exists $x_n \in S_n$ such that $x \in B_{1/n}(x_n)$ thus $d(x_n, x) < 1/n$, so the sequence (x_n) in S converges to x. □

Corollary 17.3.7 (Lebesgue number)**.** *Suppose that (X, d) is a compact metric space. Then, for any open cover $(U_i)_{i\in I}$ of X there exists a positive number r such that, for any $x \in X$ the ball $B_r(x)$ is contained in one of the open sets U_i. Such a number is called a* Lebesgue number *of the open cover.*

Proof. Any $x \in X$ is contained in one of the open sets U_i so there exists $r_x > 0$ such that the ball $B_{r_x}(x)$ is contained in one of the open sets U_i. The collection of open balls

$$\left\{ B_{r_x/2}(x) \right\}_{x\in X}$$

is, tautologically, an open cover of X. Since X is compact, there exist finitely many points $x_1, \dots, x_n$ in X such that the balls $B_{r_{x_1}/2}(x_1), \dots, B_{r_{x_n}/2}(x_n)$ cover X. We set

$$r := \min_{1\leqslant k\leqslant n} r_{x_k}/2.$$

Note that any $x \in X$ belongs to one of the balls $B_{r_{x_k}/2}(x_k)$ so

$$B_r(x) \subset B_{r_{x_k}}(x_k),$$

and, by construction, $B_{r_{x_k}}(x)$ is contained in one of the open sets U_i. □

Corollary 17.3.8. *Suppose* (X, d) *and* $(Y, \bar{d})$ *are metric spaces and* X *is compact. Then any continuous map* $F : X \to Y$ *is* uniformly continuous*, i.e.,*

$$\begin{gathered} \forall \varepsilon > 0,\ \exists \delta = \delta(\varepsilon) > 0 \text{ such that} \\ \forall x, x' \in X,\ d(x, x') < \delta \Rightarrow \bar{d}\big(F(x), F(x') \big) < \varepsilon. \end{gathered} \tag{17.3.1}$$

Proof. We argue by contradiction. Assume (17.3.1) is false. Thus, there exists $\varepsilon_0 > 0$ such that, for any $n \in \mathbb{N}$ there exist x_n, x_n' satisfying

$$d(x_n, x_n') < \frac{1}{n} \text{ and } \bar{d}\big(F(x_n), F(x_n') \big) \geqslant \varepsilon_0.$$

Since X is compact, the sequence (x_n) contains a convergent subsequence $(x_{n_k})_{k\geqslant 1}$

$$x_* = \lim_{k\to\infty} x_{n_k}$$

Since

$$d\big(x_{n_k}, x_{n_k}' \big) < \frac{1}{n_k} \to 0 \text{ as } k \to \infty,$$

we deduce

$$x_* = \lim_{k\to\infty} x_{n_k}'.$$

Hence, since F is continuous

$$\varepsilon_0 \leqslant \lim_{k\to\infty} \bar{d}\big(F(x_{n_k}), F(x_{n_k}') \big) = \bar{d}\big(F(x_*), F(x_*) \big) = 0.$$

We have reached a contradiction. □

17.3.2. Compact subsets. A subset of a metric space (X, d) is itself a metric space with respect to the induced metric. We say that a subset $K \subset X$ is *compact* if, viewed as a metric space with the induced metric d_K, it is a compact metric space. In view of Proposition 17.1.12 a subset K is compact if any collection of open subsets of X that covers K contains a finite subcollection that also covers K. Equivalently, K is a compact subset if and only if it any sequence in K contains a subsequence that converges to a point also in K.

Proposition 17.3.9. *A compact subset K of a metric space (X, d) is a closed set.*

Proof. Let (x_n) be a convergent sequence of points in K. We have to show that its limit also belongs to K. Indeed, the sequence (x_n) is Cauchy and, since the metric space (K, d_K) is complete, it converges to a point in K. □

Corollary 17.3.10. *Suppose that (X, d) is a compact metric space and $S \subset X$. Then the following statements are equivalent.*

(i) *The set S is compact.*

(ii) *The set S is closed.*

Proof. The implication (i) $\Rightarrow$ (ii) follows from the previous proposition. To prove the converse, we assume that S is closed and we will show that it satisfies the Bolzano-Weierstrass property.

Consider a sequence (x_n) of points in S. The ambient space X is compact and thus this sequence contains a convergent subsequence. Since S is closed, the limit of this subsequence is in S. □

Arguing *exactly* as in the proof of Theorem 12.4.7 we obtain the following result.

Theorem 17.3.11 (Continuous partitions of unity)**.** *Suppose (X, d) is a metric space and that $K \subset X$ is a compact subset. Then, for any open cover $\mathfrak{U}$ of K, there exists a partition of unity on K subordinated to $\mathfrak{U}$, i.e., a finite collection of continuous functions $\chi_1, \dots, \chi_\ell : X \to [0, 1]$ with the following properties.*

(i) *For any $i = 1, \dots, \ell$, there exists an open subset U_i in the collection $\mathfrak{U}$ such that $\operatorname{supp} \chi_i \subset U_i$.*

(ii) $\chi_1(\boldsymbol{x}) + \cdots + \chi_\ell(\boldsymbol{x}) = 1$, $\forall \boldsymbol{x} \in K$.

□

Definition 17.3.12. Let (X, d) be a metric space. A subset $S \subset X$ is called *relatively compact* (or *precompact*) if its closure is compact. □

We see that a subset S of a metric space (X, d) is relatively compact if and only if any sequence of points in S contains a subsequence that converges to a point, *not necessarily in S*.

Proposition 17.3.13. *Suppose that S is a subset of the* <u>complete</u> *metric space (X, d). Then, the following are equivalent.*

(i) *The set S is relatively compact.*

(ii) *The set S is totally bounded, i.e., for any $\varepsilon > 0$ there exist finitely many points $x_1, \dots, x_n \in X$ such that*

$$S \subset \bigcup_{k=1}^{n} B_\varepsilon(x_k)$$

Proof. (i) ⇒ (ii) Clearly, for any $y \in \boldsymbol{cl}\, S$ there exists $x \in S$ such that $d(x, y) < \varepsilon$. In other words the collection $(B_\varepsilon(x))_{x \in S}$ is an open cover of the compact set $\boldsymbol{cl}\, S$ and thus it admits a finite subcover.

(ii) ⇒ (i) The closure $\boldsymbol{cl}\, S$ is a closed subset of the complete metric space X and thus it is complete as a metric subspace. We need to show that it is totally bounded. Cover S by finitely many open balls of radius $\varepsilon/4$ centered at $x_1, \dots, x_m \in X$. For every $k = 1, \dots, m$ pick $s_k \in S \cap B_{\varepsilon/2}(x_k)$. Then the collection $\{s_1, \dots, s_m\}$ is an ε-net of $\boldsymbol{cl}(S)$. Indeed, every point in $\boldsymbol{cl}\, S$ is within $\varepsilon/2$ of one the points x_j and, each of the points x_j is within $\varepsilon/2$ of the point $s_j \in S$. Thus each point in $\boldsymbol{cl}(S)$ is within ε from one of the points $s_1, \dots, s_m \in S$. □

Remark 17.3.14. If S is relatively compact then for any $\varepsilon > 0$, then the above proof shows that the set S can be covered by finitely many balls of radius $\varepsilon > 0$ *centered at points in S*. □

Theorem 17.3.15. *Suppose that (X, d) and $(Y, \bar{d})$ are metric spaces, $K \subset X$ is a compact subset and $F : X \to Y$ is a continuous map. Then $F(K)$ is a compact subset of Y.*

Proof. Let $(U_i)_{i \in I}$ be an open cover of $F(K)$. Since F is continuous, each of the sets $V_i = F^{-1}(U_i)$ is an open subset of X and the collection $(V_i)_{i \in I}$ is an open cover of K. The set K is compact so it can be covered by finitely many of the sets V_i, say $V_{i_1}, \dots, V_{i_n}$. Then the open sets $U_{i_1}, \dots, U_{i_n}$ cover $F(K)$. □

Corollary 17.3.16. *Suppose that (X, d) and $(Y, \bar{d})$ are metric spaces, $S \subset X$ is a relatively compact subset and $F : X \to Y$ is a continuous map. Then $F(S)$ is a relatively compact subset of Y.*

Proof. The set $\boldsymbol{cl}\, S$ is compact so $F(\boldsymbol{cl}\, S)$ is compact and in particular closed. Thus $\boldsymbol{cl}\, F(S) \subset F(\boldsymbol{cl}\, S)$ so $\boldsymbol{cl}\, F(S)$ is a closed subset of the compact set $F(\boldsymbol{cl}\, S)$ and thus compact. □

Theorem 17.3.17 (Weierstrass)**.** *Suppose that (X, d) is a metric space, $K \subset X$ is a compact set, and $f : K \to \mathbb{R}$ is a continuous function. Then there exist $x_*, x^* \in K$ such* *that*

$$f(x_*) = \inf_{x \in K} f(x), \quad f(x^*) = \sup_{x \in K} f(x).$$

In particular, the function f is bounded on K.

Proof. Choose a sequence of points (x_n) in K such that

$$\lim_{n \to \infty} f(x_n) = \sup_{x \in K} f(x) \in (-\infty, \infty].$$

Since K is compact, the sequence (x_n) contains a convergent subsequence (x_{n_k}). We denote by x^* is limit. Since f is continuous we deduce

$$\infty > f(x^*) = \lim_{k \to \infty} f(x_{n_k}) = \lim_{n \to \infty} f(x_n) = \sup_{x \in K} f(x).$$

The statement involving the infimum of f on K is proved in a similar fashion. □

17.4. Continuous functions on compact sets

Let (X, d) be a metric space and $K \subset X$ a compact subset.

Proposition 17.4.1. *The space $C(K)$ of continuous functions $K \to \mathbb{R}$ equipped with the sup-norm is a Banach space.*

Proof. Indeed if (f_n) is a Cauchy sequence of $C(K)$, then for any $x \in X$ the sequence of real numbers $\big(f_n(x) \big)_{n \geqslant 1}$ is Cauchy and thus convergent. We denote by $f(x)$ its limit.

Then for any $x \in X$ and any $m \in \mathbb{N}$ we have

$$|f(x) - f_m(x)| = \lim_{n \to \infty} |f_n(x) - f_m(x)| \leqslant \sup_{n \geqslant m} \|f_n - f_m\|_\infty.$$

Since (f_n) is Cauchy, for any $\varepsilon > 0$ there exists $m = m(\varepsilon)$ such that

$$\sup_{n \geqslant m} \|f_n - f_m\|_\infty < \frac{\varepsilon}{3}.$$

The function f_m is uniformly continuous so there exists $\delta = \delta(m, \varepsilon) > 0$ such that

$$d(x, x') < \delta \Rightarrow |f_m(x) - f_m(x')| < \frac{\varepsilon}{3}.$$

We deduce that if $d(x, x') < \delta$ then

$$|f(x) - f(x')| \leqslant |f(x) - f_m(x)| + |f_m(x) - f_m(x')| + |f_m(x') - f(x)| < \varepsilon.$$

This proves that $f \in C(K)$. On the other hand,

$$\|f - f_m\|_\infty \leqslant \sup_{n \geqslant m} \|f_n - f_m\|_\infty$$

so that

$$\lim_{m \to \infty} \|f - f_m\| = 0.$$

□

In this section we investigate two properties of the space $C(K)$ that are extremely useful in practice: compactness and separability.

17.4.1. Compactness in $C(K)$. The main question we want to address in this subsection is the following: when is a family of functions $\mathcal{F} \subset C(K)$ precompact? This boils down to the even more concrete question.

> *What do we need to know about a sequence of continuous functions $f_n : K \to \mathbb{R}$ to be able to conclude that it contains a uniformly convergent subsequence.*

Clearly if this sequence contained a Cauchy (in the sup-norm) subsequence we would be able to reach this conclusion. This is however too stringent a condition. Let us first look at simpler conditions necessary for such a subsequence to exist

Recall that a function $f : K \to \mathbb{R}$ is continuous at $x \in K$ if

$$\forall \varepsilon > 0,\ \ \exists \delta = \delta_f(x, \varepsilon) > 0,\ \ \forall x' \in K:\ \ d(x, x') < \delta \Rightarrow |f(x) - f(x')| < \varepsilon. \tag{17.4.1}$$

We will refer to a function $\varepsilon \mapsto \delta_f(x, \varepsilon)$ satisfying (17.4.1) as *a continuity rate*[3] of the function f at the point x. At different continuity points x, x' it is possible that $\delta(\varepsilon, x) \neq \delta(\varepsilon, x')$. However, if $f : K \to \mathbb{R}$ is continuous everywhere, then it is also *uniformly* continuous so

$$\forall \varepsilon > 0,\ \ \exists \delta = \delta_f(\varepsilon) > 0,\ \ \forall x, x' \in K:\ \ d(x, x') < \delta \Rightarrow |f(x) - f(x')| < \varepsilon.$$

In other words, for a uniformly continuous function we can choose a function $\varepsilon \mapsto \delta_f(\varepsilon)$ so that (17.4.1) works for all $x \in K$, not just a specific x. We will refer to such a function as a *rate of uniform continuity.*

Suppose now that the sequence of continuous functions $f_n : K \to \mathbb{R}$ converges uniformly to the continuous function $f_\infty : K \to \mathbb{R}$. Then we can choose a rate of uniform continuity that works for *all* the functions f_n. More precisely

$$\begin{split}\forall \varepsilon > 0,\ \exists \delta = \delta(\varepsilon) > 0 :\qquad\qquad\qquad\qquad\qquad\qquad\\ \forall n \in \mathbb{N},\ \forall x, x' \in K,\ \ d(x, x') < \delta \Rightarrow |f_n(x) - f_n(x')| < \varepsilon.\end{split} \tag{17.4.2}$$

Indeed, suppose that (17.4.2) were false. Then there exists $\varepsilon_0 > 0$ such that, for any $\delta > 0$, there exist $n = n(\delta) \in \mathbb{N}$, $x_\delta, x'_\delta \in K$ so that

$$d(x_\delta, x'_\delta) < \delta \text{ and } |f_{n(\delta)}(x_\delta) - f_{n(\delta)}(x'_\delta)| \geqslant \varepsilon_0.$$

Now choose $\delta = 1/m$ with $m \to \infty$. We get sequences $n_m \in \mathbb{N}$ $x_m, x'_m \in K$ with the above property. Since the set K is compact, there exists a subsequence m_k such that of x_{m_k} is convergent to a point x_* in K and

$$\lim_{k\to\infty} m_k = m_\infty \in \mathbb{N} \cup \{\infty\},\ \ d(x_{m_k}, x'_{m_k}) < \frac{1}{m_k},\ \ \forall k.$$

Then, for any $k \in \mathbb{N}$

$$\begin{split}0 < \varepsilon_0 \leqslant \big|f_{m_k}(x_{m_k}) - f_{m_k}(x'_{m_k})\big|\\ \leqslant \big|f_{m_k}(x_{m_k}) - f_{m_\infty}(x_{m_k})\big| + \big|f_\infty(x_{m_k}) - f_\infty(x'_k)\big| + \big|f_\infty(x'_{m_k}) - f_{m_k}(x'_{m_k})\big|\end{split}$$

[3] This is also referred as *modulus of continuity.*

$$\leqslant \big\|f_{m_k} - f_{m_\infty}\big\|_\infty + \big|f_{m_\infty}(x_{m_k}) - f_{m_\infty}(x'_{m_k})\big| + \big\|f_{m_k} - f_{m_\infty}\big\|_\infty.$$

Note that

$$\lim_{k\to\infty} \big\|f_{m_k} - f_{m_\infty}\big\|_\infty = 0$$

since f_{m_k} converges uniformly to f_{m_∞}. Since f_{m_∞} is uniformly continuous and

$$\lim_{k\to\infty} d(x_{m_k}, x'_{m_k}) = 0,$$

we deduce

$$\lim_{k\to\infty} |f_{m_\infty}(x_{m_k}) - f_{m_\infty}(x'_{m_k})| = 0.$$

Definition 17.4.2. Let (X,d) be a metric space, K a compact set and $\mathcal{F} \subset C(K)$ be a family of continuous functions on K.

(i) We say that the family is *bounded* if $\exists C > 0$ such that $\|f\|_\infty < C$, $\forall f \in \mathcal{F}$.

(ii) We say that the family is *equicontinuous* if

$$\begin{gathered} \forall \varepsilon > 0,\ \exists \delta = \delta(\varepsilon) > 0 :\ \boxed{\forall f \in \mathcal{F},\ \ \forall x, x' \in K},\\ d(x,x') < \delta \Rightarrow |f(x) - f(x')| < \varepsilon. \end{gathered} \tag{17.4.3}$$

□

The statement (17.4.2) shows that if a sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly convergent, then the family $\{f_n\}_{n\in\mathbb{N}}$ is equicontinuous. Clearly, a finite family is equicontinuous. We can now state the main result of this subsection.

Theorem 17.4.3 (Arzelà-Ascoli)**.** *Let (X,d) be a metric space, K a compact subset and $\mathcal{F} \subset C(K)$ be a family of continuous functions on K. Then the following statements are equivalent.*

(i) *The family $\mathcal{F}$ is relatively compact, i.e., any sequence of functions in $\mathcal{F}$ contains a uniformly convergent subsequence.*

(ii) *The family $\mathcal{F}$ is bounded and equicontinuous.*

Proof. We follow closely the approach in [12, Sec. VII.5].

(i) ⇒ (ii) The function

$$C(K) \ni f \mapsto \|f\|_\infty$$

is continuous and thus it is bounded on the compact subsets of $C(K)$ and thus it is bounded on $\mathcal{F}$. This proves that $\mathcal{F}$ is bounded.

Since $\mathcal{F}$ is relatively compact we deduce from Proposition 17.3.13 and Remark 17.3.14 that for any $\varepsilon > 0$ there exists a *finite* subfamily $(g_i)_{i\in I}$ of $\mathcal{F}$ such that $\forall f \in \mathcal{F}$, there exists $i = i(f) \in I$ satisfying

$$\|f - g_{i(f)}\|_\infty < \frac{\varepsilon}{3}. \tag{17.4.4}$$

Note that since the family (g_i) is finite it is equicontinuous so, for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\begin{gathered}\forall \varepsilon > 0,\ \exists \delta = \delta(\varepsilon) > 0 :\ \forall i \in I,\ \forall x, x' \in K,\\ d(x, x') < \delta \Rightarrow |g_i(x) - g_i(x')| < \frac{\varepsilon}{3}\end{gathered} \tag{17.4.5}$$

If $d(x, x') < \delta(\varepsilon/3)$ and $f \in \mathcal{F}$, then

$$\begin{aligned}|f(x) - f(x')| &\leqslant |f(x) - g_{i(f)}(x)| + |g_{i(f)}(x) - g_{i(f)}(x')| + |g_{i(f)}(x') - f(x')|\\ &\overset{(17.4.5)}{\leqslant} \|f - g_{i(f)}\|_\infty + \frac{\varepsilon}{3} + \|f - g_{i(f)}\|_\infty \overset{(17.4.4)}{\leqslant} \varepsilon.\end{aligned}$$

This proves that $\mathcal{F}$ is equicontinuous.

(ii) $\Rightarrow$ (i). The normed space $C(K)$ is complete and thus, in view of Proposition 17.3.13, it suffices to prove that $\mathcal{F}$ is totally bounded.

Since $\mathcal{F}$ is equicontinous

$$\forall \varepsilon > 0,\ \exists \delta = \delta(\varepsilon) > 0 :\ \forall f \in \mathcal{F},\ \forall x, x' \in K,\ \ d(x, x') < \delta \Rightarrow |f(x) - f(x')| < \frac{\varepsilon}{4}. \tag{17.4.6}$$

The space K is compact so there exist finitely many points $x_1, \dots, x_m \in K$ such that

$$K \subset \bigcup_{i=1}^{m} B_{\delta(\varepsilon)}(x_i).$$

Since the family $\mathcal{F}$ is bounded, there exists $C > 0$ such that

$$|f(x_i)| < C,\ \ \forall f \in \mathcal{F},\ \ i = 1, \dots, m.$$

Choose $n \in \mathbb{N}$ sufficiently large so that $\frac{2C}{n} < \frac{\varepsilon}{4}$ and define c_k, $k = 0, 1, \dots, n$

$$c_k = -C + \frac{2kC}{n}.$$

More precisely, c_k are the points dividing the interval $[-C, C]$ into n equal parts. Note that for any $f \in \mathcal{F}$ and any $i = 1, \dots, m$, the value of f at x_i belongs to one of the intervals $(c_{k-1}, c_k]$

Denote by Φ the finite collection of functions $\{1, \dots, m\} \to \{1, \dots, n\}$. For $\varphi \in \Phi$ we denote by $\mathcal{F}_\varphi$ the subfamily of $\mathcal{F}$ consisting of functions f such that

$$f(x_i) \in \big(\, c_{\varphi(i)-1}, c_{\varphi(i)}\, \big].$$

Clearly

$$\mathcal{F} = \bigcup_{\varphi \in \Phi} \mathcal{F}_\varphi.$$

We will show that for any $\varphi \in \Phi$ the set $\mathcal{F}_\varphi$ has diameter $< \varepsilon$, more precisely

$$\|f - g\|_\infty < \varepsilon,\ \ \forall f, g \in \mathcal{F}_\varphi.$$

If this happens, then

$$\mathcal{F}_\varphi \subset B_\varepsilon(f),\ \ \forall f \in \mathcal{F}_\varphi.$$

For each $\varphi \in \Phi$ choose a function $f_\varphi \in \mathcal{F}_\varphi$ and we deduce

$$\mathcal{F} = \bigcup_{\varphi \in \Phi} \mathcal{F}_\varphi \subset \bigcup_{\varphi \in \Phi} B_\varepsilon(f_\varphi).$$

Let $f, g \in \mathcal{F}_\varphi$. Then, for any $x \in K$ there exists x_i such that $d(x, x_i) < \delta(\varepsilon)$ and

$$|f(x) - g(x)| \leqslant |f(x) - f(x_i)| + |f(x_i) - g(x_i)| + |g(x_i) - g(x)|.$$

From (17.4.6) we deduce that

$$|f(x) - f(x_i)|,\ |g(x_i) - g(x)| < \frac{\varepsilon}{4}.$$

Since $f, g \in \mathcal{F}_\varphi$ we have

$$|f(x_i) - g(x_i)| < c_{\varphi(i)} - c_{\varphi(i)-1} < \frac{\varepsilon}{4}.$$

Hence

$$|f(x) - g(x)| < \frac{3\varepsilon}{4},\ \ \forall x \in K,\ f, g \in \mathcal{F}_\varphi$$

so that

$$\|f - g\|_\infty < \varepsilon,\ \ \forall f, g \in \mathcal{F}_\varphi.$$

□

Corollary 17.4.4. *Let (K, d) be a compact metric space and $\mathcal{F} \subset C(K)$ a bounded family of functions such that*

$$\exists L > 0,\ \ \forall f \in \mathcal{F},\ \ \forall x, x' \in K:\ \ \big|f(x) - f(x')\big| \leqslant L d(x, x').$$

Then $\mathcal{F}$ is precompact.

Proof. The family satisfies (17.4.3) with $\delta(\varepsilon) = \frac{\varepsilon}{L}$ so it is equicontinuous. □

17.4.2. Approximations of continuous functions. Suppose that (X, d) is a metric space and $K \subset X$ is a compact subset. The main goal of this subsection is to produce examples of relatively small dense vector subspaces of $C(K)$. We begin with a rather exotic result guaranteeing *uniform* convergence of a sequence of functions.

We say that a sequence of functions $f_n : K \to \mathbb{R}$ converges *pointwisely* to a function $f : X \to \mathbb{R}$ if

$$\forall x \in K\ \ \lim_{n \to \infty} f_n(x) = f(x).$$

This means that

$$\boxed{\forall x \in K},\ \forall \varepsilon > 0\ \exists N = N(\varepsilon, x) > 0:\ \ \forall n > N\ \ |f_n(x) - f(x)| < \varepsilon.$$

The convergence is *uniform* if

$$\forall \varepsilon > 0,\ \exists N = N(\varepsilon) > 0:\ \boxed{\forall x \in K},\ \ \forall n > N\ \ |f_n(x) - f(x)| < \varepsilon.$$

Clearly a sequence that converges uniformly also converges pointwisely, but the converse is not necessarily true. We also know that if a sequence (f_n) of continuous functions on K converges uniformly to a function f, then the limit is continuous.

Suppose that (f_n) is a sequence of continuous functions on K that converges *pointwisely* to a *continuous* function $f : K \to \mathbb{R}$. Can we conclude that the converges is actually uniform? Exercise 17.41 describes an example of a sequence of continuous functions converging pointwisely but not uniformly to a continuous function. The discussion in the previous subsection shows that f_n converges uniformly if and only if the family $(f_n)_{n\in\mathbb{N}}$ is equicontinuous. The next result describes one other situation when the answer to the above question is positive.

Theorem 17.4.5 (Dini)**.** *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence in $C(K)$, i.e.,*

$$\forall x \in K,\ \forall n \in \mathbb{N}: \quad f_n(x) \leqslant f_{n+1}(x).$$

Assume that for any $x \in K$ the limit $f(x)$ of the nondecreasing sequence of real numbers $\big(f_n(x)\big)_{n\in\mathbb{N}}$ is finite and the resulting function

$$K \ni x \mapsto f(x) \in \mathbb{R}$$

is continuous. Then the sequence (f_n) converges uniformly *to f.*

Proof. We argue by contradiction, Thus we assume that there exists $\varepsilon_0 > 0$ such that for any $N > 0$ there exists $x_N \in K$ and $\nu = \nu(N) > N$ such that $|f_\nu(x_N) - f(x_N)| > \varepsilon_0$, i.e.,

$$\forall N > 0\ \exists x_N \in K, \quad f_N(x_N) \leqslant f_{\nu(N)}(x_N) \leqslant f(x_N) - \varepsilon_0. \tag{17.4.7}$$

Since K is compact, upon extracting a subsequence we can assume that

$$\lim_{N\to\infty} x_N = x_* \in K.$$

Choose $n_0 > 0$ such that

$$f(x_*) - \frac{\varepsilon_0}{2} < f_{n_0}(x_*) \leqslant f(x_*). \tag{17.4.8}$$

Now observe that for $N > n_0$ we have

$$f_{n_0}(x_N) \leqslant f_N(x_N) \overset{(17.4.7)}{\leqslant} f(x_N) - \varepsilon_0.$$

Letting $N \to \infty$ we deduce

$$f_{n_0}(x_*) = \lim_{N\to\infty} f_{n_0}(x_N) \leqslant \lim_{N\to\infty} f(x_N) - \varepsilon_0 = f(x_*) - \varepsilon_0.$$

This contradicts (17.4.8). □

We are now ready to state and prove the main result of this subsection.

Theorem 17.4.6 (Stone-Weierstrass)**.** *Suppose that K is a compact subset of a metric space (X, d) and $\mathcal{A} \subset C(K)$ is an algebra of continuous functions that contains the constant* *functions and is ample, i.e., separates points; see Definition 17.1.32*[4]*. Then $\mathcal{A}$ is dense in $C(K)$ with respect to the sup-norm.*

Proof. We have to show that $\boldsymbol{cl}\,\mathcal{A} = C(K)$. We follow the elegant approach in [**12**, Sec. VII.3].

Since $\mathcal{A}$ is an algebra, for any $f \in \mathcal{A}$ we have $f^n \in \mathcal{A}$. In particular, for any polynomial

$$P(t) = p_n t^n + \cdots + p_1 t + p_0$$

we have

$$P(f) = p_n f^n + \cdots + p_1 f + p_0 \in \mathcal{A}.$$

Let us also observe that the closure of $\mathcal{A}$ is also an algebra of continuous functions; see Exercise 17.40.

Lemma 17.4.7. *Consider the sequence of polynomials $P_n(t)$ defined recursively by*

$$P_1(t) = 0, \;\; P_{n+1}(t) = P_n(t) + \frac{1}{2}\big(\, t - P_n(t)^2 \,\big), \;\; \forall n \in \mathbb{N}.$$

Then, for any $t \in [0,1]$ the sequence $\big(\, P_n(t) \,\big)$ is nondecreasing and converges to $\sqrt{t}$.

Proof. Observe that $P_n(0) = 0$, $\forall n$ so the claim is true for $t = 0$. Fix $t \in (0,1]$ and consider

$$F_t : \mathbb{R} \to \mathbb{R}, \;\; F_t(x) = x + \frac{1}{2}\big(\, t - x^2 \,\big).$$

Note that

$$\frac{d}{dx} F_t(x) = 1 - x \geqslant 0, \;\; \forall x \in [0,1].$$

Hence F_t is increasing on $[0,1]$. Now observe that for any $t \in [0,1]$ we have

$$F_t(0) = \frac{1}{2} t < \sqrt{t}, \;\; F_t(1) = \frac{1+t}{2} \leqslant 1, \;\; F_t(\sqrt{t}) = \sqrt{t}.$$

Hence

$$F_t\big(\, [0,1] \,\big) \subset [0,1], \;\; \forall t \in [0,1].$$

We have

$$P_2(t) = F_t(0) = \frac{1}{2} t > P_1(t) = 0.$$

Hence

$$P_1(t) \leqslant P_2(t) < \sqrt{t} \leqslant 1. \tag{17.4.9}$$

Since

$$P_{n+1}(t) = F_t\big(\, P_n(t) \,\big), \tag{17.4.10}$$

we deduce from (17.4.9) that

$$F_t\big(\, P_1(t) \,\big) \leqslant F_t\big(\, P_2(t) \,\big) \leqslant F_t(\sqrt{t}) = \sqrt{t},$$

i.e.

$$0 \leqslant P_2(t) \leqslant P_3(t) \leqslant \sqrt{t}.$$

[4] This means that for any $x_0, x_1 \in X$, $x_0 \neq x_1$, there exists a function $g \in \mathcal{A}$ such that $g(x_0) \neq g(x_1)$.

Arguing inductively we deduce

$$0 \leqslant P_n(t) \leqslant P_{n+1}(t) \leqslant \sqrt{t}.$$

Hence the sequence $P_n(t)$ is nondecreasing and bounded above by $\sqrt{t}$ and thus converges to a limit $0 \leqslant \ell_t \leqslant \sqrt{t}$. Using (17.4.10) we deduce that $F_t(\ell_t) = \ell_t$ so $\ell_t = \sqrt{t}$. □

Using Dini's Theorem we deduce that the above sequence $P_n(t)$ converges to $\sqrt{t}$ uniformly on $[0,1]$.

Lemma 17.4.8. *For any $f \in \mathcal{A}$, $|f| \in \boldsymbol{cl}\,\mathcal{A}$.*

Proof. Since f is bounded, we deduce that there exists $c > 0$ such that

$$0 \leqslant f(x)^2/c^2 \leqslant 1, \quad \forall x \in K.$$

If $P_n(t)$ are the polynomials in Lemma 17.4.7 we deduce that $P_n\big(f^2/c^2\big) \in \mathcal{A}$ and

$$P_n\big(f(x)^2/c^2\big) \nearrow \sqrt{f(x)^2/c^2} = |f(x)|/c, \quad \forall x \in K.$$

Since $|f|/c$ is continuous we deduce from Dini's Theorem that $P_n\big(f^2/c^2\big)$ converges uniformly on K to $|f|/c$. Hence $|f|/c \in \boldsymbol{cl}\,\mathcal{A}$ so that $|f| = c(|f|/c) \in \boldsymbol{cl}\,\mathcal{A}$ since $\boldsymbol{cl}\,\mathcal{A}$ is a vector subspace of $C(X)$. □

Observe that if $f, g \in \mathcal{A}$, then $|f-g|, |f+g| \in \boldsymbol{cl}\,\mathcal{A}$ so that

$$\max(f,g) = \frac{1}{2}\big(f + g + |f-g|\big) \in \boldsymbol{cl}\,\mathcal{A}, \ \ \min(f,g) = \frac{1}{2}\big(f + g - |f-g|\big) \in \boldsymbol{cl}\,\mathcal{A}.$$

Lemma 17.4.9. *For any real numbers a_0, a_1 and any $x_0 \neq x_1 \in K$ there exists $g \in \mathcal{A}$ such that*

$$g(x_0) = a_0, \ \ g(x_1) = a_1.$$

Proof. Since $\mathcal{A}$ separates points there exists $f \in \mathcal{A}$ such that $f(x_0) \neq f(x_1)$. Now consider the function

$$g : K \to \mathbb{R}, \ \ g(x) = a_0 + \frac{a_1 - a_0}{f(x_1) - f(x_0)}\big(f(x) - f(x_0)\big), \ \ \forall x \in K.$$

Since $\mathcal{A}$ contains the constant functions we deduce that $g \in \mathcal{A}$. Clearly $g(x_i) = a_i$, $i = 0, 1$. □

Lemma 17.4.10. *For any $f \in C(K)$, any $x_0 \in K$ and any $\varepsilon > 0$ there exists a function $g \in \boldsymbol{cl}\,\mathcal{A}$ such that*

$$g(x_0) = f(x_0), \ \ g(x) \leqslant f(x) + \varepsilon, \ \ \forall x \in K.$$

Proof. For any $y \in K$ choose a function $h_y \in \mathcal{A}$ such that

$$h_y(x_0) = f(x_0), \ \ h_y(y) \leqslant f(y) + \frac{\varepsilon}{2}.$$

For every $y \in K$, there exists an open ball $B_{r_y}(y)$ such that

$$h_y(x) \leqslant f(x) + \varepsilon, \ \ \forall x \in B_{r_y}(y) \cap K.$$

The collection of open balls $\big(B_{r_y}(y)\big)_{y \in K}$ covers K and, since K is a compact set, we deduce that there exist finitely many of points $y_1, \dots, y_m \in K$ such that

$$K \subset \bigcup_{i=1}^{m} B_{r_i}(y_i), \ \ r_i := r_{y_i}.$$

Now set

$$g = \min\big(\, h_{y_1}, \dots, h_{y_m}\,\big) \in \boldsymbol{cl}\,\mathcal{A}.$$

Clearly $h_{y_i}(x_0) = f(x_0)$, $\forall i$ so $g(x_0) = f(x_0)$. Moreover, for $x \in B_{r_i}(y_i) \cap K$

$$g(x) \leqslant h_{y_i}(x) \leqslant f(x) + \varepsilon.$$

□

We can now complete the proof of Theorem 17.4.6. Since $\boldsymbol{cl}\big(\boldsymbol{cl}\,\mathcal{A}\big) = \boldsymbol{cl}\,\mathcal{A}$ (Exercise 17.3) it suffices to show that for any $f \in C(K)$ and any $\varepsilon > 0$, there exists $g \in \boldsymbol{cl}\,\mathcal{A}$ such that $\|f - g\|_\infty < \varepsilon$.

For any $x \in K$ choose a function $g_x \in \boldsymbol{cl}\,\mathcal{A}$ such that

$$g_x(x) = f(x), \ \ g_x(y) \leqslant f(y) + \varepsilon, \ \ \forall y \in K.$$

Note that for any $x \in K$ there exists $r_x > 0$ such that,

$$\forall y \in B_{r_x}(x) \cap K, \ \ g_x(y) \geqslant f(y) - \varepsilon.$$

Since K is compact we can cover it with finitely many of the above balls

$$K \subset \bigcup_{i=1}^{m} B_{r_i}(x_i), \ \ r_i := r_{x_i}.$$

Now define $g \in C(K)$ by

$$g(x) := \max\big\{\, g_{x_1}(x), \dots, g_{x_m}(x)\,\big\}, \ \ \forall x \in K.$$

Note that $g \in \boldsymbol{cl}\,\mathcal{A}$, and for $y \in B_{r_i}(x_i)$ we have

$$f(y) - \varepsilon \leqslant g_{x_j}(y) \leqslant f(y) + \varepsilon, \ \ \forall j = 1, \dots, m.$$

Hence

$$f(y) - \varepsilon \leqslant g(y) \leqslant f(y) + \varepsilon, \ \ \forall y \in X.$$

□

Corollary 17.4.11 (Weierstrass)**.** *For any continuous function $f : [a, b] \to \mathbb{R}$ there exists a sequence of polynomials $(p_n)_{n \in \mathbb{N}}$ that converges uniformly to f on $[a, b]$.*

Proof. Let $\mathcal{A} \subset C([a,b])$ denote the algebra of polynomial functions. Clearly it contains the constant functions and separates points because, for any distinct points $x_0, x_1 \in [a,b]$, the polynomial $P_1(x) = x$ takes different values at x_0 and x_1. Thus $\mathcal{A}$ is dense in $C([a,b])$. □

Let (X,d) and K be as in the Theorem 17.4.6. Denote by $C(K,\mathbb{C})$ the set of continuous functions $K \to \mathbb{C}$, This is a complex vector space. We define sup-norm as in the real case

$$\|f\|_\infty := \sup_{x \in K} \bigl|\, f(x) \,\bigr|$$

Arguing exactly as in the real case we deduce that the sup-norm is complete even in the complex vector space.

A $\mathbb{C}$-subalgebra of $C(K,\mathbb{C})$ is a complex vector subspace $\mathcal{A} \subset C(K,\mathbb{C})$ such that

$$\forall f, g \in \mathcal{A} \;\; f \cdot g \in \mathcal{A}.$$

Theorem 17.4.12 (Stone-Weierstrass: the complex version)**.** *Suppose that K is a compact subset of a metric space (X,d) and $\mathcal{A} \subset C(K,\mathbb{C})$ is a $\mathbb{C}$-subalgebra such that*

(i) $1 \in \mathcal{A}$

(ii) *$\mathcal{A}$ is closed under conjugation, i.e., $\forall f \in \mathcal{A}$, $\bar{f} \in \mathcal{A}$, and*

(iii) *$\mathcal{A}$ is ample, i.e., $\forall x0, x_1 \in K$, $x_0 \neq x_1$, and $\forall c_0, c_1 \in \mathbb{C}$ there exists $f \in \mathcal{A}$ such that $f(x_i) = c_i$, $i = 0, 1$.*

Then $\mathcal{A}$ is dense in $C\bigl(K, \mathbb{C}\bigr)$.

Proof. We set

$$\mathcal{A}_{\mathbb{R}} := \bigl\{ f \in \mathcal{A};\;\; f(x) \in \mathbb{R}, \;\; \forall x \in \mathbb{R} \bigr\}$$

note that for any $f \in \mathcal{A}$ we have

$$\mathbf{Re}\, f = \frac{1}{2}\bigl(f + \bar{f} \bigr) \in \mathcal{A}_{\mathbb{R}}, \;\; \mathbf{Im}\, f = \frac{1}{2\boldsymbol{i}}\bigl(f - \bar{f} \bigr) \in \mathcal{A}_{\mathbb{R}}.$$

Hence $\mathbf{Re}\, f, \mathbf{Im}\, f \in \mathcal{A}_{\mathbb{R}}$ for any $f \in \mathcal{A}$.

Clearly $\mathcal{A}_{\mathbb{R}}$ is an $\mathbb{R}$-subalgebra and $1 \in \mathcal{A}_{\mathbb{R}}$. Since $\mathcal{A}$ is ample we deduce that $\mathcal{A}_{\mathbb{R}}$ is ample as well and Theorem 17.4.6 implies that $\mathcal{A}_{\mathbb{R}}$ is dense in $C(K)$.

Suppose that $f \in \mathcal{A}$. Set $u = \mathbf{Re}\, f$, $v = \mathbf{Im}\, f$ so that $f = u + \boldsymbol{i}v$, $u, v \in \mathbb{R}$. There exist sequences (u_n) and (v_n) in $\mathcal{A}_{\mathbb{R}}$ such that

$$\lim_n \|u_n - u\|_\infty = \lim_n \|v_n - v\|_\infty = 0.$$

Note that $f_n = u_n + \boldsymbol{i}v_n \in \mathcal{A}$ and

$$\|f - f_n\|_\infty = \sup_{x \in K} \sqrt{|u_n(x) - u(x)|^2 + |v_n(x) - v(x)|^2} \leqslant \sqrt{\|u_n - u\|_\infty^2 + \|v_n - v\|_\infty^2}$$

so

$$\lim_n \|f_n - f\|_\infty = 0.$$

□

Example 17.4.13 (Trigonometric polynomials)**.** Denote by S^1 the unit circle

$$S^1 := \{\, z \in \mathbb{C};\ \ |z| = 1 \,\}.$$

For $n \in \mathbb{Z}$ define

$$\chi_n : S^1 \to \mathbb{C},\ \chi_n(z) = z^n$$

If we use the angular coordinate θ on S^1 so that $z = e^{\boldsymbol{i}\theta}$ for $z \in S^1$, then

$$\chi_n(\theta) = \cos n\theta + \boldsymbol{i}\sin n\theta.$$

Note that for $m, n \in \mathbb{Z}$ we have

$$\chi_n \cdot \chi_m = \chi_{m+n} \ \text{ and } \ \bar{\chi}_n = \chi_{-n}$$

Denote by $\mathcal{P}$ the complex vector space of $C\big(S^1, \mathbb{C}\big)$ spanned by the functions χ_n, $n \in \mathbb{Z}$. The elements of $\mathcal{P}$ are traditionally calllled *trigonometric polynomials*. They have the form

$$p(\theta) = \sum_{n \in bZ} c_n e^{\boldsymbol{i}n\theta}$$

where all but finitely many of the complex coefficients c_n are zero. □

The above discussion shows that $\mathcal{P}$ is a $\mathbb{C}$-subalgebra of $C\big(S^1, \mathbb{C}\big)$. By definition $1 = \chi_0 \in \mathcal{P}$. Let us observe that $\mathcal{P}$ is ample. Indeed, if $z_0, z_1 \in S^1$, $z_0 \neq z_1$, and $c_0, c_1 \in \mathbb{C}$, then the function

$$f : S^1 \to \mathbb{C}\ \ f(z) = c_0 + \frac{c_1 - c_0}{z_1 - z_0}\big(\, z - z_0 \,\big)$$

$$= \Big(\, c_0 - \frac{z_0(c_1 - c_0)}{z_1 - z_0} \,\Big)\chi_0(z) + \frac{c_1 - c_0}{z_1 - z_0}\chi_1(z) \in \mathcal{P}$$

and

$$f(z_i) = c_i,\ \ i = 0, 1.$$

We obtain the following important result.

Corollary 17.4.14 (Weierstrass)**.** *For any continuous function $f : S^1 \to \mathbb{C}$ there exists a sequence of trigonometric polynomials $p_n : S^1 \to \mathbb{C}$ that converges uniformly to f.* □

Note that a continuous function on S^1 can be identified with a continuous 2π-periodic function $\mathbb{R} \to \mathbb{C}$. The classical formulation of the above corollary is: a continuous 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$ can be arbitrarily well uniformly approximated by trigonometric polynomials.

Proposition 17.4.15. *Let (X, d) be a metric space and $K \subset X$ a compact subset. Then the Banach space $\big(\, C(K), \|-\|_\infty \,\big)$ is separable.*

Proof. The compact set K is separable. Fix a dense, countable subset $Y \subset K$. For each $y \in Y$ we define

$$\mu_y : K \to \mathbb{R},\ \ \mu_y(x) = d(x, y),\ \ \forall x \in K.$$

For any finite subset $F = \{y_1, \dots, y_n\} \subset Y$ and any function $\alpha : F \to \mathbb{N}$, $\alpha(y_i) = \alpha_i$ we define

$$\mu_{F,\alpha} = \mu_{y_1}^{\alpha_1} \cdots \mu_{y_n}^{\alpha_n} = \prod_{y \in F} \mu_y^{\alpha(y)} \in C(K).$$

We set $\mu_{\varnothing} = 1$. The collection of monomials $\mu_{F,\alpha}$, F finite subset of Y, $\alpha : F \to \mathbb{N}$ is countable and we denote by $\mathcal{A}$ the real vector subspace of $C(K)$ spanned by the monomials $\mu_{F,\alpha}$. Clearly $\mathcal{A}$ is an algebra of continuous functions. It also separates points. Indeed, if $x_0 \neq x_1$, there exists a sequence (y_n) in Y such that

$$\lim_{n\to\infty} y_n = x_0.$$

Then

$$\lim_{n\to\infty} \mu_{y_n}(x_0) = 0, \quad \lim_{n\to\infty} \mu_{y_n}(x_1) = d(x_0, x_1) > 0.$$

Thus, for n sufficiently large $\mu_{y_n}(x_0) \neq \mu_{y_n}(x_1)$. Observe that the collections of linear combinations of the form

$$\sum_{k=1}^{n} q_k \mu_{F_k,\alpha_k}, \quad q_k \in \mathbb{Q}$$

is countable and dense in $\mathcal{A}$.[5] In particular this countable collection is dense in $C(K)$.

□

Definition 17.4.16. A *Polish space* is a complete, separable metric space. □

Example 17.4.17. The Euclidean space $\mathbb{R}^n$ is a Polish space. A compact subset K of a metric space is a Polish space. The Banach space $C(K)$ is a Polish space. □

[5] Can you see why?

17.5. Exercises

Exercise 17.1. Let (X, d) be a metric space, $S \subset X$ and $x_0 \in S$. Prove that the following are equivalent.

(i) $x_0 \in \boldsymbol{int}(S)$.

(ii) There exists $r > 0$ such that $B_r(x_0) \subset S$.

Exercise 17.2. Prove Proposition 17.1.12. □

Exercise 17.3. Let (X, d) be a metric space and $S \subset X$. Prove that $\boldsymbol{cl}\big(\boldsymbol{cl}\, S\big) = \boldsymbol{cl}\, S$. □

Exercise 17.4. Let (X, d) be a metric space and $x_* \in X$. Given a sequence $(x_n)_{n\in\mathbb{N}}$ prove that the following are equivalent.

(i) The sequence $(x_n)_{n\in\mathbb{N}}$ converges to x_*.

(ii) Any subsequence of $(x_n)_{n\in\mathbb{N}}$ has a sub-subsequence that converges to x_*.

□

Exercise 17.5. Consider the space $C([0,1])$ of continuous functions $[0,1] \to \mathbb{R}$ and let

$$U := \big\{ f \in C([0,1]);\;\; f(0) > 0 \big\}.$$

(i) Prove that U is an open subset of the space $C([0,1])$ equipped with the sup-norm.

(ii) Prove that U is *not* an open subset of the space $C([0,1])$ equipped with the norm $\| - \|_1$ defined in (17.1.1).

□

Exercise 17.6. Let (X, d) be a metric space and $(x_n)_{n\in\mathbb{N}}$ a sequence of points in X. Prove that the following statements are equivalent.

(i) The sequence (x_n) converges to $x_0 \in X$.

(ii) For any open subset U of X that contains x_0 there exists $N = N_U \in \mathbb{N}$ such that, $\forall n \geqslant N_U$, the point x_n lies in U.

In other words, the concept of convergence is a topological concept. □

Exercise 17.7. Suppose that X is a set and $\mathcal{S} = (S_i)_{i\in I}$ is a collection of subsets of X. Let $U \subset X$. Prove that the following are equivalent.

(i) $U \in \mathcal{T}[\mathcal{S}]$; see Example 17.1.22(c).

(ii) For any $x \in U$ there exists a finite subset $J = J_x \subset I$ such that

$$x \in \bigcap_{j\in J_x} S_j \subset U.$$

□

Exercise 17.8. Denote by $\mathcal{D}$ the set of *dyadic* numbers inside $[0,1]$, i.e.,

$$\mathcal{D} = \bigcup_{n\in\mathbb{N}_0} \mathcal{D}_n, \quad \mathcal{D}_n = \Big\{ \frac{k}{2^n}; \;\; k\in\mathbb{N}_0, \;\; k\leqslant 2^n \Big\}.$$

Denote by $\mathcal{F}$ the collection of continuous functions $f : [0,1] \to \mathbb{R}$ with the following properties:

- $f\big(\mathcal{D}\big) \subset \mathbb{Q}$.
- There exists $n = n(f)$ such that f is linear on each of the intervals
$$\big[\,(k-1)/2^n, k/2^n\,\big], \quad k = 1,\dots,2^n.$$

(i) Prove that each function $f\in\mathcal{F}$ is uniquely determined by its restriction to $\mathcal{D}_{n(f)}$.

(ii) Prove that $\mathcal{F}$ is countable.

(iii) Prove that $\mathcal{F}$ is dense in $\big(C([0,1]), \|-\|_\infty\big)$.

□

Exercise 17.9. Suppose that (X,d) is a metric space and $A_0, A_1 \subset X$. We define the distance between A_0 and A_1 to be the number

$$\operatorname{dist}(A_0,A_1) = \inf_{(a_0,a_1)\in A_0\times A_1} d(a_0,a_1).$$

(i) Prove that
$$\operatorname{dist}(A_0,A_1) = \inf_{a_0\in A_0} \operatorname{dist}(a_0,A_1).$$

(ii) Prove that for any $x\in X$ we have
$$\operatorname{dist}(A_0,A_1) \leqslant \operatorname{dist}(x,A_0) + \operatorname{dist}(x,A_1).$$

(iii) Suppose that $X = \mathbb{R}^n$ and d is the Euclidean metric. If A_0 and A_1 are closed and A_0 is also bounded, then
$$\operatorname{dist}(A_0,A_1) = 0 \Longleftrightarrow A_0\cap A_1 \neq \varnothing.$$

(iv) Suppose that $X = \mathbb{R}^2$ and d is the Euclidean metric. Construct an example of disjoint, closed, convex subsets $A_0, A_1 \subset \mathbb{R}^2$ such that
$$\operatorname{dist}(A_0,A_1) = 0.$$

□

Exercise 17.10. Prove that an open set $U \subset \mathbb{R}^n$ is connected if and only if it is path connected. □

Exercise 17.11. Prove that a set $S\subset\mathbb{R}$ is connected in the sense of Definition 17.1.39 if and only if it is an interval. □

Exercise 17.12. Suppose that (X, d) is a metric space and $S \subset X$ is a disconnected set. Suppose that $U_0, U_1 \subset X$ are open sets such that

$$U_0 \cup U_1 \supset S, \;\; S \cap U_0 \cap U_1 = \varnothing.$$

Show that

$$U_0 \cap \boldsymbol{cl}(U_1 \cap S) = \varnothing. \qquad \square$$

Exercise 17.13. Prove Proposition 17.1.50. □

Exercise 17.14. Suppose that $(X, \|-\|_X)$ and $(Y, \|-\|_Y)$ are normed spaces. Prove that if $\dim X < \infty$, then any linear map $T : X \to Y$ is continuous. □

Exercise 17.15. Let X be a real vector space and $\|-\|_i$, $i = 0, 1$, two norms on X. Prove that the following statements are equivalent.

(i) The norms $\|-\|_0$ and $\|-\|_1$ are equivalent; see Definition 17.1.61.

(ii) A subset $U \subset X$ is open with respect to $\|-\|_0$ if and only if it is open with respect to $\|-\|_1$.

□

Exercise 17.16. Suppose that $(X, \|-\|_X)$ and $(Y, \|-\|_Y)$ are normed spaces and $T \in \boldsymbol{B}(X, Y)$. For every linear functional $\xi : Y \to \mathbb{R}$ (not necessarily continuous) we define

$$T^*\xi : X \to \mathbb{R}, \;\; T^*\xi(x) = \xi(Tx), \;\; \forall x \in X.$$

(i) Show that if ξ is continuous, so is $T^*\xi$.

(ii) Show that the induced linear operator

$$T^* : Y^* \to X^*, \;\; Y^* \ni \xi \mapsto T^*\xi \in X^*$$

is continuous.

□

Exercise 17.17. Show that the linear operator

$$T : \big(C([0,1]), \|-\|_\infty \big) \to \big(C([0,1]), \|-\|_\infty \big), \;\; f \mapsto Tf$$

given by

$$(Tf)(x) = \int_0^x f(s)\, ds$$

is continuous and $\|T\|_{\mathrm{op}} \leqslant 1$. □

Exercise 17.18. Let X be a set. Denote by $\mathbb{B}(X)$ the vector space of bounded functions $f : X \to \mathbb{R}$. For $f \in \mathbb{B}(\mathbb{X})$ we set

$$\|f\| := \sup_{x \in X} \big| f(x) \big|.$$

Suppose that $\boldsymbol{U} \subset \mathbb{B}(X)$ is a vector subspace that contains the constant functions and satisfies the following condition: for any nondecreasing sequence of nonnegative functions u_n in $\boldsymbol{U}$ such that

$$\sup_{n\in\mathbb{N}} \|u_n\| < \infty$$

the limit function

$$u_\infty(x) = \lim_{n\to\infty} u_n(x), \quad x \in X,$$

belongs to $\boldsymbol{U}$.

(i) Prove that $\big(\boldsymbol{B}(X), \|-\|\big)$ is a Banach space.

(ii) Prove that $\boldsymbol{U}$ is a closed subspace of the Banach space $\mathbb{B}(X)$.**Hint.** Suppose that (u_n) is a sequence in $\boldsymbol{U}$ that converges in the norm $\|-\|$ to a function u_∞. By extracting a subsequence you can assume $\|u_n - u_{n+1}\| < 2^{-n}$, $\forall n$. Investigate the telescoping series

$$u_1 + (u_2 - u_1) + \cdots + (u_{n+1} - u_n) + \cdots$$

to produce a nondecreasing sequence in $\boldsymbol{U}$ that converges pointwisely to u_∞.

□

Exercise 17.19. Prove that any finite dimensional vector subspace of a normed space is closed. **Hint.** Use Proposition 17.1.59 and the fact that in $\mathbb{R}^n$ any Cauchy sequence (with respect to the Euclidean norm) is convergent. □

Exercise 17.20. Denote by $C^1\big([0,1]\big)$ the space of differentiable functions $f : [0,1] \to \mathbb{R}$ with continuous derivative. For $f \in C^1\big([0,1]\big)$ we set

$$\|f\|_{C^1} := \sup_{x\in[0,1]} |f(x)| + \sup_{x\in[0,1]} |f'(x)|.$$

Prove that $\big(C^1\big([0,1]\big), \|-\|_{C^1}\big)$ is a Banach space. □

Exercise 17.21. Denote by ℓ_2 the space of sequences of real numbers

$$\underline{x} = (x_n)_{n\in\mathbb{N}} = (x_1, x_2, \dots)$$

such that

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

For $\underline{x} \in \ell_2$ we set

$$\|\underline{x}\| := \left(\sum_{n\in\mathbb{N}} x_n^2\right)^{1/2}.$$

(i) Show that $(\ell_2, \|-\|)$ is a *Banach* space.

(ii) For $n \in \mathbb{N}$ we set

$$\boldsymbol{e}_n := (\underbrace{0,\dots,0}_{n-1}, 1, 0, \dots) = (\delta_{n1}, \delta_{n2}, \dots).$$

Suppose that $\boldsymbol{\alpha} : \ell_2 \to \mathbb{R}$ is a linear functional. Set $\alpha_n := \boldsymbol{\alpha}(\boldsymbol{e}_n)$, $\forall n \in \mathbb{N}$. Prove that $\boldsymbol{\alpha}$ is continuous if and only if

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

(iii) Show that for any continuous linear functional $\boldsymbol{\alpha} : \ell_2 \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \boldsymbol{\alpha}(\boldsymbol{e}_n) = 0.$$

□

Exercise 17.22. Let $(X, \| - \|)$ be a normed spaces. A series of elements in X

$$\sum_{n\geqslant 1} x_n$$

is said to be *convergent* if the sequence of partial sums

$$S_n = \sum_{k=1}^{n} x_n$$

is convergent. The series is called *absolutely convergent* if the series of positive numbers

$$\sum_{n\geqslant 1} \|x_n\|$$

is convergent. Prove that the following are equivalent.

(i) $(X, \| - \|)$ is a Banach space.

(ii) Any absolutely convergent series of elements in X is convergent.

Hint. For (i) ⇒ (ii) have a look at the proof of Theorem 4.6.13. For (ii) ⇒ (i) use Lemma 17.2.14 and the telescoping trick: a sequence in X is convergent if the series $\sum_{n\geqslant 1}(x_n - x_{n-1})$, $x_0 = 0$, is convergent. □

Exercise 17.23. Suppose that $(X, \| - \|)$ is a normed space and $Y \subset X$ is a *finite dimensional* subspace.

(i) Prove that, for any $x_0 \in X$ there exists $y_0 \in Y$ such that

$$\|x_0 - y_0\| = \operatorname{dist}(x_0, Y) := \inf_{y\in Y} \|x_0 - y\|.$$

Hint. Consider the function $f : Y \to \mathbb{R}$, $f(y) = \|y - x_0\|$. Then argue as in Exercise 12.25 using Corollary 17.1.60.

(ii) Prove that if $Y \neq X$, then there exists $x \in X$ such that

$$1 = \|x\| = \operatorname{dist}(x, Y).$$

Hint. Choose $x_0 \in X\backslash Y$. Pick $y_0 \in Y$ such that $\|x_0 - y_0\| = \operatorname{dist}(x_0, Y)$. Show that

$$x = \frac{1}{\|x_0 - y_0\|}(x_0 - y_0),$$

will do the trick.

□

Exercise 17.24. Suppose that $(X, \| - \|)$ is an *infinite dimensional* normed space.

(i) Show that there exists a sequence of linearly independent vectors $(x_n)_{n\in\mathbb{N}}$ in X such that $\|x_n\| = 1$, $\forall n \in \mathbb{N}$. Set

$$X_0 = \{0\}, \;\; X_n := \operatorname{span}\{x_1, \dots, x_n\}, \;\; n \in \mathbb{N}.$$

(ii) Show that there exists a sequence $(\boldsymbol{e}_n)_{n\in\mathbb{N}}$ such that

$$X_n = \operatorname{span}\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}, \;\; 1 = \|\boldsymbol{e}_n\| = \operatorname{dist}(\boldsymbol{e}_n, X_{n-1}), \;\; \forall n \in \mathbb{N}.$$

Hint. Use Exercise 17.23.

(iii) Show that the sequence $(\boldsymbol{e}_n)$ contains no convergent subsequence.

□

Exercise 17.25. Suppose that $(X, \| - \|)$ is a normed space, $T \in \boldsymbol{B}(X)$ and $(T_n)_{n\in\mathbb{N}}$ is a sequence in $\boldsymbol{B}(X)$. Prove that the following are equivalent.

(i) The sequence (T_n) converges in the operator norm to T, i.e.,

$$\lim_{n\to\infty} \|T_n - T\|_{\mathrm{op}} = 0.$$

(ii) For any $\varepsilon > 0$, there exists $N = N(\varepsilon) > 0$ such that $\forall x \in X$, $\|x\| \leqslant 1$ and $\forall n > N(\varepsilon)$ we have $\|T_n x - Tx\| < \varepsilon$.

□

Exercise 17.26. Suppose that $(X, \| - \|)$ is a *Banach* space. Prove that the space $(\boldsymbol{B}(X), \| - \|_{\mathrm{op}})$ is also a Banach space. □

Exercise 17.27. Suppose that $(X, \| - \|)$ is a *Banach* space and $A \in \boldsymbol{B}(X)$.

(i) Prove that for any $t \in \mathbb{R}$ the series

$$\sum_{n\geqslant 0} \frac{t^n}{n!} A^n$$

is convergent in $\boldsymbol{B}(X)$. We denote by $E_A(t)$ its sum.[6] **Hint.** Use Exercises 17.22 and 17.26.

(ii) Prove that $E_A(t+s) = E_A(t) \cdot E_A(s)$, $\forall s, t \in \mathbb{R}$. **Hint.** Define

$$S_m(t) := \sum_{k=0}^{m} \frac{t^k}{k!} A^k.$$

Using Newton's binomial formula (3.2.4) show that

$$\| \, S_{2m}(t+s) - S_m(t)S_m(s) \, \|_{\mathrm{op}} \leqslant \sum_{n>m} \frac{(|t| + |s|)^n}{n!} \|A\|_{\mathrm{op}}^n$$

and conclude that

$$\lim_{m\to\infty} \| \, S_{2m}(t+s) - S_m(t)S_m(s) \, \|_{\mathrm{op}} = 0.$$

[6] If A were a real number then $E_A(t) = e^{tA}$.

(iii) Prove that for any $x \in X$ and any $t \in \mathbb{R}$

$$\lim_{h\to 0} \frac{1}{h}\Big(E_A(t+h)x - E_A(t)x \Big) = AE_A(t)x.$$

Hint. Use (ii).

(iv) Suppose that $\|A\|_{\mathrm{op}} < 1$. Prove that the series

$$\sum_{n\geqslant 0} A^n$$

converges in $\boldsymbol{B}(X)$ and its sum is the inverse of the operator $\mathbb{1} - A$.

(v) Prove that the set of linear homeomorphisms $X \to X$ is an open subset of $(\boldsymbol{B}(X), \|-\|_{\mathrm{op}})$.

□

Exercise 17.28. Suppose that $(X, \|-\|)$ is a *Banach* space.

(i) Prove that X is not the union of countably many *proper*[7] finite dimensional subspaces. **Hint.** Use Theorem 17.1.59 to prove that a finite dimensional subspace is closed and has empty interior if it is a proper subspace. Conclude using Theorem 17.2.19.

(ii) Prove that if X is infinite dimensional, then it does not admit a *countable* basis. **Hint.** Use (i)

□

Exercise 17.29. Suppose that $f : [0,\infty) \to \mathbb{R}$ is a continuous function such that

$$\forall x > 0, \quad \lim_{n\to\infty} f(nx) = \infty.$$

(i) Let $1 \leqslant a < b$. Show that for any $k \in \mathbb{N}$ the union

$$\bigcup_{n\geqslant k} (na, nb)$$

contains an interval of the form (r,∞) for some $r \geqslant 1$.

(ii) For $c > 0$ and $m \in \mathbb{N}$ we set

$$X_c := \big\{\, x \geqslant 1;\;\; f(x) \geqslant c \,\big\},\;\; A_{c,m} = \big\{\, x \geqslant 1;\;\; nx \in X_c,\;\; \forall n \geqslant m \,\big\}.$$

Prove that $A_{c,m}$ is closed and,

$$\bigcup_{m\in\mathbb{N}} A_{c,m} = [1,\infty).$$

(iii) Prove that for any $c > 0$, there exists $m \in \mathbb{N}$ and $1 \leqslant a < b$ such that $(a,b) \subset A_{c,m}$. **Hint.** Use Theorem 17.2.19.

[7] A subspace V of X is proper if $V \neq X$.

(iv) Prove that

$$\lim_{x\to\infty} f(x) = \infty. \tag{17.5.1}$$

Hint. Express (17.5.1) using the sets X_c.

□

Exercise 17.30. Let X denote the vector space of sequences of real numbers

$$\underline{x} = (x_1, x_2, \dots, x_n, \dots)$$

such that all but finitely many terms are 0. For $\underline{x} \in X$ we set

$$\|\underline{x}\| = \sup_n |x_n|.$$

(i) Show that $(X, \|-\|)$ is a normed space, but it is not complete.

(ii) Construct a countable basis[8] of X.

(iii) For $n \in \mathbb{N}$ define $T_n : X \to X$

$$T_n\underline{x} = (x_1, 2x_2, \dots, nx_n, 0, \dots).$$

Prove that T_n is continuous and $\|T_n\|_{\mathrm{op}} = n$.

(iv) Prove that for any $\underline{x} \in X$ the sequence $T_n\underline{x}$ converges in X. Denote by $T\underline{x}$ its limit. Show that the resulting operator $T : X \to X$ is linear but not continuous.

□

Exercise 17.31 (Banach-Steinhaus)**.** Suppose that $(X, \|-\|_X)$ and $(Y, \|-\|_Y)$ are Banach spaces and $(T_n)_{n\in\mathbb{N}}$ is a sequence of bounded linear operators $T_n : X \to Y$. Assume that $\forall x \in X$ the sequence T_nx is convergent (in Y). We denote by Tx its limit.

(i) Show that the map $T : X \to Y$, $x \mapsto Tx$ is linear.

(ii) For $x \in X$ we set

$$c(x) := \sup_{n\in\mathbb{N}} \|T_nx\|_Y.$$

Show that $c(x) < \infty$.

(iii) For $k \in \mathbb{N}$ we set

$$X_k := \big\{\, x \in X;\;\; c(x) \leqslant k \,\big\}.$$

Prove that, $\forall k \in \mathbb{N}$ the set X_k is closed and nonempty and

$$X = \bigcup_{k\in\mathbb{N}} X_k.$$

(iv) Prove that the limiting operator T is continuous. **Hint.** Show that X_k has nonempty interior for some k. Prove first that the function $x \mapsto \|Tx\|$ is bounded on some ball in X. Show that this implies that the function $x \mapsto \|Tx\|$ is also bounded on the unit ball centered at 0. Conclude that T is continuous.

[8] Compare with Exercise 17.28(ii).

□

Remark 17.5.1. The example in Exercise 17.30 shows that the conclusion (iv) of Exercise 17.31 may not hold if we do not assume that X is a *Banach space.* □

Exercise 17.32 (Riesz)**.** Suppose that $(X, \|-\|)$ is a normed space. Prove that the following are equivalent.

(i) The space X is finite dimensional.

(ii) The unit ball
$$B_1(0) := \{\, x \in X;\ \ \|x\| < 1 \,\}$$
is relatively compact.

Hint. For (ii) ⇒ (i) use Exercise 17.24. □

Exercise 17.33. Let (X_i, d_i), $i = 1, 2$, be two metric spaces and $K_i \subset oX_i$, $i = 1, 2$, are compact subsets. Prove that $K_1 \times K_2$ is a compact subset of $X_1 \times X_2$ equipped with the product metric defined in Example 17.1.2(c). □

Exercise 17.34. Let X be an *infinite dimensional* separable Banach space and $K_1, K_2 \subset X$ two nonempty compact subsets of X Prove that
$$\boldsymbol{int}\big(\, K_1 - K_2 \,\big) = \varnothing,$$
where
$$K_1 - K_2 = \{\, x_1 - x_2;\ \ x_1 \in K_1,\ \ x_2 \in K_2 \,\}.$$

□

Exercise 17.35 (Folklore)**.** Suppose that $(X, \|-\|)$ is Banach space and $S \subset X$. Prove that the following are equivalent.

(i) The set S is relatively compact.

(ii) The set S is bounded and for any $\varepsilon > 0$ there exists a *finite dimensional* subspace $Y \subset X$ such that
$$\operatorname{dist}(s, Y) < \varepsilon,\ \ \forall s \in S.$$

Hint. For (ii) ⇒ (i) use Exercise 17.23, Corollary 17.1.60 and Proposition 17.3.13. □

Exercise 17.36. Suppose that (X, d_X), (Y, d_Y) are metric spaces, (X, d) is compact. Prove that a continuous bijection $F : X \to Y$ is a homeomorphism. □

Exercise 17.37. Let (X, d) be a metric space and $K \subset X$ a compact subset. Suppose that $\mathcal{F} \subset C(K)$ is a family of continuous functions with the property that there exist $C_0, C_1 > 0$ such that
$$|f(x)| \leqslant C_0,\ \ |f(x) - f(y)| \leqslant C_1 d(x, y),\ \ \forall f \in \mathcal{F},\ \ \forall x, y \in K.$$

Prove that the family $\mathcal{F}$ is relatively compact in $C(K)$. □

Exercise 17.38. Suppose that $U \subset \mathbb{R}^m$ is a convex open set and $f_n : U \to \mathbb{R}$ is a sequence of C^1-functions such that

$$|\, f_n(\boldsymbol{x})\,| \leqslant 1, \;\; \forall \boldsymbol{x} \in U, \;\; n \in \mathbb{N}.$$

Suppose additionally that $K \subset U$ is a compact subset such that

$$\sup_{n \in \mathbb{N}} \sup_{\boldsymbol{x} \in U} \|\nabla f_n(\boldsymbol{x})\| < \infty,$$

where $\| - \|$ denotes the Euclidean norm in $\mathbb{R}^m$. Prove that (f_n) contains a subsequence that converges uniformly on K. □

Exercise 17.39. Suppose that $f, g : [0,1] \to \mathbb{R}$ are two continuous functions such that

$$\int_0^1 f(x)x^n dx = \int_0^1 g(x)x^n dx, \;\; \forall n \in \mathbb{N}_0.$$

Show that $f(x) = g(x)$, $\forall x \in [0,1]$. **Hint.** Use Corollary 17.4.11 and Exercise 9.9. □

Exercise 17.40. Suppose that (X,d) is a compact metric space and $\mathcal{A} \subset C(X)$ is an algebra of continuous functions. Prove that its closure in $C(X)$ with respect to the sup-norm is also an algebra of functions. □

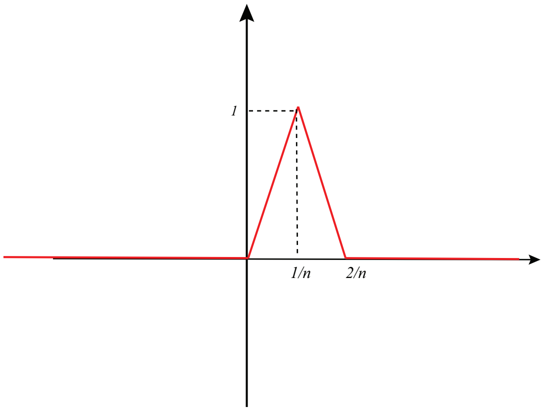

Figure 17.3. *The graph of f_n.*

Exercise 17.41. For each $n \in \mathbb{N}$ consider the continuous function $f_n : \mathbb{R} \to \mathbb{R}$ whose graph is depicted in Figure 17.3. Prove that f_n converges pointwisely to the 0 as $n \to \infty$ but the convergence is not uniform on $[0,1]$. □

Exercise 17.42 (Buchanan-Hildebrand)**.** Suppose that $f_n \in C([0,1])$ is a sequence of *continuous nondecreasing* functions that converges pointwisely to a *continuous* function $f : [0,1] \to \mathbb{R}$.

(i) Prove that if (x_n) is a sequence in $[0,1]$ such that $x_n \to x_0$, then $f_n(x_n) \to f(x_0)$.

(ii) Show that (f_n) converges *uniformly* to f.[9] **Hint.** Use (i).

□

Exercise 17.43. Suppose that (X,d) is a nonempty metric space and $f : X \to \mathbb{R}$ is a function with the following properties.

- f is *lower semicontinuous*, i.e., for any $c \in \mathbb{R}$ the set
$$\{\, f \leqslant c \,\} := \{x \in X;\ \ f(x) \leqslant c\,\}$$
is closed.
- f is *coercive*, i.e., for any $c \in \mathbb{R}$ the set $\{\, f \leqslant c \,\}$ is precompact.

Prove that there exists $x_0 \in X$ such that $f(x_0) \leqslant f(x)$, $\forall x \in X$.[10] **Hint.** Show that $\exists c \in \mathbb{R}$ such that $\{f \leqslant c\} = \varnothing$ and conclude that $\inf_{x \in X} f(x) > -\infty$. □

Exercise 17.44. Suppose that $K \in C^1\big(\mathbb{R}^2\big)$. For any $f \in C([0,1])$ denote by $T_K f$ the function
$$T_K f : [0,1] \to \mathbb{R},\ \ T_K f(t) = \int_0^1 K(t,s) f(s) ds.$$

(i) Show that T_K defines a bounded linear operator $T_K : C([0,1]) \to C([0,1])$.

(ii) Suppose that $(f_n)_{n \geqslant 1}$ is a bounded sequence in $C([0,1])$, i.e., there exists $C > 0$ such that
$$\forall n \in \mathbb{N},\ \ \sup_{t \in [0,1]} |f_n(t)| \leqslant C.$$
Prove that the sequence $g_n = T_K f_n$ admits a subsequence that converges uniformly on $[0,1]$. **Hint.** Use Corollary 17.4.4.

(iii) Show that $\ker\big(\mathbb{1} - T_K\big)$ is finite dimensional. **Hint.** Use Exercise 17.32.

□

Exercise 17.45. Suppose that (X,d) is a compact metric space. Let $C(X)$ denote the Banach space of continuous functions $X \to \mathbb{R}$ with norm
$$\|f\| = \sup_{x \in X} |f(x)|.$$
An *ideal* of $C(X)$, is a vector subspace $\mathcal{I} \subset C(X)$ such that
$$\forall f \in C(X),\ \ g \in \mathcal{I};\ \ f \cdot g \in \mathcal{I}.$$
The ideal is called *closed* if it is closed as a subset of the Banach space $C(X)$. For any ideal $\mathcal{I}$ we set
$$Z(\mathcal{I}) := \{\, x \in X;\ \ f(x) = 0,\ \ \forall f \in \mathcal{I}\,\}.$$

[9] Compare with Dini's Theorem 17.4.5.

[10] The result in Exercise 17.43 is a version of the *Fundamental Lemma of the Calculus of Variations.*

(i) Let $C \subset X$ be a closed subset and set

$$\mathcal{V}(C) := \big\{ f \in C(X);\;\; f(x) = 0,\;\; \forall x \in C \big\}.$$

Prove that $\mathcal{I}(C)$ is a closed ideal and $Z\big(\mathcal{I}(C)\big) = C$.

(ii) Prove that for any ideal $\mathcal{I}$, the set $Z(\mathcal{I}) \subset X$ is a closed subset of X and $\mathcal{V}\big(Z(\mathcal{I})\big) \supset \boldsymbol{cl}\big(\mathcal{I}\big)$.

(iii) Let $\mathcal{I}$ be an ideal. Prove that for any closed set $S \subset X$ such that $S \cap Z(\mathcal{I}) = \varnothing$ there exists a *nonnegative* function $g_S \in \mathcal{I}$ such that $g_S(x) > 0$, $\forall x \in S$. **Hint.** Use the compactness of X to show that there exist functions $g_1, \dots, g_m \in \mathcal{I}$ such that $g_1(x)^2 + \cdots + g_m(x)^2 > 0$, $\forall x \in S$.

(iv) Let $\mathcal{I}$ be an ideal and $\mathcal{V} = \mathcal{V}(\mathcal{I})$. Let $f \in \mathcal{I}(\mathcal{V})$. For $\varepsilon > 0$ we set $S_\varepsilon := \{|f| \geqslant \varepsilon\}$ and denote by g_ε the nonnegative function $g_{S_\varepsilon} \in \mathcal{I}$ found in (iii). Set $h_{\varepsilon,n} := \frac{ng_\varepsilon}{1+ng_\varepsilon}$. Prove that there exists $N = N(\varepsilon, f) > 0$ such that

$$\| f - f h_{\varepsilon,n} \| < \varepsilon,\;\; \forall n < N.$$

(v) Show that for any ideal $\mathcal{I}$ of $C(X)$ we have $\mathcal{I}\big(\mathcal{V}(\mathcal{I})\big) = \boldsymbol{cl}\big(\mathcal{I}\big)$.

□

Exercise 17.46 (Gelfand)**.** Suppose that (X, d) is a compact metric space. Let $C(X)$ denote the Banach space of continuous functions $X \to \mathbb{R}$ with norm

$$\|f\| = \sup_{x \in X} |f(x)|.$$

An ideal $\mathcal{I} \subset C(X)$ is called *maximal* if $\mathcal{I} \neq C(X)$ and there exists no ideal eJ such that $\mathcal{I} \subsetneq \mathcal{J} \subsetneq C(X)$. Denote by $\mathcal{M}$ the set of maximal ideals.

(i) For $x_0 \in X$ we denote by $\mathcal{I}_{x_0}$ the ideal consisting of continuous functions f such $f(x_0) = 0$. Prove that $\mathcal{I}_{x_0}$ is a maximal ideal.

(ii) Prove that the map

$$X \ni x_0 \mapsto \mathcal{I}_{x_0} \in \mathcal{M}$$

is a bijection.

Hint. Use Exercise 17.45. □

Remark 17.5.2. Exercise 17.45 is sometimes referred to as *topological Nullstellensatz* since it closely resembles Hilbert's famous algebraic result with the same name.

We denote by $\operatorname{Ideal}(X)$ the set of closed ideals of $C(X)$, and by $\mathscr{C}_X$ the family of closed subsets of X. The above result shows that we have a bijection

$$\mathscr{C}_X \ni C \mapsto \mathcal{I}(C) \in \operatorname{Ideal}(X)$$

with inverse

$$\operatorname{Ideal}(X) \ni \mathcal{I} \mapsto Z(\mathcal{I}) \in \mathscr{C}_X.$$

This is a Galois correspondence, i.e.,

$$C_1 \subsetneq C_2 \Longleftrightarrow \mathcal{I}(C_1) \supsetneq \mathcal{I}(C_2),\ \ \mathcal{I}_1 \subsetneq \mathcal{I}_2 \Longleftrightarrow Z(\mathcal{I}_1) \supsetneq Z(\mathcal{I}_1).$$

The concept of maximal ideal in Exercise 17.46 is purely *algebraic* because it can be defined for any commutative ring. One of the conclusions of this exercise is that the maximality in the ring $C(X)$ carries topological information. Indeed, since any maximal ideal is of the form $\mathcal{I}_{x_0}$ we deduce that any maximal ideal is a *closed* ideal. We can define the closure of a set $S \subset X$ to be

$$\boldsymbol{cl}(S) = Z\left(\bigcap_{x\in S} \mathcal{I}_x\right).$$

□

Exercise 17.47. Set $\mathbb{B} := \{0,1\}$, and denote by $\mathbb{X}$ the space of functions $f : \mathbb{N} \to \mathbb{B}$. We define a metric on $\mathbb{X}$ by setting

$$d(f,g) = \sum_{n\in\mathbb{N}} \frac{1}{2^n}\bigl|\, f(n) - g(n)\,\bigr|.$$

(i) For $r \in (0,1)$ we denote by Σ_r the *sphere* of center 0 and radius r in $\mathbb{X}$,

$$\Sigma_r = \left\{\, f \in \mathbb{X};\ \ \sum_{n\in\mathbb{N}} \frac{f(n)}{2^n} = r \,\right\}.$$

Prove that $\Sigma_{1/2}$ consists of two points, while $\Sigma_{1/3}$ consists of a single point.

(ii) Let $(f_\nu)_{\nu\in\mathbb{N}}$ be a sequence in $\mathbb{X}$. Prove that $d(f_\nu, f) \to 0$ as $\nu \to \infty$ if and only if, for any $k \in \mathbb{N}$, there exists $N = N_k \in \mathbb{N}$ such that

$$\forall \nu \geqslant N_k,\ \ \forall i \leqslant k\ \ f_\nu(i) = f(i).$$

(iii) Given $m \in \mathbb{N}$ and a subset $S \subset \mathbb{B}^m$ we set

$$\mathscr{C}_S := \{\ \ f \in \mathbb{X};\ \ \bigl(\, f(1), \dots, f(m)\,\bigr) \in S\,\}.$$

Prove that $\mathscr{C}_S$ is both closed and open in $\mathbb{X}$.

(iv) Prove that the metric space $\mathbb{X}$ is compact. **Hint.** Use (ii) to show that any sequence in $\mathbb{X}$ contains a convergent subsequence. Use Cantor's diagonal subsequence trick also employed in the proof of Theorem 17.2.13.

□

Exercise 17.48. Let $\boldsymbol{T} \in C([0,1])$ denote the vector subspace spanned by the functions

$$e_n : [0,1] \to \mathbb{R},\ \ e_n(x) = \cos(\pi n x),\ \ n = 0, 1, 2, \dots .$$

Prove that T is an $\mathbb{R}$-algebra of functions dense in the Banach space $\bigl(\, C([0,1]), \|-\|_\infty\,\bigr)$. □

Exercise 17.49. Denote by $C([0,\infty])$ the vector space of continuous functions $[0,\infty)\to\mathbb{R}$ that have finite limit at ∞. For $f\in C([0,\infty])$ we set

$$\|f\| := \sup_{t\geqslant 0}|f(t)|.$$

(i) Prove that $(C([0,\infty]),\|-\|)$ is Banach space.

(ii) Prove that the family of functions $e_\lambda\in C([0,\infty])$, $\lambda\geqslant 0$, $e_\lambda(t)=e^{-\lambda t}$, $\forall t\geqslant 0$, spans a vector subspace dense is $C([0,\infty])$ with respect to the norm $\|-\|$.

□

17.6. Exercises for extra credit

Exercise* 17.1. Consider the subset S of $\mathbb{R}^2$ defined by (see Figure 17.4)

$$S := \big\{\,\big(x,\sin(1/x)\big); x>0\,\big\}\cup\big\{(0,0),(0,1),(0,-1)\,\big\}.$$

Prove that S is connected yet it is not path connected. □

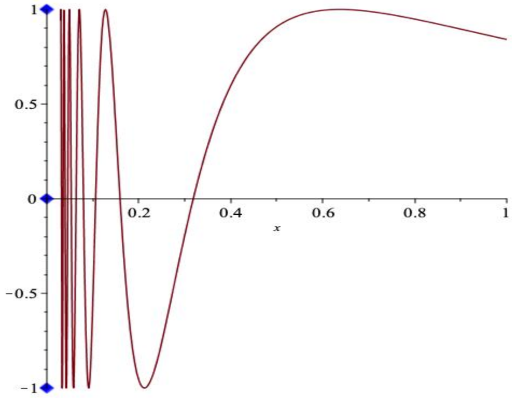

Figure 17.4. *The graph of* $\sin(1/x)$ $x>0$ *with three points added.*

Exercise* 17.2. Suppose that $(X,\|-\|)$ is a normed space and $\alpha : X\to\mathbb{R}$ a linear function. Set

$$Z := \ker\alpha = \big\{\,x\in X;\;\;\alpha(x)=0\,\big\}.$$

Prove that the following are equivalent.

(i) The function α is continuous.

(ii) The subset Z is closed.

□

Exercise* 17.3. Suppose that $(X, \| - \|)$ is a real normed space and $\alpha_1, \dots, \alpha_n \in X^*$ are continuous linear functionals on X. Suppose that $\alpha \in X^*$ is another continuous linear functional such that $\alpha(x) = 0$ for any $x \in X$ such that $\alpha_1(x) = \dots = \alpha_n(x) = 0$. Prove that there exist constants $c_1, \dots, c_n \in \mathbb{R}$ such that

$$\alpha = \sum_{k=1}^{n} c_k \alpha_k.$$

□

Exercise* 17.4. Let $(X, \| - \|)$ be a normed space. Prove that the following are equivalent.

(i) $\dim X < \infty$.

(ii) Any other norm on X is equivalent to the norm $\| - \|$.

□

Exercise* 17.5. Let $(X, \| - \|)$ be a Banach space and $A \in \boldsymbol{B}(X)$. Prove that for any $t \in \mathbb{R}$ we have

$$\lim_{n \to \infty} \left\| \left(\mathbb{1} + \frac{t}{n} A \right)^n - E_A(t) \right\|_{\mathrm{op}} = 0,$$

where $\mathbb{1}$ denotes the identity operator $X \to X$ and $E_A(t)$ is the exponential defined in Exercise 17.27. □

Chapter 18

An Introduction to Ordinary Differential Equations

Differential equations have been investigated since the dawn of calculus, as they have appeared in many questions from physics. Their study contributed in a rather substantial fashion to the development of modern analysis, and conversely, the developments in analysis provided more and more powerful tools for investigating such equations. In the meantime they found applications in other branches of mathematics, science and economics.

The field of differential equations is very broad and the many different classes of equations or types of questions require very different techniques. The present chapter has a rather modest goal, to introduce you to the *rigorous* foundations of this subject. We will cover only a few topics: local existence and uniqueness, global existence and uniqueness, continuous dependence of data, linear systems of differential equations. These topics are absolutely necessary for any more in-depth investigation of this topic.

Our presentation follows closely the wonderfully efficient book of Viorel Barbu [4], but we will cover only very few of the topics of that book.

18.1. Basic concepts and examples

18.1.1. The concept of differential equation. Loosely speaking, a *differential equation* is an equation whose unknown is a function depending on one or several variables and describing a relationship between this function and its derivatives up to a certain order. The highest order of the derivatives of the unknown function that are involved in this equation is called the *order* of the differential equation. If the unknown function depends on several variables, then the equation is called a *partial differential equation*, or *p.d.e.*. If the unknown function depends on a single variable, the equation is called an *ordinary differential equation*, or *o.d.e.*.

A first order o.d.e. has the general form

$$F(t, x, x') = 0, \tag{18.1.1}$$

where t is the argument of the unknown function $x = x(t)$, $x'(t) = \frac{dx}{dt}$ is its derivative, and F is a real valued function defined on a domain of the space $\mathbb{R}^3$.

We define a *solution* of (18.1.1) on the interval $I = (a, b)$ of the real axis to be a continuously differentiable function $x : I \to \mathbb{R}$ that verifies the equation (18.1.1) on I, i.e.,

$$F\big(\,t, x(t), x'(t)\,\big) = 0, \quad \forall t \in I.$$

When I is an interval of the form $[a, b]$, $[a, b)$ or $(a, b]$, the concept of solution on I is defined similarly.

In certain situations, the implicit function theorem allows us to reduce (18.1.1) to an equation of the form

$$x' = f(t, x), \tag{18.1.2}$$

where $f : \Omega \to \mathbb{R}$, with Ω an open subset of $\mathbb{R}^2$. In the sequel we will investigate exclusively equations in the form (18.1.2), henceforth referred to as *normal form*.

From a geometric viewpoint, a solution of (18.1.1) is a curve in the (t, x)-plane, having at each point a tangent line that varies continuously with the point. Such a curve is called an *integral curve* of the equation (18.1.1). In general, the set of solutions of (18.1.1) is infinite and we will (loosely) call this set the *general solution*.

We can specify a solution of (18.1.1) by imposing certain conditions. The most frequently used is the *initial condition* or *Cauchy condition*

$$x(t_0) = x_0, \tag{18.1.3}$$

where $t_0 \in I$ and $x_0 \in \mathbb{R}$ are a priori given and are called *initial values*.

The *Cauchy problem* associated to (18.1.1) asks to find a solution $x = x(t)$ of (18.1.1) satisfying the initial condition (18.1.3). Geometrically, the Cauchy problem amounts to finding an integral curve of (18.1.1) that passes through a given point $(t_0, x_0) \in \mathbb{R}^2$.

The above discussion extends naturally to first order *differential systems* of the form

$$x_i' = f_i(t, x_1, \ldots, x_n), \;\; i = 1, \ldots, n, \;\; t \in I, \tag{18.1.4}$$

where $f_1, \ldots, f_n$ are functions defined on an open subset of $\mathbb{R}^{n+1}$. By solution of the system (18.1.4) we understand a collection of continuously differentiable functions $\{x_1(t), \ldots, x_n(t)\}$ on the interval $I \subset \mathbb{R}$ that satisfy (18.1.4) on this interval, i.e.,

$$x_i'(t) = f_i\big(\,t, x_1(t), \ldots, x_n(t)\,\big), \;\; i = 1, \ldots, n, \;\; t \in I, \tag{18.1.5a}$$

$$x_i(t_0) = x_i^0, \;\; i = 1, \ldots, n, \tag{18.1.5b}$$

where $t_0 \in I$ and $(x_1^0, \dots, x_n^0)$ is a given point in $\mathbb{R}^n$. Just as in the scalar case, we will refer to (18.1.5a)-(18.1.5b) as the *Cauchy problem* associated to (18.1.4). The above system can be written more succinctly by considering the map

$$\boldsymbol{F} : I \times \mathbb{R}^n \to \mathbb{R}^n, \;\; \boldsymbol{F}(t, \boldsymbol{x}) = \begin{bmatrix} f_1(t, \boldsymbol{x}) \\ \vdots \\ f_n(t, \boldsymbol{x}) \end{bmatrix}, \;\; \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Then (18.1.5a)-(18.1.5b) can be rewritten as

$$\boldsymbol{x}'(t) = \boldsymbol{F}\bigl(\, t, \boldsymbol{x}(t)\,\bigr), \;\; \boldsymbol{x}(t_0) = \boldsymbol{x}^0, \;\; \boldsymbol{x} : I \to \mathbb{R}^n. \tag{18.1.6}$$

When the map $\boldsymbol{F}$ is independent of t, the system of equations is called *autonomous*. Any non-autonomous system

$$\boldsymbol{x}'(t) = \boldsymbol{F}\bigl(\, t, \boldsymbol{x}(t)\,\bigr)$$

can be converted to an autonomous one using the following simple trick. Introduce new variables

$$\boldsymbol{y} = (y_0, y_1, \dots, y_n)$$

and the new map

$$\widehat{\boldsymbol{F}}(y) = \begin{bmatrix} \hat{F}_0(\boldsymbol{y}) \\ \hat{F}_1(\boldsymbol{y}) \\ \vdots \\ \hat{F}_n(\boldsymbol{y}) \end{bmatrix} = \begin{bmatrix} 1 \\ f_1(y_0, y_1, \dots, y_n) \\ \vdots \\ f_n(y_0, y_1, \dots, y_n) \end{bmatrix}.$$

Then $\boldsymbol{x}(t)$ is a solution of (18.1.6) if and only iff $\boldsymbol{y}(t) = \bigl(\, t, \boldsymbol{x}(t)\,\bigr)$ is a solution of

$$\boldsymbol{y}'(t) = \widehat{F}(\boldsymbol{y}), \;\; y_0(t_0) = t_0, \;\; y_k(t_0) = x_k^0, \;\; \forall k = 1, \dots, k.$$

Let us mention a simple fact that we will use frequently in the sequel namely that the Cauchy problem (18.1.6) is equivalent to the *integral equation*

$$\boldsymbol{x}(t) = \boldsymbol{x}^0 + \int_{t_0}^{t} \boldsymbol{F}\bigl(\, s, \boldsymbol{x}(s)\,\bigr) ds.$$

Indeed, this is a simple application of the Fundamental Theorem of Calculus.

From a geometric point of view, a solution of the system (18.1.4) is a path in the space $\mathbb{R}^n$. In many situations or phenomena modeled by differential systems of the type (18.1.4), the collection $\bigl(\, x_1(t), \dots, x_n(t)\,\bigr)$ represents the coordinates of the state of a system at time t, and thus the trajectory $t \mapsto \bigl(\, x_1(t), \dots, x_n(t)\,\bigr)$ describes the evolutions of that particular system. For this reasons, the solutions of a differential system are often called the *trajectories* of the system.

Consider now ordinary differential equations of order n, that is, having the form

$$F\bigl(\, t, x, x', \dots, x^{(n)}\,\bigr) = 0, \tag{18.1.7}$$

where F is a given function. Assuming it is possible to solve for $x^{(n)}$, we can reduce the above equation to its *normal form*

$$x^{(n)} = f\big(t, x, \dots, x^{(n-1)}\big). \tag{18.1.8}$$

By solution of (18.1.8) on the interval I we understand a function of class C^n on I (that is, a function n-times differentiable on I with continuous derivatives up to order n) that verifies (18.1.8) at every $t \in I$. The Cauchy problem associated to (18.1.8) asks to find a solution of (18.1.8) that satisfies the conditions

$$x(t_0) = x_0^0, \;\; x'(t_0) = x_1^0, \dots, x^{(n-1)}(t_0) = x_{n-1}^0, \tag{18.1.9}$$

where $t_0 \in I$ and $x_0^0, x_1^0, \dots, x_{n-1}^0$ are given.

Via a simple transformation we can reduce the equation (18.1.8) to a system of type (18.1.4). To this aim, we introduce the new unknown functions $x_1, \dots, x_n$ using the unknown function x by setting

$$x_1 := x, \;\; x_2 := x', \dots, x_n := x^{(n-1)}. \tag{18.1.10}$$

With these notations, the equation (18.1.8) becomes the differential system

$$\begin{array}{rcl} x_1' & = & x_2 \\ x_2' & = & x_3 \\ \vdots & \vdots & \vdots \\ x_n' & = & f(t, x_1, \dots, x_n). \end{array} \tag{18.1.11}$$

Conversely, any solution of (18.1.11) defines via (18.1.10) a solution of (18.1.8). The change in variables (18.1.10) transforms the initial conditions (18.1.9) into

$$x_i(t_0) = x_{i-1}^0, \;\; i = 1, \dots, n.$$

The above procedure can also be used to transform differential systems of order n (that is, differential systems containing derivatives up to order n) into differential systems of order 1.

Most differential equations cannot be solved explicitly. There are though a few classical classes of differential equations whose solutions can be determined "by hand". We describe below some of these situations.

18.1.2. Separable equations. These are equations of the form

$$\frac{dx}{dt} = f(t)g(x), \;\; x = x(t), \;\; t \in I = (a, b), \tag{18.1.12}$$

where f is a continuous function on (a, b) and g is a continuous function on a, possibly unbounded, interval (x_1, x_2).

Here is the classical approach to this type of equations. Multiplying both sides by dt we can rewrite (18.1.12) as

$$\frac{dx}{g(x)} = f(t)dt.$$

Integrating from t_0 to t, where t_0 is an arbitrary point in I we deduce

$$\int_{x_0}^{x(t)} \frac{du}{g(u)} \overset{u=x(s)}{=} \int_{t_0}^{t} \frac{x'(s)ds}{g(x(s))} = \int_{t_0}^{t} f(s)ds. \tag{18.1.13}$$

We set

$$G(x) := \int_{x_0}^{x} \frac{du}{g(u)}. \tag{18.1.14}$$

The function G is obviously continuous and monotone on the interval (x_1, x_2). It is thus invertible and its inverse has the same properties. We can rewrite (18.1.13) as

$$x(t) = G^{-1}\left(\int_{t_0}^{t} f(s)ds \right), \quad t \in I. \tag{18.1.15}$$

We have thus obtained a formula describing the solution of (18.1.12) satisfying the Cauchy condition $x(t_0) = x_0$.

If the above argument sounds fishy[1], here is an alternate one. Note that (18.1.12) can be rewritten as

$$\frac{d}{dt}G\big(x(t)\big) = G'\big(x(t)\big)\frac{dx}{dt} = f(t).$$

Hence $G(x(t))$ is an antiderivative of $f(t)$.

Conversely, a function x given by the equality (18.1.15) is continuously differentiable on I and its derivative satisfies

$$x'(t) = \frac{f(t)}{G'(x)} = f(t)g(x).$$

In other words, x is a solution of (18.1.12). Of course, $x(t)$ is only defined for those values of t such that $\int_{t_0}^{t} f(s)ds$ lies in the range of G.

By way of illustration consider the o.d.e.

$$x' = (2-x)\tan t, \quad t \in \left(0, \frac{\pi}{2}\right).$$

Arguing as in the general case, we rewrite this equation in the form

$$\frac{dx}{2-x} = \tan t\, dt.$$

We integrate the above equality

$$\int_{x_0}^{x} \frac{d\theta}{2-\theta} = \int_{t_0}^{t} \tan s\, ds, \quad t_0, t \in \left(0, \frac{\pi}{2}\right), \quad x(t_0) = x_0,$$

and we deduce

$$-\ln\frac{|x(t)-2|}{|x_0-2|} = -\ln\frac{|\cos t|}{|\cos t_0|}.$$

If we set $C := (x_0-2)\cos t_0$ we deduce that the general solution is given by

$$x(t) = C\cos t + 2, \quad t \in \left(0, \frac{\pi}{2}\right),$$

[1] Why can one treat the derivative $\frac{dx}{dt}$ as if it were a genuine fraction?

where C is an arbitrary constant.

Let us consider the initial value problem

$$x'(t) = 1 + x^2, \;\; x(0) = 0.$$

We have

$$d\big(\arctan x\,\big) = \frac{dx}{1+x^2} = dt$$

so

$$\arctan x(t) = t + C.$$

Since $x(0) = 0$ we deduce $C = \arctan(0) = 0$ so $x(t) = \tan t$. Note that

$$\lim_{t\to\pm\pi/2} x(t) = \pm\infty, \tag{18.1.16}$$

so the solution $x(t)$ is well defined only on the interval $(-\pi/2, \pi/2)$. The blow-up phenomenon described by (18.1.16) is not an oddity. It occurs in many other situations and it is rather something to be expected.

18.1.3. Homogeneous equations. Consider the differential equation

$$x' = h(x/t), \tag{18.1.17}$$

where h is a continuous function defined on an interval (h_1, h_2). We will assume that $h(r) \neq r$ for any $r \in (h_1, h_2)$. The equation (18.1.17) is called a *homogeneous* differential equation. It can be solved by introducing a new unknown function u defined by the equality $x = tu$. The new function u satisfies the separable differential equation

$$tu' = h(u) - u$$

which can be solved by the method described in Subsection 18.1.2. We have to mention that many first order o.d.e.-s can be reduced by simple substitutions to separated or homogeneous differential equations.

Consider for example the differential equation

$$x' = \frac{at + bx + c}{a_1t + b_1x + c_1},$$

where a, b, c and a_1, b_1, c_1 are constants. This equation can be reduced to a homogeneous equation of the form

$$\frac{dy}{ds} = \frac{as + by}{a_1s + b_1y}$$

by making the change in variables

$$s := t - t_0, \;\; y := x - x_0,$$

where (t_0, x_0) is a solution of the linear algebraic system

$$at_0 + bx_0 + c = a_1t_0 + b_1x_0 + c_1 = 0.$$

18.1.4. First order linear differential equations. Consider the differential equation

$$x' = a(t)x + b(t), \tag{18.1.18}$$

where a and b are continuous functions on the, possibly unbounded, interval (t_1, t_2). To solve (18.1.18) we multiply both sides of this equation by

$$\exp\left(-\int_{t_0}^{t} a(s)ds\right),$$

where t_0 is some point in (t_1, t_2). We obtain

$$\frac{d}{dt}\left(\exp\left(-\int_{t_0}^{t} a(s)ds\right)x(t)\right) = b(t)\exp\left(-\int_{t_0}^{t} a(s)ds\right).$$

Hence, the general solution of (18.1.18) is given by

$$x(t) = \exp\left(\int_{t_0}^{t} a(s)ds\right)\left(x_0 + \int_{t_0}^{t} b(s)\exp\left(-\int_{t_0}^{s} a(\tau)d\tau\right)ds\right), \tag{18.1.19}$$

where x_0 is an arbitrary real number. Conversely, derivating (18.1.19) we deduce that the function x defined by this equality is the solution of (18.1.18) satisfying the Cauchy condition $x(t_0) = x_0$.

Consider now the differential equation

$$x' = a(t)x + b(t)x^{\alpha}, \tag{18.1.20}$$

where α is a real number not equal to 0 or 1. The equation (18.1.20) is called a *Bernoulli type* equation and can be reduced to a linear equation using the substitution $y = x^{1-\alpha}$.

Indeed, $x = y^{1/(1-\alpha)}$ so

$$x' = \frac{1}{1-\alpha} y^{\frac{\alpha}{1-\alpha}} y'$$

$$ax + bx^{\alpha} = ay^{\frac{1}{1-\alpha}} + by^{\frac{\alpha}{1-\alpha}}.$$

Hence

$$\frac{1}{1-\alpha} y^{\frac{\alpha}{1-\alpha}} y' = ay^{\frac{1}{1-\alpha}} + by^{\frac{\alpha}{1-\alpha}},$$

and we conclude

$$y' = (1-\alpha)ay + (1-\alpha)b.$$

18.1.5. Riccati equations. Named after *Jacopo Riccati* (1676-1754), these equations have the general form

$$x' = a(t)x + b(t)x^2 + c(t), \quad t \in I, \tag{18.1.21}$$

where a, b, c are continuous functions on the interval I. In general, the equation (18.1.21) is not solvable by quadratures but it enjoys several interesting properties which we will dwell upon later. Here we only want to mention that if we know a particular solution $\varphi(t)$ of (18.1.21), then using the substitution $y = x - \varphi$ we can reduce the equation (18.1.21) to a Bernoulli type equation (18.1.20) in y.

Indeed, we have $x = y + \varphi$ so

$$(y+\varphi)' = a(y+\varphi) + b(y+\varphi)^2 + c = a\varphi + by^2 + 2b\varphi y + b\varphi^2 + c.$$

Using the equality $\varphi' = a\varphi + b\varphi^2 + c$ we deduce

$$y' = \underbrace{(a+2b\varphi)}_{A} y + b\varphi^2 = Ay + b\varphi^2.$$

We leave to the reader the task of verifying this fact.

18.1.6. Lagrange equations. These are equations of the form

$$x = t\varphi(x') + \psi(x'), \tag{18.1.22}$$

where φ and ψ are two continuously differentiable functions defined on a certain interval of the real axis such that $\varphi(p) \neq p$, $\forall p$. Assuming that x is a solution of (18.1.22) on the interval $I \subset \mathbb{R}$, we deduce after differentiating that

$$x' = \varphi(x') + t\varphi'(x')x'' + \psi'(x')x'', \tag{18.1.23}$$

where $x'' = \frac{d^2x}{dt^2}$. We denote by p the function x' and we observe that (18.1.23) implies that

$$p = \varphi(p) + t\varphi'(p)\frac{dp}{dt} + \psi'(p)\frac{dp}{dt},$$

$$\frac{dp}{dt}\big(\,t\varphi'(p) + \psi'(p)\,\big) = p - \varphi(p),$$

$$\frac{dt}{dp} = \frac{1}{\frac{dp}{dt}} = \frac{\varphi'(p)}{p-\varphi(p)}t + \frac{\psi'(p)}{p-\varphi(p)}. \tag{18.1.24}$$

We can interpret (18.1.24) as a linear o.d.e. with unknown t, viewed as a function of p. Solving this equation using formula (18.1.19) we obtain for t an expression of the form

$$t = A(p, C), \tag{18.1.25}$$

where C is an arbitrary constant. Using this in (18.1.22) we deduce that

$$x = A(p, C)\varphi(p) + \psi(p). \tag{18.1.26}$$

If we interpret p as a parameter, the equalities (18.1.25) and (18.1.26) define a parametrization of the curve in the (t, x)-plane described by the graph of the function x. In other words, the above method leads to a parametric representation of the solution of (18.1.22).

18.1.7. Clairaut equations. Named after *Alexis C. Clairaut* (1713-1765), these equations correspond to the degenerate case $\varphi(p) \equiv p$ of (18.1.22) and they have the form

$$x = tx' + \psi(x'). \tag{18.1.27}$$

Derivating the above equality we deduce

$$x' = tx'' + x' + \psi'(x')x''$$

and thus

$$x''\big(\,t + \psi'(x')\,\big) = 0. \tag{18.1.28}$$

We distinguish two types of solutions. The first type is defined by the equation $x'' = 0$. Hence

$$x = C_1 t + C_2, \tag{18.1.29}$$

where C_1 and C_2 are arbitrary constants. Using (18.1.29) in (18.1.27) we see that C_1 and C_2 are not independent but are related by the equality

$$C_2 = \psi(C_1).$$

Therefore

$$x = C_1 t + \psi(C_1), \tag{18.1.30}$$

where C_1 is an arbitrary constant. This is the *general solution* of the Clairaut equation. .

A second type of solutions is obtained from (18.1.28),

$$t + \psi'(x') = 0. \tag{18.1.31}$$

Proceeding as in the case of Lagrange equations, we set $p := x'$ and we obtain from (18.1.31) and (18.1.27) the parametric equations

$$\begin{array}{rcl} t &=& -\psi'(p) \\ x &=& -\psi'(p)p + \psi(p) \end{array} \tag{18.1.32}$$

that describe a function called the *singular solution* of the Clairaut equation (18.1.27). It is not difficult to see that the solution (18.1.32) does not belong to the family of solutions (18.1.30). Geometrically, the curve defined by (18.1.32) is the envelope of the family of lines described by (18.1.30).

18.1.8. Integral inequalities. This subsection is devoted to the investigation of the following ubiquitous linear integral inequality

$$x(t) \leqslant b(t) + \int_{t_0}^{t} a(s)x(s)ds, \quad t \in [a, b]. \tag{18.1.33}$$

We assume that

- the functions $xa, b : [t_0, t_1] \to \mathbb{R}$ are continuous on $[t_0, t_1]$ and
- $a(t) \geqslant 0$, $\forall t \in [t_0, t_1]$.

The next result, usually referred to as *Gronwall's inequality* is the main tool for obtaining a priori estimates of solutions of o.d.e.-s.

Lemma 18.1.1 (Gronwall). *If the above conditions are satisfied, then $x(t)$ satisfies the inequality*

$$x(t) \leqslant b(t) + \int_{t_0}^{t} a(s)b(s)e^{(\,A(t)-A(s)\,)}ds, \tag{18.1.34}$$

where

$$A(t) = \int_{t_0}^{t} a(\tau)d\tau.$$

Proof. We set

$$y(t) := \int_{t_0}^{t} a(s)x(s)ds.$$

Then $y(t_0) = 0$ $y'(t) = a(t)x(t)$ and the inequality (18.1.33) can be restated as $x(t) \leqslant b(t)+y(t)$. Since $a(t) \geqslant 0$, we have

$$a(t)x(t) \leqslant a(t)b(t) + a(t)y(t),$$

and we deduce that

$$y'(t) = a(t)x(t) \leqslant a(t)b(t) + a(t)y(t).$$

We multiply both sides of the above inequality with $e^{-A(t)}$ to obtain

$$\frac{d}{dt}\Big(y(t)e^{-A(t)}\Big) \leqslant a(t)b(t)e^{-A(t)}.$$

Integrating we obtain

$$y(t) \leqslant e^{A(t)} \int_{t_0}^{t} a(s)b(s)e^{-A(s)}ds = \int_{t_0}^{t} a(s)b(s)e^{(\,A(t)-A(s)\,)}ds. \tag{18.1.35}$$

We reach the desired conclusion by recalling that $x(t) \leqslant b(t) + y(t)$. □

Corollary 18.1.2. *Let $x : [t_0, t_1] \to [0, \infty)$ be a continuous nonnegative function satisfying the inequality*

$$x(t) \leqslant M + \int_{t_0}^{t} a(s)x(s)ds, \tag{18.1.36}$$

where M is a nonnegative constant and $a : [t_0, t_1] \to \mathbb{R}$ is a continuous nonnegative function. Then

$$x(t) \leqslant Me^{A(t)},\ A(t) = \int_{t_0}^{t} a(s)ds,\ \ \forall t \in [t_0, t_1]. \tag{18.1.37}$$

Remark 18.1.3. The above inequality is optimal in the following sense: if we have equality in (18.1.36), then we have equality in (18.1.37) as well. Note also that we can identify the right-hand-side of (18.1.37) as the unique solution of the linear Cauchy problem

$$u'(t) = a(t)u(t),\ \ u(t_0) = M.$$

Thus we have equality in(18.1.37) if we have equality in (18.1.36). □

Proof. Observe first that

$$a(s)e^{\left(A(t)-A(s)\right)} = -\frac{d}{ds}e^{\left(A(t)-A(s)\right)}.$$

When $b(t) = M$, $\forall t \in [a,b]$ the right-hand side of (18.1.34) becomes

$$M + M\int_{t_0}^{t} a(s)e^{\left(A(t)-A(s)\right)}ds = M - Me^{\left(A(t)-A(s)\right)}\Big|_{s=t_0}^{s=t} = Me^{A(t)}.$$

□

We will frequently use Gronwall's inequality to produce a priori estimates of solutions of o.d.e.-s and systems of o.d.e.-s. In the remainder of this section we will discuss two slight generalizations of this inequality.

Proposition 18.1.4. *Let $x : [a,b] \to \mathbb{R}$ be a continuous function that satisfies the inequality*

$$\frac{1}{2}x(t)^2 \leqslant \frac{1}{2}x_0^2 + \int_a^t \psi(s)|x(s)|ds, \quad \forall t \in [a,b], \tag{18.1.38}$$

where $\psi : [a,b] \to (0,\infty)$ is a continuous positive function. Then $x(t)$ satisfies the inequality

$$|x(t)| \leqslant |x_0| + \int_a^t \psi(s)ds, \quad \forall t \in [a,b]. \tag{18.1.39}$$

Proof. For $\varepsilon > 0$ we define

$$y_\varepsilon(t) := \frac{1}{2}(x_0^2 + \varepsilon^2) + \int_a^t \psi(s)|x(s)|ds, \quad \forall t \in [a,b].$$

Using (18.1.38) we deduce

$$x(t)^2 \leqslant 2y_\varepsilon(t), \quad \forall t \in [a,b]. \tag{18.1.40}$$

Combining this with the equality

$$y_\varepsilon'(t) = \psi(t)|x(t)|$$

and (18.1.38) we conclude that

$$y_\varepsilon'(t) \leqslant \sqrt{2y_\varepsilon(t)}\psi(t), \quad \forall t \in [a,b].$$

Integrating from a to t we deduce

$$\sqrt{2y_\varepsilon(t)} \leqslant \sqrt{2y_\varepsilon(a)} + \int_a^t \psi(s)ds, \quad \forall t \in [a,b].$$

Using (18.1.40) we deduce

$$|x(t)| \leqslant \sqrt{2y_\varepsilon(a)} + + \int_a^t \psi(s)ds \leqslant |x_0| + \varepsilon + \int_a^t \psi(s)ds, \quad \forall t \in [a,b].$$

Letting $\varepsilon \to 0$ in the above inequality we obtain (18.1.39). □

18.2. Existence and uniqueness for the Cauchy problem

In this section we will use the sup-norm on $\mathbb{R}^n$. Thus if $\boldsymbol{x} = (x_1, \dots, x_n)$ then

$$\|\boldsymbol{x}\| := \max\big(\,|x_k|,\ k = 1, \dots, n\,\big),$$

while $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm

$$\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}.$$

18.2.1. Picard's existence theorem. Let $\Omega \subset \mathbb{R} \times \mathbb{R}^n = \mathbb{R}^{n+1}$ be an open set, $I \subset \mathbb{R}$ and open interval. Consider a continuous map

$$\boldsymbol{F} : \Omega \to \mathbb{R}^n,\ \ \Omega \ni (t, \boldsymbol{x}) \mapsto \boldsymbol{F}(t, \boldsymbol{x}) \in \mathbb{R}^n.$$

Fix $(t_0, \boldsymbol{x}^0) \in \Omega$. We want to describe a simple and very useful condition that guarantees *local* existence and unicity of the Cauchy problem

$$\boldsymbol{x}'(t) = \boldsymbol{F}\big(\,t, \boldsymbol{x}(t)\,\big),\ \ \boldsymbol{x}(t_0) = \boldsymbol{x}^0. \tag{18.2.1}$$

The actual precise statement is a bit of a mouthful.

Theorem 18.2.1 (E. Picard)**.** *For $a, b > 0$ we denote by $\Delta_{a,b}$ the $(n+1)$-dimensional box*

$$\big\{\,t \in \mathbb{R};\ \ |t - t_0| \leqslant a\,\big\} \times \big\{\,\boldsymbol{x} \in \mathbb{R}^n;\ \ \|\boldsymbol{x} - \boldsymbol{x}^0\| \leqslant b\,\big\}.$$

Fix a, b such that $\Delta_{a,b} \subset \Omega$. Suppose that the following hold.

(i) *The map $\boldsymbol{F}$ is continuous.*

(ii) *The map $\boldsymbol{F}$ is* locally Lipschitz *in the $\boldsymbol{x}$-variable, i.e., for any compact subset $K \subset \Omega$ there exists $L = L_K > 0$ such that if $(t, \boldsymbol{x}), (t, \boldsymbol{y}) \in K$, then*

$$\|\boldsymbol{F}(t, \boldsymbol{x}) - \boldsymbol{F}(t, \boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{18.2.2}$$

Set

$$\begin{gathered} L_0 := L_{\Delta_{a,b}},\ \ M := \sup_{(t,\boldsymbol{x}) \in \Delta_{a,b}} \|\boldsymbol{F}(t, \boldsymbol{x})\|, \\ \delta := \min\left(a, \frac{b}{M}\right),\ \ J := [t_0 - \delta, t_0 + \delta]. \end{gathered} \tag{18.2.3}$$

Then there $\underline{\text{exists}}$ a $\underline{\text{unique}}$ solution of (18.2.1) defined on the interval J.

Proof. We present a modern proof based on Banach's fixed point theorem.

We consider the set

$$\mathcal{X} := \big\{\boldsymbol{x} \in C(J, \mathbb{R}^n);\ \ \|\boldsymbol{x}(t) - \boldsymbol{x}^0\| \leqslant b,\ \ \forall t \in J\,\big\}. \tag{18.2.4}$$

Note that for any $\boldsymbol{x} \in \mathcal{X}$ we have

$$\big(\,t, \boldsymbol{x}(t)\,\big) \in \Delta_{a,b} \subset \Omega,\ \ \forall t \in J.$$

We equip $\mathcal{X}$ with the metric

$$d_*(\boldsymbol{x},\boldsymbol{y}) = \sup_{t\in J} \|x(t)-y(t)\| e^{-2L_0|t-t_0|}, \tag{18.2.5}$$

where L_0 is the Lipschitz constant in (18.2.2). The set $\mathcal{X}$ equipped with this metric is *complete* as a closed subset of the Banach space $C(J,\mathbb{R}^n)$ equipped with the norm

$$\|\boldsymbol{x}\|_* := \sup_{t\in J} \|x(t)\| e^{-2L_0|t-t_0|}.$$

We define the *nonlinear* operator

$$\Gamma : \mathcal{X} \to C\big(J,\mathbb{R}^n\big),\ \ \Gamma[\boldsymbol{x}](t) = \boldsymbol{x}^0 + \int_{t_0}^t \boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)ds.$$

Step 1. The set $\mathcal{X}$ is Γ-invariant, i.e., $\Gamma\big[\mathcal{X}\big]\subset\mathcal{X}$. Indeed if $\boldsymbol{x}\in\mathcal{X}$, then for any $t\in J$ we have

$$\|\Gamma[\boldsymbol{x}](t)-\boldsymbol{x}^0\| = \left\|\int_{t_0}^t \boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)ds\right\| \leqslant \left|\int_{t_0}^t \|\boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)\|ds\right|$$
$$\overset{(18.2.3)}{\leqslant} \left|\int_{t_0}^t Mds\right| = M|t-t_0| \leqslant \delta M \overset{(18.2.3)}{\leqslant} b$$

so that $\Gamma[\boldsymbol{x}]\in\mathcal{X}$, $\forall \boldsymbol{x}\in\mathcal{X}$. Thus Γ is a map $\mathcal{X}\to\mathcal{X}$.

Step 2 The induced map $\Gamma:\mathcal{X}\to\mathcal{X}$ is a contraction. For $\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}$ we have

$$\|\,\Gamma[\boldsymbol{x}](t)-\Gamma[\boldsymbol{y}](t)\,\| = \left\|\int_{t_0}^t \big(\boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)-\boldsymbol{F}\big(s,\boldsymbol{y}(s)\big)ds\right\|$$
$$\leqslant \left|\int_{t_0}^t \|\,\boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)-\boldsymbol{F}\big(s,\boldsymbol{y}(s)\big)\,\|ds\right| \overset{(18.2.2)}{\leqslant} L_0\left|\int_{t_0}^t \|\boldsymbol{x}(s)-\boldsymbol{y}(s)\|ds\right|.$$

Hence

$$e^{-2L_0|t-t_0|}\|\,\Gamma[\boldsymbol{x}](t)-\Gamma[\boldsymbol{y}](t)\,\|$$
$$\leqslant L_0 e^{-2L_0|t-t_0|}\left|\int_{t_0}^t e^{2L_0|s-t_0|}\underbrace{\big(\,\|\,\boldsymbol{x}(s)-\boldsymbol{y}(s)\,\|e^{-2L_0|s-t_0|}\,\big)}_{\leqslant d_*(\boldsymbol{x},\boldsymbol{y})}ds\right|$$
$$\overset{(18.2.5)}{\leqslant} L_0e^{-2L_0|t-t_0|}\left|\int_{t_0}^t e^{2L_0|s-t_0|}ds\right| d_*(\boldsymbol{x},\boldsymbol{y}) = L_0e^{-2L_0|t-t_0|}d_*(\boldsymbol{x},\boldsymbol{y})\left|\int_0^{t-t_0} e^{2L_0|\tau|}d\tau\right|$$
$$=\frac{1}{2}e^{-2L_0|t-t_0|}\big(\,e^{2L_0|t-t_0|}-1\,\big)d_*(\boldsymbol{x},\boldsymbol{y}) \leqslant \frac{1}{2}d_*(\boldsymbol{x},\boldsymbol{y}).$$

Hence

$$d_*\big(\,\Gamma[\boldsymbol{x}],\Gamma[\boldsymbol{y}]\,\big) = \sup_{t\in J} e^{-2L_0|t-t_0|}\|\,\Gamma[\boldsymbol{x}](t)-\Gamma[\boldsymbol{y}](t)\,\| \leqslant \frac{1}{2}d_*(\boldsymbol{x},\boldsymbol{y}).$$

Hence $\Gamma:\mathcal{X}\to\mathcal{X}$ is a contraction.

Banach's fixed point theorem implies that there exists a function $\boldsymbol{x}*\in\mathfrak{X}$ such that $\boldsymbol{x}_*=\Gamma[\boldsymbol{x}_*]$, i.e.,

$$\boldsymbol{x}_*(t)=\boldsymbol{x}^0+\int_{t_0}^t\boldsymbol{F}\big(s,\boldsymbol{x}_*(s)\big)ds,\ \ \forall t\in J.$$

We deduce from the above equality that $\boldsymbol{x}_*(t_0)=\boldsymbol{x}^0$. Since $\boldsymbol{F}$ is continuous, we deduce from the Fundamental Theorem of Calculus that the right-hand-side of the above equality is C^1 and thus $\boldsymbol{x}_*$ is also C^1 and satisfies

$$\boldsymbol{x}_*'(t)=\boldsymbol{F}\big(t,\boldsymbol{x}_*(t)\big),\ \ \forall t\in J.$$

Thus $\boldsymbol{x}_*$ is a solution of (18.2.1). This establishes the existence part of Picard's theorem.

Step 3. Uniqueness. Suppose that $\boldsymbol{x}:J\to\Omega$ is another solution of (18.2.1). Let us observe that, a priori, the function $\boldsymbol{x}$ need not belong to $\mathfrak{X}$ so it need not be a fixed point of $\Gamma:\mathfrak{X}\to\mathfrak{X}$.

Fix a compact subset $K\subset\Omega$ that contains the graphs of both $\boldsymbol{x}_*$ and $\boldsymbol{x}$. For example K could be the union of these two graphs. Let $L=L_K$. Then

$$\underbrace{\|\boldsymbol{x}(t)-\boldsymbol{x}_*(t)\|}_{=:u(t)}\leqslant\left|\int_{t_0}^t\|\boldsymbol{F}\big(s,\boldsymbol{x}_*(s)\big)-\boldsymbol{F}\big(s,\boldsymbol{x}_*(s)\big)\|ds\right|\leqslant L\left|\int_{t_0}^t u(s)ds\right|.$$

Gronwall's inequality (18.1.37) implies that $u(t)=0$, $\forall t\in J$, i.e., $\boldsymbol{x}=\boldsymbol{x}_*$. This proves the uniqueness part of Picard's theorem. □

Remark 18.2.2. (a) Let us observe that the locally Lipschitz condition is automatically satisfied if Ω is convex and $\boldsymbol{F}$ is C^1.

(b) Picard's theorem establishes only *local* existence. This is not a limitation of the proof, it is a feature of the theory of o.d.e.-s. Consider the Cauchy problem

$$x'=x^2,\ \ x(0)=1.$$

Here $n=1$ and $\boldsymbol{F}(t,x)=x^2$ is locally Lipschitz.

The above equation is separable and we deduce

$$\frac{x'}{x^2}=1\Rightarrow\frac{1}{x(0)}-\frac{1}{x(t)}=\int_0^t\frac{x'}{x^2}dt=t$$

so that

$$x(t)=\frac{1}{1-t}$$

Note that the solution $x(t)$ explodes in finite time and exists only on the interval $(-\infty,1)$.

(c) The local Lipschitz condition guaranteed the uniqueness of the solution of the Cauchy problem via Gronwall's inequality. Without this condition the uniqueness in not guaranteed. Consider for example the Cauchy problem

$$x'=\frac{3}{2}x^{1/3},\ \ x(0)=0.$$

This problem has a trivial solution $x(t) = 0$, $\forall t$ and a nontrivial one $x(t) = \max(t,0)^{3/2}$. □

18.2.2. Peano's existence theorem. The local Lipschitz condition is not necessary for the existence of solutions of Cauchy problems. We will prove an existence result for the Cauchy problem due to G. Peano (1858-1932). Roughly speaking, it states that the continuity of $\boldsymbol{F}$ alone suffices to guarantee that the Cauchy problem (18.2.1) has a solution in a neighborhood of the initial point. Beyond its theoretical significance, this result will offer us the opportunity to discuss another important technique of investigating and approximating the solutions of an o.d.e.. We are talking about the polygonal method, due essentially to L. Euler.

We continue using the same notations as in Theorem 18.2.1.

Theorem 18.2.3. *Let $\boldsymbol{F} : \Delta_{a,b} \to \mathbb{R}^n$ be a continuous function defined on*

$$\Delta := \big\{ (t,\boldsymbol{x}) \in \mathbb{R}^{n+1};\ \ |t-t_0| \leqslant a,\ \ \|\boldsymbol{x}-\boldsymbol{x}_0\| \leqslant b \big\}.$$

Then the Cauchy problem (18.2.1) admits at least one solution on the interval

$$J := [t_0-\delta, t_0+\delta],\ \ \delta := \min\left(a, \frac{b}{M}\right),\ \ M := \sup_{(t,\boldsymbol{x})\in\Delta_{a,b}} \|\boldsymbol{f}(t,\boldsymbol{x})\|.$$

Proof. We will prove the existence on the interval $[t_0, t_0+\delta]$. The existence on $[t_0-\delta, t_0]$ follows from a similar argument. Set $\Delta = \Delta_{a,b}$.

Fix $\varepsilon > 0$. Since $\boldsymbol{F}$ is uniformly continuous on Δ, there exists $\eta(\varepsilon) > 0$ such that

$$\|\boldsymbol{F}(t,\boldsymbol{x}) - \boldsymbol{F}(s,\boldsymbol{y})\| \leqslant \varepsilon,$$

for any $(t,\boldsymbol{x}), (s,\boldsymbol{y}) \in \Delta$ such that

$$|t-s| \leqslant \eta(\varepsilon),\ \ \|\boldsymbol{x}-\boldsymbol{y}\| \leqslant \eta(\varepsilon).$$

Consider the uniform subdivision $t_0 < t_1 < \cdots < t_{N(\varepsilon)} = t_0+\delta$, where $t_j = t_0 + jh_\varepsilon$, for $j = 0,\dots,N(\varepsilon)$, and $N(\varepsilon)$ is chosen large enough so that

$$h = h_\varepsilon := \frac{\delta}{N(\varepsilon)} \leqslant \min\left(\eta(\varepsilon), \frac{\eta(\varepsilon)}{M}\right). \tag{18.2.6}$$

We consider the polygonal, i.e., the continuous piecewise linear function

$$\boldsymbol{u}_\varepsilon : [t_0, t_0+\delta] \to \mathbb{R}^n$$

defined by

$$\begin{gathered} \boldsymbol{u}_\varepsilon(t) = \boldsymbol{u}_\varepsilon(t_j) + (t-t_j)\boldsymbol{F}\big(t_j, \boldsymbol{u}_\varepsilon(t_j)\big),\ \ t_j < t \leqslant t_{j+1} \\ \boldsymbol{u}_\varepsilon(t_0) = \boldsymbol{x}_0. \end{gathered} \tag{18.2.7}$$

Note that the function $\boldsymbol{u}_\varepsilon$ is uniquely determined by its values at the nodes t_i. This values are determined using the recurrence relation

$$\boldsymbol{u}_\varepsilon\big(t_{j+1}\big) = \boldsymbol{u}_\varepsilon\big(t_j\big) + h\boldsymbol{F}\big(t_j, \boldsymbol{u}_\varepsilon(t_j)\big),\ \ \forall j = 0,1,\dots,N(\varepsilon)-1. \tag{18.2.8}$$

To understand the origin of the first equality in (18.2.7) note that it can be rewritten as

$$u'(t_j) \approx \frac{\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}_\varepsilon(t_j)}{t - t_j} = \boldsymbol{F}\big(\, t_j, \boldsymbol{u}_\varepsilon(t_j) \,\big).$$

This implies that

$$\|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}_\varepsilon(t_j)\| \leqslant M(t - t_j) \leqslant Mh(\varepsilon) \leqslant \eta(\varepsilon), \quad \forall t \in [t_j, t_{j+1}]. \tag{18.2.9}$$

We deduce that if $t \in [t_0, t_0 + \delta]$, then

$$\|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{x}_0\| \leqslant M\delta \leqslant b.$$

Indeed, if $t \in [t_j, t_{j+1}]$ then

$$\|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{x}_0\| = \|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}_\varepsilon(t_0\| \leqslant \|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}_\varepsilon(t_j)\| + \sum_{i=1}^{j} \|\boldsymbol{u}_\varepsilon(t_i) - \boldsymbol{u}_\varepsilon(t_{i-1})\|$$

$$\leqslant M(t - t_j) + M\sum_{i=1}^{j}(t_i - t_{i-1}) = M(t - t_0) \leqslant M\delta.$$

Thus $(t, \boldsymbol{u}_\varepsilon(t)\,) \in \Delta$, $\forall t \in [t_0, t_0 + \delta]$, so that the equalities (18.2.7) are consistent. The equalities (18.2.7) also imply the estimates

$$\|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}_\varepsilon(s)\| \leqslant M|t - s|, \quad \forall t, x \in [t_0, t_0 + \delta]. \tag{18.2.10}$$

In particular, the inequality (18.2.10) shows that the family of functions $(\boldsymbol{u}_\varepsilon)_{\varepsilon>0}$ is uniformly bounded and equicontinuous on the interval $[t_0, t_0 + \delta]$. Arzelà-Ascoli's Theorem 17.4.3 shows that there exist a continuous function $\boldsymbol{u} : [t_0, t_0 + \delta] \to \mathbb{R}^n$ and a subsequence $(\boldsymbol{u}_{\varepsilon_\nu})$, $\varepsilon_\nu \searrow 0$, such that

$$\lim_{\nu\to\infty} \boldsymbol{u}_{\varepsilon_\nu}(t) = \boldsymbol{u}(t) \;\text{ uniformly on } [t_0, t_0 + \delta]. \tag{18.2.11}$$

We will prove that $\boldsymbol{u}(t)$ is a solution of the Cauchy problem (18.2.1).

With this goal in mind, we consider the family of functions

$$\boldsymbol{v}_\varepsilon(t) := x_0 + \int_{t_0}^{t} \boldsymbol{F}(s, \boldsymbol{u}_\varepsilon(s))ds, \quad t \in [t_0, t_0 + \delta]$$

Let $t \in [t_j, t_{j+1}]$. We deduce from(18.2.9)

$$|t - t_j|, \quad \|\boldsymbol{u}_\varepsilon(t) - \boldsymbol{u}(t_j)\| \leqslant \eta(\varepsilon).$$

Hence,

$$\|\boldsymbol{F}(s, \boldsymbol{u}_\varepsilon(s)) - \boldsymbol{F}(t_j, \boldsymbol{u}_\varepsilon(t_j))\| \leqslant \varepsilon$$

so that

$$\left\|\int_{t_j}^{t} \big(\, \boldsymbol{F}(s, \boldsymbol{u}_\varepsilon(s)) - \boldsymbol{F}(t_j, \boldsymbol{u}_\varepsilon(t_j)) \,\big) ds\right\| \leqslant \int_{t_j}^{t} \|\, \boldsymbol{F}(s, \boldsymbol{u}_\varepsilon(s)) - \boldsymbol{F}(t_j, \boldsymbol{u}_\varepsilon(t_j)) \,\| ds \leqslant \varepsilon(t - t_j).$$

Now observe that

$$\int_{t_j}^{t} \boldsymbol{F}(s, \boldsymbol{u}_\varepsilon(s))ds = \boldsymbol{v}_\varepsilon(t) - \boldsymbol{v}_\varepsilon(t_j)$$

and

$$\int_{t_j}^{t} \boldsymbol{F}(t_j,\boldsymbol{u}_\varepsilon(t_j))ds = (t-t_j)\boldsymbol{F}(t_j,\boldsymbol{u}_\varepsilon(t_j)) = \boldsymbol{u}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t_j).$$

Hence $\forall j$, $\forall t\in[t_j,t_{j+1}]$ we have

$$\Big\|\big(\boldsymbol{v}_\varepsilon(t)-\boldsymbol{v}_\varepsilon(t_j)\big)-\big(\boldsymbol{u}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t_j)\big)\Big\| \leqslant \varepsilon(t-t_j). \tag{18.2.12}$$

On the other hand, $\forall t\in[t_j,t_{j+1}]$

$$\begin{aligned}\boldsymbol{v}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t) &= \big((\boldsymbol{v}_\varepsilon(t)-\boldsymbol{v}_\varepsilon(t_j)\big)-\big(\boldsymbol{u}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t_j)\big)\\ &\quad+\sum_{i=0}^{j-1}\Big(\big(\boldsymbol{v}_\varepsilon(t_{i+1})-\boldsymbol{v}_\varepsilon(t_i)\big)-\big(\boldsymbol{u}_\varepsilon(t_{i+1})-\boldsymbol{u}_\varepsilon(t_i)\big)\Big).\end{aligned}$$

We deduce

$$\begin{aligned}\|\boldsymbol{v}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t)\| &\leqslant \big\|\big(\boldsymbol{v}_\varepsilon(t)-\boldsymbol{v}_\varepsilon(t_j)\big)-\big(\boldsymbol{u}_\varepsilon(t)-\boldsymbol{u}_\varepsilon(t_j)\big)\big\|\\ &\quad+\sum_{i=0}^{j-1}\Big\|\big(\boldsymbol{v}_\varepsilon(t_{i+1})-\boldsymbol{v}_\varepsilon(t_i)\big)-\big(\boldsymbol{u}_\varepsilon(t_{i+1})-\boldsymbol{u}_\varepsilon(t_i)\big)\Big\|.\\ &\overset{(18.2.12)}{\leqslant} \varepsilon(t-t_j)+\varepsilon\sum_{i=0}^{j-1}\big(t_{i+1}-t_i\big)=\varepsilon(t-t_0)\leqslant\varepsilon\delta.\end{aligned}$$

Hence

$$\sup_{t_0\leqslant t\leqslant t_0+\delta}\|\boldsymbol{u}_\varepsilon(t)-\boldsymbol{v}_\varepsilon(t)\|\leqslant\varepsilon\delta,\quad\forall\varepsilon>0.$$

Thus $\boldsymbol{v}_{\varepsilon_\nu}$ converges uniformly to $\boldsymbol{u}$ as $\nu\to\infty$. Letting $\nu\to\infty$ in the equality

$$\boldsymbol{v}_{\varepsilon_\nu}(t)=x_0+\int_{t_0}^{t}\boldsymbol{F}\big(s,\boldsymbol{u}_{\varepsilon_\nu}(s)\big)ds,\quad t\in[t_0,t_0+\delta]$$

we deduce

$$\boldsymbol{u}(t)=x_0+\int_{t_0}^{t}\boldsymbol{F}(s,\boldsymbol{u}(s))ds,\quad t\in[t_0,t_0+\delta],$$

so that $\boldsymbol{u}$ is a solution of (18.2.1) on the interval $[t_0,t_0+\delta]$. This completes the proof of Theorem 18.2.3. □

Remark 18.2.4. If $\boldsymbol{F}$ is locally Lipschitz in the $\boldsymbol{x}$ variable then, using the result in Exercise 17.4, we deduce that $\boldsymbol{u}_\varepsilon$ converges uniformly to the *unique* solution of (18.2.1). The piecewise linear function $\boldsymbol{u}_\varepsilon$ is thus an approximation of the real solution. This approximation is uniquely determined by finitely many data $\boldsymbol{u}_\varepsilon(t_j)$ that can be determined explicitly using the recurrence (18.2.8).

This recurrence is easily implementable on a computer. This numerical scheme for approximating the solution of an initial value problem is commonly referred to as the *Euler method.* □

18.2.3. Global existence and uniqueness. We consider the system of differential equations described in vector notation by

$$\boldsymbol{x}' = \boldsymbol{F}(t, \boldsymbol{x}), \tag{18.2.13}$$

where the function $\boldsymbol{F} : \Omega \to \mathbb{R}^n$ is continuous on the open subset $\Omega \subset \mathbb{R}^{n+1}$. Additionally, we will assume that $\boldsymbol{F}$ is *locally Lipschitz* in $\boldsymbol{x}$ on Ω.

We recall that, if A, B are subsets of $\mathbb{R}^m$, then the distance between them is defined by

$$\operatorname{dist}(A, B) = \inf\big\{ \|\boldsymbol{a} - \boldsymbol{b}\|; \ \ \boldsymbol{a} \in A, \ \ \boldsymbol{b} \in B \big\}.$$

Above and in the sequel $\| - \|$ denotes the sup-norm on $\mathbb{R}^m$.

Remark 18.2.5. It is useful to observe that if K is a compact subset of Ω, then the distance $\operatorname{dist}(K, \partial\Omega)$ from K to the boundary $\partial\Omega$ of Ω is strictly positive. Indeed, suppose that $(\boldsymbol{x}_\nu)$ is a sequence in K and $(\boldsymbol{y}_\nu)$ is a sequence in $\partial\Omega$ such that

$$\lim_{\nu\to\infty} \|\boldsymbol{x}_\nu - \boldsymbol{y}_\nu\| = \operatorname{dist}(K, \partial\Omega). \tag{18.2.14}$$

Since K is compact, the sequence $(\boldsymbol{x}_\nu)$ is bounded. Using (18.2.14) we deduce that the sequence $(\boldsymbol{y}_\nu)$ is also bounded. The Bolzano-Weierstrass theorem now implies that there exist subsequences $(\boldsymbol{x}_{\nu_k})$ and $(\boldsymbol{y}_{\nu_k})$ converging to $\boldsymbol{x}_0$ and respectively $\boldsymbol{y}_0$. Since both K and $\partial\Omega$ are closed, we deduce that $\boldsymbol{x}_0 \in K$, $\boldsymbol{y}_0 \in \partial\Omega$, and

$$\|\boldsymbol{x}_0 - \boldsymbol{y}_0\| = \lim_{k\to\infty} \|\boldsymbol{x}_{\nu_k} - \boldsymbol{y}_{\nu_k}\| = \operatorname{dist}(K, \partial\Omega).$$

Since $K \cap \partial\Omega = \varnothing$ we conclude that $\operatorname{dist}(K, \partial\Omega) > 0$. □

Returning to the differential system (18.2.13), consider $(t_0, \boldsymbol{x}_0) \in \Omega$ and $\Delta \subset \Omega$ a parallelepiped of the form

$$\Delta = \Delta_{a,b} := \big\{ (t, \boldsymbol{x}) \in \mathbb{R}^{n+1}; \ \ |t - t_0| \leqslant a, \ \ \|\boldsymbol{x} - \boldsymbol{x}_0\| \leqslant b \big\}.$$

(Since Ω is open, $\Delta_{a,b} \subset \Omega$ if a and b are sufficiently small.)

Fix a positive number M such that

$$M \geqslant \sup_{(t,\boldsymbol{x})\in\Delta} \|\boldsymbol{F}(t, \boldsymbol{x})\|.$$

Applying Picard's Theorem 18.2.1 to the system (18.2.13) restricted to Δ we deduce that the existence and uniqueness of a solution $\boldsymbol{x} = \boldsymbol{u}(t)$ satisfying the initial condition $\boldsymbol{u}(t_0) = \boldsymbol{x}_0$ and defined on an interval $[t_0 - \delta, t_0 + \delta]$, where

$$\delta = \min\left(a, \frac{b}{M}\right).$$

In other words, we have the following local existence result.

Theorem 18.2.6. *Let $\Omega \subset \mathbb{R}^{n+1}$ be an open set and assume that the function*

$$\boldsymbol{F} = \boldsymbol{F}(t, \boldsymbol{x}) : \Omega \to \mathbb{R}^n$$

is continuous and locally Lipschitz as a function of $\boldsymbol{x}$. Then for any $(t_0, \boldsymbol{x}_0) \in \Omega)$ there exists a unique solution $\boldsymbol{x}(t) = \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$ of (18.2.13) defined on a neighborhood of t_0 and satisfying the initial condition

$$\boldsymbol{x}(t; t_0, \boldsymbol{x}_0)\big|_{t=t_0} = \boldsymbol{x}_0.$$

□

We must emphasize the local character of the above result. Both the existence and the uniqueness of the Cauchy problem take place in a neighborhood of the initial moment t_0. However, we expect uniqueness to have a global nature, that is, if two solutions $\boldsymbol{x} = \boldsymbol{x}(t)$ and $\boldsymbol{y} = \boldsymbol{y}(t)$ of (18.2.13) are equal at a point t_0, then they should coincide on the common interval of existence. (Their equality on a neighborhood of t_0 follows from the local uniqueness result.)

The next theorem, which is known in literature as the *global uniqueness theorem*, states that, the global uniqueness holds under the assumptions of Theorem 18.2.6.

Theorem 18.2.7. *Assume that $\boldsymbol{F} : \Omega \to \mathbb{R}^n$ satisfies the assumptions in Theorem 18.2.6. If $\boldsymbol{x}, \boldsymbol{y}$ are two solutions of (18.2.13) defined on the open intervals I and respectively J. If $\boldsymbol{x}(t_0) = \boldsymbol{y}(t_0)$ for some $t_0 \in I \cap J$, then $\boldsymbol{x}(t) = \boldsymbol{y}(t)$, $\forall t \in I \cap J$.*

Proof. Let $(t_1, t_2) = I \cap J$. We will prove that $\boldsymbol{x}(t) = \boldsymbol{y}(t)$, $\forall t \in [t_0, t_2)$. The equality to the left of t_0 is proved in a similar fashion. Let

$$\mathcal{T} := \big\{\, \tau \in [t_0, t_2);\ \ \boldsymbol{x}(t) = \boldsymbol{y}(t);\ \ \forall t \in [t_0, \tau]\,\big\}.$$

Then $\mathcal{T} \neq \varnothing$ and we set $T := \sup \mathcal{T}$. We claim that $T = t_2$.

To prove the claim we argue by contradiction. Assume that $T < t_2$. Then

$$x(T) = \lim_{t \nearrow T} x(t) = \lim_{t \nearrow T} y(t) = y(T),$$

so $\boldsymbol{x}(t) = \boldsymbol{y}(t)$, $\forall t \in [t_0, T]$. Since $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are both solutions of (18.2.13) we deduce from Theorem 18.2.6 that there exists $\varepsilon > 0$ such that $\boldsymbol{x}(t) = \boldsymbol{y}(t)$, $\forall t \in [T, T + \varepsilon]$. This contradicts the maximality of T and concludes the proof of the theorem. □

A solution $\boldsymbol{u} = \boldsymbol{u}(t)$ of (18.2.13) defined on the interval $I = (a, b)$ is called *right-extendible* if there exists $b' > b$ and a solution $\boldsymbol{\psi}$ of (18.2.13), defined on (a, b') such that $\boldsymbol{\psi} = \boldsymbol{u}$ on (a, b). The notion of *left-extendible* solutions is defined analogously. A solution is called *extendible* if it is right-extendible or left-extendible or both. A solution that is not extendible is called *saturated*. In other words, a solution $\boldsymbol{u}$ defined on an interval I is saturated if I is the maximal domain of existence. Similarly, a solution that is not right-extendible (respectively left-extendible) is called *right-saturated* (respectively *left-saturated*).

Theorem 18.2.6 implies that maximal interval on which a saturated solution is defined must be an open interval. If a solution $\boldsymbol{u}$ is right-saturated, then the interval on which it is defined is open on the right. Similarly, if a solution $\boldsymbol{u}$ is left-saturated, then the interval on which it is defined is open on the left.

Indeed, if $\boldsymbol{u} : [a, b) \to \mathbb{R}^n$ is solution of (18.2.13) defined on an interval that is not open on the left, then Theorem 18.2.6 implies that there exists a solution $\widetilde{\boldsymbol{u}}(t)$ defined on an interval $[a - \delta, a + \delta]$ as satisfying the initial condition $\widetilde{\boldsymbol{u}}(a) = \boldsymbol{u}(a)$. The local uniqueness theorem implies that $\widetilde{\boldsymbol{u}} = \boldsymbol{u}$ on $[a, a + \delta]$ and thus the function

$$\widehat{\boldsymbol{u}}_0(t) = \begin{cases} \boldsymbol{u}(t), & t \in [a, b), \\ \widetilde{\boldsymbol{u}}(t), & t \in [a - \delta, a], \end{cases}$$

is a solution of (18.2.13) on $[a - \delta, b)$ that extends $\boldsymbol{u}$, showing that $\boldsymbol{u}$ is not left-saturated.

As an illustration, consider the o.d.e.

$$x' = x^2 + 1,$$

with initial condition $x(t_0) = x_0$. This is a separable o.d.e., and we find that

$$x(t) = \tan\big(\, t - t_0 + \arctan x_0 \,\big).$$

It follows that, on the right, the maximal existence interval is $[t_0,\ t_0 + \frac{\pi}{2} - \arctan x_0)$, while on the left, the maximal existence interval is $(t_0 - \frac{\pi}{2} - \arctan x_0, t_0]$. Thus, the saturated solution is defined on the interval $(t_0 - \frac{\pi}{2} - \arctan x_0, t_0 + \frac{\pi}{2} - \arctan x_0)$.

Our next result characterizes the right-saturated solutions. In the remainder of this section we will assume that $\Omega \subset \mathbb{R}^{n+1}$ is an open subset and $\boldsymbol{F} : \Omega \to \mathbb{R}^n$ is a continuous map that is also locally Lipschitz in the variable $\boldsymbol{x} \in \mathbb{R}^n$.

Theorem 18.2.8. *Let $\boldsymbol{u} : [t_0, t_1) \to \mathbb{R}^n$ be a solution of (18.2.13). Then the following are equivalent.*

(i) *The solution $\boldsymbol{u}$ is right-extendible.*

(ii) *The graph of $\boldsymbol{u}$,*

$$\Gamma := \big\{\, (t, \boldsymbol{u}(t)\,);\ \ t \in [t_0, t_1)\,\big\},$$

is contained in a compact subset of Ω.

Proof. (i) $\Rightarrow$ (ii). Assume that $\boldsymbol{u}$ is right-extendible. Thus, there exists a solution $\boldsymbol{\psi}(t)$ of (18.2.13) defined on an interval $[t_0, t_1 + \delta)$, $\delta > 0$, and such that

$$\boldsymbol{\psi}(t) = \boldsymbol{u}(t),\ \ \forall t \in [t_0, t_1).$$

In particular, it follows that Γ is contained in $\widehat{\Gamma}$, the graph of the restriction of $\boldsymbol{\psi}$ to $[t_0, t_1]$. Now observe that $\widehat{\Gamma}$ is a compact subset of Ω because it is image of the compact interval $[t_0, t_1]$ via the continuous map $t \mapsto (t, \boldsymbol{\psi}(t))$.

(ii) $\Rightarrow$ (i) Assume that $\Gamma \subset K$, where K is a compact subset of Ω. We will prove that $\boldsymbol{u}(t)$ can be extended to a solution of (18.2.13) on an interval of the form $[t_0, t_1+\delta]$, for some $\delta > 0$.

Since $\boldsymbol{u}(t)$ is a solution, we have

$$\boldsymbol{u}(t) = \boldsymbol{u}(t_0) + \int_{t_0}^{t} \boldsymbol{F}\big(\, s, \boldsymbol{u}(s)\,\big) ds, \;\; \forall t \in [t_0, t_1).$$

We deduce

$$\|\boldsymbol{u}(t) - \boldsymbol{u}(t')\| \leqslant \left| \int_{t'}^{t} \big\| \boldsymbol{F}\big(\, s, \boldsymbol{u}(s)\,\big) \big\| ds \right| \leqslant M_K |t-t'|, \;\; \forall t, t' \in [t_0, t_1),$$

where

$$M_K := \sup_{(s,\boldsymbol{x}) \in K} \|\boldsymbol{F}(s, \boldsymbol{x})\|.$$

Cauchy's characterization of convergence now shows that $\boldsymbol{u}(t)$ has a (finite) limit as $t \nearrow t_1$ and we set

$$\boldsymbol{u}(t_1) := \lim_{t \nearrow t_1} \boldsymbol{u}(t).$$

We have thus extended $\boldsymbol{u}$ to a continuous function on $[t_0, t_1]$ that we continue to denote by $\boldsymbol{u}$. The continuity of $\boldsymbol{F}$ implies that

$$\boldsymbol{u}'(t_1 - 0) = \lim_{t \nearrow t_1} \boldsymbol{u}'(t) = \lim_{t \nearrow t_1} \boldsymbol{F}(t, \boldsymbol{u}(t)\,) = \boldsymbol{F}(t_1, \boldsymbol{u}(t_1)\,). \tag{18.2.15}$$

On the other hand, according to Theorem 18.2.6, there exists a solution $\boldsymbol{\psi}(t)$ of (18.2.13) defined on an interval $[t_1 - \delta, t_1 + \delta]$ and satisfying the initial condition $\boldsymbol{\psi}(t_1) = \boldsymbol{u}(t_1)$. Consider the function

$$\widetilde{\boldsymbol{u}}(t) = \begin{cases} \boldsymbol{u}(t), & t \in [t_0, t_1], \\ \boldsymbol{\psi}(t), & t \in (t_1, t_1 + \delta]. \end{cases}$$

Obviously

$$\widetilde{\boldsymbol{u}}'(t_1 + 0) = \boldsymbol{\psi}'(t_1) = \boldsymbol{F}(t_1, \boldsymbol{\psi}(t_1)) = \boldsymbol{F}(t_1, \boldsymbol{u}(t_1)) \overset{(18.2.15)}{=} \boldsymbol{u}'(t_1 - 0).$$

This proves that $\widetilde{\boldsymbol{u}}$ is C^1, and satisfies the differential equation (18.2.13). Clearly $\widetilde{\boldsymbol{u}}$ extends $\boldsymbol{u}$ to the right. □

The next result shows that any solution can be extended to a saturated solution.

Theorem 18.2.9. *Any solution $\boldsymbol{u}$ of (18.2.13) admits a unique extension to a saturated solution.*

Proof. The uniqueness is a consequence of Theorem 18.2.7 on global uniqueness. To prove the extendibility to a saturated solution we will limit ourself to proving the extendibility to a right-saturated solution.

We denote by $\mathscr{A}$ the set of all solutions $\boldsymbol{\psi}$ of (18.2.13) that extend $\boldsymbol{u}$ to the right. The set $\mathscr{A}$ is totally ordered by the inclusion of the domains of definition of the solutions

$\boldsymbol{\psi}$ and, as such, the set $\mathscr{A}$ has an upper bound, $\tilde{\boldsymbol{u}}$. This is a right-saturated solution of (18.2.13). □

We will next investigate the behavior of the saturated solutions of (18.2.13) in a neighborhood of the boundary $\partial\Omega$ of the domain Ω where (18.2.13) is defined. For simplicity we only discuss the case of right-saturated solutions. The case of left-saturated solutions is identical.

Theorem 18.2.10. *Let $\boldsymbol{\varphi}(t)$ be a right-saturated solution of (18.2.13) defined on the interval $[t_0, T)$. Then any limit point as $t \nearrow T$ of the graph*

$$\Gamma := \big\{\, (t, \boldsymbol{\varphi}(t));\ \ t_0 \leqslant t < T \,\big\}$$

is either the point at infinity of $\mathbb{R}^{n+1}$, or a point on $\partial\Omega$.

Proof. The theorem states that, if (τ_ν) is a sequence in $[t_0, T)$ such that the limit

$$\lim_{\nu\to\infty} (\tau_\nu, \boldsymbol{\varphi}(\tau_\nu))$$

exists, then

(i) either $T = \infty$,

(ii) or $\lim_{\nu\to\infty} \|\boldsymbol{\varphi}(\tau_\nu)\| = \infty$,

(iii) or $T < \infty$, $\boldsymbol{x}^* = \lim_{\nu\to\infty} \boldsymbol{\varphi}(\tau_\nu) \in \mathbb{R}^n$ and $(T, \boldsymbol{x}^*) \in \partial\Omega$.

We argue by contradiction. Assume that all the three options are violated. Since (i), (ii) do not hold we deduce that $T < \infty$ and that the limit $\lim_{\nu\to\infty} \boldsymbol{\varphi}(\tau_\nu)$ exists and it is a point $\boldsymbol{x}^* \in \mathbb{R}^n$. The point $(T, \boldsymbol{x}^*)$ is in the closure of Ω and, since (iii) is also violated, we deduce that $(T, \boldsymbol{x}^*) \notin \partial\Omega$, i.e., $(T, \boldsymbol{x}^*) \in \Omega$. Let

$$\eta := \operatorname{dist}_\infty\big(\, (T, \boldsymbol{x}^*), \partial\Omega \,\big) := \inf_{(t,\boldsymbol{y})\in\partial\Omega} \max\big(\, |T - t|, \|\boldsymbol{x}^*, \boldsymbol{y}\| \,\big),$$

where $\|-\|$ denotes the sup-norm in $\mathbb{R}^n$, and the sup-norm of $(s, \boldsymbol{y}) \in \mathbb{R}^{n+1}$ is $\max(|s|, \|\boldsymbol{x}\|)$. Note that $\eta > 0$ since $(T, \boldsymbol{x}^*) \notin \partial\Omega$ so the closed box

$$S := \big\{\, (t, \boldsymbol{x}) \in \mathbb{R}^{n+1};\ \ |t - T| \leqslant \eta/2,\ \ \|\boldsymbol{x} - \boldsymbol{x}^*\| \leqslant \eta/2 \,\big\}$$

is contained in Ω; see Figure 18.1.

We deduce that, for any $(s_0, \boldsymbol{y}_0) \in S$, the parallelepiped

$$\Delta_{s_0,\boldsymbol{y}_0} := \left\{\, (t, \boldsymbol{x}) \in \mathbb{R}^{n+1};\ \ |t - s_0| \leqslant \frac{\eta}{4},\ \ \|\boldsymbol{x} - \boldsymbol{y}_0\| \leqslant \frac{\eta}{4} \,\right\} \tag{18.2.16}$$

is contained in the compact subset of Ω (see Figure 18.1),

$$K = \left\{\, (t, \boldsymbol{x}) \in \mathbb{R}^{n+1};\ \ |t - T| \leqslant \frac{3\eta}{4},\ \ \|\boldsymbol{x} - \boldsymbol{x}^*\| \leqslant \frac{3\eta}{4} \,\right\}.$$

We set

$$\delta := \min\left\{ \frac{\eta}{4}, \frac{\eta}{4M} \right\},\ \ M := \sup_{(t,\boldsymbol{x})\in K} \|\boldsymbol{F}(t, \boldsymbol{x})\|.$$

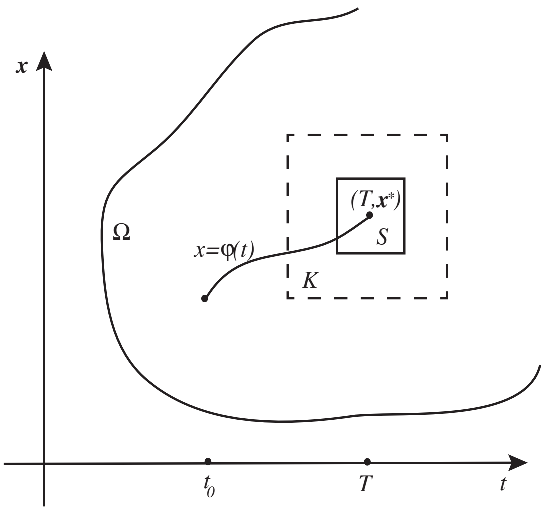

Figure 18.1. The behavior of a right-saturated solution.

Appealing to Picard's existence and uniqueness theorem (Theorem 18.2.1), where Δ is defined in (18.2.16), it follows that <u>*for any*</u> $(s_0, \boldsymbol{y}_0) \in S$, there exists a unique solution $\boldsymbol{\psi}_{s_0,\boldsymbol{y}_0}(t)$ of (18.2.13) defined on the interval $[s_0 - \delta, s_0 + \delta]$ and satisfying the initial condition $\boldsymbol{\psi}(s_0) = \boldsymbol{y}_0$. Moreover, its graph is contained in $\Delta_{s_0,\boldsymbol{y}_0}$.

Fix ν is sufficiently large so that

$$(\tau_\nu, \boldsymbol{\varphi}(\tau_\nu)) \in S \ \text{ and } \ |\tau_\nu - T| \leqslant \frac{\delta}{2},$$

and define $\boldsymbol{y}_\nu := \boldsymbol{\varphi}(\tau_\nu)$,

$$\widetilde{\boldsymbol{\varphi}}(t) := \begin{cases} \boldsymbol{\varphi}(t), & t_0 \leqslant t \leqslant \tau_\nu, \\ \\ \boldsymbol{\psi}_{\tau_\nu,\boldsymbol{y}_\nu}(t), & \tau_\nu < t \leqslant \tau_\nu + \delta. \end{cases}$$

Then $\widetilde{\boldsymbol{u}}(t)$ is a solution of (18.2.13) defined on the interval $[t_0, \tau_\nu + \delta]$. This interval strictly contains the interval $[t_0, T]$ and $\widetilde{\boldsymbol{\varphi}} = \boldsymbol{\varphi}$ on $[t_0, T)$. This contradicts our assumption that $\boldsymbol{\varphi}$ is a right-saturated solution, and completes the proof of Theorem 18.2.10. □

Example 18.2.11. Suppose $n = 1$, $\Omega = (-\infty, 0) \times \mathbb{R}$ and

$$F : \Omega \to \mathbb{R}, \ \ F(t,x) = -\frac{1}{t}x - \frac{1}{t^2}\cos(1/t).$$

The function $x(t) = \frac{1}{t}\sin(\frac{1}{t})$, $t < 0$, satisfies the differential equation $x'(t) = F(t,x)$. Indeed,

$$x'(t) = -\frac{1}{t^2}\sin(1/t) - \frac{1}{t^2}\cos(1/t) = -\frac{1}{t}x(t) - \frac{1}{t^2}\cos 1/t).$$

This solution is right-saturated, and any point on $\{0\}\times[-\infty,\infty]$ is a limit point of $(t, x(t))$ as $t\nearrow 0$. □

Theorem 18.2.12. *Let $\Omega=\mathbb{R}^{n+1}$ and $\boldsymbol{u}(t)$ a right-saturated solution of (18.2.13) defined on $[t_0,T)$. Then*

(i) *either $T<\infty$ and*

$$\lim_{t\nearrow T}\|\boldsymbol{u}(t)\|=\infty,$$

(ii) *or $T=\infty$.*

Proof. From Theorem 18.2.10 it follows that any limit point as $t\nearrow T$ on the graph Γ of $\boldsymbol{u}$ is the point at infinity since $\partial\Omega=\varnothing$ in this case. If $T<\infty$, then necessarily

$$\lim_{t\nearrow T}\|\boldsymbol{u}(t)\|=\infty.$$

□

Theorem 18.2.10 takes a simpler form in the case of autonomous systems.

Corollary 18.2.13. *Supppose that $U\subset\mathbb{R}^n$ is an open set, $\boldsymbol{F}:U\to\mathbb{R}^n$ is a locally Lipschitz map and $\boldsymbol{x}:[t_0,T)\to U$ a right saturated solution of the o.d.e.*

$$\boldsymbol{x}'(t)=\boldsymbol{F}\big(\boldsymbol{x}(t)\big).$$

Then

(i) *either $T=\infty$,*

(ii) *or $T<\infty$ and any limit point of $\boldsymbol{x}(t)$ as $t\nearrow T$ is either a point of ∂U, or the point at ∞ of $\mathbb{R}^n$, i.e.,*

$$\lim_{t\nearrow T}\|\boldsymbol{x}(t)\|=\infty.$$

□

Theorems 18.2.10 and 18.2.12 are useful in determining the maximal existence interval of a solution. Loosely speaking, Theorem 18.2.12 states that a solution $\boldsymbol{u}$ is either defined on the whole positive semiaxis, or it "blows up" in finite time. This phenomenon is commonly referred to as the finite-time *blowup* phenomenon.

Example 18.2.14. To illustrate Theorem 18.2.12 consider the Cauchy problem

$$x'=x^2,\;\; x(0)=2.$$

Using separation of variables we deduce

$$\frac{dx}{x^2}=dt\Rightarrow\frac{1}{x}=-t+C\Rightarrow x=\frac{1}{C-t}.$$

Since $x(0)=2$ we deduce $C=1/2$ so $x=\frac{2}{1-2t}$. Its graph is depicted in Figure 18.2.

Its maximal existence interval on the right is $[0,1/2)$. □

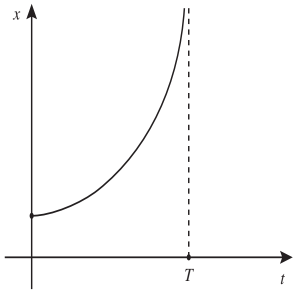

Figure 18.2. A finite-time blowup.

Corollary 18.2.15. *Suppose* $\boldsymbol{F}:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n$, $\boldsymbol{F}=\boldsymbol{F}(t,\boldsymbol{x})$, *is a continuous map, locally Lipschitz in* $\boldsymbol{x}$. *If* $\boldsymbol{x}:(T_-,T_+)\to\mathbb{R}^n$ *a saturated solution of the o.d.e.* $\boldsymbol{x}'(t)=\boldsymbol{F}\big(t,\boldsymbol{x}(t)\big)$ *such that there exists a continuous function* $M:\mathbb{R}\to[0,\infty)$ *so that*

$$\|u(t)\|\leqslant M(t),\ \ \forall t\in(T_-,T_+),$$

then $T_\pm=\mp\infty$ □

Corollary 18.2.16. *Suppose* $\boldsymbol{F}:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n$, $F=F(t,\boldsymbol{x})$*is a continuous map, locally Lipschitz in* $\boldsymbol{x}$, *such that there exists a continuous function*

$$a:\mathbb{R}\to(0,\infty),\ \ t\mapsto a(t)$$

such that

$$\forall(t,\boldsymbol{x})\in\mathbb{R}\times\mathbb{R}^n:\ \ \|\boldsymbol{F}(t,\boldsymbol{x})\|\leqslant a(t)\big(1+\|\boldsymbol{x}\|\big).$$

If $\boldsymbol{x}:(T_-,T_+)\to\mathbb{R}^n$ *is a saturated solution of the o.d.e.* $\boldsymbol{x}'(t)=\boldsymbol{F}\big(\boldsymbol{x}(t)\big)$, *then* $T_\pm=\mp\infty$.

Proof. We prove only that $T_+=\infty$. The case $T_-=-\infty$ can be dealt with in a similar fashion. Pick $t_0\in(T_-,T_+)$. We have

$$\boldsymbol{x}(t)=\boldsymbol{x}(t_0)+\int_{t_0}^t\boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)ds$$

so that, $\forall t\in[t_0,T_+)$

$$\|\boldsymbol{x}(t)\|\leqslant\|\boldsymbol{x}(t_0)\|+\int_{t_0}^t\big\|\boldsymbol{F}\big(s,\boldsymbol{x}(s)\big)\big\|ds\leqslant\|\boldsymbol{x}(t_0)\|+\int_{t_0}^t a(s)\big(1+\|\boldsymbol{x}(s)\|\big)ds$$

$$=\underbrace{\|\boldsymbol{x}(t_0)\|+\int_{t_0}^t a(s)ds}_{=:b(t)}+\int_{t_0}^t a(s)\|\boldsymbol{x}(s)\|ds.$$

Gronwall's inequality (18.1.34) implies

$$\|\boldsymbol{x}(t)\|\leqslant\underbrace{b(t)+\int_{t_0}^t b(s)e^{A(t)-A(s)}ds}_{=:M(t)},\ \ A(t)=\int_{t_0}^t a(s)ds.$$

The conclusion now follows from Corollary 18.2.15. □

Example 18.2.17. Consider the system of o.d.e.-s

$$\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x}), \tag{18.2.17}$$

where $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz and satisfies

$$\big(\,\boldsymbol{x}, \boldsymbol{F}(\boldsymbol{x})\,\big) \leqslant 0, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{18.2.18}$$

Above, $(-,-)$ is the natural inner product on $\mathbb{R}^n$. Recall that $\|-\|_2$ denotes the canonical Euclidean norm on $\mathbb{R}^n$, $\|\boldsymbol{x}\|_2 = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$.

According to the existence and uniqueness theorem, for any $(t_0, \boldsymbol{x}_0) \in \mathbb{R}^{n+1}$ there exists a unique solution $\varphi(t) = \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$ of (18.2.17) satisfying $\varphi(t_0) = \boldsymbol{x}_0$ and defined on a maximal interval $[t_0, T)$. We want to prove that, under the above assumptions, we have $T = \infty$.

To show this, take the inner product of both sides of (18.2.17) with $\varphi(t)$. Using (18.2.18) we deduce

$$\frac{1}{2}\frac{d}{dt}\big(\,\varphi(t), \varphi(t)\,\big) = \big(\,\varphi(t), \varphi'(t)\,\big) = \big(\,\boldsymbol{F}\varphi(t)\,,\,\varphi(t)\,\big) \leqslant 0, \;\; \forall t \in [t_0, T),$$

and therefore

$$\|\varphi(t)\|_2^2 \leqslant \|\varphi(t_0)\|_2^2, \;\; \forall t \in [t_0, T).$$

Thus, the solution $\varphi(t)$ is bounded, there is no blowup, so $T = \infty$.

Suppose now that $\boldsymbol{F}$ satisfies only the constraint

$$\big(\,\boldsymbol{x}, \boldsymbol{F}(\boldsymbol{x})\,\big) < 0, \;\; \forall \|\boldsymbol{x}\|_2 = 1. \tag{18.2.19}$$

The initial value problem

$$\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x}), \;\; \boldsymbol{x}(0) = \boldsymbol{x}_0$$

admits a unique right saturated solution defined on a maximal interval $[0, T)$. We want to prove that $T = \infty$ if the initial condition $\boldsymbol{x}_0$ is sufficiently small, $\|\boldsymbol{x}_0\|_2 < 1$.

From (18.2.19) we deduce that there exists a small $\eta > 0$ such that

$$\big(\,\boldsymbol{x}, \boldsymbol{F}(\boldsymbol{x})\,\big) < 0, \;\; \forall 1 - \eta \leqslant \|\boldsymbol{x}\|_2 \leqslant 1 + \eta.$$

Set $\rho(t) := \|\boldsymbol{x}(t)\|_2^2$.

We argue by contradiction. Suppose that $T < \infty$. We deduce from Corollary 18.2.13 that

$$\lim_{t \nearrow T} \rho(t) \in \{\, 1, \infty \,\}.$$

In either case, since $\rho(0) < 1$, there exists $T_1 \in (0, T]$ such that

$$\lim_{t \nearrow T_1} \rho(t) = 1, \;\; \rho(t) < 1, \;\; \forall t \in [0, T_1)$$

We deduce that there exists $T_2 \in [0, T_1)$ such that $\rho(t) \in [1 - \eta, 1]$, $\forall t \in (T_2, T_1)$.

From the Lagrange Mean Value Theorem we deduce that for any $t < T_1$ there exists $\xi_t \in (t, T_1)$ such that

$$\rho'(\xi_t) = \frac{1 - \rho(t)}{T_1 - t} > 0.$$

Note that $\rho'(t) = 2\big(\,\boldsymbol{x}(t), \boldsymbol{F}(\boldsymbol{x}(t))\,\big) < 0$ if $\boldsymbol{x}(t) \in [1 - \eta, 1]$. Hence $\rho'(\xi_t) < 0$, $\forall t \in (T_2, T_1)$. This contradiction shows that $T = \infty$. □

Definition 18.2.18. Let $U \subset \mathbb{R}^n$ be an open set, $\boldsymbol{F} : U \to \mathbb{R}^n$ a continuous map (vector field) and $\alpha : U \to \mathbb{R}$ a continuous function.

(i) The function α is said to be a *Lyapunov function* of the differential system $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ if for any solution $\boldsymbol{x}(t)$ of this system, the function $t \mapsto \alpha\big(\boldsymbol{x}(t)\big)$ is nonincreasing.

(ii) The function α is said to be a *prime integral* of the differential system $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ if for any solution $\boldsymbol{x}(t)$ of this system, the function $t \mapsto \alpha\big(\boldsymbol{x}(t)\big)$ is constant.

□

Proposition 18.2.19. *Let $U \subset \mathbb{R}^n$ be an open set, $\boldsymbol{F} : U \to \mathbb{R}^n$ a locally Lipschitz map (vector field) and $\alpha : U \to \mathbb{R}$ a C^1-function. Then the following are equivalent.*

(i) *The function α is a Lyapunov function of $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$.*

(ii) $\big(\boldsymbol{F}(\boldsymbol{x}), \nabla\alpha(\boldsymbol{x})\big) \leqslant 0$, $\forall \boldsymbol{x} \in U$, *where $(-,-)$ denotes the canonical inner product in $\mathbb{R}^n$.*

Proof. (i) ⇒ (ii) Suppose that α is a Lyapunov function. Fix an arbitrary $\boldsymbol{x}_0 \in U$ and denote by $\boldsymbol{x}(t)$ the saturated solution of the initial value problem

$$\boldsymbol{x}'(t) = \boldsymbol{f}\big(\,(\boldsymbol{x}(t)\,\big),\;\; \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

The function $u(t) = \alpha\big(\boldsymbol{x}(t)\big)$ is nonincreasing and thus $u'(0) \leqslant 0$. The chain rule implies

$$0 \geqslant u'(0) = \big(\,\nabla\alpha(\boldsymbol{x}(0)), \boldsymbol{x}'(0)\,\big) = \big(\,\nabla\alpha(\boldsymbol{x}_0), \boldsymbol{F}(\boldsymbol{x}_0)\,\big).$$

Conversely, if (ii) holds, then for any solution $\boldsymbol{x}(t)$ of $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ we have

$$\frac{d}{dt}\alpha\big(\,\boldsymbol{x}(t)\,\big) = \big(\,\nabla\alpha(\boldsymbol{x}(t)), \boldsymbol{x}'(t)\,\big) = \big(\,\nabla\alpha(\boldsymbol{x}(t)), \boldsymbol{F}(\boldsymbol{x}(t))\,\big) \leqslant 0$$

so α is a Lyapunov function. □

Observe that a function α is a prime integral of $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ if and only if both α and $-\alpha$ are Lyapunov functions of this differential system. If α and $\boldsymbol{F}$ are as in Proposition 18.2.19, then α *is a prime integral of this system if and only if*

$$\big(\,\boldsymbol{F}(\boldsymbol{x}), \nabla\alpha(\boldsymbol{x})\,\big) = 0,\;\; \forall \boldsymbol{x} \in U. \tag{18.2.20}$$

Proposition 18.2.20. *Let $U \subset \mathbb{R}^n$ be an open set, $\boldsymbol{F} : U \to \mathbb{R}^n$ locally Lipschitz and $\alpha : U \to \mathbb{R}$ a Lyapunov function of the differential system $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$. Suppose that α is* coercive, *i.e., for any $c \in \mathbb{R}$ the sublevel set*

$$\{\alpha \leqslant c\} = \big\{\boldsymbol{x} \in U;\;\; \alpha(\boldsymbol{x}) \leqslant c\,\big\}$$

is compact. Then any right saturated solution $\boldsymbol{x} : [0, T) \to U$ of $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ is well defined for any $t \geqslant 0$, i.e., $T = \infty$.

Proof. Set $c_0 = \alpha\big(\boldsymbol{x}(0)\big)$. Since α is a Lyapunov functions we deduce

$$\alpha\big(\boldsymbol{x}(t)\big) \leqslant \alpha\big(\boldsymbol{x}(0)\big) = c_0, \;\; \forall t \in [0,T)$$

so that $\alpha(t) \in K_0 := \{\alpha \leqslant 0\}$.

We argue by contradiction. Suppose that $T < \infty$. Then the graph

$$\big\{(t, \boldsymbol{x}(t)); \;\; t \in [0,T)\big\} \subset \mathbb{R} \times U$$

is contained in the compact subset $[0,T] \times K_0$ proving that the solution is right-extendible contradicting the fact that $\boldsymbol{x}$ is right-saturated. □

Let us observe that condition (18.2.18) in Example 18.2.17 shows that the function $\alpha(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ is a coercive Lyapunov function of the differential system $\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x})$ thus proving that the right saturated solutions exist up to $T = \infty$.

18.2.4. Dissipative systems of differential equations. A continuous map

$$\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$$

is called *dissipative* if it satisfies the *dissipativity* or *monotonicity* condition

$$\big(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\big) \leqslant 0, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \tag{18.2.21}$$

where we denoted by $(-,-)$ the canonical Euclidean inner product in $\mathbb{R}^n$.

Example 18.2.21. (a) A continuous map $f : \mathbb{R} \to \mathbb{R}$ is dissipative if and only if f is nion-increasing.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Exercise 14.7 shows that the map $G : \mathbb{R}^n \to \mathbb{R}^n$, $G(\boldsymbol{x}) = -\nabla f(\boldsymbol{x})$ is dissipative iff f is convex. □

In this subsection we consider *dissipative, autonomous differential systems*, i.e., systems of ordinary differential equations of the form

$$\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x}), \tag{18.2.22}$$

where $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous dissipative map. We associate to (18.2.22) the initial condition

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0, \tag{18.2.23}$$

where $(t_0, \boldsymbol{x}_0)$ is a given point in $\mathbb{R}^{n+1}$.

The mathematical models of a large class of physical phenomena, such as diffusion, leads to dissipative differential systems. For dissipative systems we have the following interesting existence and uniqueness result.

Theorem 18.2.22. *If the continuous map $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is dissipative, then for any $(t_0, \boldsymbol{x}_0) \in \mathbb{R}^{n+1}$ the Cauchy problem (18.2.22)+(18.2.23) admits a unique solution defined on $[t_0, \infty)$. WE denote it by $\boldsymbol{x} = \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$. Moreover, the map*

$$S : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^n, \;\; (t, \boldsymbol{x}_0) \mapsto S(t)\boldsymbol{x}_0 := \boldsymbol{x}(t; 0, \boldsymbol{x}_0),$$

satisfies the following properties.

$$S(0)\boldsymbol{x}_0 = \boldsymbol{x}_0, \ \ \forall \boldsymbol{x}_0 \in \mathbb{R}^n, \tag{18.2.24a}$$

$$S(t+s)\boldsymbol{x}_0 = S(t)S(s)\boldsymbol{x}_0, \ \ \forall \boldsymbol{x}_0 \in \mathbb{R}^n, \ \ t,s \geqslant 0, \tag{18.2.24b}$$

$$\|S(t)\boldsymbol{x}_0 - S(t)\boldsymbol{y}_0\|_2 \leqslant \|\boldsymbol{x}_0 - \boldsymbol{y}_0\|_2, \ \ \forall t \geqslant 0, \ \ \boldsymbol{x}_0, \boldsymbol{y}_0 \in \mathbb{R}^n. \tag{18.2.24c}$$

Proof. According to Peano's theorem, for any $(t_0, \boldsymbol{x}_0) \in \mathbb{R}^n$ the Cauchy problem (18.2.22)-(18.2.21) admits local solutions. The dissipativity of $\boldsymbol{F}$ forces the *uniqueness* of the Cauchy problem.

To see this we argue by contradiction and assume that this Cauchy problem admits another solution $\boldsymbol{x} = \widetilde{\boldsymbol{\varphi}}(t)$. On their common domain of existence $[t_0, t_1)$ the functions $\boldsymbol{\varphi}$ and $\widetilde{\boldsymbol{\varphi}}$ satisfy the differential system

$$\big(\,\boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\,\big)' = \boldsymbol{F}(\,\boldsymbol{\varphi}(t)\,) - \boldsymbol{F}(\,\widetilde{\boldsymbol{\varphi}}(t)\,) \tag{18.2.25}$$

Taking the inner product of both sides of (18.2.25) with $\boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)$ we deduce

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\|\boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\|_2^2 \ = \ \Big(\,\big(\,\boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\,\big)',\ \boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\,\Big)\\ = \big(\,\boldsymbol{f}(\,\boldsymbol{\varphi}(t)\,) - \boldsymbol{f}(\,\widetilde{\boldsymbol{\varphi}}(t)\,), \boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\,\big) \overset{(18.2.21)}{\leqslant} 0, \ \ \forall t \in [t_0, t_1).\end{gathered} \tag{18.2.26}$$

Thus

$$\|\boldsymbol{\varphi}(t) - \widetilde{\boldsymbol{\varphi}}(t)\|_2^2 \leqslant \|\boldsymbol{\varphi}(t_0) - \widetilde{\boldsymbol{\varphi}}(t_0)\|_2^2, \ \ \forall t \in [t_0, t_1). \tag{18.2.27}$$

This proves that $\boldsymbol{\varphi} = \widetilde{\boldsymbol{\varphi}}$ on $[t_0, t_1)$ since $\widetilde{\boldsymbol{\varphi}}(t_0) = \boldsymbol{\varphi}(t_0)$.

Let $\boldsymbol{\varphi}$ be the unique solution of the Cauchy problem (18.2.22)-(18.2.21) defined on a *maximal* interval $[t_0, T)$. To prove that $\boldsymbol{\varphi}$ is defined on the entire semi axis $[t_0, \infty)$ we first prove that it is bounded on $[t_0, T)$. To achieve this we take the inner product of

$$\boldsymbol{\varphi}'(t) = \boldsymbol{F}(\,\boldsymbol{\varphi}(t)\,)$$

with $\psi(t) = \boldsymbol{\varphi}(t) - \boldsymbol{x}_0$ and we deduce

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\|\psi(t)\|_2^2 = \big(\,\boldsymbol{F}(\,\boldsymbol{\varphi}(t)\,), \boldsymbol{\varphi}(t) - \boldsymbol{x}_0\,\big)\\ = \big(\,\boldsymbol{F}(\,\boldsymbol{\varphi}(t)\,) - \boldsymbol{F}(\boldsymbol{x}_0), \boldsymbol{\varphi}(t) - \boldsymbol{x}_0\,\big) + \big(\,\boldsymbol{F}(\boldsymbol{x}_0), \boldsymbol{\varphi}(t) - \boldsymbol{x}_0\,\big)\\ \overset{(18.2.21)}{\leqslant} \ \|\boldsymbol{F}(0)\|_2 \cdot \|\psi(t)\|_2, \ \ \forall t \in [t_0, T).\end{gathered}$$

Integrating this inequality on $[t_0, t]$ and setting $u(t) := \|\boldsymbol{\varphi}(t) - \boldsymbol{x}_0\|_2$, $C = \|\boldsymbol{F}(\boldsymbol{x}_0)\|_2$ we deduce

$$\frac{1}{2}u(t)^2 \leqslant C\int_{t_0}^t u(s)ds, \ \ \forall t \in [t_0, T).$$

From Proposition 18.1.4 we deduce

$$\|\boldsymbol{\varphi}(t) - \boldsymbol{x}_0\|_2 = u(t) \leqslant \|\boldsymbol{F}(\boldsymbol{x}_0)\|_2(t - t_0), \ \ \forall t \in [t_0, T). \tag{18.2.28}$$

Since we have not assumed that the function $\boldsymbol{F}$ is locally Lipschitz, we cannot invoke directly Theorem 18.2.10 or Theorem 18.2.12. However, the inequality (18.2.28) implies in a similar fashion the equality $T = \infty$. Here are the details.

We argue by contradiction and we assume that $T < \infty$. The inequality (18.2.28) implies that there exists an increasing sequence (t_k) and $\boldsymbol{v} \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} t_k = T, \;\; \lim_{k\to\infty} \boldsymbol{\varphi}(t_k) = \boldsymbol{v}.$$

According to the facts established so far there exists a solution $\boldsymbol{\psi}$ of (18.2.22) defined on $[T-\delta, T+\delta]$ and satisfying the initial condition $\boldsymbol{\psi}(T) = \boldsymbol{v}$.

On the interval $[T-\delta, T)$ we have

$$\boldsymbol{\varphi}'(t) - \boldsymbol{\psi}'(t) = \boldsymbol{F}(\,\boldsymbol{\varphi}(t)\,) - \boldsymbol{F}(\,\boldsymbol{\psi}(t)\,).$$

Taking the inner product of this equality with $\boldsymbol{\varphi}(t) - \boldsymbol{\psi}(t)$ and using the dissipativity condition (18.2.21) we deduce as before that

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{\varphi}(t) - \boldsymbol{\psi}(t)\|^2 \leqslant 0, \;\; \forall t \in [T-\delta, T).$$

Hence

$$\|\boldsymbol{\varphi}(t) - \boldsymbol{\psi}(t)\|^2 \leqslant \|\boldsymbol{\varphi}(t_k) - \boldsymbol{\psi}(t_k)\|^2, \;\; \forall t \in [t_k, T).$$

Since

$$\lim_{k\to\infty} \|\boldsymbol{\varphi}(t_k) - \boldsymbol{\psi}(t_k)\| = 0,$$

we conclude that

$$\lim_{t\nearrow T} \boldsymbol{\varphi}(t) = \boldsymbol{v}.$$

We can thus extend $\boldsymbol{\varphi}$ to a continuous function in the closed interval $[t_0, T]$. For $t \in [t_0, T)$ we have

$$\boldsymbol{\varphi}(t) = \boldsymbol{\varphi}(t_0) + \int_{t_0}^{t} \boldsymbol{F}(\,\boldsymbol{\varphi}(s)\,)ds, \;\; \forall t_0 \leqslant t < T$$

Letting $t \nearrow T$ we deduce

$$\boldsymbol{v} = \boldsymbol{\varphi}(T) = \boldsymbol{\varphi}(t_0) + \int_{t_0}^{T} \boldsymbol{F}\big(\,\varphi(s)\,\big)ds.$$

Now define

$$\tilde{\boldsymbol{\varphi}} : [t_0, T+\delta) \to \mathbb{R}^n, \;\; \tilde{\boldsymbol{\varphi}}(t) = \begin{cases} \boldsymbol{\varphi}(t), & 0 \leqslant t \leqslant T, \\ \boldsymbol{\psi}(t), & T < t < T - \delta. \end{cases}$$

For $t \in [0, T]$ this function satisfies the integral equation

$$\tilde{\boldsymbol{\varphi}}(t) - \tilde{\boldsymbol{\varphi}}(t_0) = \int_{t_0}^{t} \boldsymbol{F}\big(\,\tilde{\boldsymbol{\varphi}}(s)\,\big)ds,$$

while for $T < t < T + \delta$ it satisfies

$$\tilde{\boldsymbol{\varphi}}(t) - \tilde{\boldsymbol{\varphi}}(t_0) = \tilde{\boldsymbol{\varphi}}(t) - \tilde{\boldsymbol{\varphi}}(T) + \tilde{\boldsymbol{\varphi}}(T) - \tilde{\boldsymbol{\varphi}}(t_0)$$

$$= \int_T^t \boldsymbol{F}\big(\tilde{\boldsymbol{\varphi}}(s)\big)ds + \int_{t_0}^T \boldsymbol{F}\big(\tilde{\boldsymbol{\varphi}}(s)\big)ds = \int_{t_0}^t \boldsymbol{F}\big(\tilde{\boldsymbol{\varphi}}(s)\big)ds.$$

Thus $\tilde{\boldsymbol{\varphi}}(t)$ is a solution of the Cauchy problem (18.2.22)+(18.2.23) on the interval $[t_0, T+\delta)$ containing strictly the assumed maximal existence interval $[t_0, T)$. This proves that $T = \infty$.

To prove (18.2.24b) we observe that both functions

$$\boldsymbol{y}_1(t) = S(t+s)\boldsymbol{x}_0 \ \text{ and } \ \boldsymbol{y}_2(t) = S(t)S(s)\boldsymbol{x}_0,$$

satisfy the equations (18.2.22) and have identical values at $t = 0$. The uniqueness of the Cauchy problems for (18.2.22) now implies that $\boldsymbol{y}_1(t) = \boldsymbol{y}_2(t)$, $\forall t \geqslant 0$.

The inequality (18.2.24c) now follows from (18.2.27) where $\widetilde{\boldsymbol{\varphi}}(t) = \boldsymbol{x}(t; 0, \boldsymbol{y}_0)$.

□

Remark 18.2.23. (a) A family of maps $S(t) : \mathbb{R}^n \to \mathbb{R}^n$, $t \geqslant 0$, satisfying (18.2.24a), (18.2.24b), (18.2.24c) is called a *continous semigroup of contractions* in the space $\mathbb{R}^n$. The function $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is called the *generator* of the semigroup $S(t)$.

(b) The forward uniqueness of the Cauchy problem (18.2.22)+(18.2.23) does not signify that one should expect backwards uniqueness as well. Consider the continuous decreasing function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \begin{cases} -x^{1/3}, & x \geqslant 0, \\ |x|^{1/3}, & x < 0. \end{cases}$$

The function $x : \mathbb{R} \to \mathbb{R}$ $x(t) = 0$, $\forall t \in \mathbb{R}$ is a solution of the Cauchy problem

$$x'(t) = f\big(x(t)\big), \quad \forall t \in \mathbb{R}, \quad x(0) = 0. \tag{18.2.29}$$

Any other solution $y(t)$ of this Cauchy problem satisfies $y(t) = x(t) = 0$, $\forall t \geqslant 0$. There exists however solutions $y(t)$ such that $y(t) \neq x(t)$ for $t < 0$. We can find such a $y(t)$ using separation of variables.

For $t < 0$ we have $y' = |y|^{1/3}$ so y is increasing. In particular $y(t) \leqslant 0$, $\forall t \leqslant 0$. We have

$$\int_t^0 y'(s)|y(s)|^{-1/3}(s)ds = -t$$

so

$$-\frac{3}{2}|y|^{2/3}\Big|_{s=t}^{s=0} = -t = |t|, \quad |y(t)|^{2/3} = \frac{2|t|}{3}, \quad |y(t)| = \left(\frac{2|t|}{3}\right)^{3/2}.$$

Since $y(t) \leqslant 0$ we deduce

$$y(t) = -\left(\frac{2|t|}{3}\right)^{3/2}.$$

$$y(t) = \begin{cases} 0, & t \geqslant 0 \\ -\left(\frac{2|t|}{3}\right)^{3/2}, & t < 0 \end{cases}$$

is a nontrivial solution of (18.2.29). □

18.2.5. Continuous dependence on initial conditions and parameters. We now return to the differential system (18.2.13) defined on the open subset $\Omega \subset \mathbb{R}^{n+1}$. We will assume as in the previous section that the function $\boldsymbol{F} : \Omega \to \mathbb{R}^n$ is continuous in the variables $(t, \boldsymbol{x})$, and locally Lipschitz in the variable $\boldsymbol{x}$. Theorem 18.2.6 shows that for any $(t_0, \boldsymbol{x}_0) \in \Omega$ there exists a unique solution $\boldsymbol{x} = \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$ of the system (18.2.13) that verifies the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$. The solution $\boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$, which we will assume to be saturated, is defined on an interval typically dependent of the point $(t_0, \boldsymbol{x}_0)$. For simplicity we will assume the initial moment t_0 to be fixed.

Denote by $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ the right-saturated solution of the initial value problem

$$x'(t) = \boldsymbol{F}\big(t, \boldsymbol{x}(t)\big), \;\; \boldsymbol{x}(t_0 = \boldsymbol{v}.$$

Denote by $\big[\, t_0, T(t_0, \boldsymbol{v})\,\big)$ the maximal existence intervals of the right-saturated solution $x(t; t_0, v)$. It is reasonable to expect that as $\boldsymbol{v}$ varies in a neighborhood of $\boldsymbol{x}_0$, the corresponding right-saturated solution $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ will not stray too far from the solution $\boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$. The next theorem confirms that this is the case, in a rather precise form. To state this result, let us denote by $S(\boldsymbol{x}_0, \eta)$ the open box/ball of center $\boldsymbol{x}_0$ and radius η in $\mathbb{R}^n$, i.e.,

$$S(\boldsymbol{x}_0, \eta) := \big\{\, \boldsymbol{v} \in \mathbb{R}^n; \;\; \|\boldsymbol{v} - \boldsymbol{x}_0\| < \eta \,\big\}.$$

Theorem 18.2.24 (Continuous dependence on initial data)**.** *Fix $(t_0, \boldsymbol{x}_0) \in \Omega$ and set $T = T(t_0, \boldsymbol{x}_0)$. Then, for any $T' \in [t_0, T)$, there exists $\rho = \rho(T') > 0$ such that the following hold.*

(i) *For any $\boldsymbol{v} \in S(\boldsymbol{x}_0, \rho)$, we have $T(t_0, v) > T'$, i.e., the right-saturated solution $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ is defined on the interval $[t_0, T']$.*

(ii) *The correspondence*

$$S(\boldsymbol{x}_0, \rho) \ni \boldsymbol{v} \mapsto \boldsymbol{x}(t; t_0, \boldsymbol{v}) \in C\big([t_0, T']; \mathbb{R}^n\,\big)$$

is a continuous map from the box $S(\boldsymbol{x}_0, \rho)$ to the space Banach space of continuous maps from $[t_0, T']$ to $\mathbb{R}^n$ equipped with the sup-norm. In other words, for any sequence $(\boldsymbol{v}_k)$ in $S(\boldsymbol{x}_0, \rho)$ that converges to $\boldsymbol{v} \in S(\boldsymbol{x}_0, \rho)$, the functions $\boldsymbol{x}(t; t_0, \boldsymbol{v}_k)$ are defined on $[t_0, T']$ and converge uniformly on $[t_0, T']$ to $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ as $k \to \infty$.

Proof. Fix $T' \in [t_0, T)$. Denote by Γ the graph of the solution $\boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$,

$$\Gamma := \big\{(t, \boldsymbol{x}(t; t_0, \boldsymbol{x}_0); \;\; t \in [t_0, T']\,\big\}.$$

Since $t \to \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$ is continuous we deduce that Γ is a compact subset of Ω so that

$$\eta := \operatorname{dist}(\Gamma, \partial\Omega) > 0.$$

For any $r > 0$ we denote by $\Gamma(r)$ the open set

$$\Gamma(r) := \big\{\, (s, \boldsymbol{y}) \in \mathbb{R} \times \mathbb{R}^n; \;\; \operatorname{dist}\big(\, (s, \boldsymbol{y}), \Gamma\,\big) < r \,\big\}$$

For $r < \eta$, the closure of $\Gamma(r)$ is compact and contained in Ω. We denote by K the closure of $\Gamma(\eta/2)$.

For any $(t_0, \boldsymbol{v}) \in \Gamma(\eta/2)$, there exists a maximal $T_v > t_0$ such that the solution $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ exists for all $t \in [t_0, T_v)$ and

$$\left\{ \,(\, t, \boldsymbol{x}(t; t_0, \boldsymbol{v})\,);\;\; t_0 \leqslant t < T_v \,\right\} \subset \Gamma(\eta/2) \subset K.$$

Set $T'_v := \min(T_v, T')$. On the interval $\big[\, t_0, T'_v\big)$ we have the equality

$$\boldsymbol{x}(t; t_0, \boldsymbol{x}_0) - \boldsymbol{x}(t; t_0, \boldsymbol{v}) = \int_{t_0}^{t} \Big(\boldsymbol{F}\big(\, s, \boldsymbol{x}(s; t_0, \boldsymbol{x}_0)\,\big) - \boldsymbol{F}\big(\, s, \boldsymbol{x}(s; t_0, \boldsymbol{v})\,\big) \Big) ds.$$

Because the graphs of $\boldsymbol{x}(s; t_0, \boldsymbol{x}_0)$ and $\boldsymbol{x}(s; t_0, \boldsymbol{v})$ over $\big[\, t_0, T'_v\big)$ are contained in the compact set K, the locally Lipschitz assumption implies that there exists a constant $L_K > 0$ such that

$$\big\| \boldsymbol{x}(t; t_0, \boldsymbol{x}_0) - \boldsymbol{x}(t; t_0, \boldsymbol{v}) \big\| \leqslant \| \boldsymbol{x}_0 - \boldsymbol{v} \| + L_K \int_{t_0}^{t} \big\|\, \boldsymbol{x}(s; t_0, \boldsymbol{x}_0) - \boldsymbol{x}(s; t_0, \boldsymbol{v})\, \big\| ds.$$

Gronwall's inequality now implies

$$\big\| \boldsymbol{x}(t; t_0, \boldsymbol{x}_0) - \boldsymbol{x}(t; t_0, \boldsymbol{v}) \big\| \leqslant e^{L_K(t-t_0)} \| \boldsymbol{x}_0 - \boldsymbol{v} \|, \;\; \forall t \in [t_0, T'_v). \tag{18.2.30}$$

Set

$$\rho = \rho(T') := \frac{\eta e^{-L_K(T'-t_0)}}{4}.$$

If $\| \boldsymbol{x}_0 - \boldsymbol{v} \| \leqslant \rho$ we deduce from (18.2.30) that

$$\forall t \in \big[\, t_0, T'_v\,\big), \;\; \big\| \boldsymbol{x}(t; t_0, \boldsymbol{x}_0) - \boldsymbol{x}(t; t_0, \boldsymbol{v}) \big\| < \frac{\eta}{4}.$$

This proves that the graph of the restriction of $\boldsymbol{x}(t; t_0, \boldsymbol{v})$ to $[t_0, T'_v)$ is contained in a compact subset of $\Gamma(\eta/2)$ so this solution is right-extendible, i.e., $T_v > T'_v = \min(T_v, T')$. In other words $T_v > T'$ if $\| \boldsymbol{x}_0 - \boldsymbol{v} \| \leqslant \rho$.

More generally, if $\| \boldsymbol{v} - \boldsymbol{x}_0 \|, \| \boldsymbol{v}' - \boldsymbol{x}_0 \| < \rho$, then arguing as in the proof of (18.2.30) we deduce

$$\big\| \boldsymbol{x}(t; t_0, \boldsymbol{v}') - \boldsymbol{x}(t; t_0, \boldsymbol{v}) \big\| \leqslant e^{L_K(t-t_0)} \| \boldsymbol{v}' - \boldsymbol{v} \|, \;\; \forall t \in [0, T'],$$

i.e.,

$$\sup_{t \in [t_0, T']} \big\| \boldsymbol{x}(t; t_0, \boldsymbol{v}') - \boldsymbol{x}(t; t_0, \boldsymbol{v}) \big\| \leqslant e^{L_K(T'-t_0)} \| \boldsymbol{v}' - \boldsymbol{v} \|.$$

Thus the map

$$S(\boldsymbol{x}_0, \rho) \ni \boldsymbol{v} \mapsto \boldsymbol{x}(t; t_0, \boldsymbol{v}) \in C\big([t_0, T']; \mathbb{R}^n \,\big)$$

is Lipschitz, hence continuous. This completes the proof of Theorem 18.2.24. □

Let us now consider the special case when the system (18.2.13) is autonomous, i.e., the map $\boldsymbol{F}$ is independent of t. More precisely, we assume that $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitz function. One should think of $\boldsymbol{F}$ as a vector field on $\mathbb{R}^n$: to each point $\boldsymbol{x} \in \mathbb{R}^n$ we attach the vector $\boldsymbol{F}(\boldsymbol{x})$. The solutions of the o.d.e. $\boldsymbol{x}'(t) = \boldsymbol{F}\big(\boldsymbol{x}(t)\big)$ are called the *flow lines* of the vector field; see Figure 18.3.

For any $\boldsymbol{u} \in \mathbb{R}^n$ we set

$$\Phi^t(\boldsymbol{u}) := \boldsymbol{x}(t; \boldsymbol{u}),$$

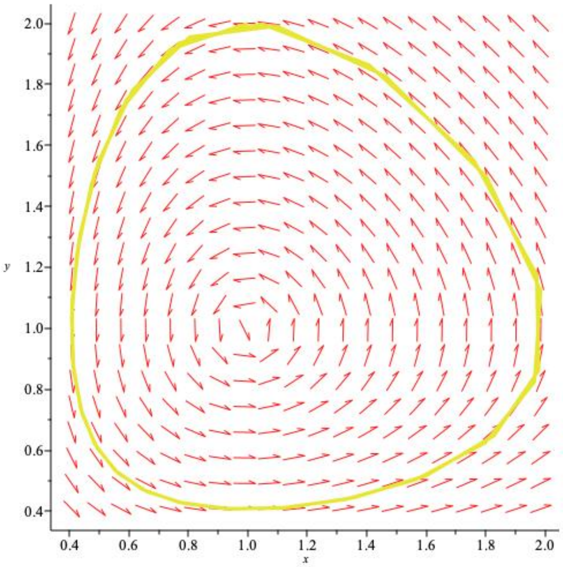

Figure 18.3. *The planar vector field $\boldsymbol{F}(x,y) = \big(-x+xy, y-xy\big)$ and one of its flow lines.*

where $\boldsymbol{x}(t;0,\boldsymbol{u})$ is the unique saturated solution of the system

$$\boldsymbol{x}' = \boldsymbol{F}(\boldsymbol{x}), \tag{18.2.31}$$

satisfying the initial condition $\boldsymbol{x}(0) = \boldsymbol{u}$.

Suppose for simplicity that $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is such that, for any $\boldsymbol{u} \in \mathbb{R}^n$, the saturated solution $\boldsymbol{x}(t;\boldsymbol{u})$ exists for all $t \in \mathbb{R}$. We obtain a collection of maps

$$\Phi^t_{\boldsymbol{F}} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n,\;\; \mathbb{R}\times\Omega \ni (t,\boldsymbol{x}_0) \mapsto \Phi^t_{\boldsymbol{F}}(\boldsymbol{x}_0) = \boldsymbol{x}(t;\boldsymbol{x}_0) \tag{18.2.32}$$

Proposition 18.2.25. *The collection of maps $(\Phi^t)_{t\in\mathbb{R}}$ satisfies the following conditions.*

(i) $\Phi^t_{\boldsymbol{F}}$ *is a continuous map* $\mathbb{R}^n \to \mathbb{R}^n$ *for any* $t \in \mathbb{R}$.

(ii) $\Phi^0_{\boldsymbol{F}} = \mathbb{1}_\Omega$.

(iii) $\Phi^t_{\boldsymbol{F}} \circ \Phi^s_{\boldsymbol{F}} = \Phi^{t+s}_{\boldsymbol{F}}(\boldsymbol{x}_0)$, $\forall s,t \in \mathbb{R}$.

(iv) $\Phi^t_{\boldsymbol{F}}$ *is a homeomorphism with inverse* $\Phi^{-t}_{\boldsymbol{F}}$.

(v) *For any* $\boldsymbol{x}_0 \in \mathbb{R}^n$

$$\boldsymbol{F}(\boldsymbol{x}_0) = \lim_{t\to 0}\frac{1}{t}\big(\,\Phi^t_{\boldsymbol{F}}(\boldsymbol{x}_0) - \boldsymbol{x}_0\,\big)$$

Proof. The continuity condition (i) follows from the continuous dependence of solutions on initial data. Condition (ii) is equivalent with the tautological equality $\boldsymbol{x}(0;\boldsymbol{x}_0) = \boldsymbol{x}_0$, $\forall \boldsymbol{x}_0 \in \Omega$. To prove (iii) fix $\boldsymbol{x}_0 \in \Omega$ and set $\boldsymbol{x}(t) := \boldsymbol{x}(t;\boldsymbol{x}_0)$. Note that the function

$\boldsymbol{y}(t) = \boldsymbol{x}(t+s)$ is the solution of the Cauchy problem

$$\boldsymbol{y}(0) = \boldsymbol{x}(s), \ \ \boldsymbol{y}'(t) = \frac{d\boldsymbol{x}}{dt}(t+s) = \boldsymbol{F}\big(\boldsymbol{x}(t+s)\big) = F\big(\boldsymbol{y}(t)\big).$$

The function $\boldsymbol{z}(t) = \boldsymbol{x}\big(t; \boldsymbol{x}(s)\big)$ satisfies the same initial value problem and the global uniqueness implies$\boldsymbol{y}(t) = \boldsymbol{z}(t)$, $\forall t$, i.e.,

$$\Phi^{t+s}_{\boldsymbol{F}}(\boldsymbol{x}_0) = \boldsymbol{y}(t) = \boldsymbol{x}\big(t; \boldsymbol{x}(s)\big) = \Phi^t_{\boldsymbol{F}}\big(\boldsymbol{x}(s)\big) = \Phi^t_{\boldsymbol{F}}\big(\Phi^s_{\boldsymbol{F}}(\boldsymbol{x}_0)\big).$$

The equality (iv) follows from (ii) and (iii). Finally (v) is equivalent with the equality

$$\boldsymbol{x}'\big(0; \boldsymbol{x}_0\big) = \boldsymbol{F}(\boldsymbol{x}_0).$$

□

Definition 18.2.26. Let (X, d) be a metric space. A *topological dynamical system* or *flow* on X is a collection of maps $\Phi^t : X \to X$, $t \in \mathbb{R}$, satisfying the conditions (i)-(iii) in Proposition 18.2.25.

The dynamical system $\Phi_{\boldsymbol{F}}$ defined in (18.2.32) is called the *flow generated by the vector field* $\boldsymbol{F}$. □

Example 18.2.27. Suppose that A is an $n \times n$ matrix with real entries. We view A as a continuous linear operator$A : \mathbb{R}^n \to \mathbb{R}^n$. It has a well defined exponential (see Exercise 17.26)

$$e^{tA} : \mathbb{R}^n \to \mathbb{R}^n, \ \ e^{tA}\boldsymbol{x} = \sum_{n \geqslant 0} \frac{t^n}{n!} A^n \boldsymbol{x}, \ \ t \in \mathbb{R}.$$

As shown in Exercise 17.26 the exponential map satisfies

$$e^{tA} e^{sA} = e^{(t+s)A}, \ \ \forall s, t \in \mathbb{R}$$

so the collection of linear homeomorphisms $e^{tA} : \mathbb{R}^n \to \mathbb{R}^n$ is a dynamical system on $\mathbb{R}^n$. Its called the linear flow generated by A.

Fix $\boldsymbol{x}_0 \in \mathbb{R}^n$ and set $\boldsymbol{x}(t) = e^{tA}\boldsymbol{x}_0$. Then $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and, as shown in Exercise 17.26,

$$\dot{\boldsymbol{x}}(t) = \frac{d}{dt} e^{tA}\boldsymbol{x}_0 = A e^{tA}\boldsymbol{x}_0 = A\boldsymbol{x}(t).$$

In other words, $\boldsymbol{x}(t)$ is the solution of the initial value problem

$$\dot{\boldsymbol{x}} = A\boldsymbol{x}, \ \ \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Thus e^{tA} is the flow generated by the linear vector field $\boldsymbol{F}(\boldsymbol{x}) = A\boldsymbol{x}$. □

Consider now the differential system

$$\boldsymbol{x}' = \boldsymbol{F}(t, \boldsymbol{x}, \lambda), \ \ \lambda \in \Lambda \subset \mathbb{R}^m, \tag{18.2.33}$$

where $\boldsymbol{F} : \Omega \times \Lambda \to \mathbb{R}^n$ is a continuous function, Ω is an open subset of $\mathbb{R}^{n+1}$, and Λ is an open subset of $\mathbb{R}^m$. Additionally, we will assume that $\boldsymbol{F}$ is locally Lipschitz in $(\boldsymbol{x}, \lambda)$ on

$\Omega \times \Lambda$. In other words, for any compact sets $K_1 \subset \Omega$ and $K_2 \subset \Lambda$ there exists a positive constant L such that

$$\begin{gathered}\|\boldsymbol{F}(t,\boldsymbol{x},\lambda) - \boldsymbol{F}(t,\boldsymbol{y},\mu)\| \leqslant L\big(\,\|\boldsymbol{x}-\boldsymbol{y}\| + \|\lambda-\mu\|\,\big),\\ \forall (t,\boldsymbol{x}),(t,\boldsymbol{y}) \in K_1,\ \ \lambda,\mu \in K_2.\end{gathered} \tag{18.2.34}$$

Above, we denoted by the same symbol the norms $\|-\|$ in $\mathbb{R}^m$ and $\mathbb{R}^n$.

For any $(t_0,\boldsymbol{x}_0) \in \Omega$, and $\lambda \in \Lambda$, the system (18.2.33) admits a unique solution $\boldsymbol{x} = \boldsymbol{x}(t;t_0,\boldsymbol{x}_0,\lambda)$ satisfying the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$. Loosely speaking, our next result states that the correspondence $\lambda \mapsto \boldsymbol{x}(-;t_0,\boldsymbol{x}_0,\lambda)$ is continuous.

Theorem 18.2.28 (Continuous dependence on parameters)**.** *Fix a point* $(t_0,\boldsymbol{x}_0,\lambda_0) \in \Omega\times\Lambda$. *Let* $[t_0,T)$ *be the maximal interval of existence on the right of the solution* $\boldsymbol{x}(t;t_0,\boldsymbol{x}_0,\lambda_0)$. *Then, for any* $T' \in [t_0,T)$ *there exists* $\eta = \eta(T') > 0$ *such that for any* $\lambda \in S(\lambda_0,\eta)$ *the solution* $\boldsymbol{x}(t;t_0,\boldsymbol{x}_0,\lambda)$ *is defined on* $[t_0,T']$. *Moreover, the application*

$$S(\lambda_0,\eta) \ni \lambda \mapsto \boldsymbol{x}(-;t_0,\boldsymbol{x}_0,\lambda) \in C\big(\,[t_0,T'],\mathbb{R}^n\big)$$

is continuous.

Proof. The above result is a special case of Theorem 18.2.24 on the continuous dependence on initial data.

We denote by $\boldsymbol{z}$ the $(n+m)$-dimensional vector $\boldsymbol{z} = (\boldsymbol{x},\lambda) \in \mathbb{R}^{n+m}$, and we define

$$\widetilde{\boldsymbol{F}} : \Omega\times\Lambda \to \mathbb{R}^{n+m},\ \ \widetilde{\boldsymbol{F}}(t,\boldsymbol{x},\lambda) = \big(\,\boldsymbol{F}(t,\boldsymbol{x},\lambda),0\,\big) \in \mathbb{R}^n\times\mathbb{R}^m.$$

The system (18.2.33) can be rewritten as

$$\dot{\boldsymbol{x}} = \boldsymbol{F}(t,\boldsymbol{x},\lambda),\ \ \dot{\lambda} = 0,$$

or equivalently,

$$\boldsymbol{z}'(t) = \widetilde{\boldsymbol{F}}\big(\,t,\boldsymbol{z}(t)\,\big). \tag{18.2.35}$$

The initial condition becomes

$$\boldsymbol{z}(t_0) = \boldsymbol{z}_0 := (\boldsymbol{x}_0,\lambda_0). \tag{18.2.36}$$

We have thus reduced the problem to investigating the dependence of the solutions $\boldsymbol{z}(t)$ of (18.2.35) on the initial data. Our assumptions on $\boldsymbol{F}$ show that $\widetilde{\boldsymbol{F}}$ satisfies the assumptions of Theorem 18.2.24. □

18.3. A brief introduction to the calculus of variations

Let $n \in \mathbb{N}$ and suppose that $\mathcal{O} \subset \mathbb{R}^n$ be an open subset. We will denote the points in $\mathbb{R}^n\times\mathbb{R}^n$ by $(\boldsymbol{x},\boldsymbol{v})$,

$$\boldsymbol{q} = (\boldsymbol{q}^1,\dots,\boldsymbol{q}^n),\ \ \boldsymbol{v} = (v^1,\dots,v^n)$$

You should think of $\boldsymbol{q}$ as position and of $\boldsymbol{v}$ as velocity.

Fix a C^1-function $L : \mathcal{O} \times \mathbb{R}^n \to \mathbb{R}$. We will refer to L as *Lagrangian function* or simply *Lagrangian.*

Example 18.3.1 (Classical mechanics Lagrangian)**.** Suppose that $U : \mathcal{O} \to \mathbb{R}$ is a C^2-function. The function U is known classically as the potential function. To U we associate the Lagrangian

$$L(\boldsymbol{q}, \boldsymbol{v}) = \frac{1}{2}|\boldsymbol{v}|^2 - U(\boldsymbol{q}),$$

where $| - |$ denotes the Euclidean norm on $\mathbb{R}^n$. □

Fix two points $\boldsymbol{q}_0, \boldsymbol{q}_1 \in \mathcal{O}$ and denote by $\mathcal{P}(\boldsymbol{q}^0, \boldsymbol{q}^2)$ the space of C^1-paths $\gamma : [0,1] \to \mathcal{O}$, such that $\gamma(0) = \boldsymbol{q}_0$, $\gamma(1) = \boldsymbol{q}_1$.

The Lagrangian L defines a functional

$$S_L : \mathcal{P}(\boldsymbol{q}^0, \boldsymbol{q}^1) \to \mathbb{R}, \;\; S_L\big[\,\gamma\,\big] = \int_0^1 L\big(\,\boldsymbol{\gamma}(t), \dot{\gamma}(t)\,\big)dt.$$

The scalar $S_L\big[\,\gamma\,\big]$ is known as the *action* of the path γ determined by the Lagrangian L. The functional S_L is usually referred as the *action functional.*

The Main Problem of the Calculus of Variation *Find the least action paths among all the paths in* $\mathcal{P}(\boldsymbol{q}_0, \boldsymbol{q}_1)$*, i.e., find the paths* $\alpha \in \mathcal{P}(\boldsymbol{q}_0, \boldsymbol{q}_1)$ *such that*

$$S_L\big[\,\alpha\,\big] \leqslant S_L\big[\,\gamma\,\big], \;\; \forall \gamma \in \mathcal{P}(\boldsymbol{x}_0, \boldsymbol{x}_1).$$

18.3.1. The Euler-Lagrange equations. Let us observe that a priori the action functional $S_L : \mathcal{P}(\boldsymbol{q}^0, \boldsymbol{q}^1) \to \mathbb{R}$ may not bounded from below so there might not exists least action paths $\alpha \in \mathcal{P}(\boldsymbol{q}^0, \boldsymbol{q}^1)$. Leaving this problem aside, assume that such minimizers exist. Is there a procedure that will give us a reasonable chance of finding them?

You should compare this with a similar problem in one variable calculus: give a C^1-function $f : \mathbb{R} \to \mathbb{R}$, find $x_0 \in \mathbb{R}$ such that $f(x_0) \leqslant f(x)$, $\forall x \in \mathbb{R}$.

In that case we had Fermat's principle that states that if x_0 is a minimizer, then $f'(x_0) = 0$. Thus, if the minimizers exist, they are located in the much smaller set of critical points of f. This does does not mean that some critical point must be a minimizer, but we have gained a lot: we really need to investigate only a (typically) discrete set of points to solve the minimization problem.

A similar situation, with some extra twists, occurs in infinite dimensions when trying to solve the main problem of the calculus of variations.

Theorem 18.3.2 (Euler-Lagrange)**.** *Suppose that* $L : \mathcal{O} \times \mathbb{R}^n \to \mathbb{R}$ *is a* C^2 *function. Fix* $\boldsymbol{q}_0, \boldsymbol{q}_1 \in \mathcal{O}$ *and consider the action functional*

$$S :: \mathcal{P}(\boldsymbol{q}_0, \boldsymbol{q}_1) \to \mathbb{R}, \;\; S_L[\gamma] = \int_0^1 L\big(\,\gamma(t), \dot{\gamma}t\,\big).$$

Suppose that $\gamma_0 \in \mathcal{P}(\boldsymbol{q}_0, \boldsymbol{q}_1)$ *is a* C^2 *path that minimizes* S_L. *Then*

$$\gamma_0(t) = \big(\, \gamma_0^1(t), \dots \gamma_0^n(t)\,\big)$$

satisfies the Euler-Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial v^i}\big(\,\gamma(t), \dot{\gamma}(t)\,\big) = \frac{\partial L}{\partial q^i}\big(\,\gamma(t), \dot{\gamma}(t)\,\big), \;\; i = 1, \dots, n. \tag{18.3.1}$$

We can rewrite these in a more compact form

$$\frac{d}{dt}\frac{\partial L}{\partial \boldsymbol{v}}(\gamma(t), \dot{\gamma}(t)\,\big) = \frac{\partial L}{\partial \boldsymbol{q}}\big(\,\gamma(t), \dot{\gamma}(t)\,\big), \;\; i = 1, \dots, n. \tag{18.3.2}$$

Proof. Consider the curve C traced by the path γ_0

$$C := \gamma_0\big(\,[0,1]\,\big) \subset \mathcal{O}.$$

The curve C is a compact subset of the open set $\mathcal{O}$ and thus

$$r_0 = \operatorname{dist}\big(\,C_0, \partial\mathcal{O}\,\big) > 0.$$

Denote by $\mathcal{V}_0$ the space of C^1 paths $\alpha : [0,1] \to \mathbb{R}$ such that $\alpha(0) = \alpha(1) = 0$. Classically the paths in $\mathcal{V}_0$ are known as variations (of γ_0). Physicists would use the notation $\delta\gamma_0$ to denote a variation α.

Clearly, for any variation α, there exists $S = S_\alpha > 0$ such that

$$\big|\, s\alpha(t)\,\big| < r_0/2, \;\; \forall |s| < S_\alpha, \;\; t \in [0,1].$$

Fix a variation α and set $\gamma_s(t) = \gamma_0 + s\alpha(t)$, $|s| < S_\alpha$. Note that $\gamma_s \in \mathcal{P}(\boldsymbol{q}_0, \boldsymbol{q}_1)$, $\forall |s| < S_\alpha$. The family $(\gamma_s)_{|s|<S_\alpha}$ is usually referred to as a *deformation* of the path γ_0. We denote by $f_\alpha(s)$ the action of γ_s, i.e.,

$$f_\alpha(s) = S_L\big[\,\gamma_s\,\big] = \int_0^1 L\big(\,\gamma_s, \dot{\gamma}_s\,\big)dt.$$

Note that $f_\alpha(0) = S_L\big[\,\gamma_0\,\big]$, and since γ_0 is an action minimizer we deduce that

$$f_\alpha(0) \leqslant f_\alpha(s), \;\; \forall |s| < S_\alpha.$$

Invoking Fermat's principle (Theorem 7.4.2) we deduce that

$$f_\alpha'(0) = 0, \;\; \forall \alpha \in \mathcal{V}_0. \tag{18.3.3}$$

The equality (18.3.3) imposes infinitely many constraints on γ_0, one constraint for each variation α. We have

$$0 = \frac{d}{ds}\,|_{s=0} \int_0^1 L\big(\,\gamma_s, \dot{\gamma}_s\,\big)dt = \int_0^1 \frac{d}{ds}\,|_{s=0} L\big(\,\gamma_s, \dot{\gamma}_s\,\big)dt.$$

Note that

$$\frac{d}{ds}\,|_{s=0}\gamma_s = \alpha, \;\; \frac{d}{ds}\,|_{s=0}\dot{\gamma}_s = \dot{\alpha} = \frac{d}{dt}\alpha.$$

If we denote by $\bullet$ the canonical inner product on $\mathbb{R}^n$, then we deduce

$$\frac{d}{ds}\,|_{s=0} L\big(\,\gamma_s, \dot{\gamma}_s\,\big) = \alpha(t) \bullet \nabla_{\boldsymbol{q}} L\big(\,\gamma_0(t), \dot{\gamma}_0(t)\,\big) + \dot{\alpha}(t) \bullet \nabla_{\boldsymbol{v}} L\big(\,\gamma_0(t), \dot{\gamma}_0(t)\,\big).$$

Integrating by parts[2] and using the fact that $\alpha(0)=\alpha(1)=0$ we deduce

$$\int_0^1 \dot{\alpha}(t)\bullet\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)dt=-\int_0^1\alpha(t)\bullet\Big(\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)\Big)dt$$

Hence

$$\begin{aligned}\int_0^1\frac{d}{ds}\Big|_{s=0}L\big(\gamma_s,\dot{\gamma}_s\big)dt&=\int_0^1\alpha(t)\bullet\nabla_{\boldsymbol{q}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)dt\\&\quad-\int_0^1\alpha(t)\bullet\Big(\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)\Big)dt\\&=\int_0^1\alpha(t)\bullet\Big(\nabla_{\boldsymbol{q}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)-\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)\Big)dt.\end{aligned}$$

Thus, if γ_0 is a C^2 action minimizer[3], then <u>*for any*</u> variation $\alpha\in\mathcal{V}_0$ we have

$$\int_0^1\alpha(t)\bullet\Big(\nabla_{\boldsymbol{q}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)-\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)\Big)dt=0.$$

This implies (see Exercise 9.9) that

$$\nabla_{\boldsymbol{q}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)-\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\gamma_0(t),\dot{\gamma}_0(t)\big)=0,\;\;\forall t\in[0,1].$$

These are precisely the Euler-Lagrange equations (18.3.1). □

Definition 18.3.3. A solution of the Euler-Lagrange equations (18.3.1) is called an *extremal* of the Lagrangian L. □

Example 18.3.4 (Conservation of energy)**.** Consider the classical mechanics Lagrangian in Example 18.3.4 ,

$$L(\boldsymbol{q},\boldsymbol{v})=\frac{1}{2}\big|\,\boldsymbol{v}\,\big|^2-U(\boldsymbol{q}).$$

Then

$$\nabla_{\boldsymbol{v}}L=\boldsymbol{v},\;\;\nabla_{\boldsymbol{q}}L=-\nabla U(\boldsymbol{q}).$$

If $\gamma(t)$ is an extremal of S_L

$$\nabla_{\boldsymbol{v}}L\big(\gamma(t),\dot{\gamma}(t)\big)=\dot{\gamma}(t),\;\;\nabla_{\boldsymbol{q}}L\big(\gamma(t),\dot{\gamma}(t)\big)=-\nabla_{\boldsymbol{q}}U(\boldsymbol{q}).$$

The Euler-Lagrange equations take the form

$$\ddot{\gamma}(t)=-\nabla_{\boldsymbol{q}}U\big(\gamma(t)\big).\tag{18.3.4}$$

If we interpret $\boldsymbol{F}(\boldsymbol{q}):=-\nabla U(\boldsymbol{q})$ as a force field, then the above equations become Newton's law of motion.

We define the energy of the Lagrangian function L to be the function

$$E:\mathcal{O}\times\mathbb{R}^n\to\mathbb{R},\;\;E(\boldsymbol{q},\boldsymbol{v})=\frac{1}{2}\big|\,\boldsymbol{v}\,\big|^2+U(\boldsymbol{q}).$$

[2] If the inner product $\bullet$ confuses you, think of $\bullet$ as multiplication of two scalars a.k.a. numbers.

[3] Can you see why the C^2 assumption on γ_0 and L was important in the above arguments?

The first term in the above sum is the kinetic energy and the second term is the potential energy. The energy is conserved along any extremal $\gamma(t)$, i.e.,

$$\frac{d}{dt}E\big(\,\gamma(t),\dot{\gamma}(t)\,\big)=0.$$

Indeed

$$E\big(\,\gamma(t),\dot{\gamma}(t)\,\big)=\frac{1}{2}\big|\,\dot{\gamma}(t)\,\big|^2+U\big(\,\gamma(t)\,\big)=\boldsymbol{v}\bullet\nabla_{\boldsymbol{v}}L\big(\,\boldsymbol{q},\boldsymbol{v}\,\big)-L\big(\,\boldsymbol{q},\boldsymbol{v}\,\big),$$

$$\frac{d}{dt}E\big(\,\gamma(t),\dot{\gamma}(t)\,\big)=\dot{\gamma}(t)\bullet\ddot{\gamma}(t)+\dot{\gamma}(t)\bullet\nabla U\big(\,\gamma(t)\,\big)=\dot{\gamma}(t)\bullet\underbrace{\Big(\,\ddot{\gamma}(t)+\nabla U\big(\,\gamma(t)\,\big)\,\Big)}_{=0}.$$

□

18.3.2. Noether's conservation principle. In a groundbreaking 1918 paper [**33**], Emmy Noether described a fundamental connection between the symmetries of a Lagrangian and conservation laws satisfied by its extremals. I want to discuss a baby case of this principle, yet strong enough to have nontrivial consequences.

Suppose that $A:\mathbb{R}^n\to\mathbb{R}^n$ is a linear operator. It defines a linear flow on $\mathbb{R}^n\times\mathbb{R}^n$

$$(\boldsymbol{q},\boldsymbol{v})\mapsto\big(\,e^{tA}\boldsymbol{q},e^{tA}\boldsymbol{v}\,\big),\;\;t\in\mathbb{R}.$$

Suppose that $\mathcal{O}\subset\mathbb{R}^n$ is an open set invariant with respect to this flow, i.e.,

$$\forall\boldsymbol{q}\in\mathcal{O},\;\;\forall t\in\mathbb{R},\;\;e^{tA}\boldsymbol{q}\in\mathcal{O}.$$

Let $L:\mathcal{O}\times\mathbb{R}^n\to\mathbb{R}$ be an A-invariant lagrangian, i..e.,

$$\big(\,e^{tA}\boldsymbol{q},e^{tA}\boldsymbol{v}\,\big)=L(\boldsymbol{q},\boldsymbol{v}),\;\;\forall(\boldsymbol{q},\boldsymbol{v})\in\mathcal{O}\times\mathbb{R}^n,\;\;t\in\mathbb{R}.$$

differentiating with respect to t the above equality we deduce that

$$0=\frac{d}{dt}\Big|_{t=0}L\big(\,e^{tA}\boldsymbol{q},e^{tA}\boldsymbol{v}\,\big)=(A\boldsymbol{q})\bullet\nabla_{\boldsymbol{q}}L(\boldsymbol{q},\boldsymbol{v})+(A\boldsymbol{v})\bullet\nabla_{\boldsymbol{v}}L(\boldsymbol{q},\boldsymbol{v}).\tag{18.3.5}$$

Define A-*momentum* of L to be the function

$$p_A:\mathcal{O}\times\mathbb{R}^n\to\mathbb{R},\;\;p_A(\boldsymbol{q})=(A\boldsymbol{q})\bullet\nabla_{\boldsymbol{v}}L(\boldsymbol{q},\boldsymbol{v})$$

Theorem 18.3.5 (Noether's conservation principle)**.** *If the Lagrangian L is A-invariant, then the A-momentum is conserved along any extremal of L. More precisely, if $\gamma(t)$ is an extremal of L, then*

$$\frac{d}{dt}p_A\big(\,\gamma(t),\dot{\gamma}(t)\,\big)=0,\;\;\forall t.$$

Proof. We have

$$\frac{d}{dt}p_A\big(\,\gamma(t),\dot{\gamma}(t)\,\big)=\frac{d}{dt}\Big(\,\big(\,A\gamma(t)\,\big)\bullet\nabla_{\boldsymbol{v}}L\big(\,\gamma(t),\dot{\gamma}(t)\,\big)\,\Big)$$

$$=\big(\,A\dot{\gamma}(t)\,\big)\bullet\nabla_{\boldsymbol{v}}L\big(\,\gamma(t),\dot{\gamma}(t)\,\big)+\big(\,A\gamma(t)\,\big)\bullet\Big(\,\frac{d}{dt}\nabla_{\boldsymbol{v}}L\big(\,\gamma(t),\dot{\gamma}(t)\,\big)\,\Big)$$

$$\overset{(18.3.1)}{=}\big(\,A\dot{\gamma}(t)\,\big)\bullet\nabla_{\boldsymbol{v}}L\big(\,\gamma(t),\dot{\gamma}(t)\,\big)+\big(\,A\gamma(t)\,\big)\bullet\nabla_{\boldsymbol{q}}L\big(\,\gamma(t),\dot{\gamma}(t)\,\big)\overset{(18.3.5)}{=}0.$$

□

Example 18.3.6. Suppose that $n = 3$. Denote by $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ the canonical basis of $\mathbb{R}^3$. Recall that $\mathbb{R}^3$ is equipped with a cross product (see Example 11.2.12)

$$\times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3.$$

Each $\boldsymbol{v} \in \mathbb{R}^3$ determines a linear operator

$$\sigma_{\boldsymbol{v}} : \mathbb{R}^3 \to \mathbb{R}^3, \;\; \sigma_{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{v} \times bx.$$

This operator is represented by a 3×3 skew-symmetric matrix that, for simplicity, we denote also by $\sigma_{\boldsymbol{v}}$. If $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ is the canonical basis of $\mathbb{R}^3$, then

$$\boldsymbol{i} \times \boldsymbol{j} = \boldsymbol{k}, \;\; \boldsymbol{j} \times \boldsymbol{k}, \;\; \boldsymbol{k} \times \boldsymbol{i} = \boldsymbol{j}.$$

The matrices $\sigma_{\boldsymbol{i}}, \sigma_{\boldsymbol{j}}, \sigma_{\boldsymbol{k}}$ are known in theoretical physics as the *Pauli matrices.* . For example,

$$\sigma_{\boldsymbol{k}} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then

$$e^{t\sigma_{\boldsymbol{k}}} = \begin{bmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

so that $e^{t\sigma_{\boldsymbol{k}}}$ describes the 1-parameter group of rotations about the $\boldsymbol{k}$-axis of $\mathbb{R}^3$. Suppose that

$$U : \mathbb{R}^3 \backslash \{0\} \to \mathbb{R}, \;\; U(\boldsymbol{q}) = -\frac{1}{|q|}.$$

and

$$L(\boldsymbol{q}, \boldsymbol{v}) = \frac{1}{2}|\boldsymbol{v}|^2 - U(\boldsymbol{q}).$$

This Lagrangian describes the motion of a planet under the attraction of the Sun that is located at the origin of $\mathbb{R}^3$.

Then

$$\nabla_{\boldsymbol{q}} L = -\frac{1}{|\boldsymbol{q}|^3}\boldsymbol{q}, \;\; \nabla_{\boldsymbol{v}} L = \boldsymbol{v}$$

and the Euler-Lagrangian equations are

$$\ddot{\gamma}(t) = -\frac{1}{|\gamma(t)|^3}\gamma(t).$$

The total energy is

$$E = \frac{1}{2}|\boldsymbol{v}|^2 - \frac{1}{|q|}$$

and it is conserved during the motion. We write

$$\boldsymbol{q} = (x, y, z), \;\; \boldsymbol{v} = (v_x, v_y, v_z)$$

The Lagrangian L is obviously $\sigma_{\boldsymbol{k}}$-invariant and its $\sigma_{\boldsymbol{k}}$-momentum is

$$p_{\boldsymbol{k}}(\boldsymbol{q},\boldsymbol{v}) = \big(\,\sigma_{\boldsymbol{k}}\boldsymbol{q}\,\big)\bullet\boldsymbol{v} = v_y x - v_x y$$

If $\gamma(t) = (x(t), y(t), z(t))$ is an extremal of S_L, then the $\sigma_{\boldsymbol{k}}$ momentul is constat along γ, i.e.,

$$p_{\boldsymbol{k}}\big(\,\gamma(t),\dot{\gamma}(t)\,\big) = x(t)\dot{y}(t) - y(t)\dot{x}(t) = \big(\,\gamma(t)\times\dot{\gamma}(t)\,\big)\bullet\boldsymbol{k}.$$

is independent of t. This is the angular momentum with respect to the $\boldsymbol{k}$-axis

Using rotations about the $\boldsymbol{i}$-axis and rotations about the $\boldsymbol{j}$-axis one obtains similarly

$$\frac{d}{dt}\big(\,\gamma(t)\times\dot{\gamma}(t)\,\big) = 0.$$

In other words, the vector $\gamma(t)\times\dot{\gamma}(t)$ is constant along an extremal. This is implies one of Kepler's laws, namely that the trajectory of a planet is planar and during the motion it sweeps equal areas in equal amount of times. The plane of motion is the plane perpendicular to the constant angular momentum or, equivalently, the plane spanned by $\gamma(0)$ and $\dot{\gamma}(0)$.

With a bit more effort one can deduce all of Kepler's laws from the equations of motion. For details we refer to [1, Sec. 2.8]. For a particularly nice explanation why the trajectory is an ellipse I refer to Feynman's lost lecture.[4] □

We close here this brief introduction to calculus of variations. The reader curious about how this story evolved can do no wrong by consulting Chapter 4 of [9], a classic gem by R. Courant and D. Hilbert.

18.4. Systems of linear differential equations

The study of the linear systems of o.d.e.-s offers an example of a well-put-together theory, based on methods and results from linear algebra. As we will see, there exist many similarities between the theory of systems of linear algebraic equations and the theory of systems of linear o.d.e.-s. In applications, linear systems appears most often as "first approximations" of more complex processes.

18.4.1. Notation and some general results. A system of first order linear o.d.e.-s has the form

$$x_i'(t) = \sum_{j=1}^{n} a_{ij}(t)x_j(t) + b_i(t),\;\; i = 1,\dots,n,\;\; t\in I, \tag{18.4.1}$$

where I is an interval of the real axis and $a_{ij}, b_i : I \to \mathbb{R}$ are continuous functions. The system (18.4.1) is called *nonhomogeneous*. If $b_i(t)\equiv 0$, $\forall i$, then the system is called *homogeneous*. In this case it has the form

$$x_i'(t) = \sum_{j=1}^{n} a_{ij}(t)x_i(t),\;\; i = 1,\dots,n,\;\; t\in I. \tag{18.4.2}$$

[4] https://youtu.be/xdIjYBtnvZU?si=1LRYSxizqitTqHrP

Using the vector notation we can rewrite (18.4.1) and (18.4.2) in the form

$$\boldsymbol{x}'(t) = A(t)\boldsymbol{x}(t) + \boldsymbol{b}(t), \;\; t \in I \tag{18.4.3a}$$

$$\boldsymbol{x}'(t) = A(t)\boldsymbol{x}(t), \;\; t \in I, \tag{18.4.3b}$$

where

$$\boldsymbol{x}(t) := \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}, \;\; \boldsymbol{b}(t) := \begin{bmatrix} b_1(t) \\ \vdots \\ b_n(t) \end{bmatrix},$$

and $A(t)$ is the $n \times n$, time dependent matrix $A(t) := \big(a_{ij}(t)\big)_{1 \leqslant i,j \leqslant n}$. We denote by $\|A(t)\|$ the norm of $A(t)$ viewed as a continuous linear operator $\mathbb{R}^n \to \mathbb{R}^n$, where $\mathbb{R}^n$ is equipped with the sup-norm $\|-\|$ as in the previous sections. For simplicity, in the sequel we will assume that $I = \mathbb{R}$.

Obviously, the local existence and uniqueness theorem (Theorem 18.2.1), as well as the results concerning global existence and uniqueness apply to the system (18.4.3a). Thus for any $t_0 \in \mathbb{R}$, and any $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exists a unique saturated solution of (18.4.1) satisfying the initial condition

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0. \tag{18.4.4}$$

In this case, the domain of existence of the saturated solution coincides with the interval I. In other words, we have the following result.

Theorem 18.4.1. *The saturated solution $\boldsymbol{x} = \boldsymbol{u}(t)$ of the Cauchy problem (18.4.1), (18.4.4) is defined on the entire interval $I = \mathbb{R}$.*

Proof. Set $\boldsymbol{F}(t, \boldsymbol{x}) = b(t) + A(t)\boldsymbol{x}$, $c(t) := \max\big(\,\|\underline{(}t)\|, \|A(t)\|_{\mathrm{op}}$.

$$\|\boldsymbol{F}(t, \boldsymbol{x})\| \leqslant \|b(t)\| + \|A(t)\|_{\mathrm{op}} \cdot \|\boldsymbol{x}\| \leqslant c(t)\big(1 + \|\boldsymbol{x}|\,\big).$$

The conclusion now follows from Corollary 18.2.16. □

18.4.2. Homogeneous systems of linear differential equations. In this section we will investigate the system (18.4.2) (equivalently, (18.4.3b)). We begin with a theorem on the structure of the set of solutions.

Theorem 18.4.2. *The set of solutions of the system (18.4.2) is a real vector space of dimension n.*

Proof. The set of solutions is obviously a real vector space. Indeed, the sum of two solutions of (18.4.2) and the multiplication by a scalar of a solution are also solutions of this system.

We will show that there exists a linear isomorphism between the space E of solutions of (18.4.2) and the space $\mathbb{R}^n$. Fix a point $t_0 \in \mathbb{R}$ and denote by Γ_{t_0} the map $E \to \mathbb{R}^n$ that associates to a solution $\boldsymbol{x} \in E$ its value at $t_0 \in E$, i.e.,

$$\Gamma_{t_0}(\boldsymbol{x}) = \boldsymbol{x}(t_0) \in \mathbb{R}^n, \;\; \forall \boldsymbol{x} \in E.$$

The map Γ_{t_0} is obviously linear. The existence and uniqueness theorem concerning the Cauchy problems associated to (18.4.2) implies that Γ_{t_0} is also surjective and injective. This completes the proof of Theorem 18.4.2. □

The above theorem shows that the space E of solutions of (18.4.2) admits a basis consisting of n solutions. Let

$$\{\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^n\},$$

be one such basis. In particular, $\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^n$ are n linearly independent solutions of (18.4.2), i.e., the only constants $c_1, c_2, \dots, c_n$ such that

$$c_1\boldsymbol{x}^1(t) + c_2\boldsymbol{x}^2(t) + \cdots + c_n\boldsymbol{x}^n(t) = 0, \quad \forall t \in I,$$

are the null ones, $c_1 = c_2 = \cdots = c_n = 0$. The matrix $\boldsymbol{X}(t)$ whose columns are given by the function $\boldsymbol{x}^1(t), \boldsymbol{x}^2(t), \dots, \boldsymbol{x}^n(t)$,

$$\boldsymbol{X}(t) := \big[\,\boldsymbol{x}^1(t), \boldsymbol{x}^2(t), \dots, \boldsymbol{x}^n(t)\,\big], \quad t \in I,$$

is called a *fundamental matrix* It is easy to see that the matrix $\boldsymbol{X}(t)$ is a solution of the differential equation

$$\boldsymbol{X}'(t) = A(t)\boldsymbol{X}(t), \quad t \in I, \tag{18.4.5}$$

where we denoted by $\boldsymbol{X}'(t)$ the matrix whose entries are the derivatives of the corresponding entries of $\boldsymbol{X}(t)$.

A fundamental matrix $X(t)$ is not unique, but it is uniquely determined by its value at a point t_0. The matrix $X(t_0)$ is invertible since the map Γ_{t_0} maps linearly independent solutions to linearly independent vectors in $\mathbb{R}^n$. Note that $\boldsymbol{X}(t)\boldsymbol{X}(t_0)^{-1}$ is the (unique!) fundamental solution determined by the solutions $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_n(t)$ satisfying

$$\boldsymbol{x}_i(t_0) = e_i, \quad i = 1, \dots, n,$$

where $e_1, \dots, e_n$ denotes the canonical basis of $\mathbb{R}^n$. This shows that if $\boldsymbol{Y}(t)$ is another fundamental solution, then

$$\boldsymbol{Y}(t)\boldsymbol{Y}(t_0)^{-1} = \boldsymbol{X}(t)\boldsymbol{X}(t_0)^{-1}$$

Thus the matrix $\boldsymbol{X}(t)\boldsymbol{X}(t_0)^{-1}$ is independent of the choice of the fundamental solution $\boldsymbol{X}(t)$. We set

$$\boxed{S(t, t_0) = S_A(t, t_0) = \boldsymbol{X}(t)\boldsymbol{X}(t_0)^{-1}}.$$

We will refer to the family of operators $S_A(t, t_0) : \mathbb{R}^n \to \mathbb{R}^n$, $t_0, t \in \mathbb{R}$ as the *propagator* or the *scattering matrix* of the homogeneous system (18.4.2). The proof of the following result is left to the reader as an exercise.

Proposition 18.4.3. *For any $t_0, t_1, t_2 \in \mathbb{R}$ we have*

$$\begin{gathered} S(t_0, t_0) = \mathbb{1}_{\mathbb{R}^n}, \\ S(t_2, t_0) = S(t_2, t_1)S(t_1, t_0), \\ S(t_0, t_1) = S(t_1, t_0)^{-1}, \quad S(t_1, t_0) = S(t_1, 0)S(0, t_0) = S(t_1, 0)S(t_0, 0)^{-1} \end{gathered} \tag{18.4.6}$$

$$\frac{d}{dt}S(t,t_0) = A(t)S(t,t_0).$$

Proof. The key identity is (18.4.6) since all the others follow from it. Set

$$U(t) := S(t,t_0),\ \ V(t) = S(t,t_1)S(t_1,t_0).$$

We have to show that $U(t) = V(t)$, $\forall t \in \mathbb{R}$. Clearly

$$U(t_1) = S(t_1,t_0) = S(t_1,t_1)S(t_1,t_0) = V(t_1).$$

Observe next that

$$U'(t) = A(t)U(t),\ \ V'(t) = A(t)V(t).$$

The equality (18.4.6) is now a consequence of the uniqueness of solutions of initial value problems for homogeneous linear systems of ode-s. □

Corollary 18.4.4. *The unique solution of the initial value problem*

$$\boldsymbol{x}'(t) = A(t)\boldsymbol{x}(t),\ \ \boldsymbol{x}(t_0) = \boldsymbol{c}_0$$

is $\boldsymbol{x}(t) = S(t,t_0)\boldsymbol{c}_0$. □

Given a collection $\{\boldsymbol{x}^1,\dots,\boldsymbol{x}^n\}$ of solutions of (18.4.2), we define the *Wronskian* of this collection to be the determinant

$$W(t) := \det \boldsymbol{X}(t), \tag{18.4.7}$$

where $\boldsymbol{X}(t)$ denotes the matrix with columns $\boldsymbol{x}^1(y),\dots,\boldsymbol{x}^n(t)$. The next result, due to the Polish mathematician *H. Wronski* (1778-1853) explains the relevance of the quantity $W(t)$.

Theorem 18.4.5. *The collection of solutions $\{\boldsymbol{x}^1,\dots,\boldsymbol{x}^n\}$ of (18.4.2) is linearly independent if and only if its Wronskian $W(t)$ is nonzero at a point of the interval I (equivalently, on the entire interval I).*

Proof. As we know, for any $t_0 \in \mathbb{R}$, the linear map Γ_{t_0} is a linear isomorphism from the space of solutions of (18.4.2) to $\mathbb{R}^n$, so the solutions $\boldsymbol{x}_1(t),\dots,\boldsymbol{x}_n(t)$ are linearly independent iff the vectors $\boldsymbol{x}_1(t_0),\dots,,\boldsymbol{x}_n(t_0)$ are linearly independent in $\mathbb{R}^n$, i.e., iff the matrix $\boldsymbol{X}(t_0)$ is nonsingular. □

Theorem 18.4.6 (Liouville)**.** *Let $W(t)$ be the Wronskian of a collection of n solutions of the system (18.4.2). Then we have the equality*

$$W(t) = W(t_0)\exp\left(\int_{t_0}^t \operatorname{tr} A(s)ds\right),\ \ \forall t_0,t \in I, \tag{18.4.8}$$

where $\operatorname{tr} A(t)$ *denotes the trace of the matrix* $A(t)$,

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

Proof. Without loss of generality we can assume that $W(t)$ is the Wronskian of a linearly independent collection of solutions $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. (Otherwise, the equality (18.4.8) would follow trivially from Theorem 18.4.5.) Denote by $\boldsymbol{X}(t)$ the fundamental matrix with columns $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_n(t)$.

From the definition of the derivative we deduce that for any $t \in I$ we have

$$\boldsymbol{X}(t+\varepsilon) = \boldsymbol{X}(t) + \varepsilon \boldsymbol{X}'(t) + o(\varepsilon), \ \text{ as } \varepsilon \to 0,$$

where $o(\varepsilon)$ denotes a quantity such that

$$\lim_{\varepsilon \searrow 0} \frac{o(\varepsilon)}{\varepsilon} = 0.$$

From (18.4.5) we deduce that

$$\boldsymbol{X}(t+\varepsilon) = X(t) + \varepsilon A(t)\boldsymbol{X}(t) + o(\varepsilon) = \big(\mathbb{1} + \varepsilon A(t) + o(\varepsilon)\big) X(t), \ \ \forall t \in I. \tag{18.4.9}$$

In (18.4.9) we take the determinant of both sides and we deduce

$$W(t+\varepsilon) = \det\big(\mathbb{1} + \varepsilon A(t) + o(\varepsilon)\big) W(t).$$

At this point we recall a classical linear algebra fact. Namely, if B is a $n \times n$ matrix then $\det(\mathbb{1} + sB)$ is a polynomial in s of degree $\leqslant n$ of the form

$$\det(\mathbb{1} + sB) = 1 + s(\operatorname{tr} B) + \cdots + s^n \det B.$$

If we let $B = A(t) + \frac{o(\varepsilon)}{\varepsilon}$, then we deduce

$$\det(1 + \varepsilon A(t) + \varepsilon) = \det(1 + \varepsilon B) = 1 + \varepsilon \operatorname{tr} A(t) + o(\varepsilon).$$

Hence

$$W(t+\varepsilon) = \big(1 + \varepsilon \operatorname{tr} A(t) + o(\varepsilon)\big) W(t),$$

so that

$$\frac{W(t+\varepsilon) - W(t)}{\varepsilon} = \left(\operatorname{tr} A(t) + \frac{o(\varepsilon)}{\varepsilon}\right) W(t).$$

Letting $\varepsilon \to 0$ in the above equality we obtain

$$W'(t) = \big(\operatorname{tr} A(t)\big) W(t).$$

Integrating the above linear differential equation we obtain (18.4.8).

□

Remark 18.4.7. From Liouville's theorem (1809-1882) we deduce in particular the fact that if the Wronskian is nonzero at a point, then it is nonzero everywhere.

Taking into account that the determinant of a matrix is the oriented volume of the parallelepiped determined by its columns, Liouville's formula (18.4.8) describes the variation of the volume of the parallelepiped determined by $\{\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_n(t)\}$. In particular, if $\operatorname{tr} A(t) = 0$, then this volume is conserved along the trajectories of the system (18.4.2).

This fact admits a generalization to nonlinear differential systems of the form

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}), \ \ t \in I, \tag{18.4.10}$$

where $\boldsymbol{f} : I \times \mathbb{R}^n \to \mathbb{R}^n$ is a C^1-map such that

$$\operatorname{div}_{\boldsymbol{x}} \boldsymbol{f}(t, \boldsymbol{x}) \equiv 0. \tag{18.4.11}$$

Assume that for any $\boldsymbol{x}_0 \in \mathbb{R}^n$ there exists a solution $S(t; t_0 \boldsymbol{x}_0) = \boldsymbol{x}(t; t_0, \boldsymbol{x}_0)$ of the system (18.4.10) satisfying the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and defined on the interval I. Let D be a domain in $\mathbb{R}^n$ and set

$$D(t) = S(t)D := \big\{ S(t; t_0, \boldsymbol{x}_0);\;\; \boldsymbol{x}_0 \in D \big\}.$$

Liouville's theorem from statistical physics states that *the volume of $D(t)$ is constant.* □

18.4.3. Nonhomogeneous systems of linear differential equations. In this section we will investigate the nonhomogeneous system (18.4.1) or, equivalently, the system (18.4.3a). Our first result concerns the structure of the set of solutions.

Theorem 18.4.8. *Let $\boldsymbol{X}(t)$ be a fundamental solution of the homogeneous system (18.4.2) and $\widetilde{\boldsymbol{x}}(t)$ a given solution of the nonhomogeneous system (18.4.3a). Then the general solution of the system (18.4.3a) has the form*

$$\boldsymbol{x}(t) = \boldsymbol{X}(t)\boldsymbol{c} + \widetilde{\boldsymbol{x}}(t), \;\; t \in I, \tag{18.4.12}$$

where $\boldsymbol{c}$ is an arbitrary vector in $\mathbb{R}^n$.

Proof. Obviously, any function $\boldsymbol{x}(t)$ of the form (18.4.12) is a solution of (18.4.3a). Conversely, let $\boldsymbol{y}(t)$ be an arbitrary solution of the system (18.4.3a) determined by its initial condition $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, where $t_0 \in I$ and $\boldsymbol{y}_0 \in \mathbb{R}^n$. Consider the linear algebraic system

$$\boldsymbol{X}(t_0)\boldsymbol{c} = \boldsymbol{y}_0 - \widetilde{\boldsymbol{x}}(t_0).$$

Since $\det \boldsymbol{X}(t_0) \neq 0$, the above system has a unique solution $\boldsymbol{c}_0$. Then the function $\boldsymbol{X}(t_0)\boldsymbol{c}_0 + \widetilde{\boldsymbol{x}}(t)$ is a solution of (18.4.3a) and has the value $\boldsymbol{y}_0$ at t_0. The existence and uniqueness theorem then implies that

$$\boldsymbol{y}(t) = \boldsymbol{X}(t_0)\boldsymbol{c}_0 + \widetilde{\boldsymbol{x}}(t). \;\; \forall t \in I.$$

In other words, the arbitrary solution $\boldsymbol{y}(t)$ has the form (18.4.12). □

The next result, sometimes referred to as *Duhamel's formula*, clarifies the statement of Theorem 18.4.8 by offering an explicit representation for a particular solution of (18.4.3a).

Theorem 18.4.9 (Duhamel formula)**.** *Let $S_A(t,s)$ be the propagator of the homogeneous system (18.4.2). Then the general solution of the nonhomogeneous system (18.4.3a) admits the integral representation*

$$\boldsymbol{x}(t) = S_A(t, t_0)\boldsymbol{c} + \int_{t_0}^{t} S_A(t,s)\boldsymbol{b}(s)ds, \;\; t \in I, \tag{18.4.13}$$

where $t_0 \in I$, $\boldsymbol{c} \in \mathbb{R}^n$.

Proof. Set $\boldsymbol{X}(t) := S_A(t, t_0)$. Then $\boldsymbol{X}(t)$ is a fundamental solution of the homogeneous system. We seek a particular solution $\widetilde{\boldsymbol{x}}(t)$ of (18.4.3a) of the form

$$\widetilde{\boldsymbol{x}}(t) = \boldsymbol{X}(t)\boldsymbol{\gamma}(t), \;\; t \in I, \tag{18.4.14}$$

where $\boldsymbol{\gamma} : I \to \mathbb{R}^n$ is a function to be determined. Since $\widetilde{\boldsymbol{x}}(t)$ is supposed to be a solution of (18.4.3a) we deduce

$$\boldsymbol{X}'(t)\boldsymbol{\gamma}(t) + \boldsymbol{X}(t)\boldsymbol{\gamma}'(t) = A(t)\boldsymbol{X}(t) + \boldsymbol{b}(t).$$

Using the equality (18.4.5), we have $\boldsymbol{X}'(t) = A(t)\boldsymbol{X}(t)$ and we deduce that

$$\boldsymbol{\gamma}'(t) = \boldsymbol{X}(t)^{-1}\boldsymbol{b}(t), \;\; \forall t \in I,$$

and thus we can choose $\boldsymbol{\gamma}(t)$ of the form

$$\boldsymbol{\gamma}(t) = \int_{t_0}^{t} \boldsymbol{X}(s)^{-1}\boldsymbol{b}(s)ds, \;\; t \in I, \tag{18.4.15}$$

where $t_0 \in I$ is some fixed point in I. Hence

$$\tilde{\boldsymbol{x}}(t) = \boldsymbol{X}(t)\left(\int_{t_0}^{t} \boldsymbol{X}(s)^{-1}\boldsymbol{b}(s)ds\right) = \int_{t_0}^{t} \underbrace{\boldsymbol{X}(t)\boldsymbol{X}(s)^{-1}}_{S_A(t,s)}\boldsymbol{b}(s)ds$$

is a solution of (18.4.3a). The representation formula (18.4.13) now follows from Theorem 18.4.8. □

Remark 18.4.10. The solution of the initial value problem

$$\boldsymbol{x}' = A(t)\boldsymbol{x}(t) + \boldsymbol{b}(t), \;\; \boldsymbol{x}(t_0) = \boldsymbol{x}^0$$

is

$$\boxed{\boldsymbol{x}(t) = S_A(t, t_0)\boldsymbol{x}^0 + \int_{t_0}^{t} S_A(t, s)\boldsymbol{b}(s)ds}\,.$$

□

18.4.4. Higher order linear differential equations. Consider the linear homogeneous differential equation of order n

$$x^{(n)} + a_1(t)x^{(n-1)}(t) + \cdots + a_n(t)x(t) = 0, \;\; t \in I, \tag{18.4.16}$$

and the associated nonhomogeneous one

$$x^{(n)} + a_1(t)x^{(n-1)}(t) + \cdots + a_n(t)x(t) = f(t), \;\; t \in I, \tag{18.4.17}$$

where a_i, $i = 1, \ldots, n$ and f are continuous functions on an interval I.

Using the general procedure of reducing a higher order o.d.e. to a system of first order o.d.e.-s we set

$$x_1 := x, \;\; x_2 := x', \;\ldots,\; x_n := x^{(n-1)}.$$

The homogeneous equation (18.4.16) is equivalent with the first order linear differential system

$$\begin{array}{ccl} x_1' & = & x_2 \\ x_2' & = & x_3 \\ \vdots & \vdots & \vdots \\ x_n' & = & -a_n x_1 - a_{n-1}x_2 - \cdots - a_1 x_n. \end{array} \tag{18.4.18}$$

In other words, the map Λ defined by

$$x \mapsto \Lambda x := \begin{bmatrix} x \\ x' \\ \vdots \\ x^{(n-1)} \end{bmatrix},$$

defines a linear isomorphism between the set of solutions of (18.4.16) and the set of solutions of the linear system (18.4.18). From Theorem 18.4.2 we deduce the following result.

Theorem 18.4.11. *The set of solutions of (18.4.16) is a real vector space of dimension n.* □

Let us fix a basis $\{x_1, \dots, x_n\}$ of the space of solutions of (18.4.16).

Corollary 18.4.12. *The general solution of (18.4.16) has the form*

$$x(t) = c_1 x_1(t) + \cdots + c_n x_n(t), \tag{18.4.19}$$

where $c_1, \dots, c_n$ are arbitrary constants. □

Just like in the case of linear differential systems, a collection of n linearly independent solutions of (18.4.16) is called a *fundamental* system (or collection) of solutions.

Using the isomorphism Λ we can define the concept of Wronskian of a collection of n solutions of (18.4.16). If $\{x_1, \dots, x_n\}$ is such a collection, then its *Wronskian* is the function $W : I \to \mathbb{R}$ defined by

$$W(t) := \det \begin{bmatrix} x_1 & \cdots & \cdots & x_n \\ x_1' & \cdots & \cdots & x_n' \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(n-1)} & \cdots & \cdots & x_n^{(n-1)} \end{bmatrix}. \tag{18.4.20}$$

Theorem 18.4.5 has the following immediate consequence.

Theorem 18.4.13. *The collection of solutions $\{x_1, \dots, x_n\}$ of (18.4.16) is fundamental if and only if its Wronskian is nonzero at a point or, equivalently, everywhere on I.* □

Taking into account the special form of the matrix $A(t)$ corresponding to the system (18.4.18) we have the following consequence of Liouville's theorem

Theorem 18.4.14. *For any $t_0, t \in I$ we have*

$$W(t) = W(t_0)\exp\left(-\int_{t_0}^{t} a_1(s)ds\right), \tag{18.4.21}$$

where $W(t)$ is the Wronskian of a collection of solutions. □

Theorem 18.4.8 shows that the general solution of the nonhomogeneous equation (18.4.17) has the form

$$x(t) = c_1x_1(t) + \cdots + c_nx_n(t) + \widetilde{x}(t), \tag{18.4.22}$$

where $\{x_1, \ldots, x_n\}$ is a fundamental collection of solutions of the homogeneous equation (18.4.16) , and $\widetilde{x}(t)$ is a particular solution of the nonhomogeneous equation (18.4.17).

We seek the particular solution using the method of *variation of constants*. In other words, we seek $\widetilde{x}(t)$ of the form

$$\widetilde{x}(t) = c_1(t)x_1(t) + \cdots + c_n(t)x_n(t), \tag{18.4.23}$$

where $\{x_1, \ldots, x_n\}$ is a fundamental collection of solutions of the homogeneous equation (18.4.16), and $c_1, \ldots, c_n$ are unknown functions determined from the system

$$\begin{array}{ccc} c_1'x_1 + \cdots + c_n'x_n & = & 0 \\ c_1'x_1' + \cdots + c_n'x_n' & = & 0 \\ \vdots & \vdots & \vdots \\ c_1'x_1^{(n-1)} + \cdots + c_n'x_n^{(n-1)} & = & f(t). \end{array} \tag{18.4.24}$$

The determinant of the above system is the Wronskian of the collection $\{x_1, \ldots, x_n\}$ and it is nonzero since this is a fundamental collection. Thus the above system has a unique solution $\{c_1', \ldots, c_n'\}$. It is now easy to verify that the function $\widetilde{x}$ given by (18.4.23) and (18.4.24) is indeed a solution of (18.4.17).

18.4.5. Higher order linear differential equations with constant coefficients. In this subsection we will deal with the problem of finding a fundamental collection of solutions for the differential equation

$$x^{(n)} + a_1x^{(n-1)} + \cdots + a_{n-1}x' + a_nx = 0, \tag{18.4.25}$$

where $a_1, \ldots, a_n$ are *real* constants. The *characteristic polynomial* of the differential equation (18.4.25) is the algebraic polynomial

$$L(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n. \tag{18.4.26}$$

To any polynomial of degree $\leqslant n$

$$P(\lambda) = \sum_{k=0}^{n} p_k\lambda^k$$

we associate the differential operator

$$P(D)=\sum_{k=0}^{n}p_kD^k,\ \ D^k:=\frac{d^k}{dt^k}. \tag{18.4.27}$$

This acts on the space $C^n(\mathbb{R})$ of functions n-times differentiable on $\mathbb{R}$ with continuous n-th order derivatives according to the rule

$$x\mapsto P(D)x:=\sum_{k=0}^{n}p_kD^kx=\sum_{k=0}^{n}p_k\frac{d^kx}{dt^k}. \tag{18.4.28}$$

Note that (18.4.25) can be rewritten in the compact form

$$L(D)x=0.$$

The key fact for the problem at hand is the following equality

$$L(D)e^{\lambda t}=L(\lambda)e^{\lambda t},\ \ \forall t\in\mathbb{R},\ \ \lambda\in\mathbb{C}. \tag{18.4.29}$$

Indeed, the equality (18.4.29) holds if $L(\lambda)=\lambda^k$ and thus, by linearity, it holds for any polynomial $L\in\mathbb{C}[\lambda]$.

From the equality (18.4.29) it follows that if λ is a root of the characteristic polynomial, then $e^{\lambda t}$ is a solution of (18.4.25). If λ is a root of multiplicity $m(\lambda)$ of $L(\lambda)$, then we define

$$S_\lambda^{\mathbb{C}}:=\big\{\,e^{\lambda t},\dots t^{m(\lambda)-1}e^{\lambda t}\,\big\}.$$

If $L(\lambda)$ has *real* coefficients, then $L(\lambda)=0\Longleftrightarrow L(\bar{\lambda})=0$, where $\bar{\lambda}$ denotes the complex conjugate of λ. In this case we define

$$S_\lambda^{\mathbb{R}}:=\begin{cases}\big\{\,e^{\lambda t},\dots t^{m(\lambda)-1}e^{\lambda t}\,\big\}, & \text{if }\ \lambda\in\mathbb{R},\\ \\ \big\{\,\mathbf{Re}\,e^{\lambda t},\ \mathbf{Im}\,e^{\lambda t},\dots t^{m(\lambda)-1}\,\mathbf{Re}\,e^{\lambda t},\ t^{m(\lambda)-1}\,\mathbf{Im}\,e^{\lambda t}\,\big\}, & \text{if }\ \lambda\in\mathbb{C}\backslash\mathbb{R},\end{cases}$$

where **Re** and respectively **Im** denote the *real* and respectively *imaginary* part of a complex number. Note that since the coefficients $a_1,\dots,a_n$ are real we have

$$S_\lambda=S_{\bar{\lambda}},$$

for any root λ of L. Moreover, if $\lambda=a+\boldsymbol{i}b$, $b\neq 0$, is a root with multiplicity $m(\lambda)$, then

$$S_\lambda=\big\{e^a\cos bt,e^a\sin bt,\dots,t^{m(\lambda)-1}e^a\cos bt,t^{m(\lambda)-1}e^a\sin bt\,\big\}.$$

Theorem 18.4.15. *Let*

$$\mathcal{R}_L=\big\{\lambda\in\mathbb{C};\ \ \mathbf{Im}\,\lambda\geqslant 0,\ \ P(\lambda)=0\,\big\}.$$

For each $\lambda\in\mathcal{R}_L$ we denote by $m(\lambda)$ its multiplicity. Then the collection

$$S^{\mathbb{R}}:=\bigcup_{\lambda\in\mathcal{R}_L}S_\lambda^{\mathbb{R}} \tag{18.4.30}$$

is a fundamental collection of solutions of the equation (18.4.25).

Proof. The proof relies on the following generalization of the product formula.

Lemma 18.4.16. *For any $x, y \in C^n(\mathbb{R})$ and any polynomial $L \in \mathbb{C}[\lambda]$ of degree at most n we have*

$$L(D)(xy) = \sum_{\ell=0}^{n} \frac{1}{\ell!}\big(L^{(\ell)}(D)x\big)\big(D^\ell y\big), \tag{18.4.31}$$

where $L^{(\ell)}(D)$ is the differential operator associated to the polynomial $L^{(\ell)}(\lambda) := \frac{d^\ell}{d\lambda^\ell}L(\lambda)$.

Proof. We present two proofs.

1st Method. Denote by $\mathbb{C}[\lambda]_n$ the subspace of $\mathbb{C}[\lambda]$ consisting of polynomials of degree $\leqslant n$ and by $\mathcal{L}$ the subset of $\mathbb{C}[\lambda]_n$ consisting of polynomials L satisfying (18.4.31) for any $x, y \in C^n(\mathbb{R})$. Observe first that $\mathcal{L}$ is a vector subspace of $\mathbb{C}[\lambda]_n$ containing the constant one. For $L(\lambda) = \lambda^k$, $k = 1, \dots, n$ we have $L(D) = D^k$ and we have

$$L^{(\ell)}(\lambda) = k(k-1)\cdots(k-\ell+1)\lambda^{k-\ell},$$

$$L(D)(xy) = D^k(xy) \overset{(7.6.1)}{=} \sum_{\ell=0}^{k}\binom{k}{\ell}\big(D^{k-\ell}x\big)\big(D^\ell y\big)$$

$$= \sum_{\ell=0}^{k}\frac{k(k-1)\cdots(k-\ell+1)}{\ell!}\big(D^{k-\ell}x\big)\big(D^\ell y\big) = \sum_{\ell=0}^{n}\frac{1}{\ell!}\big(L^{(\ell)}(D)x\big)\big(D^\ell y\big).$$

Hence $1, \lambda, \dots, \lambda^n \in \mathcal{L}$ proving that $\mathcal{L} = \mathbb{C}[\lambda]_n$.

2nd Method. Using the product formula we deduce that $L(D)(xy)$ has the form

$$L(D)(xy) = \sum_{\ell=0}^{n}\big(L_\ell(D)x\big)\big(D^\ell y\big), \tag{18.4.32}$$

where $L_\ell(\lambda)$ are certain polynomials of degree $\leqslant n-\ell$. In (18.4.32) we let $x = e^{\lambda t}$, $y = e^{\mu t}$ where λ, μ are arbitrary complex numbers. Then

$$L_\ell(D)x = L_\ell(\lambda)e^{\lambda t},\;\; D^\ell y = \mu^\ell e^{\mu t},\;\; \big(L_\ell(D)x\big)\big(D^\ell y\big) = e^{(\lambda+\mu)t}L_\ell(\lambda)\mu^\ell.$$

From (18.4.29) and (18.4.32) we obtain the equality

$$L(\lambda+\mu) = e^{-(\lambda+\mu)t}L(D)e^{(\lambda+\mu)t} = \sum_{\ell=0}^{n}L_\ell(\lambda)\mu^\ell. \tag{18.4.33}$$

On the other hand, Taylor's expansion for L at λ implies

$$L(\lambda+\mu) = \sum_{\ell=0}^{n}\frac{1}{\ell!}L^{(\ell)}(\lambda)\mu^\ell,\;\; \forall \lambda, mu \in \mathbb{C}.$$

Comparing the last equality with (18.4.33) we deduce $L_\ell(\lambda) = \frac{1}{\ell!}L^{(\ell)}(\lambda)$. □

Let us now prove that any function in the collection (18.4.30) is indeed a solution of (18.4.25). Let

$$x(t) = t^r e^{\lambda t},\;\; \lambda \in \mathcal{R}_L,\;\; 0 \leqslant r < m(\lambda).$$

Lemma 18.4.16 implies that

$$L(D)x = \sum_{\ell=0}^{n}\frac{1}{\ell!}\big(L^{(\ell)}(D)e^{\lambda t}\big)\big(D^\ell t^r\big)$$

$$= \sum_{\ell=0}^{r} \frac{1}{\ell!}\big(\underbrace{L^{(\ell)}(\lambda)e^{\lambda t}}_{=0}\big)\big(D^{\ell}t^{r}\big) + \sum_{\ell=r+1}^{n}\frac{1}{\ell!}\big(L^{(\ell)}(D)e^{\lambda t}\big)\big(\underbrace{D^{\ell}t^{r}}_{=0}\big) = 0.$$

If λ is a complex number, then the above equality also implies $L(D)\,\mathbf{Re}\,x = L(D)\,\mathbf{Im}\,x = 0$.

Since the complex roots of L come in conjugate pairs, we conclude that the set $S^{\mathbb{R}}$ in (18.4.30) consists of exactly n real solutions of (18.4.25). To prove the theorem it suffices to show that the functions in $S^{\mathbb{R}}$ are linearly independent <u>over $\mathbb{R}$.</u>

Equivalently, it suffices the show that the collection of functions

$$S^{\mathbb{C}} := \bigcup_{\lambda\in\mathcal{R}_L} S^{\mathbb{C}}_{\lambda}$$

is linearly independent <u>over $\mathbb{C}$</u> since the roots of L come in conjugate pairs and

$$\mathbf{Re}\,t^{r}e^{\lambda} = \frac{t^{r}}{2}\big(e^{\lambda t} + e^{\bar{\lambda}t}\big),\ \ \mathbf{Im}\,t^{r}e^{\lambda} = \frac{t^{r}}{2\boldsymbol{i}}\big(e^{\lambda t} - e^{\bar{\lambda}t}\big),$$

We argue by contradiction and we assume that they are linearly dependent over $\mathbb{C}$. We deduce there exists a collection of complex polynomials $P_{\lambda}(t)$, $\lambda\in\mathcal{R}_L$, *not all trivial*, such that and

$$\sum_{\lambda\in\mathcal{R}_L} P_{\lambda}(t)e^{\lambda t} = 0.$$

The following elementary result shows that such polynomials do not exist.

Lemma 18.4.17. *Suppose that $\mu_1,\dots,\mu_k$ are pairwise distinct complex numbers. If*

$$P_1(t),\dots,P_k(t)$$

are complex polynomials such that

$$P_1(t)e^{\mu_1 t} + \cdots P_k(t)e^{\mu_k t} = 0,\ \ \forall t\in\mathbb{R},$$

then $P_1(t)\equiv\cdots\equiv P_k(t)\equiv 0$.

Proof. We argue by induction on k. The result is obviously true for $k=1$. Assume that the result is true and we prove that it is true for $k+1$. Suppose that $\mu_0,\mu_1,\dots,\mu_k$ are pairwise distinct complex numbers and $P_0(t),P_1(t),\dots,P_k(t)$ are complex polynomials such that

$$P_0(t)e^{\mu_0 t} + P_1(t)e^{\mu_1 t} + \cdots P_k(t)e^{\mu_k t} = 0,\ \ \forall t\in\mathbb{R}. \tag{18.4.34}$$

Set

$$m := \max\big\{\deg P_0, \deg P_1,\dots,\deg P_k\big\}. \tag{18.4.35}$$

For simplicity we assume that $\deg P_0 = m$. Dividing both sides of (18.4.34) by $e^{\mu_0 t}$ we deduce

$$P_0(t) + \sum_{j=1}^{k} P_k(t)e^{z_j t} = 0,\ \ \forall t\in\mathbb{R},$$

where $z_j := \mu_j - \mu_0 \neq 0$, $\forall j = 1, \dots, m$. We derivate the above equality $(m+1)$-times. We have $P_0^{(m+1)}(t) = 0$ and, using Lemma 18.4.16 with $L(D) = D^{m+1}$ we deduce that for any $t \in \mathbb{R}$ we have

$$0 = \sum_{j=1}^{k} \underbrace{\left(\sum_{\ell=0}^{m+1} \binom{m+1}{\ell} z_j^{\ell} D^{m+1-\ell} P_j(t) \right)}_{=:Q_j(t)} e^{z_j t}.$$

The induction assumption implies that $\forall j = 1, \dots, k$

$$Q_j(t) = z_j^{m+1} P_j(t) + \binom{m}{1} z_j^m DP_j(t) + \binom{m+1}{2} z_j^{m-1} D^2 P_j(t) + \cdots = 0. \tag{18.4.36}$$

We claim that this implies that $P_j = 0$.

To see this assume, that $P_j \neq 0$ and set $r = \deg P_j = \deg Q_j$. Then $D^{r+1} P_j = 0$ and $D^r P_j \neq 0$. Using (18.4.36) we deduce

$$0 = D^r Q_j(t) = z_j^{m+1} D^r P_j(t) + \underbrace{\binom{m}{1} z_j^m \underbrace{D^{r+1} P_j(t)}_{=0} + \binom{m+1}{2} z_j^{m-1} \underbrace{D^{r+2} P_j(t)}_{=0} + \cdots}_{=0}.$$

Since $z_j \neq 0$ we deduce $D^r P_j = 0$. This contradiction proves the lemma. □

This completes the proof of Theorem 18.4.15. □

Let us briefly discuss the nonhomogeneous equation associated to (18.4.25), i.e., the equation

$$x^{(n)} + a_1 x^{(n-1)} + \cdots + a_n x = f(t), \quad t \in I. \tag{18.4.37}$$

We have seen that the knowledge of a fundamental collection of solutions of the homogeneous equations allows us to determine a solution of the non homogeneous equation by using the method of variation of constants. When the equation has constant coefficients and $f(t)$ has the special form detailed below, this process simplifies somewhat.

A complex valued function $f : I \to \mathbb{C}$ is called a *quasipolynomial* if it is a linear combination with complex coefficients of functions of the form $t^k e^{\mu t}$, where $k \in \mathbb{Z}_{\geqslant 0}$, $\mu \in \mathbb{C}$. A real valued function $f : I \to \mathbb{R}$ is called a *quasipolynomial* if it is the real part of a complex polynomial. For example, the functions $t^k e^{at} \cos bt$ and $t^k e^{at} \sin bt$ are real quasipolynomials.

We want to explain how to find a complex valued solution $x(t)$ of the differential equation

$$L(D)x = f(t)$$

where $f(t)$ is a complex quasipolynomial. Since $L(D)$ has *real* coefficients, and $x(t)$ is a solution of the above equation, then

$$L(D) \, \mathbf{Re}\, x = \mathbf{Re}\, f(t).$$

By linearity, we can reduce the problem to the special situation when

$$f(t) = P(t)e^{\gamma t}, \tag{18.4.38}$$

where $P(t)$ is a complex polynomial and $\gamma \in \mathbb{C}$.

Suppose that γ is a root of order ℓ of the characteristic polynomial $L(\lambda)$. (When $\ell = 0$ this means that $L(\gamma) \neq 0$.) We seek a solution of the form

$$x(t) = t^\ell Q(t)e^{\gamma t}, \tag{18.4.39}$$

where Q is a complex polynomial of degree $\leqslant \deg P$ to be determined. Using Lemma 18.4.16 we deduce from the equality $L(D)x = f(t)$ that

$$P(t) = \sum_{k=0}^{n} \frac{1}{k!} L^{(k)}(\gamma) D^k\big(t^\ell Q(t)\big) = \sum_{k=\ell}^{n} \frac{1}{k!} L^{(k)}(\gamma) D^k\big(t^\ell Q(t)\big). \tag{18.4.40}$$

The last equality leads to an upper triangular linear system in the coefficients of $Q(t)$ which can then be determined in terms of the coefficients of $P(t)$.

We will illustrate the above general considerations on some physical models described by a second order linear differential equation.

18.4.6. The harmonic oscillator. Consider the equation of the harmonic oscillator in the presence of friction

$$mx''(t) + \beta x'(t) + kx = f(t), \quad t \in \mathbb{R}, \tag{18.4.41}$$

where m, β, k are *positive* constants. Think of a bead of mass m allowed to slide along a linear rod embedded in a fluid and attached to an elastic spring with elasticity constant k. Its location along the rod at time t is $x(t)$. The constant β measures the resistance to motion due to the fluid. The function $F(t)$ is an external force acting on the moving bead. The Second Law of Dynamics then implies

$$mx''(t) = -\beta x'(t) - kx + F(t)$$

which is precisely (18.4.41). Dividing (18.4.41) by m and setting $b := \frac{\beta}{m}$, $\omega^2 := \frac{k}{m}$, we obtain the equation

$$x''(t) + bx'(t) + \omega^2 x(t) = f(t) := \frac{1}{m}F(t). \tag{18.4.42}$$

The associated characteristic equation

$$\lambda^2 + b\lambda + \omega^2 = 0,$$

has roots

$$\lambda_{1,2} = -\frac{b}{2} \pm \sqrt{\left(\frac{b}{2}\right)^2 - \omega^2}.$$

We distinguish several cases.

1. $b^2 - 4\omega^2 > 0$. This corresponds to the case where the friction coefficient b is "large", λ_1 and λ_2 are *negative* real numbers, and the general solution of (18.4.41) has the form

$$x(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + \widetilde{x}(t),$$

C_1 and C_2 are arbitrary constants, and $\widetilde{x}(t)$ is a particular solution of the nonhomogeneous equation.

The function $\widetilde{x}(t)$ is called a "*forced solution*" of the equation. Since λ_1 and λ_2 are negative, in the absence of the external force f, the motion dies down fast, converging exponentially to 0.

2. $b^2 - 4\omega^2 = 0$. In this case

$$\lambda_1 = \lambda_2 = -\frac{b}{2},$$

and the general solution of the equation (18.4.41) has the form

$$x(t) = C_1 e^{-\frac{bt}{2m}} + C_2 t e^{-\frac{bt}{2m}} + \widetilde{x}(t).$$

3. $b^2 - 4m\omega^2 < 0$. This is the most interesting case from a physics viewpoint. In this case

$$\lambda_1 = -\frac{b}{2} + \boldsymbol{i}u, \quad \lambda_2 = -\frac{b}{2} - \boldsymbol{i}u$$

where

$$u^2 = -\left(\frac{b}{2}\right)^2 + \omega^2.$$

According to the general theory, the general solution of (18.4.41) has the form

$$\big(\, C_1 \cos ut + C_2 \sin ut \,\big) e^{-\frac{bt}{2m}} + \widetilde{x}(t). \tag{18.4.43}$$

Let us assume that the external force has a harmonic character as well, i.e.,

$$f(t) = a \cos \nu t \ \text{ or } \ f(t) = a \sin \nu t,$$

where the frequency ν and the amplitude a are real, nonzero, constants.

Suppose first that friction is present, i.e., $b \neq 0$. We seek a particular solution of the form

$$\widetilde{x}(t) = a_1 \cos \nu t + a_2 \sin \nu t.$$

Then

$$\begin{aligned} \widetilde{x}''(t) + b\widetilde{x}'(t) + \omega^2 \widetilde{x}(t) &= \big(\, -a_1\nu^2 + b\nu a_2 + \omega^2 a_1 \,\big) \cos \nu t \\ &\quad + \big(\, -\nu^2 a_2 - b\nu a_1 + \omega^2 a_2 \,\big) \sin \nu t \\ &= \big(\, (\omega^2 - \nu^2) a_1 + b\nu a_2 \,\big) \cos \nu t + \big(\, -b\nu a_1 + (\omega^2 - \nu^2) a_2 \,\big) \sin \nu t. \end{aligned}$$

When $f = a \cos \nu t$ we deduce

$$(\omega^2 - \nu^2) a_1 + b\nu a_2 = a, \quad -b\nu a_1 + (\omega^2 - \nu^2) a_2 = 0.$$

Thus, if we write $d(\nu) := (\omega^2 - \nu^2)$, we deduce

$$a_2 = \frac{b\nu}{d(\nu)} a_1, \quad d(\nu) a_1 + \frac{b^2\nu^2}{d(\nu)} a_1 = a$$

so that

$$a_1 = \frac{a d(\nu)}{d(\nu)^2 + b^2\nu^2 d(\nu)}, \quad a_2 = \frac{ab\nu}{d(\nu)^2 + b^2\nu^2 d(\nu)},$$

the function

$$\widetilde{x}(t) = \frac{a(\omega^2 - \nu^2)}{(\omega^2 - \nu^2)^2 + b^2\nu^2}\cos\nu t + \frac{ab\nu}{(\omega^2 - \nu^2)^2 + b^2\nu^2}\sin\nu t \tag{18.4.44}$$

is a solution of (18.4.42). Interestingly, as $t \to \infty$, the general solution (18.4.43) is asymptotic to the particular solution (18.4.44), i.e.,

$$\lim_{t\to\infty}\big(\,x(t) - \widetilde{x}(t)\,\big) = 0,$$

so that for t sufficiently large, the general solution is practically indistinguishable from the particular forced solution $\widetilde{x}$.

Consider now the case when $b = 0$, i.e., the friction is nonexistent. Assume $\nu \neq \omega$, i.e., we are in the *nonresonant situation*. Then

$$\widetilde{x}(t) = \widetilde{x}_\nu(t) = \frac{a}{\omega^2 - \nu^2}\cos\nu t.$$

If the frequency ν of the external perturbation is very close to the characteristic frequency of the oscillatory system, i.e.,

$$\nu \approx \omega. \tag{18.4.45}$$

then the amplitude

$$\left|\frac{a}{\omega^2 - \nu^2}\right|$$

is huge. This is the *resonance* phenomenon encountered often in oscillatory mechanical systems.

Practically, it manifests itself when the friction is negligible and the frequency of the external force is very close to the characteristic frequency of the system. In such cases oscillatory systems perform oscillations with large amplitudes. This can have both beneficial applications (think of tuning in a radio broadcast) and devastating consequences, such as the Tacoma bridge disaster. Note that, $\forall t \in \mathbb{R}$, $\widetilde{x}_\nu(t)$ has no limit as $\nu \to \omega$.

On the other hand

$$\widetilde{y}_\nu(t) = \widetilde{x}_\nu(t) - \frac{a}{\omega^2 - \nu^2}\cos\omega t = \frac{a}{\omega^2 - \nu^2}\big(\cos\nu t - \cos\omega t\,\big),$$

is also the solution of

$$x'' + \omega^2 x(t) = f_\nu(t) := a\cos\nu t, \tag{$\mathbf{E}_\nu$}$$

satisfying the initial conditions $\widetilde{y}_\nu(0) = \widetilde{y}_\nu'(0) = 0$. L'Hôpital's rule shows that

$$\forall t \in \mathbb{R}\quad \lim_{\nu\to\omega}\widetilde{y}_\nu(t) = \lim_{\nu\to\omega} a\frac{\cos\nu t - \cos\omega t}{\omega^2 - \nu^2} = \frac{at}{2\omega}\sin\omega t =: \widetilde{y}_\omega(t).$$

Let us observe that when $b = 0$, $\nu = \omega$, $f = a\cos\omega t$, then the general solution of

$$x'' + \omega^2 x = f(t)$$

has the form

$$a_1\cos\omega t + a_2\sin\omega t + B_1 t\cos\omega t + B_2 t\sin\omega t.$$

where the coefficients B_1, B_2 are uniquely determined from the equality

$$\frac{d^2}{dt^2}\big(B_1t\cos\omega t + B_2t\sin\omega t\big) + \omega^2\big(B_1t\cos\omega t + B_2t\sin\omega t\big) = a\cos\omega t.$$

We have

$$\begin{aligned}\frac{d}{dt}\big(B_1t\cos\omega t + B_2t\sin\omega t\big) &= B_1\cos\omega t - \omega B_1t\sin\omega t + B_2\sin\omega t + \omega B_2t\cos\omega t\\ &= (\omega B_2t + B_1)\cos\omega t + (-\omega B_1t + B_2)\sin\omega t\\ \frac{d^2}{dt^2}\big(B_1t\cos\omega t + B_2t\sin\omega t\big) &= B_2\omega\cos\omega t - \omega(B_2\omega t + B_1)\sin\omega t\\ &\quad -B_1\omega\sin\omega t + \omega(-B_1\omega t + B_2)\cos\omega t\\ &= (-\omega^2B_1 + 2B_2\omega)\cos\omega t + (-\omega^2B_2 - 2B_1\omega)\sin\omega t.\end{aligned}$$

We deduce that

$$\begin{aligned}a\cos\omega t &= \frac{d^2}{dt^2}\big(B_1t\cos\omega t + B_2t\sin\omega t\big) + \omega^2\big(B_1t\cos\omega t + B_2t\sin\omega t\big)\\ &= 2B_2\omega\cos\omega t - 2B_1\omega\sin\omega t,\ \ \forall t.\end{aligned}$$

Hence

$$B_1 = 0,\ \ B_2 = \frac{a}{2\omega}$$

Thus, the general solution of $x'' + \omega^2x = a\cos\omega t$ is

$$x(t) = a_1\cos\omega t + a_2\sin\omega t + \frac{at}{2\omega}\sin\omega t = a_1\cos\omega t + a_2\sin\omega t + \widetilde{y}_\omega(t).$$

In particular, this shows that $\widetilde{y}_\omega(t)$ is a solution of ($\mathbf{E}_\nu$) for $\nu = \omega$.

The oscillations of $x(t)$ are increasingly bigger and bigger as t increases. In Figure 18.4 we have depicted $x(t)$ corresponding to $a_1 = a_2 = \omega = 1$, $a = 2$.

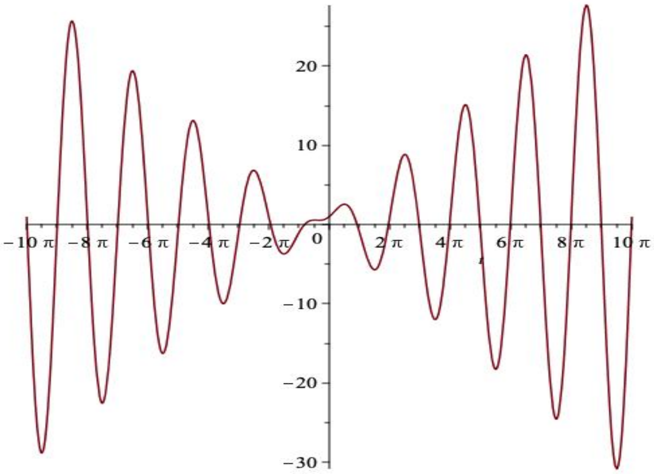

Figure 18.4. *The graph of* $\sin(t) + \cos(t) + t\sin(t)$.

18.4.7. Approximating shocks. For each $n \in \mathbb{N}$, $n \geqslant 2$ define

$$F_n : \mathbb{R} \to \mathbb{R},\ \ F_n(x) = \begin{cases} 0, & |x| > \frac{1}{n}, \\ n^2(x + 1/n), & -\frac{1}{n} \leqslant x \leqslant 0, \\ n - n^2 x, & 0 < x \leqslant \frac{1}{n} \end{cases}$$

In Figure 18.5 We have depicted the graph of $F_4(x)$.

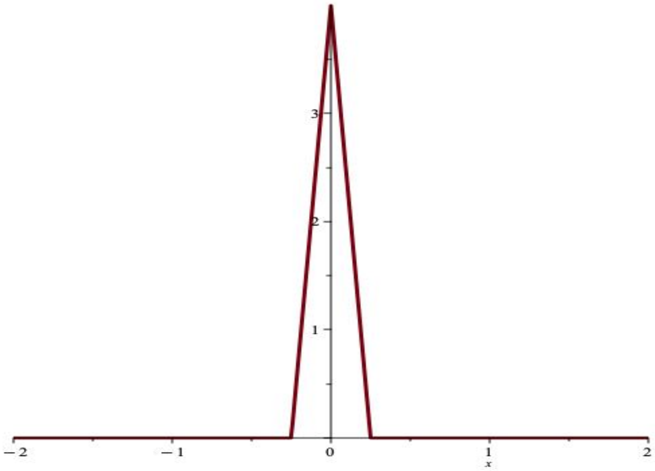

Figure 18.5. *The graph of $F_4(x)$.*

The sequence of functions F_n approximates Dirac's delta function $\delta(x)$ in the sense that

$$\int_{\mathbb{R}} F_n(t)dt = 1,\ \ \forall n,\ \ \lim_{n \to \infty} F_n(t) = 0,\ \ \forall t \neq 0,\ \ \lim_{n \to \infty} F_n(0) = \infty.$$

We want to investigate the behavior as $n \to \infty$ of the solution $x_n(t)$ of the initial value problem

$$\ddot{x}_n(t) = F_n(x),\ \ x_n(0) = -1,\ \ \dot{x}_n(0) = v_0 > 0. \tag{18.4.46}$$

This equation describes the motion on the x-axis of a particle of mass $m = 1$ that starts at $x = -1$ with initial velocity 1. Initially there is no force acting on it by when it approaches the origin it enters a narrow region where the force field acting on it is immense: it suffers a shock near the origin. Set

The particle reaches the location $x = 1/n$ at time $t_n = (1 - \frac{1}{n})/v_0$ because it the region $[-1, 1/n]$ it travels with constant velocity v_0. In the region $[-1/n, 0]$ the laws of motion change to

$$\ddot{x}_n = n^2 x_n + n,\ \ x_n(t_n) = -1/n,\ \ \dot{x}_n(t_n) = v_0.$$

The first equation can be rewritten as

$$\ddot{x}_n - n^2 x_n = n.$$

We seek a solution of the form

$$x_n(t) = A_n e^{n(t-t_n)} + B_n e^{-n(t-t_n)} + C_n$$

The constant C_n is determined from the equality $-n^2 C_n = n$ so $C_n = -\frac{1}{n}$. The constants A_n, B_n are determined from the initial conditions

$$A_n + B_n + C_n = -\frac{1}{n} = x_n(t_n), \;\; nA_n - nB_n = v_0 = \dot{x}_n(t_n).$$

Hence $A_n + B_n = 0$ $A_n - B_n = \frac{1}{nv_0}$ so that

$$A_n = \frac{1}{2nv_0} = -B_n.$$

Thus starting at the moment t_n the position $x_n(t)$ has the form

$$x_n(t) = \frac{1}{2nv_0}\big(e^{n(t-t_n)} - e^{-n(t-t_n)} \big) = \frac{1}{nv_0} \sinh n(t-t_n) - \frac{1}{n}.$$

The particle reaches the origin at time s_n determined by the equation

$$\sinh n(s_n - t_n) = v_0.$$

If we set $r := e^{n(s_n-t_n)}$ we deduce $r - r^{-1} = 2v_0$. or equivalently

$$r^2 - 2v_0 r - 1 = 0, \;\; r = v_0 + \sqrt{v_0^2 + 1}.$$

At the moment s_n the velocity the particle is

$$v_1 := \dot{x}_n(s_n) = \cosh n(s_n - t_n) = \sqrt{1 + \sinh^2 n(s_n - \tau_n)} = \sqrt{1 + v_0^2}.$$

In the region $[0, 1/n]$ the position of the particle satisfies the initial value problem

$$\ddot{x}_n = n - n^2 x_n, \;\; x_n(s_n) = 0, \;\; \dot{x}_n(s_n) = v_1.$$

The first equation above can be rewritten as $\ddot{x}_n + n^2 x_n = n$ and we seek a solution of the form

$$x_n(t) = A_n e^{\boldsymbol{i}n(t-s_n)} + B_n e^{-\boldsymbol{i}n(t-s_n)} + C_n.$$

The constant C_n is obtained from the equality $n^2 C_n = n$ so $C_n = \frac{1}{n}$. From the initial conditions we deduce

$$A_n + B_n + C_n = x_n(s_n) = 0, \;\; n\boldsymbol{i}\big(A_n - B_n \big) = \dot{x}_n(s_n) = v_1,$$

i.e.,

$$A_n + B_n = -C_n = \frac{1}{n}, \;\; A_n + B_n = \frac{v_1}{n\boldsymbol{i}} = -\frac{v_1 \boldsymbol{i}}{n},$$

so that

$$2A_n = -\frac{1 + v_1 \boldsymbol{i}}{n}, \;\; B_n = \bar{A}_n = \frac{1 - v_1 \boldsymbol{i}}{2n}$$

Hence

$$x_n(t) = -\frac{1}{2n}\big((1 + v_1\boldsymbol{i}) e^{\boldsymbol{i}n(t-s_n)} + (1 - v_1\boldsymbol{i}) e^{-\boldsymbol{i}n(t-s_n)} \big) + \frac{1}{n}.$$

Note that

$$1+v_1\boldsymbol{i}=\sqrt{1+v_1^2}\,e^{\boldsymbol{i}\varphi},\ \ \cos\varphi=\frac{1}{\sqrt{1+v_1^2}},\ \ \sin\varphi=\frac{v_1}{\sqrt{1+v_1^2}}$$

so $\varphi=\arccos\big(\,(1+v_0^2)^{-1/2}\,\big)$

$$\begin{aligned}x_n(t)&=-\frac{\sqrt{1+v_1^2}}{2n}\big(\,e^{\boldsymbol{i}n(t-s_n)+\boldsymbol{i}\varphi}+e^{-\boldsymbol{i}n(t-s_n)-\boldsymbol{i}\varphi}\,\big)+\frac{1}{n}\\&=-\frac{\sqrt{1+v_1^2}}{n}\cos\big(\,n(t-s_n)+\varphi\,\big).\end{aligned}$$

The moment $\tau_n>s_n$ when $x(\tau_n)=\frac{1}{n}$ is determined from the equality

$$n(\tau_n-s_n)+\varphi=\pi/2,$$

so

$$\tau_n-s_n=\frac{1}{n}\big(\,\pi/2-\varphi\big),\ \ \varphi=\arccos\big(\,(1+v_0^2)^{-1/2}\,\big).$$

Observe that

$$\dot{x}_n(\tau_n)=\sqrt{1+v_1^2}\sin\big(\,n(\tau_n-s_n)+\varphi\,\big)=\sqrt{1+v_1^2}.$$

For $t\geqslant\tau_n$ we have

$$x(t)=\big(\,1+v_1^2\,\big)^{1/2}(t-\tau_n)+\frac{1}{n}=\big(\,2+v_0^2\,\big)^{1/2}(t-\tau_n)+\frac{1}{n}.$$

The interval $[t_n,s_n]$ is of size $O(1/n)$ and during this interval the speed of the particle increases to $v_1=\sqrt{1+v_0^2}$. The interval $[s_n,\tau_n]$ is also a size $O(1/n)$ and over this time interval the speed increases from v_1 to $v_2=\sqrt{1+v_1^2}$. Thus in the very short time interval $[t_n,\tau_n]$ the speed has increased to $v_2=\sqrt{2+v_0^2}>v_0$. Note that the terminal velocity v_2 *is independent of* n. Outside the interval $[t_n,\tau_n]$ the velocity $\dot{x}_n(t)$ is constant, equal to v_0, if $t\leqslant t_n$, and equal to v_2, if $t\geqslant\tau_n$. For n large this is a rather dramatic change.

We can compute the terminal velocity v_2 by using the conservation of energy principle. Consider the function

$$W_n(x)=\int_{-\infty}^x F_n(s).$$

Note that $W(x)=0$ for $x\leqslant -1/n$ and $W(x)=1$, for $x\geqslant 1/n$. We have $W_n'(x)=F_n(x)$ and from the equality

$$\ddot{x}_n=W_n'(x_n)$$

we deduce

$$0=\ddot{x}_n\dot{x}_n-W_n(x_n)\dot{x}_n=\frac{d}{dt}\Big(\,\frac{1}{2}\dot{x}_n^2-W(x_n)\,\Big).$$

Thus the quantity

$$E(t)=\frac{1}{2}\dot{x}_n(t)^2-W\big(\,x_n(t)\,\big)$$

is independent of time. This is the total energy of the particle that is conserved during motion.

For $t \leqslant 0$, $W_n\big(x_n(t)\big) = 0$ so $E(0) = \frac{1}{2}v_0^2$. For $t \gg 0$ we have $W_n\big(x_n(t)\big) = 1$ and thus for $t \gg 0$ we have

$$\frac{1}{2}\dot{x}_n(t)^2 - 1 = E(0) = \frac{1}{2}v_0^2.$$

Thus, for $t \gg 0$ we have

$$\dot{x}_n(t)^2 = 2 + v_0^2 = v_2^2.$$

Using the Arzelà-Ascoli Theorem 17.4.3 or Exercise 17.42, one can show that as $n \to \infty$ the functions $x_n(t)$ converge uniformly on compacts to

$$x_\infty(t) = \begin{cases} v_0 t - 1, & t \leqslant \frac{1}{v_0}, \\ v_2\big(t - 1/v_0\big), & t > 1/v_0. \end{cases}$$

Let us observe that $x_\infty(1/v_0) = 0$ and $\ddot{x}_\infty(1/v_0) = (v_2 - v_0)\delta(1/v_0)$. The particle under goes a dramatic shock near the moment $1/v_0$.

For $n = 4$, and $v_0 = 1$, we have $t_n = 0.75$, $s_n \approx 0.7970$, $\tau_n \approx 0.9508$. The graph of $x_4(t)$ is depicted in Figure 18.6. The horizontal coordinate is time, and the vertical coordinate indicate the location. The behavior in the interval $[t_4, \tau_4]$ is so dramatic that it cannot be visualized accurately on the time scale $[0, 2]$.

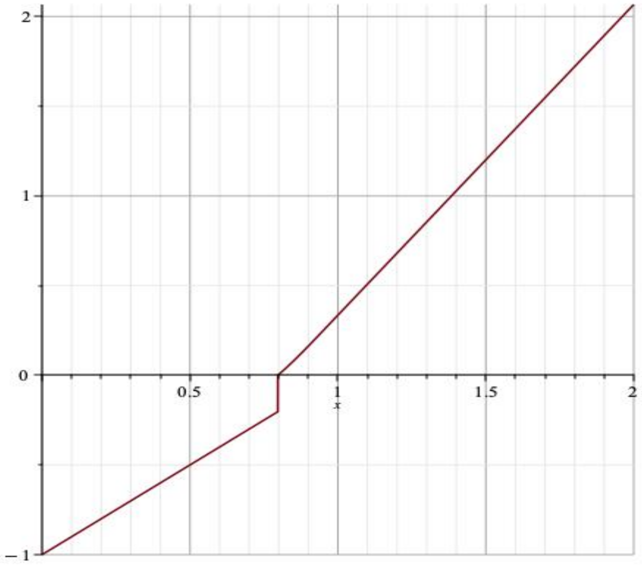

Figure 18.6. *The graph of $x_4(t)$, $0 \leqslant t \leqslant 2$.*

18.4.8. Linear differential systems with constant coefficients. We will investigate the first order differential system

$$\boldsymbol{x}' = A\boldsymbol{x}, \quad t \in \mathbb{R}, \tag{18.4.47}$$

where $A = (a_{ij})_{1\leqslant i,j\leqslant n}$ is a *constant*, real, $n \times n$-matrix. We denote by $S_A(t)$ the fundamental matrix of (18.4.47) uniquely determined by the initial condition

$$S_A(0) = \mathbb{1},$$

where $\mathbb{1}$ denotes the identity matrix. More concretely, for every $\boldsymbol{x}_0 \in \mathbb{R}^n$ the function $\boldsymbol{x}(t) = S_A(t)\boldsymbol{x}_0$ is the solution of the Cauchy problem

$$\boldsymbol{x}'(t) = A\boldsymbol{x}(t), \;\; \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

This property characterizes $S_A(t)$.

Proposition 18.4.18. *The family $\{\, S_A(t);\;\; t \in \mathbb{R}\,\}$ satisfies the following properties.*

(i) $S_A(t+s) = S_A(t)S_A(s)$, $\forall t, s \in \mathbb{R}$.

(ii) $S_A(0) = \mathbb{1}$.

(iii)
$$\lim_{t\to t_0} S_A(t)\boldsymbol{x} = S_A(t_0)\boldsymbol{x}, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n, \;\; t_0 \in \mathbb{R}.$$

Proof. The group property was already established in Proposition 18.2.25. It follows from the uniqueness of solutions of the Cauchy problems associated to (18.4.47): the functions $Z(t) = S_A(t)S_A(s)$ and $Y(t) = SA(t+s)$ both satisfy (18.4.47) with initial condition

$$Y(0) = Z(0) = S_A(s).$$

Property (ii) follows from the definition, while (iii) follows from the fact that the function $t \mapsto S_A(t)\boldsymbol{x}$ is a solution of (18.4.47) and in particular it is continuous. □

Proposition 18.4.18 expresses the fact that the family $\{\, S_A(t);\;\; t \in \mathbb{R}\,\}$ is a *one-parameter group* of linear transformations of the space $\mathbb{R}^n$. The equality (iii), which can be easily seen to hold in the stronger sense of the norm of the spaces of $n \times n$-matrices, expresses the continuity property of the group $S_A(t)$. The map $t \to S_A(t)$ satisfies the differential equation

$$\frac{d}{dt}S_A(t)\boldsymbol{x} = AS_A(t)\boldsymbol{x}, \;\; \forall t \in \mathbb{R}, \;\; \forall \boldsymbol{x} \in \mathbb{R}^n,$$

and thus

$$Ax = \frac{d}{dt}\Big|_{t=0} S_A(t)\boldsymbol{x} = \lim_{t\to 0}\frac{1}{t}\big(\, S_A(t)\boldsymbol{x} - \boldsymbol{x}\,\big). \tag{18.4.48}$$

The equality (18.4.48) expresses the fact that A is the generator of the one-parameter group $S_A(t)$. Note that in this case the propagator $S_A(t,s)$ is

$$S_A(t,s) = S_A(t)S_A(s)^{-1} = e^{(t-s)A},$$

where

$$e^{tA} = \sum_{n\geqslant 0}\frac{t^n}{n!}A^n. \tag{18.4.49}$$

Indeed, as shown in Exercise 17.27, the family of operators $E_A(t) = e^{tA}$ is a solution of the initial value problem

$$\frac{d}{dt}E_A(t) = AE_A(t), \;\; E_A(0) = \mathbb{1}.$$

The family $S_A(t)$ is also a solution of this initial value problem and we deduce $S_A(t) = E_A(t)$, $\forall t$.

Using (18.4.13) we deduce that for any continuous function $f : \mathbb{R} \to \mathbb{R}^n$, the solution of the non-homogeneous initial value problem

$$\boldsymbol{x}'(t) = A\boldsymbol{x}(t) + \boldsymbol{f}(t), \;\; \boldsymbol{x}(t_0) = \boldsymbol{x}_0$$

is given by the Duhamel's formula

$$\boldsymbol{x}(t) = e^{(t-t_0)A}\boldsymbol{x}_0 + \int_{t_0}^{t} e^{(t-s)A}\boldsymbol{f}(s)ds. \tag{18.4.50}$$

We next investigate the structure of the fundamental matrix $S_A(t)$ and the ways we can compute it. We rely on the results described in Exercise 17.27.

We consider a slightly more general case that makes the algebraic manipulations more transparent. Suppose that A is a *complex* $n \times n$ matrix and V denotes the *complex* vector space $\boldsymbol{V} := \mathbb{C}^n$. We denote by $\boldsymbol{V}_{\mathbb{R}}$ its real part

$$\boldsymbol{V}_{\mathbb{R}} := \big\{\, \boldsymbol{z} = (z_1, \dots, z_n) \in \mathbb{C}^n; \;\; \mathbf{Im}\, z_k = 0, \;\; \forall k = 1, \dots, n \,\big\}.$$

Observe that the matrix A has real entries if and only if $A\boldsymbol{V}_{\mathbb{R}} \subset \boldsymbol{V}_{\mathbb{R}}$. (Can you prove this?)

We equip $\boldsymbol{V}$ with the sup-norm

$$\|\boldsymbol{z}\| := \max\big(\,|z_1|, \dots, |z_n|\,\big), \;\; \forall \boldsymbol{z} = (z_1, \dots, z_n) \in \mathbb{C}^n.$$

As such, $\boldsymbol{V}$ is a complex Banach space and $A : \boldsymbol{V} \to \boldsymbol{V}$ is a bounded linear operator. For every $t \in \mathbb{R}$, consider as in Exercise 17.27 the operator e^{tA} defined by the convergent series

$$e^{tA} = \mathbb{1} + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \cdots . \tag{18.4.51}$$

As claimed in Exercise 17.27, for every $\boldsymbol{z}_0 \in V$, the function $\boldsymbol{z}(t) = e^{tA}\boldsymbol{z}_0$ is a solution of the Cauchy problem

$$\frac{d\boldsymbol{z}}{dt} = A\boldsymbol{z}(t), \;\; \boldsymbol{z}(0) = \boldsymbol{z}_0.$$

In other words, $S_A(t) = e^{tA}$ for any complex matrix. If A is real, then all the terms of the series (18.4.51) are real matrices so e^{tA} is also a real matrix. Thus in this case, the exponential of A as a real $n \times n$ matrix coincides with the restriction to $\boldsymbol{V}_{\mathbb{R}}$ of the exponential of A viewed as a complex $n \times n$ matrix. We will identify the bases of $\boldsymbol{V}$ with linear isomorphisms (gauge)

$$G : \mathbb{C}^n \to \boldsymbol{V}.$$

More precisely, if $\underline{\boldsymbol{e}} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$ is the canonical basis of $\mathbb{C}^n$, then $\boldsymbol{g}_1 = G\boldsymbol{e}_1, \dots, \boldsymbol{g}_n = G\boldsymbol{e}_n$ is a basis of $\boldsymbol{V}$. In the basis determined by G, the operator A is represented by the complex $n \times n$ matrix A_G such that the diagram below is commutative

$$\begin{tikzcd} \mathbb{C}^n \arrow[r, "A_G"] \arrow[d, "G"] & \mathbb{C}^n \arrow[d, "G"'] \\ \boldsymbol{V} \arrow[r, "A"'] & \boldsymbol{V} \end{tikzcd}$$

This means that $AG = GA_G$ so that

$$A = GA_GG^{-1}.$$

If $\boldsymbol{B}(\boldsymbol{V})$ denotes the Banach space of bounded linear operators $\boldsymbol{V} \to \boldsymbol{V}$, then the map

$$\boldsymbol{B}(\mathbb{C}^n) \ni T \mapsto GTG^{-1} \in \boldsymbol{B}(\boldsymbol{V})$$

is continuous and we deduce that

$$e^{tA} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{t^k}{k!}A^n = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{t^k}{k!}GA_G^nG^{-1} = G\left(\lim_{n\to\infty} \sum_{k=0}^{n} \frac{t^k}{k!}A_G^n\right)U^{-1} = Ge^{tA_G}G^{-1}.$$

Here is how one should interpret the above equality: if we can explicitly compute G, G^{-1} and e^{tA_G}, then we can explicitly compute e^{tA}. We will show how to use this principle, but first let us describe some simple examples of matrices A for which e^{tA} can be computed explicitly in finite time.

Example 18.4.19. (a) Suppose that A is a diagonal $m \times m$ matrix

$$A = \operatorname{Diag}(a_1, \dots, a_m).$$

Then $e^{tA} = \operatorname{Diag}\big(e^{ta_1}, \dots, e^{ta_m}\big)$.

(b) Suppose that N is a nilpotent $m \times m$ matrix, i.e., $N^{p+1} = 0$ for some $p \in \mathbb{N}_0$. Then

$$e^{tN} = \mathbb{1} + \frac{t}{1!}N + \frac{t^2}{2!}N^2 + \cdots + \frac{t^p}{p!}N^p.$$

(c) Suppose that A, B are complex $m \times m$ matrices and we know how to compute e^{tA} and e^{tB}. If A and B commute, $AB = BA$, then

$$e^{t(A+B)} = e^{tA}e^{tB}$$

so we also know how to compute $e^{t(A+B)}$.

(d) Consider the nilpotent $m \times m$ matrix

$$N = N_m = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

For each $\lambda \in \mathbb{C}$ we obtain a Jordan cell $C_\lambda = \lambda \boldsymbol{I}_m + N$. Clearly $\lambda \boldsymbol{I}_m$ and N commute and we deduce from (b) and (c) that

$$e^{tC_\lambda} = e^{t\lambda}\left(\mathbb{1} + \frac{t}{1!}N + \frac{t^2}{2!}N^2 + \cdots + \frac{t^m}{m!}N^m\right).$$

(e) Suppose thatA_1 and A_2 are square matrices of sizes m_1 and respectively m_2 and we can compute the exponentials e^{tA_k}, $k = 1, 2$. Then we can also compute the exponential of the matrix

$$A_1 \oplus A_2 := \left[\begin{array}{cc} A_1 & 0 \\ 0 & A_2 \end{array}\right].$$

More precisely,

$$e^{t(A_1 \oplus A_2)} = e^{tA_1} \oplus e^{tA_2}.$$

(f) The theory of Jordan decomposition informs us that for any square matrix A there exists a Jordan basis G such that

$$A_G = C_{\lambda_1} \oplus \cdots \oplus C_{\lambda_k}$$

where $\lambda_1, \dots, \lambda_k$ are the eigenvalues of A, and the matrices C_{λ_j} are Jordan cells of various sizes. In other words, for any matrix A we can find a basis G such that e^{tA_G} is computable. Thus, theoretically, e^{tA} is computable for any complex matrix. In practice this is a bridge too far. If the size of A is very large ($\geqslant 5$) finding the eigenvalues accurately is improbable. The best one can hope in such cases is to find reasonable approximations of the eigenvalues. If the matrix admits nontrivial Jordan cells, i.e., it is not diagonalizable, then it is possible to miss this fact using approximations: with probability 1, a small perturbation of a matrix renders it diagonalizable. □

Example 18.4.20. Suppose that the matrix A is diagonalizable. This happens for example when A is Hermitian or has distinct eigenvalues.

Thus $\boldsymbol{V}$ admits a basis/gauge G consisting of eigenvectors of A. More precisely, the columns $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$ of G are eigenvectors of A

$$A\boldsymbol{g}_k = \lambda_k \boldsymbol{g}_k, \;\; k = 1, \dots, n.$$

In this case

$$A_G = \operatorname{Diag}(\lambda_1, \dots, \lambda^n), \;\; e^{tA_G} = \operatorname{Diag}\big(\, e^{t\lambda_1}, \dots, e^{t\lambda_n}\,\big)$$
$$e^{tA} = G \operatorname{Diag}\big(\, e^{t\lambda_1}, \dots, e^{t\lambda_n}\,\big) G^{-1}.$$

Consider for example the 2×2-matrix

$$A = \left[\begin{array}{cc} a & -b \\ b & a \end{array}\right], \;\; a, b \in \mathbb{R}.$$

When $b = 0$ the matrix A is diagonal and the computation of e^{tA} is immediate. We assume $b \neq 0$. The characteristic polynomial of A is

$$P(\lambda) = \lambda^2 - (\operatorname{tr} A)\lambda + \det A = \lambda^2 - 2a\lambda + a^2 + b^2.$$

Its roots are $\lambda = a + b\boldsymbol{i}$ and $\bar{\lambda} = a - b\boldsymbol{i}$. Since $b \neq 0$, we have $\lambda \neq \bar{\lambda}$ so the matrix is diagonalizable.

The vector

$$\boldsymbol{g} = \begin{bmatrix} 1 \\ -\boldsymbol{i} \end{bmatrix}$$

is a complex eigenvector of $\lambda = a + b\boldsymbol{i}$. Indeed,

$$A\boldsymbol{g} = \begin{bmatrix} a + b\boldsymbol{i} \\ b - a\boldsymbol{i} \end{bmatrix} = (a + b\boldsymbol{i})\boldsymbol{g}.$$

The conjugate eigenvector

$$\bar{\boldsymbol{g}} = \begin{bmatrix} 1 \\ \boldsymbol{i} \end{bmatrix}$$

is an eigenvector of $\bar{\lambda} = a - b\boldsymbol{i}$. In this case

$$G = \begin{bmatrix} 1 & 1 \\ -\boldsymbol{i} & \boldsymbol{i} \end{bmatrix}, \;\; G^{-1} = \frac{1}{2i}\begin{bmatrix} \boldsymbol{i} & -1 \\ \boldsymbol{i} & 1 \end{bmatrix}$$

so

$$e^{tA} = \frac{1}{2\boldsymbol{i}}\begin{bmatrix} 1 & 1 \\ -\boldsymbol{i} & \boldsymbol{i} \end{bmatrix} \cdot \begin{bmatrix} e^{t\lambda} & 0 \\ 0 & e^{t\bar{\lambda}} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{i} & -1 \\ \boldsymbol{i} & 1 \end{bmatrix} = \frac{1}{2\boldsymbol{i}}\begin{bmatrix} e^{t\lambda} & e^{t\bar{\lambda}} \\ -\boldsymbol{i}e^{t\lambda} & \boldsymbol{i}e^{t\bar{\lambda}} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{i} & -1 \\ \boldsymbol{i} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}\big(e^{t\lambda} + e^{t\bar{\lambda}}\big) & \frac{1}{2\boldsymbol{i}}\big(e^{t\bar{\lambda}} - e^{t}\lambda\big) \\ \frac{1}{2\boldsymbol{i}}\big(e^{t\lambda} - e^{t\bar{\lambda}}\big) & \frac{1}{2}\big(e^{t\lambda} + e^{t\bar{\lambda}}\big) \end{bmatrix} = \begin{bmatrix} \mathbf{Re}\, e^{t\lambda} & -\mathbf{Im}\, e^{t\lambda} \\ \mathbf{Im}\, e^{t\lambda} & \mathbf{Re}\, e^{t\lambda} \end{bmatrix}$$

$$= \begin{bmatrix} e^{ta}\cos(tb) & -e^{ta}\sin(tb) \\ e^{ta}\sin(tb) & e^{ta}\cos(tb) \end{bmatrix} = e^{ta}\begin{bmatrix} \cos(tb) & -\sin(tb) \\ \sin(tb) & \cos(tb) \end{bmatrix}.$$

Altenatively, we can write $A = a\mathbb{1} + bJ$ where

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Clearly the matrices $A_0 = a\mathbb{1}_2$ and $A_1 = bJ$ commute so

$$e^{tA} = e^{tA_0}e^{tA_1} = e^{ta}e^{tbJ}.$$

Now observe that

$$J^2 = -\mathbb{1}, \;\; J^3 = -J, \;\; J^4 = \mathbb{1}, \;\; J^5 = J, \dots$$

Hence

$$e^{sJ} = \mathbb{1} + \frac{s}{1!}J - \frac{s^2}{2!}\mathbb{1} - \frac{s^3}{3!}J + \frac{s^4}{4!}\mathbb{1} - \cdots$$

$$= \Big(1 - \frac{s^2}{2!} + \frac{s^4}{4!} - \cdots\Big)\mathbb{1} + \Big(\frac{s}{1!} - \frac{s^3}{3!} + \frac{s^5}{5!} - \cdots\Big)J = \big(\cos s\big)\mathbb{1} + \big(\sin s\big)J.$$

Hence

$$e^{tbJ} = \begin{bmatrix} \cos(tb) & -\sin(tb) \\ \sin(tb) & \cos(tb) \end{bmatrix}.$$

□

Example 18.4.21. Suppose that

$$G=\begin{bmatrix}1&0&0\\1&1&0\\0&0&1\end{bmatrix}\ \text{ and }\ A_G=C_1=\mathbb{1}_3+N_3.$$

Then

$$G^{-1}=\begin{bmatrix}1&0&0\\-1&1&0\\0&0&1\end{bmatrix},\ N_3^2=\begin{bmatrix}0&0&1\\0&0&0\\0&0&0\end{bmatrix},$$

$$e^{tN_3}=\mathbb{1}+\begin{bmatrix}0&t&0\\0&0&t\\0&0&0\end{bmatrix}+\frac{1}{2}\begin{bmatrix}0&0&t^2\\0&0&0\\0&0&0\end{bmatrix}=\begin{bmatrix}1&t&t^2/2\\0&1&t\\0&0&1\end{bmatrix},$$

$$e^{tA_G}=e^t\begin{bmatrix}1&t&t^2/2\\0&1&t\\0&0&1\end{bmatrix}$$

$$e^{tA}=e^t\begin{bmatrix}1&0&0\\1&1&0\\0&0&1\end{bmatrix}\cdot\begin{bmatrix}1&t&t^2/2\\0&1&t\\0&0&1\end{bmatrix}\cdot\begin{bmatrix}1&0&0\\-1&1&0\\0&0&1\end{bmatrix}.$$

18.4.9. Differentiability of the solutions with respect to initial data and parameters. In this section we have a new look at the problem investigated in Section 18.2.5. Consider the Cauchy problem

$$\begin{aligned}&\boldsymbol{x}'=\boldsymbol{F}(t,\boldsymbol{x}),\ \ (t,\boldsymbol{x})\in\Omega\subset\mathbb{R}^{n+1},\\&\boldsymbol{x}(t_0)=\boldsymbol{x}_0,\end{aligned}\tag{18.4.52}$$

where $\boldsymbol{F}:\Omega\to\mathbb{R}^n$ is continuous in $(t,\boldsymbol{x})$ and locally Lipschitz in $\boldsymbol{x}$. We denote by $\boldsymbol{x}(t;t_0,\boldsymbol{x}_0)$ the right-saturated solution of the Cauchy problem (18.4.52) defined on the right-maximal existence interval $[t_0,T_+)$. We proved in Theorem 18.2.24 that for any $T\in[t_0,T_+)$ there exists $\eta>0$ such that for any

$$\boldsymbol{\xi}\in S(\boldsymbol{x}_0,\eta)=\big\{\boldsymbol{x}\in\mathbb{R}^n;\ \ \|\boldsymbol{x}-\boldsymbol{x}_0\|<\eta\big\}$$

the solution $\boldsymbol{x}(t,\xi)=\boldsymbol{x}(t;t_0,\boldsymbol{\xi})$ the solution of the initial value problem

$$\begin{aligned}&\boldsymbol{x}'=\boldsymbol{F}(t,\boldsymbol{x}),\ \ (t,\boldsymbol{x})\in\Omega\subset\mathbb{R}^{n+1},\\&\boldsymbol{x}(t_0)=\boldsymbol{\xi},\end{aligned}$$

is defined in $[t_0,T]$ and the resulting map

$$S(\boldsymbol{x}_0,\eta)\ni\boldsymbol{\xi}\mapsto\boldsymbol{x}(t,\boldsymbol{\xi})\in C\big([t_0,T];\mathbb{R}^n\big)$$

is continuous. We now investigate the differentiability of the above map. We begin by recalling some facts about Fréchet differentiability.

Suppose that X, Y are Banach spaces with norms $\| - \|_X$ and respectively $\| - \|_Y$, $U \subset X$ is an open subset and $T : U \to Y$ a map. We say that T is Fréchet differentiable at $x_0 \in X$ if there exists a bounded linear operator $L \in \boldsymbol{B}\big(X, Y \big)$ such that

$$\lim_{x \to x_0} \frac{\|T(x) - T(x_0) - L(x - x_0)\|_Y}{\|x - x_0\|_X} = 0. \tag{18.4.53}$$

Using Landau's notation we can rewrite the last condition as

$$\|T(x) - T(x_0) - L(x - x_0)\|_Y = o\big(\|x - x_0\|_X \big) \;\text{ as } \boldsymbol{x} \to \boldsymbol{x}_0. \tag{18.4.54}$$

Traditionally, one writes $x = x_0 + h$ and the above equality becomes

$$\lim_{h \to 0} \frac{\|T(x_0 + h) - T(x_0) - Lh\|_Y}{\|h\|_X} = 0. \tag{18.4.55}$$

The operator L is denoted by $T'(x_0)$ and it is called the *Fréchet derivative* of T at x_0. Observe that is T is differentiable at a point x_0 and $T'(x_0)$ is its differential, then the action of the operator $T'(x_0)$ on a vector $h \in X$ is given by an ordinary derivative

$$T'(x_0)h = \frac{d}{d\tau}\big|_{\tau=0} T(x_0 + \tau h).$$

The map T is said to be differentiable on U if it is differentiable at any $u \in U$, and it is called C^1 if it is differentiable on U and the resulting map

$$U \ni u \mapsto T'(u) \in \boldsymbol{B}(X, Y)$$

is continuous, where the vector space $\mathbb{B}(X, Y)$ is equipped with the operator norm $\| - \|_{op}$. When $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ we recover our earlier definition of differentiability, Definition 13.1.1.

The function

$$X \ni x \mapsto T(x_0) + T'(x_0)(x - x_0) \in Y$$

is called the *linear approximation* of T at x_0. The error of this approximation is the remainder

$$R(x, x_0) = T(x) - T(x_0) - T'(x_0)(x - x_0)$$

Theorem 18.4.22. *Under the same assumptions as in Theorem 18.2.24 assume additionally that the function $\boldsymbol{F}$ is differentiable with respect to $\boldsymbol{x}$ and the differential $\boldsymbol{F}_{\boldsymbol{x}}$ is continuous with respect to $(t, \boldsymbol{x})$. Fix $\boldsymbol{\xi}_0 \in S(\boldsymbol{x}_0, \eta)$. For any $t \in [t_0, T]$ we denote by $A(t)$ the $\boldsymbol{x}$-derivative of $\boldsymbol{F}$ at the point $(t, \boldsymbol{x}(t, \boldsymbol{x}_0))$,*

$$A(t) = \boldsymbol{F}_{\boldsymbol{x}}\big(t, \boldsymbol{x}(t, \boldsymbol{x}_0) \big).$$

*Then for any $t \in [t_0, T]$ the function $\boldsymbol{\xi} \mapsto \boldsymbol{x}(t, \boldsymbol{\xi})$ is differentiable with respect to $\boldsymbol{\xi}$ at $\boldsymbol{\xi}_0$ and its differential $X(t) := \boldsymbol{x}_{\boldsymbol{\xi}}(t, \boldsymbol{\xi}_0)$, viewed as a function of t, is the fundamental matrix of the linear system (*variation equation*)*

$$\boldsymbol{y}' = A(t)\boldsymbol{y}, \;\; t_0 \leqslant t \leqslant T, \tag{18.4.56}$$

satisfying the initial condition

$$X(t_0) = \mathbb{1}. \tag{18.4.57}$$

In other words

$$X'(t) = A(t)X(t),\;\; X(t_0) = \mathbb{1}.$$

Proof. The reason why (18.4.56) determines $X(t)$ can be explained heuristically as follows. We have the equality

$$\boldsymbol{x}(t, \boldsymbol{\xi}_0 + \tau\boldsymbol{h}) = \boldsymbol{\xi}_0 + \tau\boldsymbol{h} + \int_{t_0}^{t} \boldsymbol{F}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}_0 + \tau\boldsymbol{h})\,\big) ds.$$

Differentiating (formally) the above equality with respect to τ at $\tau = 0$, using the chain rule and recalling that

$$X(s)\boldsymbol{h} = \frac{d}{d\tau}\big|_{\tau=0}\boldsymbol{x}(s, \boldsymbol{\xi}_0 + \tau\boldsymbol{h})$$

we deduce

$$X(t)\boldsymbol{h} = \boldsymbol{h} + \int_{t_0}^{t} \boldsymbol{F}_{\boldsymbol{x}}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big) X(s)\boldsymbol{h}\; ds.$$

This is equivalent with (18.4.56) + (18.4.57). Let us supply the precise arguments.

Let $\boldsymbol{\xi} \in S(\boldsymbol{\xi}_0, \eta)$. (Think $\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \tau\boldsymbol{h}$.) Fix $t_1 \in [0, T]$. We need to estimate the remainder

$$\rho(t_1, \boldsymbol{\xi}, \boldsymbol{\xi}_0) := \boldsymbol{x}(t_1, \boldsymbol{\xi}) - \boldsymbol{x}(t_1, \boldsymbol{\xi}_0) - X(t_1)\big(\,\boldsymbol{\xi} - \boldsymbol{\xi}_0\,\big).$$

Note that

$$\rho(t_1, \boldsymbol{\xi}, \boldsymbol{\xi}_0) = \int_{t_0}^{t_1} \Big(\, \boldsymbol{F}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi})\,\big) - \boldsymbol{F}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big) - A(s)X(s)\big(\,\boldsymbol{\xi} - \boldsymbol{\xi}_0\,\big)\Big) ds, \tag{18.4.58}$$

where $X(t)$ is the fundamental matrix of the linear system (18.4.56) that verifies the initial condition (18.4.57). On the other hand,

$$\begin{aligned}\boldsymbol{F}(s, \boldsymbol{u}) - \boldsymbol{F}\big(\, s, \boldsymbol{u}_0)\,\big) &= \int_0^1 \frac{d}{d\tau}\boldsymbol{F}\big(\, s, \boldsymbol{u}_0 + \tau(\boldsymbol{u} - \boldsymbol{u})\,\big) d\tau \\ &= \int_0^1 \boldsymbol{F}_{\boldsymbol{x}}\big(\, s, \boldsymbol{u}_0 + \tau(\boldsymbol{u} - \boldsymbol{u}_0)\,\big)\big(\,\boldsymbol{u} - \boldsymbol{u}_0\,\big) d\tau.\end{aligned}$$

Hence

$$\begin{aligned}&\boldsymbol{F}(s, \boldsymbol{u}) - \boldsymbol{F}\big(\, s, \boldsymbol{u}_0)\,\big) - \boldsymbol{F}_x(s, \boldsymbol{u}_0)\big(\,\boldsymbol{u} - \boldsymbol{u}_0\,\big) \\ &= \underbrace{\int_0^1 \Big(\, \boldsymbol{F}_{\boldsymbol{x}}\big(\, s, \boldsymbol{u} + \tau(\boldsymbol{u} - \boldsymbol{u}_0)\,\big)\big(\,\boldsymbol{u} - \boldsymbol{u}_0\,\big) - \boldsymbol{F}_x(s, \boldsymbol{u}_0)\big(\,\boldsymbol{u} - \boldsymbol{u}_0\,\big)\Big) d\tau}_{=D(s,\boldsymbol{u},\boldsymbol{u}_0)}\,.\end{aligned}$$

Note that

$$\|D(s, \boldsymbol{u}, \boldsymbol{u}_0)\| \leqslant \underbrace{\Big(\int_0^1 \big\|\boldsymbol{F}_{\boldsymbol{x}}\big(\, s, \boldsymbol{u}_0 + \tau(\boldsymbol{u} - \boldsymbol{u}_0)\,\big)\;\; - \boldsymbol{F}_x(s, \boldsymbol{u}_))\big\| d\tau\Big)}_{=:\omega_s(\boldsymbol{u},\boldsymbol{u}_0)}\|\boldsymbol{u} - \boldsymbol{u}_0\|. \tag{18.4.59}$$

If we let $\boldsymbol{u}_0 = \boldsymbol{x}(s, \boldsymbol{\xi})$ and $\boldsymbol{u} = \boldsymbol{x}(s, \boldsymbol{\xi})$ we deduce that

$$\boldsymbol{F}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi})\,\big) - \boldsymbol{F}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big)$$

$$\begin{aligned} &= \boldsymbol{F}_{\boldsymbol{x}}\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big)\big(\, \boldsymbol{x}(s, \boldsymbol{\xi}) - \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big) + R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big) \\ &\qquad = A(s)\big(\, \boldsymbol{x}(s, \boldsymbol{\xi}) - \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big) + R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big), \end{aligned} \tag{18.4.60}$$

where

$$R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big) = D\big(\, s, \boldsymbol{x}(s, \boldsymbol{\xi}), \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big).$$

Hence

$$\rho(t_1) = \int_{t_0}^{t_1} A(s)\rho(s)ds + \int_{t_0}^{t_1} R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big)ds. \tag{18.4.61}$$

Since $\boldsymbol{F}$ is locally Lipschitz, there exists a constant $L > 0$ such that $\forall t \in [t_0, T]$ we have

$$\|\boldsymbol{x}(t, \boldsymbol{\xi}) - \boldsymbol{x}(t, \boldsymbol{\xi}_0)\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| + L\int_{t_0}^{t} \|\boldsymbol{x}(s, \boldsymbol{\xi}) - \boldsymbol{x}(s, \boldsymbol{\xi}_0)\|ds.$$

Invoking Gronwall's inequality we deduce

$$\|\boldsymbol{x}(t, \boldsymbol{\xi}) - \boldsymbol{x}(t, \boldsymbol{\xi}_0)\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|e^{L(t-t_0)}, \quad \forall t \in [t_0, T]. \tag{18.4.62}$$

Hence, for any $s \in [t_0, t]$

$$\begin{aligned} \|R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big)\| &\overset{(18.4.59)}{\leqslant} \omega_s\big(\, \boldsymbol{x}(s, \boldsymbol{\xi}), \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big)\|\boldsymbol{x}(s, \boldsymbol{\xi}) - \boldsymbol{x}(s, \boldsymbol{\xi}_0)\| \\ &\overset{(18.4.62)}{\leqslant} \underbrace{e^{L(s-t_0)}\omega_s\big(\, \boldsymbol{x}(s, \boldsymbol{\xi}), \boldsymbol{x}(s, \boldsymbol{\xi}_0)\,\big)}_{=:\Omega_s(\boldsymbol{\xi}, \boldsymbol{\xi}_0)}\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|. \end{aligned}$$

As we know, the map

$$S(\boldsymbol{x}_0, \eta) \ni \xi \mapsto \boldsymbol{x}(-, \boldsymbol{\xi}) \in C\big(\, [t_0, T], \mathbb{R}^n\big)$$

is continuous so

$$\lim_{\boldsymbol{\xi}\to\boldsymbol{\xi}_0} \boldsymbol{x}(s, \boldsymbol{\xi}) = \boldsymbol{x}(s, \boldsymbol{\xi}_0)$$

uniformly in $s \in [t_0, T]$. The inequality (18.4.62) and the *uniform* continuity of the differential $\boldsymbol{F}_{\boldsymbol{x}}$ on the compacts of Ω imply that the remainder R in (18.4.60) satisfies the estimate

$$\|R\big(\, s, \boldsymbol{\xi}, \boldsymbol{\xi}_0\,\big)\| \leqslant \Omega_s(\boldsymbol{\xi}, \boldsymbol{\xi}_0)\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|, \tag{18.4.63}$$

where

$$\lim_{\boldsymbol{\xi}\to\boldsymbol{\xi}_0} \Omega_s(\boldsymbol{\xi}, \boldsymbol{\xi}_0) = 0,$$

uniformly in s. Set

$$z(t) := \|\rho(t)\|, \;\; L_1 = \sup_{s\in[t_0,T]} \|A(s)\|_{\mathrm{op}}, \;\; \delta(\boldsymbol{\xi}, \boldsymbol{\xi}_0) = \sup_{s\in[t_0,T]} \Omega_s(\boldsymbol{\xi}, \boldsymbol{\xi}_0).$$

Using (18.4.63) in (18.4.61) we obtain the estimate

$$z(t) \leqslant (T - t_0)\delta(\boldsymbol{\xi}, \boldsymbol{\xi}_0)\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| + L_1\int_{t_0}^{t} z(s)ds, \quad \forall t \in [t_0, T']. \tag{18.4.64}$$

Invoking Gronwall's inequality again we deduce that

$$\|\, \boldsymbol{x}(t, \boldsymbol{\xi}) - \boldsymbol{x}(t, \boldsymbol{\xi}_0) - X(t)\big(\, \boldsymbol{\xi} - \boldsymbol{\xi}_0\,\big)\, \|$$

$$\leqslant (T-t_0)\delta(\boldsymbol{\xi},\boldsymbol{\xi}_0)\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}_0\|e^{L_1(T-t_0)} = o\big(\,\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\,\big). \tag{18.4.65}$$

The last inequality implies (see Remark 13.1.4 or (18.4.54)) that

$$\boldsymbol{x}_{\boldsymbol{\xi}}(t,\boldsymbol{\xi}_0) = X(t).$$

□

Example 18.4.23. Consider a simple gravitational pendulum. A material point of mass m is attached to a rigid light arm of length L whose other end is attached to a frictionless pivot; see Figure 18.7. The angular displacement θ is a function of time satisfying the differential equation

$$\theta+\omega^2\sin\theta=0,\ \ \omega^2=\frac{g}{L}.$$

Denote by $\theta(t,s)$ the solution of the above equation satisfying the initial conditions

$$\theta(0)=s,\ \ \theta'(0)=0.$$

Note that $\theta(t,0)=0$, $\forall t$. Theorem 18.4.22 shows that, for every t, the function $s\mapsto\theta(t,s)$ is differentiable. We set

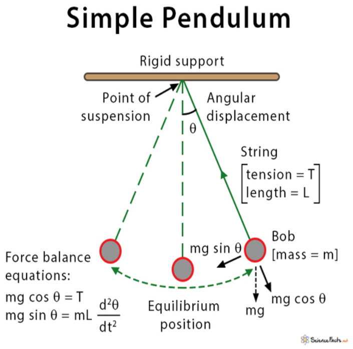

Figure 18.7. *Simple gravitational pendulum.*

$$y(t):=\partial_s\theta(t,s)\big|_{s=0}.$$

Derivating with respect to s the equality

$$\partial_t^2\theta(t,s)+\omega^2\sin\theta(t,s)=0$$

we deduce that $y(t)$ is the unique solution of the differential equation

$$y''(t)+\omega^2y(t)=0,\ \ y(0)=1,\ \ y'(0)=0.$$

Hence

$$y(t)=\cos(\omega t).$$

The function y is periodic, with period $T=\frac{2\pi}{\omega}$ and thus

$$\theta(T,s)=sy(T)+O(s^2)=\theta(0,s)+O(s^2)$$

Hence, if the initial angular displacement $\theta(0) = s$ is small, then, after T seconds the pendulum is very close to the initial position. Hence we can take $T = 2\pi\sqrt{\frac{L}{g}}$ to be a reasonably good approximation of the period of the oscillation of a simple pendulum. □

We next investigate the differentiability with respect to a parameter $\boldsymbol{\lambda}$ of the solution $\boldsymbol{x}(t, \lambda)$ of the Cauchy problem

$$\boldsymbol{x}' = \boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{\lambda}), \;\; (t, \boldsymbol{x}) \in \Omega \subset \mathbb{R}^{n+1}, \;\; \boldsymbol{\lambda} \in U \subset \mathbb{R}^m, \tag{18.4.66a}$$

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0. \tag{18.4.66b}$$

The parameter $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ varies in a bounded open subset U on $\mathbb{R}^m$. Fix $\boldsymbol{\lambda}_0 \in U$. Assume that the right-saturated solution $\boldsymbol{x}(t, \boldsymbol{\lambda}_0)$ the solution of (18.4.66a)-(18.4.66b) corresponding to $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ is defined on the maximal-to-the-right interval $[t_0, T_+)$.

Theorem 18.4.24. *Let*

$$\boldsymbol{F} : \Omega \times U \to \mathbb{R}^n$$

be a continuous function, differentiable in the $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ variables, and with the differentials $\boldsymbol{F}_{\boldsymbol{x}}$, $\boldsymbol{F}_{\boldsymbol{\lambda}}$ continuous in $(t, \boldsymbol{x}, \boldsymbol{\lambda})$. Then, for any $T \in [t_0, T_+)$, there exists $\delta > 0$ such that the following hold.

(i) *The solution $\boldsymbol{x}(t, \boldsymbol{\lambda})$ is defined on $[t_0, T]$ for any*

$$\boldsymbol{\lambda} \in S(\boldsymbol{\lambda}_0, \delta) := \big\{\, \boldsymbol{\lambda} \in \mathbb{R}^m; \;\; \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\| < \delta \,\big\}.$$

(ii) *For any $t \in [t_0, T]$ the map*

$$S(\boldsymbol{\lambda}_0, \delta) \ni \boldsymbol{\lambda} \mapsto \boldsymbol{x}(t, \boldsymbol{\lambda}) \in \mathbb{R}^n$$

is differentiable and the differential $\boldsymbol{y}(t) := \boldsymbol{x}_{\boldsymbol{\lambda}}(t, \boldsymbol{\lambda}_0)$ at $\boldsymbol{\lambda}_0$ is uniquely determined by the (matrix valued) linear nonhomogeneous Cauchy problem

$$\boldsymbol{y}'(t) = \boldsymbol{F}_{\boldsymbol{x}}\big(t, \boldsymbol{x}(t, \boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0\big)\boldsymbol{y}(t) + \boldsymbol{F}_{\boldsymbol{\lambda}}\big(t, \boldsymbol{x}(t, \boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0\big), \;\; \forall t \in [t_0, T], \tag{18.4.67}$$

$$\boldsymbol{y}(t_0) = 0. \tag{18.4.68}$$

Proof. As in the proof of Theorem 18.2.28 we form a new Cauchy problem,

$$\begin{gathered} \boldsymbol{x}' = \widehat{\boldsymbol{F}}(t, \boldsymbol{z}), \\ \boldsymbol{z}(t_0) = \boldsymbol{\zeta} := (\boldsymbol{\xi}, \boldsymbol{\lambda}). \end{gathered} \tag{18.4.69}$$

where

$$\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\lambda}), \;\; \widehat{\boldsymbol{F}}(t, \boldsymbol{z}) = \big(\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{\lambda}), 0\big) \in \mathbb{R}^n \times \mathbb{R}^m.$$

According to Theorem 18.4.22, the map

$$\boldsymbol{\zeta} \mapsto \big(\boldsymbol{x}(t, \boldsymbol{\xi}, \boldsymbol{\lambda}), \boldsymbol{\lambda}\big) =: \boldsymbol{z}(t, \boldsymbol{\zeta})$$

is differentiable and its differential

$$Z(t) := \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\zeta}} = \begin{bmatrix} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\xi}}(t; t_0, \boldsymbol{\xi}, \boldsymbol{\lambda}) & \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\lambda}}(t; t_0, \boldsymbol{\xi}, \boldsymbol{\lambda}) \\ 0 & \mathbb{1}_m \end{bmatrix}$$

($\mathbb{1}_m$ is the identity $m \times m$ matrix) satisfies the differential equation

$$Z'(t) = \widehat{\boldsymbol{F}}_{\boldsymbol{z}}(t, \boldsymbol{z})Z(t), \quad Z(t_0) = \mathbb{1}_{n+m}. \tag{18.4.70}$$

Taking into account the description of $Z(t)$ and the equality

$$\widehat{\boldsymbol{F}}_{\boldsymbol{z}}(t, \boldsymbol{z}) = \begin{bmatrix} \boldsymbol{F}_{\boldsymbol{x}}(t, \boldsymbol{x}, \boldsymbol{\lambda}) & \boldsymbol{F}_{\boldsymbol{\lambda}}(t, \boldsymbol{x}, \boldsymbol{\lambda}) \\ 0 & 0 \end{bmatrix},$$

we conclude from (18.4.70) that $\boldsymbol{y}(t) := \boldsymbol{x}_{\boldsymbol{\lambda}}(t; t_0, \boldsymbol{x}_0, \boldsymbol{\lambda})$ satisfies the Cauchy problem (18.4.67)-(18.4.68).

□

Remark 18.4.25. The matrix $\boldsymbol{x}_{\boldsymbol{\lambda}}(t, \boldsymbol{\lambda}_0)$ is sometimes called the *sensitivity matrix* and its entries are known as *sensitivity functions*. Measuring the changes in the solution under small variations of the parameter $\boldsymbol{\lambda}$, this matrix is an indicator of the robustness of the system. □

Theorem 18.4.24 is especially useful in the approximation of the solutions of the differential systems via the so called *small-parameter method*.

Let us denote by $\boldsymbol{x}(t, \boldsymbol{\lambda})$ the solution $\boldsymbol{x}(t; t_0, \boldsymbol{x}_0, \boldsymbol{\lambda})$ of the Cauchy problem (18.4.66a)-(18.4.66b). We then have a first order approximation

$$\boldsymbol{x}(t, \boldsymbol{\lambda}) = \boldsymbol{x}(t, \boldsymbol{\lambda}_0) + \boldsymbol{x}_{\boldsymbol{\lambda}}(t, \boldsymbol{\lambda}_0)(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) + o(\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|), \tag{18.4.71}$$

where $\boldsymbol{y}(t) = \boldsymbol{x}_{\boldsymbol{\lambda}}(t, \boldsymbol{\lambda}_0)$ is the solution of the variation equation (18.4.67)-(18.4.68). Thus, in a neighborhood of the parameter $\boldsymbol{\lambda}_0$ we have

$$\boldsymbol{x}(t, \boldsymbol{\lambda}) \approx \boldsymbol{x}(t, \boldsymbol{\lambda}_0) + \boldsymbol{x}_{\boldsymbol{\lambda}}(t, \boldsymbol{\lambda}_0)(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0).$$

We have thus reduced the approximation problem to solving a linear differential system.

Example 18.4.26. Let us illustrate the technique on the following example

$$x' = x + \lambda t x^2 + 1, \quad x(0) = 1, \tag{18.4.72}$$

where λ is a sufficiently small parameter. The equation (18.4.72) is a Riccati type equation and cannot be solved by quadratures. However, for $\lambda = 0$ it reduces to a linear equation and its solution is

$$x(t, 0) = 2e^t - 1.$$

According to formula (18.4.71) the solution $x(t, \lambda)$ admits an approximation

$$x(t, \lambda) = 2e^t - 1 + \lambda y(t) + o(|\lambda|),$$

where $y(t) = x_\lambda(t, 0)$ is the solution of the variation equation

$$y' = y + t(2e^t - 1)^2, \quad y(0) = 0.$$

Hence

$$y(t) = \int_0^t s(2e^s - 1)^2 e^{t-s} ds.$$

Thus, for small values of the parameter λ the solution of the problem (18.4.72) is well approximated by

$$2e^t - 1 + \lambda e^t \big(4te^t - 2t^2 - 4e^t + e^{-t} + 3 - te^{-t} \big).$$

□

18.5. Exercises

Exercise 18.1. A reservoir contains ℓ liters of salt water with the concentration c_0. Salt water is flowing in the reservoir at a rate of ℓ_0-liters per minute and with a concentration α_0. The same amount of salt water is leaving the reservoir every minute. Assuming that the salt in water is uniformly distributed, find the time evolution of the concentration of salt in water.**Hint.** Let the state of the system be the concentration $x(t)$ of salt in water. Show that

$$\ell x'(t) = (\alpha_0 - x(t))\ell_0,$$

and the initial condition $x(0) = c_0$. □

Exercise 18.2. Prove that any solution of the o.d.e.

$$x' = \sqrt[3]{\frac{x^2+1}{t^4+1}}$$

has two horizontal asymptotes. □

Exercise 18.3. Prove that there exists a unique solution of the od.e.

$$tx' = (2t^2)x = t^2,$$

that has a finite limit as $t \to \infty$. Find this solution. □

Exercise 18.4. Find the plane curve with the property that the distance from the origin to any tangent line to the curve is equal to the x-coordinate of the tangency point. □

Exercise 18.5. Find the solution of the o.d.e.

$$3x^2x' + 16t = 2tx^3$$

that is bounded on the positive semiaxis $[0, \infty)$. □

Exercise 18.6. Prove that the o.d.e.

$$x' + \omega x = f(t), \tag{18.5.1}$$

where ω is a positive constat and $f : \mathbb{R} \to \mathbb{R}$ is continuous and bounded has a unique solution that is bounded on $\mathbb{R}$. Find this solution x and prove that if f is periodic, then x is also periodic, with the same period. □

Exercise 18.7. Consider the o.d.e.

$$tx' + ax = f(t),$$

where a is a positive constant and

$$\lim_{t\to 0} f(t) = \alpha.$$

Prove that there exists a unique solution of this equation that has finite limit as $t \to 0$ and then find this solution. □

Exercise 18.8. According to Newton's heating and cooling law the rate of decrease in temperature of a body that is cooling is proportional to the difference between the temperature of the body and the temperature of the ambient surrounding. Find the equation that models the cooling phenomenon. □

Exercise 18.9. Let $f : [0, \infty) \to \mathbb{R}$ be a continuous function such that $\lim_{t\to\infty} f(t) = 0$. Prove that any solution of (18.5.1) goes to 0 as $t \to \infty$. □

Exercise 18.10. Let $f : [0, \infty) \to \mathbb{R}$ be a continuous function such that

$$\int_0^\infty |f(t)|dt < \infty.$$

Prove that the solutions of the o.d.e.

$$x' + \big(\omega + f(t)\big)x = 0, \;\; \omega > 0,$$

converge to 0 as $t \to \infty$. □

Exercise 18.11. Solve the (Lotka-Volterra) o.d.e.

$$\frac{dy}{dx} = \frac{y(ux - v)}{x(a - by)}, \;\; a, b, u, v > 0.$$

□

Exercise 18.12. Solve the o.d.e.

$$x' = k(a - x)(b - x).$$

(Such an equation models certain chemical reactions.) □

Exercise 18.13. Find the solution of the o.d.e.

$$x' \sin t = 2(x + \cos t)$$

that stays bounded as $t \to \infty$. □

Exercise 18.14. Prove that by using the substitution $y = \frac{x'}{x}$ we can reduce the second order o.d.e. $x'' = a(t)x$ to the Riccati-type equation

$$y' = -y^2 + a(t). \tag{18.5.2}$$

□

Exercise 18.15. Prove that if $x_1(t), x_2(t), x_3(t), x_4(t)$ are solutions of a Riccati-type o.d.e., then the cross-ratio

$$\frac{x_3(t) - x_1(t)}{x_3(t) - x_2(t)} \div \frac{x_4(t) - x_1(t)}{x_4(t) - x_2(t)}$$

is independent of t.**Hint.** Use the remark in Subsection 18.1.5 to show that there exist constants $C_1, C_2 \in \mathbb{R}$ such that

$$\frac{1}{x_4} - \frac{1}{x_2} = C_1\left(\frac{1}{x_3} - \frac{1}{x_2}\right), \;\; \frac{1}{x_4} - \frac{1}{x_1} = C_2\left(\frac{1}{x_3} - \frac{1}{x_1}\right).$$

□

Exercise 18.16. Find the plane curves such that the area of the triangle formed by any tangent with the coordinate axes is a given constant a^2. □

Exercise 18.17. Consider the family of curves in the (t, x)-plane described by

$$F(t, x, \lambda) = 0, \quad \lambda \in \mathbb{R}. \tag{18.5.3}$$

(a) Find the curves that are orthogonal to all the curves in this family.

(b) Find the curves orthogonal the all the curves in the family $x = \lambda e^t$.

Hint. Since the tangent line to a curve is parallel to the vector $\left(1, -\frac{F_x}{F_t}\right)$, the orthogonal curves are solutions of the differential equation

$$x' + \frac{F_x}{F_t} = 0,$$

where λ was replaced by its value $\lambda = \lambda(t, x)$ determined from (18.5.3). □

Exercise 18.18. Find the solitons of the Klein-Gordon equation

$$u_{tt} - u_{xx} + u + u^3 = 0. \tag{18.5.4}$$

□

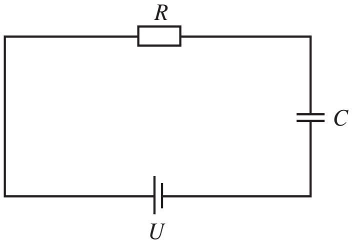

Figure 18.8. An *RC* circuit.

Exercise 18.19. Find the differential equation that models the behavior of an *RC* electric circuit as in Figure 18.8.

Hint. If we denote by Q the electric charge of the capacitor, then we have $C^{-1}Q + RI = U$, where I denotes the electric current. Thus Q satisfies the o.d.e.

$$R\frac{dQ}{dt} + C^{-1}Q = U. \tag{18.5.5}$$

□

Exercise 18.20. Find the system of o.d.e.-s that models the behavior of the electrical circuit in Figure 18.9.

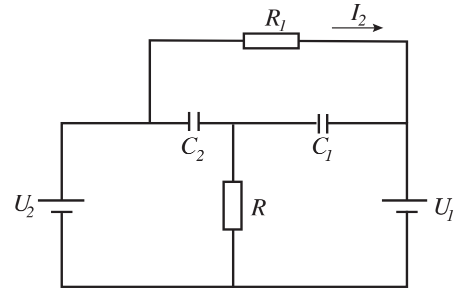

Figure 18.9. A more complex RC circuit.

Hint. Denote by Q_i the electrical charge of the capacitor C_i, $i = 1, 2$, and by I_i the corresponding electrical currents. Kirchoff's laws yield the equations

$$\begin{aligned} C_2^{-1}Q_2 + RI_1 &= U_2, \\ -C_2^{-1}Q_2 + R_1 I_2 + C_1^{-1}Q_1 &= 0, \\ \frac{dQ_1}{dt} = I_2, \ \ \frac{dQ_2}{dt} &= I_1 - I_2. \end{aligned} \tag{18.5.6}$$

Using as state of the system the pair $x_1 = Q_1$ and $x_2 = Q_2$ we obtain a system of first order o.d.e.-s in (x_1, x_2). □

Exercise 18.21. Let $(-T_-, T_+)$, $0 < T_\pm \leqslant \infty$, denote the maximal interval of existence for the solution of the Cauchy problem

$$x' = -x^2 + t + 1, \ \ x(0) = 1.$$

(i) Prove that $T_+ = \infty$.

(ii) Prove that $1 \leqslant T_- < \infty$.

Hint. (i) For $t > 0$ we have $x' \leqslant t + 1$. (ii) Consider the new function $y(t) = x(-t) - 1$, $t \in [0, T_-)$. Show that for $t \in (0, 1)$ it satisfies $y' \leqslant (y + 1)^2$ while for $t > 1$ is satisfies $y' \geqslant (y + 1)^2$. □

Exercise 18.22. Consider the Cauchy problem

$$x' = f(t, x), \ \ x(t_0) = x_0, \ \ (t_0, x_0) \in \Omega \subset \mathbb{R}^2, \tag{18.5.7}$$

where the function f is continuous in (t, x) and locally Lipschitz in x. Prove that if $x_0 \geqslant 0$ and $f(t, 0) > 0$ for any $t \geqslant t_0$, then the saturated solution $x(t; t_0, x_0)$ is nonnegative for any $t \geqslant t_0$ in the existence interval. □

Exercise 18.23. Consider the system

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}), \ \ \boldsymbol{x}(t_0) = \boldsymbol{x}_0, \ \ t_0 \geqslant 0,$$

where the function $\boldsymbol{f} : [0, \infty[\times \mathbb{R}^n \to \mathbb{R}^n$ is continuous in $(t, \boldsymbol{x})$, locally Lipschitz in $\boldsymbol{x}$ and satisfies the condition

$$\big(\boldsymbol{f}(t, \boldsymbol{x}), P\boldsymbol{x} \big) \leqslant 0, \ \ \forall t \geqslant 0, \ \ \boldsymbol{x} \in \mathbb{R}^n, \tag{18.5.8}$$

where P is a real, symmetric positive definite $n \times n$ matrix. Prove that any right-saturated solution of the system is defined on the semi-axis $[t_0, \infty)$.

Hint. Note that $(\boldsymbol{x}, \boldsymbol{y}_P := (\boldsymbol{x}, P\boldsymbol{y})$ is a new inner product on $\mathbb{R}^n$. □

Exercise 18.24. Consider the Cauchy problem

$$x'' + ax + f(x') = 0, \quad x(t_0), \quad x'(t_0) = x_1, \tag{18.5.9}$$

where a is a positive constant, and $f : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz function satisfying

$$yf(y) \geqslant 0, \quad \forall y \in \mathbb{R}.$$

Prove that any right-saturated solution of (18.5.9) is defined on the semi-axis $[t_0, \infty[$.

Hint. Multiply (18.5.9) by x' and prove

$$\frac{1}{2}\frac{d}{dt}\Big(\,|x'(t)|^2 + a|x(t)|^2\,\Big) \leqslant 0, \quad \forall t \geqslant t_0.$$

□

Exercise 18.25. In the anisotropic theory of relativity due to V.G. Boltyanski, the propagation of light in a neighborhood of a mass m located at the origin of $\mathbb{R}^3$ is described by the equation

$$\boldsymbol{x}' = -\frac{m\gamma}{\|\boldsymbol{x}\|2^3}\boldsymbol{x} + \boldsymbol{u}(t), \tag{18.5.10}$$

where γ is a positive constant $\boldsymbol{u} : [0, \infty[\to \mathbb{R}^3$ is a continuous and *bounded* function, i.e.,

$$\exists C > 0; \quad \|\boldsymbol{u}(t)\|_2 \leqslant C, \quad \forall t \geqslant 0,$$

and $\boldsymbol{x}(t) \in \mathbb{R}^3$ is the location of the photon at time t. Prove that there exists $r > 0$ such that all the trajectories of (18.5.10) which start at $t = 0$ in the ball

$$B_r := \big\{\boldsymbol{x} \in \mathbb{R}^3; \quad \|\boldsymbol{x}\|_2 < r\,\big\}$$

will stay inside the ball as long as they are defined. (Such a ball is called a *black hole* in astrophysics.)

Hint. Take the inner product of (18.5.10) with $\boldsymbol{x}(t)$ to obtain the differential inequality

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{x}(t)\|_2^2 = -\frac{m\gamma}{\|\boldsymbol{x}(t)\|_2} + \big(\,\boldsymbol{u}(t), \boldsymbol{x}(t)\,\big) \leqslant -\frac{m\gamma}{\|\boldsymbol{x}(t)\|_2} + C\|\boldsymbol{x}(t)\|_2.$$

Use this differential inequality to obtain an upper estimate for $\|\boldsymbol{x}(t)\|_2$. □

Exercise 18.26. Prove that the saturated solution of the Cauchy problem

$$x' = e^{-x^2} + t^2, \quad x(0) = 1, \tag{18.5.11}$$

is defined on the interval $\big[0, \frac{1}{2}\big]$. Use Euler's method with step size $h = 10^{-2}$ to find an approximation of this solutions at the nodes $t_j = jh$, $j = 1, \dots, 50$.

You can use computers in which case name the software you used and include the code. □

Exercise 18.27. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous and nonincreasing function. Consider the Cauchy problem

$$\begin{aligned} x'(t) &= f(x), \quad \forall t \geqslant 0, \\ x(0) &= x_0. \end{aligned} \tag{18.5.12}$$

According to Theorem 18.2.22, this problem has a unique solution $x(t)$ which exists on $[0, \infty[$.

(i) Prove that for any $\lambda > 0$ the function $\mathbb{1} - \lambda f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x - \lambda f(x)$, is bijective. We denote by $(\mathbb{1} - \lambda f)^{-1}$ its inverse

(ii) Prove that the inverse $J_\lambda := (\mathbb{1} - \lambda f)^{-1}$ is 1-Lipschitz.

(iii) Prove that for any $x \in \mathbb{R}$ we have

$$\lim_{\lambda \searrow 0} \frac{J_\lambda x - x}{\lambda} = x.$$

(iv) For any integer $n > 0$ we set

$$(\mathbb{1} - \lambda f)^{-n} := \underbrace{(\mathbb{1} - \lambda f)^{-1} \circ \cdots \circ (\mathbb{1} - \lambda f)^{-1}}_{n}.$$

Prove that $x(t)$ is given by the formula

$$x(t) = \lim_{n \to \infty} \left(\mathbb{1} - \frac{t}{n} f \right)^{-n} x_0, \quad \forall t \geqslant 0. \tag{18.5.13}$$

Hint. Fix $t > 0$, $n > 0$, set $h_n := \frac{t}{n}$ and define iteratively

$$x_0^n = x_0, \quad \frac{x_i^n - x_{i-1}^n}{h_n} = f(x_i^n), \quad i = 1, \dots n,$$

i.e.,

$$x_i^n = \left(\mathbb{1} - \frac{t}{n} f \right)^{-1} x_{i-1}^n = \left(\mathbb{1} - \frac{t}{n} f \right)^{-i} x_0. \tag{18.5.14}$$

Let $x^n : [0, t] \to \mathbb{R}$ be the unique continuous function which is linear on each of the intervals $[(i-1)h_n, ih_n]$ and satisfies

$$x^n(ih_n) = x_i^n, \quad \forall i = 0, \dots, n.$$

Argue as in the proof of Peano's theorem to show that the family of functions $(x^n)_{n \in \mathbb{N}}$ in $C([0, t])$ is bounded and equicontinuous. Deduce that it converges uniformly to x on $[0, t]$ as $n \to \infty$. The equality (18.5.13) now follows from (18.5.14). This problem is closely related to the situation described in Subsection 18.2.4. □

Exercise 18.28. Consider the Cauchy problem

$$\begin{aligned} x' &= f(x), \quad t \geqslant 0 \\ x(0) &= x_0, \end{aligned} \tag{18.5.15}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a continuous nonincreasing function. Let $x = \varphi(t)$ be a solution of (18.5.15). Prove that if the set

$$F := \big\{ y \in \mathbb{R}; \ f(y) = 0 \big\}$$

is nonempty then the following hold.

(i) The function $t \mapsto |x'(t)|$ is nonincreasing on $[0,\infty)$.

(ii) $\lim_{t\to\infty} |x'(t)| = 0$.

(iii) There exists $x_\infty \in F$ such that

$$\lim_{t\to\infty} x(t) = x_\infty.$$

(iv) Fix $y \in \mathbb{R}$ and $\lambda > 0$ and consider the nonincreasing function $g : \mathbb{R} \to \mathbb{R}$ $g(x) = f(x) - \lambda x + y$. Denote by $\varphi(t)$ the solution of the initial value problem

$$\varphi'(t) = g\big(\varphi(t)\big), \quad \varphi(0) = 0.$$

Prove that

$$\lim_{t\to\infty} \varphi(t) = (\mathbb{1} - \lambda f)^{-1}(y).$$

Hint. (i) Since f is nonincreasing we have

$$\frac{1}{2}\frac{d}{dt}\big(x(t+h) - x(t)\big)^2 = \big(x'(t+h) - x'(t)\big)\big(x(t+h) - x(t)\big)$$

(ii) Multiply both sides of (18.5.15) with $x(t) - y_0$, where $y_0 \in F$. Conclude similarly that

$$\frac{d}{dt}\big(x(t) - y_0\big)^2 \leqslant 0.$$

showing that $\lim_{t\to\infty}(x(t) - y_0)^2$ exists. Show that

$$\int_0^t |x'(s)|^2 ds = F(x(t)) - F(x_0) < \infty,$$

where F is an antiderivative of f. Prove that this, combined with (i), yields (ii). Prove that there exists a subsequence $t_n \to \infty$ such that $x(t_n)$ is convergent and show that its limit y_∞ is in F. Conclude that

$$\lim_{t\to\infty} x(t) = y_\infty.$$

□

Exercise 18.29. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a C^1 convex function such that

(i) $g(\boldsymbol{x}) \geqslant 0$, $\forall \boldsymbol{x} \in \mathbb{R}^n$,

(ii) The set $C_g := \big\{\, \boldsymbol{x};\ ]\ \nabla g(\boldsymbol{x}) = 0 \,\big\}$ is nonempty.

Set $\boldsymbol{F} = -\nabla g$. Show that if $\boldsymbol{\varphi} : [0,\infty) \to \mathbb{R}^n$ is a solution of $\boldsymbol{\varphi}' = \boldsymbol{F}\big(\boldsymbol{\varphi}\big)$, then

$$\lim_{t\to\infty} \boldsymbol{\varphi}(t)$$

exists and it is a point in C_g. **Hint.** Imitate the approach in Exercise 18.28. Have a look at Subsection 18.2.4. □

Exercise 18.30. Consider the equation $x' = f(t,x)$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is a function continuous in (t,x) and locally Lipschitz in x and satisfies the growth constraint

$$\big|\, f(t,x) \,\big| \leqslant \alpha(t)|x|, \quad \forall (t,x) \in \mathbb{R}^2,$$

where $\alpha : (0,\infty) \to (0,\infty)$ is a continuous function and, for some $t_0 > 0$,

$$\int_{t_0}^{\infty} \alpha(t)dt < \infty.$$

(i) Prove that any saturated solution of the equation exists for any $t > 0$.

(ii) Prove that any saturated solution of the equation has a finite limit as $t \to \infty$.

(iii) Prove that if, additionally, f satisfies the Lipschitz condition

$$\big| f(t,x) - f(t,y) \big| \leqslant \alpha(t)|x-y|, \quad \forall t \in \mathbb{R}, \quad x, y \in \mathbb{R},$$

then there exists a bijective correspondence between the initial values of the solutions and their limits at ∞.

□

Exercise 18.31. Prove that the maximal existence interval of the Cauchy problem

$$x' = ax^2 + t^2, \quad x(0) = x_0, \tag{18.5.16}$$

(a is a positive constant) is bounded above. **Hint.**

$$x(t;0,x_0) \geqslant x_0 + \frac{t^3}{3}, \quad \frac{1}{x(t;0,x_0)} \leqslant \frac{1}{x_0} - at.$$

□

Exercise 18.32. Consider the differential system

$$\boldsymbol{x}' = A(t)\boldsymbol{x}, \quad t \geqslant 0, \tag{18.5.17}$$

where $A(t)$ is an $n \times n$ matrix whose entries depend continuously on $t \in [0,\infty)$ and satisfies the condition

$$\liminf_{t\to\infty} \int_0^t \operatorname{tr} A(s)ds > -\infty. \tag{18.5.18}$$

Let $X(t)$ be a fundamental matrix of the system (18.5.17) that is bounded as a function of $t \in [0,\infty)$. Prove that the function

$$[0,\infty) \ni t \mapsto \| X(t)^{-1} \| \in]0,\infty)$$

is also bounded.**Hint.** Use Liouville's theorem. □

Exercise 18.33. Prove that if all the solutions of the system (18.5.17) are bounded on $[0,\infty[$ and (18.5.18) holds, then any solution of the system

$$\boldsymbol{x}' = B(t)\boldsymbol{x}, \quad t \geqslant 0, \tag{18.5.19}$$

is bounded on $[0,\infty[$. Above, $B(t)$ is an $n \times n$ matrix, whose entries depend continuously on $t \geqslant 0$, and satisfying the condition

$$\int_0^\infty \|B(s) - A(s)\|ds < \infty.$$

Hint. Rewrite the system (18.5.19) in the form

$$\boldsymbol{x}' = A(t)\boldsymbol{x} + \big(B(t) - A(t)\big)\boldsymbol{x},$$

and then use the formula of variation of constants. □

Exercise 18.34. Prove that all the solutions of the differential equation

$$x' + \left(1 + \frac{2}{t(1+t^2)}\right) x = 0$$

are bounded on $[0, \infty[$.

Hint. Interpret the function $f(t) = -2x(\,t(1+t^2)\,)^{-1}$ as nonhomogeneous term and then use the formula of variaton of constants. □

Exercise 18.35. Express as a power series the solution of the Cauchy problem

$$x'' - tx = 0, \quad x(0) = 0, \quad x'(0) = 1.$$ □

Exercise 18.36. Consider the linear second order equation

$$x'' + a_1(t)x' + a_2(t)x = 0, \quad t \in I := [\alpha, \beta], \tag{18.5.20}$$

where $a_i : I \to \mathbb{R}$, $i = 1, 2$, are continuous functions. A *zero* of a solution $x(t)$ is a point $t_0 \in I$ such that $x(t_0) = 0$. Prove that the following hold.

(i) The set of zeros of a nonzero solution is at most countable, and contains only isolated points.

(ii) The zero sets of two linearly independent solutions x, y separate each other, i.e. between any two consecutive zeros of x there exists precisely one zero of y. (This is due to *Jacques Sturm* (1803-1855).)

Hint. (ii) Let t_1, t_2 be two consecutive zeros of x. $y(t) \neq 0$ on $[t_1, t_2]$, the function $\varphi(t) = \frac{x(t)}{y(t)}$ is C^1 on $[t_1, t_2]$. Use Rolle's theorem to reach a contradiction. □

Exercise 18.37. Using Problem 18.14, prove that the equation $x'' = a(t)x$ is non-oscillatory, i.e., it admits solutions with only finitely many zeros, it is necessary and sufficient that the Riccati equation

$$y' = -y^2 + a(t)$$

admits solutions defined on the entire semi-axis $[0, \infty)$. □

Exercise 18.38. Consider the second order equations

$$x'' + a(t)x = 0, \tag{18.5.21a}$$

$$x'' + b(t)x = 0, \tag{18.5.21b}$$

where a, b are continuous functions on an interval $I = [t_1, t_2]$. Prove that if $\varphi(t)$ is a solution of (18.5.21a) and $\psi(t)$ is a solution of (18.5.21b), then we have the identity

$$\begin{vmatrix} \varphi(t_2) & \psi(t_2) \\ \varphi'(t_2) & \psi'(t_2) \end{vmatrix} - \begin{vmatrix} \varphi(t_1) & \psi(t_1) \\ \varphi'(t_1) & \psi'(t_1) \end{vmatrix} = \int_{t_1}^{t_2} (\,a(t) - b(t)\,)\varphi(t)\psi(t)dt. \tag{18.5.22}$$

□

Exercise 18.39 (Sturm's comparison theorem)**.** Under the same assumptions as in Problem 18.38, prove that if $a(t) \leqslant b(t)$, $\forall t \in I$, then between any two consecutive zeros of the solution $\varphi(t)$ there exists at least one zero of the solution $\psi(t)$. □

Exercise 18.40. Find all the values of the complex parameter λ such that the boundary value problem

$$x'' + \lambda x = 0, \quad t \in [0, 1], \tag{18.5.23a}$$

$$x(0) = x(1) = 0, \tag{18.5.23b}$$

admits nontrivial solutions. (A boundary value problem as above is known as a *Sturm-Liouville problem*. The corresponding λ's are called the *eigenvalues* of the problem.)

Hint. Prove first that λ has to be a nonnegative real number. Solve (18.5.23a) separately in the case $\lambda = 0$ and $\lambda > 0$ and then impose the condition (18.5.23b) to find than $\lambda = (n\pi)^2$, $n \in \mathbb{Z}_{>0}$. □

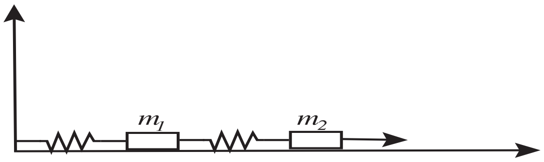

Figure 18.10. An elastic chain.

Exercise 18.41. The differential system

$$\begin{aligned} m_1 x_1'' + \omega_1 x_1 - \omega_2(x_2 - x_1) &= 0, \\ m_2 x_2'' + \omega_2(x_2 - x_1) &= f, \end{aligned} \tag{18.5.24}$$

describes the motion of a mechanical system made of two particles of masses m_1 and m_2 serially connected to a fixed point through two elastic springs with elasticity constants ω_1 and ω_2; see Figure 18.10. Solve the system in the special case $m_1 = m_2 = m$, $\omega_1 = \omega_2 = \omega$ and $f = 0$. □

Exercise 18.42. Solve the differential equation

$$\begin{aligned} x'' + \omega^2 x + \varepsilon^{-1} \min(x, 0) = 0, \quad t \geqslant 0, \\ x(0) = x_0, \quad x'(0) = 0, \end{aligned} \tag{18.5.25}$$

where $x_0 \geqslant 0$, $\varepsilon > 0$ and investigate the behavior of the solution as $\varepsilon \searrow 0$.

Hint. The limit case $\varepsilon = 0$ models the harmonic motion in the presence of an obstacle at $x = 0$. In the limit case $\varepsilon = 0$, the solution formally verifies the system

$$\begin{aligned} \big(x''(t) + \omega^2 x(t) \big) x(t) = 0, \quad \forall t \geqslant 0, \\ x(t) \geqslant 0, \quad x''(t) + \omega^2 x(t) \geqslant 0. \end{aligned}$$

□

Exercise 18.43. Prove that the matrix $X(t) = e^{(\ln t)A}$ is a fundamental matrix of the system $t\boldsymbol{x}' = \boldsymbol{x}$, $t > 0$. □

Exercise 18.44. Prove that for any $n \times n$ matrix A and any $t \in \mathbb{R}$ we have

$$\big(e^{tA}\big)^* = e^{tA^*},$$

where $*$ indicates the transpose of a matrix. □

Exercise 18.45. Prove that the solution $X(t)$ of the matrix-valued differential equation

$$X'(t) = AX(t) + X(t)B, \ \ t \in \mathbb{R},$$

satisfying the initial condition $X(0) = C$ is given by the formula

$$X(t) = e^{tA}Ce^{tB},$$

where A, B, C are $n \times n$ matrices. □

Exercise 18.46. Prove all the entries of e^{tA} are nonnegative for any $t \geqslant 0$ if and only if

$$a_{ij} \geqslant 0, \ \ \forall i \neq j, \tag{18.5.26}$$

where a_{ij} are the entries of the $n \times n$ matrix A.**Hint.** From the formula (18.4.49) we see that that the condition (18.5.26) is necessary. Conversely, if (18.5.26) holds, then there exists $\alpha > 0$ such that all the entries of $\alpha\mathbb{1} + A$ are nonnegative. Next use the equality

$$e^{tA} = e^{-\alpha t}e^{t(\alpha\mathbb{1}+A)}.$$

□

Exercise 18.47. Prove that if (18.5.26) holds, then the solution $\boldsymbol{x}$ of the Cauchy problem

$$\boldsymbol{x}' = A\boldsymbol{x} + \boldsymbol{f}(t), \ \ \boldsymbol{x}(0) = \boldsymbol{x}_0,$$

where $\boldsymbol{x}_0 \geqslant 0$ and $\boldsymbol{f}(t) \geqslant 0$, is nonnegative, i.e., all its components are positive. □

Exercise 18.48. Prove that if the $n \times n$ matrices A, B commute, i.e., $AB = BA$, then

$$e^Ae^B = e^{(A+B)} = e^Be^A.$$

□

Exercise 18.49. Let $\{A_j\}_{j\geqslant 1}$ be a sequence of $n \times n$ matrices that converge in norm to the $n \times n$ matrix A. Prove that

$$\lim_{j\to\infty} e^{tA_j} = e^{tA}, \ \ \forall t \in \mathbb{R}. \tag{18.5.27}$$

□

Exercise 18.50. Let A be a real $n \times n$ matrix, and $D \subset \mathbb{R}^n$ an invariant subspace of A, i.e., $AD \subset D$. Prove that if $\boldsymbol{x}_0 \in D$, then the solution $\boldsymbol{x}(t)$ of the Cauchy problem

$$\boldsymbol{x}' = A\boldsymbol{x}, \ \ \boldsymbol{x}(0) = \boldsymbol{x}_0$$

stays in D for any $t \in \mathbb{R}$. □

Exercise 18.51. Let A be a real $n \times n$ matrix and B an $n \times m$ matrix. Prove that if

$$\operatorname{rank}[B, AB, A^2B, \dots, A^{n-1}B] = n, \tag{18.5.28}$$

then the only vector $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$B^* e^{A^* t} \boldsymbol{x} = 0, \quad \forall t \geqslant 0, \tag{18.5.29}$$

is the null vector. **Hint.** Differentiating the equality (18.5.29) and setting $t = 0$ we deduce

$$B^*\boldsymbol{x} = B^*A^*\boldsymbol{x} = \cdots = B^*(A^*)^{n-1} = 0.$$

The condition (18.5.28) then implies $\boldsymbol{x} = 0$. The condition (18.5.28) was introduced by Kalman and plays an important role in the theory of controllability of linear differential systems. □

Exercise 18.52. Let A be an $n \times n$ matrix. Study the domain of definition of the matrix valued function

$$\sin(tA) = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k+1)!} A^{2k+1},$$

and prove that for any $\boldsymbol{x}_0 \in \mathbb{R}$ the function $\boldsymbol{x}(t) = \sin(tA)\boldsymbol{x}_0$ is a solution of the second order linear differential system $\boldsymbol{x}'' + A^2\boldsymbol{x} = 0$. □

Exercise 18.53. Compute e^{tA} when A is one of the following matrices

$$\begin{bmatrix} 2 & -1 \\ -2 & 3 \end{bmatrix}, \begin{bmatrix} -1 & 0 & 3 \\ -8 & 1 & 12 \\ -2 & 0 & 4 \end{bmatrix}, \begin{bmatrix} -1 & 2 & -1 \\ -1 & -4 & 1 \\ -1 & -2 & -1 \end{bmatrix}.$$

□

Exercise 18.54. Prove that using the substitution $t = e^\tau$ the *Euler equation*

$$t^n x^{(n)} + a_1 t^{n-1} x^{(n-1)} + \cdots + a_n x = 0,$$

where $a_1, \dots, a_n$ are real constants, reduces to a linear differential equation of order n with constant coefficients. □

Exercise 18.55. Prove that if A, B are $m \times m$ real matrices the we have *Lie-Trotter formula*

$$e^{t(A+B)} = \lim_{n\to\infty} \left(e^{\frac{t}{n}A} e^{\frac{t}{n}B} \right)^n, \quad \forall t \geqslant 0.$$

Hint. For any positive integer n the matrix

$$Y_n(t) = \left(e^{\frac{t}{n}A} e^{\frac{t}{n}B} \right)^n$$

satisfies the differential equation

$$Y_n'(t) = (A+B)Y_n(t) + \left(e^{\frac{t}{n}A} B e^{-\frac{t}{n}A} - B \right) \left(e^{\frac{t}{n}A} e^{\frac{t}{n}B} \right)^{n-1},$$

from which we can conclude that $Y_n(t) \to e^{t(A+B)}$ as $n \to \infty$. □

Chapter 19

Measure theory and integration

The goal of this chapter is to present the modern technique of integration pioneered by H. Lebesgue that considerably extends the reach of the classical Riemann integral. While the Riemann integration could be performed only over subsets of some Euclidean space, the new process can be performed over abstract sets as long as we can attach a concept of "volume" to certain of its subsets. Measure theory clarifies this last vaguely phrased requirement and the first half of this chapter is devoted to developing this theory. The second half is devoted to the construction of the new integral and describing some of its more salient features.

For any set Ω we denote by 2^{Ω} the collection of all the subsets of Ω. For $S \subset \Omega$ we denote by $\boldsymbol{I}_S$ its indicator function

$$\boldsymbol{I}_S : \Omega \to \{0,1\},\;\; \boldsymbol{I}_S(\omega) = \begin{cases} 1, & \omega \in S, \\ 0, & \omega \in \Omega \backslash S. \end{cases}$$

For any $S \subset \Omega$ we denote by S^c its complement, $S^c := \Omega \backslash S$.

If $F : X \to Y$ is a map between two sets X, Y and $A \subset Y$ then we set

$$\{F \in A\} := F^{-1}(A) \subset X.$$

Note that

$$\boldsymbol{I}_{\{F \in A\}}(x) = \boldsymbol{I}_A \circ F(x) = \boldsymbol{I}_A\big(F(x) \big).$$

In particular, if $F : X \to \mathbb{R}$ and $a, b \in \mathbb{R}$, then

$$\{F \leqslant a\} = \{F \in (-\infty, a]\},\;\; \{a \leqslant F \leqslant b\} = \{F \in [a,b]\} \subset X \;\text{ etc.}$$

19.1. Measurable spaces and measures

The Lebesgue technique of integration requires a choice of a measure. Intuitively, a measure assigns to a subset of a given set a nonnegative real number; think length, area, cardinality. This section is devoted to clarifying the concept of measure.

19.1.1. Sigma-algebras.

Fix a nonempty set Ω.

Definition 19.1.1. (a) A collection $\mathcal{R}$ of subsets of Ω is called a *ring of subsets* of Ω if it satisfies the following conditions

(i) $\forall A, B \in \mathcal{R}$, $A \cap B, A \cup B \in \mathcal{R}$.

(ii) $\forall A, B \in \mathcal{R}$, $A \subset B \Rightarrow B \backslash A \in \mathcal{R}$.

(b) A collection $\mathcal{A}$ of subsets of Ω is called an *algebra* of subsets of Ω if it is a ring and $\Omega \in \mathcal{A}$.

(c) A collection $\mathcal{S}$ of subsets of Ω is called a *σ-algebra* (or *sigma-algebra*) of Ω if it is an algebra of Ω and the union of any countable subfamily of $\mathcal{S}$ is a set in $\mathcal{S}$, i.e.,

$$\forall (A_n)_{n\in\mathbb{N}} \in \mathcal{S}^{\mathbb{N}}, \quad \bigcup_{n\geqslant 1} A_n \in \mathcal{S}. \tag{19.1.1}$$

(d) A *measurable space* is a pair $(\Omega, \mathcal{S})$ where $\mathcal{S}$ is a sigma-algebra of subsets of Ω. A set $S \in \mathcal{S}$ is called $\mathcal{S}$-*measurable.* □

Remark 19.1.2. To prove that an algebra $\mathcal{S}$ is a σ-algebra is suffices to verify (19.1.1) *only for increasing* sequence of subsets $A_n \in \mathcal{S}$. Indeed, if $(A_n)_{n\in\mathbb{N}}$ is any sequence of subsets in $\mathcal{S}$, then for any n we have

$$\bar{A}_n = A_1 \cup \cdots \cup A_n \in \mathcal{S}$$

since $\mathcal{S}$ is an algebra. The new sequence $(\bar{A}_n)_{n\in\mathbb{N}}$ is increasing and

$$\bigcup_{n\in\mathbb{N}} \bar{A}_n = \bigcup_{n\in\mathbb{N}} A_n.$$

□

Example 19.1.3. Let us describe a few classical examples of sigma algebras.

(i) The collection 2^{Ω} of all subsets of Ω is obviously a σ-algebra.

(ii) Suppose that $\mathcal{S}$ is a (σ-)algebra of a set Ω and $F : \widehat{\Omega} \to \Omega$ is a map. Then the preimage

$$F^{-1}(\mathcal{S}) = \big\{ F^{-1}(S); \quad S \in \mathcal{S} \big\}$$

is a (σ-)algebra of subsets of $\widehat{\Omega}$. The σ-algebra $F^{-1}(\mathcal{S})$ is denoted by $\sigma(F)$ and it is called the *σ-algebra generated by* F or the *pullback of* $\mathcal{S}$ *via* F.

(iii) Given $A \subset \Omega$ we denote by $\mathcal{S}_A$ the *σ-algebra generated by* A, i.e.,

$$\mathcal{S}_A = \big\{ \varnothing, A, A^c, \Omega \big\}.$$

We will refer to it as the *Bernoulli algebra* with success A. Note that $\mathcal{S}_A$ is the pullback of $\mathscr{2}^{\{0,1\}}$ via the indicator function $\boldsymbol{I}_A : \Omega \to \{0,1\}$.

(iv) If $(\mathcal{S}_i)_{i\in I}$ is a family of (σ-)algebras of Ω, then their intersection

$$\bigcap_{i\in I} \mathcal{S}_i \subset \mathscr{2}^{\Omega}$$

is a (σ-)algebra of Ω.

(v) If $\mathscr{C} \subset \mathscr{2}^{\Omega}$ is a family of subsets of Ω, then we denote by $\sigma(\mathscr{C})$ the σ-algebra generated by $\mathscr{C}$, i.e., the intersection of all σ-algebras that contain $\mathscr{C}$. In particular, if $\mathcal{S}_1, \mathcal{S}_2$ are σ-algebras of Ω, then we set

$$\mathcal{S}_1 \vee \mathcal{S}_2 := \sigma(\mathcal{S}_1 \cup \mathcal{S}_2).$$

More generally, for any family $(\mathcal{S}_i)_{i\in I}$ of σ-algebras we set

$$\bigvee_{i\in I} \mathcal{S}_i := \sigma\left(\bigcup_{i\in I} \mathcal{S}_i\right).$$

(vi) Suppose that we are given a countable partition $\boldsymbol{P} = \{A_n\}_{n\in\mathbb{N}}$ of Ω,

$$\Omega = \bigsqcup_{n\in\mathbb{N}} A_n.$$

Then the σ-algebra generated by this partition, denoted by $\sigma(\boldsymbol{P})$, is the sigma consisting of all the subsets of Ω that are unions of the sets A_n. This σ-algebra can be viewed as the σ-algebra generated by the map

$$X : \Omega \to \mathbb{N}, \;\; X = \sum_{n\in\mathbb{N}} n\boldsymbol{I}_{A_n}.$$

More precisely $\sigma(\boldsymbol{P}) = X^{-1}\big(\mathscr{2}^{\mathbb{N}}\big)$, so that $A_n = X^{-1}\big(\{n\}\big)$.

(vii) If $(\Omega_1, \mathcal{S}_1)$ and $(\Omega_2, \mathcal{S}_2)$ are two measurable spaces, then we denote by $\mathcal{S}_1 \otimes \mathcal{S}_2$ the sigma algebra of $\Omega_1 \times \Omega_2$ generated by the collection of *rectangles*

$$\big\{S_1 \times S_2 : \;\; S_1 \in \mathcal{S}_1, \; S_2 \in \mathcal{S}_2\,\big\} \subset \mathscr{2}^{\Omega_1\times\Omega_2}.$$

(viii) If X is a metric space and $\mathcal{T}_X \subset \mathscr{2}^X$ denotes the family of open subsets, then the *Borel* σ-algebra of X, denoted by $\mathcal{B}_X$, is the σ-algebra generated by $\mathcal{T}_X$. The sets in $\mathcal{B}_X$ are called the *Borel subsets of* X.

The real axis $\mathbb{R}$ is naturally a metric space. The associated Borel algebra $\mathcal{B}_{\mathbb{R}}$ can equivalently described as the sigma-algebra generated by the collection of semiaxes

$$(-\infty, a], \;\; a \in \mathbb{R}$$

Note that since any open set in $\mathbb{R}^n$ is a countable union of open cubes we have

$$\mathcal{B}_{\mathbb{R}^n} = \mathcal{B}_{\mathbb{R}}^{\otimes n}. \tag{19.1.2}$$

(ix) We set $\bar{\mathbb{R}} = [-\infty, \infty]$. The Borel algebra of $\bar{\mathbb{R}}$ is the sigma-algebra generated by the intervals

$$[-\infty, a], \quad a \in [-\infty, \infty].$$

For simplicity we will refer to the Borel subsets of $\bar{\mathbb{R}}$ simply as *Borel sets.*

(x) If $(\Omega, \mathcal{S})$ is a measurable space and $X \subset \Omega$, then the collection

$$\mathcal{S}|_X := \{\, S \cap X : \quad S \in \mathcal{S} \,\} \subset 2^X$$

is a σ-algebra of X called *the trace of* $\mathcal{S}$ *on* X.

Equivalently, consider the natural inclusion $i_X : X \to \Omega$, $i_X(x) = x$, $\forall x \in X$. Then $\mathcal{S}|_X$ coincides with the pullback of $\mathcal{S}$ via i_X; see (ii).

□

Remark 19.1.4 (Human language conversion)**.** Suppose that we are given a family of subsets $(S_i)_{i \in I}$ of a set Ω. Let us observe that the statement

$$\omega \in \bigcap_{i \in I} S_i$$

translates into the formula $\forall i \in I$, $\omega \in S_i$ or, in human language, "ω belongs to *all* of the sets in the family". The statement

$$\omega \in \bigcup_{i \in I} S_i$$

translates into the formula $\exists i \in I$, $\omega \in S_i$ or, in human language, "ω belongs to *at least one* of the sets S_i".

For example, the statement

$$\omega \in \bigcup_{n \in \mathbb{N}} \bigcap_{k \geqslant n} S_k$$

translates into

$$\exists n \in \mathbb{N}, \quad \forall k \geqslant n : \quad \omega \in S_k.$$

Equivalently, this means that ω belongs to all but finitely many of the sets S_k.

Conversely, statements involving the quantifiers $\exists, \forall$ can be translated into set theoretic statements using the conversion rules

$$\exists \to \cup, \quad \forall \to \cap.$$

□

Definition 19.1.5. Let $\mathscr{C}$ be a collection of subsets of a set Ω. We say that $\mathscr{C}$ is a *π-system* if it is closed under finite intersections, i.e.,

$$\forall A, B \in \mathscr{C} : \quad A \cap B \in \mathscr{C}.$$

The collection $\mathscr{C}$ is called a *λ-system* if it satisfies the following conditions.

(i) $\varnothing, \Omega \in \mathscr{C}$.

(ii) if $A, B \in \mathscr{C}$ and $A \subset B$, then $B \backslash A \in \mathscr{C}$.

(iii) If $A_1 \subset A_2 \subset \cdots$ belong to $\mathscr{C}$, then so does their union.

□

Lemma 19.1.6. *Suppose that $\mathscr{C}$ is a collection of subsets of a set Ω. Then the following are equivalent.*

(i) *$\mathscr{C}$ is a σ-algebra.*

(ii) *$\mathscr{C}$ is both a λ- and a π-system.*

Proof. The implication (i) $\Rightarrow$ (ii) is obvious. Let us prove the converse. It suffices to prove that if $\mathscr{C}$ is both a λ- and π-system and $(A_n)_{n\in\mathbb{N}}$ is a sequence in $\mathscr{C}$, then

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{C}.$$

Indeed, set

$$B_n = A_1 \cup \cdots \cup A_n.$$

Note that $A_k^c = \Omega\backslash A_k \in \mathscr{C}$. Hence $A_1^c \cap \cdots \cap A_n^c \in \mathscr{C}$, since $\mathscr{C}$ is a π-system. We deduce that

$$B_n = A_1 \cup \cdots \cup A_n = \big(A_1^c \cap \cdots \cap A_n^c \big)^c = \Omega\backslash\big(A_1^c \cap \cdots \cap A_n^c \big) \in \mathscr{C}.$$

Now observe that $B_1 \subset B_2 \subset \cdots$ and, since $\mathscr{C}$ is a λ-system, we have

$$\bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n \in \mathscr{C}.$$

□

Since the intersection of any family of λ-systems is a λ-system we deduce that for any collection $\mathscr{C} \subset 2^{\Omega}$ there exists a smallest λ-system containing $\mathscr{C}$. We denote this system by $\Lambda(\mathscr{C})$ and we will refer to it as the *λ-system generated by $\mathscr{C}$*.

Example 19.1.7. Suppose that $\mathcal{H}$ is the collection of half-infinite intervals

$$(-\infty, x], \quad x \in \mathbb{R}.$$

Then $\mathcal{H}$ is π-system of $\mathbb{R}$. The λ-system generated by $\mathcal{H}$ contains all the open intervals. Since any open subset of $\mathbb{R}$ is a countable union of open intervals we deduce that $\Lambda(\mathcal{H})$ coincides with the Borel σ-algebra $\mathcal{B}_{\mathbb{R}}$. □

Theorem 19.1.8 (Dynkin's $\pi - \lambda$ theorem)**.** *Suppose that $\mathcal{P}$ is a π-system. Then $\Lambda(\mathcal{P}) = \sigma(\mathcal{P})$. In other words, any λ-system that contains $\mathcal{P}$, also contains the σ-algebra generated by $\mathcal{P}$.*

Proof. Since any σ-algebra is a λ-system we deduce $\Lambda(\mathcal{P}) \subset \sigma(\mathcal{P})$. Thus it suffices to show that $\sigma(\mathcal{P}) \subset \Lambda(\mathcal{P})$. Equivalently, it suffices to show that $\Lambda(\mathcal{P})$ is a σ-algebra. This happens if and only if the λ-system $\Lambda(\mathcal{P})$ is also a π-system. Hence it suffices to show that $\Lambda(\mathcal{P})$ is closed under (finite) intersections.

For $A \in \Lambda(\mathscr{P})$ we set

$$\mathscr{L}_A := \big\{ B \in \mathscr{2}^{\Omega} : \;\; B \cap A \in \Lambda(\mathscr{P}) \big\}.$$

Note $\Lambda(\mathscr{P})$ is a π-system iff

$$\Lambda(\mathscr{P}) \subset \mathscr{L}_A, \;\; \forall A \in \Lambda(\mathscr{P}). \tag{19.1.3}$$

Observe first that $\mathscr{L}_A$ is a λ-system. Indeed, $\Omega \in \mathscr{L}_A$ since $A \in \Lambda(\mathscr{P})$. Obviously $\varnothing \in \mathscr{L}_A$ since

$$\varnothing \cap A = \varnothing = A \backslash A \in \Lambda(\mathscr{P}).$$

If $B_1 \subset B_2$ are in $\mathscr{L}_A$, then $B_1 \cap A \subset B_2 \cap A$ are in $\Lambda(\mathscr{P})$. Since $\Lambda(\mathscr{P})$ is a λ-system we deduce

$$(B_2 \backslash B_1) \cap A = (B_2 \cap A) \backslash (B_1 \cap A) \in \Lambda(\mathscr{P}).$$

This proves that $\mathscr{L}_A$ satisfies property (ii) in the definition of a λ-system. The property (iii) is proved in a similar fashion relying on the fact that $\Lambda(\mathscr{P})$ is a λ-system.

Since $\mathscr{P}$ is a π-system, if $A \in \mathscr{P}$, then $\mathscr{P} \subset \mathscr{L}_A$. Hence In particular $\Lambda(\mathscr{P})$ is contained in the λ-system $\mathscr{L}_A$, $\forall A \in \mathscr{P}$. Thus, if $A \in \mathscr{P}$ and $B \in \Lambda(\mathscr{P})$, then $B \in \mathscr{L}_A$, i.e., $A \cap B \in \Lambda(\mathscr{P})$. Hence

$$\mathscr{P} \subset \mathscr{L}_B, \;\; \forall B \in \Lambda(\mathscr{P}) \Rightarrow \Lambda(\mathscr{P}) \subset \mathscr{L}_B, \;\; \forall B \in \Lambda(\mathscr{P}).$$

This proves (19.1.3) and completes the proof of the $\pi - \lambda$ theorem. □

Remark 19.1.9. The typical applications of the $\pi - \lambda$ theorem are of the following form. We need to show that all the sets of a σ-algebra $\mathscr{S}$ of Ω of a property $\mathscr{X}$. Denote by $\mathscr{S}(\mathscr{X})$ the collection of subsets $S \subset \Omega$ that enjoy the property $\mathscr{X}$. According to the $\pi - \lambda$ theorem, in order to prove that $\mathscr{S}(\mathscr{X}) \supset \mathscr{S}$ it suffices to show two things.

(i) $\mathscr{S}(\mathscr{X})$ is a λ-system.

(ii) There exists a π-system $\mathscr{P}$ generating $\mathscr{S}$ and contained in $\mathscr{S}(\mathscr{X})$.

Indeed, if (i) and (ii) are satisfied, then $\mathscr{S}(X) \supset \Lambda(\mathscr{P}) = \sigma(\mathscr{P}) = \mathscr{S}$. The essence of (i) is that the collection of sets in $\mathscr{S}$ satisfying property $\mathscr{X}$ is closed under monotone limits. To establish (ii) we need to supply enough examples of sets in $\mathscr{S}$ satisfying this property. □

19.1.2. Measurable maps.

Definition 19.1.10. A map $F : \Omega_1 \to \Omega_2$ is called *measurable* with respect to the σ-algebras $\mathscr{S}_i$ on Ω_i, $i = 1, 2$ or $(\mathscr{S}_1, \mathscr{S}_2)$-*measurable* if $F^{-1}(\mathscr{S}_2) \subset \mathscr{S}_1$, i.e.,

$$F^{-1}(S_2) \in \mathscr{S}_1, \;\; \forall S_2 \in \mathscr{S}_2.$$

Two measurable spaces $(\Omega_i, \mathscr{S}_i)$, $i = 1, 2$, are called *isomorphic* if there exists a bijection $F : \Omega_1 \to \Omega_2$ such that $F^{-1}(\mathscr{S}_2) = \mathscr{S}_1$ or, equivalently, both F and its inverse F^{-1} are measurable. □

Definition 19.1.11. Suppose that $(\Omega, \mathscr{S})$ is a measurable space. A function $f : \Omega \to \bar{\mathbb{R}}$ is called $\mathscr{S}$-*measurable* if, for any Borel subset $B \subset \bar{\mathbb{R}}$ we have $f^{-1}(B) \in \mathscr{S}$. □

Example 19.1.12. (a) The composition of two measurable maps is a measurable map.

(b) A subset $S \subset \Omega$ is $\mathcal{S}$-measurable if and only if the indicator function $\boldsymbol{I}_S$ is a measurable function.

(c) If $\mathcal{A}$ is the σ-algebra generated by a finite or countable partition

$$\Omega = \bigsqcup_{i \in I} A_i, \quad I \subset \mathbb{N},$$

then a function $f : \Omega \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is $\mathcal{A}$-measurable if and only if it is constant in the chambers A_i of this partition. □

Proposition 19.1.13. *Consider a map $F : (\Omega_1, \mathcal{S}_1) \to (\Omega_2, \mathcal{S}_2)$ between two measurable spaces. Suppose that $\mathscr{C}_2$ is a π-system of Ω_2 such that $\sigma(\mathscr{C}_2) = \mathcal{S}_2$. Then the following statements are equivalent.*

(i) *The map F is measurable.*

(ii) *$F^{-1}(C) \in \mathcal{S}_1$, $\forall C \in \mathscr{C}_2$.*

Proof. Clearly (i) ⇒ (ii). The opposite implication follows from the $\pi - \lambda$ theorem since the set

$$\{ C \in \mathcal{S}_2; \;\; F^{-1}(C) \in \mathcal{S}_1 \}$$

is a λ-system containing the π-system $\mathscr{C}_2$. □

Corollary 19.1.14. *If $F : X \to Y$ is a continuous map between metric spaces, then it is $(\mathcal{B}_X, \mathcal{B}_Y)$ measurable.*

Proof. Denote by $\mathcal{T}_Y$ the collection of open subsets of Y. Then $\mathcal{T}_Y$ is a π-system and, by definition, it generates $\mathcal{B}_Y$. Since F is continuous, for any $U \in \mathcal{T}_Y$ the set $F^{-1}(U)$ is open in X and thus belongs to $\mathcal{B}_X$. □

Corollary 19.1.15. *Let $(\Omega, \mathcal{S})$ be a measurable space. A function $X : \Omega \to \mathbb{R}$ is $(\mathcal{S}, \mathcal{B}_{\mathbb{R}})$-measurable if and only if the sets $X^{-1}\big((-\infty, x] \big)$ are $\mathcal{S}$-measurable for any $x \in \mathbb{R}$.*

Proof. It follows from Proposition 19.1.13 by observing that the collection

$$\{ (-\infty, x]; \;\; x \in \mathbb{R} \} \subset 2^{\mathbb{R}}$$

is a π-system and the σ-algebra it generates is $\mathcal{B}_{\mathbb{R}}$.

□

Corollary 19.1.16. *Consider a pair of maps between measurable spaces*

$$F_i : (\Omega, \mathcal{S}) \to (\Omega_i, \mathcal{S}_i), \quad i = 1, 2.$$

Then the following statements are equivalent.

(i) *The maps F_i are measurable.*

(ii) *The map*

$$F_1 \times F_2 : \Omega \to \Omega_1 \times \Omega_2,\ \ \omega \mapsto \big(\, F_1(\omega), F_2(\omega)\,\big)$$

is $(\mathcal{S}, \mathcal{S}_1 \otimes \mathcal{S}_2)$*-measurable.*

Proof. (i) ⇒ (ii) Observe that if the maps F_1, F_2 are measurable then

$$F_1^{-1}(S_1),\ F_2^{-1}(S_2) \in \mathcal{S},\ \ \forall S_1 \in \mathcal{S}_1,\ \ S_2 \in \mathcal{S}_2$$

$$\Rightarrow\ (F_1 \times F_2)^{-1}(S_1 \times S_2) = F_1^{-1}(S_1) \cap F_2^{-1}(S_2) \in \mathcal{S},\ \ \forall S_1 \in \mathcal{S}_1,\ \ S_2 \in \mathcal{S}_2.$$

Since the collection $S_1 \times S_2$, $S_i \in \mathcal{S}_i$, $i = 1, 2$, is a π-system that, by definition, generates $\mathcal{S}_1 \otimes \mathcal{S}_2$ we see that the last statement is equivalent with the measurability of $F_1 \times F_2$.

(ii) ⇒ (i) For $i = 1, 2$ we denote by π_i the natural projection

$$\Omega_1 \times \Omega_2 \to \Omega_i,\ \ (\omega_1, \omega_2) \mapsto \omega_i.$$

The maps π_i are $(\mathcal{S}_1 \otimes \mathcal{S}_2, \mathcal{S}_i)$ measurable and

$$F_i = \pi_i \circ (F_1 \times F_2).$$

□

Definition 19.1.17. For any measurable space $(\Omega, \mathcal{S})$ we denote by $\bar{\mathcal{L}}^0(\Omega) = \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ the space of measurable functions $\Omega \to \bar{\mathbb{R}}$, and by $\mathcal{L}^0(\Omega, \mathcal{S})$ the space of measurable functions $(\Omega, \mathcal{S}) \to \mathbb{R}$.

The subset of $\bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ (resp. $\mathcal{L}^0_+(\Omega, \mathcal{S})$) consisting of nonnegative functions is denoted by $\bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ (resp $\mathcal{L}^0_+(\Omega, \mathcal{S})$), while the subspace of $\mathcal{L}^0(\Omega, \mathcal{S})$ consisting of bounded measurable functions is denoted $\mathcal{L}^\infty(\Omega, \mathcal{S})$. □

Remark 19.1.18. The algebraic operations "+" and "·" on $\mathbb{R}$ admit (partial) extensions to $\bar{\mathbb{R}}$,

$$c \pm \infty = \pm\infty, \infty + \infty = \infty,\ \ c \cdot \infty = \infty,\ \ \forall c > 0.$$

As we know, there are a few "illegal" operations

$$\infty - \infty,\ \ 0 \cdot \infty,\ \ \frac{0}{0}\ \text{ etc.}$$ □

Proposition 19.1.19. *Fix a measurable space* $(\Omega, \mathcal{S})$*. Then the following hold.*

(i) *For any* $f, g \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ *and any* $c \in \mathbb{R}$ *we have*

$$f + g,\ fg,\ cf \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S}),$$

whenever these functions are well defined.

(ii) *Let* $(f_n)_{n \in \mathbb{N}}$ *be a sequence in* $\bar{\mathcal{L}}^0(\Omega, \mathcal{S})$*. For any* $\omega \in \Omega$ *we set*

$$m(\omega) = \inf_{n \in \mathbb{N}} f_n(\omega) \in [-\infty, \infty],\ \ M(\omega) = \sup_{n \in \mathbb{N}} f_n(\omega) \in (-\infty, \infty].$$

Then $m, M \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$*.*

(iii) *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\bar{\mathcal{L}}^0(\Omega,\mathcal{S})$. Then*

$$\liminf_n f_n := \sup_m \inf_{n\geqslant m} f_n \;\; and \;\; \limsup_n f_n = \inf_m \sup_{n\geqslant m} f_n$$

are measurable.

(iv) *If $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\bar{\mathcal{L}}^0(\Omega,\mathcal{S})$ such that, for any $\omega\in\Omega$ the limit*

$$f_\infty(\omega) = \lim_{n\to\infty} f_n(\omega)$$

exists, then $f_\infty : \Omega \to \bar{\mathbb{R}}$ is also $\mathcal{S}$-measurable.

Proof. (i) Denote by $\mathcal{D}$ the subset of $\bar{\mathbb{R}}^2$ consisting of the pairs (x,y) for which $x+y$ is well defined. Observe that $f+g$ is the composition of two measurable maps

$$\Omega \to \mathcal{D} \subset \bar{\mathbb{R}}^2,\;\; \omega \mapsto \big(f(\omega), g(\omega) \big),\;\; \mathcal{D} \to \bar{\mathbb{R}},\;\; (x,y) \mapsto x+y.$$

Above, the first map is measurable according to Corollary 19.1.16 and the second map is Borel measurable since it is continuous. The measurability of fg and cf is established in a similar fashion.

(ii) Let us prove first that M is measurable. It suffices to prove that for any $x\in\bar{R}$ the set $\{M(\omega)\leqslant x\}$ is $\mathcal{S}$-measurable. To achieve this we will use the human language conversion procedure discussed in Remark 19.1.4,

$$\exists \to \cup \;\text{ and }\; \forall \to \cap.$$

Note that

$$M(\omega)\leqslant x \Longleftrightarrow \forall n\in\mathbb{N}:\;\; f_n(\omega)\leqslant x$$

Equivalently

$$\big\{ M\leqslant x \big\} = \bigcap_{n\in\mathbb{N}} \{f_n\leqslant x\} \in \mathcal{S}.$$

Similarly, to prove that m is measurable it suffices to show that the sets $\{m\geqslant x\}$ are measurbale for any $x\in\bar{\mathbb{R}}$. The details are left to the reader.

(iii) We set

$$A_m := \inf_{n\geqslant m} f_n,\;\; B_m := \sup_{n\geqslant m} f_n.$$

By (ii), these are measurable functions and so are

$$\limsup_n f_n = \sup_m A_m,\;\; \liminf_b f_n = \inf_m B_m.$$

(iv) In this case

$$f_\infty = \liminf_n f_n = \limsup_n f_n$$

and (iii) implies that f_∞ is measurable □

Corollary 19.1.20. *For any function $f\in\bar{\mathcal{L}}^0(\Omega,\mathcal{S})$, its positive and negative parts,*

$$f^+ := \max(f,0),\;\; f^- := \max(-f,0)$$

belong to $\bar{\mathcal{L}}^0_+(\Omega,\mathcal{S})$ as well. □

Proof. The function f^+ is the composition of two measurable functions

$$\Omega \xrightarrow{f} \bar{\mathbb{R}},\ \ \bar{\mathbb{R}} \ni x \mapsto \max(x,0) \in \bar{\mathbb{R}},$$

so it is measurable. A similar argument shows that f^- is measurable. □

A function $f \in \mathcal{L}^0(\Omega, \mathcal{S})$ is called *elementary* or *step function* if its range is a *finite* subset of $\mathbb{R}$. More concretely, this means there exist *finitely many disjoint measurable sets*

$$A_1, \dots, A_N \in \mathcal{S}$$

and real numbers $c_1, \dots, c_N$ such that

$$f(\omega) = \sum_{k=1}^{N} c_k \boldsymbol{I}_{A_k}(\omega),\ \ \forall \omega \in \Omega. \tag{19.1.4}$$

We denote by $\mathscr{E}(\Omega, \mathcal{S})$ the space of elementary functions and by $\mathscr{E}_+(\Omega, \mathcal{S})$ the subspace consisting of nonnegative ones.

Define $D_n : [0,\infty] \to [0,\infty]$,

$$D_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \boldsymbol{I}_{[(k-1)2^{-n}, k2^{-n})}(r) + n\boldsymbol{I}_{[n,\infty)}(r) = \min\left(\frac{\lfloor 2^n r \rfloor}{2^n}, n \right), \tag{19.1.5}$$

where $\lfloor x \rfloor$ denotes the largest integer $\leqslant x$. Let us observe that if $r \in [0,1)$ has binary expansion

$$r = 0.\epsilon_1 \epsilon_2 \cdots \epsilon_n \cdots,\ \ \epsilon_i = 0,1,$$

where binary expansion does not have an infinite tail of consecutive 1-s. Then

$$D_n(r) = 0.\epsilon_1 \cdots \epsilon_n = \sum_{k=1}^{n} \frac{\epsilon_n}{2^n}.$$

Lemma 19.1.21. *The function D_n is right-continuous and nondecreasing. Moreover*

$$D_n(r) \leqslant D_{n+1}(r),\ \ \forall n \in \mathbb{N},\ \ r \in [0,\infty]. \tag{19.1.6}$$

and

$$\lim_{n\to\infty} D_n(r) = r,\ \ \forall r \in [0,\infty].$$

Proof. The first claim follows from the definition (19.1.5). Let us shows that the sequence $(\,D_n(r)\,)$ is nondecreasing. We distinguish several cases.

Case 1. If $r \in [(k-1)2^{-n}, k2^{-n})$, then either

$$r \in \big[\,(k-1)2^{-n}, (k-1)2^{-n} + 2^{-(n+1)}\,\big),$$

or

$$r \in [(k-1)2^{-n} + 2^{-(n+1)}, k2^{-n}).$$

Hence

$$D_{n+1}(r) = D_n(r) \text{ or } D_{n+1}(r) = D_n(r) + 2^{-(n+1)}.$$

Case 2. If $r \in [n, n+1)$, then $D_n(r) = n \leqslant D_{n+1}(r)$.

Case 3. If $r \geqslant n+2$, then $D_n(r) = n < n+1 = D_{n+1}(r)$. □

Using the function D_n we obtain a sequence of transformations

$$D_n : \mathcal{L}^0_+(\Omega, \mathcal{S}) \to \mathscr{E}_+(\Omega, \mathcal{S}), \quad f \mapsto D_n[f], \quad D_n[f](\omega) = D_n\big(f(\omega)\big), \quad \forall \omega \in \Omega.$$

Lemma 19.1.22. *The transformations D_n enjoy the following properties.*

(i) *For any $n \in \mathbb{N}$ the transformation $D_n[-]$ is* monotone*, i.e., for any $f, g \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ such that $f \leqslant g$ we have*

$$D_n[f] \leqslant D_n[g].$$

(ii) *For any $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ the sequence of elementary functions $(D_n[f])$ is nondecreasing and converges everywhere to f, i.e.,*

$$\lim_{n\to\infty} D_n[f](\omega) = f(\omega), \quad \forall \omega \in \Omega.$$

(iii) *For any $n \in \mathbb{N}$ the transformation $D_n[-]$ is* local*, i.e., for any $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ and any $S \in \mathcal{S}$ we have*

$$D_n[f\boldsymbol{I}_S] = D_n[f]\boldsymbol{I}_S.$$

Proof. The properties (i) and (ii) follow directly from Lemma 19.1.21. To prove (iii) observe that if $S \in \mathcal{S}$, then for $\omega \in S$ we have

$$D_n[f\boldsymbol{I}_S](\omega) = D_n\big(f(\omega)\big) = I_S(\omega)D_n[f](\omega),$$

and, for $\omega \in S^c$ we have

$$D_n[f\boldsymbol{I}_S](\omega) = D_n(0) = 0 = I_S(\omega)D_n[f](\omega).$$

□

Corollary 19.1.23. *Any nonnegative measurable function $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ is the limit of a nondecreasing sequence of elementary functions.* □

Definition 19.1.24. Let $(\Omega, \mathcal{S})$ be a measurable space. A collection $\mathcal{M}$ of $\mathcal{S}$-measurable functions $f : \Omega \to (-\infty, \infty]$ is called a *monotone class* of $(\Omega, \mathcal{S})$ if it satisfies the following conditions.

(i) $\boldsymbol{I}_\Omega \in \mathcal{M}$.

(ii) If $f, g \in \mathcal{M}$ are *bounded* and $a, b \in \mathbb{R}$, then $af + bg \in \mathcal{M}$.

(iii) If (f_n) is an increasing sequence of nonnegative random variables in $\mathcal{M}$ with finite limit f_∞, then $f_\infty \in \mathcal{M}$.

□

Theorem 19.1.25 (Monotone Class Theorem)**.** *Suppose that $\mathcal{M}$ is a monotone class of the measurable space $(\Omega, \mathcal{S})$ and $\mathscr{C}$ is a π-system that generates $\mathcal{S}$ and such that $\boldsymbol{I}_C \in \mathcal{M}$, $\forall C \in \mathscr{C}$. Then $\mathcal{M}$ contains $\mathcal{L}^\infty(\Omega, \mathcal{S})$ and all the nonnegative $\mathcal{S}$-measurable functions $\Omega \to [0, \infty]$.*

Proof. Observe that the collection

$$\mathcal{A} := \left\{ A \in \mathcal{S} : \ \boldsymbol{I}_A \in \mathcal{M} \right\}$$

is a λ-system.

Indeed $\boldsymbol{I}_\Omega$ and $\boldsymbol{I}_\varnothing = \boldsymbol{I}_\Omega - \boldsymbol{I}_\Omega \in \mathcal{M}$. If $A \subset B$ are in $\mathcal{A}$, then $B \backslash A \in \mathcal{A}$ since $\boldsymbol{I}_{B\backslash A} = \boldsymbol{I}_B - \boldsymbol{I}_A \in \mathcal{M}$. Finally, if $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence in $\mathcal{A}$ and

$$A_\infty = \bigcup_{n\in\mathbb{N}} A_n,$$

then $(\boldsymbol{I}_{A_n})$ is an increasing sequence of nonnegative functions in $\mathcal{M}$ and thus

$$\boldsymbol{I}_{A_\infty} = \lim_{n\to\infty} \boldsymbol{I}_{A_n} \in \mathcal{M},$$

so that $A_\infty \in \mathcal{A}$.

Hence $\mathcal{A}$ is a λ-system containing $\mathscr{C}$ and the $\pi - \lambda$ theorem implies that $\mathcal{A}$ contains $\sigma(\mathscr{C}) = \mathcal{S}$.

Thus $\mathcal{M}$ contains all the elementary functions. Since any nonnegative measurable function is an increasing pointwise limit of elementary functions we deduce that $\mathcal{M}$ contains all the nonnegative measurable functions. Finally, if f is a bounded measurable function, then f^+, f^- are nonnegative and bounded measurable functions so $f^+, f^- \in \mathcal{M}$ and thus $f = f^+ - f^- \in \mathcal{M}$.

□

Remark 19.1.26. The Monotone Class Theorem is a very versatile tool for proving general statements about measurable functions. Suppose that we want to prove that all the finite measurable functions satisfy a certain property **P**. Then it suffices to show the following.

(i) The constant function satisfies **P**

(ii) If f, g are nonnegative and satisfy **P** then $-f$ and $af + bg$ satisfy **P**, $\forall a, b \geqslant 0$.

(iii) The limit of an increasing sequence of functions satisfying **P** also satisfies **P**.

(iv) For any set A of a π-system $\mathscr{C}$ that generates $\mathcal{S}$, the indicator $\boldsymbol{I}_A$ satisfies **P**.

□

Theorem 19.1.27 (Dynkin)**.** *Suppose that $F : (\Omega, \mathcal{S}) \to (\Omega', \mathcal{S}')$ is a measurable map. Let $X : \Omega \to \mathbb{R}$ be an $\mathcal{S}$-measurable function. Then the following are equivalent.*

(i) *The function X is $\big(\sigma(F), \mathcal{B}_{\mathbb{R}}\big)$-measurable, where we recall from Example 19.1.3 (ii) that $\sigma(F) = F^{-1}(\mathcal{S}')$ is the σ-algebra generated by F.*

(ii) *There exists an $(\mathcal{S}', \mathcal{B}_{\mathbb{R}})$-measurable function $X' : \Omega' \to \mathbb{R}$ such that $X = X' \circ F$.*

Proof. Clearly, (ii) ⇒ (i). To prove that (i) ⇒ (ii) consider the family $\mathcal{M}$ of $\sigma(\mathcal{F})$-measurable functions of the form $X' \circ F$, $X' \in \mathcal{L}^0(\Omega', \mathcal{S}')$. We will prove that $\mathcal{M} = \mathcal{L}^0\big(\Omega, \sigma(\mathcal{F})\big)$. We will achieve using the monotone class theorem.

Step 1. $\boldsymbol{I}_\Omega \in \mathcal{M}$.

Step 2. $\mathcal{M}$ is a vector space. Indeed if $X, Y \in \mathcal{M}$ and $a, b \in \mathbb{R}$, then there exist $\mathcal{S}'$-measurable functions X', Y' such that

$$X = X' \circ F,\;\; Y = Y' \circ F,\;\; aX + bY = (aX' + bY') \circ F.$$

Hence $aX + bY \in \mathcal{M}$.

Step 3. $\boldsymbol{I}_A \in \mathcal{M}$, $\forall A \in \sigma(F)$. Indeed, since $A \in \sigma(F)$ there exists $A' \in \mathcal{S}'$ such that

$$A = F^{-1}(A')$$

so $\boldsymbol{I}_A = \boldsymbol{I}_{A'} \circ F$. Hence $\mathcal{M}$ contains all the $\sigma(F)$-measurable elementary functions.

Step 4. Suppose now that $X \in \mathcal{L}^0\big(\Omega, \sigma(F)\big)$ is nonnegative. Then there exists an increasing sequence $(X_n)_{n \in \mathbb{N}}$ of $\sigma(F)$-measurable nonnegative elementary functions that converges pointwise to X. For every $n \in \mathbb{N}$ there exists an $\mathcal{S}$-measurable elementary function $X_n' : \Omega' \to \mathbb{R}$ such that

$$X_n(\omega) = X_n'\big(F(\omega)\big),\;\; \forall \omega \in \Omega$$

Define

$$\Omega_0' := \big\{\, \omega' \in \Omega';\;\; \text{the limit } \lim_{n\to\infty} X_n'(\omega') \text{ exists and it is finite}\,\big\}$$

Let us observe that Ω_0' is $\mathcal{S}'$-measurable because

$$\omega' \in \Omega_0' \Longleftrightarrow \forall \nu \geqslant 1,\;\; \exists N \geqslant 1,\;\; \forall m, n \geqslant N :\;\; |X_n'(\omega') - X_m'(\omega')| < 1/\nu,$$

i.e.,

$$\Omega_0' = \bigcap_{\nu \in \mathbb{N}} \bigcup_{N \geqslant 1} \bigcap_{m,n > N} \Big\{ |X_n'(\omega') - X_m'(\omega')| < 1/\nu \Big\}.$$

Clearly, $F(\Omega) \subset \Omega_0'$. For any $\omega' \in \Omega'$ we set

$$X_\infty'(\omega') := \begin{cases} \lim_{n\to\infty} X_n'(\omega'), & \omega' \in \Omega_0', \\ \\ 0, & \omega' \in \Omega' \backslash \Omega_0'. \end{cases}$$

Arguing as in the proof of Proposition 19.1.19(ii) we deduce that X_∞' is $\mathcal{S}'$-measurable. For any $\omega \in \Omega$ the sequence $X_n'\big(F(\omega)\big) = X_n(\omega)$ is increasing and the limit

$$\lim_{n\to\infty} X_n'\big(F(\omega)\big)$$

exists and it is finite. Hence

$$X_\infty'\big(F(\omega)\big) = X(\omega),\;\; \forall \omega \in \Omega.$$

This proves that $\mathcal{M}$ is a monotone class in $\mathcal{L}^0\big(\Omega, \sigma(F)\big)$ that is also a vector space so it coincides with $\mathcal{L}^0\big(\Omega, \sigma(F)\big)$. □

Corollary 19.1.28. *Suppose that $X_1, \dots, X_n : (\Omega, \mathcal{S}) \to \mathbb{R}$ are $\mathcal{S}$-measurable functions. Then the function $X : \Omega \to \mathbb{R}$ is $\sigma(X_1, \dots, X_n)$-measurable if and only if there exists a $(\mathcal{B}_{\mathbb{R}^n}, \mathcal{B}_{\mathbb{R}})$-measurable function $u : \mathbb{R}^n \to \mathbb{R}$ such that*

$$X = u\big(X_1, \dots, X_n\big).$$

Proof. Apply the above theorem with $(\Omega', \mathcal{S}') = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ and

$$F(\omega) = \big(X_1(\omega), \dots, X_n(\omega)\big).$$

□

Remark 19.1.29. We see that, in its simplest form, Corollary 19.1.28 describes a measure theoretic form of functional dependence. Thus, if in a given experiment we can measure the quantities $X_1, \dots, X_n$ and we know that the information $X \leqslant c$ can be decided only by measuring the quantities $X_1, \dots, X_n$, then X is in fact a (measurable) function of $X_1, \dots, X_n$. In plain English this sounds tautological. In particular, this justifies the choice of term "measurable". □

19.1.3. Measures. The next crucial ingredient needed to define the Lebesgue integral is the concept of measure. This assigns a "size" to a measurable set: think of the area of a region in the plane. This concept should satisfy two desirable properties.

- **Additivity.** The measure (area) of the union of two regions A, B is the sum of the measures of the region from which we subtract the measure of the overlap.
- **Continuity.** If A is "close" to B, then the measure of A is close to that of B.

Here is the precise definition.

Definition 19.1.30. Suppose that $(\Omega, \mathcal{S})$ is a measurable space.

(i) A *measure* on $(\Omega, \mathcal{S})$ is a function

$$\mu : \mathcal{S} \to [0, \infty], \quad \mathcal{S} \ni S \mapsto \mu\big[\, S \,\big] \in [0, \infty]$$

such that, $\mu\big[\, \varnothing \,\big] = 0$, and it is *countably additive* or σ*-additive*, i.e., for any sequence of pairwise disjoint $\mathcal{S}$-measurable sets $(A_n)_{n \in \mathbb{N}}$ we have

$$\mu\left[\, \bigcup_{n \in \mathbb{N}} A_n \,\right] = \sum_{n \geqslant 1} \mu\big[\, A_n \,\big]. \tag{19.1.7}$$

We will denote by $\operatorname{Meas}(\Omega, \mathcal{S})$ the set of measures on $(\Omega, \mathcal{S})$.

(ii) The measure is called σ*-finite* if there exists an increasing sequence of $\mathcal{S}$-measurable sets

$$A_1 \subset A_2 \subset \cdots$$

such that

$$\bigcup_{n \in \mathbb{N}} A_n = \Omega \;\text{ and }\; \mu\big[\, A_n \,\big] < \infty, \;\; \forall n \in \mathbb{N}.$$

(iii) The measure is called *finite* if $\mu\big[\, \Omega \,\big] < \infty$. A *probability measure* is a measure $\mathbb{P}$ such that $\mathbb{P}\big[\, \Omega \,\big] = 1$.

(iv) A *measured space* is a triplet $(\Omega, \mathcal{S}, \mu)$, where $(\Omega, \mathcal{S})$ is a measurable space and $\mu : \mathcal{S} \to [0, \infty]$ is a measure.

□

Remark 19.1.31. The σ-additivity condition (19.1.7) is equivalent to a pair of conditions that are more convenient to verify in concrete situations.

(i) μ is *finitely additive*, i.e., for any finite collection of disjoint $\mathcal{S}$-measurable sets $A_1, \dots, A_n$ we have

$$\mu\left[\bigcup_{k=1}^{n} A_k\right] = \sum_{k=1}^{n} \mu\big[\, A_k \,\big].$$

(ii) μ is *increasingly continuous* i.e., for any increasing sequence of $\mathcal{S}$-measurable sets $A_1 \subset A_2 \subset \cdots$

$$\mu\left[\bigcup_{n\in\mathbb{N}} A_n\right] = \lim_{n\to\infty} \mu\big[\, A_n \,\big]. \tag{19.1.8}$$

Indeed, if μ is countably additive, we set

$$S := \bigcup_{n\in\mathbb{N}} A_n,\ \ S_1 = A_1,\ \ S_2 = A_2 \backslash S_1, \dots,$$

and we observe that the sets (S_n) are disjoint

$$\mu\big[\, S \,\big] = \lim_{n\to\infty} \sum_{k=1}^{n} \mu\big[\, S_k \,\big].$$

On the other hand,

$$\bigcup_{k=1}^{n} S_k = A_n$$

so

$$\sum_{k=1}^{n} \mu\big[\, S_k \,\big] = \mu\big[\, A_n \,\big].$$

Conversely if (19.1.8) and $(S_n)_{n\geqslant 1}$ is a sequence of disjoint sets, then

$$A_n = \bigcup_{k=1}^{n} S_k$$

is an increasing sequence of sets and the countable additivity follows by running the above arguments in reverse.

If $\mu\big[\, \Omega \,\big] < \infty$ and μ is finitely additive, then the increasing continuity condition (ii) is equivalent with the *decreasing continuity* condition, i.e., for any decreasing sequence of $\mathcal{S}$-measurable sets $B_1 \supset B_2 \supset \cdots$

$$\mu\left[\bigcap_{n\in\mathbb{N}} B_n\right] = \lim_{n\to\infty} \mu\big[\, B_n \,\big]. \tag{19.1.9}$$

Indeed, the sequence $B_n^c = \Omega \backslash B_n$ is increasing and $\mu\big[\, B_n^c \,\big] = \mu\big[\, \Omega \,\big] - \mu\big[\, B_n \,\big]$.

□

The next result describes some elementary but useful monotonicity properties of measure.

Proposition 19.1.32. *Suppose that* $\big(\Omega, \mathcal{S}, \mu\big)$ *is a measured space.*

(i) *If* $S_1, S_2 \in \mathcal{S}$ *and* $S_1 \subset S_2$*, then* $\mu\big[\,S_1\,\big] \leqslant \mu\big[\,S_2\,\big]$.

(ii) *If* $S_1, S_2 \in \mathcal{S}$*, then*

$$\mu\big[\,S_1 \cup S_2\,\big] \leqslant \mu\big[\,S_1\,\big] + \mu\big[\,S_2\,\big].$$

This inequality is known as the union bound*.*

Proof. (i) We have $S_2 = S_1 \cup (S_2 \backslash S_1)$ and

$$\mu\big[\,S_2\,\big] = \mu\big[\,S_1\,\big] + \mu\big[\,S_2 \backslash S_1\,\big] \geqslant \mu\big[\,S_1\,\big].$$

(ii) We have

$$\mu\big[\,S_1 \cup S_2\,\big] = \mu\big[\,S_1 \backslash (S_1 \cap S_2)\,\big] + \mu\big[\,S_2\,\big] \leqslant \mu\big[\,S_1\,\big] + \mu\big[\,S_2\,\big].$$

□

Example 19.1.33. Here are some simple examples of measures. In the next section we will describe a very important technique of producing measures.

(i) If $(\Omega, \mathcal{S})$ is a measurable space, then for any $\omega_0 \in \Omega$, the *Dirac measure* concentrated at ω_0 is the probability measure

$$\delta_{\omega_0} : \mathcal{S} \to [0, \infty), \;\; \delta_{\omega_0}\big[\,S\,\big] = \begin{cases} 1, & \omega_0 \in S, \\ 0, & \omega_0 \notin S. \end{cases}$$

(ii) Suppose that S is a finite or countable set. To any function $w : S \to [0, \infty]$ we associate a measure $\mu = \mu_w$ on $(S, 2^S)$ uniquely determined by the condition

$$\mu\big[\,\{s\}\,\big] = w(s), \;\; \forall s \in S.$$

We say that $\mu\big[\,\{s\}\,\big]$ is the *mass* of s with respect to μ. Often, for simplicity we will write

$$\mu\big[\,s\,\big] := \mu\big[\,\{s\}\,\big].$$

Note that for any $A \subset S$ we have

$$\mu_w\big[\,A\,\big] = \sum_{a \in A} w(a).$$

The associated measure μ_w is a probability measure if

$$\sum_{s \in S} w(s) = 1.$$

When S is finite and

$$w(s) = \frac{1}{|S|}, \;\; \forall s \in S,$$

then the associated probability measure μ_w is called the *uniform probability measure* on the finite set S.

(iii) Suppose that Ω is a set equipped with a partition

$$\Omega = \bigsqcup_{n\in\mathbb{N}} A_n$$

Denote by $\mathcal{S}$ the sigma-algebra generated by this partition, i.e., $\mathcal{S}$ consists of countable unions of the chambers A_k. Any function $w : \mathbb{N} \to [0,\infty)$, $n \mapsto w_n$, defines a measure μ_w on $\mathcal{S}$ uniquely determined by the conditions

$$\mu_w\big[\, A_n \,\big] = w_n, \;\; \forall n \in \mathbb{N}.$$

(iv) Suppose that $F : (\Omega, \mathcal{S}) \to (\Omega', \mathcal{S}')$ is a measurable map between measurable spaces. Then F induces a map

$$F_\# : \operatorname{Meas}(\Omega, \mathcal{S}) \to \operatorname{Meas}(\Omega', \mathcal{S}'), \;\; \mu \mapsto F_\#\mu. \tag{19.1.10}$$

More precisely, for any measure μ on $\mathcal{S}$ we set

$$F_\#\mu\big[\, S' \,\big] := \mu\big[\, F^{-1}(S') \,\big], \;\; \forall S' \in \mathcal{S}'.$$

The measure $F_\#\mu \in \operatorname{Meas}(\Omega', \mathcal{S}')$ is called the *pushforward of μ via F*.

To appreciate the complexity of this operation let us consider a very simple case. Suppose that A, B are finite sets and $\Phi : A \to B$. We can view Φ as a measurable map

$$\Phi : \big(\, A, 2^A \,\big) \to \big(\, B, 2^B \,\big).$$

Suppose that $\mu : 2^A \to [0,\infty)$ is a measure such that the mass of $a \in A$ is $\mu_a := \mu\big[\, \{a\} \,\big]$. Set $\nu := \Phi_\#\mu$. Then the ν-mass of $b \in B$ is

$$\nu\big[\, \{b\} \,\big] = \mu\big[\, \Phi^{-1}(b) \,\big] = \sum_{\Phi(a)=b} \mu_a.$$

In other words, the ν-mass of b is the sum of μ-masses of the points $a \in A$ that are mapped to b by Φ.

□

Proposition 19.1.34. *Consider a measurable space $(\Omega, \mathcal{S})$ and two finite measures*

$$\mu_1, \mu_2 : \mathcal{S} \to [0,\infty)$$

such that $\mu_1\big[\, \Omega \,\big] = \mu_2\big[\, \Omega \,\big]$. Then the collection

$$\mathscr{C} := \big\{\, C \in \mathcal{S};\;\; \mu_1\big[\, C \,\big] = \mu_2\big[\, C \,\big] \,\big\}$$

is a λ-system. In particular, if μ_1, μ_2 coincide on a π-system $\mathcal{P} \subset \mathcal{S}$, then they coincide on the sigma-algebra generated by $\mathcal{P}$.

Proof. Clearly $\emptyset, \Omega \in \mathscr{C}$. If $A, B \in \mathscr{C}$ and $A \subset B$, then

$$\mu_1\big[\,A\,\big] = \mu_2\big[\,A\,\big] < \infty,\ \ \mu_1\big[\,B\,\big] = \mu_2\big[\,B\,\big] < \infty$$

so

$$\mu_1\big[\,B\backslash A\,\big] = \mu_1\big[\,B\,\big] - \mu_1\big[\,A\,\big] = \mu_2\big[\,B\,\big] - \mu_2\big[\,A\,\big] = \mu_2\big[\,B\backslash A\,\big],$$

so $B\backslash A \in \mathscr{C}$. The condition (iii) in the Definition 19.1.5 of a λ-system follows from the σ-additivity of the measures μ_1, μ_2. □

Definition 19.1.35. Suppose that μ is a measure on the measurable space $(\Omega, \mathcal{S})$.

(i) A set $N \subset \Omega$ is called *μ-negligible* if there exists a set $S \in \mathcal{S}$ such that

$$N \subset S \text{ and } \mu\big[\,S\,\big] = 0.$$

We denote by $\mathcal{N}_\mu$ the collection of μ-negligible sets.

(ii) The σ-algebra $\mathcal{S}$ is said to be *complete* with respect to μ (or μ-complete) if it contains all the μ-negligible subsets. In this case we also say that the measured space $(\Omega, \mathcal{S}, \mu)$ is *complete.*

(iii) The *μ-completion* of $\mathcal{S}$ is the σ-algebra $\mathcal{S}^\mu := \sigma(\mathcal{S}, \mathcal{N}_\mu)$.

□

Clearly $\mathcal{S}^\mu$ is the smallest μ-complete σ-algebra containing $\mathcal{S}$. The proof of the following result is left to the reader as an exercise.

Proposition 19.1.36. *Suppose that μ is a σ-finite measure on the σ-algebra $\mathcal{S} \subset 2^\Omega$.*

(i) *The completion $\mathcal{S}^\mu$ has the alternate description*

$$\mathcal{S}^\mu = \big\{\,S \cup N;\ \ S \in \mathcal{S},\ \ N \in \mathcal{N}_\mu\,\big\} \subset 2^\Omega.$$

(ii) *The measure μ admits a unique extension to a measure $\bar{\mu} : \mathcal{S}^\mu \to [0, \infty)$. More precisely*

$$\forall S \in \mathcal{S},\ \ N \in \mathcal{N}_\mu,\ \ \bar{\mu}\big[\,S \cup N\,\big] = \mu\big[\,S\,\big].$$

□

19.1.4. The "almost everywhere" terminology. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space. We say that a property $\boldsymbol{P}$ of elements $\omega \in \Omega$ is satisfied *μ-almost everywhere* (or μ-a.e. for brevity) if there exists a set $N \in \mathcal{S}$ such that $\mu\big[\,N\,\big] = 0$ and any $\omega \in \Omega\backslash N$ satisfies the property $\boldsymbol{P}$. When the measure μ is clear from the context we will write simply a.e..

For example, a sequence of functions $f_n : \Omega \to \mathbb{R}$ is said to converge to $f : \Omega \to \mathbb{R}$ a.e. if the property

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

is satisfied a.e..

Let us observe that if $f : \Omega \to \mathbb{R}$ is $\mathcal{S}$-measurable and $g = f$ a.e., then g need not be $\mathcal{S}$-measurable. For example if N is negligible but not measurable, then the indicator function $\boldsymbol{I}_N$ is zero a.e., but not measurable. On the other hand we have the following result.

Proposition 19.1.37. *Suppose that* $(\Omega, \mathcal{S}, \mu)$ *is a* <u>*complete*</u> *measured space and*

$$f, g : \Omega \to (-\infty, \infty]$$

are two functions that are equal a.e. Then f *is* $\mathcal{S}$*-measurable if and only if* g *is* $\mathcal{S}$*-measurable.*

Proof. Suppose that f is measurable. Fix a μ-negligible set $N \in \mathcal{S}$ such that $f(\omega) = g(\omega)$, $\forall \omega \in \Omega \backslash N$. Then

$$g = g\boldsymbol{I}_{\Omega\backslash N} + g\boldsymbol{I}_N = f\boldsymbol{I}_{\Omega\backslash N} + g\boldsymbol{I}_N.$$

Now observe that $h := g\boldsymbol{I}_N$ is measurable. Indeed , for any $c \in (-\infty, \infty]$

$$\{h \leqslant c\} = \begin{cases} \{g \leqslant c\} \cap N, & c < 0, \\ (\Omega\backslash N) \cup (\{g \leqslant c\} \cap N), & c \geqslant 0. \end{cases}$$

The set $\{g \leqslant c\} \cap N$ is negligible since it is contained in the negligible set N. In particular, it is measurable since $\mathcal{S}$ is complete. Since $f\boldsymbol{I}_{\Omega\backslash N}$ is measurable as a product of two measurable sets we deduce that g is measurable as sum of measurable functions. □

Corollary 19.1.38. *Suppose that* $(\Omega, \mathcal{S}, \mu)$ *is a complete measured space and*

$$f_n : \Omega \to (-\infty, \infty], \;\; n \in \mathbb{N},$$

is a sequence of measurable functions that converges a.e. to a function $f : \Omega \to (-\infty, \infty]$*. Then* f *is* $\mathcal{S}$*-measurable.*

Proof. Fix a μ-negligible set $N \in \mathcal{S}$ such that

$$f(\omega) = \lim_{n\to\infty} f_n(\omega), \;\; \forall \omega \in \Omega\backslash N.$$

Then $f_n\boldsymbol{I}_{\Omega\backslash N}$ is measurable and converges everywhere to $f\boldsymbol{I}_{\Omega\backslash N}$. Thus $f\boldsymbol{I}_{\Omega\backslash N}$ is measurable and, since $f\boldsymbol{I}_{\Omega\backslash N} = f$ a.e., we deduce that f is also measurable. □

It turns out that the convergence a.e. is very close to uniform convergence.

Theorem 19.1.39 (Egorov)**.** *Suppose that* μ *is a* <u>finite</u> *measure on the measurable space* $(\Omega, \mathcal{S})$ *and* f, $(f_n)_{n\in\mathbb{N}}$ *are measurable functions such that* $f_n \to f$ μ*-a.e.. Then, for any* $\varepsilon > 0$ *there exists a set* $E \in \mathcal{S}$ *with the following properties.*

(i) $\mu\big[\, E \,\big] < \varepsilon$.

(ii) *The functions* f_n *converge uniformly to* f *on* $\Omega\backslash E$*.*

Proof. Without loss of generality we can assume $f_n(\omega) \to f(\omega)$ for any $\omega \in \Omega$. Hence, for $k \in \mathbb{N}$ and any $\omega \in \Omega$

$$\exists N \in \mathbb{N}: \ \ \forall n > N \ \ |f_n(\omega) - f(\omega)| < 1/k.$$

In other words, $\forall k \in \mathbb{N}$

$$\Omega = \bigcup_{N \in \mathbb{N}} \underbrace{\bigcap_{n > N} \left\{ |f_n - f| < 1/k \right\}}_{S_{N,k}}.$$

Note that $S_{N,k} \subset S_{N+1,k}, \ \ \forall k, N \in \mathbb{N}$. Thus, $\forall k \in \mathbb{N}$

$$\mu\big[\,\Omega\,\big] = \lim_{N \to \infty} \mu\big[\, S_{N,k}\,\big].$$

Hence, $\forall k > 0$, $\exists N_k > 0$ such that

$$\mu\big[\,\underbrace{\Omega \backslash S_{N_k,k}}_{=:E_k}\,\big] < \frac{\varepsilon}{2^k}.$$

Set

$$E := \bigcup_{k \in \mathbb{N}} E_k = \Omega \backslash \bigcap_{k \in \mathbb{N}} S_{N_k,k},$$

so that

$$\mu\big[\,E\,\big] \leqslant \sum_{k \in \mathbb{N}} \mu\big[\,E_k\,\big] < \varepsilon.$$

Note that

$$\mu\big[\,\Omega \backslash E\,\big] = \mu\big[\,\Omega\,\big] - \mu\big[\,E\,\big] > \mu\big[\,\Omega\,\big] - \varepsilon.$$

Note that

$$\omega \in \Omega \backslash E \Longleftrightarrow \omega \in \bigcap_{k \geqslant 1} S_{N_k,k} \Longleftrightarrow \forall k \in \mathbb{N}, \ \omega \in S_{N_k,k}$$

$(S_{n,k} \supset S_{N_k,k}$ if $n \geqslant N_k)$

$$\Rightarrow \ \forall k \geqslant 1, \ \ |f_n(\omega) - f(\omega)| < 1/k, \ \ \forall n \geqslant N_k,$$

so that f_n converges uniformly to f on $\Omega \backslash E$.

□

Consider a measure μ on the measurable space $(\Omega, \mathcal{S})$. The a.e. equality of measurable functions is an equivalence relation on the space $\bar{\mathcal{L}}^0(\Omega, \mathcal{S})$. We denote the quotient space by $\bar{L}^0(\Omega, \mathcal{S}, \mu)$. The quotient depends on the choice of measure μ. In the sequel, for simplicity we will refer to the elements of $L^0(\Omega, \mathcal{S}, \mu)$ as *functions* although they really are *equivalence classes of functions*.

Note that if $f, f' \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ and $f = f'$ μ-a.e., then $|f| < \infty$ μ-a.e. if and only if $|f'| < \infty$ μ-a.e.. We denote by $L^0(\Omega, \mathcal{S}, \mu)$ the space of equivalence classes of measurable functions that are finite μ-a.e..

Additionally, $f, f', g, g' \in \mathcal{L}^0(\Omega, \mathcal{S}, \mu)$, $f = f'$ and $g = g'$ μ-a.e., then $f + g = f' + g'$ and $cf = cf'$ μ-a.e., $\forall c \in \mathbb{R}$. This proves that $L^0(\Omega, \mathcal{S}, \mu)$ has a structure of vector space inherited from $\mathcal{L}^0\big(\Omega, \mathcal{S}\big)$.

We denote by $\bar{L}^0_+(\Omega, \mathcal{S}, \mu)$ the subset of equivalence classes of measurable functions that are nonnegative μ-a.e..

19.1.5. Premeasures. The measures in Example 19.1.33 are mostly confined to measured defined on the sigma-algebra determined by a partition. Constructing (nontrivial) measures on more general classes of sigma-algebras, e.g., the Borel sigma-algebra of a metric space, requires considerably more effort. The technology most frequently used to construct measures was proposed by Constantin Carathéodory (1873-1950) and it starts with a close relative of the concept of measure, namely the concept of *premeasure*.

Definition 19.1.40. Fix a set Ω and a *ring* of subsets $\mathcal{F} \subset 2^{\Omega}$.

(i) A function $\mu : \mathcal{F} \to [0, \infty]$ is called a *premeasure* if it satisfies the following conditions.
 (a) $\mu\big[\,\varnothing\,\big] = 0$
 (b) μ is *finitely additive*, i.e., for any finite collection of disjoint sets $A_1, \dots, A_n \in \mathcal{F}$ we have
$$\mu\left[\,\bigcup_{k=1}^{n} A_k\,\right] = \sum_{k=1}^{n} \mu\big[\,A_k\,\big].$$
 (c) μ is (conditionally) *countably additive*, i.e., if $(A_n)_{n\in\mathbb{N}}$ is a sequence of disjoint sets in $\mathcal{F}$ whose union is a set A also in $\mathcal{F}$, then
$$\mu\big[\,A\,\big] = \sum_{n\geqslant 1} \mu\big[\,A_n\,\big].$$

(ii) The premeasure μ is called *σ-finite* if there exists a sequence of sets $(\Omega_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ such that
$$\Omega = \bigcup_{n\in\mathbb{N}} \Omega_n, \quad \mu\big[\,\Omega_n\,\big] < \infty, \quad \forall n \in \mathbb{N}.$$

□

Proposition 19.1.41. *Suppose that $\mathcal{F}$ is an* algebra *of subsets of Ω and $\mu : \mathcal{F} \to [0, \infty]$ is an finitely additive function such that $\mu\big[\,\Omega\,\big] < \infty$. Then the following are equivalent.*

(i) *μ is conditionally countably additive.*

(ii) *For any descreasing seqeunce of sets $(F_n)_{n\geqslant 1}$ in $\mathcal{F}$ such that*
$$\bigcap_{n\geqslant 1} F_n = \varnothing$$
we have
$$\lim_{n\to\infty} \mu\big[\,F_n\,\big] = 0.$$

Proof. Note that $\mu\big[\,\varnothing\,\big] = \mu\big[\,\varnothing\,\big] + \mu\big[\,\varnothing\,\big]$ so $\mu\big[\,\varnothing\,\big] = 0$.

(i) $\Longrightarrow$ (ii). Let (F_n) be a decreasing sequence of subsets in $\mathcal{F}$ with empty intersection. Set $B_n = F_n^c = \Omega\backslash F_n$. Note that $\mu\big[\,B_n\,\big] = \mu\Omega\,\big] - \mu\big[\,F_n\,\big]$. Set $B_0 = \varnothing$, $A_n = B_n\backslash B_{n-1}$. The sets $(A_n)_{n\geqslant 1}$ are pairwise disjoint, $\mu\big[\,A_n\,\big] = \mu\big[\,B_n\,\big] - \mu\big[\,B_{n-1}\,\big]$, and

$$\bigcup_n A_n = \bigcup_n B_n = \Omega.$$

Since μ is conditionally countably additive we deduce that

$$\mu\big[\,\Omega\,\big] = \sum_{k\geqslant 1}\mu\big[\,A_k\,\big] = \lim_{n\to\infty}\underbrace{\sum_{k=1}^{n}\mu\big[\,A_k\,\big]}_{=\mu\big[\,B_n\,\big]} = \lim_{n\to\infty}\mu\big[\,B_n\,\big] = \mu\big[\,\Omega\,\big] - \lim_{n\to\infty}\mu\big[\,F_n\,\big].$$

Hence

$$\lim_{n\to\infty}\mu\big[\,F_n\,\big] = 0.$$

(ii) $\Longrightarrow$ (i) Let $(A_n)_{n\geqslant 1}$ be sequence of pairwise disjoint sets in $\mathcal{F}$ such that the union

$$A = \bigcup_{n\geqslant 1} A_n \in \mathcal{F}.$$

Set

$$B_n := \bigcup_{k=1}^{n} A_n,\ \ F_n := A\backslash B_n.$$

Then $(F_n)_{n\geqslant 1}$ is a decreasing sequence of sets in $\mathcal{F}$ with empty intersection so

$$0 = \lim_{n\to\infty}\mu\big[\,F_n\,\big] = \lim_{n\to\infty}\big(\,\mu\big[\,A\,\big] - \mu\big[\,B_n\,\big]\,\big) = \mu\big[\,A\,\big] - \lim_{n\to\infty}\sum_{k=1}^{n}\mu\big[\,\big[\,A_k\,\big],$$

so

$$\mu\big[\,A\,\big] = \sum_{n\geqslant 1}\mu\big[\,A_n\,\big].$$

□

Example 19.1.42. Suppose that $\mu : (\Omega, \mathcal{S}) \to [0,\infty]$ is a measure and $\mathcal{R}$ is a ring generating the sigma-algebra $\mathcal{S}$. Then the restriction of μ to $\mathcal{R}$ is a premeasure. Less obvious is the fact that any premeasure on $\mathcal{R}$ is obtained in this fashion. This follows from Carathédory's work that we will describe in the next section. □

In practice, the countable additivity condition (i.c) above is the most difficult to verify and often is the consequence of some hidden compactness condition. The next fundamental example will illustrate this point.

Theorem 19.1.43 (Alexandrov)**.** *Suppose that K is a compact metric space, $\mathcal{F}$ is an* <u>*algebra*</u> *of subsets of K and $\mu : \mathcal{F} \to [0,1]$ is a finitely additive function satisfying the* *following regularity property: for any $F \in \mathcal{F}$ and any $\varepsilon > 0$ there exists a set $F_- \in \mathcal{F}$ such that*

$$\boldsymbol{cl}(F_-) \subset F, \;\; \mu\big[\, F \backslash F_- \,\big] < \varepsilon.$$

Then μ is a premeasure. □

Proof. Let us introduce a convenient terminology. For $\varepsilon > 0$ we define an *ε-squeeze* of a set $F \in \mathcal{F}$ to be a set $G \in \mathcal{F}$ such that $\boldsymbol{cl}(G) \subset F$ and $\mu\big[\, F \backslash G \,\big] < \varepsilon$.

Lemma 19.1.44. *Suppose that $F_1, F_2 \in \mathcal{F}$, $F_2 \subset F_1$, and for $i = 1, 2$, G_i is an ε_i-squeeze of F_i. Then $G_1 \cap G_2$ is an $(\varepsilon_1 + \varepsilon_2)$-squeeze of F_2.*

Proof of Lemma 19.1.44. Clearly

$$\boldsymbol{cl}(G_1 \cap G_2) \subset \boldsymbol{cl}(G_2) \subset F_2,$$

$$F_2 \backslash (G_1 \cap G_2) = F_2 \cap (G_1 \cap G_2)^c = F_2 \cap (G_1^c \cup G_2^c) = (F_2 \cap G_1^c) \cup (F_2 \cap G_2^c),$$

and

$$\mu\big[\, F_2 \backslash (G_1 \cap G_2) \,\big] = \mu\big[\, (F_2 \backslash G_1) \cup (F_2 \backslash G_2) \,\big]$$

$$\leqslant \mu\big[\, F_2 \backslash G_1 \,\big] + \mu\big[\, F_2 \backslash G_2 \,\big] \leqslant \mu\big[\, F_1 \backslash G_1 \,\big] + \mu\big[\, F_2 \backslash G_2 \,\big] \leqslant \varepsilon_1 + \varepsilon_2.$$

□

To prove that μ is a premeasure it suffices to show that if $(F_n)_{n \in \mathbb{N}}$ is a decreasing sequence in $\mathcal{F}$ with empty intersection, then

$$\lim_{n \to \infty} \mu\big[\, F_n \,\big] = 0.$$

Fix $\varepsilon > 0$. For $n \in \mathbb{N}$, fix an $\frac{\varepsilon}{2^n}$-squeeze G_n of F_n. Define

$$H_n := \bigcap_{k=1}^{n} G_n, \;\; \varepsilon_n := \sum_{k=1}^{n} \frac{\varepsilon}{2^k} = \varepsilon\big(\, 1 - 2^{-n} \,\big).$$

Applying Lemma 19.1.44 iteratively we deduce that H_n is an ε_n-squeeze of F_n. By construction the sequence H_n is decreasing and thus the sequence of closures $\boldsymbol{cl}(H_n)$ is decreasing as well. Note that

$$\bigcap_n \boldsymbol{cl}(H_n) \subset \bigcap F_n = \varnothing.$$

Since K is compact we deduce that there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $\boldsymbol{cl}(H_N) = \varnothing$. Hence $H_N = \varnothing$ and since H_N is an ε_N-squeeze we deduce that, $\forall n \geqslant N$

$$\mu\big[\, F_n \,\big] \leqslant \mu\big[\, F_N \,\big] = \mu\big[\, F_N \backslash H_N \,\big] \leqslant \varepsilon_N < \varepsilon.$$

□

19.1.6. The Lebesgue premeasure. Denote by $\mathcal{I}$ the collection of intervals of $\mathbb{R}$. By *interval* we mean a connected subset of $\mathbb{R}$. In other words, the intervals are the sets of the form

- $\varnothing$, $\mathbb{R}$,
- intervals of the form $(a,b]$, $-\infty \leqslant a < b < \infty$,
- intervals of the form (a,∞), $a \in \mathbb{R}$.

Denote $\mathcal{F}$ the subsets of $\mathbb{R}$ that are disjoint unions of intervals in $\mathcal{I}$, i.e., sets S of the form

$$S = I_1 \cup \cdots \cup I_n,\ \ I_1,\dots,I_n \in \mathcal{I},\ \ I_j \cap I_k = \varnothing,\ \ \forall j \neq k. \tag{19.1.11}$$

Let us observe that a set $S \in \mathcal{F}$ admits many decompositions of the type (19.1.11). For example,

$$(a,b] = (a,c] \cup (c,b],\ \ \forall c \in (a,b).$$

Lemma 19.1.45. *The collection $\mathcal{F}$ is an algebra of sets.*

Proof. Observe that the disjoint union of two sets in $\mathcal{F}$ is a set in $\mathcal{F}$. Observe next that the intersection of two intervals I,J in $\mathcal{I}$ is an interval in $\mathcal{I}$. More generally, if $I \in \mathcal{I}$ and

$$F = \bigsqcup_{k=1}^{n} J_k \in \mathcal{F},\ \ J_k \in \mathcal{I},$$

then we have

$$I \cap F = \bigsqcup_{k=1}^{n} I \cap J_k \in \mathcal{F}.$$

Finally if

$$F' = \bigsqcup_{j=1}^{m} I_j,\ \ I_j \in \mathcal{I},$$

then

$$F' \cap F = \bigsqcup_{j=1}^{m} \underbrace{I_j \cap F}_{\in \mathcal{F}} \in \mathcal{F}.$$

Thus, $\mathcal{F}$ is closed under taking intersections.

Note that if $I \in \mathcal{I}$, then $I^c = \mathbb{R} \backslash I \in \mathcal{F}$. Thus the complement of a set in $\mathcal{F}$ is the intersection of sets in $\mathcal{F}$ and thus it is in $\mathcal{F}$.

A union of sets in $\mathcal{F}$ is in $\mathcal{F}$ because its complement is an intersection of sets in $\mathcal{F}$. This proves that $\mathcal{F}$ is an algebra of sets. □

Define $\boldsymbol{\lambda} : \mathcal{I} \to [0,\infty]$, $\boldsymbol{\lambda}(I) =$ the length of the interval I. More precisely,

$$\begin{gathered}\boldsymbol{\lambda}\big[\,\varnothing\,\big] = 0,\ \ \boldsymbol{\lambda}\big[\,\mathbb{R}\,\big] = \boldsymbol{\lambda}\big[\,(-\infty,b)\,\big] = \boldsymbol{\lambda}\big[\,(-\infty,b]\,\big]\\ = \boldsymbol{\lambda}\big[\,(a,\infty)\,\big] = \boldsymbol{\lambda}\big[\,[a,\infty)\,\big] = \infty,\\ \boldsymbol{\lambda}\big[\,[a,b)\,\big] = \boldsymbol{\lambda}\big[\,(a,b]\,\big] = \boldsymbol{\lambda}\big[\,[a,b]\,\big] = b-a,\ \ \forall -\infty < a \leqslant b < \infty.\end{gathered}$$

Observe that if an interval $I \in \mathcal{I}$ is the disjoint union of $n > 1$ intervals $I_1, \dots, I_n \in \mathcal{I}$,

$$I = \bigsqcup_{j=1}^{n} I_j,$$

then

$$\boldsymbol{\lambda}\big[\, I \,\big] = \sum_{j=1}^{n} \boldsymbol{\lambda}\big[\, I_j \,\big]. \tag{19.1.12}$$

Suppose that $S \in \mathcal{F}$ decomposes in two different ways as disjoint unions of intervals in $\mathcal{I}$

$$S = \bigsqcup_{j=1}^{m} I_j = \bigsqcup_{k=1}^{n} J_k,$$

then

$$\sum_{j=1}^{m} \boldsymbol{\lambda}\big[\, I_j \,\big] = \sum_{k=1}^{n} \boldsymbol{\lambda}\big[\, J_k \,\big]. \tag{19.1.13}$$

Indeed, set

$$U_{jk} := I_j \cap J_k.$$

Note that the sets U_{jk} are pairwise disjoint, they belong to the collection $\mathcal{I}$ and

$$I_j = \bigsqcup_{k=1}^{n} U_{jk}, \quad J_k = \bigsqcup_{j=1}^{m} U_{jk}.$$

We deduce from (19.1.12)

$$\boldsymbol{\lambda}\big[\, I_j \,\big] = \sum_{k=1}^{n} \boldsymbol{\lambda}\big[\, U_{jk} \,\big], \quad \boldsymbol{\lambda}\big[\, J_k \,\big] = \sum_{j=1}^{m} \boldsymbol{\lambda}\big[\, U_{jk} \,\big], \quad \forall j, k$$

so that

$$\sum_{j=1}^{m} \boldsymbol{\lambda}\big[\, I_j \,\big] = \sum_{j=1}^{m} \sum_{k=1}^{n} \boldsymbol{\lambda}\big[\, U_{jk} \,\big] = \sum_{k=1}^{n} \sum_{j=1}^{m} \boldsymbol{\lambda}\big[\, U_{jk} \,\big] = \sum_{k=1}^{n} \boldsymbol{\lambda}\big[\, J_k \,\big].$$

For every $S \in \mathcal{F}$ represented as a disjoint union of intervals in $\mathcal{I}$,

$$S = \bigsqcup_{j=1}^{m} I_j$$

we set

$$\boldsymbol{\lambda}\big[\, S \,\big] := \sum_{j=1}^{m} \boldsymbol{\lambda}\big[\, I_j \,\big].$$

The equality (19.1.13) shows that the above definition is independent of the decomposition of S as a disjoint union of intervals in $\mathcal{I}$.

The map $\boldsymbol{\lambda}$ is obviously finitely additive. Indeed if $S, S' \in \mathcal{F}$ are disjoint

$$S = \bigsqcup_{k=1}^{m} I_k, \quad S' = \bigsqcup_{k=m+1}^{m+n} I_k, \quad S \cup S' = \bigsqcup_{k=1}^{m+n} I_k$$

and

$$\boldsymbol{\lambda}\big[\, S \cup S' \,\big] = \sum_{k=1}^{m+n} \boldsymbol{\lambda}\big[\, I_k \,\big] = \boldsymbol{\lambda}\big[\, S \,\big] + \boldsymbol{\lambda}\big[\, S' \,\big].$$

For each $N \in \mathbb{N}$ we denote by $\mathcal{F}_N$ the collection

$$\mathcal{F}_N = F \cap [-N,N] = \big\{ F \cap [-N,N];\ \ F \in \mathcal{F} \big\}$$

We set $K_N := [-N,N]$ so $\mathcal{F}_N$ is an algebra of subsets of the compact interval K_N. It consists of *finite* disjoint union of intervals of in the collection

$$\mathcal{I}_N := \mathcal{I} \cap K_N = \big\{ \, (a,b],\ \ [-N,c];\ \ -N \leqslant a < b \leqslant N,\ \ -N \leqslant c \leqslant N \, \big\}.$$

We define as before an additive measure

$$\boldsymbol{\lambda}_N : \mathcal{F}_N \to [0,\infty)$$

uniquely determined by the equalities

$$\boldsymbol{\lambda}_N\big[\,(a,b]\,\big] = b-a,\ \ \boldsymbol{\lambda}_N\big[\,[-N,c]\,\big] = c+N = c-(-N).$$

Note that for any interval $(a,b]$, $-\infty \leqslant a < b < \infty$, we have

$$b-a = \lim_{N\to\infty} \boldsymbol{\lambda}\big[\,(a,b] \cap K_N \,\big].$$

We deduce that

$$\forall F \in \mathcal{F},\ \ \boldsymbol{\lambda}\big[\, F \,\big] = \lim_{N\to\infty} \boldsymbol{\lambda}_N\big[\, F \cap K_N \,\big]. \tag{19.1.14}$$

Theorem 19.1.46. *The additive measure of $\boldsymbol{\lambda}_N$ is a premeasure.*

Proof. Since K_N is compact it suffices to show that $\boldsymbol{\lambda}_N$ satisfies the regularity property in Alexandrov's Theorem 19.1.43.

Let $F \in \mathcal{F}_N$ and $\varepsilon > 0$. Then F is a disjoint union of n intervals

$$F = \bigsqcup_{k=1}^{n} I_k,\ \ I_k \in \mathcal{I}_N$$

For each interval I_k we can find an interval $I_k^- \subset I_k$, $I_k^- \in \mathcal{I}_N$ such that

$$\boldsymbol{cl}\big(\, I_k^- \,\big) \subset I_k,\ \ \boldsymbol{\lambda}_N\big[\, I_k \setminus I_k^- \,\big] < \frac{\varepsilon}{n}.$$

We set

$$F^- := \bigsqcup_{k=1}^{n} I_k^-.$$

Note that F^- is closed, $F^- \subset F$ and $\boldsymbol{\lambda}_N\big[\, F \setminus F^- \,\big] < \varepsilon$. □

Theorem 19.1.47. *The above map $\boldsymbol{\lambda} : \mathcal{F} \to [0,\infty]$ is a premeasure. We will refer to it as the* Lebesgue premeasure.

Proof. Suppose that (S_n) is an increasing family of subsets in $\mathcal{F}$ such that

$$S = \bigcup_{n\in\mathbb{N}} S_n \in \mathcal{F}.$$

We will show that

$$\boldsymbol{\lambda}\big[\, S \,\big] = \lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big]. \tag{19.1.15}$$

Clearly it suffices to prove only that

$$\boldsymbol{\lambda}\big[\, S \,\big] \leqslant \lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big].$$

We distinguish two cases.

Case 1. $\mu\big[\, S \,\big] < \infty$. We will prove that for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big] \geqslant \boldsymbol{\lambda}\big[\, S \,\big] - \varepsilon.$$

Acording to (19.1.14) we can find N sufficiently large such that

$$\boldsymbol{\lambda}_N\big[\, S \cap K_N \,\big] \geqslant \boldsymbol{\lambda}\big[\, S \,\big] - \varepsilon,$$

where we recall that $K_N = [-N, N]$. Set $S^N := S \cap K_N$, $S^N_n = S_n \cap K_N$. Then $S_N, S^N_n \in \mathcal{F}_N$ and

$$S^N_n \nearrow S^N.$$

Theorem 19.1.46 implies

$$\lim_{n\to\infty} \boldsymbol{\lambda}_N\big[\, S^N_n \,\big] = \boldsymbol{\lambda}_N\big[\, S^N \,\big] \geqslant \boldsymbol{\lambda}\big[\, S \,\big] - \varepsilon.$$

Obviously

$$\lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big] \geqslant \lim_{n\to\infty} \boldsymbol{\lambda}_N\big[\, S^N_n \,\big].$$

Case 2. $\boldsymbol{\lambda}\big[\, S \,\big] = \infty$. We will prove that for any $C > 0$

$$\lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big] \geqslant C.$$

Fix $C > 0$ and choose N sufficiently large such that $\boldsymbol{\lambda}_N\big[\, S^N \,\big] \geqslant C$. Using Theorem 19.1.46 again we deduce

$$\lim_{n\to\infty} \boldsymbol{\lambda}\big[\, S_n \,\big] \geqslant \lim_{n\to\infty} \boldsymbol{\lambda}_N\big[\, S^N_n \,\big] = \boldsymbol{\lambda}\big[\, S^N \,\big] \geqslant C.$$

This completes the proof of Theorem 19.1.47. □

Remark 19.1.48 (Lebesgue-Stieltjes premeasures)**.** The above result extends with no conceptual modifications to the following more general situation. Fix a *gauge* function, i.e., a nondecreasing, right-continuous function $F : \mathbb{R} \to \mathbb{R}$. Recall that right-continuity signifies that for any $x_0 \in \mathbb{R}$.

$$F(x_0) = \lim_{x\searrow x_0} F(x).$$

We set

$$F(\pm\infty) = \lim_{x\to\infty} F(x), \;\; F(x_0-) := \lim_{x\nearrow x_0} F(x), \;\; \forall x_0 \in \mathbb{R}.$$

For $\infty \leqslant a \leqslant b \leqslant \infty \in \mathscr{J}$ we set

$$\boldsymbol{\lambda}_F\big[\,[a,b]\,\big] = F(b) - F(a-),\;\; \boldsymbol{\lambda}\big[\,(a,b)\,\big] = F(b-) - F(a),$$

$$\boldsymbol{\lambda}_F\big[\,(a,b]\,\big] = F(b) - F(a),\;\; \boldsymbol{\lambda}_F\big[\,[a,b)\,\big] = F(b-) - F(a-),$$

Note that

$$\bar{\lambda}_F\big[\,\mathbb{R}\,\big] = F(\infty) - F(-\infty),\;\; \boldsymbol{\lambda}_F\big[\,\{a\}\,\big] = F(a) - F(a-).$$

Let us observe that $F(\infty) - F(-\infty)$ is well defined since $F(-\infty) \leqslant F(0) < \infty$. Arguing exactly as above we deduce that $\boldsymbol{\lambda}_F$ extends to a *sigma-finite* premeasure

$$\boldsymbol{\lambda}_F : \mathscr{F} \to [0,\infty]$$

called the *Lebesgue-Stieltjes premeasure* associated to the gauge function F. The Lebesgue premeasure corresponds to the gauge function $F_0(x) = x$, $\forall x \in \mathbb{R}$. Note that for any $c \in \mathbb{R}$, $\boldsymbol{\lambda}_F = \boldsymbol{\lambda}_{F+c}$ □

19.2. Construction of measures

In this section we describe a very general and very versatile method of producing measures. This technique, pioneered by Constantin Carathéodory (1873-1950) is based on two conceptually distinct results, both due to Carathéodory.

- The first result shows that to a rather vague notion of volume (or measure) called *outer measure* we can associate a sigma-algebra and a measure on it. However, a priori it is not clear what the resulting sigma-algebra is. What if we want to produce measures on a *given* sigma-algebra?
- The second result shows that if the above process is applied to a *premeasure* on an algebra $\mathcal{A}$, then we obtain a measure of the sigma-algebra generated by $\mathcal{A}$.

19.2.1. The Carathéodory construction. Let Ω be a nonempty set.

Definition 19.2.1. An *outer measure* on Ω is a function

$$\mu : 2^{\Omega} \to [0,\infty]$$

satisfying the following conditions.

(i) $\mu\big[\,\varnothing\,\big] = 0$.

(ii) The function μ is *monotone*, i.e., for any $A \subset B \subset \Omega$ we have $\mu\big[\,A\,\big] \leqslant \mu\big[\,B\,\big]$.

(iii) The function μ is *countably subadditive*, i.e., for any a finite or countable family $(A_i)_{i\in I}$ of subsets of Ω, then

$$\mu\Big[\,\bigcup_{i\in I} A_i\,\Big] \leqslant \sum_{i\in I} \mu\big[\,A_i\,\big].$$

Given an outer measure μ on Ω we say that a set $S \subset \Omega$ is *μ-measurable* if

$$\mu\big[\,A\,\big] = \mu\big[\,A \cap S\,\big] + \mu\big[\,A \cap S^c\,\big],\;\; \forall A \subset \Omega. \tag{19.2.1}$$

We will denote by $\mathscr{S}_\mu$ the collection of μ-measurable subsets of Ω. □

Observe that for any $A, S \subset \Omega$ we have $A = (A \cap S^c) \cup (A \cap S)$. If μ is an outer measure, we have

$$\mu\big[\,A\,\big] \leqslant \mu\big[\,A \cap S^c\,\big] + \mu\big[\,A \cap S\,\big].$$

Thus, a set $S \subset \Omega$ is μ-measurable if and only if

$$\mu\big[\,A\,\big] \geqslant \mu\big[\,A \cap S\,\big] + \mu\big[\,A \cap S^c\,\big], \ \ \forall A \subset \Omega. \tag{19.2.2}$$

Another way of stating the measurability of S is in terms of *separation*. The sets U, V are said to be separated by S is $U \subset S$ and $V \subset \Omega \backslash S$. The set S is μ-measurable if and only if, for any sets U, V separated by S we have

$$\mu\big[\,U \cup V\,\big] \geqslant \mu\big[\,U\,\big] + \mu\big[\,V\,\big]. \tag{19.2.3}$$

Indeed, (19.2.3) follows from (19.2.2) by letting $A = U \cap V$ so that $U = A \cap S$ and $V = A \cap S^c$. Conversely, if (19.2.3) holds and $A \subset \Omega$, we set $U = A \cap S$, $V = A \cap S^c$ and then S separates U and V, $A \supset U \cup V$ so

$$\mu\big[\,A\,\big] \geqslant \big[\,U \cup V\,\big] \geqslant \mu\big[\,U\,\big] + \mu\big[\,V\,\big].$$

As the next result shows, outer measures are a dime a dozen. Before we state it we need to introduce some terminology.

Definition 19.2.2. Let Ω be a nonempty set.

(i) A *class of models* or *paving* in Ω is a family $\mathcal{M}$ of subsets of Ω (called models) such that $\varnothing \in \mathcal{M}$.

(ii) If $\mathcal{M}$ is a class of models in Ω, then an $\mathcal{M}$-*cover* of a subset $S \subset \Omega$ is a countable family of models $(M_n)_{n\in\mathbb{N}}$ in $\mathcal{M}$ such that

$$S \subset \bigcup_{n\in\mathbb{N}} M_n.$$

(iii) A *gauge* on a class of models $\mathcal{M}$ is a function $\rho : \mathcal{M} \to [0,\infty]$ such that $\rho\big[\,\varnothing\,\big] = 0$.

□

Proposition 19.2.3. *Let Ω be a nonempty set, $\mathcal{M}$ a class of models in Ω and ρ a gauge on $\mathcal{M}$. For any $S \subset \Omega$ we define $\mu\big[\,S\,\big] = \mu_\rho\big[\,S\,\big]$ to be the infimum of the sums*

$$\sum_{n\in\mathbb{N}} \rho\big[\,M_n\,\big],$$

over all the $\mathcal{M}$-covers $(M_n)_{n\in\mathbb{N}}$ of S. If S admits no $\mathcal{M}$-cover we set $\mu\big[\,S\,\big] := \infty$. Then μ is an outer measure on Ω called the outer measure determined by the gauge ρ.

Proof. Indeed, (i) follows from $\rho[\varnothing\,\big] = 0$. Let $A \subset B$. If B admits no $\mathcal{M}$ cover then obvisouly $\mu\big[\,A\,\big] \leqslant \infty = \mu\big[\,B\,\big]$. Otherwise, any $\mathcal{M}$-cover of B is also an $\mathcal{M}$-cover of A, and from the definition of μ as an infimum over $\mathcal{M}$-covers we deduce that $\mu\big[\,A\,\big] \leqslant \mu\big[\,B\,\big]$.

Suppose that $(A_n)_{n\in\mathbb{N}}$ is a countable family of subsets of Ω. Set

$$A=\bigcup_{n\in\mathbb{N}}A_n.$$

If $\mu\big[\,A_n\,\big]=\infty$ for some n, then the subadditivity is obvious. Suppose that $\mu\big[\,A_n\,\big]<\infty$, $\forall n$. Then for any $\varepsilon>0$ there exists an $\mathcal{M}$-cover $(M_{n,k})_{k\in\mathbb{N}}$ of A_n such that

$$\mu\big[\,A_n\,\big]+\frac{\varepsilon}{2^n}\geqslant\sum_{k\in\mathbb{N}}\rho\big[\,M_{nk}\,\big].$$

Then $(M_{n,k})_{n,k\in\mathbb{N}}$ is an $\mathcal{M}$ cover of A and thus, $\forall\varepsilon>0$

$$\mu\big[\,A\,\big]\leqslant\sum_{k,n\in\mathbb{N}}\rho\big[\,M_{n,k}\,\big]+\sum_{n\in\mathbb{N}}\frac{\varepsilon}{2^n}\leqslant\sum_{n\in\mathbb{N}}\mu\big[\,A_n\,\big]+\varepsilon.$$

This proves the countable subadditivity of μ. □

Theorem 19.2.4 (Carathéodory construction)**.** *If μ is an outer measure on the nonempty set Ω, then the following hold.*

(i) *The collection $\mathcal{S}_\mu$ of μ-measurable sets is a sigma-algebra.*

(ii) *The restriction of μ to $\mathcal{S}_\mu$ is a measure.*

(iii) *The sigma-algebra $\mathcal{S}_\mu$ is μ-complete, i.e.,*

$$\forall S\subset\Omega,\;\;\mu\big[\,S\,\big]=0\Rightarrow S\in\mathcal{S}_\mu.$$

Proof. 1. Clearly $\varnothing\in\mathcal{S}_\mu$. Since $(S^c)^c=S$ we deduce from the definition (19.2.1) that

$$\forall S,\;\;S\in\mathcal{S}_\mu\Longleftrightarrow S^c\in\mathcal{S}_\mu.$$

In particular, $\Omega\in\mathcal{S}_\mu$.

2. Let us show that

$$S_1,S_2\in\mathcal{S}_\mu\Rightarrow S_1\cup S_2\in\mathcal{S}_\mu.$$

We have to prove that for any $A\subset\Omega$ we have

$$\mu\big[\,A\cap(S_1\cup S_2)\,\big]+\mu\big[\,A\cap(S_1\cup S_2)^c\,\big]\leqslant\mu\big[\,A\,\big].$$

We have

$$(S_1\cup S_2)=S_1\cup(S_2\cap S_1^c)$$

so

$$A\cap(S_1\cup S_2)=(A\cap S_1)\cup(A\cap S_1^c\cap S_2)$$

so that

$$\mu\big[\,A\cap(S_1\cup S_2)\,\big]+\mu\big[\,A\cap(S_1\cup S_2)^c\,\big]$$
$$=\mu\big[\,(A\cap S_1)\cup(A\cap S_1^c\cap S_2)\,\big]+\mu\big[\,(A\cap S_1^c)\cap S_2^c\,\big]$$

(μ is subadditive)

$$\leqslant\mu\big[\,A\cap S_1\,\big]+\mu\big[\,(A\cap S_1^c)\cap S_2\,\big]+\mu\big[\,(A\cap S_1^c)\cap S_2^c\,\big]$$

(S_2 is μ-measurable)

$$=\mu\big[\,A\cap S_1\,\big]+\mu\big[\,A\cap S_1^c\,\big]=\mu\big[\,A\,\big],$$

where at the last step we have used the μ-measurability of S_1. Since

$$\Omega\in\mathcal{S}_\mu\;\;\text{and}\;\;S\in\mathcal{S}_\mu\Longleftrightarrow S^c\in\mathcal{S}_\mu,$$

we deduce that $\mathcal{S}_\mu$ is an *algebra*.

3. Observe that if $S_1, S_2 \in \mathcal{S}_\mu$ are disjoint, then using (19.2.1) with $A = S_1 \cup S_2$ we deduce

$$\mu\big[\, S_1 \cup S_2 \,\big] = \mu\big[\, S_1 \,\big] + \mu\big[\, S_2 \,\big].$$

This proves that μ is *finitely additive* on $\mathcal{S}_\mu$.

4. Let us show that $\mathcal{S}_\mu$ is a *sigma-algebra.* Let $(S_n)_{n\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{S}_\mu$. We denote by S_∞ their union and we set $T_n := S_n \backslash S_{n-1}$, $\forall n \in \mathbb{N}$. The sets T_n are pairwise disjoint and

$$S_n = \bigcup_{k=1}^n T_k.$$

The finite additivity implies

$$\mu\Big[\, S_n \,\Big] = \sum_{k=1}^n \mu\big[\, T_k \,\big].$$

We will show that $S_\infty \in \mathcal{S}_\mu$. Since $T_n \in \mathcal{S}_\mu$ we deduce that for any set $A \subset \Omega$ we have

$$\mu\big[\, A \cap S_n \,\big] = \mu\big[\, A \cap S_n \cap T_n \,\big] + \mu\big[\, A \cap S_n \cap T_n^c \,\big] = \mu\big[\, A \cap S_n \,\big] + \mu\big[\, A \cap S_{n-1} \,\big].$$

We conclude inductively that

$$\mu\big[\, A \cap S_n \,\big] = \sum_{j=1}^n \mu\big[\, A \cap T_j \,\big].$$

On the other hand, $S_n \in \mathcal{S}_\mu$ and we have

$$\mu\big[\, A \,\big] = \mu\big[\, A \cap S_n \,\big] + \mu\big[\, A \cap S_n^c \,\big] = \sum_{j=1}^n \mu\big[\, A \cap T_j \,\big] + \mu\big[\, A \cap S_n^c \,\big]$$

$$\geqslant \sum_{j=1}^n \mu\big[\, A \cap T_j \,\big] + \mu\big[\, A \cap S_\infty^c \,\big].$$

Letting $n \to \infty$ we deduce

$$\mu\big[\, A \,\big] \geqslant \sum_{n=1}^\infty \mu\big[\, A \cap T_n \,\big] + \mu\big[\, A \cap S_\infty^c \,\big],$$

$$\geqslant \mu\Big[\, \bigcup_{n=1}^\infty (\, A \cap T_j \,) \,\Big] + \mu\big[\, A \cap S_\infty^c \,\big] = \mu\big[\, A \cap S_\infty \,\big] + \mu\big[\, A \cap S_\infty^c \,\big].$$

This proves that S_∞ is μ-measurable.

5. We will show that μ is *countably additive* on $\mathcal{S}_\mu$. Let $(S_n)_{n\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{S}_\mu$. We denote by S_∞ their union and we set $T_n := S_n \backslash S_{n-1}$. The monotonicity of μ implies $\mu\big[\, S_\infty \,\big] \geqslant \mu\big[\, S_n \,\big]$, $\forall n$. As in **4** we have

$$\mu\big[\, S_\infty \,\big] \geqslant \mu\Big[\, S_n \,\Big] = \sum_{k=1}^n \mu\big[\, T_k \,\big].$$

Letting $n \to \infty$ we deduce

$$\mu\big[\, S_\infty \,\big] \geqslant \lim_{n\to\infty} \mu\big[\, S_n \,\big] = \sum_{k=1}^\infty \mu\big[\, T_k \,\big].$$

On the other hand, the countable subadditivity of the outer measure μ shows that the opposite inequality is true, i.e.,

$$\mu\big[\, S_\infty \,\big] \leqslant \sum_{k=1}^\infty \mu\big[\, T_k \,\big],$$

and thus

$$\mu\Big[\, S_\infty \,\Big] = \sum_{k=1}^\infty \mu\big[\, T_k \,\big] = \lim_{n\to\infty} \mu\big[\, S_n \,\big].$$

6. Finally, let us show that $\mathcal{S}_\mu$ is μ-complete. Suppose that $S \subset N \in \mathcal{S}_\mu$ and $\mu[\,N\,] = 0$. We want to show that S is also μ-measurable. Indeed, note first that $\mu[\,S\,] = 0$. For any $A \subset \Omega$ we have

$$\begin{aligned}
&\underbrace{\mu[\,A \cap S\,]}_{=0} + \mu[\,A \cap S^c\,] = \mu[\,A \cap S^c\,]\\
&(N \in \mathcal{S}_\mu,\ \mu[\,N\,] = 0) \qquad = \underbrace{\mu[\,A \cap S^c \cap N\,]}_{=0} + \mu[\,A \cap S^c \cap N^c\,]\\
&(A \cap S^c \cap N^c \subset A \cap N^c) \qquad \leqslant \mu[\,A \cap N^c\,] = \underbrace{\mu[\,A \cap N\,]}_{=0} + \mu[\,A \cap N^c\,] = \mu[\,A\,].
\end{aligned}$$

This completes the proof of Theorem 19.2.4. □

Without any additional knowledge about the outer measure μ we cannot say too much about the sigma algebra of μ-measurable sets. We want to describe two instances when we can provide additional information about $\mathcal{S}_\mu$.

Consider first the case when the outer measure is obtained from a gauge that is a premeasure on an algebra of sets.

Theorem 19.2.5 (Carathéodory extension)**.** *Let Ω be a nonempty set and*

$$\mu : \mathcal{F} \to [0, \infty]$$

is a premeasure on the ring of subsets $\mathcal{F}$. Think of $\mathcal{F}$ as a class of models and of μ as a gauge. Denote by $\widehat{\mu}$ the associated outer measure. Then the following hold.

(i) $\mu[\,F\,] = \widehat{\mu}[\,F\,]$, $\forall F \in \mathcal{F}$.

(ii) $\sigma(\mathcal{F}) \subset \mathcal{S}_{\widehat{\mu}}$.

In other words, the restriction of $\widehat{\mu}$ to $\sigma(\mathcal{F})$ is a measure that extends μ.

Moreover, if μ is σ-finite, then any measure on $\sigma(\mathcal{F})$ that extends μ coincides with $\widehat{\mu}\big|_{\sigma(\mathcal{F})}$.

Proof. We will need a general fact about premeasures.

Lemma 19.2.6. *For any $F \in \mathcal{F}$, and any sequence $(F_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ such that*

$$F \subset \bigcup_{n\in\mathbb{N}} F_n,$$

we have

$$\mu[\,F\,] \leqslant \sum_{n\in\mathbb{N}} \mu[\,F_n\,].$$

Proof of Lemma 19.2.6. Set

$$B_n = \bigcup_{k=1}^{n} F_n, \quad B = \bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} F_n.$$

Then $B \supset F$,

$$F \cap B_1 \subset F \cap B_2 \subset \cdots \quad \text{and} \quad F \cap B = F.$$

We have $F \cap B_n \in \mathcal{F}$ and the finite additivity of μ implies that

$$\mu\big[\,F \cap B_n\,\big] \leqslant \mu\big[\,B_n\,\big] \leqslant \sum_{k=1}^{n} \mu\big[\,F_k\,\big].$$

The *conditional countable additivity* of μ implies that

$$\mu\big[\,F\,\big] = \lim_{n\to\infty} \mu\big[\,F \cap B_n\,\big] \leqslant \sum_{k=1}^{\infty} \mu\big[\,F_k\,\big].$$

□

Let us mention that the conditional countable additivity assumption is only needed in the prove of the above lemma. The arguments to follow do not require this assumption, if the validity of Lemma 19.2.6 is granted.

(i) Let $F \in \mathcal{F}$. Observe that $\widehat{\mu}\big[\,F\,\big] \leqslant \mu\big[\,F\,\big]$ since F covers itself.

On the other hand, Lemma 19.2.6 implies that for any $\mathcal{F}$-cover $(F_n)_{n\in\mathbb{N}}$ of F we have

$$\mu\big[\,F\,\big] \leqslant \sum_{n\in\mathbb{N}} \mu\big[\,F_n\,\big]$$

so that $\mu\big[\,F\,\big] \leqslant \widehat{\mu}\big[\,F\,\big]$. This proves (i).

(ii) It suffices to show that $\mathcal{F} \subset \mathcal{S}_{\widehat{\mu}}$. Let $F \in \mathcal{F}$. We want to prove that for any $A \subset \Omega$ we have

$$\widehat{\mu}\big[\,A\,\big] \geqslant \widehat{\mu}\big[\,A \cap F\,\big] + \widehat{\mu}\big[\,A \cap F^c\,\big]. \tag{19.2.4}$$

Fix $A \subset \Omega$. The above equality is true if $\widehat{\mu}\big[\,A\,\big] = \infty$ so we only need to consider the case $\widehat{\mu}\big[\,A\,\big] < \infty$. For any $\varepsilon > 0$ there exists an $\mathcal{F}$-cover $(F_n)_{n\in\mathbb{N}}$ of A such that

$$\widehat{\mu}\big[\,A\,\big] + \varepsilon \geqslant \sum_{n\in\mathbb{N}} \mu\big[\,F_n\,\big] \geqslant \widehat{\mu}\big[\,A\,\big].$$

Since μ is additive and $F_n = F_n \cap F \sqcup F_n \cap F^c$ we deduce

$$\sum_{n\in\mathbb{N}} \mu\big[\,F_n\,\big] = \sum_{n\in\mathbb{N}} \mu\big[\,F_n \cap F\,\big] + \sum_{n\in\mathbb{N}} \mu\big[\,F_n \cap F^c\,\big] \geqslant \widehat{\mu}\big[\,A \cap F\,\big] + \widehat{\mu}\big[\,A \cap F^c\,\big],$$

where the last inequality follows from the fact that the families $(F_n \cap F)$ and $(F_n \cap F^c)$ are $\mathcal{F}$-covers of $A \cap F$ and respectively $A \cap F^c$. Hence, for any $\varepsilon \geqslant 0$

$$\widehat{\mu}\big[\,A\,\big] + \varepsilon \geqslant \widehat{\mu}\big[\,A \cap F\,\big] + \widehat{\mu}\big[\,A \cap F^c\,\big].$$

Letting $\varepsilon \searrow 0$ we obtain (19.2.4).

If μ is finite, then Proposition 19.1.34 shows that it admits a unique extension to a measure on $\sigma(\mathcal{F})$.

Suppose now that μ is sigma-finite and $\bar{\mu} : \sigma(\mathcal{F}) \to [0,\infty)$ is a measure extending μ. Fix an increasing sequence $(F_n)_{n\in\mathbb{N}}$ of sets in $\mathcal{F}$ such that

$$\mu\big[\,F_n\,\big] < \infty,\ \ \forall n \in \mathbb{N},$$

and

$$\Omega = \bigcup_{n\in\mathbb{N}} F_n.$$

For $n \in \mathbb{N}$ we define

$$\bar{\mu}_n,\ \ \widehat{\mu}_n : \sigma(\mathcal{F}) \to [0,\infty],$$
$$\bar{\mu}_n[\,S\,] := \bar{\mu}[\,S \cap F_n\,],\ \ \widehat{\mu}_n[\,S\,] := \widehat{\mu}[\,S \cap F_n\,],\ \ \forall S \in \sigma(\mathcal{F})$$

Note that $\bar{\mu}_n$ and $\widehat{\mu}_n$ are finite measures that coincide on $\mathcal{F}$, and $\bar{\mu}_n[\,\Omega_n\,] = \mu[\,F_n\,] = \widehat{\mu}_n[\,\Omega\,]$. We deduce from Proposition 19.1.34 that $\bar{\mu}_n = \widehat{\mu}_n$, $\forall n$, so that

$$\forall S \in \sigma(\mathcal{F}) :\ \ \bar{\mu}[\,S \cap F_n\,] = \widehat{\mu}[\,S \cap F_n\,].$$

Letting $n \to \infty$ in the above equality we deduce that

$$\bar{\mu}[\,S\,] = \widehat{\mu}[\,S\,],\ \ \forall S \in \sigma(\mathcal{F}).$$

□

In Exercise 19.21 we describe a generalization of Theorem 19.2.5.

Definition 19.2.7. Suppose that (X,d) is a metric space. An outer measure $\mu : 2^X \to [0,\infty]$ is said to be *metric* if for any $S_1, S_2 \subset X$,

$$\operatorname{dist}(S_1,S_2) > 0 \Rightarrow \mu[\,S_1 \cup S_2\,] = \mu[\,S_1\,] + \mu[\,S_2\,], \tag{19.2.5}$$

where

$$\operatorname{dist}(S_1,S_2) := \inf_{(x_1,x_2)\in S_1\times S_2} d(x_1,x_2).$$

□

Theorem 19.2.8 (Carathéodory)**.** *Suppose that μ is a* metric *outer measure on the metric space (X,d). Then any Borel subset of X is μ-measurable, so μ defines a measure on the Borel sigma-algebra $\mathcal{B}_X$. Moreover, $\mathcal{S}_\mu$ contains the μ-completion of $\mathcal{B}_X$.*

Proof. It suffices to show that any closed subset $C \subset X$ is μ-measurable, i.e.,

$$\mu[\,A\,] \geqslant \mu[\,A \cap C\,] + \mu[\,A \backslash C\,],\ \ \forall A \subset X. \tag{19.2.6}$$

Let $A \subset X$. The equality (19.2.6) is obviously true if $\mu[\,A\,] = \infty$ so we assume that $\mu[\,A\,] < \infty$. Set

$$A_n := \{\, a \in A;\ \ \operatorname{dist}(a,C) \geqslant 1/n \,\}.$$

Note that

$$A_1 \subset A_2 \subset \cdots \ \text{ and } \bigcup_{n\mathbb{N}} A_n = \{\, a \in A;\ \ \operatorname{dist}(a,C) > 0 \,\} = A\backslash C.$$

Since μ is a metric outer measure we deduce from (19.2.5) that

$$\mu[\,A \cap C\,] + \mu[\,A_n\,] = \mu[\,(A \cap C) \cup A_n\,] \leqslant \mu[\,A\,].$$

To prove (19.2.6) it suffices to show that

$$\mu[\,A_n\,] \to \mu[\,A\backslash C\,] \tag{19.2.7}$$

Set $C_n := A_{m+1}\backslash A_n$, $\forall n$. The clincher is the simple observation that $\operatorname{dist}(C_m, C_n) > 0$ if $|m-n| \geqslant 2$. Using this (19.2.5) we deduce

$$\sum_{j=1}^J \mu[\,C_{2j}\,] = \mu\Big[\,\bigcup_{j=1}^J C_{2j}\,\Big] \leqslant \mu[\,A\,] < \infty$$
$$\sum_{j=1}^J \mu[\,C_{2j-1}\,] = \mu\Big[\,\bigcup_{j=1}^J C_{2j-1}\,\Big] \leqslant \mu[\,A\,] < \infty.$$

Hence

$$\sum_{n\geqslant 1} \mu[\,C_n\,] < \infty.$$

From the countable subadditivity of μ we deduce

$$\mu[\,A\backslash C\,] \leqslant \mu[\,A_m\,] + \underbrace{\sum_{n\geqslant m}\mu[\,C_n\,]}_{=:t_m} \leqslant \mu[\,A\,] + t_m$$

The equality (19.2.7) follows by letting $m \to \infty$ and observing that $t_m \to 0$. □

19.2.2. The Lebesgue measure on the real line.

In Subsection 19.1.6 we constructed the Lebesque premeasure $\boldsymbol{\lambda}$ defined on the algebra of sets $\mathcal{F}$ generated by the set $\mathcal{I}$ of intervals of the form

$$\mathbb{R},\ \ (a,b],\ \ -\infty \leqslant a < b < \infty.$$

It is uniquely determined by the conditions

$$\boldsymbol{\lambda}\big[\,(a,b]\,\big] = b-a,\ \ \forall-\infty < a < b \leqslant \infty.$$

The Carathéodory extension of $\boldsymbol{\lambda}$ is defined on a sigma-algebra $\mathcal{S}_{\boldsymbol{\lambda}}$ containing $\sigma(\mathcal{F})$ and the resulting measure

$$\boldsymbol{\lambda} : \mathcal{S}_{\boldsymbol{\lambda}} \to [0,\infty]$$

is called the *Lebesgue measure* on the real axis. The subsets in $\mathcal{S}_{\boldsymbol{\lambda}}$ are called *Lebesgue measurable*. Observe that the sigma-algebra generated by $\mathcal{F}$ is the Borel sigma algebra $\mathcal{B}_{\mathbb{R}}$ generated by the open subsets of $\mathbb{R}$. The Carathéodory construction shows that the $\boldsymbol{\lambda}$-completion of $\mathcal{B}_{\mathbb{R}}$ is contained in $\mathcal{S}_{\boldsymbol{\lambda}}$,

$$\mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}} \subset \mathcal{S}_{\boldsymbol{\lambda}}.$$

Remark 19.2.9. Let us recall main steps in the construction of the Lebesgue measure. Using $\mathcal{F}$ as class of models and $\boldsymbol{\lambda} : \mathcal{F} \to [0,\infty]$ as gauge we construct the outer measure

$$\widehat{\boldsymbol{\lambda}} : 2^{\mathbb{R}} \to [0,\infty].$$

More explicitly, given $S \subset \mathbb{R}$, then $\widehat{\boldsymbol{\lambda}}\big[\,S\,\big] = c$ if and only if the following hold.

(i) For any countable cover $(I_n)_{n\in\mathbb{N}}$ of S by intervals $I_n = (a_n, b_n]$ we have

$$c \leqslant \sum_{n\in\mathbb{N}} \boldsymbol{\lambda}\big[\,I_n\,\big] = \sum_{n\in\mathbb{N}}(b_n - a_n).$$

(ii) For any $\varepsilon > 0$ there exists a countable cover $(I_n)_{n\in\mathbb{N}}$ of S by intervals $I_n = (a_n, b_n]$ such that

$$\sum_{n\in\mathbb{N}}(b_n - a_n) \leqslant c + \varepsilon.$$

The Lebesgue measure of a Borel subset $B \subset \mathbb{R}$ is then $\widehat{\boldsymbol{\lambda}}\big[\,B\,\big]$. Moreover, if S happens to be in $\mathcal{F}$, then $\widehat{\boldsymbol{\lambda}}\big[\,S\,\big] = \boldsymbol{\lambda}\big[\,S\,\big]$. □

Example 19.2.10 (The Cantor set)**.** For each closed and bounded interval $I = [a,b]$ we denote by $C(I)$ the union of intervals obtained by removing from I the open middle third

$$C(I) := L(I) \cup R(I),\ \ L(I) := [a, a+(b-a)/3], R(I) := [b-(b-a)/3, b].$$

Thus $L(I)$ is the left third of I, while $R(I)$ is the right third of I.

More generally, if S is a union of disjoint compact intervals,

$$S = I_1 \cup \cdots \cup I_n,$$

we set

$$C(S) := C(I_1) \cup \cdots \cup C(I_n).$$

Note that $C(S)$ is itself a union of disjoint compact intervals and $C(S) \subset S$.

We denote by $\mathfrak{X}$ the collection of subsets that are unions of finitely many disjoint compact intervals. We have thus obtained a map

$$C : \mathfrak{X} \to \mathfrak{X}, \;\; S \mapsto C(S).$$

Note that $C(S) \subset S$ and

$$\boldsymbol{\lambda}\big[\, C(S) \,\big] = \frac{2}{3}\boldsymbol{\lambda}\big[\, S \,\big].$$

Consider the sequence of subsets $S_n \in \mathfrak{X}$ defined recursively as

$$S_0 := [0,1], \;\; S_n = C(S_{n-1}), \;\; \forall n \in \mathbb{N}.$$

Observe that

$$S_0 \supset S_1 \supset \cdots \supset S_n \cdots, \;\; \boldsymbol{\lambda}\big[\, S_n \,\big] = \left(\frac{2}{3}\right)^n.$$

The sets S_n are nonempty and compact so

$$S_\infty := \bigcap_{n \geqslant 0} S_n$$

is a nonempty compact subset called the *Cantor set*. It is a negligible set since

$$\boldsymbol{\lambda}\big[\, S_\infty \,\big] = \lim_{n \to \infty} \boldsymbol{\lambda}\big[\, S_n \,\big] = 0.$$

On the other hand S_∞ is very large. To see this we consider the following infinite binary tree.

It has a root labelled S_0. It has two successors, the components of $C(S_0)$, $L(S_0)$ and $R(S_0)$. We obtain the first generation of vertices consiting of two vertices labelled $L(S_0)$ and $R(S_0)$. Inductively, the n-th generation consists of 2^n vertices (the components of S_n). Each vertex has two successors, a left and a right successor; see Figure 19.1.

Observe that there is a bijection

$$\{L, R\}^{\mathbb{N}} \to S_\infty, \;\; \{L, R\}^{\mathbb{N}} \ni \underline{A} \mapsto x(\underline{A}) \in S_\infty. \tag{19.2.8}$$

More precisely, given a sequence $\underline{A} = (A_1, A_2, A_3, \dots,) \in \{L, R\}^{\mathbb{N}}$ ($A_n = L, R$) we obtain a nested sequence of intervals $I_n = I_n(\underline{A})$ defined by

$$I_1 = A_1(S_0), \;\; I_{n+1} = A_{n+1}(I_n) \subset I_n, \;\; n \in \mathbb{N}.$$

The n-th interval I_n has length 3^{-n}. The intersection of this sequence of nested intervals consists of a single point $x(\underline{A})$ that obviously belongs to the Cantor set.

Note that if $\underline{A} \neq \underline{B}$, then there exists $n \in \mathbb{N}$ such that $I_n(\underline{A}) \cap I_n(\underline{B}) = \varnothing$ so that $x(\underline{A}) \neq x(\underline{B})$. Thus the map (19.2.8) is injective.

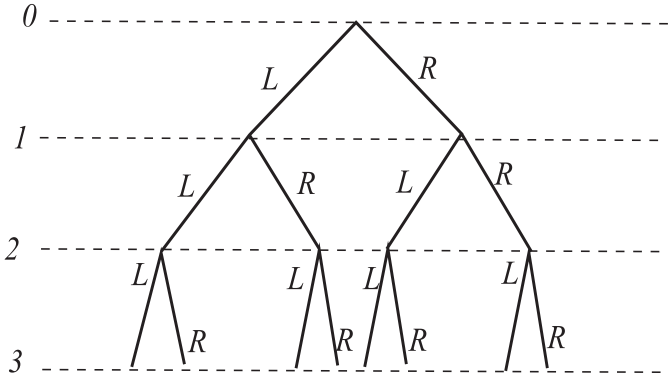

Figure 19.1. *Three generations of an infinite binary tree.*

By construction this map is surjective because any x in the Cantor set lives in the intersection of such a sequence of intervals. Thus

$$\operatorname{card} S_\infty = 2^{\aleph_0} = \aleph_c = \operatorname{card}[0,1],$$

where at the last step we invoked Theorem A.3.7. □

In the remainder of this subsection we will try to understand how large is the collection of Lebesgue measurable subsets of the real axis. By construction, $\mathcal{S}_{\boldsymbol{\lambda}}$ contains the Borel sigma-algebra $\mathcal{B}_{\mathbb{R}}$. Also by construction, the sigma-algebra $\mathcal{S}_{\boldsymbol{\lambda}}$ is $\boldsymbol{\lambda}$-complete and thus it must also contain the $\boldsymbol{\lambda}$-completion $\mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}}$ of $\mathcal{B}_{\mathbb{R}}$, i.e., $\mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}} \subset \mathcal{S}_{\boldsymbol{\lambda}}$.

Proposition 19.2.11. $\mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}} = \mathcal{S}_{\boldsymbol{\lambda}}$.

Proof. Let us introduce some classical terminology. A *G_δ-subset* of $\mathbb{R}$ is defined to be the intersection of countably many open subsets. An *F_σ-subset* is the complement of a G_δ-set or, equivalently, the union of countably many closed sets. Clearly the G_δ-sets and F_σ-sets are Borel measurable.

We will show that for any Lebesgue measurable set S there exist Borel sets A, B such that

$$A \subset S \subset B \ \text{ and } \ \boldsymbol{\lambda}\big[\,S\,\big] = \boldsymbol{\lambda}\big[\,A\,\big] = \boldsymbol{\lambda}\big[\,B\,\big].$$

We carry the proof in two steps.

Step 1. We prove that the claim is true when S is bounded, i.e., there exists $R > 0$ such that

$$S \subset (-R, R).$$

We first show that there exists a G_δ-set G containing S such that

$$\boldsymbol{\lambda}\big[\,S\,\big] = \boldsymbol{\lambda}\big[\,G\,\big].$$

For any $\varepsilon > 0$ there exists a countable family of intervals $(I_n)_{n\in\mathbb{N}}$ in $\mathcal{I}$ such that

$$I_n \subset (-R-1, R+1), \ \ \forall n, \ \ S \subset \bigcup_{n\in\mathbb{N}} I_n$$

and

$$\sum_{n\in\boldsymbol{n}} \boldsymbol{\lambda}\big[\,I_n\,\big] \leqslant \boldsymbol{\lambda}\big[\,S\,\big] + \varepsilon.$$

The intervals I_n are not open, but each is contained in an *open* interval $J_n = J_n^\varepsilon \subset (-R-1, R+1)$ such that

$$\boldsymbol{\lambda}\big[\, J_n \,\big] \leqslant \boldsymbol{\lambda}\big[\, I_n \,\big] + \frac{\varepsilon}{2^n}.$$

Set

$$J^\varepsilon := \bigcup_{n\in\mathbb{N}} J_n.$$

Then $S \subset J^\varepsilon$ and

$$\boldsymbol{\lambda}\big[\, S \,\big] \leqslant \boldsymbol{\lambda}\big[\, J^\varepsilon \,\big] \leqslant \sum_{n\in\boldsymbol{n}} \boldsymbol{\lambda}\big[\, I_n \,\big] + \varepsilon \leqslant \boldsymbol{\lambda}\big[\, S \,\big] + 2\varepsilon.$$

Set

$$G := \bigcap_{n\in\mathbb{N}} J^{1/n}.$$

Then G is a G_δ-set containing S and $\boldsymbol{\lambda}\big[\, G \,\big] = \boldsymbol{\lambda}\big[\, S \,\big]$.

Consider the set $T = [-R, \times R]\backslash S$. It is a bounded Lebesgue measurable and thus there exists a G_δ set $G \supset T$ such that $\boldsymbol{\lambda}\big[\, G \,\big] = \boldsymbol{\lambda}\big[\, T \,\big]$. The set $F = [-R, R]\backslash G$ is an F_σ set contained in S that has the same Lebesgue measure as S.

Step 2. The general case. For every $n \in \mathbb{N}$ set

$$S_n = \big(\, (-n, n-1] \cup [n-1, n) \,\big) \cap S.$$

From Step 1 we deduce that for any $n \in \mathbb{N}$ there exists a G_δ-set $G_n \supset S_n$ and an F_σ-set $F_n \subset S_n$ such that

$$\boldsymbol{\lambda}\big[\, F_n \,\big] = \boldsymbol{\lambda}\big[\, S_n \,\big] = \boldsymbol{\lambda}\big[\, G_n \,\big].$$

Set

$$A = \bigcup_{n\in\mathbb{N}} F_n, \quad B = \bigcup_{n\in\mathbb{N}} G_n$$

Clearly B is a Borel set and

$$A \subset S = \bigcup_{n\in\mathbb{N}} S_n \subset B$$

Note that

$$\boldsymbol{\lambda}\big[\, S\backslash A \,\big] = \sum_{n\in\mathbb{N}} \boldsymbol{\lambda}\big[\, S_n\backslash F_n \,\big] = 0$$

and

$$\boldsymbol{\lambda}\big[\, B\backslash S \,\big] \leqslant \sum_{n\in\mathbb{N}} \boldsymbol{\lambda}\big[\, G_n\backslash S_n \,\big] = 0.$$

This shows that any Lebesgue measurable set differs from a Borel set by a Lebesgue negligible subset.

□

Let us observe another property of Lebesgue measurable sets. For any $S \subset \mathbb{R}$ and $r \in \mathbb{R}$ we denote by $S + r$ the set

$$S + r := \big\{\, s + r;\;\; s \in S \,\big\}.$$

Equivalently

$$x \in S + r \Longleftrightarrow x - r \in S.$$

In other words $S+r$ is the preimage of S via the homeomorphism $h_r : \mathbb{R} \to \mathbb{R}$, $h_r(x) = x-r$. Since the preimage of a Borel set via a continuous map $\mathbb{R} \to \mathbb{R}$ is also Borel we deduce

$$S \in \mathcal{B}_{\mathbb{R}} \Rightarrow S + r \in \mathcal{B}_{\mathbb{R}}, \;\; \forall r \in \mathbb{R}.$$

Proposition 19.2.12. *For any $S \in \mathcal{S}_{\boldsymbol{\lambda}}$ and any $r \in \mathbb{R}$ we have $S + r \in \mathcal{S}_{\boldsymbol{\lambda}}$ and*

$$\boldsymbol{\lambda}\big[\, S + r \,\big] = \boldsymbol{\lambda}\big[\, S \,\big].$$

Proof. The proof is a simple application of Dynkin's $\pi - \lambda$ theorem. For simplicity we write

$$\boldsymbol{\lambda}_r[\,S\,] := \boldsymbol{\lambda}[\,S + r\,],\ \ \forall r \in \mathbb{R}.$$

Let

$$\mathcal{A} := \big\{\, S \in \mathcal{S}_{\boldsymbol{\lambda}};\ \ \boldsymbol{\lambda}_r[\,A\,] = \boldsymbol{\lambda}[\,A\,], \big\},\ \ A \subset [-n,n]\,\big\}.$$

We have to show that $\mathcal{A} = \mathcal{S}_{\boldsymbol{\lambda}}$. For $n \in \mathbb{N}$ we set

$$\mathcal{A}_n := \big\{\, A \in \mathcal{A};\ \ A \subset [-n,n]\,\big\}.$$

We will show that all the Borel subsets of $[-n,n]$ are contained in $\mathcal{A}_n$.

Observe first that $\mathcal{A}_n$ contains all the intervals of the form $[-n,a]$, $a \in [-n,n]$. The collection $\mathcal{P}$ of these intervals is a π-system of subsets of $[-n,n]$, i.e., it is closed under intersections. Obviously $\varnothing, [-n,n] \in \mathcal{A}_n$.

Next, observe that if $A, B \in \mathcal{A}_n$, $A \subset B$, then $B \backslash A \in \mathcal{A}_n$. Indeed

$$(B\backslash A) + r = (B + r)\backslash(A + r)$$

so that

$$\boldsymbol{\lambda}_r[\,B\backslash A\,] = \boldsymbol{\lambda}[\,(B+r)\backslash(A+r)\,] = \boldsymbol{\lambda}[\,B + r\,] - \boldsymbol{\lambda}[\,A + r[$$

$$= \boldsymbol{\lambda}[\,B\,] - \boldsymbol{\lambda}[\,A\,] = \boldsymbol{\lambda}[\,B\backslash A\,].$$

Finally, observe that if $(A_k)_{k\in\mathbb{N}}$ is an increasing sequence of sets in $\mathcal{A}_n$, and

$$A = \bigcup_{n\in\mathbb{N}} A_n,$$

then

$$\bigcup_{k\in\mathbb{N}} (A_k + r) = A + r$$

and we have

$$\boldsymbol{\lambda}[\,A + r\,] = \lim_{n\to\infty} \boldsymbol{\lambda}[\,A_n + r\,] = \lim_{n\to\infty} \boldsymbol{\lambda}[\,A_n\,] = \boldsymbol{\lambda}[\,A\,]$$

so that $A \in \mathcal{A}_n$. This proves that $\mathcal{A}_n$ is also λ-system of subsets of $[-n,n]$ and the $\pi - \lambda$-theorem implies that is contains the sigma-algebra generated by $\mathcal{P}$. This is the Borel algebra $\mathcal{B}_{[-C,C]}$. Thus for any Borel subset $B \subset \mathbb{R}$ we have

$$\boldsymbol{\lambda}[\,B \cap [-n,n]\,] = \boldsymbol{\lambda}_r[\,B \cap [-n,n]\,].$$

Letting $n \to \infty$ we deduce $\boldsymbol{\lambda}[\,B\,] = \boldsymbol{\lambda}_r[\,B\,]$. Thus $\mathcal{B}_{\mathbb{R}} \subset \mathcal{A}$ so that $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}$ on $\mathcal{B}_R$.

If $S \subset \mathbb{R}$ is $\boldsymbol{\lambda}$-negligible, then there exists $N \in \mathcal{B}_{\mathbb{R}}$ such that $\boldsymbol{\lambda}[\,N\,] = 0$ and $S \subset N$. Note that

$$\boldsymbol{\lambda}[\,N + r\,] = \boldsymbol{\lambda}_r[\,N\,] = \boldsymbol{\lambda}[\,N\,] = 0.$$

Hence S is also $\boldsymbol{\lambda}_r$ negligible. Thus $\mathcal{B}^{\boldsymbol{\lambda}} = \mathcal{B}^{\boldsymbol{\lambda}_r}$ and thus $\boldsymbol{\lambda}$ coincides with $\mathcal{B}^{\boldsymbol{\lambda}} = \mathcal{S}_{\boldsymbol{\lambda}}$. □

As we have seen the Cantor set is negligible and has the same cardinality as $\mathbb{R}$. Hence, any subset of the Cantor set is negligible so that $\operatorname{card} \mathcal{S}_{\boldsymbol{\lambda}} \geqslant 2^{\operatorname{card}\mathbb{R}} = 2^{\aleph_c}$. Obviously since $\mathcal{S}_{\boldsymbol{\lambda}} \subset 2^{\mathbb{R}}$ we deduce $\operatorname{card} \mathcal{S}_{\boldsymbol{\lambda}} \leqslant 2^{\operatorname{card}\mathbb{R}}$. Hence $\operatorname{card} \mathcal{S}_{\boldsymbol{\lambda}} = 2^{\operatorname{card}\mathbb{R}}$. Is it possible that $\mathcal{S}_{\boldsymbol{\lambda}} = 2^{\mathbb{R}}$, i.e., any subset of $\mathbb{R}$ is Lebesgue measurable? Our next result shows that this is not the case.

Proposition 19.2.13 (G. Vitali)**.** *There exist subsets of* $\mathbb{R}$ *that are not Lebesgue measurable.*

Proof. Define a binary relation "$\sim$" on $\mathbb{R}$ by declaring $x \sim y$ if $y - x \in \mathbb{Q}$. Clearly $x \sim x$. The relation is symmetric since

$$x \sim y \Longleftrightarrow y - x \in \mathbb{Q} \Longleftrightarrow x - y \in \mathbb{Q} \Longleftrightarrow y \sim x.$$

Finally, the relation is transitive because if $x \sim y$ and $y \sim z$ then $(z-y), (y-x) \in \mathbb{Q}$ so that

$$z - x = (z-y) + (y-x) \in \mathbb{Q}$$

i.e., $x \sim z$. Using the axiom of choice (see page 1016) there exists a complete set of representatives of this equivalence relation, i.e., a subset $S \subset \mathbb{R}$ such that any real number is equivalent with exactly one element of S. By replacing each element $s \in S$ by its fractional part $\{s\} = s - \lfloor s \rfloor$ we can assume that $S \subset [0,1)$.

Consider the countable family of translates

$$(S+q)_{q \in \mathbb{Q}}.$$

Observe that any two sets in this collection are disjoint. Indeed, if $x \in (S+q_1) \cap (S+q_2)$ then there exist $x_1, x_2 \in S$ such that $x = x_1 + q_1 = x_2 + q_2$. Thus $x_2 - x_1 = q_1 - q_2 \in \mathbb{Q}$. Since no two distinct elements in S are equivalent we conclude that $x_1 = x_2$ so $q_1 = q_2$.

Observe next that

$$\mathbb{R} = \bigcup_{q \in \mathbb{Q}} (S+q).$$

Indeed, for any $x \in \mathbb{R}$ there exists $s \in S$ such that $s \sim x$, i.e., $x - s \in \mathbb{Q}$. If we write $q := x - s$, then $x = s + q$ so that $x \in S + q$.

We claim that the set S is not Lebesgue measurable. We argue by contradiction.

Suppose that S is Lebesgue measurable. We deduce that $S+q$ is Lebesgue measurable $\forall q \in \mathbb{Q}$ and thus

$$\infty = \boldsymbol{\lambda}\big[\,\mathbb{R}\,\big] = \sum_{q \in \mathbb{Q}} \boldsymbol{\lambda}\big[\,S+q\,\big].$$

Hence, there exists $q_0 \in \mathbb{Q}$ such that

$$\boldsymbol{\lambda}\big[\,S+q_0\,\big] > 0.$$

Since $\boldsymbol{\lambda}\big[\,S\,\big] = \boldsymbol{\lambda}\big[\,S+q_0\,\big]$ we deduce

$$\boldsymbol{\lambda}\big[\,S\,\big] > 0.$$

Now observe that

$$\bigcup_{q \in [0,1] \cap \mathbb{Q}} (S+q) \subset [0,2],$$

so that

$$\sum_{q \in [0,1] \cap \mathbb{Q}} \boldsymbol{\lambda}\big[\,S\,\big] = \sum_{q \in [0,1] \cap \mathbb{Q}} \boldsymbol{\lambda}\big[\,(S+q)\,\big] < 2.$$

This shows that $\boldsymbol{\lambda}\big[\,S\,\big] = 0$. This contradiction shows that S is not Lebesgue measurable. □

Remark 19.2.14. We have three important sigma-algebras of subsets of $\mathbb{R}$: the sigma-algebra $\mathcal{B}_{\mathbb{R}}$ of all Borel subsets, the sigma-algebra $\mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}}$ of Lebesgue measurable subsets, and the sigma-algebra $2^{\mathbb{R}}$ of all the subsets of $\mathbb{R}$. We have inclusions

$$\mathcal{B}_{\mathbb{R}} \subset \mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}} \subset 2^{\mathbb{R}}.$$

We have shown that the second inclusion is strict.

One can show that (see [**25**, Chap.10])

$$\operatorname{card} \mathcal{B}_{\mathbb{R}} = 2^{\aleph_0} = \aleph_c < 2^{\aleph_c} = \operatorname{card} \mathcal{B}_{\mathbb{R}}^{\boldsymbol{\lambda}}.$$

This shows that "most" Lebesgue measurable sets are *not* Borel measurable.

The Lebesgue measure $\boldsymbol{\lambda}$ was constructed in a rather special way. It is defined on a sigma-algebra containing all the intervals $(a,b]$ and satisfies

$$\boldsymbol{\lambda}\big[\,(a,b]\,\big] = b-a.$$

Is it possible that, by some other method, we could construct a measure μ on the sigma-algebra of *all the subsets* of $\mathbb{R}$ such that $\mu\big[\,(a,b]\,\big] = b-a$, $\forall a<b$?

The surprising answer is NO, this is not possible! This fact is a consequence of the Axiom of Choice. For details we refer to [**25**, Sec. 8.2]. □

19.2.3. Lebesgue-Stiltjes measures. Suppose that $F:\mathbb{R}\to\mathbb{R}$ is a gauge function, i.e., a nondecreasing, right-continuous function. As explained in Remark 19.1.48, the function F defines a sigma-additive premeasure $\boldsymbol{\lambda}_F$ and thus extends to a measure

$$\boldsymbol{\lambda}_F : (\mathbb{R},\mathcal{B}_{\mathbb{R}})\to[0,\infty]$$

called the *Lebesgue-Stiltjes measure* associated to the gauge function F. Observe that this measure satisfies the finiteness condition

$$\mu_F\big[\,(a,b]\,\big] = F(b)-F(a)<\infty,\ \ \forall a<b.$$

Clearly this is equivalent with the condition $\mu_F\big[\,K\,\big]<\infty$, for any compact subset of $\mathbb{R}$.

When $F(\infty)=1$ and $F(-\infty)=0$ the resulting measure $\boldsymbol{\lambda}_F$ is a probability measure on $\mathcal{B}_{\mathbb{R}}$ and the function F is classically known as the *cumulative distribution function* of this probability measure. In this case the function $F(x)$ is uniquely determined by the equality $F(x)=\boldsymbol{\lambda}_F\big[\,(-\infty,x]\,\big]$.

Conversely, suppose that $\mu:(\mathbb{R},\mathcal{B}_{\mathbb{R}})\to[0,\infty]$ is a finite measure satisfying the above finiteness condition. We want to show that there exists a gauge function F such that $\mu=\boldsymbol{\lambda}_F$.

We begin by defining $G:(0,\infty)\to[0,\infty)$, $G(x)=\mu\big[\,(0,x]\,\big]$. The function G is nondecreasing and thus it has at most countably many points of discontinuity.[1] Thus G is continuous at some point $x_0\in(0,\infty)$. Define $F:\mathbb{R}\to[0,\infty)$

$$F(x)=\begin{cases}\mu\big[\,(x_0,x]\,\big], & x>x_0,\\ 0, & x=x_0,\\ -\mu\big[\,(x,x_0]\,\big], & x<x_0.\end{cases}$$

Clearly F is nondecreasing and

$$\mu\big[\,(a,b]\,\big]=F(b)-F(a),\ \ \forall a,b.$$

Let show that

$$F(x)=\lim_{y\searrow x}F(y),\ \ \forall x\in\mathbb{R}.$$

This is obviously true for $x>x_0$. Next observe that

$$0\leqslant F(x_0+\varepsilon)\leqslant F(x_0+\varepsilon)-F(x_0-\varepsilon)$$

[1] Can you prove this?

$$= \mu\big[\,(x_0-\varepsilon, x_0+\varepsilon]\,\big] = G(x_0+\varepsilon) - G(x_0-\varepsilon) \to 0 \ \text{ as } \varepsilon \searrow 0$$

since G is continuous at x_0. Hence F is right continuous at x_0.

Suppose now that $x < x_0$. For $y \in (x, x_0]$ we have

$$F(y) - F(x) = \mu\big[\,(x,y]\,\big] \to 0 \ \text{ as } y \searrow 0.$$

Clearly $\boldsymbol{\lambda}_F = \mu$ since

$$\boldsymbol{\lambda}_F\big[\,(a,b]\,\big] = \mu\big[\,(a,b]\,\big] = F(b) - F(a), \ \ \forall a,b.$$

Note that if $\mu\big[\,\mathbb{R}\,\big] < \infty$, then as gauge function we can choose

$$F(x) = \mu\big[\,(-\infty, x]\,\big].$$

It satisfies

$$F(-\infty) := \lim_{x\to-\infty} F(x) = 0, \ \ F(\infty) = \mu\big[\,\mathbb{R}\,\big] < \infty. \tag{19.2.9}$$

Suppose that we have a gauge function F satisfying the conditions (19.2.9) above. We set $M := F(\infty)$. The *quantile* function of F is the function

$$Q : [0,M] \to [-\infty,\infty], \ \ Q(y) = \inf\big\{\, x \in \mathbb{R};\ \ F(x) \geqslant y \,\big\}. \tag{19.2.10}$$

Proposition 19.2.15. *Suppose that F is a gauge function satisfying (19.2.9). Denote by Q its quantile function Q.*

(i) *The function Q is nondecreasing, $Q(F(x)) \leqslant x$, $\forall x \in [-\infty,\infty]$.*

(ii) $Q^{-1}\big(\,[-\infty,x]\,\big) = [0, F(x)]$.

(iii) *The function Q is left continuous, i.e., for any $y_0 \in [0,M]$ we have*

$$\lim_{y\nearrow y_0} Q(y) = Q(y_0).$$

(iv) $Q_{\#}\boldsymbol{\lambda}_{[0,M]} = \boldsymbol{\lambda}_F$ *where $\boldsymbol{\lambda}_{[0,M]}$ is the Lebesgue measure on the interval $[0,M]$ and $Q_{\#}$ denotes the pushforward by Q defined in Example 19.1.33(iv)*

□

The proof is left to you as an exercise.

19.3. The Lebesgue integral

In this section we describe a method of integration pioneered by H. Lebesgue. To contrast it with the Riemann method of integration consider the following experiment. Suppose you have a large pile of coins consisting of pennies, nickels, dimes and quarters and you want to find out the total worth of that pile. There are two ways to do this.

The Riemann way is to successively add the values the coins, one by one. Lebesgue's way is to separate the coins into piles according to their values, the penny pile, the nickel pile etc., find the value of each of these piles by counting the number of coins in it and adding the results.

The counting part of Lebesgue's approach is abstractly encoded by a measure, so each choice of measure leads to a different process of integration. Our presentation is greatly inspired from the presentation in [6, I.4].

19.3.1. Definition and fundamental properties. Fix a measurable space $(\Omega, \mathcal{S})$. Recall some notation.

- $\bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ denotes the space of measurable functions $(\Omega, \mathcal{S}) \to [-\infty, \infty]$ and $\bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ denotes the set of nonnegative ones.
- $\mathcal{L}^0(\Omega, \mathcal{S})$ denotes the space of measurable functions $(\Omega, \mathcal{S}) \to (-\infty, \infty)$ and $\mathcal{L}^0_+(\Omega, \mathcal{S})$ denotes the set of nonnegative ones.

A measurable function $f \in \mathcal{L}^0(\Omega, \mathcal{S})$ is called *elementary* or *step function* if there exist finitely many *disjoint* measurable sets $A_1, \dots, A_N \in \mathcal{S}$ and real numbers $c_1, \dots, c_N$ such that

$$f(\omega) = \sum_{k=1}^{N} c_k \boldsymbol{I}_{A_k}(\omega), \quad \forall \omega \in \Omega. \tag{19.3.1}$$

We denote by $\mathscr{E}(\Omega, \mathcal{S})$ the set of elementary functions and by $\mathscr{E}_+(\Omega, \mathcal{S})$ the subspace of nonnegative elementary functions.

Let us observe $f \in \mathscr{E}(\Omega, \mathcal{S})$ if and only if there exists a finite measurable partition partition $A_0, A_1, \dots, A_n$ and real numbers $c_0, c_1, \dots, c_n$ such that

$$f = \sum_{k=0}^{n} c_k \boldsymbol{I}_{A_k}.$$

Indeed, suppose that f is elementary. Then there exist disjoint measurable sets $A_1, \dots, A_n$ and real numbers $c_1, \dots, c_n$ such that

$$f = \sum_{k=1}^{n} c_k \boldsymbol{I}_{A_k}.$$

We set $A_0 = \Omega \backslash (A_1 \cup \cdots \cup A_n)$. Then $A_0, A_1, \dots, A_n$ is a measurable partition of Ω and

$$f = 0 \cdot \boldsymbol{I}_{A_0} + \sum_{k=1}^{n} c_k \boldsymbol{I}_{A_k}.$$

Lemma 19.3.1. *The set $\mathscr{E}(\Omega, \mathcal{S})$ is a vector subspace of the space of all measurable functions $\Omega \to \mathbb{R}$.*

Proof. Suppose $f, g \in \mathscr{E}(\Omega, \mathcal{S})$,

$$f = \sum_{i=1}^{m} a_i \boldsymbol{I}_{A_i}, \quad g = \sum_{j=1}^{n} b_j \boldsymbol{I}_{B_j}$$

where the measurable sets $(A_i)_{1 \leqslant i \leqslant m}$ and $(B_j)_{1 \leqslant j \leqslant n}$ define partitions of Ω. We set

$$C_{ij} = A_i \cap B_j, \quad \forall 1 \leqslant i \leqslant m, \; 1 \leqslant j \leqslant n.$$

Then sets (C_{ij}) define a measurable partition of Ω and

$$f+g=\sum_{i,j}(a_i+b_j)\boldsymbol{I}_{C_{ij}}\in\mathscr{E}(\Omega,\mathcal{S}).$$

Clearly, for every $c\in\mathbb{R}$, the function cf is also elementary. □

Remark 19.3.2. Let us point out that $\mathscr{E}(\Omega,\mathcal{S})$ is also a subalgebra of the algebra of measurable functions since for any $A,B\in\mathcal{S}$ we have $\boldsymbol{I}_A\cdot\boldsymbol{I}_B=\boldsymbol{I}_{A\cap B}$, $A\cap B\in\mathcal{S}$. □

The construction of the integral with respect to the measure μ begins by constructing an extension of μ from $\mathcal{S}$ to $\mathscr{E}_+(\Omega,\mathcal{S})$ and preserving the additivity properties of μ. More precisely, if

$$f=\sum_{i=1}^m a_i\boldsymbol{I}_{A_i}\in\mathscr{E}_+(\Omega,\mathcal{S}),$$

then we set

$$\mu\big[\,f\,\big]=\int_\Omega f(\omega)\mu\big[\,d\omega\,\big]:=\sum_{i=1}^m a_i\mu\big[\,A_i\,\big]\in[0,\infty].$$

Lemma 19.3.3. *Let $f,g\in\mathscr{E}_+(\Omega,\mathcal{S})$ and $c\in[0,\infty)$. Then*

$$f+g,\;\; cf,\;\; \min(f,g),\;\; \max(f,g)\in\mathscr{E}_+(\Omega,\mathcal{S}),$$

and

$$\mu\big[\,f+g\,\big]=\mu\big[\,f\,\big]+\mu\big[\,g\,\big],\;\;\mu\big[\,cf\,\big]=c\mu\big[\,f\,\big].$$

Moreover, if $f\leqslant g$, then $\mu\big[\,f\,\big]\leqslant\mu\big[\,g\,\big]$.

Proof. As in the proof of Lemma 19.3.1 we write

$$f=\sum_i a_i\boldsymbol{I}_{A_i},\;\; g=\sum_j b_j\boldsymbol{I}_{B_j},$$

where $(A_i)_{1\leqslant i\leqslant m}$ and $(B_j)_{1\leqslant j\leqslant n}$ are measurable partitions of Ω. Then

$$f=\sum_{i,j}a_i\boldsymbol{I}_{A_i\cap B_j},\;\; g=\sum_{i,j}b_j\boldsymbol{I}_{A_i\cap B_j}$$

$$f+g=\sum_{i,j}(a_i+b_j)\boldsymbol{I}_{A_i\cap B_j}\in\mathscr{E}_+(\Omega,\mathcal{S}),$$

$$\min(f,g)=\sum_{i,j}\min(a_i,b_j)\boldsymbol{I}_{A_i\cap B_j}\in\mathscr{E}_+(\Omega,\mathcal{S}),$$

$$\max(f,g)=\sum_{i,j}\max(a_i,b_j)\boldsymbol{I}_{A_i\cap B_j}\in\mathscr{E}_+(\Omega,\mathcal{S}).$$

We have

$$\mu\big[\,f+g\,\big]=\sum_{i,j}(a_i+b_j)\mu\big[\,A_i\cap B_j\,\big]$$

$$
\begin{aligned}
&= \sum_i \sum_j a_i \mu\big[\, A_i \cap B_j \,\big] + \sum_j \sum_i b_j \mu\big[\, A_i \cap B_j \,\big] \\
&= \sum_i a_i \left(\sum_j \mu\big[\, A_i \cap B_j \,\big] \right) + \sum_j b_j \left(\sum_i \mu\big[\, A_i \cap B_j \,\big] \right) \\
&= \sum_i a_i \mu\Big[\, \underbrace{\bigcup_j (A_i \cap B_j)}_{A_i} \,\Big] + \sum_j b_j \mu\Big[\, \underbrace{\bigcup_i (A_i \cap B_j)}_{B_j} \,\Big] \\
&= \sum_i a_i \mu\big[\, A_i \,\big] + \sum_j b_j \mu\big[\, B_j \,\big] = \mu\big[\, f \,\big] + \mu\big[\, g \,\big].
\end{aligned}
$$

The equality $\mu\big[\, cf \,\big] = c\mu\big[\, f \,\big]$ is obvious.

If $f \leqslant g$, then $\forall i, j$, $f(\omega) \leqslant g(\omega)$, $\forall \omega \in A_i \cap B_j$, Hence

$$
a_i \mu\big[\, A_i \cap B_j \,\big] \leqslant b_j \mu\big[\, A_i \cap B_j \,\big],
$$

so that $\mu\big[\, f \,\big] \leqslant \mu\big[\, g \,\big]$. □

For any $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ we set

$$
\mathscr{E}^f_+ := \big\{\, g \in \mathscr{E}_+(\Omega, \mathcal{S});\ \ g(\omega) \leqslant f(\omega),\ \ \forall \omega \in \Omega \,\big\}.
$$

Note that $\mathscr{E}^f_+ \neq \varnothing$ since $0 \in \mathscr{E}^f_+$. Observe that if $f \in \mathscr{E}_+(\Omega, \mathcal{S})$, then $f \in \mathscr{E}^f_+$ and Lemma 19.3.3 implies $\mu\big[\, f \,\big] \geqslant \mu\big[\, g \,\big]$, $\forall g \in \mathscr{E}^f_+$. Hence

$$
\mu\big[\, f \,\big] = \sup_{g \in \mathscr{E}^f_+} \mu\big[\, g \,\big],\ \ \forall f \in \mathscr{E}_+(\Omega, \mathcal{S}).
$$

Motivated by this fact, for any $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$ we set

$$
\boxed{\mu\big[\, f \,\big] := \sup_{g \in \mathscr{E}^f_+} \mu\big[\, g \,\big] \in [0, \infty].}
$$

We say that $\mu\big[\, f\big]$ is the *Lebesgue integral of the nonnegative measurable function f with respect to the measure μ* and we will use the alternate notation

$$
\boxed{\mu\big[\, f \,\big] = \int_\Omega f d\mu = \int_\Omega f(\omega) \mu\big[\, d\omega \,\big].}
$$

Definition 19.3.4. A measurable function $f \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ is called *μ-integrable* if $\mu\big[\, f_+ \,\big],\ \ \mu\big[\, f_- \,\big] < \infty$. In this case we define its *Lebesgue integral* to be

$$
\int_\Omega f d\mu = \int_\Omega f(\omega) \mu\big[\, d\omega \,\big] = \mu\big[\, f \,\big] := \mu\big[\, f_+ \,\big] - \mu\big[\, f_- \,\big].
$$

We denote by $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ the set of μ-integrable functions and by $\mathcal{L}^1_+(\Omega, \mathcal{S}, \mu)$ the set of μ-integrable nonnegative functions. □

Proposition 19.3.5. *If $f,g \in \bar{\mathscr{L}}^0_+(\Omega,\mathcal{S})$ and $f \leqslant g$, then for any $\mu \in \mathrm{Meas}(\Omega,\mathcal{S})$ we have*

$$\mu\big[\,f\,\big] \leqslant \mu\big[\,g\,\big].$$

Proof. Indeed we have $\mathscr{E}^f_+ \subset \mathscr{E}^g_+$ so

$$\mu\big[\,f\,\big] = \sup_{h\in\mathscr{E}^f_+} \mu\big[\,h\,\big] \leqslant \sup_{h\in\mathscr{E}^g_+} \mu\big[\,h\,\big] = \mu\big[\,g\,\big].$$

□

Corollary 19.3.6. *Suppose that $(\Omega,\mathcal{S},\mu)$ is a finite measured space, i.e., $\mu\big[\,\Omega\,\big] < \infty$. If $f \in \mathscr{L}^0(\Omega,\mathcal{S})$ is bounded, then it is μ-integrable.*

Proof. Let $C > 0$ such that $|f(\omega)| < C$, $\forall \omega \in \Omega$. Then, $f_\pm < C$ and for any $g \in \mathscr{E}^{f_\pm}_+$ we have $g \leqslant C\boldsymbol{I}_\Omega$. Hence

$$\mu\big[\,f_\pm\,\big] < C\mu\big[\,\Omega\,\big] < \infty$$

□

Corollary 19.3.7. *Let $a,b \in \mathbb{R}$, $a < b$, and $f \in \mathscr{L}^0\big([a,b],\mathcal{S}_{\boldsymbol{\lambda}}\big)$, where $\mathcal{S}_{\boldsymbol{\lambda}}$ denotes the sigma-algebra of Lebesgue measurable subsets. If f is bounded, then f is $\boldsymbol{\lambda}$-integrable. In particular, any continuous function of $[a,b]$ is $\boldsymbol{\lambda}$-integrable.* □

The Lebesgue integral enjoys many of the desirable properties of the Riemann integral such as the linearity of the correspondence $f \mapsto \mu\big[\,f\,\big]$. The next result is key to unlocking these nice feature out of the rather opaque Definition 19.3.4.

Theorem 19.3.8 (Monotone Convergence)**.** *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence in $\bar{\mathscr{L}}^0_+(\Omega,\mathcal{S})$. Set*

$$f(\omega) = \lim_{n\to\infty} f_n(\omega), \quad \forall \omega \in \Omega.$$

Then $f \in \mathscr{L}^0_+(\Omega,\mathcal{S})$ and

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega f d\mu.$$

Proof. From Proposition 19.3.5 we deduce that the sequence $\mu\big[\,f_n\,\big]$ is nondecreasing and is bounded above by $\mu\big[\,f\,\big]$. Hence it has a, possibly infinite, limit and

$$\lim_{n\to\infty} \mu\big[\,f_n\,\big] \leqslant \mu\big[\,f\,\big].$$

To complete the proof of the theorem we have to show that

$$\lim_{n\to\infty} \mu\big[\,f_n\,\big] \geqslant \mu\big[\,f\,\big],$$

i.e.,

$$\lim_{n\to\infty} \mu\big[\,f_n\,\big] \geqslant \mu\big[\,g\,\big], \quad \forall g \in \mathscr{E}^f_+. \tag{19.3.2}$$

We will achieve this using a clever trick. Fix $g \in \mathscr{E}_+^f$, $c \in (0,1)$, and set

$$S_n := \big\{ \omega \in \Omega;\;\; f_n(\omega) \geqslant cg(\omega) \,\big\}.$$

Since $f = \lim f_n$ and (f_n) is a nondecreasing sequence of functions we deduce that S_n is a nondecreasing sequence of measurable sets whose union is Ω. For any elementary function h the product $\boldsymbol{I}_{S_n} h$ is also elementary. For any $n \in \mathbb{N}$ we have

$$f_n \geqslant f_n \boldsymbol{I}_{S_n} \geqslant cg\boldsymbol{I}_{S_n}$$

so that

$$\mu\big[\, f_n \,\big] \geqslant \mu\big[\, \boldsymbol{I}_{S_n} f_n \,\big] \geqslant c\mu\big[\, g\boldsymbol{I}_{S_n} \,\big].$$

If we write g as a finite linear combination

$$g = \sum_j g_j \boldsymbol{I}_{A_j}$$

with A_j pairwise disjoint, then we deduce

$$\mu\big[\, f_n \,\big] \geqslant c\mu\big[\, g\boldsymbol{I}_{S_n} \,\big] = c\sum_j g_j \mu\big[\, A_j \cap S_n \,\big].$$

The sequence of sets $(A_j \cap S_n)_{n\in\mathbb{N}}$ is nondecreasing and its union is A_j so that

$$\lim_{n\to\infty} \mu\big[\, f_n \,\big] \geqslant c\sum_j g_j \lim_{n\to\infty} \mu\big[\, A_j \cap S_n \,\big] = c\sum_j g_j \mu\big[\, A_j \,\big] = c\mu\big[\, g \,\big].$$

Hence

$$\lim_{n\to\infty} \mu\big[\, f_n \,\big] \geqslant c\mu\big[\, g \,\big],\;\; \forall g \in \mathscr{E}_+^f,\;\; \forall c \in (0,1).$$

Letting $c \nearrow 1$ we deduce (19.3.2). □

Recall the function $D_n : [0,\infty] \to [0,\infty]$ defined in (19.1.5)

$$D_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \boldsymbol{I}_{[(k-1)2^{-n},k2^{-n})}(r) + n\boldsymbol{I}_{[n,\infty)}(r) = \min\Big(\frac{\lfloor 2^n r \rfloor}{2^n}, n \Big). \tag{19.3.3}$$

and the resulting sequence of transformations

$$D_n : \bar{\mathcal{L}}_+^0(\Omega,\mathcal{S}) \to \mathscr{E}_+(\Omega,\mathcal{S}),\;\; f \mapsto D_n[f],\;\; D_n[f] = D_n \circ f.$$

More precisely,

$$D_n[f](\omega) = D_n\big(\, f(\omega)\,\big),\;\; \forall \omega \in \Omega.$$

Corollary 19.3.9. *For any $f \in \bar{\mathcal{L}}_+^0(\Omega,\mathcal{S})$ we have*

$$\mu\big[\, f \,\big] = \lim_{n\to\infty} \mu\big[\, D_n[f] \,\big].$$

Proof. The sequence $D_n[f]$ is non-decreasing and converges to f. The desired conclusion follows from the Monotone Convergence theorem. □

Remark 19.3.10. Note that for any $f \in \mathcal{L}^0_+(\Omega, \mathcal{S})$ we have

$$\mu\big[\, D_n(f)\,\big] = \sum_{k=1}^{n2^n} \frac{k-1}{2^n}\mu\Big[\,\Big\{\frac{k-1}{2^n} \leqslant f < \frac{k}{2^n}\Big\}\,\Big] + n\mu\big[\,\{f \geqslant n\}\,\big].$$

The equality

$$\mu\big[\, f\,\big] = \lim_{n\to\infty} \mu\big[\, D_n[f]\,\big]$$

justifies the similarity with the Lebesgue procedure of counting the value of a pile of coins. The set Ω is the pile of coins. The value of a coin ω is $f(\omega)$. The "number" of coins with values in the interval $\big[\,(k-1)2^{-n}, k2^{-n}\,\big)$ is

$$\mu\Big[\,\Big\{\frac{k-1}{2^n} \leqslant f < \frac{k}{2^n}\Big\}\,\Big].$$

The value of this pile is approximated from below by

$$\underbrace{\frac{k-1}{2^n}}_{\text{“value” of the coin}} \qquad\times\qquad \underbrace{\mu\Big[\,\Big\{\frac{k-1}{2^n} \leqslant f < \frac{k}{2^n}\Big\}\,\Big]}_{\text{the “number” of coins of a given “value”}}\qquad. \qquad\qquad \square$$

Corollary 19.3.11. *For any $f, g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ and $a, b \in \mathbb{R}$ such that $af + bg$ is well defined we have*

$$af + bg \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$$

and

$$\int_\Omega (af + bg)d\mu = a\int_\Omega f d\mu + b\int_\Omega g d\mu. \tag{19.3.4}$$

Moreover, if $f, g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ and $f(\omega) \leqslant g(\omega)$, $\forall \omega \in \Omega$ then

$$\int_\Omega f d\mu \leqslant \int_\Omega g d\mu.$$

Proof. We will cary the proof in several stages.

A. Assume first that $f, g \in \mathcal{L}^1_+(\Omega, \mathcal{S})$ and $a, b \in [0, \infty)$.

The sequence $aD_n[f] + bD_n[g]$ is nondecreasing and converges everywhere to $af + bg$ and the Monotone Convergence Theorem implies

$$\lim_{n\to\infty} \int_\Omega \big(\, aD_n[f] + bD_n[g]\,\big)d\mu = \int_\Omega (af + bg)d\mu.$$

On the other hand, Lemma 19.3.3 implies

$$\int_\Omega \big(\, aD_n[f] + bD_n[g]\,\big)d\mu = a\int_\Omega D_n[f]d\mu + b\int_\Omega D_n[g]d\mu.$$

Letting $n \to \infty$ we deduce that

$$\mu\big[\, af + bg\,\big] = a\mu\big[\, f\,\big] + b\mu\big[\, g\,\big],$$

so $af + bg$ is integrable and (19.3.4) holds.

B. Consider now the general case $f, g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$. Note that for any $h \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ we have

$$|h| = h_+ + h_-$$

and we deduce from **A** that $\mu\big[\,|h|\,\big] = \mu\big[\,h_+\,\big] + \mu\big[\,h_-\,\big]$. Hence h is integrable if and only if $|h|$ is integrable. Observe that

$$|f+g| \leqslant |f| + |g|$$

Proposition 19.3.5 implies

$$\mu\big[\,|f+g|\,\big] \leqslant \mu\big[\,|f|+|g|\,\big]$$

and we deduce from **A** that

$$\mu\big[\,|f|+|g|\,\big] = \mu\big[\,|f|\,\big] + \mu\big[\,|g|\,\big] < \infty.$$

Hence $\mu\big[\,|f+g|\,\big] < \infty$ so $f+g$ is integrable. From the equalities

$$(-f)_+ = f_-,\ \ (-f)_- = f_+$$

we deduce

$$\mu\big[\,-f\,\big] = \mu\big[\,f_-\,\big] - \mu\big[\,f_+\,\big] = -\mu\big[\,f\,\big].$$

Observe that

$$(f+g)_+ - (f+g)_- = f+g = f_+ + f_- - (f_- + g_-)$$

so that

$$(f+g)_+ + f_- + g_- = f_+ + g_+ + (f+g)_-.$$

We deduce from part **A** that

$$\begin{aligned} &\mu\big[\,(f+g)_+\,\big] + \mu\big[\,f_-\,\big] + \mu\big[\,g_-\,\big] = \mu\big[\,f_+\,\big] + \mu\big[\,g_+\,\big] + \mu\big[\,(f+g)_-\,\big] \\ \Rightarrow\ &\mu\big[\,(f+g)_+\,\big] - \mu\big[\,(f+g)_-\,\big] = \mu\big[\,f_+\,\big] + \mu\big[\,g_+\,\big] - \big(\mu\big[\,f_-\,\big] + \mu\big[\,g_-\,\big]\big), \end{aligned}$$

i.e.,

$$\mu\big[\,f+g\,\big] = \mu\big[\,f\,\big] + \mu\big[\,g\,\big].$$

Clearly for $a \geqslant 0$

$$(af)_\pm = a(f_\pm)$$

so that

$$\mu\big[\,af\,\big] = a\mu\big[\,f\,\big],\ \ \mu\big[\,-af\,\big] = -\mu\big[\,af\,\big] = -a\mu\big[\,f\,\big].$$

Finally if $f \leqslant g$ then $0 \leqslant g - f$ so

$$0 \leqslant \int_\Omega (g-f)d\mu = \int_\Omega g d\mu - \int_\Omega f d\mu \Rightarrow \int_\Omega f d\mu \leqslant \int_\Omega g d\mu.$$

□

Let us record a useful observation we used in the above proof.

Corollary 19.3.12. *Let $f \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$. Then*

$$f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu) \Longleftrightarrow |f| \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu).$$

□

Corollary 19.3.13 (Markov's Inequality)**.** *Suppose that $f \in \mathcal{L}^1_+(\Omega, \mathcal{S}, \mu)$. Then, for any $C > 0$, we have*

$$\mu\big[\,\{f \geqslant C\}\,\big] \leqslant \frac{1}{C}\int_\Omega f d\mu. \tag{19.3.5}$$

In particular, $f < \infty$, μ-a.e.., i.e. $\mu\big[\,\{f = \infty\}\,\big] = 0$.

Proof. Note that

$$C\boldsymbol{I}_{\{f \geqslant C\}} \leqslant f \Rightarrow \; C\mu\big[\,\{f \geqslant C\}\,\big] = \int_\Omega C\boldsymbol{I}_{\{f \geqslant C\}} \leqslant \int_\Omega f d\mu.$$

Observe that

$$\{f \geqslant 1\} \supset \{f \geqslant 2\} \supset \cdots \;\; \text{and} \;\; \{f = \infty\} = \bigcap_{n \geqslant 1}\{f \geqslant n\}.$$

Since $\mu\big[\,\{f \geqslant 1\}\,\big] < \infty$ we deduce from Exercise 19.18 that

$$\mu\big[\,\{f = \infty\}\,\big] = \lim_{n\to\infty}\mu\big[\,\{f \geqslant n\}\,\big] \leqslant \lim_{n\to\infty}\frac{1}{n}\int_\Omega d\mu = 0.$$

□

Corollary 19.3.14 (Inclusion-Exclusion Principle)**.** *Suppose that $(\Omega, \mathcal{S}, \mu)$ is a* finite *measured space, i.e., $\mu\big[\,\Omega\,\big] < \infty$. Then, for any measurable sets $S_1, \dots, S_n \in \mathcal{S}$ we have*

$$\begin{aligned}\mu\Big[\bigcup_{k=1}^n S_k\Big] = \sum_{k=1}^n \mu\big[\,S_k\,\big] - \sum_{1 \leqslant i < j \leqslant n}\mu\big[\,S_i \cap S_j\,\big] + \cdots \\ +(-1)^{\ell-1}\sum_{1 \leqslant i_1 < \cdots < i_\ell \leqslant n}\mu\big[\,S_{i_1} \cap \cdots \cap S_{i_\ell}\,\big] + \cdots\end{aligned} \tag{19.3.6}$$

Proof. Note that

$$\boldsymbol{I}_{S_1^c \cap \cdots \cap S_n^c} = \boldsymbol{I}_{S_1^c} \cdots \boldsymbol{I}_{S_n^c} = \prod_{k=1}^n\big(1 - \boldsymbol{I}_{S_k}\big) = 1 + \sum_{\ell=1}^n(-1)^\ell \sum_{1 \leqslant i_1 < \cdots < i_\ell \leqslant n}\boldsymbol{I}_{S_{i_1} \cap \cdots \cap S_{i_\ell}}.$$

We deduce that

$$\boldsymbol{I}_{S_1 \cup \cdots \cup S_n} = 1 - \boldsymbol{I}_{S_1^c \cap \cdots \cap S_n^c} = \sum_{\ell=1}^n(-1)^{\ell-1}\sum_{1 \leqslant i_1 < \cdots < i_\ell \leqslant n}\boldsymbol{I}_{S_{i_1} \cap \cdots \cap S_{i_\ell}}$$

Integrating with respect to μ all sides of the above equality we obtain (19.3.6). □

Corollary 19.3.15. *Suppose that $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$. Then*

$$\left|\int_\Omega f d\mu\right| \leqslant \int_\Omega |f| d\mu. \tag{19.3.7}$$

Proof. We have $-|f| \leqslant f \leqslant |f|$ so that

$$-\int_\Omega |f| d\mu \leqslant \int_\Omega f d\mu \leqslant \int_\Omega |f| d\mu.$$

The above inequalities are equivalent to (19.3.7). □

Corollary 19.3.16. *Let $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$. Then the following are equivalent.*

(i) $f = 0$ μ-a. e..

(ii) $\mu\big[\, f\,\big] = 0$.

Proof. (i)→ (ii) Let $g \in \mathscr{E}^f_+$. Then $g = 0$ a. e. so that $\mu\big[\, g\,\big] = 0$. Hence

$$\mu\big[\, f\,\big] = \sup_{g \in \mathscr{E}^f_+} \mu\big[\, g\,\big] = 0.$$

(ii) ⇒ (i) The Markov inequality shows that

$$\mu\big[\,\{f > 1/n\}\,\big] \leqslant n\mu\big[\, f\,\big] = 0$$

so

$$\mu\big[\, f > 0\,\big] = \lim_{n\to\infty} \mu\big[\,\{f > 1/n\}\,\big] = 0.$$

□

Proposition 19.3.17. *Suppose $f, g \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S})$ and $f = g$, μ-a.e.. Then*

$$f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu) \Longleftrightarrow g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu).$$

Moreover, if one of the above equivalent conditions hold, then $\mu\big[\, f\,\big] = \mu\big[\, g\,\big]$.

Proof. It suffices to prove only the implication " ⇒". The other implication is obtained by reversing the roles of f and g. Set

$$Z := \big\{f \neq g\big\} = \big\{\, \omega \in \Omega;\;\; f(\omega) \neq g(\omega)\,\big\}.$$

The set Z is μ-negligible. From Corollary 19.3.16 we deduce

$$\mu\big[\, \boldsymbol{I}_Z|f|\,\big] = \mu\big[\, \boldsymbol{I}_Z|g|\,\big] = 0.$$

In particular $\boldsymbol{I}_Z g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$. Next observe that

$$\big(\, 1 - \boldsymbol{I}_Z\,\big) = \boldsymbol{I}_{\Omega\setminus Z} \;\text{ and }\; f(\omega) = g(\omega),\;\; \forall \omega \in \Omega\setminus Z.$$

Since $f \in \mathcal{L}^1$ and

$$0 \leqslant \boldsymbol{I}_{\Omega\setminus Z}|f| \leqslant |f|,$$

we deduce

$$\boldsymbol{I}_{\Omega\setminus Z}|f| \in \mathcal{L}^1.$$

Thus,

$$\big(\, 1 - \boldsymbol{I}_Z\,\big)g = \big(\, 1 - \boldsymbol{I}_Z\,\big)f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu).$$

We conclude that

$$g = \big(\, 1 - \boldsymbol{I}_Z\,\big)g + \boldsymbol{I}_Z g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu).$$

Suppose now that f, g are integrable and $f = g$, a.e.. From (19.3.7) we deduce

$$0 \leqslant \Big|\mu\big[\,f\,\big] - \mu\big[\,g\,\big]\Big| = \Big|\mu\big[\,f-g\,\big]\Big| \leqslant \mu\big[\,|f-g|\,\big].$$

The function $|f-g|$ is nonnegative and zero almost everywhere so $\mu\big[\,|f-g|\,\big] = 0$ by Corollary 19.3.16. □

Remark 19.3.18. The presentation so far had to tread carefully around a nagging problem: given $f, g \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$, then $f(\omega) + g(\omega)$ may not be well defined for some ω. For example, it could happen that $f(\omega) = \infty$, $g(\omega) = -\infty$. Fortunately, Corollary 19.3.13 shows that the set of such ω's is negligible. Moreover, if we redefine f and g to be equal to zero on the set where they had infinite values, then their integrals do not change. For this reason we alter the definition of $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ as follows.

$$\mathcal{L}^1(\Omega, \mathcal{S}, \mu) := \left\{ f : (\Omega, \mathcal{S}) \to \mathbb{R};\;\; f \;\text{ measurable},\;\; \int_\Omega |f| d\mu < \infty \right\}.$$

Thus, in the sequel the integrable functions will be assumed to be <u>everywhere</u> finite.

With this convention the space $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ is a vector space and the Lebesgue integral is a linear functional

$$\mu : \mathcal{L}^1(\Omega, \mathcal{S}, \mu) \to \mathbb{R}, \;\; f \mapsto \mu\big[\,f\,\big].$$ □

For any $S \in \mathcal{S}$, and any $f \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$, we set

$$\int_S f d\mu := \mu\big[\,f\boldsymbol{I}_S\,\big] = \int_\Omega f\boldsymbol{I}_S d\mu. \tag{19.3.8}$$

Since $f\boldsymbol{I}_S \leqslant f$ we deduce that

$$\int_S f d\mu \leqslant \int_\Omega f d\mu.$$

Observe that if $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$, then

$$0 \leqslant \int_S f_\pm d\mu \leqslant \int_\Omega f_\pm d\mu < \infty.$$

We set

$$\int_S f d\mu := \int_S f_+ d\mu - \int_S f_- d\mu \in \mathbb{R}.$$

Corollary 19.3.19. *For any $f \in \bar{\mathcal{L}}^0_+(\Omega)$ the function*

$$\mu_f : \mathcal{S} \to [0, \infty], \;\; \mu_f\big[\,S\,\big] := \int_S f d\mu = \mu\big[\,f\boldsymbol{I}_S\,\big] \tag{19.3.9}$$

is a measure.

Proof. Clearly $\mu_f\big[\,\varnothing\,\big] = 0$ since $f\boldsymbol{I}_\varnothing = 0$. Observe next that μ_f is finitely additive. Indeed, if $S_1, S_2 \in \mathcal{S}$ and $S_1 \cap S_2 = \varnothing$, then

$$\boldsymbol{I}_{S_1 \cup S_2} = \boldsymbol{I}_{S_1} + \boldsymbol{I}_{S_2}$$

and

$$\mu_f\big[\,S_1\cup S_2\,\big]=\mu\big[\,f\boldsymbol{I}_{S_1}+f\boldsymbol{I}_{S_2}\,\big]=\mu\big[\,f\boldsymbol{I}_{S_1}\,\big]+\mu\big[\,f\boldsymbol{I}_{S_2}\,\big]=\mu_f\big[\,S_2\,\big]+\mu_f\big[\,S_2\,\big].$$

Finally, let us check the increasing continuity. Let

$$S_1\subset S_2\subset\cdots,\ \ S_\infty=\bigcup_{n\geqslant 1}S_n$$

be a non-decreasing sequence of measurable sets. Then $(f\boldsymbol{I}_{S_n})$ is a nondecreasing sequence of nonnegative measurable functions and

$$\lim_{n\to\infty}f(\omega)\boldsymbol{I}_{S_n}(\omega)=f(\omega)\boldsymbol{I}_{S_\infty}(\omega),\ \ \forall\omega\in\Omega.$$

Hence

$$\lim_{n\to\infty}\mu_f\big[\,S_n\,\big]=\lim_{n\to\infty}\mu\big[\,f\boldsymbol{I}_{S_n}\,\big]=\mu\big[\,f\boldsymbol{I}_{S_\infty}\,\big]=\mu_f\big[\,S_\infty\,\big],$$

where the second equality follows from the Monotone Convergence Theorem.

□

For every sequence $(x_n)_{n\in\mathbb{N}}$ in $[-\infty,\infty]$ we set

$$x_k^*:=\inf_{n\geqslant k}x_n.$$

The sequence $(x_n^*)_{n\in\mathbb{N}}$ is nondecreasing and thus it has a limit. We set

$$\liminf_{n\to\infty}x_n:=\lim_{k\to\infty}x_k^*=\lim_{k\to\infty}\big(\inf_{n\geqslant k}x_n\big).$$

Equivalently, $\liminf_{n\to\infty}x_n$ is the smallest cluster point of the sequence $(x_n)_{n\in\mathbb{N}}$. We define similarly

$$\limsup_{n\to\infty}x_n:=\lim_{k\to\infty}\big(\sup_{n\geqslant k}x_n\big).$$

Equivalently, $\limsup_{n\to\infty}x_n$ is the largest cluster point of the sequence $(x_n)_{n\in\mathbb{N}}$. Clearly

$$\liminf_{n\to\infty}x_n\leqslant\limsup_{n\to\infty}x_n$$

with equality if and only if the sequence (x_n) has a limit. In this case

$$\lim_n x_n=\liminf_{n\to\infty}x_n=\limsup_{n\to\infty}x_n.$$

Note that

$$\liminf_{n\to\infty}(-x_n)=-\limsup_{n\to\infty}x_n,\ \ \limsup_{n\to\infty}(-x_n)=-\liminf_{n\to\infty}x_n.$$

Theorem 19.3.20 (Fatou's Lemma)**.** *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\bar{\mathcal{L}}_+^0(\Omega,\mathcal{S})$. Then*

$$\int_\Omega\liminf_{n\to\infty}f_n(\omega)\,\mu\big[\,d\omega\,\big]\leqslant\liminf_{n\to\infty}\int_\Omega f_n d\mu.$$

Proof. Set

$$g_k := \inf_{n \geqslant k} f_n.$$

Proposition 19.1.19(iii) implies that $g_k \in \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S})$. The sequence (g_k) is nondecreasing and

$$\liminf_{n\to\infty} f_n = \lim_{k\to\infty} g_k.$$

The Monotone Convergence Theorem implies that

$$\int_\Omega \liminf_{n\to\infty} f_n(\omega)\, \mu\big[\, d\omega\,\big] = \lim_{k\to\infty} \int_\Omega g_k d\mu.$$

Note that

$$g_k \leqslant f_n, \quad \forall n \geqslant k.$$

Hence

$$\int_\Omega g_k d\mu \leqslant \int_\Omega f_n d\mu, \quad \forall n \geqslant k,$$

i.e.,

$$\int_\Omega g_k d\mu \leqslant \inf_{n\geqslant k} \int_\Omega f_n d\mu.$$

Letting $k \to \infty$ we deduce

$$\lim_{k\to\infty} \int_\Omega g_k d\mu \leqslant \lim_{k\to\infty} \inf_{n\geqslant k} \int_\Omega f_n d\mu = \liminf_{n\to\infty} \int_\Omega f_n d\mu.$$

□

Theorem 19.3.21 (Dominated Convergence). *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ satisfying the following conditions.*

(i) *The sequence (f_n) converges everywhere to a function $f : \Omega \to \mathbb{R}$, i.e.,*

$$\lim_{n\to\infty} f_n(\omega) = f(\omega), \quad \forall \omega \in \Omega.$$

(ii) *The sequence (f_n) is dominated by an integrable function $h \in \mathcal{L}^1_+(\Omega, \mathcal{S}, \mu)$, i.e.,*

$$|f_n(\omega)| \leqslant h(\omega), \quad \forall n \in \mathbb{N}, \quad \omega \in \Omega.$$

Then $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ and

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f d\mu.$$

Proof. Note that $-h \leqslant f_n \leqslant h$, $\forall n$. Consider the sequence of *nonnegative* integrable functions $g_n = |f_n| + h$. Note that $|g_n| \leqslant 2h$ and $g_n \to |f| + h$. We deduce from Fatou's Lemma that

$$\int_\Omega (|f| + h) d\mu \leqslant \liminf_{n\to\infty} \int (|f_n| + h) d\mu \leqslant 2 \int_\Omega h d\mu < \infty$$

proving that f is integrable.

Consider now the sequences of *nonnegative* integrable functions $u_n = f_n + h$ and $v_n = h - f_n$. We deduce from Fatou's Lemma that

$$\int_\Omega (f+h)d\mu = \int_\Omega \lim_n u_n d\mu \leqslant \liminf_n \int_\Omega (f_n + h)d\mu = \liminf_n \int_\Omega f_n d\mu + \int_\Omega h d\mu.$$

Hence

$$\int_\Omega f d\mu \leqslant \liminf_n \int_\Omega f_n d\mu.$$

Invoking Fatou's lemma one more time we deduce that

$$\int_\Omega (h-f)d\mu = \int_\Omega \lim_n v_n d\mu \leqslant \liminf_n \int_\Omega (h - f_n)d\mu = \int_\Omega h d\mu + \liminf_n \int_\Omega (-f_n)d\mu.$$

Hence

$$-\int_\Omega f d\mu \leqslant \liminf_n \int_\Omega (-f_n)d\mu = -\limsup_n \int_\Omega f_n d\mu$$

so that

$$\limsup_n \int_\Omega f_n d\mu \leqslant \int_\Omega f d\mu \leqslant \liminf_n \int_\Omega f_n d\mu.$$

We conclude

$$\lim_n \int_\Omega f_n d\mu = \int_\Omega f d\mu.$$

□

Corollary 19.3.22. *Suppose that the sequence (f_n) in $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ satisfies the conditions in the Dominated Convergence Theorem. Then*

$$\lim_{n\to\infty} \int_\Omega |f_n - f| d\mu = 0.$$

Proof. Set $g_n := |f_n - f|$. Note that $g_n \to 0$ and

$$|g_n| = |f_n - f| \leqslant |f_n| + |f| \leqslant h + |f| \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu).$$

Applying the Dominated Convergence Theorem to the sequence (g_n) we deduce

$$\lim_n \int_\Omega |f_n - f| d\mu = \lim_n \int_\Omega g_n d\mu = 0.$$

□

The Dominated Convergence Theorem has an a.e. version.

Theorem 19.3.23 (Dominated Convergence: a.e. version)**.** *Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ satisfying the following conditions.*

(i) *The sequence (f_n) converges a.e. to a* *<u>measurable</u>* *function $f \in \bar{\mathcal{L}}^0(\Omega, \mathcal{S}, \mu)$.*

$$\lim_{n\to\infty} f_n(\omega) = f(\omega), \quad \forall \omega \in \Omega.$$

(ii) *The sequence* (f_n) *is a.e. dominated by an integrable function* $h \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$, $\forall n$, $\exists Z_n \in \mathcal{S}$ *such that* $\mu\big[\, Z_n \,\big] = 0$ *and* $|f_n(\omega)| \leqslant h(\omega)$, $\forall \omega \in \Omega \backslash Z_n$.

Then $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ *and*

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f d\mu.$$

Proof. There exists a negligible set $Z_\infty \in \mathcal{S}$ such that $f_n(\omega) \to f(\omega)$, $\forall \omega \in \Omega \backslash Z_\infty$. The set

$$Z := Z_\infty \cup \left(\bigcup_{n\in\mathbb{N}} Z_n \right)$$

is negligible. Define $f^* = f \boldsymbol{I}_{\Omega\backslash Z}$, $h^* = h \boldsymbol{I}_{\Omega\backslash Z}$ and $f_n^* = f_n \boldsymbol{I}_{\Omega\backslash Z}$. These functions satisfy the assumptions in Theorem 19.3.21. Since $f = f^*$ and $f_n = f_n^*$ a.e. we deduce

$$\int_\Omega f \, d\mu = \int_\Omega f^* \, d\mu = \lim_{n\to\infty} \int_\Omega f_n^* \, d\mu = \lim_{n\to\infty} \int_\Omega f_n \, d\mu.$$

□

Remark 19.3.24. If $f_n \in \mathcal{L}^0(\Omega, \mathcal{S}, \mu)$ is a sequence of measurable functions that converges μ-a.e. to a a function $f : \Omega \to \mathbb{R}$, then the function f need not be measurable if $\mathcal{S}$ is not μ-complete.

Indeed, suppose $(\omega, \mathcal{S}, \mu) = ([0,1], \mathcal{B}_{[0,1]}, \boldsymbol{\lambda})$, and S is a subset of the Cantor set C that it is not Borel. Then the constant sequence $\boldsymbol{I}_C$ converges $\boldsymbol{\lambda}$-a.e. to $\boldsymbol{I}_S$ but, by construction, $\boldsymbol{I}_S$ is not Borel measurable.

However, if $\mathcal{S}$ is μ-complete and $f_n \to f$ μ-a.e., then f is $\mathcal{S}$-measurable. □

Lemma 19.3.25. *Let* $(\Omega, \mathcal{S}, \mu)$ *be a measured space. Denote by* $\mathcal{S}^\mu$ *the* μ*-completion of* $\mathcal{S}$ *so*

$$\bar{\mathcal{L}}^0\big(\Omega, \mathcal{S}, \mu\big) \subset \bar{\mathcal{L}}^0\big(\Omega, \mathcal{S}^\mu, \mu\big).$$

For any function $f \in \bar{\mathcal{L}}^0\big(\Omega, \mathcal{S}^\mu, \mu\big)$ *there exists a function* $f' \in \bar{\mathcal{L}}^0\big(\Omega, \mathcal{S}, \mu\big)$ *such that* $f = f'$ μ*-a.e.*

Proof. It suffices to consider ony the case $f \geqslant 0$. Let $f_n \in \mathscr{E}_+\big(\Omega, \mathcal{S}^\mu\big)$ a non-0decreasing sequence of $\bar{\mathcal{S}}^\mu$-measurable functions that converge everywhere to f. For any n there exists a $\mathcal{S}$-measurable function $\bar{f_n}$ and μ-ngligible subset $Z_n \in \mathcal{S}$ such that $\bar{f_n} = f_n$ on $\Omega \backslash Z_n$. Set

$$Z_\infty = \bigcup_{n\in\mathbb{N}} Z_n \in \mathcal{S}.$$

Then Z_∞ is μ-negligible. Set

$$f_n' = \boldsymbol{I}_{\Omega\backslash Z_\infty}.$$

Then (f_n') is a nondecreasing sequence of $\mathcal{S}$- measurable functions. Its limit function is μ measurable and agrees with f off Z_∞. □

Remark 19.3.26. We have a Lebesgue integral

$$\bar{\int}_\Omega : \bar{\mathcal{L}}_0^+\big(\Omega, \mathcal{S}^\mu, \mu\big) \to [0, \infty].$$

Its restriction to $\bar{\mathscr{L}}_0^+\big(\Omega, \mathcal{S}, \mu\big) \subset \bar{\mathscr{L}}_0^+\big(\Omega, \mathcal{S}^\mu, \mu\big)$ coincides with the integral

$$\int_\Omega : \bar{\mathscr{L}}_0^+\big(\Omega, \mathcal{S}, \mu\big) \to [0, \infty].$$

Moreover If $f \in \bar{\mathscr{L}}_0^+\big(\Omega, \mathcal{S}^\mu, \mu\big)$, and $f' \in \bar{\mathscr{L}}_0^+\big(\Omega, \mathcal{S}, \mu\big)$ is such that $f = f'$ μ-a.e., then

$$\overline{\int}_\Omega f d\mu = \int_\Omega f' d\mu.$$

□

Suppose that $(\Omega_0, \mathcal{S}_0)$ and $(\Omega_1, \mathcal{S}_1)$ are two measurable spaces and $\Phi : \Omega_0 \to \Omega_1$ an $(\mathcal{S}_0, \mathcal{S}_1)$-measurable map. To any measure $\mu : \mathcal{S}_0 \to [0, \infty)$ we can associate its *pushforward* via Φ. This is the measure

$$\Phi_\# \mu : \mathcal{S}_1 \to [0, \infty), \quad \Phi_\# \mu\big[\, S_1 \,\big] = \mu\big[\, \Phi^{-1}(S_1) \,\big], \quad \forall S_1 \in \mathcal{S}_1 \tag{19.3.10}$$

The map Φ also induces a *pullback* map

$$\Phi^* : \mathscr{L}^0(\Omega_1, \mathcal{S}_1) \to \mathscr{L}^0(\Omega_0, \mathcal{S}_0), \quad \Phi^* f = f \circ \Phi.$$

The next result relates the integral with respect to $F_\# \mu$ to the integral with respect to the measure μ. It can be viewed as a very general form of the change in variables formula. In probability theory it is known under the acronym LOTUS: The Law Of The Unconscious Statistician.

Theorem 19.3.27 (Change in variables)**.** *Let $\Phi : (\Omega_0, \mathcal{S}_0) \to (\Omega_1, \mathcal{S}_1)$ be a measurable map and $\mu \in \mathrm{Meas}(\Omega_0, \mathcal{S}_0)$.*

(i) *For any $f \in \mathscr{L}_+^0(\Omega_1, \mathcal{S}_1)$ we have*

$$\Phi_\# \mu\big[\, f \,\big] = \mu\big[\, \Phi^* f \,\big]. \tag{19.3.11}$$

(ii) *If $f \in \mathscr{L}^1(\Omega_1, \mathcal{S}_1, \Phi_\# \mu)$, then $\Phi^* f \in \mathscr{L}^1(\Omega_0, \mathcal{S}_0, \mu)$ and (19.3.11) holds.*

Proof. (i) Let us observe in the special case $f = \boldsymbol{I}_{S_0}$ the equality (19.3.11) is trivially true since it becomes the definition (19.3.10) of the pushforward. By linearity we see that (19.3.11) holds for elementary functions.

Suppose that $f \in \mathscr{L}_+^0(\Omega_0, \mathcal{S}_0)$. Note that for any $n \in \mathbb{N}$ we have

$$\Phi^* D_n[f] = D_n[f] \circ \Phi = (D_n \circ f) \circ \Phi = D_n \circ (f \circ \Phi) = D_n[\Phi^* f],$$

and, since $D_n(f)$ is elementary, we have

$$\Phi_\# \mu\big[\, D_n[f] \,\big] = \mu\big[\, \Phi^* D_n[f] \,\big] = \mu\big[\, D_n[\Phi^* f] \,\big].$$

The equality (19.3.11) now follows from Corollary 19.3.9.

(ii) If $f \in \mathscr{L}^1(\Omega_1, \mathcal{S}_1, \Phi_\# \mu)$ then we write $f = f_+ - f_-$, $f_\pm \in \mathscr{L}^1(\Omega_1, \mathcal{S}_1, \Phi_\# \mu)$. The conclusion now follows from (i) applied to $f_\pm$. □

Remark 19.3.28 (Integration of complex valued functions)**.** Suppose that $(\Omega, \mathbb{S}, \mu)$ a measured space. A function $f : \Omega \to \mathbb{C}$ is said to be measurable if its real part $u = \mathbf{Re}\, f$ and its imaginary part $v = \mathbf{Im}\, f$ are measurable. We say that $f = u + \boldsymbol{i}v$ is μ-integrable if u and v are such. Note that u, v are simultaneously integrable iff $|u| + |v|$ is integrable. From the elementary inequalities

$$\frac{|u| + |v|}{\sqrt{2}} \leqslant \sqrt{u^2 + v^2} \leqslant (|u| + |v|)$$

We deduce that f is integrable if and only if $|f| = \sqrt{u^2 + v^2}$ is integrable. In this case we set

$$\int_\Omega f d\mu = \int_\Omega (u + \boldsymbol{i}v) d\mu := \int_\Omega u d\mu + \boldsymbol{i} \int_\Omega v d\mu.$$

The Dominated Convergence Theorem extends with no change to the complex case. □

19.3.2. Product measures and Fubini theorem. Suppose $(\Omega_i, \mathcal{S}_i)$, $i = 0, 1$, are two measurable spaces. Recall that $\mathcal{S}_0 \otimes \mathcal{S}_1$ is the sigma-algebra of subsets of $\Omega_0 \times \Omega_1$ generated by the collection $\mathcal{R}$ of "rectangles" of the form $S_0 \times S_1$, $S_i \in \mathcal{S}_i$, $i = 0, 1$.

The goal of this subsection is to show that two sigma-finite measures measures μ_i on $\mathcal{S}_i$, $i = 0, 1$ induce in a canonical way a measure $\mu_0 \otimes \mu_1$ uniquely determined by the condition

$$\mu_0 \otimes \mu_1\big[\, S_0 \times S_1 \,\big] = \mu_0\big[\, S_0 \,\big]\mu_1\big[\, S_1 \,\big], \;\; \forall S_i \in \mathcal{S}_i, \;\; i = 0, 1.$$

The collection $\mathcal{A}$ of subsets of $\Omega_0 \times \Omega_1$ that are finite disjoint unions of rectangles is an algebra; see Exercise 19.49. This suggests using Carathéodory's existence theorem to prove this claim.

We choose a different route that bypasses Carathéodory's existence theorem. This alternate, more efficient approach, is driven by the Monotone Class Theorem and simultaneously proves a central result in integration theory, the Fubini-Tonelli Theorem.

Lemma 19.3.29. *Suppose that*

$$f \in \mathcal{L}^0(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1) \cup \bar{\mathcal{L}}^0_+(\Omega, \mathcal{S}).$$

Then, for any $\omega_1 \in \Omega_1$ the function $f^0_{\omega_1} : \Omega_0 \to \overline{\mathbb{R}}$,

$$f^0_{\omega_1}(\omega_0) = f(\omega_0, \omega_1)$$

is $\mathcal{S}_0$-measurable and, for any $\omega_0 \in \Omega_0$, the function $f^1_{\omega_0} : (\Omega_1, \mathcal{S}_1) \to \overline{\mathbb{R}}$,

$$f^1_{\omega_0}(\omega_1) = f(\omega_0, \omega_1)$$

is $\mathcal{S}_1$-measurable.

Proof. Suppose first that $f \in \mathcal{L}^0\big(\, \Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1 \,\big)$ so that $\{f = \pm\infty\} = \varnothing$. We prove only the statement concerning $f^0_{\omega_1}$. For simplicity we will write f_{ω_1}instead of $f^0_{\omega_1}$. We will use the Monotone Class Theorem 19.1.25.

Denote by $\mathcal{M}$ the collection of functions $f \in \mathcal{L}^0(\Omega_0 \times \Omega_1, \mathcal{S}_0 \times \mathcal{S}_1)$ such that f_{ω_1} is $\mathcal{S}_0$-measurable, $\forall \omega_1 \in \Omega_1$. If $f, g \in \mathcal{M}$ are bounded, then $af + bg \in \mathcal{M}$, $\forall a, b \in \mathbb{R}$.

The collection $\mathcal{R}$ of rectangles is a π-system. Note that for any rectangle $R = S_0 \times S_1$ the function $f = \boldsymbol{I}_R$ belongs to $\mathcal{M}$. Indeed, for any $\omega_1 \in \Omega_1$ we have

$$f_{\omega_1} = \begin{cases} \boldsymbol{I}_{S_0}, & \omega_1 \in S_1, \\ 0, & \omega_1 \in \Omega_1 \backslash S_1. \end{cases}$$

If (f_n) is an increasing sequence of functions in $\mathcal{M}$ so is the sequence of slices f_{n,ω_1} so the limit f is also in $\mathcal{M}$. By the Monotone Class Theorem the collection $\mathcal{M}$ contains all the nonnegative measurable functions. Clearly, if $f \in \mathcal{M}$, then $-f \in \mathcal{M}$. If $f \in \mathcal{L}^0(\Omega, \mathcal{S})$, then $f = f_+ + (-f_-)$, $f_+, (-f_-) \in \mathcal{M}$. Hence $\mathcal{L}^0(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1) \subset \mathcal{M}$.

If $f \in \bar{\mathcal{L}}^0_+\big(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1\big)$, then for any $n \in \mathbb{N}$, $\min(n, f) \in \mathcal{L}^0\big(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1\big)$, $\min(n, f_{\omega_1}) \in \mathcal{L}^0(\Omega_0, \mathcal{S}_0)$ and $\min(n, f_{\omega_1})(\omega_0) \to f_{\omega_1}(\omega_0)$, $\forall \omega_0$, so that f_{ω_1} is $\mathcal{S}_0$-measurable. □

Let us emphasize that when $f \geqslant 0$, the conclusions of the lemma allow for f to have infinite values.

Theorem 19.3.30 (Fubini-Tonelli). *Let $(\Omega_i, \mathcal{S}_i, \mu_i)$, $i = 0, 1$ be two sigma-finite measured spaces.*

(i) *There exists a measure μ on $\mathcal{S}_0 \otimes \mathcal{S}_1$ uniquely determined by the equalities*

$$\mu\big[\, S_0 \times S_1 \,\big] = \mu_0\big[\, S_0 \,\big]\mu_1\big[\, S_1 \,\big], \quad \forall S_0 \in \mathcal{S}_0, \quad S_1 \in \mathcal{S}_1. \tag{19.3.12}$$

We will denote this measure by $\mu_0 \otimes \mu_1$.

(ii) *For each nonnegative function $f \in \bar{\mathcal{L}}^0_+(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1)$ the functions*

$$\omega_0 \mapsto \boldsymbol{I}_1\big[\, f \,\big](\omega_0) := \int_{\Omega_1} f(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big] \in [0, \infty],$$

$$\omega_1 \mapsto \boldsymbol{I}_0\big[\, f \,\big](\omega_1) := \int_{\Omega_0} f(\omega_0, \omega_1)\mu_0\big[\, d\omega_0 \,\big] \in [0, \infty]$$

are measurable and

$$\begin{aligned} &\int_{\Omega_0} \left(\int_{\Omega_1} f(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big] \right) \mu_0\big[\, d\omega_0 \,\big] \\ &\quad = \int_{\Omega_0 \times \Omega_1} f(\omega_0, \omega_1)\mu_0 \otimes \mu_1\big[\, d\omega_0 d\omega_1 \,\big] \\ &= \int_{\Omega_1} \left(\int_{\Omega_0} f(\omega_0, \omega_1)\mu_0\big[\, d\omega_0 \,\big] \right) \mu_1\big[\, d\omega_1 \,\big]. \end{aligned} \tag{19.3.13}$$

In particular, if only one of the three terms above is finite, then all three are finite and equal.

(iii) *Let $f \in \mathcal{L}^1(\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1, \mu_0 \otimes \mu_1)$. Then all the terms in (19.3.13) is well defined, finite and equal.*

Proof. Set

$$\Omega := \Omega_0 \times \Omega_1, \;\; \mathcal{S} = \mathcal{S}_0 \otimes \mathcal{S}_1, \;\; \mathcal{R} = \big\{\, S_0 \times S_1; \;\; S_i \in \mathcal{S}_i, \; i = 0,1 \,\big\} \subset \mathcal{S}$$

Assume first that the measures μ_0, μ_1 *are finite*, i.e., $\mu_i\big[\, \Omega_i \,\big] < \infty$, $i = 0,1$. Since μ_0, μ_1 are finite we deduce that there exists at most one measure $\nu : \mathcal{S} \to [0,\infty)$ satisfying (19.3.12).

Note that for any $f \in \mathcal{L}^\infty\big(\,\Omega, \mathcal{S}\,\big)$, the measurable functions $f^1_{\omega_0}$ and $f^0_{\omega_1}$ are bounded, for any $(\omega_0, \omega_1) \in \Omega_0 \times \Omega_1$. In particular these functions are also integrable because μ_0, μ_1 are finite.

Step 1. Denote by $\mathcal{M}$ the set of function $f \in \mathcal{L}^\infty\big(\,\Omega, \mathcal{S}\,\big)$ such that the function

$$\omega_0 \mapsto \boldsymbol{I}_1\big[\, f \,\big](\omega_0) = \int_{\Omega_1} f(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big] \in [0,\infty)$$

is measurable.

Observe first that $\mathcal{M}$ is a vector space and $\boldsymbol{I}_{S_0 \times S_1} \in \mathcal{M}$, $\forall S_i \in \mathcal{S}_i$. Moreover if (f_n) is a nondecreasing sequence of nonnegative sequence in $\mathcal{M}$ such that $f_n \nearrow f$ and $f \in \mathcal{L}^\infty$, then $f \in \mathcal{M}$ since the Monotone Convergence theorem implies

$$\boldsymbol{I}_1\big[\, f \,\big] = \lim_{n\to\infty} \boldsymbol{I}_1]f]$$

so $\boldsymbol{I}_1[f]$ is measurable as limit of measurable functions.

The collection of rectangles $S_0 \times S_1$ is a π-system that generates the sigma-algebra $\mathcal{S}_0 \times \mathcal{S}_1$ and the Monotone Class Theorem implies that $\mathcal{M} = \mathcal{L}^\infty\big(\,\Omega, \mathcal{S}\,\big)$.

Using increasing approximations by nonnegative elementary functions we deduce that for any $f \in \bar{\mathcal{L}}^0_+\big(\,\Omega_0 \times \Omega_1, \mathcal{S}_0 \otimes \mathcal{S}_1\,\big)$ the function

$$\omega_0 \mapsto \boldsymbol{I}_1\big[\, f \,\big](\omega_0) = \int_{\Omega_1} f(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big]$$

is measurable and thus

$$I_{1,0}\big[\, f \,\big] := \int_{\Omega_0} \left(\int_{\Omega_1} f(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big] \right) \mu_0\big[\, d\omega_0 \,\big] \in [0,\infty]$$

is well defined.

For $S \in \mathcal{S}_0 \otimes \mathcal{S}_1$ we set

$$\mu_{1,0}\big[\, S \,\big] = I_{1,0}\big[\, \boldsymbol{I}_S \,\big] = \int_{\Omega_1} \boldsymbol{I}_{\Omega_0 \times \Omega_1}(\omega_0, \omega_1)\mu_1\big[\, d\omega_1 \,\big].$$

If $\omega_0 \in \Omega_0 \backslash S_0$ the integral is 0. If $\omega_0 \in S_0$ the integral is

$$\int_{\Omega_1} \boldsymbol{I}_{S_1} d\mu_1 = \mu_1\big[\, S_1 \,\big].$$

Hence

$$\boldsymbol{I}_1\big[\, \boldsymbol{I}_{S_0 \times S_1} \,\big] = \mu_1\big[\, S_1 \,\big]\boldsymbol{I}_{S_0}.$$

We deduce

$$I_{1,0}\big[\,S_0\times S_1\,\big]=\mu_1\big[\,S_1\,\big]\int_{\Omega_0}\boldsymbol{I}_{S_0}d\mu_0=\mu_0\big[\,S_0\,\big]\cdot\mu_1\big[\,S_1\,\big].$$

Clearly if $A, A'\in\mathcal{S}$ are disjoint, then

$$\boldsymbol{I}_{A\cup A'}=\boldsymbol{I}_A+\boldsymbol{I}_{A'}$$

so

$$I_{1,0}\big[\,\boldsymbol{I}_{A\cup A'}\,\big]=I_{1,0}\big[\,\boldsymbol{I}_A\,\big]+I_{1,0}\big[\,\boldsymbol{I}'_A\,\big]$$

and

$$\mu_{1,0}\big[\,A\cup A'\,\big]=\mu_{1,0}\big[\,A\,\big]+\mu_{1,0}\big[\,A'\,\big].$$

If

$$A_1\subset A_2\subset\cdots$$

is an increasing sequence of sets in $\mathcal{S}$ and

$$A=\bigcup_{n\geqslant 1}A_n,$$

then invoking the Monotone Convergence Theorem we first deduce that $I_{1,0}\big[\,\boldsymbol{I}_{A_n}\,\big]$ is a nondecreasing sequence of measurable functions converging to $I_{1,0}\big[\,\boldsymbol{I}_A\,\big]$ and then we conclude that $\mu_{1,0}\big[\,A_n\,\big]$ converges to $\mu_{1,0}\big[\,A\,\big]$. Hence $\mu_{1,0}$ is a finite measure on $\mathcal{S}=\mathcal{S}_0\otimes\mathcal{S}_1$ satisfying (19.3.12).

Step 2. A similar argument shows that

$$\mu_{0,1}[S]=\int_{\Omega_1}\left(\int_{\Omega_0}\boldsymbol{I}_S(\omega_0,\omega_1)\mu_0\big[\,d\omega_0\,\big]\right)\mu_1\big[\,d\omega_1\,\big]$$

is also a finite measure on $\mathcal{S}=\mathcal{S}_0\otimes\mathcal{S}_1$ satisfying (19.3.12). We deduce that $\mu_{0,1}=\mu_{1,0}$. We denote this measure by $\mu_0\otimes\mu_1$.

Step 3. From **Step 2** we deduce that (19.3.13) is true for $f=\boldsymbol{I}_S$, $\forall S\in\mathcal{S}_0\otimes\mathcal{S}_1$. Using the Monotone Class Theorem we deduce (19.3.13) for $f\in\bar{\mathcal{L}}^0\big(\Omega,\mathcal{S}\big)$. This implies the equality (19.3.13) for any integrable f since f is the difference of two nonnegative integrable functions $f=f^+-f^-$ and the claim is true for $f^\pm$.

Step 4. *Suppose now that the measures μ_0,μ_1 are σ-finite.* Choose $E_i^n\in\mathcal{S}_i$ such that

$$E_i^n\subset E_i^{n+1},\ \ \mu_i\big[\,E_i^n\,\big]<\infty,\ \ \forall n\ \text{ and }\ \Omega_i=\bigcup_{n\geqslant 1}E_i^n,\ \ i=0,1.$$

For $i=0,1$, define the finite measures.

$$\mu_i^n:\mathcal{S}_i\to[0,\infty),\ \ \mu_i^n\big[\,S_i\,\big]=\mu_i\big[\,S_i\cap E_i^n\,\big].$$

We set $\nu^n:=\mu_0^n\otimes\mu_1^n$.

Proposition 19.1.34 shows that for any $S\in\mathcal{S}$, $\nu^n\big[\,S\,\big]=\nu^{n+1}\big[\,S\cap E^n\,\big]$ so that

$$\nu^n\big[\,S\,\big]\leqslant\nu^{n+1}\big[\,S\,\big],\ \ \forall S\in\mathcal{S}.$$

For $S \in \mathcal{S}$ we set

$$\nu\big[\,S\,\big] = \lim_{n\to\infty} \nu_n\big[\,S \cap E^n\,\big].$$

Note that $\nu_n\big[\,S\,\big] = \nu\big[\,S \cap E^n\,\big]$.

Clearly $\nu : \mathcal{S} \to [0,\infty]$ is finitely additive. Let us show that is ν is sigma-additive, i.e., if $S_k \nearrow S$ in $\mathcal{S}$, then

$$\lim_k \nu\big[\,S_k\,\big] = \nu\big[\,S\,\big].$$

Let $c \in \mathbb{R}$ such that $c < \nu\big[\,S\,\big]$. Choose N sufficiently large so

$$c < \nu\big[\,S \cap E^n\,\big] \leqslant \nu\big[\,S\,\big],\ \ \forall n \geqslant N.$$

Choose K sufficiently large so that

$$c < \ \nu\big[\,S_K \cap E^N\,\big] = \nu^N\big[\,S_K\,\big] \leqslant \nu^N\big[\,S\,\big] = \nu\big[\,S \cap E^N\,\big] \leqslant \nu\big[\,S\,\big].$$

Hence

$$c < \nu^N\big[\,S_K\,\big] \leqslant \nu\big[\,S_K\,\big] \leqslant \nu\big[\,S\,\big].$$

We deduce that $\forall k \geqslant K$, $c < \nu\big[\,S_k\,\big] \leqslant \nu\big[\,S\,\big]$, so that

$$c < \lim_k \nu\big[\,S_k\,\big] \leqslant \nu\big[\,S\,\big],\ \ \forall c < \nu\big[\,S\,\big]$$

i.e.,

$$\lim_k \nu\big[\,S_k\,\big] = \nu\big[\,S\,\big].$$

Note that for any rectangle $R = S_0 \times S_1$ we have $R \cap E_n = (S_0 \cap E_n^0) \times (S_1 \cap E_n^1)$ so

$$\nu\big[\,R\,\big] = \lim_{n\to\infty} \nu\big[\,R \cap E_n\,\big] = \lim_{n\to\infty} \mu_0^n \otimes \mu_1^n\big[\,R \cap E_n\,\big]$$

$$= \lim_{n\to\infty} \mu_0\big[\,S_0 \cap E_n^0\,\big]\mu_1\big[\,S_1 \cap E_n^1\,\big] = \mu_0\big[\,S_0\,\big]\mu_1\big[\,S_1\,\big].$$

In other words ν satisfies (19.3.12). If $\nu' : \mathcal{S} \to [0,\infty]$ is another measure satisfying (19.3.12), then for any $S_i \in \mathcal{S}_i$,

$$\nu'\big[\,(S_0 \times S_1) \cap E^N n\,\big] = \mu_0\big[\,S_0 \cap E_0^n\,\big]\mu_1\big[\,S_1 \cap E_1^n\,\big]$$

and Proposition 19.1.34 implies that $\nu'\big[\big[\,S \cap E^n\,\big] = \mu_0^n \otimes \mu_1^n\big[\,S\,\big] = \nu^N\big[\,S\,\big]$. Letting $n \to \infty$ we deduce $\nu' = \nu$. This proves the existence of a unique sigma-finite measure with prescribed behavior on the rectangles.

Step 5. Let $f \in \bar{\mathcal{L}}^0_+\big(\Omega, \mathcal{S}\big)$. Then for any n the function $f_n = f\boldsymbol{I}_{E^n}$ satisfies (19.3.13) because in this case these equalities follows from the corresponding equalities for the finite measures μ_i^n. Letting $n \to \infty$ and invoking the Monotone Convergence Theorem we deduce that $\mu_0 \otimes \mu_1$ satisfies (19.3.13) for any $f \in \bar{\mathcal{L}}^0_+\big(\Omega, \mathcal{S}\big)$. We deduce the equality for any integrable f by writting $f = f^+ - f^-$. The equality (19.3.13) is true for $f^\pm$ and in these cases all thre three terms of this identity are finite.

□

Remark 19.3.31. Let $f \in \bar{\mathcal{L}}^0_+(\Omega_0 \times \Omega_1, \mathcal{S}_0 \times \mathcal{S}_1, \mu_0 \otimes \mu_1)$. From (19.3.13)

$$\begin{aligned}\int_{\Omega_0}\left(\int_{\Omega_1} \big| f(\omega_0,\omega_1) \big| \mu_1\big[d\omega_1 \big]\right) \mu_0\big[d\omega_0 \big]\\ = \int_{\Omega_0\times\Omega_1} \big| f(\omega_0,\omega_1) \big| \mu_0 \otimes \mu_1\big[d\omega_0 d\omega_1 \big] \\ = \int_{\Omega_1}\left(\int_{\Omega_0} \big| f(\omega_0,\omega_1) \big| \mu_0\big[d\omega_0 \big]\right) \mu_1\big[d\omega_1 \big].\end{aligned} \tag{19.3.14}$$

Above each the terms is well defined and could be infinite. However the above equality shows that *if one of the three terms above is finite, then they all are finite and equal.* Thus, for $f \geqslant 0$ to be integrable it suffices that the first or the third term in (19.3.14) be finite. □

The above construction can be iterated. More precisely given sigma-finite measured spaces $(\Omega_k, \mathcal{S}_k, \mu_k)$, $k = 1, \dots, n$, we have a measure $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ uniquely determined by the condition

$$\mu\big[S_1 \times S_2 \times \cdots \times S_n \big] = \mu_1\big[S_1 \big]\mu_2\big[S_2 \big] \cdots \mu_n\big[S_n \big], \;\; \forall S_k \in \mathcal{S}_k, \;\; k = 1, \dots, n.$$

The Borel sigma-algebra $\mathcal{B}_{\mathbb{R}^n}$ generated by the open subsets of $\mathbb{R}^n$ satisfies

$$\mathcal{B}_{\mathbb{R}^n} = \underbrace{\mathcal{B}_{\mathbb{R}} \otimes \cdots \otimes \mathcal{B}_{\mathbb{R}}}_{n}$$

and the Lebesgue measure $\boldsymbol{\lambda}$ on $\mathcal{B}_{\mathbb{R}}$ induces a measure on $\mathbb{R}^n$

$$\boldsymbol{\lambda}_n := \underbrace{\boldsymbol{\lambda} \otimes \cdots \otimes \boldsymbol{\lambda}}_{n}.$$

We will refer to $\boldsymbol{\lambda}_n$ as the *n-dimensional Lebesgue measure.* We denote by $\mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}^n}$ the completion of the Borel algebra $\mathcal{B}_{\mathbb{R}^n}$ with respect to the Lebesgue measure $\boldsymbol{\lambda}$. We will refer to the sets in $\mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}^n}$ as *Lebesgue measurable* subsets.

Remark 19.3.32. Note that we have at our disposal another sigma-algebra

$$\underbrace{\mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}} \otimes \cdots \otimes \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}}}_{n} \supset \mathcal{B}_{\mathbb{R}^n}$$

over which $\boldsymbol{\lambda}_n$ can be defined. What is the relationship between this sigma-algebra and the completion $\mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}^n}$? For simplicity we address this question only the case $n = 2$.

Note that if $S_0, S_1 \in \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}}$, then there exist $\boldsymbol{\lambda}$-negligible Borel subsets $\mathbb{R} \supset \widetilde{S}_i \subset S_i$, $i = 0, 1$. Then

$$S_0 \times S_1 \subset \widetilde{S}_0 \times \widetilde{S}_1$$

Since $\widetilde{S}_0 \times \widetilde{S}_1$ is $\boldsymbol{\lambda}_2$-negligible we deduce that $S_0 \times S_1 \in \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}^2}$. Since the collection of rectangles $S_0 \times S_1$, $S_i \in \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}}$ is a π-system we deduce from the $\pi - \lambda$ theorem that $\mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}} \otimes \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}} \subset \mathcal{B}^{\boldsymbol{\lambda}}_{\mathbb{R}^2}$. The inclusion is strict, [7, Sec. 5.1,Exer. 5].□

Any subset $X \subset \mathbb{R}^n$ is a metric space with the induced metric and, as such, it has a Borel algebra of subsets. More precisely a subset $S \subset X$ belongs to the Borel algebra $\mathcal{B}_X$ if and only if there exists $B \in \mathcal{B}_{\mathbb{R}^n}$ such that $S = B \cap X$.

Suppose that $X \subset \mathbb{R}^n$ is itself Borel. Then any Borel subset $S \subset X$ is also a Borel subset of $\mathbb{R}^n$. The Lebesgue measure $\boldsymbol{\lambda}_n$ on $\mathbb{R}^n$ induces a measure $\boldsymbol{\lambda}_{n,X} : \mathcal{B}_X \to [0, \infty]$,

$$\boldsymbol{\lambda}_{n,X}\big[\, S \,\big] = \boldsymbol{\lambda}_n\big[\, S \,\big].$$

For simplicity, and when no confusion is possible, we will use the same notation $\boldsymbol{\lambda}_n$, when referring to $\boldsymbol{\lambda}_{n,X}$. We will denote by $\mathcal{B}_X^{\boldsymbol{\lambda}}$ the completion of $\mathcal{B}_X$ with respect to the induced Lebesgue measure, i.e., the sigma-algebra of Lebesgue measurable subsets of X.

Proposition 19.3.33. *Suppose that $B \subset \mathbb{R}^n$ is a nondegenerate box and $f : B \to \mathbb{R}$ is a Riemann integrable function. Then $f \in \mathcal{L}^1(B, \mathcal{B}_B^{\boldsymbol{\lambda}}, \boldsymbol{\lambda}_n)$ and*

$$\int_B f(x)\boldsymbol{\lambda}_n\big[\, dx \,\big] = \int_B f(x)|dx|,$$

where the integral in the right-hand side is the Riemann integral.

Proof. Using the decomposition $f = f = f_+ - f_-$ we see that it suffices to prove the result only in the case $f \geqslant 0$. We will use the terminology and notation in Subsection 15.1.1. Since f is Riemann integrable there exists a sequence $(\boldsymbol{P}_\nu)_{\nu \in \mathbb{N}}$ of partitions of B such that

$$\omega_\nu := \boldsymbol{S}^*(f, \boldsymbol{P}_\nu) - \boldsymbol{S}_*(f, \boldsymbol{P}_\nu) < 2^{-\nu}$$

For each ν we define the elementary functions

$$f_\nu = \sum_{C \in \mathscr{C}(\boldsymbol{P}_\nu)} m_C(f)\boldsymbol{I}_{C^\circ}, \quad F_\nu = \sum_{C \in \mathscr{C}(\boldsymbol{P}_\nu)} M_C(f)\boldsymbol{I}_C,$$

where C° denotes the interior of the chamber C. Observe that

$$0 \leqslant f_\nu(x) \leqslant f(x) \leqslant F_\nu(x), \quad \forall x \in B$$

and

$$\int_B f_\nu(x)\boldsymbol{\lambda}_n\big[\, dx \,\big] = \boldsymbol{S}_*(f, \boldsymbol{P}_\nu), \quad \int_B F_\nu(x)\boldsymbol{\lambda}_n\big[\, dx \,\big] = \boldsymbol{S}^*(f, \boldsymbol{P}_\nu).$$

We set $g_\nu := F_\nu - f_\nu$ so that

$$\int_B g_\nu d\boldsymbol{\lambda}_n = \omega_\nu.$$

Observe that $0 \leqslant f - f_\nu \leqslant g_\nu$. From Markov's inequality we deduce that

$$\boldsymbol{\lambda}_n\big[\, \{g_\nu \geqslant r\} \,\big] \leqslant \frac{2^{-\nu}}{r}. \tag{19.3.15}$$

Given $x \in B$, the sequence $f_\nu(x)$ does not converge to $f(\omega)$ iff

$$\exists k \in \mathbb{N}, \;\; \forall m \in \mathbb{N}, \;\; \exists \nu > m : \;\; f(x) - f_\nu(x) > 1/k.$$

Hence, if $f_\nu(x)$ does not converge to $f(x)$, then

$$x \in Z := \bigcup_k \underbrace{\bigcap_m \bigcup_{\nu > m} \big\{\, g_\nu > 1/k \,\big\}}_{=:Z_k}.$$

We have

$$\boldsymbol{\lambda}_n\left[\bigcup_{\nu>m}\{\,g_\nu>1/k\,\}\right]\leqslant\sum_{\nu>m}\boldsymbol{\lambda}_n\big[\,\{\,g_\nu>1/k\,\}\,\big]\overset{(19.3.15)}{\leqslant}\sum_{\nu>m}k2^{-\nu}=k2^{-m}.$$

Hence

$$\boldsymbol{\lambda}_n\big[\,Z_k\,\big]\leqslant k2^{-m},\quad\forall m\in\mathbb{N}.$$

This shows that $\boldsymbol{\lambda}_n\big[\,Z_k\,\big]=0$, $\forall k$ so $\boldsymbol{\lambda}_n\big[\,Z\,\big]=0$ and thus $f_\nu\to f$ a.e. Since the functions f_ν are $\mathcal{B}_B^{\boldsymbol{\lambda}}$-measurable and $\mathcal{B}_B^{\boldsymbol{\lambda}}$ is complete, we deduce from Corollary 19.1.38 that f is also $\mathcal{B}_B^{\boldsymbol{\lambda}}$-measurable. Note that

$$0\leqslant f_\nu\leqslant f\leqslant M:=\sup_B f<\infty.$$

Since the constant function M is Lebesgue integrable over B we deduce from the Dominated convergence theorem that

$$\int_B f(x)\boldsymbol{\lambda}_n\big[\,dx\,\big]=\lim_\nu\int_B f_\nu(x)\,\boldsymbol{\lambda}_n\big[\,dx\,\big]=\lim_{\nu\to\infty}\boldsymbol{S}_*(f_\nu,\boldsymbol{P}_\nu)=\int_B f(x)|dx|.$$

□

Remark 19.3.34. The above proof implies among other things that any Riemann integrable function is Lebesgue measurable. This is the best one can hope for since there exist Riemann integrable functions that are not Borel measurable. For example, for any subset S of the Cantor set C, the indicator $\boldsymbol{I}_S$ is Riemann integrable according to Lebesgue's Theorem 15.1.17. However this function is Borel measurable iff S is Borel. □

Corollary 19.3.35. *If $f:\mathbb{R}^n\to\mathbb{R}$ is Riemann integrable, then $f\in\mathcal{L}^1(\mathbb{R}^n,\mathcal{B}_{\mathbb{R}^n}^{\boldsymbol{\lambda}},\boldsymbol{\lambda}_n)$ and*

$$\int_{\mathbb{R}^n}f(x)\boldsymbol{\lambda}\big[\,dx\,\big]=\int_{\mathbb{R}^n}f(x)\,|dx|.\tag{19.3.16}$$

Proof. Since f is Riemann integrable there exists a box B such that $\operatorname{supp}f\subset B$ and $f\big|_B$ is Riemann integrable. By definition

$$\int_{\mathbb{R}^n}f(x)|dx|=\int_B f(x)\,|dx|.$$

The equality (19.3.16) now follows from Proposition 19.3.33. □

Corollary 19.3.36. *Any Jordan measurable set $S\subset\mathbb{R}^n$ is Lebesgue measurable and*

$$\operatorname{vol}_n\big[\,S\,\big]=\boldsymbol{\lambda}_n\big[\,S\,\big].$$

Proof. Since S is Jordan measurable it is contained in a box $B\subset\mathbb{R}^n$ and the induced function

$$\boldsymbol{I}_S:B\to\mathbb{R}$$

is Riemann integrable. Proposition 19.3.33 implies

$$\boldsymbol{\lambda}_n\big[\,S\,\big] = \int_B \boldsymbol{I}_S(x)\boldsymbol{\lambda}_n\big[\,dx\,\big] = \int_B \boldsymbol{I}_S(x)\,|dx| = \mathrm{vol}_n(S).$$

□

Proposition 19.3.37. *Suppose that $U \subset \mathbb{R}^n$ is an open set and $\Phi : U \to \mathbb{R}^n$ is a C^1-diffeomorphism with $V := \Phi(U)$. Denote $u = (u^1, \dots, u^n)$ the coordinates of the points in U and by $v = (v^1, \dots, v^n)$ the coordinates of the points in V.*

Then for any Borel subset $B \subset V$ we have

$$\Phi_{\#}\boldsymbol{\lambda}_V\big[\,B\,\big] = \int_B |\det J_{\Phi^{-1}}(v)|\,\boldsymbol{\lambda}_n\big[\,dv\,\big], \tag{19.3.17}$$

where $\Phi_{\#}\boldsymbol{\lambda}_U$ is the pushforward of the measure $\boldsymbol{\lambda}_U$ via the map Φ,

$$\Phi_{\#}\boldsymbol{\lambda}_U\big[\,B\,\big] := \boldsymbol{\lambda}_U\big[\,\Phi^{-1}(B)\,\big],\;\; \forall B \in \mathcal{B}_V.$$

In particular, if $T : \mathbb{R}^n \to \mathbb{R}^n$ is bijective linear map and $B \subset \mathbb{R}^n$ is a Borel subset then

$$\boldsymbol{\lambda}\big[\,T(B)\,\big] = |\det T| \cdot \boldsymbol{\lambda}\big[\,B\,\big]. \tag{19.3.18}$$

Proof. Denote by $\mathcal{B}_V$ the Borel sigma-algebra of V. We have to show that

$$\boldsymbol{\lambda}_U\big[\,\Phi^{-1}(B)\,\big] = \int_B |\det J_{\Phi^{-1}}(v)|\,\boldsymbol{\lambda}_V\big[\,dv\,\big],\;\; \forall B \in \mathcal{B}_V. \tag{19.3.19}$$

Denote by $\mathcal{F}$ the family of Borel subset of V for which (19.3.19) holds. To prove that $\mathcal{F} = \mathcal{B}_V$ we will use the $\pi - \lambda$ theorem and we will show that $\mathcal{F}$ is a λ-system that contains a π-system that generates $\mathcal{B}_V$ as a sigma-algebra.

For any Jordan measurable compact $K \subset V$ the function

$$f : \mathbb{R}^n \to \mathbb{R},;\;\; f(v) = |\det J_{\Phi^{-1}}(v)|\boldsymbol{I}_K(v).$$

is Riemann integrable. Using the change in variables formula, Theorem 15.3.1, we deduce that the function

$$f \circ \Phi : U \to \mathbb{R}$$

is Riemann integrable and

$$\int_{\Phi^{-1}(K)} f\big(\Phi(u)\big)|\det J_\Phi(u)|\,|du| \overset{(19.3.16)}{=} \int_K f(v)\,|dv| = \int_B |\det J_{\Phi^{-1}}(v)|\,\boldsymbol{\lambda}_V\big[\,dv\,\big].$$

Now observe that

$$f\big(\Phi(u)\big)|\det J_\Phi(u)| = \big|\,\det J_{\Phi^{-1}}\big(\Phi(u)\big)\,\det J_\Phi(u)\big|$$

The Chain Rule shows that we have an equality of matrices

$$\det J_{\Phi^{-1}}\big(\Phi(u)\big)J_\Phi(u) = J_{\Phi^{-1}\circ\Phi(u)} = J_{\mathbb{1}} = \mathbb{1},$$

so

$$\det J_{\Phi^{-1}}\big(\Phi(u)\big)\,\det J_\Phi(u) = 1$$

Hence (19.3.19) is true for Jordan measurable compact subsets of V. The collection of Jordan measurable compact subsets of V is a π-system that generates $\mathcal{B}_V$.

Fix a Jordan measurable compact exhaustion (K_ν) of V; see Definition 15.4.4. From the Monotone Convergence Theorem we deduce that (19.3.19) is true for

$$V = \bigcup_{\nu \geqslant 1} K_\nu.$$

Hence $V \in \mathcal{F}_V$. Obviously $S_0, S_1 \in \mathcal{F}_V$, $S_0 \subset S_1$, then $S_1 \backslash S_0$. The Monotone Convergence Theorem shows that if (S_ν) is an increasing sequence in $\mathcal{F}_V$, then so is its union.

If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a bijective linear map, then using (19.3.17) with $\Phi = T^{-1}$ we deduce

$$\boldsymbol{\lambda}\big[\, T(B)\,\big] = T^{-1}_{\#}\boldsymbol{\lambda}\big[\, B\,\big] = |\det T| \cdot \boldsymbol{\lambda}\big[\, B\,\big].$$

□

Corollary 19.3.38 (Change in variables)**.** *Suppose that $U \subset \mathbb{R}^n$ is an open set and $\Phi : U \to \mathbb{R}^n$ is a C^1-diffeomorphism with $V := \Phi(U)$. If*

$$f \in \mathcal{L}^1(V, \mathcal{B}_V, \boldsymbol{\lambda}_V),$$

then $(f \circ \Phi)|\det J_\Phi| \in \mathcal{L}^1(U, \mathcal{B}_U, \boldsymbol{\lambda})$ and

$$\int_U f\big(\Phi(u)\big)|\det J_\Phi(u)|\,\boldsymbol{\lambda}\big[\, du\,\big] = \int_V f(v)\,\boldsymbol{\lambda}\big[\, dv\,\big]$$

Proof. According to Proposition 19.3.37 we have

$$\Phi_{\#}\boldsymbol{\lambda}_U = |\det J_{\Phi^{-1}}|\boldsymbol{\lambda}_V.$$

Define

$$g : V \to (-\infty, \infty], \;\; g(v) = f(v)\big|\det J_\Phi\big(\Phi^{-1}(v)\big)\big|.$$

Note that

$$g(v)\big|\det J_{\Phi^{-1}}(v)\,\big| = f(v)\underbrace{\big|\det J_\Phi\big(\Phi^{-1}(v)\big)\big| \cdot \big|\det J_{\Phi^{-1}}(v)\,\big|}_{=|\det J_{\Phi\circ\Phi^{-1}}(v)|} = f(v).$$

Hence $g \in \mathcal{L}^1(V, \mathcal{B}_V, \Phi_{\#}\boldsymbol{\lambda}_U)$ since $f \in \mathcal{L}^1(V, \mathcal{B}_V, \boldsymbol{\lambda}_V)$. Using Theorem 19.3.27 we deduce

$$\int_V f(v)\boldsymbol{\lambda}_V\big[\, dv\,\big] = \int_V g(v)\Phi_{\#}\boldsymbol{\lambda}_V\big[\, dv\,\big]$$

$(v = \Phi(u),\, u = \Phi^{-1}(v))$

$$= \int_U g\big(\Phi(u)\big)\boldsymbol{\lambda}_U\big[\, du\,\big] = \int_U f\big(\Phi(u)\big)\big|\det J_\Phi(u)\,\big|\boldsymbol{\lambda}_U\big[\, du\,\big].$$

□

19.4. The L^p-spaces

We have developed all the technology required to introduce and investigate a class of Banach spaces that play a very important role in the modern analysis and its applications. Fix a measured space $(\Omega, \mathcal{S}, \mu)$. Recall our convention that $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ implies that $|f(\omega)| < \infty$ for any $\omega \in \Omega$.

19.4.1. Definition and Hölder inequality. For any $p \in [1, \infty)$ and $f \in \mathcal{L}^0(\Omega, \mathcal{S})$ we set

$$\|f\|_{L^p} = \|f\|_{L^p(\mu)} := \left(\int_\Omega |f(\omega)|^p \, \mu\big[\, d\omega\,\big]\right)^{\frac{1}{p}} \in [0, \infty],$$

and we define

$$\mathcal{L}^p(\Omega, \mathcal{S}, \mu) := \big\{\, f \in \mathcal{L}^0(\Omega, \mathcal{S});\;\; \|f\|_{L^p(\mu)} < \infty \,\big\},$$
$$\mathcal{L}^p_+(\Omega, \mathcal{S}, \mu) := \big\{\, f \in \mathcal{L}^p(\Omega, \mathcal{S});\;\; f \geqslant 0 \;\text{ a. e. }\,\big\}.$$

Define

$$\mathcal{L}^\infty(\Omega, \mathcal{S}, \mu) := \big\{\, f \in \mathcal{L}^0(\Omega, \mathcal{S});\;\; \exists C > 0 : \;\; |f(\omega)| \leqslant C \;\; \mu - \text{a. e.} \,\big\},$$
$$\mathcal{L}^\infty_+(\Omega, \mathcal{S}, \mu) := \big\{\, f \in \mathcal{L}^\infty(\Omega, \mathcal{S}, \mu);\;\; f \geqslant 0 \; \mu - \text{a. e.} \,\big\}.$$

For $f \in \mathcal{L}^\infty(\Omega, \mathcal{S}\mu)$ we set

$$\|f\|_{L^\infty} := \inf\big\{C \in \mathbb{Q};\;\; |f| \leqslant C \;\; \mu - \text{a. e.} \,\big\}.$$

Observe that for $p \in [1, \infty]$ we have

$$\|f\|_{L^p} = 0 \Longleftrightarrow f = 0 \;\; \mu - \text{a. e.}.$$

Clearly $\mathcal{L}^1$ and $\mathcal{L}^\infty$ are vector spaces. For $p > 1$ the function $h : (0, \infty) \to \mathbb{R}$, $h(x) = x^p$ is convex and thus

$$h\big(\,(x+y)/2\,\big) \leqslant \frac{1}{2}\big(\,h(x) + h(y)\,\big),$$

i.e.,

$$|x+y|^p \leqslant 2^{p-1}\big(\,x^p + y^p\,\big), \;\; \forall x, y > 0.$$

In particular, for any $f, g \in \mathcal{L}^p\big(\Omega, \mu\,\big)$ we have

$$\big|\, f(\omega) + g(\omega)\,\big|^p \leqslant \big|\,|f(\omega)| + |g(\omega)|\,\big|^p \leqslant 2^{p-1}\big(\,|f(\omega)|^p + |g(\omega)|^p\,\big).$$

so that

$$\int_\Omega \big|\, f(\omega) + g(\omega)\,\big|^p \, \mu\big[\, d\omega\,\big] \leqslant 2^{p-1} \int_\Omega \big(\,|f(\omega)|^p + |g(\omega)|^p\,\big)\, \mu\big[\, d\omega\,\big] < \infty$$

so that $f + g \in \mathcal{L}^p(\Omega, \mathcal{S}, \mu)$. This proves that $\mathcal{L}^p(\Omega, \mathcal{S}, \mu)$ is also vector space for $p \in (1, \infty)$.

For $p \in [1, \infty]$ we denote by p^* its conjugate exponent defined by

$$\frac{1}{p^*} + \frac{1}{p} = 1 \Longleftrightarrow p^* = \frac{p}{p-1}.$$

Note that $1^* = \infty$ and $(p^*)^* = p$.

Theorem 19.4.1 (Hölder's inequality)**.** *For any $f, g \in \mathcal{L}^0_+(\Omega, \mathcal{S})$ and any $p \in [1, \infty]$ we have*

$$\int_\Omega f(\omega)g(\omega)\mu\big[\,d\omega\,\big] \leqslant \|f\|_{L^p(\mu)} \cdot \|g\|_{L^{p^*}(\mu)}. \tag{19.4.1}$$

In particular, if $f \in \mathcal{L}^p(\Omega, \mathcal{S}, \mu)$ and $g \in \mathcal{L}^{p^*}(\Omega, \mathcal{S}, \mu)$, then $fg \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$.

Proof. Set $q := p^*$. Suppose first that f, g are elementary functions. We can then find a common measurable partition $(S_i)_{1 \geqslant i \leqslant n}$ of Ω such that

$$f = \sum_{i=1}^n f_i \boldsymbol{I}_{S_i}, \quad g = \sum_{i=1}^n g_i \boldsymbol{I}_{S_i}, \quad f_i, g_i \in [0, \infty).$$

Set

$$x_i = f_i \mu\big[\,S_i\,\big]^{1/p}, \quad y_i = g_i \mu\big[\,S_i\,\big]^{1/q}.$$

Then

$$\int_\Omega f(\omega)g(\omega)\mu\big[\,d\omega\,\big] = \sum_{i=1}^n x_i y_i,$$

$$\|f\|_{L^p} = \left(\sum_{i=1}^n x_i^p\right)^{1/p}, \quad \|g\|_{L^q} = \left(\sum_{i=1}^n y_i^q\right)^{1/q}.$$

Thus, in this case the inequality (19.4.1) becomes

$$\sum_{i=1}^n x_i y_i \leqslant \left(\sum_{i=1}^n x_i^p\right)^{1/p} \left(\sum_{i=1}^n y_i^q\right)^{1/q}$$

which is the classical Hölder inequality (8.3.15). Thus (19.4.1) holds when f, g are elementary. In general, for any $f, g \in \mathcal{L}^0_+$, we have

$$\int_\Omega D_n[f](\omega) D_n[g](\omega)\mu\big[\,d\omega\,\big] \leqslant \big\|\,D_n[f]\,\big\|_{L^p(\mu)} \cdot \big\|\,D_n[g]\,\big\|_{L^{p^*}(\mu)}, \quad \forall n \in \mathbb{N}.$$

Letting $n \to \infty$ and invoking the Monotone Convergence Theorem we obtain (19.4.1) in general.

□

When $p = 2$, we have $p^* = 2$ and Hölder's inequality specializes to the Cauchy-Schwarz inequality.

Corollary 19.4.2 (Cauchy-Schwarz inequality)**.** *For any $f, g \in \mathcal{L}^2(\Omega, \mathcal{S}\mu)$ we have*

$$\left|\int_\Omega f(\omega)g(\omega)\mu\big[\,d\omega\,\big]\right| \leqslant \int_\Omega \big|f(\omega)\big|\,\big|g(\omega)\big|\,\mu\big[\,d\omega\,\big] \leqslant \|f\|_{L^2} \cdot \|g\|_{L^2}. \tag{19.4.2}$$

□

Theorem 19.4.3 (Minkowski's inequality)**.** *Let* $p \in [1,\infty]$*. Then, for any* $f,g \in \mathscr{L}^p(\Omega,\mathcal{S},\mu)$ *we have*

$$\|f+g\|_{L^p} \leqslant \|f\|_{L^p} + \|g\|_{L^p}. \tag{19.4.3}$$

Proof. The inequality is obviously true for $p=1$ or $p=\infty$ so we will assume $p \in (1,\infty)$. Set $q = p^* = \frac{p}{p-1}$. We have

$$\boxed{\|f+g\|_{L^p}^p} = \int_\Omega |f+g|^p d\mu \leqslant \int_\Omega |f|\cdot|f+g|^{p-1}d\mu + \int_\Omega |g|\cdot|f+g|^{p-1}d\mu$$

$((|f+g|^{p-1})^q = |f+g|^p \in \mathscr{L}^1)$

$$\overset{(19.4.1)}{\leqslant} \|f\|_{L^p}\cdot\big\|\,|f+g|^{p-1}\,\big\|_{L^q} + \|g\|_{L^p}\cdot\big\|\,|f+g|^{p-1}\,\big\|_{L^q}$$

$(\big\|\,|f+g|^{p-1}\,\big\|_{L^q} = \|f+g\|_{L^p}^{p-1})$

$$= \boxed{\big(\,\|f\|_{L^p} + \|g\|_{L^p}\,\big)\cdot\|f+g\|_{L^p}^{p-1}}.$$

□

Minkowski's inequality shows that the correspondence

$$\mathscr{L}^p(\Omega,\mathcal{S},\mu) \ni f \mapsto \|f\|_{L^p} \in [0,\infty)$$

behaves almost like a norm, but with one notable exception. From the equality $\|f\|_{L^p} = 0$ we cannot conclude that $f=0$. The best that we can conclude is that $f=0$ μ-a.e.. To address this issue we introduce the relation $\sim$ on $\mathscr{L}^p(\Omega,\mathcal{S},\mu)$,

$$f \sim g \Longleftrightarrow f = g \;\; \mu - \text{a. e.}\,.$$

Clearly $\sim$ is an equivalence relation and the equality $\|f\|_{L^p} = 0$ can be rewritten as $f \sim 0$. Note that

$$f \sim f',\;\; g \sim g' \;\Rightarrow\; \|f\|_{L^p} = \|f'\|_{L^p},\;\; f+g \sim f'+g',\;\; cf \sim cf',\;\; \forall f,g \in \mathscr{L}^0,\;\; c \in \mathbb{R}.$$

This proves that the quotient space

$$L^p(\Omega,\mathcal{S},\mu) := \mathscr{L}^p(\Omega,\mathcal{S}\mu)/\sim$$

is a vector space, and the function $\|-\|_{L^p}$ descends to a genuine norm on $L^p(\Omega,\mathcal{S},\mu)$.

If we dednote by $\mathcal{S}^\mu$ the μ-completion of $\mathcal{S}$, then Remark 19.3.26 shows that

$$L^p(\Omega,\mathcal{S},\mu) = L^p(\Omega,\mathcal{S}^\mu,\mu).$$

19.4.2. The Banach space $L^p(\Omega,\mathcal{S}\mu)$**.** An important payoff of the elaborate integration theory we have been building is that the collection of functions integrable via this technology is very large and it is closed under rather flexible convergence types. The next fundamental result is a concrete consequence of these nice features.

Theorem 19.4.4 (Riesz)**.** *For any $p \in [1, \infty]$ the normed space $\big(L^p(\Omega, \mathcal{S}, \mu), \| - \|_{L^p} \big)$ is complete, i.e., it is a Banach space.*

Proof. The case $p = \infty$ is very similar to the situation described in Example 17.2.7 and we leave its proof to the reader; see Exercise 19.68. Assume that $p \in [1, \infty)$.

Suppose that $(f_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in L^p. To prove that it converges it suffices to show that it has a convergent subsequence.

Observe that there exists a subsequence $(f_{n_k})_{k \geqslant 1}$ of f_n such that

$$\| f_{n_k} - f_{n_{k+1}} \|_{L^p} < 2^{-k}.$$

For simplicity we set $g_k := f_{n_k}$. Consider the nondecreasing sequence of nonnegative measurable functions

$$S_N(\omega) = \sum_{k=1}^N |g_k(\omega) - g_{k+1}(\omega)|, \;\; \omega \in \Omega.$$

We set

$$S_\infty = \lim_{N \to \infty} S_N = \sum_{k=1}^\infty |g_k - g_{k+1}|$$

and we observe that

$$\int_\Omega |S_\infty|^p d\mu = \lim_{N \to \infty} \int_\Omega |S_N|^p d\mu = \lim_{N \to \infty} \|S_N\|_{L^p}^p$$

(use Minkowski's inequality)

$$\leqslant \lim_{N \to \infty} \left(\sum_{k=1}^N \|g_k - g_{k+1}\|_{L^p} \right)^p \leqslant \left(\sum_{k=1}^\infty 2^{-kp} \right)^p < \infty.$$

Hence $S_\infty \geqslant 0$ and

$$\int_\Omega |S_\infty|^p d\mu < \infty,$$

so that $S_\infty < \infty$ μ-a.e. Since

$$\sum_{k=1}^\infty |g_k - g_{k+1}| \leqslant S_\infty,$$

we deduce that the series $\sum_{k=1}^\infty |g_k - g_{k+1}|$ converges μ-a.e.. This implies that the series

$$g_1 + \sum_{k=1}^\infty (g_{k+1} - g_k)$$

converges μ-a.e. Note that

$$g_1 + \sum_{k=1}^m (g_{k+1} - g_k) = g_{m+1}$$

so that the sequence (g_k) converges μ-a.e.. to a function $h:\Omega\to\mathbb{R}$. By modifying all of the g_k-s on a negligible set N we can assume that this sequence converges everywhere to h so h is measurable. Observe that

$$\|g_{m+1}\|_{L^p}\leqslant\|g_1\|_{L^p}+\sum_{k=1}^{m}\underbrace{\|g_{k+1}-g_k\|_{L^p}}_{\leqslant 2^{-k}}$$

$$\leqslant\|g_1\|_{L^p}+\sum_{k=1}^{\infty}2^{-k}=\|g_1\|_{L^p}+1.$$

Using Fatou's Lemma we deduce

$$\int_\Omega|h|^pd\mu\leqslant\liminf_{m\to\infty}\int_\Omega|g_{m+1}|^pd\mu<\infty,$$

so $h\in L^p$. Finally observe that

$$|h-g_{m+1}|\leqslant|g_1|+S_m\leqslant|g_1|+S_\infty$$

so

$$|h-g_{m+1}|^p\leqslant 2^{p-1}\big(|g_1|^p+S_\infty^p\big)\in L^1.$$

From the Dominated Convergence Theorem we deduce

$$\lim_{m\to\infty}\int_\Omega|h-g_{m+1}|^pd\mu=0$$

i.e., $g_m=f_{n_m}$ converges to h in L^p.

□

Let us record here a very useful byproduct of the above proof.

Corollary 19.4.5. *Let $p\in[1,\infty]$. Any convergent sequence in $L^p(\Omega,\mu)$ has a subsequence that converges μ-a.e..* □

Example 19.4.6. Suppose that $\Omega=\mathbb{N}$, $\mathcal{S}=2^{\mathbb{N}}$ and μ is the counting measure

$$\mu\big[\{n\}\big]=1.$$

The resulting Banach space $L^p(\mathbb{N},2^{\mathbb{N}},\mu)$ is usually denoted by ℓ_p. It consists of sequences of real numbers $(x_n)_{n\in\mathbb{N}}$ such that

$$\sum_{n\geqslant 1}|x_n|^p<\infty.$$

□

☞ For any Borel subset $B\subset\mathbb{R}^n$ and any $p\in[1,\infty]$ we denote by $L^p(B)$ the space $L^p(B,\boldsymbol{L}_B,\boldsymbol{\lambda})$, where $\boldsymbol{L}_S$ is the sigma-algebra of Lebesgue measurable subsets of B.

The Dominated Convergence Theorem has an L^p-version.

Theorem 19.4.7 (Dominated Convergence theorem: L^p-version)**.** *Let $p \in [1,\infty)$. Suppose that (f_n) is a sequence in $\mathscr{L}^p(\Omega,\mathcal{S},\mu)$ satisfying the following conditions.*

(i) *The sequence (f_n) converges μ-a.e. to a function $f \in \mathscr{L}^0(\Omega,\mu)$.*

(ii) *There exists a function $h \in \mathscr{L}^p(\Omega,\mu)$ such that $|f_n| \leqslant h$ μ-a.e..*

Then $f \in \mathscr{L}^p$ abd

$$\lim_{n\to\infty} \|f_n - f\|_{L^p} = 0.$$

Proof. We deduce from Fatou's lemma that

$$\int_\Omega |\, f\,|^p d\mu \leqslant \liminf_{n\to\infty} \int_\Omega |\, f_n\,|^p d\mu \leqslant \int_\Omega |\, h\,|^p d\mu < \infty.$$

hence $f \in \mathscr{L}^p$. On the other hand,

$$|f_n - f|^p \leqslant (|h| + |f|)^p \in L^1$$

and $|f_n - f|^p \to 0$ μ-a.e.. From the Dominated Convergence Theorem we deduce

$$\lim_{n\to\infty} \int_\Omega |f_n - f|^p d\mu = 0.$$

□

19.4.3. Density results. Given a measured space $(\Omega,\mathcal{S},\mu)$ we want to describe dense subsets in $L^p(\Omega,\mu)$. Given a subfamily $\mathcal{A} \subset \mathcal{S}$ we denote by $\mathbb{R}\big[\,\mathcal{A}\,\big]$ the vector subspace of $\mathscr{L}^0(\Omega,\mathcal{S})$ spanned by the functions $\boldsymbol{I}_A$, $A \in \mathcal{A}$. We set

$$\mathbb{R}\big[\,\mathcal{A}\,\big]_+ := \mathbb{R}\big[\,\mathcal{A}\,\big] \cap \mathscr{L}^0_+(\Omega,\mathcal{S}).$$

Note that $\mathbb{R}\big[\,\mathcal{S}\,\big] = \mathscr{E}(\Omega,\mathcal{S})$. We denote by $\mathbb{Q}\big[\,\mathcal{A}\,\big]$ the subset of $\mathbb{R}\big[\,\mathcal{A}\,\big]$ consisting of linear combinations with *rational* coefficients of functions $\boldsymbol{I}_A$, $A \in \mathcal{A}$.

Proposition 19.4.8. *Suppose $\mu\big[\,\Omega\,\big] < \infty$ and $\mathcal{A} \subset \mathcal{S}$ is a π-system that generated $\mathcal{S}$ as a sigma-algebra. Then for any $p \in [1,\infty)$, the vector space $\mathbb{R}\big[\,\mathcal{A}\,\big]$ is dense in the Banach space $\big(L^p(\Omega,\mu), \|-\|_{L^p}\big)$.*

Proof. Fix $p \in [1,\infty)$. Since $\mu\big[\,\Omega\,\big] < \infty$ we deduce that $\boldsymbol{I}_S \in L^p$, $\forall S \in \mathcal{S}$. Hence $\mathbb{R}\big[\,\mathcal{A}\,\big] \subset L^p$. We denote by $\mathcal{X}$ its closure in the Banach space L^p.

Lemma 19.4.9. *The vector space $\mathbb{R}\big[\,\mathcal{S}\,\big]$ is dense in L^p.*

Proof. Let $f \in \mathscr{L}^p(\Omega,\mu)$. Then $D_n[f_\pm] \in \mathbb{R}\big[\,\mathcal{S}\,\big]$ and we will show that

$$\big\|\, f_\pm - D_n[f_\pm]\,\big\|_{L^p} \to 0.$$

Let $g \in \mathscr{L}^p_+(\Omega,\mu)$. The sequence $D_n[g]$ is nondecreasing and converges everywhere to g. Hence

$$\int_\Omega D_n[g]^p d\mu \leqslant \int_\Omega g^p d\mu < \infty$$

so $D_n[g] \in L^p$. Since $0 \leqslant D_n[g] \leqslant g$ the desired conclusion follows from the L^p-version of the Dominated Convergence Theorem. □

To prove that $\mathcal{X} = L^p(\Omega, \mathcal{S}, \mu)$ it suffices to show that

$$\mathbb{R}\big[\,\mathcal{S}\,\big] \subset \mathcal{X} = \boldsymbol{cl}_{L^p}\big(\,\mathbb{R}\big[\,\mathcal{A}\,\big]\,\big),$$

or, equivalently $\boldsymbol{I}_S \in \mathcal{X}$, $\forall S \in \mathcal{S}$. Denote by $\mathscr{C}$ the collection of subsets $S \in \mathcal{S}$ such that $\boldsymbol{I}_S \in \mathcal{X}$.

Clearly $\mathcal{A} \subset \mathscr{C}$. In particular, $\varnothing, \Omega \in \mathscr{C}$. If $A, B \in \mathscr{C}$, $A \subset B$, then $\boldsymbol{I}_A, \boldsymbol{I}_B \in \mathcal{X}$ and, because $\mathcal{X}$ is a vector space,

$$\boldsymbol{I}_{B \setminus A} = \boldsymbol{I}_B - \boldsymbol{I}_A \in \mathcal{X}.$$

If $A_1 \subset A_2 \subset \cdots$ is an increasing sequence of sets in $\mathscr{C}$ and

$$A = \bigcup_n A_n,$$

then $(\boldsymbol{I}_{A_n})_{n \in \mathbb{N}}$ is an increasing sequence of nonnegative functions in $\mathcal{X}$ that converges everywhere to $\boldsymbol{I}_A$. We deduce as in the proof of Lemma 19.4.9 that

$$\lim_{n \to \infty} \|\boldsymbol{I}_{A_n} - \boldsymbol{I}_A\|_{L^p} = 0.$$

This implies that $\boldsymbol{I}_A \in \mathcal{X}$ since $\mathcal{X}$ is closed in $L^p(\Omega, \mathcal{S}, \mu)$.

This proves that $\mathscr{C}$ is a λ-system containing $\mathcal{A}$. The $\pi - \lambda$ theorem implies $\mathcal{S} \subset \mathscr{C}$. □

Theorem 19.4.10. *Let $p \in [1, \infty)$. Suppose that μ is a finite measure on the measurable space $(\Omega, \mathcal{S})$. If $\mathcal{S}$ is generated as a sigma-algebra by a* countable *π-system $\mathcal{A}$, then the Banach space $L^p(\Omega, \mathcal{S}, \mu)$ is separable. More precisely, the collection $\mathbb{Q}\big[\,\mathcal{A}\,\big]$ is dense in $L^p(\Omega, \mu)$.* □

Corollary 19.4.11. *Suppose that μ is a sigma-finite measure on the measurable space $(\Omega, \mathcal{S})$. If $\mathcal{S}$ is generated as a sigma-algebra by a countable family $\mathcal{F}$ of sets, then for any $p \in [1, \infty)$ the Banach space $L^p(\Omega, \mu)$ is separable.*

Proof. Observe first that the π-system $\mathcal{A}$ generated by the $\mathcal{F}$ is also countable since it consists of all the finite intersections

$$F_1 \cap \cdots \cap F_n, \;\; F_1, \dots, F_n \in \mathcal{F}, \;\; n \in \mathbb{N}.$$

Fix an increasing family of measurable sets $\Omega_1 \subset \Omega_2 \subset \cdots$ such that

$$\Omega = \bigcup_{n \in \mathbb{N}} \Omega_n, \;\; \mu\big[\,\Omega_n\,\big] < \infty, \;\; \forall n.$$

Set

$$\mathcal{A}_n = \Omega_n \cap \mathcal{A} = \big\{\,\Omega_n \cap A; \;\; A \in \mathcal{A}\,\big\}.$$

We claim that

$$\bigcup_n \mathbb{Q}\big[\,\mathcal{A}_n\,\big]$$

is dense in $\mathcal{L}^p(\Omega, \mathcal{S}, \mu)$.

Fix $f \in \mathcal{L}^p_+(\Omega, \mathcal{S}, \mu)$. Then $0 \leqslant f\boldsymbol{I}_{\Omega_n} \nearrow f$ pointwisely. We conclude as before that

$$\lim_{n\to\infty} \| f\boldsymbol{I}_{\Omega_n} - f\|_{L^p} = 0.$$

For any $\varepsilon > 0$ there exists n_ε such that

$$\| f\boldsymbol{I}_{\Omega_{n_\varepsilon}} - f\|_{L^p} < \frac{\varepsilon}{2}.$$

Theorem 19.4.10 implies that there exists $g_\varepsilon \in \mathbb{Q}\big[\,\mathcal{A}_{n_\varepsilon}\,\big]$ such that

$$\Big\| f\boldsymbol{I}_{\Omega_{n_\varepsilon}} - \boldsymbol{I}_{\Omega_{n_\varepsilon}} g_\varepsilon \Big\|_{L^p(\Omega)} < \frac{\varepsilon}{2}.$$

Hence for any $f \in \mathcal{L}^p_+(\Omega, \mathcal{S}, \mu)$ and any $\varepsilon > 0$, there exist $n_\varepsilon \in \mathbb{N}$ and $g_\varepsilon \in \mathbb{Q}\big[\,\mathcal{A}_{n_\varepsilon}\,\big]$ such that

$$\| f - g_\varepsilon\|_{L^p} < \varepsilon.$$

Using the canonical decomposition $f = f_+ - f_-$ we conclude that $L^p(\Omega, \mathcal{S}, \mu)$ is separable. □

Corollary 19.4.12. *Suppose (X, d) is a separable metric space and $\mathcal{B}_X$ is its Borel algebra. If $\mu : \mathcal{B}_X \to [0, \infty]$ is a sigma-finite measure, then the spaces $L^p(X, \mathcal{B}_X, \mu)$ are separable, $\forall p \in [1, \infty)$.*

Proof. Fix a countable dense subset $Y \subset X$. The Borel algebra $\mathcal{B}_X$ is generated by the countable collection of open balls

$$B_r(y), \quad y \in Y, \quad r \in \mathbb{Q}.$$

The conclusion now follows from Corollary 19.4.11. □

Example 19.4.13. (a) The Banach spaces ℓ_p, $p \in [1, \infty)$ discussed in Example 19.4.6 are separable. Indeed, take $X = \mathbb{N}$ in the above corollary.

(b) Let $U \subset \mathbb{R}^n$ be an open set. Then $L^p(U, \mathcal{B}_U, \boldsymbol{\lambda}_n)$ is separable for any $p \in [1, \infty)$. □

Theorem 19.4.14. *Suppose that (K, d) is a compact metric space and μ is a Borel measure on K such that $\mu\big[\,K\,\big] < \infty$, then $C(K)$ is a dense subspace of $L^p(K, \mu)$, $\forall p \in [1, \infty)$. In particular, if μ_0, μ_1 are two finite Borel measures on K such that*

$$\int_K f(x)\mu_0\big[\,dx\,\big] = \int_K f(x)\mu_1\big[\,dx\,\big], \quad \forall f \in C(K) \tag{19.4.4}$$

then $\mu_0 = \mu_1$.

Proof. Fix $p \in [1, \infty)$ Note that any continuous function $f : K \to \mathbb{R}$ is p-integrable because it is bounded and the bounded measurable functions on K are p- integrable.

The collection $\mathscr{C}$ of closed subsets of K is a π-system that generate the Borel algebra $\mathcal{B}_K$ of K and, according to Proposition 19.4.8, the vector space $\mathbb{R}\big[\,\mathscr{C}\,\big]$ is dense in $L^p(K, \mu)$.

Thus it suffices to show that for any closed subset $C \subset K$ there exists a sequence of continuous functions $f_n : K \to \mathbb{R}$ such that

$$\lim_{n\to\infty} \|f_n - \boldsymbol{I}_C\|_{L^p} = 0.$$

Fix a closed set $C \subset K$. For each $\varepsilon > 0$ define

$$\mathcal{O}_\varepsilon := \big\{\, x \in K;\ \ \mathrm{dist}(x, C) < \varepsilon \,\big\},\ \ E_\varepsilon := K \backslash \mathcal{O}_\varepsilon = \big\{\, x \in K;\ \ \mathrm{dist}(x, C) \geqslant \varepsilon \,\big\}$$

The set E_ε is closed since, according to Proposition 17.1.27, the function $x \mapsto \mathrm{dist}(x, C)$ is continuous. Define

$$f_\varepsilon : K \to [0, \infty),\ \ f_\varepsilon(x) = \frac{\mathrm{dist}(x, E_\varepsilon)}{\mathrm{dist}(x, C) + \mathrm{dist}(x, E_\varepsilon)}.$$

Proposition 17.1.27 also shows that this function is well defined since $\mathrm{dist}(x, C)$ and $\mathrm{dist}(x, E_\varepsilon)$ cannot be simultaneously zero. Clearly $0 \leqslant f_\varepsilon(x) \leqslant 1$ for any $x \in K$ and $f_\varepsilon(x) = 0$, $\forall x \in E_\varepsilon$. Hence

$$\boldsymbol{I}_C(x) \leqslant f_\varepsilon(x) \leqslant \boldsymbol{I}_{\mathcal{O}_\varepsilon}(x),\ \ \forall x \in K.$$

We deduce

$$0 \leqslant f_\varepsilon - \boldsymbol{I}_C \leqslant \boldsymbol{I}_{\mathcal{O}_\varepsilon} - \boldsymbol{I}_C = \boldsymbol{I}_{\mathcal{O}_\varepsilon \backslash C},\ \ \int_K \big|\, f_\varepsilon - \boldsymbol{I}_C \,\big|^p d\mu \leqslant \int_K \boldsymbol{I}^p_{\mathcal{O}_\varepsilon \backslash C} d\mu = \mu\big[\, \mathcal{O}_\varepsilon \backslash C \,\big].$$

Now observe that

$$\bigcap_{n\in\mathbb{N}} (\mathcal{O}_{1/n} \backslash C) = \varnothing$$

so

$$\lim_{n\to\infty} \mu\big[\, \mathcal{O}_{1/n} \backslash C \,\big] = 0.$$

We deduce

$$\lim_{n\to\infty} \int_K \big|\, f_{1/n} - \boldsymbol{I}_C \,\big|^p d\mu = 0.$$

In particular

$$\lim_{n\to\infty} \int_K f_{1/n}(x) \mu\big[\, dx \,\big] = \mu\big[\, C \,\big].$$

The last equality shows that if μ_0, μ_1 are two finite Borel measures satisfying (19.4.4), then $\mu_0\big[\, C \,\big] = \mu_1\big[\, C \,\big]$, $\forall C \in \mathscr{C}$. Proposition 19.1.34 implies that $\mu_0 = \mu_1$. □

Corollary 19.4.15. *Suppose that (K, d) is a compact metric space and μ is a Borel measure on K such that $\mu\big[\, K \,\big] < \infty$. If $\mathcal{S} \subset C(K)$ is dense in $C(K)$ $\underline{\text{with respect to the sup-norm}}$, then $\mathcal{S}$ is dense in $L^p(K, \mu)$, $\underline{\text{with respect to the } L^p\text{-norm}}$, $\forall p \in [1, \infty)$.*

Proof. Let $f \in L^p(K, \mu)$. There exists a sequence $f_n \in C(K)$ such that

$$\lim_{n\to\infty} \|f_n - f\|_{L^p} = 0.$$

For any $n \in \mathbb{N}$ there exists $s_n \in \mathcal{S}$ such that $\|s_n - f_n\|_\infty < \frac{1}{n}$. Hence

$$\|s_n - f_n\|_{L^p} = \left(\int_K |s_n(x) - f_n(x)|^p \mu\big[\, dx \,\big]\right)^{1/p} \leqslant \left(\int_K \frac{1}{n^p} \mu\big[\, dx \,\big]\right)^{1/p} = \frac{\mu\big[\, K \,\big]^{1/p}}{n}.$$

We conclude that

$$\|s_n - f\|_{L^p} \leqslant \|s_n - f_n\|_{L^p} + \|f_n - f\|_{L^p} \leqslant \frac{\mu\big[\, K \,\big]^{1/p}}{n} + \|f_n - f\|_{L^p} \to 0.$$

□

Corollary 19.4.16 (Lusin). *Fix $p \in [1, \infty)$ and suppose that (K, d) is a compact metric space and μ is a Borel measure on K such that $\mu\big[\, K \,\big] < \infty$. Then for any $f \in L^p(K, \mu)$ and any $\varepsilon > 0$ there exists a Borel subset $B \subset K$ such that $\mu\big[\, K \backslash B \,\big] < \varepsilon$ and the restriction of f to B is continuous as a function $B \to \mathbb{R}$.*[2]

Proof. Fix a sequence of continuous functions $f_n : K \to \mathbb{R}$ that converges in L^p to f. We deduce from Corollary 19.4.5 that a subsequence (f_{n_k}) of (f_n) converges a.e. to f. Egorov's Theorem 19.1.39 shows that for any $\varepsilon > 0$ there exists a Borel subset $B \subset K$ such that $\mu\big[\, K \backslash B \,\big] < \varepsilon$ and the sequence (f_{n_k}) converges uniformly to f on B. This proves that $f\big|_B$ is continuous as uniform limit of continuous functions. □

Remark 19.4.17. We have seen in Example 17.2.3 that the space $C([0, 1])$ equipped with the L^1-norm is not complete. On the other hand the space $C([0, 1])$ is dense with respect to the L^1-norm in the space $L^1([0, 1])$. Hence $L^1([0, 1])$ is the completion (in the sense of Definition 17.2.16) of the space $C([0, 1])$ equipped with the L^1-norm. □

19.5. Signed measures

A *signed measure* on a measurable space $(\Omega, \mathcal{S})$ is a map

$$\mu : \mathcal{S} \to [-\infty, \infty]$$

with the following properties.

(i) $\mu\big[\, \varnothing \,\big] = 0$.

(ii) The range of μ is either contained in $(-\infty, \infty]$ or in $[-\infty, \infty)$.

(iii) It is countably additive i.e., if $\big(\, S_n \,\big)_{n \in \mathbb{N}}$ is a sequence of pairwise disjoint measurable sets, then

$$\mu\Big[\, \bigcup_{n \in \mathbb{N}} \,\Big] = \sum_{n \in \mathbb{N}} \mu\big[\, S_n \,\big].$$

Remark 19.5.1. Let us point out[3] that the the series in the right-hand side of (iii) is automatically *absolutely convergent*, even though we did not explicitly require this. Indeed,

[2] This does not eliminate the possibility that f, as a function $K \to \mathbb{R}$, is discontinuous at some points in B.

[3] Hat tip to *Zach Joseph.*

the series in (iii) *unconditionally convergent*, i.e., its sum does not change if we change the order of summation of its terms. This happens iff the series is absolutely convergent; see e.g. [**28**, Chap.IV, Sec.16, Thm.2]. □

☞ *It the sequel, to simplify the exposition, we will assume that the signed measures are valued in* $(-\infty, \infty]$

$$\mu : \mathcal{S} \to (-\infty, \infty].$$

As in the case of usual, i.e., positive measures, the countable additivity condition is equivalent with upwards continuity. More precisely, if $(S_n)_{n\in\mathbb{N}}$ is a nondecreasing family of subsets and

$$S_\infty := \bigcup_n S_n,$$

then

$$\mu\big[\, S_\infty \,\big] = \lim_{n\to\infty} \mu\big[\, S_n \,\big].$$

Observe that

$$\mu\big[\, \Omega \,\big] < \infty \Rightarrow \mu\big[\, S \,\big] < \infty, \;\; \forall S \in \mathcal{S}.$$

Indeed

$$\mu\big[\, \Omega \,\big] = \mu\big[\, S \,\big] + \mu\big[\, \Omega \backslash S \,\big], \;\; \mu\big[\, \Omega \backslash S \,\big] > -\infty.$$

Example 19.5.2. (a) If μ is a (positive) measure on $(\Omega, \mathcal{S})$ and $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$,

$$\mu_f : \mathcal{S} \to \mathbb{R}, \;\; \mu_f\big[\, S \,\big] := \int_S f d\mu = \int_\Omega f \boldsymbol{I}_S d\mu$$

is a signed measure. Sometimes we will write

$$\mu_f\big[\, d\omega \,\big] = f(\omega)\mu\big[\, d\omega \,\big].$$

(b) If μ_0, μ_1 are two (positive) measures on $(\Omega, \mathcal{S})$ such that μ_1 is finite, then their difference $\mu_0 - \mu_1$ is a signed measure. It turns out that all signed measures are obtained in this fashion. The next subsections will clarify this fact. □

19.5.1. The Hahn and Jordan decompositions. Suppose that μ is a signed measure on the measurable space $(\Omega, \mathcal{S})$.

Definition 19.5.3. A set $S \in \mathcal{S}$ is called (μ-)*negative* (resp. *positive*) if $\mu\big[\, A \,\big] \leqslant 0$ (resp. $\mu\big[\, A \,\big] \geqslant 0$) for any measurable subset $A \subset S$. We will use the notation $A <_\mu 0$ (resp. $0 <_\mu A$) to indicate that A is μ-negative (resp. μ-positive).

The measurable set $S \in \mathcal{S}$ is called a (μ-)*null set* if

$$\mu\big[\, A \,\big] = 0, \;\; \forall A \subset S, \;\; A \in \mathcal{S}.$$

□

Recall that the symmetric difference of two sets A, B is the set $A\Delta B$ defined by

$$A\Delta B = (A \cup B)\backslash(A \cap B) = (A\backslash B) \cup (B\backslash A).$$

Observe that any subset of a positive/negative set is also positive. The union of two positive/negative sets is postive/negative. Indeed if say A, B are negative and $S \subset A \cup B$, then

$$\mu\big[\,S\,\big] = \mu\big[\,S \cap A\,\big] + \mu\big[\,S \cap (B\backslash A)\,\big] \leqslant 0.$$

More generally, a countable union of negative sets $(A_n)_{n\geqslant 1}$ is also a negative set.

Indeed, let

$$S \subset \bigcup_n A_n.$$

Note that

$$\widehat{A}_n := \bigcup_{k=1}^n A_k$$

is a negative set so $S_n = S \cap \widehat{A}_n$ is negative. Then

$$\mu\big[\,S\,\big] = \lim_{n\to\infty} \mu\big[\,S_n\,\big] \leqslant 0.$$

Theorem 19.5.4 (Hahn decomposition)**.** *Suppose that μ is a signed measure on the measurable space $(\Omega, \mathcal{S})$. Then the space Ω admits a* Hahn decomposition*, i.e., a pair (P, N) of disjoint measurable subsets of Ω, where P is a positive set, N is a negative set and $\Omega = P \cup N$. Moreover, if $P' \sqcup N'$ is another Hahn decomposition, then $P\Delta P' = N\Delta N'$ is a null set.*

Proof. We denote by m the infimum of $\mu\big[\,A\,\big]$, A negative set,

$$m := \inf\big\{\,\mu\big[\,A\,\big];\;\; A <_\mu 0\,\big\}.$$

A priori, this infimum could be $-\infty$. Choose a sequence of negative sets $(A_k)_{k\geqslant 1}$ such that

$$m = \lim_{k\to\infty} \mu\big[\,A_k\,\big].$$

Set

$$N_k := \bigcup_{j=1}^k A_j.$$

Observe that since $A_k \subset N_k$ and N_k is negative we have

$$m \leqslant \mu\big[\,N_k\,\big] = \mu\big[\,N_k\backslash A_k\,\big] + \mu\big[\,A_k\,\big] \leqslant \mu\big[\,A_k\,\big]. \tag{19.5.1}$$

We set

$$N := \bigcup_k N_k.$$

Letting $k \to \infty$ in (19.5.1) we deduce

$$-\infty < \mu\big[\,N\,\big] = m \leqslant 0.$$

Thus

$$N <_\mu 0 \;\text{ and }\; \forall A <_\mu 0: \;\; \mu\big[\, N\,\big] \leqslant \mu\big[\, A\,\big]. \tag{19.5.2}$$

We claim that $P := \Omega\backslash N$ is a positive set. We argue by contradiction.

Suppose that $S_0 \subset P$ is a measurable set such that $\mu\big[\, S_0\,\big] < 0$. The set S_0 cannot be negative because then $N \cup S_0$ would be negative and $\mu\big[\, N \cup S_0\,\big] < \mu\big[\, N\,\big]$, contradicting (19.5.2). Hence S_0 contains subsets B such that $\mu\big[\, B\,\big] > 0$. Set

$$\mu_1 := \sup\big\{\mu\big[\, B\,\big];\;\; B \subset S_0,\;\; \mu\big[\, B\,\big] > 0\,\big\} = \sup\big\{\,\mu\big[\, B\,\big];\;\; B \subset S_0\,\big\}.$$

Let n_1 be the natural number such that

$$\frac{1}{n_1} < \mu_1 \leqslant \frac{1}{n_1 - 1},$$

where we set $\frac{1}{0} := \infty$. We deduce that there exists a measurable subset $B_1 \subset S_0$ with

$$\frac{1}{n_1} < \mu\big[\, B_1\,\big] \leqslant \mu_1 \leqslant \frac{1}{n_1 - 1}.$$

Set $S_1 := S_0\backslash B_1$. Then

$$\mu\big[\, S_1\,\big] = \mu\big[\, S_0\,\big] - \mu\big[\, B_1\,\big] < \mu\big[\, S_0\,\big] < 0.$$

Again the subset $S_1 \subset S_0$ cannot be negative so that there exist measurable subsets $B \subset S_1$ such that $\mu\big[\, B\,\big] > 0$. Set

$$\mu_2 = \sup\big\{\mu\big[\, B\,\big];\;\; B \subset S_1\,\big\} \leqslant \sup\big\{\mu\big[\, B\,\big];\;\; B \subset S_0\,\big\} = \mu_1.$$

We deduce that there exists a measurable subset $B_2 \subset S_1$ and a natural number $n_2 \geqslant n_1$ such that

$$\frac{1}{n_2} < \mu\big[\, B_2\,\big] \leqslant \mu_2 \leqslant \frac{1}{n_2 - 1}.$$

Iterating this procedure we obtain

- a decreasing sequence of measurable sets $(S_k)_{k\geqslant 0}$, and
- a nondecreasing sequence of natural numbers $(n_k)_{k\geqslant 1}$, such that

$$\mu\big[\, S_k\,\big] < 0,\;\; \forall k \geqslant 0,$$
$$\frac{1}{n_k} < \mu_k := \sup\big\{\,\mu\big[\, B\,\big];\;\; B \subset S_{k-1}\,\big\} \leqslant \frac{1}{n_k - 1},\;\; \forall k \geqslant 1,$$
$$\frac{1}{n_k} < \mu\big[\, \underbrace{S_{k-1}\backslash S_k}_{=:B_k}\,\big] \leqslant \mu_k \leqslant \frac{1}{n_k - 1},\;\; \forall k \geqslant 1.$$

The sets B_k are disjoint and, if B_∞ denotes their union, we deduce

$$\mu\big[\, B_\infty\,\big] = \sum_{k\geqslant 1} \mu\big[\, B_k\,\big] > \sum_{k\geqslant 1} \frac{1}{n_k} > 0.$$

Observe that $B_\infty \subset S_0$ so

$$\mu\big[\, B_\infty\,\big] = \mu\big[\, S_0\,\big] - \mu\big[\, S_0\backslash B_\infty\,\big] < \infty$$

Thus the series $\sum_k \frac{1}{n_k}$ is convergent and therefore

$$\lim_{k\to\infty} \frac{1}{n_k} = 0.$$

Now observe that

$$S_\infty := \bigcap_{k\geqslant 0} S_k = S_0\backslash B_\infty.$$

We have

$$\mu\big[\, S_\infty \,\big] = \mu\big[\, S_0 \,\big] - \mu\big[\, B_\infty \,\big] < 0.$$

The set S_∞ cannot be negative because if it were we would have

$$\mu\big[\, N \cup S_\infty \,\big] < \mu\big[\, N \,\big] = m.$$

Thus S_∞ must contain at least one measurable subset P such that $\mu\big[\, P \,\big] > 0$. Set

$$\mu_\infty := \sup\big\{\, \mu\big[\, P \,\big];\;\; P \subset S_\infty \,\big\},$$

so that $\mu_\infty > 0$. On the other hand $S_\infty \subset S_k$, $\forall k \geqslant 1$ so that

$$0 < m_\infty \leqslant \sup\big\{\, \mu\big[\, B \,\big];\;\; B \subset S_{k-1} \,\big\} = \mu_k \leqslant \frac{1}{n_k - 1},\;\; \forall k \geqslant 1.$$

Letting $k \to \infty$ we deduce

$$0 < \mu_\infty \leqslant \lim_{k\to\infty} \frac{1}{n_k - 1} = 0.$$

We have reached a contradiction. This proves that $(\Omega\backslash N, N)$ is a Hahn decomposition.

Suppose that (P', N') is another Hahn decomposition observe that obviously $P\backslash P' \subset P$ so $P\backslash P'$ is a positive set. On the other hand,

$$P = (P \cap P') \cup (P \cap N')$$

and since P' N' are disjoint we have

$$P \cap N' = P\backslash(P \cap P') = P\backslash P'$$

Hence $P\backslash P' \subset N'$ so that $P\backslash P'$ is also a negative. Hence $P\backslash P'$ is a null set. Arguing in a similar fashion we deduce that $P'\backslash P$ is also a null set. □

The Hahn decomposition(s) of Ω induced by a signed measure μ leads to a canonical description of μ as a difference of two measures.

Definition 19.5.5. Two (positive) measures μ_0, μ_1 on $(\Omega, \mathcal{S})$ are *mutually singular*, and we denote this $\mu_0 \perp \mu_1$, if there exist disjoint nonempty sets $S_0, S_1 \in \mathcal{S}$ such that

- $\Omega = S_0 \cup S_1$ and
- $\mu_0\big[\, S_1 \,\big] = 0$, $\mu_1\big[\, S_0 \,\big] = 0$.

□

Intuitively, the measure μ_0 lives on S_0 while the measure μ_1 lives on S_1. From the "point of view" of μ_1 the set S_0 is negligible, but it is non-negligible from the "point of view" of μ_0.

Example 19.5.6. (a) Suppose that (Ω_0, Ω_1) is a nontrivial measurable partition of $(\Omega, \mathcal{S})$. Denote by $\mathcal{S}_k$ the sigma-algebra inducted by $\mathcal{S}$ on Ω_k, $k = 0, 1$. Let μ_k be a measure on $(\Omega_k, \mathcal{S}_k)$, and denote by $\widehat{\mu}_k$ its extension to $(\Omega, \mathcal{S})$,

$$\widehat{\mu}_k\big[\,S\,\big] = \mu_k\big[\,S \cap \Omega_k\,\big].$$

Then $\widehat{\mu}_0 \perp \widehat{\mu}_1$.

(b) Let δ_0 denote the Dirac measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ concentrated at 0. Then $\delta_0 \perp \boldsymbol{\lambda}$. To see this it suffices to take $S_0 = \{0\}$ and $S_1 = \mathbb{R}\backslash\{0\}$. □

Theorem 19.5.7 (Jordan decomposition)**.** *Let μ be a signed measure on the measurable space $(\Omega, \mathcal{S})$. Then μ admits a unique* Jordan decomposition*, i.e., there exists a unique pair of (positive) measures $\mu_\pm$ on $(\Omega, \mathcal{S})$ satisfying the following conditions.*

(i) $\mu_+ \perp \mu_-$.

(ii) $\mu_-\big[\,\Omega\,\big] < \infty$.

(iii) $\mu = \mu_+ - \mu_-$.

Proof. Fix a Hahn decomposition (P, N) of Ω. For any $S \in \mathcal{S}$ we define

$$\mu_+\big[\,S\,\big] = \mu\big[\,S \cap P\,\big],\;\; \mu_-\big[\,S\,\big] = -\mu\big[\,S \cap N\,\big].$$

Clearly μ_+, μ_- satisfy all these conditions (i)-(iii). Suppose that (ν_+, ν_-) is another pair of measures satisfying these conditions. Choose disjoint sets $S_\pm$ such that $\Omega = S_+ \cup S_-$ and $\nu_\pm\big[\,S_\mp\,\big] = 0$. We define

$$P_\pm := P \cap S_\pm,\;\; N_\pm := N \cap S_\pm.$$

We have

$$\nu_+\big[\,N\,\big] = \nu_+\big[\,N_+\,\big],\;\; 0 \leqslant \nu_-\big[\,N_+\,\big] \leqslant \nu_-\big[\,S_+\,\big] = 0.$$

Hence

$$0 \geqslant \mu\big[\,N_+\,\big] = \nu_+\big[\,N_+\,\big] - \nu_-\big[\,N_+\,\big] \geqslant 0.$$

Hence $\nu_+\big[\,N\,\big] = \nu_+\big[\,N_+\,\big] = 0$. Arguing in a similar fashion we deduce $\nu_-\big[\,P\,\big] = 0$.

Suppose that $A \in \mathcal{S}$. We have

$$\nu_+\big[\,A\,\big] = \nu_+\big[\,A \cap P\,\big] + \nu_+\big[\,A \cap N\,\big] = \nu_+\big[\,A \cap P\,\big],\;\; \nu_-\big[\,A \cap P\,\big] = 0$$

Hence

$$\nu_+\big[\,A\,\big] = \nu_+\big[\,A \cap P\,\big] - \nu_-\big[\,A \cap P\,\big] = \mu\big[\,A \cap P\,\big] = \mu_+\big[\,A\,\big].$$

The equality $\nu_-\big[\,A\,\big] = \mu_-\big[\,A\,\big]$ is proved similarly. □

Definition 19.5.8. Let μ be a signed measure on the measurable space $(\Omega, \mathcal{S})$. If $\mu = \mu_+ - \mu_-$ is its Jordan decomposition, then the measure $\mu_+ + \mu_-$ is called the *total variation* of μ and it is denoted by $|\mu|$. □

19.5.2. The Radon-Nikodym theorem.

Definition 19.5.9. A signed measure $\nu : \mathcal{S} \to (-\infty, \infty]$ is said to be *absolutely continuous* with respect to a *positive* measure $\mu : \mathcal{S} \to [0, \infty]$, and we indicate this with the notation $\nu \ll \mu$, if

$$\forall S \in \mathcal{S},\;\; \mu\big[\,S\,\big] = 0 \Rightarrow \nu\big[\,S\,\big] = 0. \tag{19.5.3}$$

□

The proof of the next result is left to you as an exercise.

Proposition 19.5.10. *Let $(\Omega, \mathcal{S})$ be a measurable space, $\mu : \mathcal{S} \to [0, \infty]$ a measure and $\nu : \mathcal{S} \to (-\infty, \infty]$ a signed measure. If $\nu = \nu_+ - \nu_-$ is the Jordan decomposition of ν, then the following statements are equivalent.*

(i) $\nu \ll \mu$.

(ii) $\nu_\pm \ll \mu$.

(iii) $|\nu| \ll \mu$.

□

Remark 19.5.11. Suppose that ν, μ are (positive measures) such that $\nu \ll \mu$ and $\nu \neq 0$, i.e., $\nu\big[\,\Omega\,\big] \neq 0$. Then μ and ν are not mutually singular. Indeed if $\Omega = S \cup S'$, $\nu\big|\,S\,\big] == \mu\big[\,S'\,\big]$, then $\nu\big[\,S'\,\big] = 0$ as well, so $\nu\big[\,\Omega\,\big] = 0$. One can say absolute continuity is the polar opposite of mutual singularity. □

The next result justifies the term "continuity" in "absolute continuity".

Proposition 19.5.12. *Suppose that $(\Omega, \mathcal{S})$ is a measurable space, μ is a positive measure, and ν is a $\underline{\text{finite}}$ signed measure on $\mathcal{S}$. Then the following statements are equivalent.*

(i) $\nu \ll \mu$.

(ii) *For any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that*

$$\forall S \in \mathcal{S},\;\; \mu\big[\,S\,\big] < \delta \Rightarrow \big|\,\nu\big[\,S\,\big]\,\big| < \varepsilon.$$

Proof. The implication (ii) ⇒ (i) is obvious.

(i) ⇒ (ii) Proposition 19.5.10 shows that both $\nu_\pm$ satisfy (i). If we show that they both satisfy (ii) then so will ν. The upshot is that it suffices to prove this implication only in the special case when ν is a finite (positive) measure. We assume this and we argue by contradiction.

Suppose that there exists $\varepsilon_0 > 0$ such that, for any $n \in \mathbb{N}$ there exists $S_n \in \mathcal{S}_n$ with the property

$$\mu\big[\,S_n\,\big] < 2^{-n} \;\text{ and }\; \nu\big[\,S_n\,\big] \geqslant \varepsilon_0.$$

For $k \in \mathbb{N}$ we set

$$B_k := \bigcup_{n \geqslant k} S_n.$$

Observe that $B_1 \supset B_2 \supset \cdots$ and

$$\mu\big[\, B_k \,\big] \leqslant \sum_{n \geqslant k} \mu\big[\, S_n \,\big] < \sum_{n > k} 2^{-n} = 2^{-k+1}.$$

Since $B_k \supset S_k$ we deduce $\nu\big[\, B_k \,\big] \geqslant \nu\big[\, S_k \,\big] \geqslant \varepsilon_0$. We set

$$B_\infty = \bigcap_{k \geqslant 1} B_k.$$

Clearly

$$\mu\big[\, B_\infty \,\big] < \mu\big[\, B_k \,\big] < 2^{-k+1}, \;\; \forall k$$

so that $\mu\big[\, B_\infty \,\big] = 0$. On the other hand, ν *is a finite measure* and we deduce from (19.1.9) that

$$\nu\big[\, B_\infty \,\big] = \lim_{k \to \infty} \nu\big[\, B_k \,\big] \geqslant \varepsilon_0.$$

This contradicts (i). □

Proposition 19.5.13. *Suppose that $f \in \mathcal{L}^0(\Omega, \mathcal{S}, \mu)$ is a measurable function such that $f_- \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$. Then the function*

$$\mu_f : \mathcal{S} \to (-\infty, \infty], \;\; S \mapsto \mu_f\big[\, S \,\big] = \mu\big[\, f \boldsymbol{I}_S \,\big]$$

is a signed measure on $\mathcal{S}$ absolutely continuous with respect to μ, $\mu_f \ll \mu$.

Proof. As usual we write $f = f_+ - f_-$ so that $\mu_f = \mu_{f_+} - \mu_{f_-}$. We know from Corollary 19.3.19 that $\mu_{f_\pm}$ are measures so that μ_f is a signed measure. Property (19.5.3) follows from Corollary 19.3.16. □

It turns out that the above example is the most general example of measure absolutely continuous with respect to μ.

Theorem 19.5.14 (Radon-Nikodym). *Suppose that $(\Omega, \mathcal{S})$ is a measurable space and μ is a sigma-finite measure on $\mathcal{S}$ and $\nu : \mathcal{S} \to (-\infty, \infty]$ is a signed measure measure such that $|\nu|$ is sigma-finite measure. If $\nu \ll \mu$, then there exists $f \in \mathcal{L}^0(\Omega, \mathcal{S})$ such that $f_- \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ and $\nu_\pm = \mu_{f_\pm}$. Moreover if $g \in \mathcal{L}^0(\Omega, \mathcal{S})$ is another function with the above properties, then $f = g$ a.e.*

Proof. We follow the approach in [**17**, Sec. 3.2]. In view of Proposition 19.5.10 it suffices the prove the result only when ν is a (positive) measure. We carry the proof in two steps.

A. Assume first that both μ and ν are *finite* measures. The goal is clear. We have to find a nonnegative function $f \in \mathcal{L}^1\big(\, \Omega, \mathcal{S}, \mu \,\big)$ such that

$$\int_S f d\mu = \nu\big[\, S \,\big], \;\; \forall S \in \mathcal{S}.$$

Define

$$\mathcal{F} := \left\{ f \in \mathcal{L}^0_+(\Omega, \mathcal{S});\ \ \int_S f\, d\mu \leqslant \nu\big[\, S\,\big],\ \ \forall S \in \mathcal{S} \right\}.$$

Note that $0 \in \mathcal{F}$, so $\mathcal{F} \neq \varnothing$. Note also that

$$f, g \in \mathcal{F} \Rightarrow \max(f, g) \in \mathcal{F}.$$

To see this define $A := \{f > g\} \in \mathcal{S}$. Then, for any $S \in \mathcal{S}$ we have

$$\int_S \max(f, g)\, d\mu = \int_{S \cap A} f\, d\mu + \int_{S \setminus A} g\, d\mu \leqslant \nu\big[\, S \cap A\,\big] + \nu\big[\, S \setminus A\,\big] = \nu\big[\, S\,\big].$$

Set

$$M := \sup_{f \in \mathcal{F}} \int_\Omega f\, d\mu \leqslant \nu\big[\, \Omega\,\big] < \infty.$$

Now choose a sequence (f_n) in $\mathcal{F}$ such that

$$\lim_{n \to \infty} \int_\Omega f_n\, d\mu = M.$$

Set

$$g_n := \max(f_1, \dots, f_n).$$

Observe that

$$g_1 \leqslant g_2 \leqslant g_2 \leqslant \cdots,\ \ f_n \leqslant g_n,\ \ \forall n \in \mathbb{N}.$$

Hence

$$\int_\Omega f_n\, d\mu \leqslant \int_\Omega g_n\, d\mu \leqslant M.$$

Hence

$$\lim_{n \to \infty} \int_\Omega g_n\, d\mu = M.$$

We set

$$g := \lim_{n \to \infty} g_n.$$

The above limit exists since the sequence (g_n) is nondecreasing. The Monotone Convergence Theorem implies

$$\int_\Omega g\, d\mu = \lim_{n \to \infty} \int_\Omega g_n\, d\mu = M.$$

From Markov's inequality we deduce that $g < \infty$ a.e., so after modifying g on the μ-negligible set $Z = \{g = \infty\}$ we can assume that $g < \infty$ everywhere.

From the equalities

$$\int_S g_n\, d\mu \leqslant \nu\big[\, S\,\big],\ \ \forall S \in \mathcal{S},\ \ \forall n \in \mathbb{N}, \tag{19.5.4}$$

and the Monotone Convergence Theorem we deduce

$$\int_S g\, d\mu \leqslant \nu\big[\, S\,\big],\ \ \forall S \in \mathcal{S},$$

i.e., $g \in \mathcal{F}$. We claim that for any $f \in \mathcal{F}$ we have $f \leqslant g$ a.e.. Indeed, if this were not the case, then $\max(f,g) \in \mathcal{F}$, $\max(f,g) - g \geqslant 0$ and

$$M = \int_\Omega g\,d\mu \geqslant \int_\Omega \max(f,g)d\mu \leqslant M.$$

Corollary 19.3.16 implies that $g = \max(f,g)$. From (19.5.4) we deduce that the a priori signed measure $\mu' = \nu - g\mu$,

$$\mu'\big[\,S\,\big] = \nu\big[\,S\,\big] - \int_S g d\mu, \ \ \forall S \in \mathcal{S}$$

is positive measure. We will show that $\mu' = 0$, i.e.,

$$\int_S g\,d\mu = \nu\big[\,S\,\big], \ \ \forall S \in \mathcal{S}. \tag{19.5.5}$$

To achieve this we rely on the following clever trick.

Lemma 19.5.15. *Suppose that μ, μ' are two finite measures on $(\Omega, \mathcal{S})$. Then,*

- *either $\mu \perp \mu'$,*
- *or there exists $\varepsilon > 0$ and a measurable set $E \in \mathcal{S}$ such that $\mu\big[\,E\,\big] > 0$ and E is a positive set for the signed measure $\mu' - \varepsilon\mu$.*

Proof. Suppose that μ' and μ are not mutually singular. In particular, $\mu, \mu' \neq 0$. For each $k \in \mathbb{N}$ fix a Hahn decomposition (P_k, N_k) of the signed measure $\mu_k = \mu' - \frac{1}{k}\mu$. Define

$$P = \bigcup_{k \geqslant 1} P_k, \ \ N = \Omega \setminus P = \bigcap_{k \geqslant 1} N_k.$$

Observe that N is a negative set for $\mu' - \frac{1}{k}\mu$ for any k so that

$$\mu'\big[\,N\,\big] \leqslant \frac{1}{k}\mu\big[\,N\,\big], \ \ \forall k.$$

Hence $\mu'\big[\,N\,\big] = 0$. Since μ, μ' are *not mutually singular* we deduce $\mu\big[\,P\,\big] > 0$. On the other hand,

$$\mu\big[\,P\,\big] \leqslant \sum_k \mu\big[\,P_k\,\big],$$

so $\mu\big[\,P_k\,\big] > 0$ for some k and P_k is a positive set for $\mu' - \frac{1}{k}\mu$.

□

Consider the (positive) measure $\mu' = \nu - g\mu$. Note that μ' and μ are not mutually singular because $\mu\big[\,S\,\big] = 0 \Rightarrow \mu'\big[\,S\,\big] = 0$. Lemma 19.5.15 implies that there exists $E \in \mathcal{S}$ and $\varepsilon > 0$ such that

$$\mu\big[\,E\,\big] > 0 \ \ \mu'\big[\,S\,\big] - \varepsilon\mu\big[\,S\,\big] \geqslant 0, \ \ \forall S \subset E, \ \ S \in \mathcal{S},$$

i.e.,

$$\nu\big[\,S\,\big] \geqslant \int_S (g + \varepsilon)d\mu \ \ \forall S \subset E, \ \ S \in \mathcal{S}.$$

This proves that $g + \varepsilon \boldsymbol{I}_E \in \mathcal{F}$ and $g + \varepsilon \boldsymbol{I}_E > g$ on a set of positive μ-measure. This contradicts the maximality of g. This establishes the existence part.

Suppose now that $f \in \mathcal{L}^0_+(\Omega, \mathcal{S})$ is another function such that

$$\int_S f d\mu = \int_S g d\mu = \nu\big[\, S\,\big], \;\; \forall S \in \mathcal{S}.$$

Set $h := f - g$ so

$$\int_S h d\mu = 0, \;\; \forall S \in \mathcal{S}.$$

We will show that $\mu\big[\,\{h > 0\}\,\big] = \mu\big[\,\{h < 0\}\,\big] = 0$. We argue by contradiction. Suppose that $\mu\big[\,\{h > 0\}\,\big] > 0$. Since

$$\mu\big[\,\{h > 0\}\,\big] = \lim_{n\to\infty} \mu\big[\,\{h \geqslant 1/n\}\,\big]$$

we deduce that there exists $n \in \mathbb{N}$ such that $\mu\big[\,\{h \geqslant 1/n\}\,\big] > 0$. Then

$$0 = \int_{\{h \geqslant 1/n\}} h d\mu \geqslant \frac{1}{n} \int_{\{h \geqslant 1/n\}} d\mu \geqslant \frac{1}{n}\mu\big[\,\{h \geqslant 1/n\}\,\big] > 0.$$

B. Suppose that μ and ν are sigma-finite. Then there exist increasing sequences of measurable sets $(A_n)_{n\in\mathbb{N}}$, $(B_n)_{n\in\mathbb{N}}$ such that

$$\Omega = \bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n, \;\; \mu[\, A_n\,],\; \nu[\, B_n\,] < \infty, \;\; \forall n.$$

If we set $C_n := A_n \cap B_n$, then

$$\Omega = \bigcup_{n\in\mathbb{N}} C_n, \;\; \mu[\, C_n\,],\; \nu[\, C_n\,] < \infty.$$

Define finite measures

$$\mu_n[\, S\,] = \mu[\, I_{C_n} S\,], \;\; \nu_n[\, S\,] = \nu[\, I_{C_n} S\,], \;\; \forall S \in \mathcal{S}, \;\; n \in \mathbb{N}.$$

Then $\nu_n \ll \mu_n$ and we deduce from part **A** so there exist $f_n \in \mathcal{L}^0_+(\Omega, \mathcal{S})$ such that

$$\nu_n S[= \int_\Omega f \boldsymbol{I}_{C_n} d\mu.$$

From the uniqueness result in **A** we deduce that

$$f_n = \boldsymbol{I}_{C_n} f_{n+1}$$

so (f_n) is a nondecreasing sequence of measurable functions. If we denote by f its limit, we deduce from the increasing continuity of ν and the Monotone Convergence theorem that

$$\nu[\, S\,] = \lim_{n\to\infty} \nu[\, C_n \cap S\,] = \lim_{n\to\infty} \int_S f_n d\mu = \int_S f d\mu, \;\; \forall S \in \mathcal{S}.$$

To prove uniqueness, suppose that $g \in \mathcal{L}^0_+(\Omega, \mathcal{S})$ is another function such that

$$\int_S g d\mu = \int_S f d\mu, \;\; \forall S \in \mathcal{S}$$

then $g \boldsymbol{I}_{S_n} = \boldsymbol{I}_{S_n} f$ so

$$g = \lim_{n\to\infty} \boldsymbol{I}_{S_n} g = \lim_{n\to\infty} \boldsymbol{I}_{S_n} f = f.$$

□

Definition 19.5.16. If μ, ν are two sigma-finite measures on the measurable space $(\Omega, \mathcal{S})$ such that $\nu \ll \mu$, then a function $f \in L^0_+(\Omega, \mathcal{S}, \mu_f)$ such that $\nu = \mu_f$ is called a *density* of ν with respect to μ and it is denoted by $\frac{d\nu}{d\mu}$. □

Remark 19.5.17. We want to emphasize an obvious but very important point. The Radon-Nikodym is an *existence* result! It postulates the *existence* of a function given the absolute continuity condition. Heuristically, if $\nu \ll \mu$, then,

$$\frac{d\nu}{d\mu}(\omega) = \lim_{S \searrow \{\omega\}} \frac{\nu\big[\, S \,\big]}{\mu\big[\, S \,\big]} \;\; \text{for a.e. } \omega.$$

From this point of view, the Radon-Nikodym theorem is reminiscent of the Fundamental Theorem of Calculus. When μ is the Lebesgue measure this heuristics can be given a precise meaning. We will discuss this aspect in more detail in the next subsection.

□

Example 19.5.18. Suppose that $U \subset \mathbb{R}^n$ is an open set and $\Phi : U \to \mathbb{R}^n$ is a diffeomorphism with $V := \Phi(U)$. Denote $u = (u^1, \dots, u^n)$ the points in U and by $v = (v^1, \dots, v^n)$ the points in V.

Then the pushforward $\Phi_\# \boldsymbol{\lambda}_U$ is a measure on the Borel sigma-algebra of V, absolutely continuous with respect to $\boldsymbol{\lambda}_V$. Moreover

$$\frac{d\Phi_\# \boldsymbol{\lambda}_U}{d\boldsymbol{\lambda}_V} = |\det J_{\Phi^{-1}}|, \tag{19.5.6}$$

where $J_{\Phi^{-1}}$ is the Jacobian of the map $\Phi^{-1} : V \to U$. This follows from Proposition 19.3.37.

Here is how to remember this. Denote by $|du|$ the Lebesgue measure $\boldsymbol{\lambda}_U\big[\, du \,\big]$ and by $|dv|$ the Lebesgue measure $\boldsymbol{\lambda}_V\big[\, dv \,\big]$.

The map Φ is a map $v = v(u)$ and we denote its Jacobian by $\frac{dv}{du}$ the Jacobian of the inverse is $\frac{du}{dv}$. If we use the notation $|A|$ for the absolute value of the determinant of a matrix A, then we can rewrite (19.5.6) in the more suggestive form

$$\Phi_\# |du| = \left|\frac{du}{dv}\right| \cdot |dv|.$$

For example, consider the map

$$\Phi : (0,1) \to (0,\infty). \;\; u \mapsto v(u) = -\log u.$$

Then $u = \Phi^{-1}(v) = e^{-v}$ and

$$\frac{du}{dv} = -e^{-v}, \;\; \Phi_\# |du| = e^{-v} |dv|.$$ □

19.5.3. Differentiation theorems. Suppose that μ is a finite signed Borel measure on $\mathbb{R}^n$ that is absolutely continuous with respect to the Lebesgue measure $\boldsymbol{\lambda}$, $\mu \ll \boldsymbol{\lambda}$. Radon-Nikodym's theorem tells us that μ must have a special form. More precisely, there exists $f \in L^2(\mathbb{R}^m, \boldsymbol{\lambda})$ such that, for any Borel subset $B \subset \mathbb{R}^n$

$$\mu\big[\,B\,\big] = \boldsymbol{\lambda}_f\big[\,B\,\big] := \int_B f(x)\boldsymbol{\lambda}\big[\,dx\,\big].$$

We write this informally $\mu\big[\,dx\,\big] = f(x)d\boldsymbol{\lambda}\big[\,dx\,\big]$ or, abusing notation,

$$f(x) = \frac{\mu\big[\,dx\,\big]}{\boldsymbol{\lambda}\big[\,dx\,\big]}$$

We want to give a more precise meaning of the last equality.

For each $r > 0$ and $x \in \mathbb{R}^n$ we denote by $A_r\big[\,f\,\big](x)$ the average value of f over the ball $B_r(x)$, of center x and radius r. More precisely

$$A_r\big[\,f\,\big](x) := \frac{1}{v(r)}\int_{B_r(x)} f(y)\,\boldsymbol{\lambda}\big[\,dy\,\big],$$

where $v(r)$ denote the volume[4] of the n-dimensional Euclidean ball of radius r.

Lemma 19.5.19. *For any $f \in L^1(\mathbb{R}^m\boldsymbol{\lambda})$ and any $r > 0$ the function*

$$\mathbb{R}^n \ni x \mapsto A_r\big[\,f\,\big](x) \in \mathbb{R}$$

is continuous.

Proof. Denote by $\boldsymbol{\lambda}_{|f|}$ the measure associated to $|f|$

$$\boldsymbol{\lambda}_{|f|}\big[\,B\,\big] = \int_B |f(y)|\boldsymbol{\lambda}\big[\,dy\,\big].$$

Observe that for any $x_0, x \in \mathbb{R}^n$ we have

$$\big|\,A_r\big[\,f\,\big](x_0) - A_r\big[\,f\,\big](x)\,\big| = \frac{1}{v(r)}\left|\int_{B_r(x_0)} f(y)\boldsymbol{\lambda}\big[\,dy\,\big] - \int_{B_r(x)} f(y)\boldsymbol{\lambda}\big[\,dy\,\big]\right|$$

$(\Delta_r(x_0, x) = B_r(x_0) \cup B_r(x) \backslash B_r(x_0) \cap B_r(x))$

$$\leqslant \frac{1}{v(r)}\int_{\Delta_r(x_0,x)} |f(y)|\,\boldsymbol{\lambda}\big[\,dy\,\big] = \boldsymbol{\lambda}_{|f|}\big[\,\Delta_r(x_0, x)\,\big].$$

Now observe that

$$\lim_{x\to x_0} \boldsymbol{\lambda}\big[\,\Delta_r(x_0, x)\,\big] = 0,$$

and since $\boldsymbol{\lambda}_{|f|} \ll \boldsymbol{\lambda}$ we deduce from Proposition 19.5.12 that

$$\lim_{x\to x_0} \boldsymbol{\lambda}_{|f|}\big[\,\Delta_r(x_0, x)\,\big] = 0.$$

□

[4] More precisely, $v(r) = \boldsymbol{\omega}_n r^n$, where $\boldsymbol{\omega}_n$ is given by (15.3.24).

One of the main goals of this subsection is to show that there exists a Lebesgue negligible subset $\mathcal{N} \subset \mathbb{R}^n$ such that for any $x \in \mathbb{R}^n \setminus \mathcal{N}$ the averages $A_r[\,f\,](x)$ converge to $f(x)$ as $r \searrow 0$. In more compact form

$$\lim_{r \searrow 0} A_r[\,f\,](x) = f(x), \;\; \boldsymbol{\lambda} \text{ a.e.}. \tag{19.5.7}$$

The proof of this result is rather ingenious and contains a few fundamental ideas that have found many other applications in modern analysis. In fact we will prove a stronger result.

Theorem 19.5.20 (Lebesgue's differentiation theorem)**.** *Let $f \in L^1(\mathbb{R}^n)$. For $r > 0$ and $x \in \mathbb{R}^n$ we set*

$$D_r[\,f\,](x) = \frac{1}{v(r)} \int_{B_r(x)} |\, f(y) - f(x) \,|\, \boldsymbol{\lambda}[\,dy\,].$$

Then there exists a Lebesgue negligible subset $\mathcal{N} \subset \mathbb{R}^n$ such that

$$\lim_{r \searrow 0} D_r[\,f\,](x) = 0, \;\; \forall x \in \mathbb{R}^n \setminus \mathcal{N}.$$

Let us observe that Theorem 19.5.20 implies (19.5.7). Indeed, since

$$f(x) = \frac{1}{v(r)} \int_{B_r(x)} f(x) \boldsymbol{\lambda}[\,dy\,]$$

we have

$$|\, A_r[\,f\,](x) - f(x) \,| = \left| \frac{1}{v(r)} \int_{B_r(x)} (\, f(y) - f(x) \,) \boldsymbol{\lambda}[\,dy\,] \right|$$

$$\leqslant \frac{1}{v(r)} \int_{B_r(x)} |\, f(y) - f(x) \,| \boldsymbol{\lambda}[\,dy\,] = D_r[\,f\,](x).$$

The proof of Theorem 19.5.20 relies on the Maximal Theorem of Hardy and Littlewood. To state it we need to introduce an important concept.

Definition 19.5.21. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space and $f : \Omega \to \mathbb{R}$ is an $\mathcal{S}$-measurable function. We write

$$\|f\|_{1,w} := \sup_{\lambda > 0} \lambda \mu[\, \{\, |f| > \lambda \,\} \,].$$

We say that f is of *weak L^1-type* if $\|f\|_{1,w} < \infty$, i.e., there exists $C > 0$ such that

$$\mu[\, \{\, |f| > \lambda \,\} \,] \leqslant \frac{C}{\lambda}, \;\; \forall \lambda > 0.$$

We will denote by $\mathcal{L}^1_w(\Omega, \mathcal{S}, \mu)$ the set of weak L^1-type functions on $(\Omega, \mathcal{S}, \mu)$. □

Let us emphasize that $\| - \|_{1,w}$ is *not a norm*. We will denote by $\| - \|_1$ the L^1-norms.

Proposition 19.5.22. *Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space. Then the following hold.*

(i) *The set* $\mathcal{L}^1_w(\Omega, \mathcal{S}, \mu)$ *is a vector subspace of the set of measurable functions. More precisely,*

$$\forall f, g \in \mathcal{L}^1_w(\Omega, \mathcal{S}, \mu) \;\; \|f + g\|_{1,w} \leqslant 2\big(\|f\|_{1,w} + \|g\|_{1,w} \big).$$

(ii) *For any* $f \in L^1(\Omega, \mathcal{S}, \mu)$, $\|f\|_{1,w} \leqslant \|f\|_1$ *so that* $L^1(\Omega, \mathcal{S}, \mu) \subset \mathcal{L}^1_w(\Omega, \mathcal{S}, \mu)$.

Proof. (i) Note that for any $\lambda > 0$

$$\mu\big[\,\{\,|f + g| > \lambda\}\,\big] \leqslant \mu\big[\,\{\,|f| + |g| > \lambda\}\,\big]$$

$$\leqslant \mu\big[\,\{\,|f| > \lambda/2\}\,\big] + \mu\big[\,\{\,|g| > \lambda/2\}\,\big] \leqslant \frac{2}{\lambda}\big(\|f\|_{1,w} + \|g\|_{1,w} \big).$$

(ii) This is Markov's inequality (19.3.5). □

To each function $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ we associate its *Hardy-Littlewood maximal function* define

$$M\big[\,f\,\big](x) = \sup_{r>0} A_r(|f|) = \sup_{r>0} \frac{1}{v(r)} \int_{B_r(x)} |f(y)|\, \boldsymbol{\lambda}\big[\,dy\,\big].$$

Here is the key technical result of this subsection.

Theorem 19.5.23 (Hardy-Littlewood Maximal inequality)**.** *There exists* $C > 0$ *such that,* $\forall f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$,

$$\|M\big[\,f\,\big]\|_{1,w} \leqslant C\|f\|_1.$$

More explicitly, $\exists C > 0$ *such that, for any* $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ *and any* $\lambda > 0$,

$$\boldsymbol{\lambda}\big[\,\{\,M\big[\,f\,\big] > \lambda\,\}\,\big] \leqslant \frac{C}{\lambda}\|f\|_1.$$

□

We will temporarily take for granted the validity of the Maximal Theorem and show how it can be used to prove Lebesgue's differentiation theorem

Proof of Theorem 19.5.20. Let us first outline the strategy. Denote by $\mathcal{X}$ the set of functions $f \in L^2(\mathbb{R}^n, \boldsymbol{\lambda})$ for which (19.5.7) holds. We first show that $\mathcal{X}$ contains a dense subset and then we show that $\mathcal{X}$ is closed in L^1.

Observe first that any continuous compactly supported function $f : \mathbb{R}^n \to \mathbb{R}$ belongs to $\mathcal{X}$. The simple proof of this fact is left to the reader as Exercise 19.57. The set of continuous compactly supported functions $\mathbb{R}^n \to \mathbb{R}$ is dense in $L^1(\mathbb{R}^n, \boldsymbol{\lambda})$; see Exercise 19.56.

For any $f \in L^1(\mathbb{R}^n, \mathbb{R})$ and $\lambda > 0$ we set

$$S_\lambda(f) := \Big\{ x \in \mathbb{R}^n;\;\; \limsup_{r \searrow 0} D_r\big[\,f\,\big](x) > \lambda \Big\}$$

Lemma 19.5.24. *Let* $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ *and* $g \in \mathcal{X}$ *we have*

$$S_\lambda(f) = S_\lambda(f - g), \;\; \forall \lambda > 0.$$

Proof. For any $f, g \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ we have

$$D_r\big[\,f - g\,\big](x) \leqslant D_r\big[\,f\,\big](x) + D_r\big[\,g\,\big](x),$$

$$D_r\big[\,f\,\big](x) \leqslant D_r\big[\,f - g\,\big](x) + D_r\big[\,g\,\big](x).$$

Hence

$$D_r\big[\,f - g\,\big](x) \leqslant D_r\big[\,f\,\big](x) + D_r\big[\,g\,\big](x) \leqslant D_r\big[\,f - g\,\big](x) + 2D_r\big[\,g\,\big](x).$$

Since $g \in \mathcal{X}$, we have

$$\lim_{r\to 0} D_r\big[\,g\,\big](x) = 0,\ \ \forall x.$$

Hence,

$$\limsup_{r\searrow 0} D_r\big[\,f - g\,\big](x) = \limsup_{r\searrow 0} D_r\big[\,f\,\big](x).$$

□

Let $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$. To prove that $f \in \mathcal{X}$ it suffices to show that

$$\mu\big[\,S_\lambda(f)\,\big] = 0,\ \ \forall \lambda > 0. \tag{19.5.8}$$

For $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ we set

$$\widehat{M}\big[\,f\,\big](x) := \sup_{r>0} D_r\big[\,f\,\big](x).$$

Note that

$$D_r\big[\,f\,\big] \leqslant \frac{1}{v(r)}\int_{B_r(x)} |f(y)|\,\boldsymbol{\lambda}\big[\,dy\,\big] + |f(x)| = A_r\big[\,|f|\,\big](x) + f(x),$$

and we conclude that

$$\widehat{M}\big[\,f\,\big](x) \leqslant M\big[\,f\,\big](x) + |f(x)|.$$

Obviously,

$$S_\lambda(f) \subset \Big\{\, x \in \mathbb{R}^n;\ \ \widehat{M}\big[\,f\,\big](x) > \lambda \,\Big\}$$

$$\subset \Big\{\, x \in \mathbb{R}^n;\ \ M\big[\,f\,\big](x) > \frac{\lambda}{2} \,\Big\} \cup \Big\{\, x \in \mathbb{R}^n;\ \ |f(x)| > \frac{\lambda}{2} \,\Big\}.$$

Hence

$$\boldsymbol{\lambda}\big[\,S_\lambda(f)\,\big] \leqslant \boldsymbol{\lambda}\Big[\,\Big\{\, x \in \mathbb{R}^n;\ \ M\big[\,f\,\big](x) > \frac{\lambda}{2} \,\Big\}\,\Big] + \boldsymbol{\lambda}\Big[\,\Big\{\, x \in \mathbb{R}^n;\ \ |f(x)| > \frac{\lambda}{2} \,\Big\}\,\Big]$$

$$\leqslant \frac{2}{\lambda}\|M(f)\|_{1,w} + \frac{2}{\lambda}\|f\|_1.$$

Theorem 19.5.23 implies that $C > 0$ such that $\forall f \in L^1$ we have $\|M(f)\|_{1,w} \leqslant C\|f\|_1$. Hence

$$\frac{2}{\lambda}\|M(f)\|_{1,w} + \frac{2}{\lambda}\|f\|_1 \leqslant \frac{C_1}{\lambda}\|f\|_1,\ \ C_1 = 2(C+1).$$

In particular

$$\boldsymbol{\lambda}\big[\,S_\lambda(f)\,\big] = \boldsymbol{\lambda}\big[\,S_\lambda(f-g)\,\big] \leqslant \frac{C_1}{\lambda}\,\|f-g\|_1,\ \ \forall g \in \mathcal{X},$$

so that

$$\boldsymbol{\lambda}\big[\,S_\lambda(f)\,\big] \leqslant \frac{C_1}{\lambda}\inf_{g\in\mathcal{X}}\|f-g\|_1 = 0,$$

since $\mathcal{X}$ is dense in L^1. □

Proof of Theorem 19.5.23. Let

$$\begin{aligned} E_\lambda(f) &:= \big\{\,x\in\mathbb{R}^n;\;\; M\big[\,f\,\big](x)>\lambda\,\big\}\\ &= \big\{\,x\in\mathbb{R}^n;\;\; \exists r>0:\;\; A_r\big[\,|f|\,\big](x)>\lambda\,\big\} = \bigcup_{r>0}\big\{A_r\big[\,|f|\,\big]>\lambda\,\big\}. \end{aligned}$$

Lemma 19.5.19 shows that the function $A_r\big[\,f\,\big]$ is continuous. This proves that $E_\lambda(f)$ is an open subset of $\mathbb{R}^n$.

For any $x\in E_\lambda(f)$ there exists $r_x>0$ such that $A_{r_x}\big[\,f\,\big](x)>\lambda$. The collection of open balls

$$\mathscr{C} = \big(\,B_{r_x}(x)\,\big)_{x\in E_\lambda(f)}.$$

Let $v<\boldsymbol{\lambda}\big[\,E_\lambda(f)\,\big]=\boldsymbol{\lambda}\big[\,\{\,M(f)>\lambda\,\}\,\big]$. Wiener's covering theorem (Exercise 19.52) shows that there exist finitely many points $x_1,\dots,x_n\in E_\lambda(f)$ and radii $r_1,\dots,r_N$such that the balls $B_{r_j}(x_j)$ are pairwise disjoint and

$$\sum_{j=1}^N\boldsymbol{\lambda}\big[\,B_{r_j}(x_j)\,\big]\geqslant 3^{-n}v.$$

Since $A_{r_{x_j}}\big[\,f\,\big](x_j)>\lambda$ we deduce

$$3^{-n}v\leqslant\sum_{j=1}^N\boldsymbol{\lambda}\big[\,B_{r_j}(x_j)\,\big]\leqslant\frac{1}{\lambda}\sum_{j=1}^N\int_{B_{r_j}(x_j)}|f(y)|\boldsymbol{\lambda}\big[\,dy\,\big]\leqslant\frac{1}{\lambda}\|f\|_{L^1(\mathbb{R}^n)}$$

Hence

$$\frac{3^n}{\lambda}\|f\|_{L^1(\mathbb{R}^n)}\geqslant v,\;\;\forall v\leqslant\boldsymbol{\lambda}\big[\,E_\lambda(f)\,\big]$$

so that

$$\boldsymbol{\lambda}\big[\,\{\,M(f)>\lambda\,\}\,\big]\leqslant\frac{3^n\|f\|_1}{\lambda}.$$

□

Definition 19.5.25. Suppose that $f\in\mathcal{L}^1(\mathbb{R}^,\boldsymbol{\lambda})$. A point $x\in\mathbb{R}^n$ is called a *Lebesgue point* of f if

$$\lim_{r\searrow 0}D_r\big[\,f\,\big](x)=0.$$

□

Corollary 19.5.26 (Lebesgue's Fundamental Theorem of Calculus)**.** *Let* $f\in\mathcal{L}^1\big([a,b],\boldsymbol{\lambda}\,\big)$. *Define* $F:[a,b]\to\mathbb{R}$

$$F(x)=\int_a^x f(t)\boldsymbol{\lambda}\big[\,dt\,\big].$$

Then F is differentiable a. e. *and $F'(x) = f(x)$ for every x where the derivative of F exists.*

Proof. Extend f to an integrable function $\bar{f} : \mathbb{R} \to \mathbb{R}$

$$\bar{f}(x) = \begin{cases} f(x), & x \in [0,1], \\ 0, & x \in \mathbb{R}\setminus[0,1]. \end{cases}$$

Theorem 19.5.20 shows that for almost every point $x \in \mathbb{R}$ we have

$$\lim_{r \searrow 0} \frac{1}{2r} \int_{x-r}^{x+r} \big[\bar{f}(t) - \bar{f}(x)\,\big|\boldsymbol{\lambda}\big[\,dt\,\big] = 0.$$

In particular, for almost every point $x \in (0,1)$ we have

$$\frac{1}{r}\int_{x-r}^{x} \big[\, f(t) - f(x)\,\big|\boldsymbol{\lambda}\big[\,dt\,\big] \to 0, \;\; \frac{1}{r}\int_{x}^{x+r} \big(\, f(t) - f(x)\,\big|\boldsymbol{\lambda}\big[\,dt\,\big] \to 0,$$

as $r \searrow 0$. Observe that for any r sufficiently small

$$\left|\frac{F(x-r) - F(x)}{-r} - f(x)\right| = \frac{1}{r}\left|\int_{x-r}^{x} \big(\, f(t) - f(x)\,\big)\boldsymbol{\lambda}\big[\,dt\,\big]\right| \leqslant \frac{1}{r}\int_{x-r}^{x} \big|\, f(t) - f(x)\,\big|\boldsymbol{\lambda}\big[\,dt\,\big],$$

$$\left|\frac{F(x+r) - F(x)}{r} - f(x)\right| = \frac{1}{r}\left|\int_{x}^{x+r} \big(\, f(t) - f(x)\,\big)\boldsymbol{\lambda}\big[\,dt\,\big]\right| \leqslant \frac{1}{r}\int_{x}^{x+r} \big|\, f(t) - f(x)\,\big|\boldsymbol{\lambda}\big[\,dt\,\big].$$

□

19.6. Duality

Let us recall (see Definition 17.1.51) that the dual of a normed space $(X, \|-\|)$ is the space $X^* := \boldsymbol{B}(X, \mathbb{R})$ of continuous linear functions $\boldsymbol{\alpha} : X \to \mathbb{R}$. The space X^* is itself equipped with a norm $\|-\|_*$

$$\|\boldsymbol{\alpha}\|_* := \sup_{x \in X\setminus\{0\}} \frac{|\,\boldsymbol{\alpha}(x)\,|}{\|x\|}.$$

19.6.1. The dual of $C(K)$. Let (K, d) be a compact metric space. Let $C(K)$ denote the space of continuous functions $K \to \mathbb{R}$. As usual, we regard it as Banach space with norm

$$\|f\| := \sup_{x \in K} |f(x)|.$$

For ease of notation we denote by $\boldsymbol{F}$ this Banach space and by $\boldsymbol{F}_+$ the subset consisting of nonnegative functions. The goal of this subsection is to give a concrete description of the topological dual $\boldsymbol{F}^*$of X. The dual $\boldsymbol{F}^*$ contains a distinguished subset $\boldsymbol{F}^*_+$ consisting of *positive* linear functionals, i.e., continuous linear functionals $\boldsymbol{\alpha} : \boldsymbol{F} \to \mathbb{R}$ such that

$$\boldsymbol{\alpha}\big[\, f\,\big] \geqslant 0, \;\; \forall f \in \boldsymbol{F}_+.$$

Proposition 19.6.1. *If $\boldsymbol{\alpha} : \boldsymbol{F} \to \mathbb{R}$ is linear and nonnegative, i.e., $\boldsymbol{\alpha}\big(\boldsymbol{F}_+\big) \subset [0, \infty)$, then $\boldsymbol{\alpha}$ is continuous and*

$$\|\boldsymbol{\alpha}\|_* \leqslant \boldsymbol{\alpha}(1).$$

Proof. Observe first that $\boldsymbol{\alpha}$ is monontone, i.e.,

$$f \leqslant g \Rightarrow \boldsymbol{\alpha}(f) \leqslant \boldsymbol{\alpha}(g).$$

Indeed

$$f \leqslant g \Rightarrow g - f \geqslant 0 \Rightarrow \boldsymbol{\alpha}\big(\, g - f\,\big) \geqslant 0.$$

Set $C := \boldsymbol{\alpha}(1)$. For any $f \in \boldsymbol{F}$ we have

$$-\|f\| \leqslant f \leqslant \|f\| \Rightarrow -C\|f\| = -\boldsymbol{\alpha}\big(\,\|f\|\,\big) \leqslant \boldsymbol{\alpha}(f) \leqslant \boldsymbol{\alpha}\big(\,\|f\|\,\big) = C\|f\|.$$

Hence

$$\forall f \in \boldsymbol{F},\ \ \big|\,\boldsymbol{\alpha}(f)\,\big| \leqslant C\|f\|,$$

so $\boldsymbol{\alpha}$ is continuous and $\|\boldsymbol{\alpha}\|_* \leqslant C$. □

We denote by $\mathcal{M}(K)$ the space of *signed* finite Borel measures on K. Let $\mathcal{M}_+(K) \subset \mathcal{M}(K)$ denote the space of finite Borel measures on K.

Let us observe that we have a natural map

$$\mathcal{M}(K) \ni \mu \mapsto L_\mu \in \boldsymbol{F}^*,\ \ L_\mu(f) = \mu\big[\,f\,\big] := \int_K f(x)\mu\big[\,dx\,\big]$$

To see that the linear functional L_μ is continuous consider the Jordan decomposition $\mu = \mu_+ - \mu_-$. Then $L_\mu = L_{\mu_+} - L_{\mu_-}$ and $L_{\mu_\pm} : \boldsymbol{F} \to \mathbb{R}$ are linear and nonnegative maps $\boldsymbol{F} \to \mathbb{R}$. Proposition 19.6.1 implies that they are continuous and

$$\|L_{\mu_pm}\|_* \leqslant L_{\mu_\pm}(1) = \mu_\pm\big[\,K\,\big].$$

In particular,

$$\|L_\mu\|_* \leqslant |\mu|\big[\,K\,\big],\ \ \forall \mu \in \mathcal{M}(K), \tag{19.6.1}$$

where $|\mu| = \mu_+ + \mu_-$.

We have the following fundamental result due to Frygues (Frederic) Riesz (1880-1956).

Theorem 19.6.2 (Riesz representation of measures)**.** *The map*

$$\mathcal{M}(K)_+ \ni \mu \mapsto L_\mu \in \boldsymbol{F}^*_+$$

is bijective. More explicitly, for any nonnegative linear functional $\alpha : C(K) \to \mathbb{R}$ there exists a unique finite Borel measure $\mu \in \mathcal{M}(K)_+$ such that $\alpha = L_\mu$. Moreover $\|\alpha\|_* = \mu\big[\,K\,\big]$.

Proof. We will give a proof based on the concept of independent interest, namely the *Daniell integral.* Let us first gather a few elementary facts.

Definition 19.6.3. Let X be a set and $\mathcal{F}$ a vector space of bounded functions $X \to \mathbb{R}$.

(i) We say that $\mathcal{F}$ is a *vector lattice* if the following hold.
 (a) The constant function 1 belongs to $\mathcal{F}$.
 (b) If $f, g \in \mathcal{F}$, then $\max(f, g), \min(f, g) \in \mathcal{F}$.

(ii) A *Daniell integral* on $\mathcal{F}$ is a linear functional $L : \mathcal{F} \to \mathbb{R}$ satisfying the following properties.
 (a) $L\big[\,f\,\big] \geqslant 0$ for any $f \in \mathcal{F}$ such that $f \geqslant 0$.

(b) If $(f_n)_{n\geqslant 1}$ is a sequence of functions in $\mathcal{F}$ such that $f_n \geqslant f_{n+1}$, $\forall n \geqslant 1$ and
$$\forall x \in X,\quad \lim_{n\to\infty} f_n(x) = 0,$$
then $\lim_{n\to\infty} L\big[\,f_n\,\big] = 0$.

□

Lemma 19.6.4. *Let (K, d) be a compact metric space. The following hold.*

(i) *The vector space $\boldsymbol{F} = C(K)$ is a vector lattice.*

(ii) *The sigma-algebra $\sigma\big(\,\boldsymbol{F}\,\big)$ of subsets of K generated by the functions in $\boldsymbol{F}$ is the Borel sigma-algebra of K. Recall that $\sigma\big(\,\boldsymbol{F}\,\big)$ is the sigma-algebra generated by the subsets $\{f \leqslant c\} \subset K$, $f \in \boldsymbol{F}$, $c \in \mathbb{R}$.*

(iii) *If $L : \boldsymbol{F} \to \mathbb{R}$ is a continuous linear functional an $(f_n)_{n\geqslant 1}$ is a nonincreasing sequence of continuous functions on K such that*
$$\forall x \in K,\quad \lim_{n\to\infty} f_n(x) = 0,$$
then $\lim_{n\to\in\infty} L\big[\,f_n\,\big] = 0$. In particular, if $L \in \boldsymbol{F}^*_+$, then L is a Daniell integral.

Proof of Lemma 19.6.4. (i) obvious. To prove (ii) note first that any continuous function on K is Borel measurable so $\sigma\big(\,\boldsymbol{F}\,\big) \subset \mathcal{B}_K$. To prove the reverse inclusion it suffices to prove that any closed set is contained in $\sigma\big(\,\boldsymbol{F}\,\big)$. To see this, let $C \subset K$ be a closed subset. Then the function $f_C(x) = \operatorname{dist}(x, C)$ is continuous and
$$C = \big\{\, f_C \leqslant 0\,\big\} \in \sigma\big(\,\boldsymbol{F}\,\big).$$
(iii) Dini's theorem (Theorem 17.4.5) implies that f_n converges uniformly to 0 and the desired conclusion follows from the continuity of L. □

Lemma 19.6.4 shows that Theorem 19.6.2 is now a consequence of the following general results of P. J. Daniell, [10].

Theorem 19.6.5 (Daniell integral)**.** *Let Ω be an arbitrary set, $\mathcal{F}$ a vector lattice space of functions on Ω, and $L : \mathcal{F} \to \mathbb{R}$ a* Daniell integral. *Denote by $\mathcal{S}$ the sigma-algebra $\sigma\big(\,\mathcal{F}\,\big)$ generated by the functions in $\mathcal{F}$, i.e., the sigma-algebra generated by the sets*
$$\big\{\, f \leqslant c\,\big\},\quad f \in \mathcal{F},\quad c \in \mathbb{R}.$$
Then there exists a unique finite measure μ on $\mathcal{S}$ such that
$$\mathcal{F} \subset \mathcal{L}^1\big(\,\Omega, \mathcal{S}, \mu\,\big) \;\; and \;\; L\big[\,f\,\big] = \int_\Omega f d\mu,\quad \forall f \in \mathcal{F}.$$

□

Proof of Theorem 19.6.5. We follow the approach in [30, Chap.III] which is closer to the original strategy employed by Daniell, [10]. For an alternate approach we refer to [7, Sec. 7.12].

Without loss of generality we can assume that $L[\,1\,]=1$. To simplify the notation we set

$$f\vee g:=\max(f,g),\;\; f\wedge g:=\min(f,g).$$

Observe that (iii) implies that

$$f,g\in\mathcal{F},\;\; f\leqslant g\;\Rightarrow L[\,f\,]\leqslant L[\,g\,].$$

We denote by $\mathcal{F}^*$ the set of functions $f:\Omega\to(-\infty,\infty]$ such that there exists a nondecreasing sequence $(f_n)_{n\geqslant 1}$ in $\mathcal{F}$ approximating f from below, i.e.,

$$\forall\omega\in\Omega,\;\; f_n(\omega)\nearrow f(\omega)\text{ as }n\to\infty.$$

We will refer to such a sequence (f_n) as a *lower approximant* of f.

Let $f\in\mathcal{F}^*$. If (f_n) is a lower approximant of f, then the nondecreasing sequence of real numbers $L[\,f_n\,]$ has a limit in $(-\infty,\infty]$.

Lemma 19.6.6. *Let $f\in\mathcal{F}^*$. If $(f_n)_{n\geqslant 0}$ and $(g_n)_{n\geqslant 0}$ are two lower approximants of f then*

$$\lim_{n\to\infty}L[\,f_n\,]=\lim_{n\infty}L[\,g_n\,].$$

Proof. For each $n\geqslant 1$, the sequence $\big(f_n\wedge g_k\big)_{k\geqslant 1}$ is nondecreasing and converges pointwisely to f_n. Using property (iv) of L we deduce

$$L[\,f_n\,]=\lim_{k\to\infty}L[\,f_n\wedge g_k\,]\leqslant\lim_{k\to\infty}L[\,g_k\,].$$

If we now let $n\to\infty$ we deduce

$$\lim_{n\to\infty}L[\,f_n\,]\leqslant\lim_{k\infty}L[\,g_k\,].$$

Reversing the roles of the f's and the g's in the above argument we obtain the desired conclusion. □

The above lemma shows that we can extend L to a function $L:\mathcal{F}^*\to(-\infty,\infty]$ by setting

$$L[\,f\,]=\lim_{n\to\infty}L[\,f_n\,]$$

where $(f_n)_{n\geqslant 1}$ is *any* lower approximant of $f\in\mathcal{F}^*$.

Lemma 19.6.7. *Let $f,g\in\mathcal{F}^*$ and $c\geqslant 0$. Then the following hold.*

(i) $f\leqslant g\Rightarrow L[\,f\,]\leqslant L[\,g\,]$.

(ii) $f+g,\ cf\in\mathcal{F}^*$ *and* $L[\,f+g\,]=L[\,f\,]+L[\,g\,]$, $L[\,cf\,]=cL[\,f\,]$.

(iii) $f\vee g, f\wedge g\in\mathcal{F}^*$ *and*

$$L[\,f\,]+L[\,g\,]=L[\,f\vee g\,]+L[\,f\wedge g\,].$$

(iv) *If $(f_n)_{n\geqslant 0}$ is a nondecreasing sequence in $\mathcal{F}^*$ then*

$$f_\infty:=\lim_{n\to\infty}f_n\in\mathcal{F}^*\;\;\textit{and}\;\; L[\,f_\infty\,]=\lim_{n\to\infty}L[\,f_n\,].$$

Proof. (i) If (f_n) and (g_n) are lower approximants of f and respectively g, then because $f \leqslant g$, the sequences $(f_n \wedge g_n)$ and $(f_n \vee g_n)$ are lower approximants of f and respectively g and

$$L\big[\,f_n \wedge g_n\,\big] \leqslant L\big[\,f_n \vee g_n\,\big],\ \ \forall n.$$

(ii) If (f_n) and (g_n) are lower approximants of f and respectively g, then $(f_n + g_n)$ is a lower approximant of $f+g$ and (cf_n) is a lower approximant of cf.

(iii) If (f_n) and (g_n) are lower approximants of f and respectively g, then $(f_n \wedge g_n)$ and $(f_n \vee g_n)$ are lower approximants of $f \wedge g$ and respectively $f \vee g$. The rest follows from (ii) and the equality

$$f + g = f \wedge g + f \vee g.$$

(iv) Suppose that $(f_{n,k})_{k\geqslant 1}$ is a lower approximant of f_n, $\forall n$. Set

$$g_k := \max_{1\leqslant n\leqslant k} f_{n,k}.$$

Then $g_k \leqslant g_{k+1}$, $\forall k$. Indeed, for $n \leqslant k$ we have

$$f_{n,k} \leqslant f_{n,k+1},$$

and

$$g_k = \max_{1\leqslant n\leqslant k} f_{n,k} \leqslant \max_{1\leqslant n\leqslant k} f_{n,k+1}. \leqslant \max_{1\leqslant n\leqslant k+1} f_{n,k+1} = g_{k+1}$$

We have $g_k \in \mathcal{F}$

$$f_{n,k} \leqslant g_k \leqslant \max_{n\leqslant k} f_n = f_k,\ \ \forall n \leqslant k. \tag{19.6.2}$$

This shows that the nondecreasing sequence (g_k) in $\mathcal{F}$ is uniformly bounded. Hence it is a lower approximant for

$$g_\infty := \lim_{k\to\infty} g_k \in \mathcal{F}^*,$$

and

$$L\big[\,g_\infty\,\big] \leqslant \lim_{k\to\infty} L\big[\,f_k\,\big].$$

Letting $k \to \infty$ in (19.6.2) we deduce

$$f_n \leqslant g_\infty \leqslant \lim_{k\to\infty} f_k = f_\infty,\ \ \forall n.$$

Letting $n \to \infty$ we deduce

$$f_\infty = g_\infty.$$

Moreover

$$\lim_{n\to\infty} L\big[\,f_n\,\big] \leqslant L\big[\,g_\infty\,\big] \leqslant \lim_{k\to\infty} L\big[\,f_k\,\big].$$

Hence

$$\lim_{n\to\infty} L\big[\,f_n\,\big] = L\big[\,g_\infty\,\big] = L\big[\,f_\infty\,\big].$$

□

We set

$$\mathcal{F}_* := \big\{\, g : \Omega \to [-\infty, \infty);\ \ -g \in \mathcal{F}^*\,\big\}.$$

Equivalently, $g \in \mathcal{F}^*$ iff exists a nonincreasing sequence (g_n) in $\mathcal{F}$ such that

$$\forall \omega \in \Omega,\ \ g_n(\omega) \searrow g(\omega),\ \ \text{as } n \to \infty.$$

We refer to such a sequnce as an *upper approximant* Define $L : \mathcal{F}_* \to [-\infty, \infty)$ by

$$L\big[\,g\,\big] := -L\big[\,-g\,\big],\ \ \forall g \in \mathcal{F}_*$$

Equivalently

$$L\big[\,g\,\big] = \lim_{n\to\infty} L\big[\,g_n\,\big],$$

where (g_n) is any upper approximant of g. Note that if $g^* \in \mathcal{F}^*$, $g_* \in \mathcal{F}_*$ and $g^* \geqslant g_*$, then $g^* - g_* \in \mathcal{F}^*$ and

$$0 \leqslant L\big[\, g^* - g_* \,\big] = L\big[\, g^* \,\big] - L\big[\, g_* \,\big].$$

Observe also that

$$\mathcal{F} \subset \mathcal{F}^* \cap \mathcal{F}_*.$$

Denote by $\mathcal{L}^1(\mathcal{F})$ the collection of functions $f : \Omega \to [-\infty, \infty]$ with the following property:

for any $\varepsilon > 0$ there exist $f^* = f^{*,\varepsilon} \in \mathcal{F}^*$ and $f_* = f_{*,\varepsilon} \in \mathcal{F}_*$ such that

$$f_* \leqslant f \leqslant f^* \ \text{ and } \ L\big[\, f^* \,\big] - L\big[\, f_* \,\big] \leqslant \varepsilon.$$

The above condition tacitly assumes that both $L\big[\, f^* \,\big]$ and $L\big[\, f_* \,\big]$ are finite.

If we set

$$L_*\big[\, f \,\big] := \sup_{\substack{f_* \in \mathcal{F}_* \\ f_* \leqslant f}} L\big[\, f_* \,\big], \ \ L^*\big[\, f \,\big] := \inf_{\substack{f^* \in \mathcal{F}^* \\ f \leqslant f^*}} L\big[\, f^* \,\big],$$

then we see that $f \in \mathcal{L}^1(\mathcal{F})$ if and only if

$$-\infty < L_*\big[\, f \,\big] = L^*\big[\, f \,\big] < \infty.$$

Note that $\mathcal{F} \subset \mathcal{L}^1(\mathcal{F})$. Moreover,

$$\forall f \in \mathcal{F}^*, \ \ L\big[\, f \,\big] < \infty \Rightarrow f \in \mathcal{L}^1(\mathcal{F}).$$

To see this choose a lower approximant $f_n \nearrow f$. Set

$$f_{n,*} = f_n, \ \ f_n^* := f.$$

Then

$$f_{n,*} \leqslant f \leqslant f_n^* \ \text{ and } \ L\big[\, f_n^* \,\big] - L\big[\, f_{n,*} \,\big] \to 0.$$

Similarly, if $f \in \mathcal{F}_*$ and $L\big[\, f \,\big] > -\infty$, then $f \in \mathcal{L}^1(\mathcal{F})$.

We define

$$L : \mathcal{L}^1(\mathcal{F}) \to \mathbb{R}, \ \ L\big[\, f \,\big] = L_*\big[\, f \,\big] = L^*\big[\, f \,\big].$$

For $f, g : \Omega \to [-\infty, \infty]$ we define $f + g : \Omega \to [-\infty, \infty]$ by setting

$$(f+g)(\omega) := \begin{cases} f(\omega) + g(\omega), & \big(\, f(\omega), g(\omega) \,\big) \neq (\infty, -\infty), \\ 0, & \big(\, f(\omega), g(\omega) \,\big) = (\pm\infty, \mp\infty). \end{cases}$$

Note that if $f_1 \leqslant f_2$, $g_1 \leqslant g_2$, then $f_1 + g_1 \leqslant f_2 + g_2$.

Lemma 19.6.8. *The following hold.*

(i) $\mathcal{L}^1(\mathcal{F})$ *is a vector subspace of the space of functions* $\Omega \to [-\infty, \infty]$ *with respect to the addition defined as above. Moreover, L is linear.*

(ii) *If $f, g \in \mathcal{L}^1(\mathcal{F})$, then $f \wedge g$, $f \vee g \in \mathcal{L}^1(\mathcal{F})$.*

(iii) $\forall f, g \in \mathcal{L}^1(\mathcal{F})$, $f \leqslant g \Rightarrow L\big[\, f \,\big] \leqslant L\big[\, g \,\big]$.

(iv) *If (f_n) is a nondecreasing sequence in $\mathcal{L}^1(\mathcal{F})$,*

$$f_\infty(\omega) = \lim_{n\to\infty} f_n(\omega), \quad \forall \omega \in \Omega,$$

and

$$\sup_n L\big[\, f_n \,\big] < \infty,$$

then $f_\infty \in \mathcal{L}^1(\mathcal{F})$ and

$$L\big[\, f_\infty \,\big] = \lim_{n\to\infty} L\big[\, f_n \,\big].$$

Proof. (i) + (ii) We have for any $\varepsilon > 0$ there exist $f_\varepsilon^+, g_\varepsilon^+ \in \mathcal{F}^*$ and $f_\varepsilon^-, g_\varepsilon^- \in \mathcal{F}_*$ such that

$$f_\varepsilon^- \leqslant f \leqslant f_\varepsilon^+, \quad g_\varepsilon^- \leqslant g \leqslant g_\varepsilon^+,$$

$$L\big[\, f_\varepsilon^+ \,\big] - L\big[\, f_\varepsilon^- \,\big] \leqslant \frac{\varepsilon}{2}, \quad L\big[\, g^+\varepsilon \,\big] - L\big[\, g_\varepsilon^- \,\big] \leqslant \frac{\varepsilon}{2}.$$

Then $h_\varepsilon^+ = f_\varepsilon^+ + g_\varepsilon^+ \in \mathcal{F}^*$, $h_\varepsilon^- = f_\varepsilon^- + g_\varepsilon^- \in \mathcal{F}_*$. Moreover[5]

$$h_\varepsilon^- \leqslant f + g \leqslant h_\varepsilon^+.$$

Clearly

$$\big(L\big[\, f_\varepsilon^+ \,\big] + L\big[\, g_\varepsilon^+ \,\big] \big) - \big(L\big[\, f_\varepsilon^- \,\big] + L\big[\, g_\varepsilon^- \,\big] \big) \leqslant \varepsilon.$$

Hence $f + g \in \mathcal{L}^1(\mathcal{F})$. The above argument also shows that $L\big[\, f+g \,\big] = L\big[\, f \,\big] + L\big[\, g \,\big]$. Arguing in a similar fashion one shows that for any $f \in \mathcal{L}^1(\mathcal{F})$ and any $c \in \mathbb{R}$ we have $cf \in \mathcal{L}^1(\mathcal{F})$ and $L\big[\, cf \,\big] = cL\big[\, f \,\big]$. Observe next that

$$f_\varepsilon^- \wedge g_\varepsilon^- \leqslant f \wedge g \leqslant f_\varepsilon^+ \wedge g_\varepsilon^+,$$

and

$$f_\varepsilon^+ \wedge g_\varepsilon^+ - f_\varepsilon^- \wedge g_\varepsilon^- \leqslant (f_\varepsilon^+ - f_\varepsilon^-) + (g_\varepsilon^+ - g_\varepsilon^-).$$

Using Lemma 19.6.7 we deduce

$$L\big[\, f_\varepsilon^+ \wedge g_\varepsilon^+ \,\big] - L\big[\, f_\varepsilon^- \wedge g_\varepsilon^- \,\big] \leqslant L\big[\, (f_\varepsilon^+ - f_\varepsilon^-) \,\big] + L\big[\, (g_\varepsilon^+ - g_\varepsilon^-) \,\big] \leqslant \varepsilon,$$

so $f \wedge g \in \mathcal{L}^1(\mathcal{F})$. Next observe that $-(f \vee g) = (-f) \wedge (-g) \in \mathcal{L}^1(\mathcal{F})$.

(iii) If $f \in \mathcal{L}^1(\mathcal{F})$ and $f \geqslant 0$. We have $L\big[\, f \,\big] \geqslant L\big[\, f^- \,\big]$, $\forall f^- \in \mathcal{F}_*$. Then

$$L\big[\, f \,\big] \geqslant L\big[\, f^- \vee 0 \,\big] \geqslant 0.$$

If $f \geqslant g$, then $f - g \geqslant 0$ and

$$L\big[\, f \,\big] = L\big[\, g \,\big] + L\big[\, f - g \,\big] \geqslant L\big[\, g \,\big].$$

(iv) By replacing f_n with $f_n - f_1$ we can assume $f_n \geqslant 0$, $\forall n \geqslant 1$. Fix $\varepsilon > 0$.

For each $n \geqslant 1$ we can find $h_n \in \mathcal{F}^*$, $h_n \geqslant 0$ such that

$$(f_n - f_{n-1}) \leqslant h_n. \;\; L\big[\, h_n \,\big] \leqslant L\big[\, f_{n+1} - f_n \,\big] + \frac{\varepsilon}{2^n}, \quad \forall n \geqslant 1,$$

where $f_0 := 0$. The sequence

$$H_n := \sum_{k=1}^n h_k \in \mathcal{F}^*$$

is nondecreasing and

$$\lim_{n\to\infty} L\big[\, H_n \,\big] \leqslant \lim_{n\to\infty} L\big[\, f_n \,\big] + \varepsilon.$$

Hence its limit H_∞ belongs to $\mathcal{F}^* \cap \mathcal{L}^1(\mathcal{F})$. Clearly $f_\infty \leqslant H_\infty$. Choose m large so that

$$L\big[\, f_m \,\big] \geqslant \lim_{n\to\infty} L\big[\, f_n \,\big] - \varepsilon$$

[5] Pay careful attention to the situation when $f(x) = \infty$, $g(x) = -\infty$. In this case $f_\varepsilon^+(x) = \infty$ $g_\varepsilon^+(x) \in (-\infty, \infty]$ so $f_\varepsilon^+(x) + g_\varepsilon^+(x) \geqslant 0 = f(x) + g(x)$. On the other hand $f_\varepsilon^-(x) \in [-\infty, \infty)$, $g_\varepsilon^-(x) = -\infty$ so $f_\varepsilon^-(x) + g_\varepsilon^-(x) \leqslant 0$.

and then choose $f_m^- \in \mathcal{F}_*$ such that $f_m^- \leqslant f_m \leqslant f_\infty$

$$L\big[\, f_m \,\big] - L\big[\, f_m^- \,\big] < \varepsilon.$$

We deduce that $f_m^- \leqslant f_\infty \leqslant H_\infty$ and

$$L\big[\, H_\infty \,\big] - L\big[\, f_m^- \,\big] \leqslant \lim_{n\to\infty} L\big[\, f_n \,\big] - L\big[\, f_m^- \,\big] + \varepsilon \leqslant 3\varepsilon.$$

This proves (iv). □

Denote by $\mathcal{G}$ the collection of subsets $G \subset \Omega$ such that $\boldsymbol{I}_G \in \mathcal{L}^1(\mathcal{F})$. Note that $\varnothing, \Omega \in \mathcal{G}$. For $G \in \mathcal{G}$ we set

$$\mu\big[\, G \,\big] = \mu_L\big[\, G \,\big] := L\big[\, \boldsymbol{I}_G \,\big].$$

Note that

$$\boldsymbol{I}_{G_0} \wedge \boldsymbol{I}_{G_1} = \boldsymbol{I}_{G_0 \cap G_1},\ \ \boldsymbol{I}_{G_0} \vee \boldsymbol{I}_{G_1} = \boldsymbol{I}_{G_0 \cup G_1},\ \ \boldsymbol{I}_{G^c} = 1 - \boldsymbol{I}_G,$$

so that $\mathcal{G}$ is an algebra of sets. Lemma 19.6.8 shows that $\mathcal{G}$ is a sigma-algebra and μ is a measure on $\mathcal{G}$. Let us show that any $f \in \mathcal{F}$ is $\mathcal{G}$-measurable, i.e., $\sigma(\mathcal{F}) \subset eG$.

Observe that for any $f \in \mathcal{F}$ and $c > 0$ we have

$$n\big(\, f - f \wedge c \,\big) \wedge 1 \nearrow \boldsymbol{I}_{\{f > c\}}.$$

Indeed, if $f(\omega) > c$, then $\big(\, f - f \wedge c \,\big)(\omega) = f(\omega) - c > 0$,

$$\lim_{n\to\infty} n\big(\, f - f \wedge c \,\big)(\omega) = +\infty$$

and thus for n sufficiently large we have

$$n\big(\, f - f \wedge c \,\big)(\omega) \wedge 1 = 1 = \boldsymbol{I}_{f > c}(\omega).$$

if $f(\omega) \leqslant c$, then $\big(\, f - f \wedge c \,\big)(\omega) = 0$.

Hence

$$\{f > c\} \in \mathcal{G},\ \ \forall f \in \mathcal{F},\ \ c > 0.$$

In particular, this shows that

$$\{f \leqslant c\} \in \mathcal{G},\ \ \forall f \in \mathcal{F},\ \ c > 0.$$

Since

$$\{f \leqslant 0\} = \bigcap_{n \in \mathbb{N}} \{f \leqslant 1/n\}$$

we deduce that $\{f \leqslant 0\} \in \mathcal{G}$. For $c > 0$ we have

$$\{f \leqslant -c\} = \{-f \geqslant c\} = \bigcap_{n \in \mathbb{N}} \{-f > c - 1/n\} \in \mathcal{G}.$$

We deduce that

$$\{f \leqslant c\} \in \mathcal{G},\ \ \forall f \in \mathcal{F},\ \ c \in \mathbb{R},$$

so that $\sigma(\mathcal{F}) \subset \mathcal{G}$.

Observe that for any nonnegative $\sigma(\mathcal{F})$-elementary function g we have $g \in \mathcal{L}^1(\mathcal{F})$ and

$$\int_\Omega g d\mu = L\big[\, g \,\big].$$

If $f \in \mathcal{F}$, $f \geqslant 0$, then there exists a sequence of $\sigma(\mathcal{F})$-elementary functions $f_n \nearrow f$. The Monotone Convergence Theorem and Lemma 19.6.8 imply that

$$L[\,f\,] = \int_\Omega f d\mu.$$

Using the decomposition

$$f = f_+ - f_- = f \wedge 0 - (-f \wedge 0)$$

we deduce that

$$L[\,f\,] = \int_\Omega f d\mu, \;\; \forall f \in \mathcal{F}.$$

Clearly μ is uniquely determined by L. Indeed if ν is another measure with property then the above arguments show that $\nu = \mu_L$. □

Theorem 19.6.9 (The dual of $C(K)$)**.** *For any continuous linear functional $\alpha : C(K) \to \mathbb{R}$ there exists a unique finite signed Borel measure $\mu \in \mathcal{M}(K)$ such that $\alpha = L_\mu$. Moreover $\|L_\mu\|_* = |\mu|\big[\,K\,\big]$.*

Proof. I follow the presentation in [5, Thm.7.8.3]. Let α be a continuous linear functional on $\boldsymbol{F} = C(K)$. For $f \in \boldsymbol{F}_+$ we set

$$\alpha_+(f) = \sup_{g \in [0,f]} \alpha(g), \;\; [0,f] := \big\{\, g \in C(K); \;\; 0 \leqslant g \leqslant f \,\big\}.$$

Note that for any $g \in [0,f]$ we have $\|g\| \leqslant \|f\|$ and

$$\big|\,\alpha(g)\,\big| \leqslant \|\alpha\|_* \cdot \|f\|$$

so $\alpha_+(f)$ is finite. More precisely

$$\big|\,\alpha_+(f)\,\big| \leqslant \|_* \cdot \|f\|.$$

Extend α_+ to $\boldsymbol{F}$ by setting

$$\alpha_+(f) := \alpha_+(f_+) - \alpha_+(f_-).$$

Define $\alpha_- : C(K) \to \mathbb{R}$, $\alpha_- = \alpha_+ - \alpha$, i.e., $\alpha = \alpha_+ - \alpha_-$.

Lemma 19.6.10. *Both $\alpha_\pm$ are continuous positive linear functionals on $\boldsymbol{F}$.*

Proof. Observe first that $\alpha_+(f) \geqslant \alpha(0) = 0$, $\forall f \in \boldsymbol{F}_+$. Clearly

$$0 \leqslant f_1 \leqslant f_2 \Rightarrow \alpha_+(f_1) \leqslant \alpha_+(f_2).$$

Moreover, for any $c > 0$ and any $f \in \boldsymbol{F}_+$ we have

$$\alpha_+(cf) = c\alpha_+(f).$$

Set $A_\alpha := \alpha_+(1)$. From the inequalities $0 \leqslant f \leqslant \|f\|$ we deduce

$$\alpha_+(f) \leqslant A_\alpha \|f\|, \;\; \forall f \in \boldsymbol{F}_+.$$

Let $f_1, f_2 \in \boldsymbol{F}_+$. If $g_i \in [0, f_i]$, $i = 1, 2$, then $g_1 + g_2 \in [0, f_1 + f_2]$ so

$$\alpha_+(f_1) + \alpha_+(f_2) = \sup_{g_1 \in [0,f_1]} \alpha(g_1) + \sup_{g_2 \in [0,f_2]} \alpha(g_2) \leqslant \sup_{g \in [0,f_1+f_2]} \alpha(g) = \alpha_+(f_1 + f_2).$$

To prove the opposite inequality we need to show that

$$\forall g \in [0, f_1 + f_2], \;\; \alpha_+(g) \leqslant \alpha_+(f_1) + \alpha_+(f_2).$$

Let $g \in [0, f_1 + f_2]$. Set $g_1 = \min(g, f_1)$ and $g_2 = g - g_1$. Clearly $g_1 \in [0, f_1]$. Next observe that $g_2 \in [0, f_2]$ as well. Indeed, if $g_1(x) = f_1(x)$, then $g(x) \geqslant f_1(x)$ so $g_2(x) = g(x) - f_1(x) \in [0, f_2(s)]$. On the other hand if $g_1(x) = g(x)$,then $g_2(x) = 0$. Either way we have $g_2 \in [0, f_2]$. This shows that for any $g \in [0, f_1 + f_2]$,

$$\alpha(g) = \alpha(g_1) + \alpha(g_2) \leqslant \alpha_+(f_1) + \alpha_+(f_2).$$

Hence

$$\alpha_+(f_1 + f_2) = \alpha_+(f_1) + \alpha_+(f_2), \;\; \forall f_1, f_2 \in \boldsymbol{F}_+.$$

Observe next that if

$$\forall f, g, u, v \in \boldsymbol{F}_+, \;\; f - g = u - v \Rightarrow \alpha_+(f) - \alpha_+(g) = \alpha_+(u) - \alpha_+(v). \tag{19.6.3}$$

Indeed

$$f - g = u - v \Rightarrow f + v = u + g \Rightarrow \alpha_+(f) + \alpha_+(v) = \alpha_+(u) + \alpha_+(g)$$
$$\Rightarrow \alpha_+(f) - \alpha_+(g) = \alpha_+(u) - \alpha_+(v).$$

Let us show that

$$\alpha_+(f + g) = \alpha_+(f) + \alpha_+(g), \;\; \forall f, g \in \boldsymbol{F}.$$

We have to show that

$$\alpha_+\big((f + g)_+ \big) - \alpha_+\big((f + g)_- \big) = \alpha_+(f_+ + g_+) - \alpha_+(f_- + g_-).$$

This follows from (19.6.3) and the equality

$$(f + g)_+ - (f + g)_- = f + g = f_+ + g_+ - (f_- + g_-).$$

This proves that $\alpha : \boldsymbol{F} \to \mathbb{R}$ is additive. It is also homogeneous. Indeed, if $c > 0$, then

$$cf = (cf) + -(cf)- = (cf) + -(cf)- = cf_+ - cf_- = cf_+ - cf_-,$$

while if $c < 0$, then

$$cf = (cf) + -(cf)- = -cf_- + cf_+.$$

In both cases, invoking (19.6.3) we deduce $\alpha(cf) = c\alpha(f)$. Note that

$$\big|\, \alpha_+(f) \,\big| \leqslant \alpha_+(f_+) + \alpha_+(f_-) = \alpha_+(|f|) \leqslant A_\alpha \|f\|.$$

By definition $\alpha_+(f) \geqslant \alpha(f)$, $\forall f \in \boldsymbol{F}_+$ so that

$$\alpha_-(f) = \alpha_+(f) - \alpha(f) \geqslant 0, \;\; \forall f \in \boldsymbol{F}_+$$

so that $\alpha_- = \alpha_+ - \alpha$ is also a positive linear functional. □

Suppose that $\alpha \in \boldsymbol{F}^*$. Theorem 19.6.2 implies that there exist $\mu_\pm \in \mathcal{M}(K)_+$ such that $L_{\mu_\pm} = \alpha_\pm$ so that

$$L_{\mu_+ - \mu_-} = \alpha.$$

This completes the existence part of Theorem 19.6.9.

The map

$$\mathcal{M}(K) \ni \mu \mapsto L_\mu \in \boldsymbol{F}^*$$

is linear and onto. To prove that it is injective it suffices to show that its kernel is trivial, i.e., $L_\mu = 0$ iff μ. We will achieve this by proving that

$$\|L_\mu\|_* = \|\mu\| := |\mu|[\,K\,].$$

Clearly, $\forall f \in \boldsymbol{F}$

$$\left|\, L_\mu[f] \,\right| = \left| \int_K f d\mu_+ - \int_K f d\mu_- \right| \leqslant \left| \int_K f d\mu_+ \right| + \left| \int_K f d\mu_- \right|$$

$$\leqslant \int_K |f| d\mu_+ + \int_K |f| d\mu_- \leqslant \|f\| \mu_+[\,K\,] + \|f\| \mu_-[\,K\,] = \|f\| \cdot |\mu|[\,K\,],$$

so that

$$\|L_\mu\|_* \leqslant \|\mu\| = \mu_+[\,B_+\,] + \mu_-[\,B_-\,].$$

To prove that we have equality, consider a Hahn decomposition of K determined by μ, $K = B_+ \cup B_-$ where $B_\pm$ are disjoint Borel sets, B_+ is μ-positive and B_- is μ-negative. Then

$$\mu_\pm[\,f\,] = \mu[\,\boldsymbol{I}_{B_\pm} f\,],\;\; \mu[\,f\,] = \mu_+[\,\boldsymbol{I}_{B_+} f\,] - \mu_-[\,\boldsymbol{I}_{B_-} f\,]$$

and

$$\|\mu\| = \mu[\,\boldsymbol{I}_{B_+} - \boldsymbol{I}_{B_-}\,].$$

Denote by $\mathscr{C}$ the family of closed subsets of K. We deduce from Exercise 19.19 that

$$\mu_\pm[\,B_\pm\,] = \sup_{\substack{C_\pm \subset B_\pm \\ C_\pm \in \mathscr{C}}} \mu_\pm[\,C\,] = \sup_{\substack{C_\pm \subset B_\pm \\ C_\pm \in \mathscr{C}}} \mu[\,C\,].$$

Now choose sequences $g_n^\pm \in C(K)$ such that $g_n^\pm \to \boldsymbol{I}_{B_\pm}$ in $L^1(K, \mu_\pm)$. Set

$$f_n^\pm = \max\big(\min(g_n^\pm, 1), 0\big).$$

Then $f_n^\pm \in C(K)$ and Exercise 19.61 shows that $f_n^\pm$ converge in $L^1(K, \mu_\pm)$ to

$$\max\big(\min(\boldsymbol{I}_{B_\pm}, 1), 0\big) = \boldsymbol{I}_{B_\pm}.$$

A subsequence of $f_n^\pm$ converges almost everywhere to $\boldsymbol{I}_{B_\pm}$. For simplicity, assume $f_n^\pm \to \boldsymbol{I}_{B_\pm}$ a. e.. On the other hand, $0 \leqslant f_n^\pm \leqslant 1$ and $B_+ \cap B_- = \varnothing$ so that

$$-1 \leqslant f_n^+ - f_n^- \leqslant 1,\;\; \|f_n^+ - f_n^-\|_{C(K)} \to 1.$$

Observe that

$$\|L_\mu\|_* \cdot \|f_n^+ - f_n^-\|_{C(K)} \geqslant \mu[\,f_n^+ - f_n^-\,] = \mu_+[\,f_n^+\,] + \mu_-[\,f_n^-\,] - \big(\mu_-[\,f_n^+\,] + \mu_+[\,f_n^-\,]\big).$$

Note that

$$\mu_+[\,f_n^+\,] + \mu_-[\,f_n^-\,] \to \mu_+[\,B_+\,] + \mu_-[\,B_-\,] = |\mu|[\,K\,].$$

From the Dominated Convergence theorem we deduce

$$\mu_\pm\big[\, f_n^\mp \,\big] \to \mu_\pm\big[\, \boldsymbol{I}_{B_\mp} \,\big] = 0.$$

If we write $f_n = f_n^+ - f_n^-$ we deduce

$$\|L_\mu\|_* = \|L_\mu\|_* \lim_{n\to\infty} \|f_n\|_{C(K)} \geqslant \lim_{n\to\infty} \|L_\mu(f_n) = |\mu|\big[\, K \,\big] = \|\mu\|.$$

This proves that $\|L_\mu\|_* = \|\mu\|$. □

Remark 19.6.11. The space $\mathcal{M}(K)$ is a normed vector space with norm $\|\mu\| = |\mu|\big[\, K \,\big]$. The metric induced by this norm is usually referred to as the *variation distance.*

Theorem 19.6.9 can be rephrased more compactly by stating that the natural map

$$\mathcal{M}(K) \ni \mu \mapsto L_\mu \in C(K)^*$$

is a bijective isometry of normed spaces. □

Corollary 19.6.12. *Let (K, d) be a compact metric space. A family $\mathcal{F} \subset C(K)$ spans a subspace dense in $C(K)$ if and only if the only finite signed measure μ satisfying*

$$\mu\big[\, f \,\big] = 0, \;\; \forall f \in \mathcal{F}$$

is the trivial measure.

Proof. Apply Corollary 17.1.55 to the normed space $X = C(K)$ and the subspace $Z = \operatorname{span}(\mathcal{F})$. □

19.6.2. The dual of L^p. Fix a measured space $(\Omega, \mathcal{S}, \mu)$ and $p \in [1, \infty)$. Set $q := p^* = \frac{p}{p-1}$. The goal of this subsection is to give an explicit description of the dual of the Banach space $X := L^p(\Omega, \mu)$. For simplicity we denote by $\| - \|_r$ the norm of $L^r(\Omega, \mu)$, $r \in [1, \infty]$.

For any $f \in \mathcal{L}^p(\Omega, \mathcal{S}, \mu)$ and any $g \in \mathcal{L}^q(\Omega, \mathcal{S}, \mu)$, Hölder's inequality shows that the function fg is integrable and

$$\left| \int_\Omega f g d\mu \right| \leqslant \|g\|_q \|f\|_p.$$

Equivalently, we have a linear map $\boldsymbol{\alpha}_g : \mathcal{L}^p(\Omega, \mathcal{S}, \mu) \to \mathbb{R}$

$$\boldsymbol{\alpha}_g(f) = \int_\Omega f g d\mu.$$

Note that if $f = f'$ μ-a.e., then $\alpha_g(f) = \boldsymbol{\alpha}_g(f')$ so α_g induces a linear map

$$\alpha_g : L^p(\Omega, \mathcal{S}, \mu) \to \mathbb{R}.$$

Hölder's inequality implies that

$$\big|\, \boldsymbol{\alpha}_g(f) \,\big| \leqslant \|g\|_q \cdot \|f\|_p, \;\; \forall f \in L^p(\Omega, \mu), \tag{19.6.4}$$

which shows that α_g is a continuous linear functional. We have thus produced a map

$$\mathcal{L}^q(\Omega, \mathcal{S}, \mu) \ni g \mapsto \boldsymbol{\alpha}_g \in L^p(\Omega, \mu)^*.$$

Observe that if $g = g'$ μ-a.e., then $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}_{g'}$ so the above map induces a map

$$L^q(\Omega, \mathcal{S}, \mu) \ni g \mapsto \boldsymbol{\alpha}_g \in L^p(\Omega, \mu)^*.$$

Note that $\boldsymbol{\alpha}_{g+g'} = \boldsymbol{\alpha}_g + \boldsymbol{\alpha}_{g'}$ and $\boldsymbol{\alpha}_{cg} = c\boldsymbol{\alpha}_g$, $\forall g, g' \in L^1(\Omega, \mu)$, $c \in \mathbb{R}$ so the map

$$L^q(\Omega, \mathcal{S}, \mu) \ni g \mapsto \boldsymbol{\alpha}_g \in L^p(\Omega, \mu)^*$$

is linear. Set $X := L^p(\Omega, \mu)$, denote by $\| - \|$ the L^p norm on X and by $\| - \|_*$ the norm on the dual.

The inequality (19.6.4) shows that $\|\boldsymbol{\alpha}_g\|_* \leqslant \|g\|_{L^q}$, so the map $g \mapsto \boldsymbol{\alpha}_g$ is continuous. We have another representation theorem also due to F. Riesz.

Theorem 19.6.13 (Riesz representation)**.** *Suppose that $(\Omega, \mathcal{S}, \mu)$ is a sigma-finite measured space. Let $p \in (1, \infty)$, $q = p^*$,*

$$(X, \| - \|) = \big(L^p(\Omega, \mu), \| - \|_{L^p} \big).$$

The map

$$L^{p^*}(\Omega, \mu) \ni g \mapsto \boldsymbol{\alpha}_g \in X^*$$

is continuous, bijective and an isometry, i.e., $\|\boldsymbol{\alpha}_g\|_* = \|g\|_{L^{p^*}}$, $\forall g \in L^q(\Omega, \mathcal{S}, \mu)$.

Proof. We first prove that

$$\|\boldsymbol{\alpha}_g\|_* \geqslant \|g\|_q.$$

This will prove that the map is an isometry and, in particular, that it is injective.

Let $g \in L^q(\Omega, \mathcal{S}, \mu)$. Decompose as usual $g = g_+ - g_-$. Note that for any $\omega \in \Omega$, we have either $g(\omega) = g_+(\omega)$ or $g(\omega) = -g_-(\omega)$ and

$$|g(\omega)|^q = g_+(\omega)^q + g_-(\omega)^q.$$

Consider the functions $f_\pm = g_\pm^{1/(p-1)}$, $f = f_+ - f_- \in L^p$. Then

$$f, f_\pm \in L^p \;\text{ and }\; f, f_\pm g_\mp = 0, \;\; |f| = |g|^{\frac{1}{p-1}}$$

so that

$$\|\boldsymbol{\alpha}_g\|_* \cdot \|f\|_{L^p} \geqslant \boldsymbol{\alpha}_g(f) = \int_\Omega \big(g_+^{1/(p-1)} - g_-^{1/(p-1)} \big) \big(g_+ - g_- \big) d\mu$$

$$= \int_\Omega \big(g_+^{p/(p-1)} + g_-^{p/(p-1)} \big) d\mu$$

$$= \int_\Omega |g|^q d\mu = \|g\|^q_{L^q} = \|g\|_g \cdot \|g\|^{q-1}_{L^q} = \|g\|_q \left(\int_\Omega \big(|g|^{1/(p-1)} \big)^p d\mu \right)^{\frac{q-1}{q}}$$

$((q-1)/q = 1/p$, $|g|^{1/(p-1)} = |f|)$

$$= \|g\|_{L^q} \cdot \|f\|_{L^p}.$$

Hence

$$\|\boldsymbol{\alpha}_g\|_* \geqslant \|g\|_{L^q}.$$

Let us record here a simple consequence of the above computations.

$$\forall r \in (1,\infty),\;\; \forall g \in L^r_+(\Omega,\mathcal{S},\mu),\;\; f = g^{\frac{1}{r^*-1}} \Rightarrow \int_\Omega gf d\mu = \|g\|_{L^r}\cdot\|f\|_{L^{r*}} \tag{19.6.5}$$

We have to prove that for any continuous linear function $\xi \in X^*$, there exists $g \in L^q(\Omega,\mu)$ such that $\xi = \boldsymbol{\alpha}_g$. We discuss two cases.

A. *The measure μ is finite.* In this case $\boldsymbol{I}_S \in L^q$, $\forall S \in \mathcal{S}$. Since $\xi(f) = 0$ if $f = 0$ μ-a.e. we deduce

$$\xi(\boldsymbol{I}_S) = 0 \;\text{ if } \mu\big[\,S\,\big] = 0$$

Define $\mu^\xi : \mathcal{S} \to [0,\infty)$,

$$\mu^\xi\big[\,S\,\big] := \xi(\boldsymbol{I}_S).$$

From the equality

$$\boldsymbol{I}_{S_0\cup S_1} = \boldsymbol{I}_{S_0} + \boldsymbol{I}_{S_1},$$

if S_0, S_1 are disjoint, we deduce that μ^ξ is finitely additive. If $(S_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence in $\mathcal{S}$ with union S_∞ then

$$0 \leqslant \boldsymbol{I}_{S_n} \nearrow \boldsymbol{I}_{S_\infty}$$

and we deduce that $\boldsymbol{I}_{S_n} \to \boldsymbol{I}_{S_\infty}$ in $L^p(\Omega,\mathcal{S},\mu)$. Hence

$$\mu^\xi\big[\,S_n\,\big] = \xi\big(\,\boldsymbol{I}_{S_n}\,\big) \to \xi\big(\,\boldsymbol{I}_{S_\infty}\,\big) = \mu^\xi\big[\,S_\infty\,\big].$$

This proves that μ^ξ is a signed measure on $\mathcal{S}$. Moreover, if $\mu\big[\,S\,\big] = 0$

$$\mu^\xi\big[\,S\,\big] = \xi\big(\,\boldsymbol{I}_S\,\big) = 0$$

so that $\mu^\xi \ll \mu$.

The Radon-Nikodym theorem implies that there exists $u \in L^1(\Omega,\mathcal{S},\mu)$ such that $\mu^\xi\big[\,S\,\big] = \mu_u\big[\,S\,\big]$, $\forall S \in \mathcal{S}$, i.e.,

$$\xi\big(\,\boldsymbol{I}_S\,\big) = \int_\Omega u\boldsymbol{I}_S d\mu = \mu\Big[\,u\boldsymbol{I}_S\,\Big],\;\; \forall S \in \mathcal{S}. \tag{19.6.6}$$

Denote $\mathscr{E}_+(\mathcal{S},\mu)$ the space of nonnegative elementary functions. Hence

$$\xi\big(\,v\,\big) = \mu\big[\,uv\,\big],\;\; \forall v \in \mathscr{E}_+(\Omega,\mathcal{S}).$$

By linearity, we deduce

$$\xi\big(\,v\,\big) = \mu\big[\,uv\,\big],\;\; \forall v \in \mathscr{E}(\Omega,\mathcal{S}). \tag{19.6.7}$$

We claim that

$$\|u\|_{L^q} < \infty. \tag{19.6.8}$$

Theorem 19.6.13 follows immediately from this fact. Indeed, the space of elementary functions $\mathscr{E}(\Omega,\mathcal{S})$ is dense in $L^p(\Omega,\mu)$ and, since ξ is continuous we see that (19.6.7) extends to all $v \in L^p(\Omega,\mu)$. This is precisely Theorem 19.6.13.

Proof of (19.6.8) Since $\xi : L^p(\Omega,\mathcal{S},\mu) \to \mathbb{R}$ is continuous we deduce that there exists $C_\xi > 0$ such that

$$\big|\,\mu[uv]\,\big| = \big|\,\xi\big(\,v\,\big)\,\big| \leqslant C_\xi\|v\|_{L^p},\;\; \forall v \in \mathscr{E}_+(\Omega,\mathcal{S}). \tag{19.6.9}$$

Lemma 19.6.14. *Let $u \in \mathcal{L}^1_+(\Omega, \mathcal{S}, \mu)$. Then*

$$\|u\|_{L^q} = C_u := \sup_{v \in \mathscr{E}_+(\mathcal{S},\mu)\setminus 0} Q(u,v), \;\; Q(u,v) := \frac{\mu\big[\,uv\,\big]}{\|v\|_p}.$$

Proof. Suppose first that $u \in L^q(\Omega, \mu)$. Hölder's inequality shows that the map

$$L^p(\Omega,\mu)\setminus\{0\} \ni v \mapsto \mu\big[\,uv\,\big] \in \mathbb{R}$$

is continuous and $C_u \leqslant \|u\|_{L^q}$. Let $v := u^{1/(q-1)} \in L^p_+$. Then, according to (19.6.5), we have $\mu\big[\,uv\,\big] = \|u\|_{L^q}\|v\|_{L^p}$ so that

$$Q(u,v) = \|u\|_{L^q},$$

If v were an elementary function, then we could conclude $C_u \geqslant \|u\|_{L^q}$. Fortunately, the elementary functions $D_n[v] \in \mathscr{E}_+$ approximate v well, $D_n[v] \to v$ in L^p. We deduce that

$$Q\big(\,u, D_n[v]\,\big) \leqslant C_u,$$

and

$$\|u\|_{L^q} = Q\big(\,u,v\,\big) = \lim_{n\to\infty} Q\big(\,u, D_n[v]\,\big) \leqslant C_u.$$

Suppose now that $\|u\|_{L^q} = \infty$. For $k \in \mathbb{N}$, we set $u_k = \max(u,k)$. Then $u_k \leqslant u$,

$$\|u_k\|_{L^q} = \sup_{\substack{v\in\mathscr{E}_+\setminus 0,\\ \|v\|_{L^p}=1}} \mu\big[\,u_k v\,\big] = \sup_{\substack{v\in\mathscr{E}_+,\\ \|v\|_{L^p}=1}} \mu\big[\,u_k v\,\big] \leqslant \sup_{\substack{v\in\mathscr{E}_+,\\ \|v\|_{L^p}=1}} \mu\big[\,uv\,\big] = C_u.$$

Then,

$$\infty = \|u\|_{L^q} = \lim_{k\to\infty}\|u_k\|_{L^q} \leqslant C_u \leqslant \infty.$$

Hence, in this case we also have $\|u\|_{L^q} = C_u$. □

Consider the function u defined by (19.6.6). It has a decomposition $u = u_+ - u_-$. Set

$$\Omega_\pm := \big\{\,u_\pm > 0\,\big\}.$$

For any $v \in \mathscr{E}_+$ we have $v\boldsymbol{I}_{\Omega_\pm} \in \mathscr{E}_+$ and $uv\boldsymbol{I}_{\Omega_\pm} = u_\pm v$. Using (19.6.9) we deduce that for any $v \in \mathscr{E}_+$

$$\big|\,\mu\big[\,u_\pm v\,\big]\,\big| = \Big|\mu\Big[\,u\boldsymbol{I}_{\Omega_\pm}v\,\Big]\Big| \leqslant C_\xi\,\big\|\boldsymbol{I}_{\Omega_\pm}v\big\|_{L^p} \leqslant C_\xi\|v\|_{L^p}.$$

Using Lemma 19.6.14 we deduce $\|u_\pm\|_{L^p} \leqslant C_\xi < \infty$. This proves (19.6.8) and thus Theorem 19.6.13 when μ is a finite measure.

B. *The measure μ is sigma-finite.* Fix a nondecreasing sequence of measurable sets

$$\Omega_1 \subset \Omega_2 \subset \cdots$$

such that $\mu\big[\,\Omega_n\,\big] < \infty$ and

$$\Omega = \bigcup_{n\in\mathbb{N}} \Omega_n.$$

Fix $C_\xi > 0$ such that

$$\big|\,\xi\big(\,v\,\big)\,\big| \leqslant C_\xi\|v\|_{L^p}, \;\; \forall c \in L^p(\Omega,\mu).$$

We can view $L^p(\Omega_n, \mathcal{S}\cap\Omega_n,\mu)$ as a subspace of $L^p(\Omega,\mathcal{S},\mu)$: a function v on Ω_n can be viewed as a function $\widehat{v}$ on Ω vanishing outside Ω_n Equivalently, one can view $\widehat{v}$ as the extension by 0 of v to a function on Ω.

The continuous linear functional $\xi : L^p(\Omega,\mathcal{S},\mu)\to\mathbb{R}$ induces continuous linear functionals $\xi_n : L^p(\Omega_n)\to\mathbb{R}$,

$$\xi_n\big(\,v\,\big)=\xi\big(\,\widehat{v}\,\big),\;\;\forall v\in L^p(\Omega_n,\mathcal{S}\cap\Omega_n,\mu).$$

Moreover, we deduce from **A** that there exists $u_n\in L^q(\Omega_n)$ such that

$$\|u_n\|_{L^q(\Omega_n)}\leqslant C_\xi$$

and

$$\xi_n\big(\,v\,\big)=\int_{\Omega_n}u_nvd\mu,\;\;\forall v\in L^p(\Omega_n,\mathcal{S}\cap\Omega_n,\mu),$$

Note that since $L^p(\Omega_n,\mathcal{S}\cap\Omega_n,\mu)\subset L^p(\Omega_{n+1},\mathcal{S}\cap\Omega_{n+1},\mu)$ we deduce

$$u_{n+1}\big|_{\Omega_n}=u_n\;\;\mu-\text{a. e.}\,.$$

Modifiying the functions u_n on a negligible set we can assune that the above equality holds everywhere for every n. Define

$$\widehat{u}:\Omega\to\mathbb{R},\;\;\widehat{u}(\omega)=u_n(\omega)\;\;\text{if }\omega\in\Omega_n$$

Clearly $\widehat{u}(\omega)$ does not depend on the choice of n in its definition. Denote by $\widehat{u}_n$ the extension by 0 of u_n to a function on Ω_n. Equivalently, $\widehat{u}_n=\widehat{u}\boldsymbol{I}_{\Omega_n}$. We have

$$\lim_{n\to\infty}\widehat{u}_n(\omega)=\widehat{u}(\omega),\;\;\forall\omega\in\Omega,$$

and

$$\big|\widehat{u}_n\big|\leqslant\big|\widehat{u}_{n+1}\big|.$$

The Monotone Convergence Theorem implies

$$\int_\Omega\big|\widehat{u}\big|^qd\mu=\lim_{n\to\infty}\int_{\Omega_n}\big|\widehat{u}_n\big|^qd\mu\leqslant C_\xi.$$

Hence $\widehat{u}\in L^q(\Omega,\mathcal{S},\mu)$. Moreover, given $v\in L^p(\Omega,\mathcal{S},\mu)$, the sequence $v_n=v\boldsymbol{I}_{\Omega_n}$ converges in L^p to v and we have

$$\xi\big(\,v\,\big)=\lim_{n\to\infty}\xi\big(\,v_n\,\big)=\lim_{n\to\infty}\int_\Omega\widehat{u}_nv_nd\mu=\int_\Omega\widehat{u}vd\mu,$$

where at the last step we used the Dominated Convergence Theorem. This completes the proof of Theorem 19.6.13.

□

19.7. Exercises

Exercise 19.1. Let C denote the cube $[0,1]^n \subset \mathbb{R}^n$, $n \in \mathbb{N}$. Denote by $\mathcal{J}_C$ the collection of Jordan measurable subsets of C; see Definition 15.1.27. Prove that $\mathcal{J}_C$ is an algebra of subsets of C, but it *is not* a sigma-algebra. □

.

Exercise 19.2. Suppose that $(\Omega, \mathcal{S})$ is a measurable space and $(S_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{S}$. Let S be the subset of Ω consisting of the points ω that belong to infinitely many of the sets S_n. Prove that $S \in \mathcal{S}$. **Hint.** Use Remark 19.1.4. □

Exercise 19.3. Prove that the sigma-algebra $\mathcal{B}_{\mathbb{R}}$ of Borel subsets of $\mathbb{R}$ is generated by the collection of intervals

$$\Big[\,\frac{k}{2^n}, \frac{k+1}{2^n}\,\Big],\ \ k \in \mathbb{Z},\ \ n \in \mathbb{N}.$$ □

Exercise 19.4. Construct a bijection $\Phi : [0,1) \to (0,1)$ with the property that $B \subset (0,1)$ is Borel if and only if $\Phi^{-1}(B)$ is a Borel set. □

Exercise 19.5. Let $\mathcal{S}_1, \mathcal{S}_2$ be two sigma-algebras of subsets of Ω. Prove that the following are equivalent.

(i) $\mathcal{S}_1 \cup \mathcal{S}_2$ is a sigma algebra.

(ii) Either $\mathcal{S}_1 \subset \mathcal{S}_2$ or $\mathcal{S}_2 \subset \mathcal{S}_1$.

□

Exercise 19.6. Let Ω be a set. A collection $\mathcal{S}$ of subsets of Ω is called a *semiring of subsets* if

- $\varnothing \in \mathcal{S}$,
- for all $A, B \in \mathcal{S}$, $A \cap B \in \mathcal{S}$, and $A \backslash B$ is a union of finitely many and disjoint subsets $S_1, \dots, S_n \in \mathcal{S}$.

A collection $\mathcal{S}$ is called a *ring* if for any $A, B \in \mathcal{S}$, $A \cap B, A \cup B, A\backslash B \in \mathcal{S}$.

(i) Prove that the collection of subsets of $\mathbb{R}$ of the form $(a,b]$, $-\infty \leqslant a \leqslant b < \infty$, is a semiring.

(ii) Let $\mathcal{S}$ be a semiring of subsets of Ω and let $\mathcal{R}$ denote the collection of all finite disjoint unions of sets in $\mathcal{S}$. Show that $\mathcal{R}$ is a ring. It is called the *ring generated by the semiring* $\mathcal{S}$.

(iii) Suppose that $\mathcal{R}$ is a ring of subsets of Ω. Prove that the collection

$$\mathcal{A} = \mathcal{R} \cup \big\{\, \Omega\backslash R;\ \ R \in \mathcal{R} \,\big\}.$$

is an algebra of subsets of Ω.

□

Exercise 19.7. Suppose that (X,d) is a metric space. For every continuous function $f : X \to \mathbb{R}$ we set

$$H_f := \big\{\, x \in X;\;\; f(x) \leqslant 0 \,\big\}.$$

Prove that the sigma-algebra generated by the sets H_f, $f \in C(X)$, coincides with the Borel sigma-algebra of X defined in Example 19.1.3(viii). **Hint.** Use Proposition 17.1.27. □

Exercise 19.8. Let Ω be a set and $(A_n)_{n\in\mathbb{N}}$ be a partition of Ω, i.e.,

$$\Omega = \bigcup_{n\in\mathbb{N}} A_n,\;\; A_n \cap A_m = \varnothing,\;\; \forall m \neq n.$$

Denote by $\mathcal{S}$ the sigma-algebra generated by the sets $(A_n)_{n\in\mathbb{N}}$. Describe all the $\mathcal{S}$-measurable functions $f : \Omega \to \mathbb{R}$. □

Exercise 19.9. Suppose that Ω is a set and $f_i : \Omega \to \mathbb{R}$, $i \in I$, is a family of functions. For each $i \in I$ we denote by $\mathcal{S}_i$ the sigma-algebra generated by f_i,

$$\mathcal{S}_i = f_i^{-1}\big(\mathcal{B}_{\mathbb{R}}\big).$$

We set

$$\mathcal{S} := \bigvee_{i\in I} \mathcal{S}_i.$$

(i) Prove that for any $i \in I$ the function f_i is $(\mathcal{S}, \mathcal{B}_{\mathbb{R}})$-measurable.

(ii) Suppose that $\mathcal{S}' \subset 2^{\Omega}$ is a sigma-algebra such that all the functions f_i are $(\mathcal{S}', \mathcal{B}_{\mathbb{R}})$-measurable. Prove that $\mathcal{S}' \supset \mathcal{S}$.

□

Exercise 19.10. Suppose that (X,d) is a *compact* metric space. We denote by F the Banach space $C(X)$ equipped with the sup-norm. We denote by $\mathcal{B}_F$ the Borel sigma-algebra of F. For each $x \in X$ we define $E_x : F \to \mathbb{R}$, $E_x(f) = f(x)$, for any continuous function $f : X \to \mathbb{R}$. We set (see Example 19.1.3)

$$\mathcal{S}_x = E_x^{-1}\big(\mathcal{B}_{\mathbb{R}}\big),\;\; x \in X,\;\; \mathcal{S} = \bigvee_{x\in X} \mathcal{S}_x.$$

(i) Prove that $E_x : F \to \mathbb{R}$ is a continuous, $\forall x \in X$.

(ii) Prove that $\mathcal{B}_F = \mathcal{S}$. **Hint.** Show that if $S \subset X$ is a dense subset in X, then

$$\|f\| = \sup_{s\in S} |f(s)| = \sup_{s\in S} \big| E_s(f) \big|.$$

Next use Corollary 17.3.6.

(iii) For $n \in \mathbb{N}$, $\vec{r} \in \mathbb{R}^n$ and $x_1, \dots, x_n \in X$ we set

$$C_{x_1,\dots,x_n}\big(\vec{r}\big) := \big\{ f \in C(X);\;\; f(x_i) \leqslant r_i,\;\; \forall i = 1, \dots, n \,\big\} \subset F.$$

Prove that the collection

$$\mathscr{C} := \big\{ C_{x_1,\dots,x_n}\big(\vec{r}\big);\;\; n \in \mathbb{N},\; \vec{r} \in \mathbb{R}^n,\;\;\; x_1, \dots x_n \in X \,\big\}$$

is a π-system that generates $\mathcal{B}_F$. **Hint.** Observe that

$$C_{x_1,\dots,x_n}(\vec{r}) = \bigcap_{i=1}^{n} \{E_{x_i} \leqslant r_i\},$$

and then use (ii).

(iv) Suppose that $(\Omega, \mathcal{A})$ is a measurable space and $T : \Omega \to F$ is a map

$$\Omega \ni \omega \mapsto T_\omega \in F.$$

Prove that T is $(\mathcal{A}, \mathcal{B}_F)$-measurable if and only if for any $x \in X$ the function

$$T^x : \Omega \to \mathbb{R}, \;\; \omega \mapsto T_\omega(x)$$

is measurable. **Hint.** Use (iii) and Proposition 19.1.13.

□

Exercise 19.11. Prove Proposition 19.1.19(iii). □

Exercise 19.12. Prove that a monotone function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable, i.e., for any Borel subset $B \subset \mathbb{R}$ the preimage $f^{-1}(B)$ is also Borel measurable. □

Exercise 19.13. Let Ω be a set and $\mathcal{S}$ a semiring of subsets of Ω; see Exercise 19.6. Fix an additive measure on $\mathcal{S}$, i.e., a function $\mu : \mathcal{S} \to [0, \infty]$ such that, for any $A, B \in \mathcal{S}$ such that $A \cup B \in \mathcal{S}$, $\mu\big[\, A \cup B \,\big] + \mu\big[\, A \cap B \,\big] = \mu\big[\, A \,\big] + \mu\big[\, B \,\big]$.

(i) Prove that there exists a unique additive measure $\bar{\mu}$ on the ring generated by $\mathcal{S}$ such that $\bar{\mu}\big[\, S \,\big] = \mu\big[\, S \,\big]$, $\forall S \in \mathcal{S}$. □

(ii) Suppose that μ is conditionally sigma-additive meaning that if $(S_n)_{n \geqslant 1}$ is a sequence of disjoint subsets of $\mathcal{S}$ whose union is a set S also in $\mathcal{S}$, then

$$\mu\big[\, S \,\big] = \sum_{n \geqslant 1} \mu\big[\, S_n \,\big].$$

Prove that the extension $\bar{\mu}$ of μ postulated in (i) is also conditionally sigma-additive.

□

Exercise 19.14. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space. Prove that for any sequence $(S_n)_{n \in \mathbb{N}}$ in $\mathcal{S}$ we have

$$\mu\Big[\, \bigcup_{n \in \mathbb{N}} S_n \,\Big] \leqslant \sum_{n \in \mathbb{N}} \mu\big[\, S_n \,\big].$$

This inequality is known as the *union bound*. **Hint.** Use Proposition 19.1.32. □

Exercise 19.15. Prove Proposition 19.1.36. □

Exercise 19.16. Consider the measure $\mu : \big(\mathbb{N}, 2^{\mathbb{N}}\big) \to [0,\infty)$ determined by the conditions

$$\mu\big[\,\{n\}\,\big] = \frac{1}{2^n},\quad \forall n \in \mathbb{N}.$$

Consider the function $f : \big(\mathbb{N}, 2^{\mathbb{N}}\big) \to \big(\mathbb{R}, \mathcal{B}_{\mathbb{R}}\big)$, $f(n) = \cos n\pi$, $\forall n \in \mathbb{N}$. Describe the measure $\nu = f_{\#}\mu : \mathcal{B}_{\mathbb{R}} \to [0,\infty)$, $\nu\big[\,B\,\big] = \mu\big[\,f^{-1}(B)\,\big]$, $\forall B \in \mathcal{B}_{\mathbb{R}}$. □

Exercise 19.17. Suppose that $(\mu_n)_{n\in\mathbb{N}}$ is a sequence of measures on the measurable space $(\Omega, \mathcal{S})$ and $(w_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative real numbers. Prove that the sum

$$\mu = \sum_{n\in\mathbb{N}} w_n\mu_n : \mathcal{S} \to [0,\infty],\quad \mu\big[\,S\,\big] = \sum_{n\in\mathbb{N}} w_n\mu_n\big[\,S\,\big],\quad \forall S$$

is also a measure on $(\Omega, \mathcal{S})$. **Warning.** The sigma-additivity of μ is not as obvious as it appears. □

Exercise 19.18. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space and $(S_n)_{n\geqslant 1}$ is a decreasing sequence of measurable sets

$$S_1 \supset S_2 \supset S_3 \supset \cdots .$$

Prove that if $\mu\big[\,S_1\,\big] < \infty$, then

$$\mu\left[\,\bigcap_{n\geqslant 1} S_n\,\right] = \lim_{n\to\infty} \mu\big[\,S_n\,\big].$$

Show using a concrete example that the above equality need not hold if $\mu\big[\,S_1\,\big] = \infty$. □

Exercise 19.19. Suppose that μ is a finite Borel measure on the metric space (X,d). Denote by $\mathscr{C}$ the collection of Borel subsets S of X satisfying the regularity property: for any $\varepsilon > 0$ there exists a closed subset $C_\varepsilon \subset S$ and an open subset $\mathcal{O}_\varepsilon \supset S$ such that

$$\mu\big[\,\mathcal{O}_\varepsilon \backslash C_\varepsilon\,\big] < \varepsilon.$$

(i) Show that $S \in \mathscr{C} \Rightarrow S^c := X\backslash S \in \mathscr{C}$.

(ii) Show that any closed set belongs to $\mathscr{C}$. **Hint.** Use Proposition 17.1.27.

(iii) Show that $\mathscr{C}$ is a π-system.

(iv) Show that $\mathscr{C}$ is a λ-system.

(v) Show that $\mathscr{C}$ coincides with the family of Borel subsets.

□

Exercise 19.20. Suppose that $\mu : \big(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}\big) \to [0,\infty]$ is a *finite* measure defined on the Borel sigma-algebra generated by the open subsets of $\mathbb{R}^n$. Prove that for any Borel set $B \subset \mathbb{R}^n$ and any $\varepsilon > 0$ there exists an open set $U \supset B$ and a compact set $K \subset B$ such that $\mu\big[\,U\backslash K\,\big] < \varepsilon$. **Hint.** Prove that any closed set in $\mathbb{R}^n$ is the union of countably many compact subsets. Conclude using Exercise 19.19. □

Exercise 19.21. Fix a set Ω and suppose that $\mathcal{M} \subset \Omega$ is a class of models (see Definition 19.2.2) that is also a *semiring* and $\Omega \in \mathcal{M}$., i.e., it satisfies the following additional properties: $\forall M_1,\ M_2 \in \mathcal{M}$, $M_1 \cap M_2 \in \mathcal{M}$ and $M_2 \backslash M_1$ is a disjoint union of finitely many elements in $\mathcal{M}$.[6]

Suppose that $\rho : \mathcal{M} \to [0, \infty)$ is a gauge on $\mathcal{M}$ satisfying the conditions

- If $M_1, M_2 \in \mathcal{M}$, and $M_1 \cup M_2 \in \mathcal{M}$, then
$$\rho\big[\, M_1 \cup M_2 \,\big] = \rho\big[\, M_1 \,\big] + \rho\big[\, M_2 \,\big] - \rho\big[\, M_1 \cap M_2 \,\big].$$
- If $M_n \in \mathcal{M}$, $\forall n \in \mathbb{N}$ and $\cup_n M_n \in \mathcal{M}$, then
$$\rho\Big[\, \bigcup_n M_n \,\Big] \leqslant \sum_n \rho\big[\, M_n \,\big].$$

Denote by μ_ρ the outer measure determined by ρ as in Proposition 19.2.3 and let $\mathcal{S}_\rho$ be the sigma algebra of μ_ρ-measurable sets; see Definition 19.2.1 .

(i) Prove that $\mathcal{M} \subset \mathcal{S}_\rho$.

(ii) Prove that $\mu_\rho\big[\, M \,\big] = \rho\big[\, M \,\big]$, $\forall M \in \mathcal{M}$.

Hint. Use the same strategy as in the proof of Theorem 19.2.5. □

Exercise 19.22. Suppose that (X, d) is a *separable* metric space. For any subset $S \subset X$ we set
$$\operatorname{diam}(S) := \sup_{x,y \in S} d(x, y).$$
Consider the collection of subsets
$$\mathcal{M}_\delta := \big\{\, S \subset X;\ \ \operatorname{diam}(S) < \delta \,\big\}.$$
For $t \geqslant 0$ denote by χ_t^δ the outer measure obtained by using class of models $\mathcal{M}_\delta$ and the gauge function (see Proposition 19.2.3)
$$\rho_t : \mathcal{M}_\delta \to [0, \infty),\ \ \rho_t(S) = \operatorname{diam}(S)^t.$$
Note that $\chi_t^\delta \geqslant \chi_t^{\delta'}$, $\forall 0 < \delta \leqslant \delta'$. We set
$$\chi_t(S) = \sup_{\delta > 0} \chi_t^\delta = \lim_{\delta \searrow 0} \chi_t^\delta(S),\ \ \forall S \subset X.$$
Prove that χ_t is a metric outer measure (Definition 19.2.7). This shows that the restriction of χ_t to the Borel sigma-algebra of X is a measure. It is called the *t-dimensional Hausdorff measure.* □

Exercise 19.23. Suppose that $\mu : (\mathbb{R}, \mathcal{B}_\mathbb{R}) \to [0, \infty]$ is a measure on the sigma-algebra of Borel subsets of $\mathbb{R}$ satisfying the following conditions.

(i) For any $B \in \mathcal{B}_\mathbb{R}$ and any $r \in \mathbb{R}$, $\mu\big[\, B + r \,\big] = \mu\big[\, B \,\big]$, where $B + r := \{b + r;\ \ b \in B\}$.

(ii) $\mu\big[\, (0, 1] \,\big] = 1$.

[6] The collection of semiintervals $(a, b]$ $a, b \in \mathbb{R}$ is a semi-algebra.

Prove that μ coincides with the Lebesgue measure. **Hint.** Compute first $\mu\big[\,(0,1/n]\,\big]$, $n\in\mathbb{N}$. □

Exercise 19.24 (H. Steinhaus)**.** Suppose that $E\subset\mathbb{R}$ is Lebesgue measurable and

$$0<\boldsymbol{\lambda}\big[\,E\,\big]<\infty.$$

(i) Prove that for any $\varepsilon\in(0,1)$ there exists an interval J such that

$$\boldsymbol{\lambda}\big[\,E_J\,\big]>(1-\varepsilon)\boldsymbol{\lambda}\big[\,J\,\big]>0,\;\;E_J:=E\cap J$$

Hint. Prove that there exists a sequence of intervals $(J_n)_{n\in\mathbb{N}}$ that cover E such that

$$\sum_{n\in\mathbb{N}}\boldsymbol{\lambda}[\,J_n\,]<\boldsymbol{\lambda}[\,E\,]/(1-\varepsilon)$$

.

(ii) Prove that there exists $r>0$ such that the set

$$E-E:=\big\{\,x-y;\;\;x,y\in E\,\big\}$$

contains the interval $\big[-r,r\big]$.

Hint. Note that $\boldsymbol{\lambda}\big[\,E\cap(E+t)\,\big]>0\Rightarrow t\in E-E$. Fix $c\in(1/2,1)$ and choose an interval J such that $\boldsymbol{\lambda}[\,E_J\,]>c\boldsymbol{\lambda}[\,J\,]$. Prove that if $|t|<\boldsymbol{\lambda}[\,J\,]$, then $\boldsymbol{\lambda}\big[\,E_J\cap(E_J+t)\,\big]\geqslant(2c-1)\boldsymbol{\lambda}[\,J\,]-|t|$.

□

Exercise 19.25 (H. Steinhaus)**.** Suppose that $f:\mathbb{R}\to\mathbb{R}$ is a Borel measurable function such that

$$f(1)=1,\;\;f(x+y)=f(x)+f(y),\;\;\forall x,y\in\mathbb{R}.$$

(i) Prove that f is bounded on a set of positive Lebesgue measure. **Hint.** Look at the sets $\{|f|\leqslant n\}$, $n\in\mathbb{N}$.

(ii) Prove that f is bounded on some open interval containing 0. **Hint.** Use Exercise 19.24.

(iii) Prove that $f(x)=x$, $\forall x\in\mathbb{R}$. **Hint.** Prove first that $f(x)=x$, $\forall x\in\mathbb{Q}$. Conclude using (ii).

□

Exercise 19.26. Let $F:[0,1]\to[0,1]$, $F(x)=2x-\lfloor 2x\rfloor$, where $\lfloor r\rfloor$ denotes the integer part of the real number r.

(i) Prove that F is Borel measurable.

(ii) Let $I_0:=[1/3,2/3]$, $I_n:=F^{-1}(I_{n-1})$, $\forall n\in\mathbb{N}$. Draw pictures of the sets I_0,I_1,I_2 and then compute their Lebesgue measures.

(iii) Compute $F_{\#}\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ denotes the Lebesgue measure on $[0,1]$.

Hint. Draw the graph of the function $F(x)$ and use it to visualize $F^{-1}(I)$, where $I\subset[0,1]$ is an interval. □

Exercise 19.27. Suppose that $S\subset\mathbb{R}$ is Lebesgue measurable and $\boldsymbol{\lambda}\big[\,S\,\big]=0$. Prove that $\mathbb{R}\backslash S$ is dense in $\mathbb{R}$. □

Exercise 19.28 (E. Borel)**.** Set $\mathbb{B} := \{0,1\}$ ($\mathbb{B}$ is the set of *bits*) and denote by $\mathbb{X}$ the space of functions $f : \mathbb{N} \to \mathbb{B}$. We define a metric

$$d : \mathbb{X} \times \mathbb{X} \to [0,\infty),\ \ d\big(f, g\big) = \sum_{n\in\mathbb{N}} \frac{|f(n) - g(n)|}{2^n}.$$

Given $m \in \mathbb{N}$ and a subset $S \subset \mathbb{B}^m$ we set

$$C_S := \big\{\ \ f \in \mathbb{X};\ \ \big(f(1), \dots, f(m)\big) \in S\,\big\}.$$

We will refer to a set of this form as m-cylinder and we denote by $\mathscr{C}_m$ the collection of m-cylinders. Define

$$\mathscr{C} := \bigcup_{m\geqslant 1} \mathscr{C}_m$$

and

$$\mu_m : \mathscr{C}_m \to [0,1],\ \ \mu_m\big[\,C_S\,\big] = \frac{|S|}{2^m}.$$

(i) Prove that $\mathscr{C}_m$ is an algebra of subsets of $\mathbb{X}$ and μ_m is finitely additive.

(ii) Prove that $\mathscr{C}_m \subset \mathscr{C}_{m+1}$ and

$$\mu_{m+1}\big|_{\mathscr{C}_m} = \mu_m.$$

(iii) Define $\mu : \mathscr{C} \to [0,1]$, by setting $\mu\big|_{\mathscr{C}_m} = \mu_m$. Prove that μ is well defined and it is a premeasure. We denote by $\bar{\mu}$ its extension as a measure to $\bar{\mathscr{C}} := \sigma(\mathscr{C})$. **Hint.** Use Theorem 19.1.43. You can assume the conclusions of Exercise 17.47.

(iv) Define $T : \mathbb{X} \to [0,1]$,

$$T(f) = \sum_{n\in\mathbb{N}} \frac{f(n)}{2^n},\ \ \forall f : \mathbb{N} \to \mathbb{B}.$$

Prove that T is a Lipschitz map and $T^{-1}(B) \subset \bar{\mathscr{C}}$ for any Borel subset $B \subset [0,1]$. **Hint.** Start by showing that $T^{-1}\big([0,k/2^n]\big) \in \bar{\mathscr{C}}$, $\forall n \in \mathbb{N}$, $k = 0,1,2,\dots,2^n$.

(v) Prove that $T_\#\bar{\mu} = \boldsymbol{\lambda}$ - the Lebesgue measure on $[0,1]$. **Hint.** Start by computing

$$\bar{\mu}\big[\,T^{-1}\big(\,[(k-1)/2^n, k/2^n]\,\big)\,\big],\ \ k = 1,\dots,2^n.$$

□

Exercise 19.29 (Cantor-Vitali)**.** Let

$$X := \big\{\,f \in C\big([0,1]\big);\ \ f(0) = 0,\ \ f(1) = 1\,\big\}.$$

Note that X is a closed subset of the Banach space $C\big([0,1]\big)$ equipped with the sup-norm, $\|-\|$. For any function $f:[0,1]\to\mathbb{R}$ we define $Tf:[0,1]\to\mathbb{R}$

$$(Tf)(x)=\begin{cases}\frac{f(3x)}{2}, & 0\leqslant x\leqslant\frac{1}{3},\\ \\ \frac{1}{2}, & \frac{1}{3}<x<\frac{2}{3},\\ \\ \frac{1}{2}+\frac{f(3x-2)}{2}, & \frac{2}{3}\leqslant x\leqslant 1.\end{cases}$$

(i) Show that if $f\in X$, then $Tf\in X$ and $\|Tf-Tg\|\leqslant\frac{1}{2}\|f-g\|$, $\forall f,g\in X$.

(ii) Define inductively a sequence (f_n) in X, $f_0(x)=x$, $f_n=Tf_{n-1}$, $\forall n\in\mathbb{N}$. Sketch the graphs of the functions f_0,f_1,f_2.

(iii) Show that the functions f_n converge in the norm $\|-\|$ to a function $f_\infty\in X$ satisfying $Tf_\infty=f_\infty$.

(iv) Show that f_∞ is nondecreasing and constant on the connected components of the complement of the Cantor set S_∞; see Example 19.2.10. (The function f_∞ is also known as "*Devil's staircase*".)

(v) Deduce that the continuous function f_∞ is differentiable almost everywhere with derivative 0, yet it is not constant.

□

Exercise 19.30. Prove Proposition 19.2.15. □

Exercise 19.31. Fix an integer $n\geqslant 2$ and define $F_n:\mathbb{R}\to\mathbb{R}$

$$F_n(x)=\begin{cases}0, & x<0,\\ \frac{k}{n}, & x\in\big[(k-1)/n,k/n\big),\ 1\leqslant k\leqslant n,\\ 1, & x\geqslant 1.\end{cases}$$

(i) Prove that F_n is a gauge function satisfying the finiteness condition (19.2.9). Denote by $\boldsymbol{\lambda}_n$ the associated Lebesgue-Stiltjes measure.

(ii) Show that

$$\boldsymbol{\lambda}_n=\frac{1}{n}\sum_{k=1}^{n}\delta_{(k-1)/n},$$

where δ_t is the Dirac measure on $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$ concentrated at t.

(iii) Find the quantile Q_n of F_n.

(iv) Graph the functions F_n and Q_n for $n=3$.

(v) Prove that for any continuous function $f:[0,1]\to\mathbb{R}$ we have

$$\lim_{n\to\infty}\int_{[0,1]}f(t)\boldsymbol{\lambda}_n\big[\,dt\,\big]=\int_0^1 f(t)dt,$$

where the integral on the right-hand side is the *Riemann* integral of f.

□

Exercise 19.32. Suppose that $F : \mathbb{R} \to \mathbb{R}$ is continuous, strictly increasing and

$$F(-\infty) = 0, \;\; F(\infty) = M.$$

(i) Prove that the induced map $F : \mathbb{R} \to (0, M)$ is bijective.

(ii) Denote by Q the quantile function of F. Prove that $Q(y) = F^{-1}(y)$, $\forall y \in (0, M)$.

□

Exercise 19.33. Suppose that $\mu : \big(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}\big) \to [0,\infty]$ is a measure defined on the Borel sigma-algebra generated by the open subsets of $\mathbb{R}^n$. Prove that the following statements are equivalent.

(i) $\mu\big[\, K \,\big] < \infty$ for any compact subset $K \subset \mathbb{R}^n$.

(ii) For any $x \in \mathbb{R}^n$ there exists $r > 0$ such that $\mu\big[\, B_r(x) \,\big] < \infty$, where $B_r(x)$ denotes the open ball of radius r centered at x.

□

Exercise 19.34. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space such that $\mu\big[\, \Omega \,\big] < \infty$ and $f_n : (\Omega, \mathcal{S}, \mu) \to \overline{\mathbb{R}}$ is a sequence of measurable functions satisfying

$$\forall \varepsilon > 0, \;\; \sum_{n \in \mathbb{N}} \mu\big[\, \{|f_n| \geqslant \varepsilon\} \,\big] < \infty. \tag{19.7.1}$$

(i) Prove that, for any $\varepsilon > 0$, the set of $\omega \in \Omega$ such that $|f_k(\omega)| \geqslant \varepsilon$, for infinitely many k's, is μ-negligible. **Hint.** Use Exercise 19.2.

(ii) Prove that $f_n \to 0$, μ-a.e..

Exercise 19.35. Define $r_n : [0,1] \to \mathbb{R}$, $n = 0, 1, 2, \ldots,$

$$r_0 = \boldsymbol{I}_{(0,1]}, \;\; r_n = \sum_{k=1}^{2^n} (-1)^{k+1} \boldsymbol{I}_{((k-1)2^{-n}, k2^{-n}]}, \;\; \forall n \in \mathbb{N}.$$

(i) Draw the graphs of r_0, r_1, r_2.

(ii) Draw the graphs of R_0, R_1, R_2, where

$$R_k(x) = \int_0^x r_k(t)dt, \;\; k = 0, 1, 2, 3, \ldots .$$

(iii) Show that for any $n > m \geqslant 0$ we have

$$\int_{[0,1]} r_m(x) r_n(x) \, \boldsymbol{\lambda}\big[\, dx \,\big] = 0.$$

□

Exercise 19.36. Consider the measure $\mu : \big(\mathbb{N}, 2^{\mathbb{N}}\big) \to [0, \infty]$

$$\mu\big[\, S \,\big] = \# S, \;\; \forall S \subset \mathbb{N}.$$

Let $f : \mathbb{N} \to \mathbb{R}$. Prove that the following statements are equivalent.

(i) $f \in \mathcal{L}^1(\mathbb{N}, 2^{\mathbb{N}}, \mu)$.

(ii) The series

$$f(1) + f(2) + \cdots + f(n) + \cdots$$

is absolutely convergent; see Definition 4.6.12.

Deduce that if $f \in \mathcal{L}^1\big(\mathbb{N}, 2^{\mathbb{N}}, \mu\big)$, then

$$\int_{\mathbb{N}} f d\mu = \sum_{n \in \mathbb{N}} f(n) := \lim_{n \to \infty} \big(f(1) + \cdots + f(n) \big). \qquad \square$$

Exercise 19.37. Consider the function

$$f : [0,1] \to \mathbb{R}, \;\; f(x) = \begin{cases} 1, & x \in [0,1] \backslash \mathbb{Q}, \\ 0, & x \in \mathbb{Q} \cap [0,1]. \end{cases}$$

(i) Show that f is Borel measurable and

$$\int_{[0,1]} f(x) \boldsymbol{\lambda}\big[\, dx \,\big] = 1.$$

(ii) Prove that f is not Riemann integrable.

□

Exercise 19.38. Let $L, T, \Phi : [0,1] \to \mathbb{R}$

$$L(x) = 4x(1-x), \;\; T(x) = \begin{cases} 2x, & x \in [0, 1/2], \\ 2 - 2x, & x \in (1/2, 1], \end{cases},$$

$$\Phi(x) = \frac{1}{2}\big(1 - \cos(\pi x) \big).$$

(i) Prove that L, T are Borel measurable and

$$L\big([0,1] \big), \;\; T\big([0,1] \big) \subset [0,1].$$

(ii) Denote by $\boldsymbol{\lambda}$ the Lebesgue measure on $[0,1]$. Show that for any Borel subset $B \subset [0,1]$ we have

$$T_{\#}\boldsymbol{\lambda}\big[\, B \,\big] = \boldsymbol{\lambda}\big[\, B \,\big],$$

and describe explicitly $L_{\#}\boldsymbol{\lambda}\big[\, B \,\big]$.

(iii) Prove that Φ is a homeomorphism $[0,1] \to [0,1]$ and

$$L = \Phi \circ T \circ \Phi^{-1}.$$

(iv) Set $\nu := \Phi_{\#}\boldsymbol{\lambda}$. Prove that for any Borel subset $B \subset [0,1]$ we have

$$\nu\big[\,B\,\big] = \frac{1}{\pi}\int_B \frac{1}{\sqrt{x(1-x)}}\boldsymbol{\lambda}\big[\,dx\,\big]$$

and show that $L_{\#}\nu = \nu$.

□

Exercise 19.39. Suppose that $\big(\Omega, \mathcal{S}, \mu\big)$ is a measured space such that $\mu\big(\,\Omega\,\big) < \infty$. Let $f \in \mathcal{L}^1\big(\,\Omega, \mathcal{S}, \mu\big)$. Prove that the following are equivalent.

(i) $f \geqslant 0$, μ-a.e..

(ii) For any $S \in \mathcal{S}$

$$\int_\Omega f\boldsymbol{I}_S d\mu \geqslant 0.$$

Hint. (ii) ⇒ (i) Choose S of the form $S = \{f \leqslant c\}$ for appropriate c.

□

Exercise 19.40. Let $\big(\,\Omega, \mathcal{S}, \mu\,\big))$ be. a measured space and suppose that $(f_n)_{n\in bN}$ is a sequence of functions in $\mathcal{L}^1\big(\,\Omega, \mathcal{S}, \mu\,\big)$ that converges μ-a.e. to a function $f \in \mathcal{L}^1\big(\,\Omega, \mathcal{S}, \mu\,\big)$. Prove that the following are equivalent.

(i) $\|\,f - f_n\,\|_{L^1} \to 0$ as $n \to \infty$.

(ii) $\|\,f_n\,\|_{L^1} \to \|\,f\,\|_{L^1}$ as $n \to \infty$.

Hint. (ii) ⇒ (i) Apply Fatou's lemma to the sequence of functions $g_n = |f_n| + |f| - |f - f_n|$. □

Exercise 19.41. Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measured space and (S, d) is a metric space. Consider a function

$$F : S \times \Omega \to \mathbb{R},\;\; (s, \omega) \mapsto F_s(\omega)$$

satisfying the following properties.

(i) For any $s \in S$ the function $\Omega \ni \omega \mapsto F_s(\omega) \in \mathbb{R}$ is measurable.

(ii) For any $\omega \in \Omega$ the function $S \ni s \mapsto F_s(\omega) \in \mathbb{R}$ is continuous.

(iii) There exists $h \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ such that $|F_s(\omega)| \leqslant h(\omega)$, $\forall (s, \omega) \in S \times \Omega$.

Prove that $F_s \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$, $\forall s \in S$, and the resulting function

$$S \ni s \mapsto \int_\Omega F_s(\omega)\mu\big[\,d\omega\,\big] \in \mathbb{R}$$

is continuous. **Hint.** Use the Dominated Convergence Theorem. □

Exercise 19.42. Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measured space and $I \subset \mathbb{R}$ is an open interval. Consider a function

$$F : I \times \Omega \to \mathbb{R},\;\; (t, \omega) \mapsto F(t, \omega)$$

satisfying the following properties.

(i) For any $t \in I$ the function $F(t,-) : \Omega \to \mathbb{R}$ is integrable,

$$\int_\Omega |F(t,\omega)|\,\mu\big[\,d\omega\big] < \infty.$$

(ii) For any $\omega \in \Omega$ the function $I \ni t \mapsto F(t,\omega) \in \mathbb{R}$ is differentiable at $t_0 \in I$. We denote by $F'(t_0,\omega)$ its derivative.

(iii) There exists $h \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ such that

$$|F(t,\omega) - F(t_0,\omega)| \leqslant h(\omega)|t - t_0|, \quad \forall (t,\omega) \in I \times \Omega.$$

Prove that the function

$$I \ni t \mapsto \int_\Omega F(t,\omega)\mu\big[\,d\omega\,\big] \in \mathbb{R}$$

is differentiable at t_0 and

$$\frac{d}{dt}\Big|_{t=t_0}\left(\int_\Omega F(t,\omega)\mu\big[\,d\omega\,\big]\right) = \int_\Omega F'(t_0,\omega)\mu\big[\,d\omega\,\big].$$

□

Exercise 19.43. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuous function. Prove that the following are equivalent.

(i) $f \in \mathcal{L}^1\big(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \boldsymbol{\lambda}\big)$.

(ii) The improper Riemann integral

$$\int_{-\infty}^{\infty} f(x)dx$$

is absolutely convergent.

□

Exercise 19.44. Consider the function

$$J : \mathbb{R} \to \mathbb{R}, \quad J(a) = \int_{-\infty}^{\infty} e^{-x^2/2}\cos(ax)\,dx.$$

(i) Prove that $J \in C^1(\mathbb{R})$.

(ii) Show that

$$J'(a) = -aJ(a), \quad \forall a \in \mathbb{R}.$$

(iii) Compute $J(a)$.

Hint. For (i) and (ii) use Exercises 19.43 and 19.42. In (iii) you will need (15.4.4) at some point. □

Exercise 19.45. Consider the functions $f_n, f : \mathbb{R} \to \mathbb{R}$, $n \in \mathbb{N}$

$$f_n(x) = \begin{cases} \frac{\sin(\pi x)}{x}, & |x| \leqslant n, \\ 0, & |x| > n, \end{cases}, \quad f(x) = \frac{\sin(\pi x)}{x}.$$

At 0 we have

$$f_n(0) = f(0) = \lim_{x\to 0} \frac{\sin(\pi x)}{x} = \pi.$$

(i) Show that $f_n(x) \to f(x)$, $\forall x \in \mathbb{R}$.

(ii) Show that $f_n \in \mathcal{L}^1(\mathbb{R}, \boldsymbol{\lambda})$ and

$$\lim_{n\to\infty} \int_{\mathbb{R}} f_n(x) \boldsymbol{\lambda}\big[\, dx \,\big]$$

exists and it is finite.

(iii) Show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(x)|\, \boldsymbol{\lambda}\big[\, dx \,\big] = \infty.$$

(iv) Show that f is not Lebesgue integrable.

□

Exercise 19.46. Suppose that Ω is a finite set and μ is a measure on 2^{Ω} such that $\mu\big[\,\Omega\,\big] = 1$. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Prove that for any function $f : \Omega \to \mathbb{R}$

$$\Phi\Big(\int_{\Omega} f d\mu \Big) \leqslant \int_{\Omega} \Phi(f) d\mu.$$

□

Exercise 19.47. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space such that $\mu\big[\,\Omega\,\big] = 1$ and $\Phi : \mathbb{R} \to \mathbb{R}$ is a C^1 convex function. Let $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mathbb{P})$ such that $\Phi(f) \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$.

(i) Prove that for any $x_0 \in \mathbb{R}$ there exists a linear function $L = L_{x_0,\Phi} : \mathbb{R} \to \mathbb{R}$ such that

$$L(x_0) = \Phi(x_0), \;\; L(x) \leqslant \Phi(x), \;\; \forall x \in \mathbb{R}.$$

(ii) Prove that

$$\Phi\Big[\int_{\Omega} f d\mu \Big] \leqslant \int_{\Omega} \Phi(f) d\mu.$$

Hint. Use (i) with $x_0 = \int_{\Omega} f d\mu$.

(iii) Prove that if $1 \leqslant p_0 < p_1 < \infty$ then

$$\mathcal{L}^{p_1}(\Omega, \mathcal{S}, \mu) \subset \mathcal{L}^{p_0}(\Omega, \mathcal{S}, \mu).$$

Hint. Use (ii) with $\Phi(x) = |x|^{p_1/p_0}$.

□

Exercise 19.48. Construct a sequence of continuous functions $f_n : \mathbb{R} \to [0, \infty)$ with the following properties.

(i)

$$\int_{\mathbb{R}} f_n(x) \lambda\big[\, dx \,\big] = 1, \;\; \forall n \in \mathbb{N}.$$

(ii)

$$\lim_{n\to\infty} f_n(x) = 0, \;\; \forall x \in \mathbb{R}.$$

□

Exercise 19.49. Suppose that $(\Omega_i, \mathcal{S}_i)$, $i = 0, 1$, are two measurable spaces. Denote by $\mathcal{A}$ the collection of subsets of $\Omega_0 \times \Omega_1$ that are finite disjoint unions of rectangles $S_0 \times S_1$, $S_i \in \mathcal{S}_i$. Prove that $\mathcal{A}$ is an algebra of sets.

Hint. Suppose that $A := S_0 \times S_1 \subset \Omega_0$, $B := T_0 \times T_1$ are two rectangles. Prove that $A \cap B$ is a rectangle and $A^c = (\Omega_0 \times \Omega_1)\backslash A \in \mathcal{A}$. Conclude by observing that $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$. □

Exercise 19.50. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space and $f \in \mathcal{L}^1_+(\Omega, \mathcal{S}, \mu)$. Define

$$R_f := \big\{ (\omega, x) \in \Omega \times \mathbb{R};\;\; 0 \leqslant x \leqslant f(\omega) \big\}.$$

Intuitively, R_f is the region below the graph of f and above the "horizontal axis".

(i) Prove that $R_f \in \mathcal{S} \otimes \mathcal{B}_{\mathbb{R}}$.

(ii) Show that

$$\mu \otimes \boldsymbol{\lambda}\big[\, R_f \,\big] = \int_{\Omega\times\mathbb{R}} \boldsymbol{I}_{R_f}(\omega, x)\mu \otimes \boldsymbol{\lambda}\big[\, d\omega dx \,\big] = \int_\Omega f(\omega)\mu\big[\, d\omega \,\big].$$

(iii) Show that

$$\int_\Omega f(\omega)\mu\big[\, d\omega \,\big] = \int_0^\infty \mu\big[\, \{f > x\} \,\big]\,\boldsymbol{\lambda}\big[\, dx \,\big].$$

(iv) Suppose that $\Phi : [0, \infty) \to [0, \infty)$ is a nondecreasing C^1 function such that $\Phi(0) = 0$. Prove that

$$\int_\Omega \Phi\big(\, f(\omega)\,\big)\mu\big[\, d\omega \,\big] = \int_0^\infty \Phi'(x)\mu\big[\, \{f > x\} \,\big]\,\boldsymbol{\lambda}\big[\, dx \,\big].$$

Hint.(ii)+(ii) use Fubini Theorem in two different ways. For (iii) Note that $\Phi(f(\omega)) = \int_0^{f(\omega)} \Phi'(x)\boldsymbol{\lambda}[dx]$. □

Exercise 19.51. Suppose that $S \subset \mathbb{R}^2$ is Borel measurable and its intersection with any vertical line $\{x\} \times \mathbb{R}$ has 1-dimensional Lebesgue measure zero. In other, words the

$$S_x := \big\{\, y \in \mathbb{R};\;\; (x, y) \in S \,\big\} \subset \mathbb{R}$$

is negligible, $\forall x \in \mathbb{R}$. Prove that the 2-dimensional Lebesgue measure of S is also zero. □

Exercise 19.52 (N. Wiener)**.** Consider a family $\mathcal{B} := \big(B_{r_i}(x_i)\big)_{i\in I}$ of open balls in $\mathbb{R}^n$. Let U be an open set contained in their union. Fix $c < \boldsymbol{\lambda}\big[\, U \,\big]$.

(i) Show that there exists a compact subset $K \subset U$ such that $\boldsymbol{\lambda}\big[\, K \,\big] > c$.

(ii) Prove that there exists a *finite* subfamily $\big(B_{r_j}(x_j)\big)_{j\in J}$ of the family $\mathcal{B}$ such that any two balls in this subfamily are disjoint and

$$\sum_{j\in J}\boldsymbol{\lambda}\big[\,B_{r_j}(x_j)\,\big]>3^{-n}c.$$

Hint. Prove that there exists a finite subfamily of pairwise disjoint balls $\big(B_{r_j}(x_j)\big)_{j\in J}$ such that the family $\big(B_{3r_j}(x_j)\big)_{j\in J}$ covers the compact set K found in (i). Look at some special cases first. Understand what happens when K is covered by two balls from $\mathcal{B}$ and then what happens when K is covered by three balls from $\mathcal{B}$.

□

Exercise 19.53. Fix $p\in(0,1)$, set $q=1-p$. Let $\Omega=\mathbb{N}$, $\mathcal{S}=2^{\mathbb{N}}$ and $\mu:\mathcal{S}\to[0,\infty]$ determined by

$$\mu\big[\,\{n\}\,\big]=pq^{n-1}.$$

Let $f:\mathbb{N}\to[0,\infty)$, $f(n)=n$, $\forall n\in\mathbb{N}$.

(i) Prove that

$$\sum_{n\in\mathbb{N}}npq^{n-1}=\int_{\mathbb{N}}fd\mu$$

(ii) Use the equality in Exercise 19.50(iii) to compute $\int_{\mathbb{N}}fd\mu$.

Hint. (ii) Show that $\mu\big[\,\{f>x\}\,\big]=q^{\lfloor x\rfloor}$ where $\lfloor x\rfloor$ denotes the integer part of x. □

Exercise 19.54. Fix $p\in(0,1)$ and set $q=1-p$. Consider the set $\Omega=\{0,1\}$ and the measure $\mu:2^{\Omega}\to[0,\infty)$ defined by

$$\mu\big[\,\{0\}\,\big]=q,\;\;\mu\big[\,\{1\}\,\big]=p.$$

Fix $n\in\mathbb{N}$ and consider the product Ω^n. The elements of Ω^n are strings $\vec{\epsilon}=(\epsilon_1,\dots,\epsilon_n)$ consisting of 0's and 1's. We have coordinate functions

$$u_k:\Omega^n\to\{0,1\},\;\;u_k(\vec{\epsilon})=\epsilon_k.$$

(i) Prove that we have an equality of sigma-algebras

$$2^{\Omega^n}=\underbrace{2^{\Omega}\otimes\cdots\otimes 2^{\Omega}}_{n}.$$

(ii) Denote by μ_n the product measure

$$\mu_n:=\underbrace{\mu\otimes\cdots\otimes\mu}_{n}.$$

Compute

$$\int_{\Omega^n}e^{tu_k}d\mu_n,\;\;\forall k=1,\dots,n,\;\;t\in\mathbb{R}.$$

(iii) Show that for $j \neq k$ we have

$$\int_{\Omega^n} e^{tu_j} e^{tu_k} d\mu_n = \left(\int_{\Omega^n} e^{tu_j} d\mu_n\right)\left(\int_{\Omega^n} e^{tu_k} d\mu_n\right), \quad t \in \mathbb{R}.$$

(iv) Set $s = u_1 + \cdots + u_n$. Compute

$$\int_{\Omega^n} s d\mu_n, \quad \int_{\Omega^n} e^{ts} d\mu_n.$$

(v) Observe that the range of s is $R_n := \{0, 1, \dots, n\}$. Denote by β_n the pushforward $s_\# \mu_n$ so that β_n is a measure on R_n. Describe the measure $\beta_n : 2^{R_n} \to [0, \infty]$.

(vi) Let $f : R_n \to \mathbb{R}$, $f(k) = k$. Use Theorem 19.3.27 and (iv) above to show

$$\sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \int_{R_n} f d\beta_n = \int_{\Omega^n} s d\mu_n$$

□

Exercise 19.55. Suppose that $f : \mathbb{R}^n \to \mathbb{C}$ is Lebesgue integrable. Denote by $\langle -, - \rangle$ the standard inner product in $\mathbb{R}^n$ and by $\| - \|$ the Euclidean norm.

(i) Prove that for any $\xi \in \mathbb{R}^n$ the function $x \mapsto e^{\boldsymbol{i}\langle x, \xi \rangle} f(x)$ is also Lebesgue integrable and the resulting function

$$\mathbb{R}^n \ni \xi \mapsto \widehat{f}(\xi) := \int_{\mathbb{R}^n} e^{\boldsymbol{i}\langle x, \xi \rangle} f(x) \boldsymbol{\lambda}\big[\, dx \,\big] \in \mathbb{C}$$

is bounded and continuous. **Hint.** Use Exercise 19.41.

(ii) Compute $\widehat{f}(\xi)$ when $f(x) = e^{-\|x\|^2}$. **Hint.** Use Fubini and Exercise 19.44.

□

Exercise 19.56. Let $n \in \mathbb{N}$. For $\boldsymbol{p} \in \mathbb{R}^n$ and $r > 0$ denote by $\bar{B}_r$ the closed ball in $\mathbb{R}^n$ of radius r center at 0

$$\bar{B}_r(\boldsymbol{p}) = \big\{\, \boldsymbol{x} \in \mathbb{R}^n; \ \ \|\boldsymbol{x}\| \leqslant r \,\big\}.$$

(i) Construct a sequence f_ν of compactly supported continuous functions $f_\nu : \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{\nu \to \infty} \int_{\mathbb{R}^n} \big|\, f_\nu - \boldsymbol{I}_{\bar{B}_1(0)} \,\big|\, d\boldsymbol{\lambda}_n = 0.$$

(ii) Prove that the space $C_{\mathrm{cpt}}(\mathbb{R}^n)$ of compactly supported continuous functions $\mathbb{R}^n \to \mathbb{R}$ is dense in $L^1(\mathbb{R}^n)$. **Hint.** Use (i) and Proposition 19.4.8.

□

Exercise 19.57. Let $n \in \mathbb{N}$. For $x \in \mathbb{R}^n$ and $r > 0$ denote by $B_r(x)$ the open ball in $\mathbb{R}^n$ of radius r center at x. For $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ and $r > 0$ we define

$$D_r f : \mathbb{R}^n \to [0, \infty), \quad D_r f(x) = \frac{1}{\boldsymbol{\lambda}\big[\, B_r(x) \,\big]} \int_{B_r(x)} \big|\, f(y) - f(x) \,\big|\, \boldsymbol{\lambda}\big[\, dy \,\big].$$

Show that if f is a *uniformly* continuous function then

$$\lim_{r\searrow 0}\sup_{x\in\mathbb{R}^n} D_r f(x) = 0.$$

□

Exercise 19.58. Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is a continuous function. Prove that $f\in L^1(\mathbb{R}^n,\boldsymbol{\lambda}_n)$ if and only if the improper integral

$$\int_{\mathbb{R}^n} f(x)\,|dx|$$

is absolutely integrable in the sense of Definition 15.4.8. □

Exercise 19.59. Denote by C_n the cube $[0,1]^n\subset\mathbb{R}^n$ and define

$$a_n : C_n\to\mathbb{R},\ \ a_n(x) = \frac{1}{n}\big(x_1+\cdots+x_n\big)$$

(i) Prove that $0\leqslant a_n(x)\leqslant 1$, $\forall x\in C_n$.

(ii) Show that

$$\int_{C_n} a_n(x)\boldsymbol{\lambda}\big[\,dx\,\big] = \frac{1}{2}.$$

(iii) Set $\bar{a}_n(x) = a_n(x)-\frac{1}{2}$. Show that

$$\int_{C_n}\bar{a}_n(x)^2\boldsymbol{\lambda}\big[\,dx\,\big] = \frac{1}{12n}.$$

(iv) Fix $\varepsilon>0$. Prove that[7]

$$\lim_{n\to\infty}\boldsymbol{\lambda}_n\big[\,\{|\bar{a}_n|\geqslant\varepsilon\}\cap C_n\,\big] = 0.$$

Hint. Use Markov's inequality.

Exercise 19.60. Suppose that $(\Omega,\mathcal{S},\mu)$ is a measured space and $p,q,r\in(1,\infty)$ satisfy

$$\frac{1}{r} = \frac{1}{p}+\frac{1}{q}.$$

Prove that for any $f\in L^p(\Omega,\mathcal{S},\mu)$ and any $g\in L^q(\Omega,\mathcal{S},\mu)$ we have $fg\in L^r(\Omega,\mathcal{S},\mu)$ and

$$\|fg\|_{L^r}\leqslant\|f\|_{L^p}\|g\|_{L^q}.$$

Hint. Use Hölder's inequality. □

Exercise 19.61. Suppose that $(\Omega,\mathcal{S},\mu)$ is a measured space, $p\in[1,\infty)$ and $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are sequences that converge in $L^p(\Omega,\mathcal{S},\mu)$ to $f\in L^p(\Omega,\mathcal{S},\mu)$ and respectively $g\in L^p(\Omega,\mathcal{S},\mu)$.

[7] The equality (iv) is a special case of the Law of Large Numbers and is a manifestation of high dimensional concentration phenomenon: for n very large, most of the points in the cube C_n are concentrated near the median hyperplane $\{\bar{a}_n(x) = 1/2\}$.

(i) Prove that the sequences $(|f_n|)$, $\max(f_n, g_n)$, $\min(f_n, g_n)$ converge in $L^p(\Omega, \mathcal{S}, \mu)$ to $\max(f, g)$ and respectively $\min(f, g)$

(ii) Suppose that $p > 2$. Prove that $(f_n g_n)_{n\in\mathbb{N}}$ converges in $L^{p/2}(\Omega, \mathcal{S}, \mu)$ to fg.

□

Exercise 19.62. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space such that $\mu\big[\,\Omega\,\big] < \infty$ prove that if $p_0, p_1 \in [1, \infty]$ and $p_0 \leqslant p_1$, then

$$L^{p_0}(\Omega, \mathcal{S}, \mu) \supset L^{p_1}(\Omega, \mathcal{S}, \mu).$$

and the induced linear map $L^{p_1}(\Omega, \mathcal{S}, \mu) \to L^{p_0}(\Omega, \mathcal{S}, \mu)$ is continuous. **Hint.** Use Hölder's inequality. □

Exercise 19.63. Let $f \in L^1\big(\,[0,1]\,\big) = L^1\big(\,[0,1], \mathcal{B}_{[0,1]}, \boldsymbol{\lambda}\,\big)$. For $k, n \in \mathbb{N}$, $k \leqslant n$, we set

$$C_{k,n} := [(k-1)/n, k/n],\;\; f_{k,n} := n\int_{C_{k,n}} f(x)dx,\;\; f_n := \sum_{k=1}^n f_{k,n}\boldsymbol{I}_{C_{k,n}} \in L^1\big(\,[0,1]\,\big).$$

Prove that

$$\lim_{n\to\infty}\int_{[0,1]}\big|\,f_n(x) - f(x)\,\big|dx = 0.$$

Hint. Begin by proving the statement in the special case when f is continues. Conclude using Theorem 19.4.14.

□

Exercise 19.64. $(\Omega, \mathcal{S}, \mu)$ is a measured space such that $\mu\big[\,\Omega\,\big] < \infty$. Define

$$\rho : \mathbb{R} \to [0, \infty),\;\; \rho(x) = \min(|x|, 1)$$

and

$$d = d_\mu : L^0(\Omega, \mathcal{S}) \times L^0(\Omega, \mathcal{S}) \to [0, \infty),\;\; d(f, g) = \int_\Omega \rho(f - g)d\mu.$$

(i) Prove that $(L^0(\Omega, \mathcal{S}), d)$ is a metric space. We denote by $L^0(\Omega, \mathcal{S}, \mu)$ this metric space.

(ii) Prove that the natural map $L^1(\Omega, \mathcal{S}, \mu) \to L^0(\Omega, \mathcal{S}, \mu)$ is continuous, i.e.,

$$\lim_{n\to\infty}\|f_n - f\|_{L^1} = 0 \Rightarrow \lim_{n\to\infty} d_\mu(f_n, f) = 0.$$

(iii) Prove that if $f_n \to f$ μ-a.e., then $\lim_{n\to\infty} d_\mu(f_n, f) = 0$.

(iv) Suppose that $f, f_1, f_2, \ldots \in L^0(\Omega, \mathcal{S}, \mu)$. Prove that the following are equivalent.

(a) $\lim_{n\to\infty} d_\mu(f_n, f) = 0$.

(b) The sequence $(f_n)_{n\in\mathbb{N}}$ *converges to f in measure*, i.e.,

$$\forall\varepsilon > 0,\;\lim_{n\to\infty}\mu\big[\,\{|f_n - f| > \varepsilon\,\big] = 0.$$

□

Exercise 19.65. Suppose that $f, g \in L^1(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \boldsymbol{\lambda})$.

(i) Show that the function $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $h(x,y) = f(x-y)g(y)$ is measurable.

(ii) Prove that the function

$$y \mapsto f(x-y)g(y)$$

is integrable for almost every $x \in \mathbb{R}^n$. We write

$$f * g(x) := \int_{\mathbb{R}^n} f(x-y)g(y)\boldsymbol{\lambda}\big[\,dy\,\big]$$

(iii) Show that $f * g \in L^1(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \boldsymbol{\lambda})$, $f * g = g * f$ a.e., and

$$\|f * g\|_{L^1} \leqslant \|f\|_{L^1}\|g\|_{L^1}. \tag{19.7.2}$$

□

Exercise 19.66. Suppose that $w : \mathbb{R}^n \to [0,\infty)$ is a nonnegative continuous function such that

$$w(x) = 0,\ \ \forall \|x\| > 1 \ \text{ and } \int_{\mathbb{R}^n} w(x)\boldsymbol{\lambda}_n\big[\,dx\,\big] = 1.$$

For $\nu \in \mathbb{N}$ we set

$$w_\nu : \mathbb{R}^n \to \mathbb{R},\ \ w_\nu(x) = \nu^n w(\nu x).$$

(i) Prove that

$$\int_{\mathbb{R}^n} w_\nu(x-y)d\boldsymbol{\lambda}\big[\,dy\,\big] = 1,\ \ \forall \nu \in \mathbb{N},\ \ x \in \mathbb{R}^n.$$

(ii) Prove for any $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$ and any $\nu \in \mathbb{N}$ the function $w_\nu * f : \mathbb{R}^n \to \mathbb{R}$ is continuous. **Hint.** Use Exercise 19.41.

(iii) Prove that for any $f \in C_{\mathrm{cpt}}(\mathbb{R}^n)$ and any $\nu \in \mathbb{N}$ the function $w_\nu * f$ has compact support and

$$\lim_{\nu\to\infty} \sup_{x\in\mathbb{R}^n} \big|\, w_\nu * f(x) - f(x)\,\big| = 0.$$

Hint. Show that

$$f(x) = \int_{\mathbb{R}^n} f(x)w_\delta(x-y)\boldsymbol{\lambda}\big[\,dy\,\big],$$

$$\big|\, w_\nu * f(x) - f(x)\,\big| \leqslant \int_{\mathbb{R}^n} \big|\, f(y) - f(x)\,\big| w_\nu(x-y)\boldsymbol{\lambda}\big[\,dy\,\big]$$

$$= \int_{B_{1/\nu}(x)} \big|\, f(y) - f(x)\,\big|\, w_\nu(x-y)\boldsymbol{\lambda}\big[\,dy\,\big].$$

Conclude using the uniform continuity of f.

(iv) Prove that for any $f \in C_{\mathrm{cpt}}(\mathbb{R}^n)$ and any $\nu \in \mathbb{N}$ we have

$$\lim_{\nu\to\infty} \|w_\nu * f - f\|_{L^1} = 0.$$

Hint. Use (iii) above.

(v) Prove that for any $f \in L^1(\mathbb{R}^n, \boldsymbol{\lambda})$

$$\lim_{\nu\to\infty} \|w_\nu * f - f\|_{L^1} = 0.$$

Hint. Use Exercise 19.56, (19.7.2), and (iv) above.

□

Exercise 19.67 (The Moment Problem)**.** For any finite Borel measure μ on $[0,1]$ we associate its *momenta*

$$M_n(\mu) := \int_{[0,1]} x^n \, \mu\big[\, dx \,\big], \;\; n = 0, 1, 2, \dots .$$

(i) Prove that for any $t \in \mathbb{R}$ the series

$$\sum_{n \geqslant 0} M_n(\mu) \frac{t^n}{n!}$$

is absolutely convergent. Denote by $M_\mu(t)$ its sum.

(ii) Prove that

$$M_\mu(t) = \int_{[0,1]} e^{tx} \mu\big[\, dx \,\big].$$

(iii) Suppose that μ, ν are two finite Borel measures. Prove that

$$\mu = \nu \Longleftrightarrow M_\mu(t) = M_\nu(t), \;\; \forall t \in \mathbb{R}.$$

Hint. Use Exercise 19.42 to show that $M_\mu(t) = M_\nu(t)$ implies $M_n(\mu) = M_n(\nu)$, $\forall n \geqslant 0$. Prove next that $\mu\big[\, P \,\big] = \nu\big[\, P \,\big]$ for any polynomial P. Conclude using Corollary 17.4.11 and 19.4.15.

□

Exercise 19.68. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space. Prove that the normed space $L^\infty(\Omega, \mathcal{S}, \mu)$ is complete. □

Exercise 19.69. Suppose that $\mu \in \mathrm{Meas}(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ is a measure such that

$$\mu\big[\, \mathbb{R}^n \,\big] = 1 \;\text{ and }\; C := \int_{\mathbb{R}^n} \|x\| \mu\big[\, x \,\big] < \infty,$$

where $\|x\|$ denotes the Euclidean norm of $x \in \mathbb{R}^n$. For any $\varepsilon > 0$ define $S_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$, $S_\varepsilon(x) = \varepsilon x$. Set $\mu_\varepsilon := (S_\varepsilon)_{\#}\mu$.

(i) Express

$$\int_{\mathbb{R}^n} \|x\| \mu_\varepsilon\big[\, dx \,\big]$$

in term of C. **Hint.** Use Theorem 19.3.27.

(ii) For each $r > 0$ we denote by B_r the open ball of radius r centered at 0. Prove that for any $r > 0$,

$$\lim_{\varepsilon \searrow 0} \mu_\varepsilon\big[\, \mathbb{R}^n \backslash B_r \,\big] = 0.$$

(iii) Suppose that μ is absolutely continuous with respect to the Lebesgue measure $\boldsymbol{\lambda}_n$ and the corresponding density is $\rho = \frac{d\mu}{d\boldsymbol{\lambda}_n}$. Show that $\mu_\varepsilon \ll \boldsymbol{\lambda}_n$ and compute the density, $\rho_\varepsilon = \frac{d\mu_\varepsilon}{d\boldsymbol{\lambda}_n}$.

□

Exercise 19.70. Prove Proposition 19.5.10. **Hint.** Use the Hahn decomposition of ν. □

Exercise 19.71. Suppose that $(\Omega, \mathcal{S})$ is a measurable spaces and μ_0, μ_1 are two probability measures on $\mathcal{S}$, $\mu_0\big[\,\Omega\,\big] = \mu_1\Omega\,\big] = 1$. Consider the new probability measure

$$\nu := \frac{1}{2}\big(\,\mu_0 + \mu_1\,\big).$$

(i) Prove that $\mu_0, \mu_1 \ll \nu$. For $i = 0, 1$ we denote by $\frac{d\mu_i}{d\nu} \in L^1_+(\Omega, \mathcal{S}, \mu)$ the density of μ_i with respect to ν; see Definition 19.5.16.

(ii) Define

$$H(\mu_0, \mu_1) = \int_\Omega \left(\frac{d\mu_0}{d\nu}\right)^{\frac{1}{2}} \left(\frac{d\mu_1}{d\nu}\right)^{\frac{1}{2}} d\nu.$$

Prove that $0 \leqslant H(\mu_0, \mu_1) \leqslant 1$.

(iii) Show that if $H(\mu_0, \mu_1) = 0$, then $\mu_0 \perp \mu_1$; see Definition 19.5.5

□

Exercise 19.72. Suppose that $(\Omega, \mathcal{S}, \mu)$ is a finite measured space and $f \in \mathcal{L}^1\big(\,\Omega, \mathcal{S}, \mu\big)$. Fix a sigma-subalgebra $\mathcal{A} \subset \mathcal{S}$.

(i) Prove that there exists a function $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$

$$\int_A f(x)\boldsymbol{\lambda}\big[\,dx\,\big] = \int_A g(x)\boldsymbol{\lambda}\big[\,dx\,\big], \quad \forall A \in \mathcal{A}. \tag{19.7.3}$$

(ii) Prove that if $g_1, g_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ satisfy (19.7.3), then $g_1 = g_2$ μ-a. e..

□

Exercise 19.73. Denote by $\boldsymbol{\lambda}$ the Lebesgue measure on $[0, 1]$. Fix a partition $\boldsymbol{P}$ of $[0, 1]$

$$\boldsymbol{P} : 0 = a_0 < a_1 < \cdots < a_{n-1} < a_n = 1.$$

Denote by $\mathcal{S}(\boldsymbol{P})$ the sigma-algebra generated by the intervals

$$A_i = [a_{i-1}, a_i), i = 1, \dots, n-1, \quad A_n = [a_{n-1}, 1].$$

Fix a *Borel measurable* function $f \in L^1\big(\,[0, 1], \boldsymbol{\lambda}\,\big)$.

(i) Prove that a Borel measurable function $g : [0, 1] \to \mathbb{R}$ is $\mathcal{S}(\mathbb{P})$-measurable if and only if there exist constants $C_1, \dots, C_n \in \mathbb{R}$ such that

$$g = \sum_{k-1}^{n} C_k \boldsymbol{I}_{A_k}.$$

(ii) Describe explicitly an $\mathcal{S}(\mathbb{P})$-measurable function $g : [0, 1] \to \mathbb{R}$ such that

$$\int_S f(x)\boldsymbol{\lambda}\big[\,dx\,\big] = \int_S g(x)\boldsymbol{\lambda}\big[\,dx\,\big], \quad \forall S \in \mathcal{S}(\mathbb{P}).$$

□

Exercise 19.74 (Vitali-Hahn-Saks)**.** Suppose that $(\Omega, \mathcal{S}, \mu)$ is a *finite* measured space. Define an equivalence relation on $\mathcal{S}$ by setting $S \sim S'$ if $\mu\big[\, S\Delta S' \,\big] = 0$, where Δ denotes the symmetric difference

$$S\Delta S' = \big(\, S\backslash S' \,\big) \cup \big(\, S' \cup S \,\big).$$

Define $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$

$$d\big(\, S_0, S_1 \,\big) = \mu\big[\, S_0 \Delta S_1 \,\big].$$

(i) Prove that $\forall S_0, S_1, S_2 \in \mathcal{S}$ we have

$$d\big(\, S_0, S_1 \,\big) = d\big(\, S_1, S_0\big),\;\; d\big(\, S_0, S_2 \,\big) \leqslant d\big(\, S_0, S_1 \,\big) + d\big(\, S_1, S_2 \,\big)$$

and $d\big(\, S_0, S_1 \,\big) = 0$ iff $S_0 \sim S_1$.

(ii) Prove that d defines a *complete* metric d on $\bar{\mathcal{S}} := \mathcal{S}/\sim$.

(iii) Suppose that $\lambda : \mathcal{S} \to \mathbb{R}$ is a signed measure that is absolutely continuous with respect to μ. Hence $\lambda\big[\, S_0 \,\big] = \lambda\big[\, S_1 \,\big] = 0$ if $S_0 \sim S_1$. Prove that the induced function

$$\lambda : \bar{\mathcal{S}} \to \mathbb{R}$$

is continuous with respect to the metric d.

(iv) Suppose that (λ_n) is a sequence of finite signed measures on $\mathcal{S}$ such that $\lambda_n \ll \mu, \forall n$ and, $\forall S \in \mathcal{S}$, the sequence $\lambda_n\big[\, S \,\big]$ has a finite limit λ. Prove that $\lambda : \mathcal{S} \to \mathbb{R}$ is finitely additive and $\lambda\big[\, S \,\big] = 0$ if $\mu\big[\, S \,\big] = 0$.

(v) For any $\varepsilon > 0$ and $k \in \mathbb{N}$ we set

$$\bar{\mathcal{S}}_{k,\varepsilon} := \Big\{ S \in \bar{\mathcal{S}};\; \sup_{m \in \mathbb{N}} \big|\, \lambda_l\big[\, S \,\big] - \lambda_{k+n}\big[\, S \,\big] \,\big| \leqslant \varepsilon \Big\}.$$

Prove that the sets $\bar{\mathcal{S}}_{k,\varepsilon} \subset \bar{\mathcal{S}}$ are closed with respect to the metric d and

$$\bar{\mathcal{S}} = \bigcup_{k \in \mathbb{N}} \bar{\mathcal{S}}_{k,\varepsilon},\;\; \forall \varepsilon > 0.$$

(vi) Prove that the induced function $\bar{\lambda} : \bar{\mathcal{S}} \to \mathbb{R}$ is continuous with respect to the metric d and deduce that λ is countably additive. **Hint.** It suffice to show that for any decreasing sequence (S_n) in $\mathcal{S}$ with empty intersection we have $\lim \lambda\big[\, S_n \,\big] = 0$. Deduce this from (v) and Baire's theorem.

□

19.8. Exercises for extra credit

Exercise* 19.1. Construct an example of a set Ω and an increasing sequence of sigma-algebras of Ω

$$\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots \subset 2^{\Omega},$$

such that the union

$$\bigcup_{n\in\mathbb{N}} \mathcal{S}_n$$

is *not* a sigma-algebra. □

Exercise* 19.2. Let $(\Omega, \mathcal{S}, \mu)$ be a finite measured space and $(F_n)_{n\in\mathbb{N}}$ a sequence in $\mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ that converges everywhere to a function $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$. Prove that

$$\lim_{n\to\infty} \int_\Omega \bigl|\, f_n(\omega) - f(\omega) \,\bigr| \mu\bigl[\, d\omega \,\bigr] = 0 \iff \lim_{n\to\infty} \int_\Omega \bigl|\, f_n(\omega) \,\bigr| \mu\bigl[\, d\omega \,\bigr] = \int_\Omega \bigl|\, f(\omega) \,\bigr| \mu\bigl[\, d\omega \,\bigr].$$

□

Exercise* 19.3. Suppose that $\bigl|\, X, \|-\| \,\bigr|$ is an *infinite dimensional* Banach space and $C_1, C_2 \subset X$ are compacts subsets. Prove that the set

$$C_1 - C_2 := \bigl\{\, x_1 - x_2;\;\; x_i \in C_i,\;\; i = 1, 2 \,\bigr\}.$$

has empty interior. □

Elements of functional analysis

20.1. Hilbert spaces

The Euclidean geometry and topology we have discussed Sections 11.2 and 11.3 have infinite dimensional counterparts with wide ranging applications.

20.1.1. Inner products and their associated norms. Suppose that $\mathbb{K}=\mathbb{R},\mathbb{C}$ and X is a $\mathbb{K}$-vector space, possibly infinite dimensional. For $\lambda\in\mathbb{C}$ we denote by $\bar{\lambda}$ its conjugate. Note that when $\lambda\in\mathbb{R}$ we have $\bar{\lambda}=\lambda$.

Definition 20.1.1. A $\mathbb{K}$-*inner product* on X is a map

$$\langle -,-\rangle : X\times X\to\mathbb{K}.$$

satisfying the following conditions.

(i) For any $x,y\in X$ we have $\langle y,x\rangle=\overline{\langle x,y\rangle}$, where $\bar{\lambda}$ denotes the conjugate of the complex number λ.

(ii) $\forall x,y,z\in X$,

$$\langle x+y,z\rangle=\langle x,z\rangle+\langle y,z\rangle \;\text{ and }\; \langle z,x+y\rangle=\langle z,x\rangle+\langle z,y\rangle.$$

$\forall x,y\in X$ and $\lambda\in\mathbb{K}$ we have

$$\langle\lambda x,y\rangle=\lambda\langle x,y\rangle,\;\;\langle x,\lambda y\rangle=\bar{\lambda}\langle x,y\rangle.$$

(iii) For any $x\in X$, $\langle x,x\rangle\geqslant 0$, with equality iff $x=0$.

A *pre-Hilbert space* over $\mathbb{K}$ is a pair $\big(X,\langle -,-\rangle\big)$, where X is $\mathbb{K}$-vector space and $\langle -,-\rangle$ is a $\mathbb{K}$-inner product. □

Example 20.1.2. (a) Let $X = \mathbb{C}^n$. For $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in $\mathbb{C}^n$ we set

$$\langle x, y \rangle := x_1\bar{y}_1 + \cdots + x_n\bar{y}_n.$$

Then $\langle -, - \rangle$ defined as above is a complex inner product.

(b) Suppose that $(\Omega, \mathcal{S}, \mu)$ is a measured space. Then the space $L^2(\Omega, \mu)$ is a real inner pre-Hilbert space with the inner product

$$\langle f, g \rangle := \int_\Omega fg d\mu.$$

Denote by $\mathcal{L}^2(\Omega, \mu, \mathbb{C})$ the space of measurable functions $(\Omega, \mathcal{S}) \to (\mathbb{C}, \mathcal{B}_\mathbb{C})$ such that $|f| \in \mathcal{L}^2(\Omega, \mu)$

$$\int_\Omega |f(\omega)|^2 \mu\big[\, d\omega\, \big].$$

Note that if $f, g \in \mathcal{L}^2(\Omega, \mu, \mathbb{C})$, then $\big| f\bar{g} \big| = |f| \cdot |fg| \in \mathcal{L}^1(\Omega, \mu)$. We define $L^2(\Omega, \mu, \mathbb{C})$ the quotient of $\mathcal{L}^2(\Omega, \mu, \mathbb{C})$ by the μ-a.e. equality. The map

$$\langle -, - \rangle : L^2(\Omega, \mu, \mathbb{C}) \times L^2(\Omega, \mu, \mathbb{C}) \to \mathbb{C},\ \ \langle f, g \rangle = \int_\Omega f(\omega)\overline{g(\omega)}\mu\big[\, d\omega\, \big]$$

is a complex inner product. □

Let $(X, \langle -, - \rangle)$ be a pre-Hilbert space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. For $x \in X$ the inner product $\langle x, x \rangle$ is a real nonnegative number. We set

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

Observe that for any $\lambda \in \mathbb{K}$ and $x \in \mathbb{R}$ we have

$$\|\lambda x\|^2 = \langle \lambda x, \lambda x \rangle = \lambda\bar{\lambda}\langle x, x \rangle = |\lambda|^2 \|x\|^2$$

so that

$$\|\lambda x\| = |\lambda| \cdot \|x\|, \ \ \forall x \in X, \ \ \lambda \in \mathbb{K}.$$

Theorem 20.1.3 (Cauchy-Schwarz inequality). *Suppose that $\big(X, \langle -, - \rangle\big)$ is a pre-Hilbert space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Then, for any $x, y \in X$ we have*

$$\big|\langle x, y \rangle\big| \leqslant \|x\| \cdot \|y\|. \tag{20.1.1}$$

Moreover

$$\big|\langle x, y \rangle\big| = \|x\| \cdot \|y\|$$

if and only if there exists $\lambda \in \mathbb{K}$ such that $y = \lambda x$.

Proof. We prove the inequality in the case $\mathbb{K} = \mathbb{C}$. The proof in the case $\mathbb{K} = \mathbb{R}$ is identical.

The inequality is obviously true if $x = 0$ so we assume $x \neq 0$. For any complex number λ we have

$$\begin{aligned} 0 \leqslant \langle \lambda x - y, \lambda x - y\rangle &= \langle \lambda x, \lambda x\rangle - \lambda\langle x, y\rangle - \bar{\lambda}\langle y, x\rangle + \langle y, y\rangle \\ &= |\lambda|^2\|x\|^2 - 2\,\mathbf{Re}\,\lambda\langle x, y\rangle + \|y\|^2, \end{aligned}$$

where $\mathbf{Re}\,z$ denotes the real part of a complex number $z = a + \boldsymbol{i}b$,

$$\mathbf{Re}\,z = a = \frac{1}{2}\big(\,z + \bar{z}\,\big).$$

We write $\langle x, y\rangle \in \mathbb{C}$ in polar coordinates,

$$\langle x, y\rangle = re^{\boldsymbol{i}\theta}, \quad r := \big|\langle x, y\rangle\big|.$$

Now choose $\lambda = te^{-\boldsymbol{i}\theta}$, $t \in \mathbb{R}$ so that

$$\lambda\langle x, y\rangle = tr \in \mathbb{R} \;\Rightarrow \mathbf{Re}\,\lambda\langle x, y\rangle = tr.$$

We deduce

$$\|te^{-\boldsymbol{i}\theta}x - y\|^2 = t^2\underbrace{\|x\|^2}_{A} - \underbrace{2r}_{B}\,t + \underbrace{\|y\|^2}_{C}.$$

The quadratic function

$$f(t) = At^2 - Bt + C$$

is nonnegative for any $t \in \mathbb{R}$ and since $A > 0$ this can happen if and only if

$$B^2 \leqslant 4AC \Longleftrightarrow \big|\langle x, y\rangle\big|^2 \leqslant \|x\|^2\|y\|^2.$$

This is precisely (20.1.1). Observe that we have equality iff $(2B)^2 - 4A^2C^2 = 0$, i.e., $f(t) = \|te^{-\boldsymbol{i}\theta}x - y\|^2$ has one real root t_0, i.e., there exists $\lambda = t_0e^{-\boldsymbol{i}\theta} \in \mathbb{C}$ such that $y = \lambda x$. □

Theorem 20.1.4. *Suppose that* $\big(X, \langle -, -\rangle\big)$ *is a pre-Hilbert space over* $\mathbb{K} = \mathbb{R}, \mathbb{C}$. *Then the function*

$$X \ni x \mapsto \|x\| = \sqrt{\langle x, x\rangle} \in [0, \infty)$$

is a norm on X.

Proof. We already know that

$$\|\lambda x\| = |\lambda| \cdot \|x\|, \quad \forall \lambda \in K,\ \forall x \in X,$$

so all that remains to show is

$$\|x + y\| \leqslant \|x\| + \|y\|, \quad \forall x, y \in X.$$

We observe as in the proof of (20.1.1) that

$$\begin{aligned} \|x + y\|^2 = \|x\|^2 + 2\,\mathbf{Re}\,\langle x, y\rangle + \|y\|^2 &\overset{(20.1.1)}{\leqslant} \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 \\ &= \big(\,\|x\| + \|y\|\,\big)^2. \end{aligned}$$

□

Theorem 20.1.5 (Parallelogram Law)**.** *Suppose that* $\big(X, \langle -, - \rangle\big)$ *is a pre-Hilbert space with associated norm* $\|-\|$*. Then*

$$\forall x, y \in X, \quad \|x+y\|^2 + \|x-y\|^2 = 2\big(\,\|x\|^2 + \|y\|^2\,\big). \tag{20.1.2}$$

Proof. The computation in the proof of the Cauchy-Schwarz inequality shows that

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

$$\|x-y\|^2 = \|x\|^2 - 2\,\mathbf{Re}\,\langle x, y\rangle + \|y\|^2.$$

Adding up the two equalities we obtain the parallelogram law. □

Remark 20.1.6. The parallelogram law characterizes pre-Hilbert spaces. More precisely if the normed $(X, \|-\|)$ satisfies (20.1.2), then the norm $\|-\|$ is associated to an inner product. For example, if X is a real vector space, then the inner product is

$$\frac{1}{4}\big(\,\|x+y\|^2 - \|x+y\|^2\,\big).$$

For a proof we refer to K. Yosida [**43**]. □

Proposition 20.1.7. *Suppose that* $\big(H, \langle -, - \rangle\big)$ *is a (real) pre-Hilbert space. Then the inner product*

$$\langle -, - \rangle : H \times H \to \mathbb{R}$$

is continuous, i.e., if $(x_n)_{n\in\mathbb{N}}$ *and* $(y_n)_{n\in\mathbb{N}}$ *are sequences in* H *converging to* x *and respectively* y*, then*

$$\lim_{n\to\infty} \langle x_n, y_n \rangle = \langle x, y \rangle$$

Proof. Observe first that since (x_n) and (y_n) are convergent they are bounded so there exists $C > 0$ such that

$$\|x_n\|, \quad \|y_n\| < C, \quad \forall n \in \mathbb{N}.$$

We have

$$\begin{gathered}\big|\langle x_n, y_n\rangle - \langle x, y\rangle\big| = \big|\langle x_n, y_n\rangle - \langle x, y_n\rangle + \langle x, y_n\rangle - \langle x, y\rangle\big| \\ = \big|\langle x_n - x, y_n\rangle + \langle x, y_n - y\rangle\big| \leqslant \big|\langle x_n - x, y_n\rangle\big| + \big|\langle x, y_n - y\rangle\big| \\ \overset{(20.1.1)}{\leqslant} \|x_n - x\| \cdot \|y_n\| + \|x\| \cdot \|y_n - y\| \leqslant C\|x_n - x\| + \|x\| \cdot \|y_n - y\| \to \infty.\end{gathered}$$

□

Definition 20.1.8. A *Hilbert space* is a pre-Hilbert space $(H, \langle -, - \rangle)$ such that the associated normed space $(H, \|-\|)$ is complete. □

Example 20.1.9. If $(\Omega, \mathcal{S})$ is a measurable space and $\mu : \mathcal{S} \to [0, \infty]$ is a sigma-additive measure then $L^2(\Omega, \mu)$ is a Hilbert space. In particular ℓ_2 is a Hilbert space. □

Definition 20.1.10. Suppose that $\big(H_i, \langle -,-\rangle_i\big)$, $i=0,1$, be two pre-Hilbert spaces over $\mathbb{K}$. A linear map $T : H_0 \to H_1$ is called an *isomorphism of pre-Hilbert spaces* if it is bijective and

$$\langle Tx, Ty\rangle_1 = \langle x, y\rangle_0, \;\; \forall x, y \in H_0. \tag{20.1.3}$$

□

Proposition 20.1.11. *Suppose that* $\big(H_i, \langle -,-\rangle_i\big)$, $i=0,1$ *be two pre-Hilbert spaces over* $\mathbb{K}$ *and* $T : H_0 \to H_1$ *is a linear bijection. Then the following conditions are equivalent.*

(i) *The map* T *is a pre-Hilbert space isomorphism.*

(ii) *The map* T *is an isometry, i.e.,*

$$\|Tx\|_1 = \|x\|_0, \;\; \forall x \in H_0.$$

Proof. The implication (i) ⇒ (ii) follows by setting $x=y$ in (20.1.3).

Conversely, let us assume that T is an isometry. Then, $\forall x, y \in H_0$ we have

$$\|Tx\|_1^2 + 2\,\mathbf{Re}\,\langle Tx, Ty\rangle_1 + \|Ty\|_1^2 = \|T(x+y)\|_1^2 = \|x+y\|_0^2$$
$$= \|x\|_0^2 + 2\,\mathbf{Re}\,\langle x, y\rangle_0 + \|y\|_0^2 = \|Tx\|_1^2 + 2\,\mathbf{Re}\,\langle x, y\rangle_0 + \|Ty\|_1^2.$$

We deduce that

$$\mathbf{Re}\,\langle Tx, Ty\rangle_1 = \mathbf{Re}\,\langle x, y\rangle_0, \;\; \forall x, y \in H_0.$$

This proves the implication (ii) ⇒ (i) when $\mathbb{K} = \mathbb{R}$. If $\mathbb{K} = \mathbb{C}$ and $\boldsymbol{i} = \sqrt{-1}$, then we observe that

$$\mathbf{Im}\,\langle Tx, Ty\rangle_1 = \mathbf{Re}\,\langle T(-\boldsymbol{i}x), Ty\rangle_1 = \mathbf{Re}\,\langle -\boldsymbol{i}x, y\rangle_0 = \mathbf{Im}\,\langle x, y\rangle_0.$$

□

20.1.2. Orthogonal projections. Let H be a pre-Hilbert space. As in the finite dimensional case we observe that if $x, y \in H\setminus\{0\}$, then

$$\frac{\mathbf{Re}\,\langle x, y\rangle}{\|x\|\cdot\|y\|} \in [-1,1]$$

so there exists a unique $\theta \in [0,\pi]$ such that

$$\cos\theta = \frac{\mathbf{Re}\,\langle x, y\rangle}{\|x\|\cdot\|y\|}.$$

We will refer to this θ as the angle between the vectors x, y and we will denote it by $\measuredangle(x,y)$. Note that

$$\boxed{\cos\measuredangle(x,y) = \frac{\mathbf{Re}\,\langle x, y\rangle}{\|x\|\cdot\|y\|}} \;\text{ and }\; \boxed{\mathbf{Re}\,\langle x, y\rangle = \|x\|\cdot\|y\|\cos\measuredangle(x,y)}.$$

Two vectors x, y in a pre-Hilbert space H are called *orthogonal*, and we denote this $x \perp y$, if $\langle x, y\rangle = 0$. Clearly

$$x \perp y \Longleftrightarrow y \perp x.$$

Note that if $x, y \neq 0$, then

$$x \perp y \Longleftrightarrow \measuredangle(x, y) = \frac{\pi}{2}$$

Theorem 20.1.12 (Pythagoras)**.** *If x, y are orthogonal vectors in a pre-Hilbert space $\big(H, \langle -, - \rangle\big)$, then*

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof. The proof is identical to the one of Theorem 11.2.8. We have

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + \underbrace{2\,\mathbf{Re}\langle x, y \rangle}_{=0} + \|y\|^2.$$

□

☞ In the remainder of this section, for simplicity of presentation, we will assume that all Hilbert spaces are real. The complex case is only notationally more complicated.

Definition 20.1.13. For any nonempty subset X of a *Hilbert* space $\big(H, \langle -, - \rangle\big)$ we define its *orthogonal complement* to be the set

$$X^\perp := \big\{ y \in H;\;\; y \perp x,\;\; \forall x \in X \big\}.$$

□

Observe that

$$X^\perp = \bigcap_{x \in X} \{x\}^\perp. \tag{20.1.4}$$

In particular this shows that

$$X_1 \subset X_2 \Rightarrow X_1^\perp \supset X_2^\perp. \tag{20.1.5}$$

Proposition 20.1.14. *For any nonempty subset X of a Hilbert space $\big(H, \langle -, - \rangle\big)$ its orthogonal complement $X^\perp$ is a $\underline{\text{closed vector subspace}}$ of H. Moreover*

$$X^\perp = \big(\boldsymbol{cl}(X)\big)^\perp,$$

where $\boldsymbol{cl}(X)$ denotes the closure of X in H.

Proof. To show that $X^\perp$ is a closed vector subspace it suffices to show that for any $x \in X$ the set $\{x\}^\perp$ is a closed vector subspace of H. Observe first that $\{x\}^\perp$ is a vector subspace. Indeed, if $y, z \perp x$, then

$$\langle y + z, x \rangle = \langle y, x \rangle + \langle z, x \rangle = 0 \Rightarrow (y + z) \perp x$$

Similarly, if $y \perp x$ and $t \in \mathbb{R}$, then $(ty) \perp x$. Thus $\{x\}^\perp$ is a vector subspace.

To prove that $\{x\}^{\perp}$ is closed consider a sequence (y_n) in $\{x\}^{\perp}$ that converges to y. We have to show that $y \perp x$. We have

$$\langle y, x\rangle = \lim_{n\to\infty} \underbrace{\langle y_n, x\rangle}_{=0} = 0.$$

Hence $y \perp x$.

Since $X \subset \boldsymbol{cl}(X)$ we deduce that $\boldsymbol{cl}(X)^{\perp} \subset X^{\perp}$. Conversely let $y \in X^{\perp}$ and $x_* \in \boldsymbol{cl}(X)$. There exists a sequence (x_n) in X such that $x_n \to x_*$. Hence

$$\langle y, x_*\rangle = \lim_{n\to\infty} \langle y, x_n\rangle = 0$$

proving that $y \in \boldsymbol{cl}(X)^{\perp}$. □

Theorem 20.1.15 (Orthogonal projection)**.** *Suppose that U is a* closed *vector subspace of the* Hilbert *space $(H, \langle -,-\rangle)$. Then, for any $x \in H$, there* exists a unique *$x_* \in U$ such that*

$$\|x - x_*\| \leqslant \|x - u\|, \quad \forall u \in U.$$

Moreover, x_* is the unique point in U such that $(x - x_*) \in U^{\perp}$, i.e.,

$$(x - x_*) \perp u, \quad \forall u \in U. \tag{20.1.6}$$

The vector x_* is called the orthogonal projection *of x on U and it is denoted by $P_U x$.*

Proof. Step 1. Existence. Suppose that (u_n) is a sequence in U such that

$$\lim_{n\to\infty} \|x - u_n\| = d := \inf_{u\in U} \|x - u\|.$$

From the parallelogram law we deduce that for any $m, n \in \mathbb{N}$ we have

$$\left\|\frac{1}{2}(x - u_n) + \frac{1}{2}(x - u_m)\right\|^2 + \left\|\frac{1}{2}(u_n - u_m)\right\|^2 = \frac{1}{2}\big(\|x - u_n\|^2 + \|x - u_m\|^2\big).$$

Now observe that

$$\frac{1}{2}(u_n + u_m) \in U, \;\; \frac{1}{2}(x - u_n) + \frac{1}{2}(x - u_m) = x - \frac{1}{2}(u_n + u_m),$$

so

$$d^2 \leqslant \left\|x - \frac{1}{2}(u_n + u_m)\right\|^2 = \left\|\frac{1}{2}(x - u_n) + \frac{1}{2}(x - u_m)\right\|^2.$$

Hence

$$d^2 + \left\|\frac{1}{2}(u_n - n_m)\right\|^2 \leqslant \frac{1}{2}\big(\|x - u_n\|^2 + \|x - u_m\|^2\big).$$

By construction

$$\lim_{m,n\to\infty} \frac{1}{2}\big(\|x - u_n\|^2 + \|x - u_m\|^2\big) = d^2$$

so

$$\lim_{m,n\to\infty} \|u_n - u_m\| = 0.$$

Hence the sequence $(u_n)_{n\in\mathbb{N}}$ is Cauchy and, since H is complete, it converges to a point x_*. Note that $x_* \in U$ because U is closed.

Step 2. Suppose that

$$x_* \in U \ \text{ and } \ \|x - x_*\| = \inf_{u\in U} \|x - u\|.$$

We will show that x_* satisfies (20.1.10). Let $u \in U$. Set

$$f : \mathbb{R} \to \mathbb{R}, \ \ f(t) = \|x - (x_* + tu)\|^2.$$

Note that $f(0) = \|x - x_*\|^2$, ao $f(0) = d^2 \leqslant f(t)$, $\forall t \in \mathbb{R}$, so 0 is a minimum point of f On the other hand

$$f(t) = \|(x - x_*) - tu\|^2 = \|x - x_*\|^2 - 2t\langle x - x_*, u\rangle + t^2\|u\|^2$$

Hence, the function f is differentiable so that

$$0 = f'(0) = -2\langle x - x_*, u\rangle = 0, \ \ \forall u \in U.$$

Step 3. Uniqueness. Suppose there exist $x_*, y_* \in U$ such that

$$\|x - x_*\| = \|x - y_*\| = \inf_{u\in U} \|x - u\|.$$

Then, according to **Step 2** x_*, y_* satisfy (20.1.10). Then $y_* - x_* \in U$

$$\langle x - x_*, y_* - x_*\rangle = \langle x - y_*, y_* - x_*\rangle = 0.$$

Subtracting the two equalities we deduce

$$\|y_* - x_*\|^2 = \langle y_* - x_*, y_* - x_*\rangle = 0,$$

so $x_* = y_*$. □

To any closed vector subspace $U \subset H$ we have associated a map $P_U : H \to H$ such that

$$P_U x \in U, \ \ \forall x \in X,$$

$$\|x - P_U x\| \leqslant \|x - u\|, \ \ \forall u \in U,$$

and $x - P_U x \in U^\perp$.

Proposition 20.1.16. *Let H be a Hilbert space. For any* closed *subspace $U \subset H$ the following hold.*

(i) *The map $P_U : H \to H$ is linear and continuous. Moreover $\|P_U\|_{\mathrm{op}} \leqslant 1$.*

(ii) *$U = \ker(\mathbb{1} - P_U)$, $P_U^2 = P_U$.*

(iii) *$(U^\perp)^\perp = U$.*

(iv) *$P_{U^\perp} = \mathbb{1} - P_U$.*

Proof. (i) Let $x, y \in H$. Set $x_* = P_U x$, $y_* = P_U y$. We have to show that $P_U(x+y) = x_*+y_*$, i.e., $(x+y) - (x_* + y_*) \in U^\perp$. To see this note that $(x - x^*), (y - y_*) \in U^\perp$. On the other hand, $U^\perp$ is a vector subspace cccording to Proposition 20.1.14. Hence

$$(x+y) - (x^* + y^*) = (x - x^*) + (y - y^*) \in U^\perp.$$

A similar argument shows that $P_U(tx) = tP_U x$, $\forall t \in \mathbb{R}$, $x \in H$.

Observe that for any $x \in H$ we have

$$x = (x - P_U x) + P_U x, \;\; P_U x \perp (x - P_U x)$$

and from Pythagoras' Theorem we deduce

$$\|x\|^2 = \|x - P_U x\|^2 + \|P_U x\|^2 \geqslant \|P_U x\|^2, \;\; \forall x \in X.$$

Hence

$$\|P_U x\| \leqslant \|x\|, \;\; \forall x \in X.$$

Theorem 17.1.48 imples that P_U is continuous and $\|P_U\|_{\mathrm{op}} \leqslant 1$.

(ii) Clearly if $x \in U$, then $P_U x = x$ since x is the point is U closest to x. Hence $U \subset \ker(\mathbb{1} - P_U)$. Conversely, if $x \in \ker(\mathbb{1} - P_U)$ then

$$x = P_U x \in U.$$

Hence $U = \ker(\mathbb{1} - P_U)$. We deduce that

$$P_U^2 x = P_U(\, P_U x\,) = P_U x$$

since $P_U x \in U$, $\forall x \in X$.

(iii) Note that

$$\langle u^\perp, u\rangle = 0, \;\; \forall u \in U, \;\; u^\perp \in U^\perp.$$

This proves that $u \in (U^\perp)^\perp$, $\forall u \in U$, i.e., $U \subset (U^\perp)^\perp$.

To prove the opposite inclusion, $(U^\perp)^\perp \subset U$, let $v \in (U^\perp)^\perp$. Set $v_* = P_U v$. Then $v_* \in U$ and $(v - v_*) \in U^\perp$ so that

$$0 = \langle v, v - v_*\rangle = \|v\|^2 - \langle v, v_*\rangle \geqslant \|v\|^2 - \|v\| \cdot \|v_*\| = \|v\|\big(\,\|v\| - \|v_*\|\,\big)$$

Hence $\|v\| \leqslant \|v_*\|$. Pythagoras' Theorem implies

$$\|v_*\|^2 \geqslant \|v\|^2 = \|v - v_*\|^2 + \|v_*\|^2.$$

Hence $\|v - v_*\| = 0$, so $v = v_* \in U$.

(iv) Let $x \in H$. To prove that $P_{U^\perp} x = x - P_U x$ it suffices to show that $x - (x - P_u x) \in (U^\perp)^\perp$. Indeed,

$$x - (x - P_U x) = P_U x \in U = (U^\perp)^\perp.$$

□

Corollary 20.1.17. *Suppose that U is a closed subspace. For any $x \in H$ there exists a unique $u \in U$ and a unique $v \in U^\perp$ such that*

$$x = u + v.$$

Proof. The existence follows from the equality $\mathbb{1} = P_U + P_{U^\perp}$ which yields $x = P_U x + P_{U^\perp} x$. If

$$x = u + v = u' + v', \quad u, u' \in U, \quad v, v' \in U^\perp$$

then $u - u' = v' - v$ so that $u - u' \in U \cap U^\perp$. Hence $(u - u') \perp (u - u')$, i.e.,

$$\|u - u'\|^2 = \langle u - u', u - u' \rangle = 0.$$

□

Corollary 20.1.18. *Suppose that $U \subset H$ is a vector subspace of H, not necessarily closed. Then*

$$\boldsymbol{cl}(U) = (U^\perp)^\perp.$$

In particular, U is dense if and only if $U^\perp = \{0\}$.

Proof. We have

$$U^\perp = \boldsymbol{cl}(U)^\perp$$

and Proposition 20.1.16 (iii) implies

$$\boldsymbol{cl}(U) = \big(\boldsymbol{cl}(U)^\perp\big)^\perp = (U^\perp)^\perp.$$

Note that U is dense iff $\boldsymbol{cl}(U) = H$ or, equivalently, $U^\perp = \boldsymbol{cl}(U)^\perp = H^\perp = 0$.

□

Example 20.1.19. Suppose that U is a finite dimensional subspace of the Hilbert space H. It is then a closed subspace of H; see Exercise 17.19. Fix a basis $e_1, e_2, \dots, e_n$ of U and set

$$U_k := \operatorname{span}\{e_1, \dots, e_k\}$$

Then $\dim U_k = k$. The classical Gram-Schmidt procedure can be described as follows. Define

$$f_1 := \frac{1}{\|e_1\|} e_1$$

so that $\|f_1\| = 1$ and $\operatorname{span}\{f_1\} = U_1$. Next, define

$$u_2 = e_2 - P_{U_1} e_2, \quad f_2 = \frac{1}{\|u_2\|} u_2.$$

Then

$$\|f_2\| = 1, \quad f_2 \perp U_1, \quad \operatorname{span}\{f_1, f_2\} = U_2.$$

Iterating this procedure we obtain a basis $f_1, \dots, f_n$ of U such that

$$f_{k+1} = \frac{1}{\|u_{k+1}\|} u_{k+1}, \quad u_{k+1} = e_{k+1} - P_{U_k} e_{k+1},$$

$$\operatorname{span}\{f_1, \dots, f_k\} = U_k, \quad \forall k = 1, \dots, n,$$

$$\|f_k\| = 1, \quad \forall k = 1, \dots, n, \quad f_i \perp f_j, \quad \forall 1 \leqslant i, j \leqslant n. \tag{20.1.7}$$

We recall that a basis that satisfies (20.1.7) is called an *orthonormal basis*. Note that (20.1.7) can be rewritten as

$$\langle f_j, f_k\rangle = \delta_{jk} = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases} \tag{20.1.8}$$

Thus, the Gram-Schmidt procedure

$$\{e_1, \dots, e_n\} \to \{f_1, \dots, f_n\}$$

converts a basis $\{e_1, \dots, e_n\}$ into an orthonormal basis $\{f_1, \dots, f_n\}$ such that

$$\operatorname{span}\{e_1, \dots, e_k\} = \operatorname{span}\{f_1, \dots, f_k\}, \quad \forall k = 1, \dots, n.$$

If $(e_k)_{1 \leqslant k \leqslant n}$ is any *orthonormal* basis of U, then

$$P_U x = \sum_{k=1}^{n} \langle x, e_k\rangle e_k. \tag{20.1.9}$$

Indeed, note that

$$\left\langle x - \sum_{k=1}^{n} \langle x, e_k\rangle e_k, e_j \right\rangle = \langle x, e_j\rangle - \sum_{k=1}^{n} \langle x, e_k\rangle\langle e_k, e_j\rangle$$

$$\overset{(20.1.8)}{=} \langle x, e_j\rangle - \sum_{k=1}^{n} \langle x, e_k\rangle \delta_{kj} = 0.$$

This proves that

$$\left(x - \sum_{k=1}^{n} \langle x, e_k\rangle e_k \right) \perp e_j, \quad \forall j = 1, \dots, n \Rightarrow \left(x - \sum_{k=1}^{n} \langle x, e_k\rangle e_k \right) \perp \underbrace{\operatorname{span}\{e_1, \dots, e_n\}}_{U}.$$

Hence

$$P_U x = \sum_{k=1}^{n} \langle x, e_k\rangle e_k. \tag{20.1.10}$$

20.1.3. Duality. Recall that H^* denotes the dual of the real Hilbert space H and consists of continuous linear functionals $L : H \to \mathbb{R}$. The norm of such a functional is

$$\|L\|_* := \inf\big\{\, C > 0; \;\; |L(h)| \leqslant C\|h\|, \;\; \forall h \in H \,\big\}.$$

A vector $x \in H$ defines a linear functional

$$x^{\downarrow} : H \to \mathbb{R}, \;\; x^{\downarrow}(h) = \langle h, x\rangle.$$

The Cauchy-Schwarz inequality shows that

$$|x^{\downarrow}(h)| = \big|\langle h, x\rangle\big| \leqslant \|x\| \cdot \|h\|, \;\; \forall h \in H.$$

Hence $x^{\downarrow}$ is a continuous linear functional. We have thus obtained a linear map

$$H \ni x \mapsto x^{\downarrow} \in H^*.$$

This is injective because

$$x^{\downarrow} = 0 \Rightarrow 0 = x^{\downarrow}(x) = \langle x, x\rangle = \|x\|^2 \Rightarrow x = 0.$$

Theorem 20.1.20 (Riesz representation)**.** *Let H be a real Hilbert space. Then the map*

$$H \ni x \mapsto x^{\downarrow} \in H^*$$

is a surjective isometry, i.e., for any continuous linear functional $L \in H^*$ there exists a unique $x \in H$ such that $L = x^{\downarrow}$. Moreover $\|L\|_* = \|x^{\downarrow}\|_* = \|x\|$. We will write $x = L_{\uparrow}$.

Proof. Let $L \in H^*\backslash\{0\}$. Set $U = \ker L$ so U is a closed subspace of H since L is continuous. Moreover $U \neq H$ since $L \neq 0$. Choose $x_0 \in U^{\perp}$, $\|x_0\| = 1$ and set $c_0 = L(x_0)$. Set

$$x := c_0 x_0.$$

Let us show that $U^{\perp} = \operatorname{span}\{x_0\}$. Note that L induces a linear map

$$L : U^{\perp} \to \mathbb{R}$$

This map is injective because $L(h) = 0$ and $h \in U^{\perp}$ implies $h \in U \cap U^{\perp} = \{0\}$. Hence

$$1 = \dim \operatorname{span}\{x_0\} \leqslant \dim U^{\perp} \leqslant 1.$$

Thus x_0 is a basis of $U^{\perp}$.

We claim that $L = x^{\downarrow}$. Let $h \in H$. We decompose it as a sum

$$h = h_0 + h^{\perp}, \;\; h_0 \in U, \;\; h^{\perp} \in U^{\perp}$$

Then

$$L(h) = L(h^{\perp}) \;\text{ and }\; x^{\downarrow}(h) = \langle h, x\rangle = \langle h^{\perp}, x\rangle = x^{\downarrow}(h^{\perp})$$

so it suffices to show that $L(h) = x^{\downarrow}(h)$, $\forall h \in U^{\perp}$. By construction

$$L(x_0) = c_0 = \langle c_0 x_0, x_0\rangle = \langle x, x_0\rangle = x^{\downarrow}(x_0)$$

and since x_0 is a basis of $U^{\perp}$ this proves our claim. We know that $\|L\|_* \leqslant \|x\| = |c_0|$. On the other hand

$$\|L\|_* = \|L\|_* \cdot \|x_0\| \geqslant |L(x_0)| = c_0.$$

□

Remark 20.1.21. The Riesz representation theorem as a complex counterpart. If H is a *complex* Hilbert space, we denote by H^* the space of continuous linear maps $H \to \mathbb{C}$. To each $x \in H$ we associate $x^{\downarrow} \in H^*$ as before,

$$x^{\downarrow}(h) = \langle h, x\rangle, \;\; \forall h \in H.$$

The map $H \ni x \to x^{\downarrow} \in H^*$ is *conjugate linear*, i.e., it is additive and

$$\forall \lambda \in \mathbb{C}, \;\; \forall x \in H \;\; (\lambda x)^{\downarrow} = \bar{\lambda}(x^{\downarrow}).$$

The same argument we used in the real case shows that the map $H \ni x \to x^{\downarrow} \in H^*$ is bijective. □

Remark 20.1.22 (Bra-ket notation)**.** In his foundations of quantum mechanics P. A. M. Dirac (1902-1984) introduced to types of vectors *bra* vectors, denoted $\langle x|$, and *ket* vectors, denoted $|x\rangle$. The bra vectors form a Hilbert space[1] H and the ket vectors are vectors in its topological dual H^*. Thus, if $|\alpha\rangle$ is a linear functional $H \to \mathbb{K}$, then its value on a bra vector $\langle x|$ is denoted $\langle x|\alpha\rangle$. In Dirac's notation, the Riesz map

$$H \ni x \mapsto x^{\downarrow} \in H^*$$

takes the form $\langle x| \mapsto |x\rangle$.

If U is a finite dimensional subspace of H and $e_1, \dots, e_n$ is an orthonormal basis of U, then, using the bra-ket notation, we can rewrite (20.1.10) in the form

$$P_U = \sum_k |e_k\rangle\langle e_k| \Longleftrightarrow P_U\langle x| = \sum_k \langle x|e_k\rangle\langle e_k|.$$

□

Remark 20.1.23. If $(\Omega, \mathcal{S}, \mu)$ is a measured space, then the Riesz representation theorem in the case $H = L^2(\Omega, \mu)$ yields Theorem 19.6.13 in the case $p = 2$, without the assumption of sigma-finiteness on μ. □

Conventions. Let H be a real Hilbert space.

- We will denote the elements of the dual H^* with small cap Greek letters α, β, ξ etc.
- For $\alpha \in H^*$ we will denote by $\alpha_{\uparrow}$ the unique vector $x \in H$ such that $x^{\downarrow} = \alpha$. More explicitly $\alpha_{\uparrow}$ is uniquely determined by the condition

$$\alpha(h) = \langle \alpha_{\uparrow}, h\rangle, \quad \forall h \in H.$$

Suppose that $(H_i, \langle -, -\rangle_i)$, $i = 0, 1$ are two real Hilbert spaces and $B : H_0 \to H_1$ is a bounded linear operator; see Definition 17.1.49. Note that for any continuous linear functional $\xi : H_1 \to \mathbb{R}$ the composition $\xi \circ B : H_0 \to \mathbb{R}$ is a continuous linear functional. We have thus obtained a linear map

$$B^{\vee} : H_1^* \to H_0^*, \quad H_1^* \ni \xi \mapsto B^{\vee}(\xi) := \xi \circ B.$$

We deduce that

$$\|B^{\vee}(\xi)\|_* = \|\xi \circ B\|_* \overset{(17.1.9)}{\leqslant} \|\xi\|_{\mathrm{op}} \cdot \|B\|_{\mathrm{op}} = \|B\|_{\mathrm{op}} \cdot \|\xi\|_*, \tag{20.1.11}$$

so $B^{\vee}$ is a bounded linear operator and

$$\|B^{\vee}\|0 \leqslant \|B\|_{\mathrm{op}}.$$

Using the Riesz representation theorem we can identify H_i^* with H_i and $B^{\vee}$ with a bounded linear operator $B^* : H_1 \to H_0$. More precisely

$$\forall x \in H_1, \quad B^* x = (B^{\vee} x^{\downarrow})_{\uparrow}. \tag{20.1.12}$$

[1] Physicists actually work with complex vector spaces

Let us unwrap the above equality. This means that

$$\boxed{\forall y \in H_0,\ \ \forall x \in H_1 :\ \ \langle B^* x, y\rangle = (B^\vee x^\downarrow)(y) = x^\downarrow(By) = \langle x, By\rangle.} \tag{20.1.13}$$

Definition 20.1.24 (The adjoint)**.** Let $B : H_0 \to H_1$ be a bounded linear operator between the Hilbert spaces H_0 and H_1. The linear operator $B^* : H_1 \to H_0$ defined the equality (20.1.13) is called the *adjoint* of B. When $H_0 = H_1 = H$ and $B^* = B$ we say that the operator B is *self-adjoint* or *symmetric*. □

From the equality (20.1.12) and the Riesz Representation Theorem we deduce

$$\|B^* x\|_0 = \|B^\vee x^\downarrow\|_* \overset{(20.1.11)}{\leqslant} \|B^\vee\|_{\mathrm{op}} \cdot \|x^\downarrow\|_* \leqslant \|B\|_{\mathrm{op}} \cdot \|x\|_1.$$

This shows that B^* is a continuous linear operator and

$$\|B^*\|_{\mathrm{op}} \leqslant \|B\|_{\mathrm{op}}.$$

From the equality (20.1.13) we deduce

$$\forall y \in H_0,\ \ \forall x \in H_1 :\ \ \langle By, x\rangle = \langle y, B^* x\rangle = \langle (B^*)^* y, x\rangle$$

Which shows that

$$B = (B^*)^*.$$

Hence

$$\|B\|_{\mathrm{op}} = \|(B^*)^*\|_{\mathrm{op}} \leqslant \|B^*\|_{\mathrm{op}}.$$

This proves that

$$\|B^*\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}. \tag{20.1.14}$$

20.1.4. Abstract Fourier decompositions. The separable Hilbert spaces resemble very much finite dimensional ones. The goal of this subsection is to present in detail some of their features.

Definition 20.1.25. Suppose that H a Hilbert space and $(e_i)_{i \in I}$ is a collection of vectors in H. The collection is said to be an *orthogonal collection* if it satisfies the following conditions

$$\langle e_i, e_j\rangle = 0,\ \ \forall i, j \in I,\ \ i \neq j.$$

The collection is called an *orthonormal collection* if it is orthogonal and

$$\|e_i\| = 1,\ \ \forall i \in I.$$

An orthogonal collection is called *complete* if $\operatorname{span}\{ e_i;\ \ i \in I \}$ is dense in H. A *Hilbert basis* is a complete orthonormal collection. □

We want to emphasize that $\operatorname{span}\{ e_i;\ \ i \in I \}$ consists of linear combinations of *finitely many* of the vectors e_i.

Theorem 20.1.26. *Any separable Hilbert space admits a Hilbert basis consisting of at most countably many vectors.*

Proof. Suppose that H is a separable Hilbert space of infinite[2] dimension. Fix a countable dense subset of H, $(x_n)_{n\in\mathbb{N}}$. We set

$$U_n = \operatorname{span}\{x_1, \dots, x_n\},\ \ n \in \mathbb{N}.$$

Note that $\dim U_{n+1} \leqslant \dim U_n + 1$ and

$$\operatorname{span}\{x_n;\ \ n\in\mathbb{N}\} = U := \bigcup_{n\in\mathbb{N}} U_n.$$

Note that

$$\lim_{n\to\infty} \dim U_n = \infty.$$

Indeed, if this limit were finite, then $\dim U < \infty$. In particular, U is finite dimensional, thus closed. We deduce that U contains the closure of the set $\{x_n\}_{n\in\mathbb{N}}$ which is the entire Hilbert space H. This is impossible since H is infinite dimensional.

There exists an increasing sequence of natural numbers $(n_k)_{k\in\mathbb{N}}$ such that $\dim U_{n_k} = k$. We set $H_0 = 0$ $H_k := U_{n_k}$, $k \geqslant 1$. We construct a sequence of vectors $(e_k)_{k\in\mathbb{N}}$ as follows.

Choose vectors $h_k \in H_k \backslash H_{k-1}$, $k \in \mathbb{N}$. Define

$$u_k = h_k - P_{H_{k-1}} h_k,\ \ e_k = \frac{1}{\|u_k\|} u_k.$$

Note that for any $k \in \mathbb{N}$ we have

$$e_k \in H_k \backslash H_{k-1},\ \ \|e_k\| = 1,\ \ e_k \perp H_{k-1}.$$

Thus the collection $\{e_k\}_{k\in\mathbb{N}}$ is an orthonormal system such that

$$H_k = \operatorname{span}\{e_1, \dots, e_k\},\ \ \forall k.$$

Note that

$$\operatorname{span}\{e_k;\ \ k\in\mathbb{N}\} = \bigcup_{k\in\mathbb{N}} H_k = \bigcup_{n\in\mathbb{N}} U_n \supset \{x_n;\ \ n\in\mathbb{N}\}$$

so $\operatorname{span}\{e_k;\ \ k\in\mathbb{N}\}$ is dense in H. This shows that $\{e_k\}_{k\in\mathbb{N}}$ is a Hilbert basis. □

Theorem 20.1.27. *Suppose that H is a separable Hilbert space and $(e_n)_{n\in\mathbb{N}}$ is a Hilbert basis. For any $x \in H$ we have*

$$x = \sum_{n\in\mathbb{N}} \langle x, e_n\rangle e_n,\ \ \textit{i.e.},\ \ \lim_{n\to\infty} \left\| x - \sum_{k=1}^{n} \langle x, e_n\rangle e_n \right\| = 0, \tag{20.1.15a}$$

and

$$\|x\|^2 = \sum_{n\in\mathbb{N}} \big|\langle x, e_n\rangle\big|^2. \tag{20.1.15b}$$

[2] The finite dimensional ones are dealt with similarly and this case is discussed in most linear algebra books such as [40].

Proof. For $n \in \mathbb{N}$ define

$$U_n := \operatorname{span}\{\, e_1, \dots, e_n \,\}.$$

Then $U_1 \subset U_2 \subset \cdots$ and their union

$$U = \bigcup_{n \in \mathbb{N}} U_n$$

is dense in U.

Fix $x \in H$. Set

$$x_n = P_{U_n} x = \sum_{k=1}^{n} \langle x, e_k \rangle e_k.$$

Note that

$$\| x - -x_n \| = \operatorname{dist}\big(\, x, U_n \,\big) \geqslant \operatorname{dist}\big(\, x, U_n \,\big).$$

The nonincreasing sequence of nonnegative numbers $d_n = \operatorname{dist}\big(x, U_n \big)$ has a finite limit $d_\infty \geqslant 0$. We claim that this limit is 0.

Indeed, since U is dense in H, there exists an increasing sequence of natural numbers

$$n_1 < n_2 < \cdots$$

and vectors $u_{n_k} \in U_{n_k}$ such that

$$\lim_{k \to \infty} u_{n_k} = x.$$

Hence

$$d_{n_k} = \operatorname{dist}\big(\, x, U_{n_k} \,\big) \leqslant \| x - u_{n_k} \| \to 0 \ \text{ as } k \to \infty.$$

Thus, $d_\infty = 0$ and therefore

$$x_n = \sum_{k=1}^{n} \langle x, e_k \rangle e_k \to x.$$

This proves (20.1.15a).

Next, observe that Pythagoras' Theorem implies

$$\| x \|^2 = \| x - x_n \|^2 + \| x_n \|^2 = \| x - x_n \|^2 + \sum_{k=1}^{n} \big| \langle x, e_n \rangle \big|^2.$$

The equality (20.1.15b) is obtained by letting $n \to \infty$ in the above equalities. □

The equality (20.1.15a) is called the *abstract Fourier decomposition* of x with respect to the Hilbert basis $(e_n)_{n \in \mathbb{N}}$, and the numbers

$$\langle x, e_n \rangle, \quad n \in \mathbb{N},$$

are called the *Fourier coefficients* of x with respect to the basis $(e_n)_{n \in \mathbb{N}}$. The identity (20.1.15b) is known as *Parseval identity*.

Theorem 20.1.28. *Let H be a separable Hilbert space. Then any Hilbert basis has at most countably many vectors.*

Proof. The case $\dim H < \infty$ is a standard linear algebra fact. Assume $\dim H = \infty$. Since H is separable it admits a countable Hilbert basis $(e_n)_{n\in\mathbb{N}}$. Suppose that $(f_i)_{i\in I}$ is another Hilbert basis. For each $n \in \mathbb{N}$ we set

$$I_n := \bigl\{\, i \in I;\ \langle f_i, e_n\rangle \neq 0 \,\bigr\}.$$

Observe that since

$$f_i \neq 0 \ \text{ and } \ f_i = \sum_{n\in\mathbb{N}} \langle f_i, e_n\rangle e_n, \ \ \forall i$$

we deduce that

$$\forall i \in I,\ \ \exists n \in \mathbb{N}\ \ \langle f_i, e_n\rangle \neq 0.$$

Hence

$$I = \bigcup_{n\in\mathbb{N}} I_n.$$

We will prove that each of the sets I_n is at most countable.

Observe that

$$I_n = \bigcup_{k\in\mathbb{N}} I_n^k, \quad I_n^k := \Bigl\{\, i \in I;\ \ |\langle f_i, e_n\rangle| \geqslant \frac{1}{k} \,\Bigr\},$$

Note that

$$I_n^1 \subset I_n^2 \subset I_n^3 \subset \cdots$$

We will show that

$$|I_n^k| \leqslant k^2, \ \ \forall n, k \in \mathbb{N}. \tag{20.1.16}$$

More precisely, we will show that if $J \subset I_n^k$ is a finite subset, then $|J| \leqslant k^2$. Set

$$H_J := \operatorname{span}\bigl\{\, f_j;\ \ j \in J \,\bigr\}.$$

Denote by e_n^J the orthogonal projection of e_n on H_J. Then

$$1 = \|e_n\|^2 \geqslant \|e_n^J\|^2 = \Biggl\| \sum_{j\in J} \langle e_n, f_j\rangle f_j \Biggr\|^2 = \sum_{j\in J} |\langle e_n, f_j\rangle|^2 \overset{J\subset I_n^k}{\geqslant} \sum_{j\in J} \frac{1}{k^2} = \frac{|J|}{k^2}.$$

This shows that I_n is at most countable for any n, so I is at most countable. Since $\dim H = \infty$ we deduce that I is countable. □

Suppose that H is a separable real Hilbert space. Fix a Hilbert basis $(e_n)_{n\in\mathbb{N}}$ of H. To a vector $x \in H$ we associate the sequence of Fourier coefficients with respect to this basis

$$x \mapsto \underline{x} = (x_n)_{n\in\mathbb{N}}, \ \ x_n = \langle x, e_n\rangle.$$

From the Parseval identity we deduce

$$\sum_{n\in\mathbb{N}} x_n^2 = \|x\|^2 < \infty$$

and thus we have a map

$$H \ni x \mapsto \underline{x} \in \ell_2.$$

This map is an isometry, i.e., $\|x\|_H = \|\underline{x}\|_{\ell_2}$. We want to show that this map is surjective.

Let $\underline{y} \in \ell_2$. We want to show that there exists $x \in H$ such that $\underline{x} = \underline{y}$. More precisely we will show that the series

$$\sum_{n\in\mathbb{N}} y_n e_n$$

converges in H. Its sum is then a vector x such that $\underline{x} = \underline{y}$.

Since $\underline{y} \in \ell_2$ we deduce

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

Set

$$Y_n = \sum_{k=1}^{n} y_n e_n, \;\; s_n = \sum_{k=1}^{n} y_k^2.$$

Note that for any $m < n$ we have

$$\|Y_n - Y_m\|^2 = \left\| \sum_{k=m+1}^{n} y_k e_k \right\|^2 = \sum_{k=m+1}^{n} y_k^2 = s_n - s_m.$$

Since the sequence s_n is convergent, it is Cauchy, so

$$\lim_{m,n\to\infty} |s_n - s_m| = 0.$$

Hence

$$\lim_{m,n\to\infty} \|Y_n - Y_m\| = \lim_{m,n\to\infty} \sqrt{|s_n - s_m|} = 0.$$

Thus, the sequence of partial sums $(Y_n)_{n\in\mathbb{N}}$ is Cauchy and, since H is complete, this sequence is convergent. We have thus proved the following result.

Theorem 20.1.29. *Suppose that $(H, \langle -,-\rangle)$ is a separable real Hilbert space and $(e_n)_{n\in\mathbb{N}}$ is a Hilbert basis. Then the correspondence*

$$H \in x \mapsto \big(\langle x, e_n \rangle \big)_{n\in\mathbb{N}} \in \ell_2$$

is an isomorphism of Hilbert spaces. □

20.2. A taste of harmonic analysis

We want to take a brief side trip in our journey through functional analysis to discuss a piece of classical analysis that is responsible for many of the developments in functional analysis, partial differential equations, representation theory and number theory. We barely scratch the surface of this branch of mathematics. For a more in depth presentation of this subject we refer to [**26, 46**], two classic sources in this area of mathematics.

20.2.1. Trigonometric series: L^2-theory. Denote by $\mathbb{T}$ the unit circle in $\mathbb{R}^2$,

$$\mathbb{T} := \big\{ (x,y) \in \mathbb{R}^2; \;\; x^2 + y^2 = 1 \big\}.$$

We think of it as a compact metric subspace of $\mathbb{R}^2$ equipped with the Euclidean metric. On the other hand, we can also think of $\mathbb{T}$ as a closed C^1-curve in the plane with parametrization

$$\theta \ni (-\pi, \pi] \ni\mapsto \boldsymbol{\alpha}(\theta) = \big(\cos\theta, \sin\theta \big)$$

then we have an integral along this curve (see Section 16.1.1)

$$L : C\big(\mathbb{T}\big) \to \mathbb{R}, \;\; f \mapsto L\big[\, f \,\big] = \int_{-\pi}^{\pi} f(\theta) d\theta.$$

This is a nonnegative continuus linear functional on $C(\mathbb{T})$ and thus it corresponds to a finite Borel measure $\boldsymbol{\sigma} : \mathcal{B}_{\mathbb{T}} \to [0, \infty)$.

If we set $\mathbb{I} := (-\pi, \pi] \subset \mathbb{R}$, then we can view $\boldsymbol{\alpha}$ as a continuous bijection

$$\boldsymbol{\alpha} : \mathbb{I} \to \mathbb{T}, \;\; (-\pi, \pi] \ni \theta \mapsto \big(\cos\theta, \sin\theta \big) \in \mathbb{T}.$$

The measure $\boldsymbol{\sigma}$ is then the pushforward measure

$$\boldsymbol{\sigma} := \boldsymbol{\alpha}_{\#}\boldsymbol{\lambda} : \mathcal{B}_{\mathbb{T}} \to [0, \infty)$$

Note that a continuous function $f : \mathbb{T} \to \mathbb{R}$ can be identified with a continuous function $f : [-\pi, \pi] \to \mathbb{R}$ such that $f(-\pi) = f(\pi)$ or, equivalently, with a 2π-periodic continuous function

$$f : \mathbb{R} \to \mathbb{R}, \;\; f(x + 2\pi) = f(x), \;\; \forall x \in \mathbb{R}.$$

For any $n \in \mathbb{N}$ we define

$$u_n, v_n : \mathbb{T} \to \mathbb{R}, \;\; u_n(\theta) = \cos n\theta, \;\; v_n(\theta) = \sin n\theta, \;\; \theta \in (-\pi, \pi].$$

We set $u_0 : \mathbb{T} \to \mathbb{R}$, $u_0(\theta) = 1$, $\forall\theta$. We denote by $\mathcal{T} \subset C(\mathbb{T})$ the subspace spanned by the functions u_m, v_n, $n \geqslant 0$, $m > 0$. As mentioned in Example 17.4.13, the functions in $\mathcal{T}$ are called *trigonometric polynomials*, and have the form

$$p(\theta) = a_0 + \sum_{k=1}^{n} \big(a_n \cos k\theta + b_n \sin k\theta \big).$$

As shown in Example 17.4.13, the Stone-Weierstrass theorem implies that the space $\mathcal{T}$ is actually an algebra of functions dense in $C(\mathbb{T})$ with respect to the sup-norm. Corollary 19.4.15 implies that the space $\mathcal{T}$ is dense in $L^2(\mathbb{T}, \boldsymbol{\sigma})$.

Lemma 20.2.1. *The collection of functions*

$$\big\{ u_m, v_n; \;\; m \geqslant 0, \;\; n > 0 \big\}$$

is a complete orthogonal collection of $L^2(\mathbb{T})$. More precisely, we have

$$\|u_0\|^2_{L^2} = 2\pi, \;\; \langle u_0, u_m \rangle = \langle u_0, v_n \rangle = 0, \;\; \forall m, n > 0, \tag{20.2.1a}$$

$$\|u_m\|^2_{L^2} = \|v_n\|^2_{L^2} = \pi, \;\; \forall m, n > 0, \tag{20.2.1b}$$

$$\langle u_m, u_n \rangle = \langle v_m, v_n \rangle = 0, \;\; \forall m, n > 0, \;\; m \neq 0. \tag{20.2.1c}$$

$$\langle u_m, v_n \rangle = 0, \;\; \forall m \geqslant 0, \;\; n > 0. \tag{20.2.1d}$$

Proof. Note that

$$\|u_0\|^2_{L^2} = \int_{-\pi}^{\pi} d\theta = 2\pi, \;\; \langle u_0, u_n \rangle = \int_{-\pi}^{\pi} \cos n\theta d\theta = 0, \;\; \forall n > 0.$$

The equality $\langle u_0, v_n \rangle = 0$ is proved in a similar fashion.

Next observe that for $n > 0$ we have

$$u_n(\theta)^2 - v_n(\theta)^2 = (\cos n\theta)^2 - (\sin n\theta)^2 = \cos(2n\theta)$$

and thus

$$\|u_n\|^2_{L^2} - \|v_n\|^2_{L^2} = \int_{-\pi}^{\pi} \big(u_n^2 - v_n^2 \big)\, d\theta = \int_{-\pi}^{\pi} \cos(2n\theta)\, d\theta = 0$$

Hence $\|u_n\|^2_{L^2} = \|v_n\|^2_{L^2}$. On the other hand,

$$\|u_n\|^2_{L^2} + \|v_n\|^2_{L^2} = \int_{-\pi}^{\pi} \underbrace{\big(u_n^2 + v_n^2 \big)}_{=1}\, d\theta = 2\pi.$$

Observe that

$$\langle u_m, u_n\rangle - \langle v_m, v_n\rangle = \int_{-\pi}^{\pi} \big(\cos m\theta \cos n\theta - \sin m\theta \sin n\theta \big)\, d\theta = \int_{-\pi}^{\pi} \cos(m+n)\theta\, d\theta = 0,$$

$$\langle u_m, u_n\rangle + \langle v_m, v_n\rangle = \int_{-\pi}^{\pi} \big(\cos m\theta \cos n\theta + \sin m\theta \sin n\theta \big)\, d\theta = \int_{-\pi}^{\pi} \cos(m-n)\theta\, d\theta = 0.$$

This proves (20.2.1c). The equalities (20.2.1d) are proved in a similar fashion using the trigonometric identities in Section 5.7. □

We set

$$\bar{u}_0 := \frac{1}{\sqrt{2\pi}},\;\; \bar{u}_n = \frac{1}{\sqrt{\pi}} u_n,\;\; \bar{v}_n = \frac{1}{\sqrt{\pi}} v_n$$

Lemma 20.2.1 shows that the collection

$$\big\{\, \bar{u}_n,\; \bar{v}_n,\;\; m \geqslant 0,\; n > 0 \,\big\}$$

is an orthonormal system that spans $\mathfrak{T}$. Since $\mathfrak{T}$ is dense in $L^2(\mathbb{T}, \boldsymbol{\sigma})$ we deduce that the above collection is a Hilbert basis of $L^2(\mathbb{T}, \boldsymbol{\sigma})$. Thus, to any $f \in L^2(\mathbb{T}, \boldsymbol{\sigma})$ there is an associated Fourier series

$$\langle f, \bar{u}_0\rangle \bar{u}_0 + \sum_{n\in\mathbb{N}} \big(\langle f, \bar{u}_n\rangle \bar{u}_n + \langle f, \bar{v}_n\rangle \bar{v}_n \;\big)$$

that converges in L^2 to f. Let us describe this series more explicitly. For $m \geqslant 0$ and $n > 0$ we set

$$a_m = a_m(f) = \langle f, u_m\rangle = \int_{-\pi}^{\pi} f(\theta) \cos m\theta\, d\theta,\;\; b_n = b_n(f) = \int_{-\pi}^{\pi} f(\theta) \sin n\theta\, d\theta.$$

Observe that

$$\langle f, \bar{u}_0\rangle \bar{u}_0 = \frac{1}{2\pi} \langle f, u_0\rangle u_0 = \frac{1}{2\pi} a_0 u_0,$$
$$\langle f, \bar{u}_n\rangle \bar{u}_n = \frac{1}{\pi} \langle f, u_n\rangle u_n = \frac{1}{\pi} a_n u_n,$$

and, similarly

$$\langle f, \bar{v}_n\rangle \bar{v}_n = \frac{1}{\pi} b_n v_n.$$

Thus the associated Fourier series is

$$\frac{a_0}{2\pi} + \frac{1}{\pi} \sum_{n\in\mathbb{N}} \big(a_n \cos n\theta + b_n \sin n\theta \big).$$

Its partial sums,

$$S_n[\,f\,] = \frac{a_0(f)}{2\pi} + \frac{1}{\pi}\sum_{k=1}^{n}\big(\,a_k(f)\cos k\theta + b_k(f)\sin k\theta\,\big),\ \ n \in \mathbb{N},$$

converge in L^2 to $f(\theta)$. Moreover, Parseval's identity takes the form

$$\int_{-\pi}^{\pi} f(\theta)^2 d\theta = |\langle f, \bar{u}_0\rangle|^2 + \sum_{n\in\mathbb{N}}\big(\,|\langle f, \bar{u}_n\rangle|^2 + |\langle f, \bar{v}_n\rangle|^2\,\big) = \frac{a_0^2}{2\pi} + \frac{1}{\pi}\sum_{n\in\mathbb{N}}\big(\,a_n^2 + b_n^2\,\big). \tag{20.2.2}$$

This identity has surprising consequences.

Example 20.2.2 (**A computation of Euler).** Riemann's zeta function is

$$\zeta : (1,\infty) \to \mathbb{R},\ \ \zeta(r) = \sum_{n\in\mathbb{N}}\frac{1}{n^r}$$

As explained in Example 4.6.7(b), the above series is convergent for $r > 1$. We want to compute $\zeta(2k)$, $k \in \mathbb{N}$.

The function $p_1 : (-\pi,\pi] \to \mathbb{R}$, $p_1(x) = x$ is bounded and thus defines a (discontinuous) L^2 function on $\mathbb{T}$. Note that

$$\|\mu_1\|_{L^2}^2 = \int_{-\pi}^{\pi} x^2 dx = \frac{2\pi^3}{3}.$$

Observe that for any $n \in \mathbb{N}$ the function $x\cos nx$ is odd so

$$a_n(p_1) = \int_{-\pi}^{\pi} x\cos nx\, dx = 0,\ \ \forall n.$$

On the other hand, using the equality $\cos n\pi = (-1)^n$, we deduce

$$b_n(p_1) = \int_{-\pi}^{\pi} x\sin nx\, dx = \frac{1}{n}\big(-x\cos nx\,\big)\Big|_{-\pi}^{\pi} + \underbrace{\frac{1}{n}\int_{-\pi}^{\pi}\cos nx\, dx}_{=0} = \frac{2\pi(-1)^{n+1}}{n}. \tag{20.2.3}$$

Hence

$$\frac{1}{\pi}b_n^2 = \frac{4\pi}{n^2},$$

and

$$\frac{2\pi^3}{3} = 4\pi\sum_{n\in\mathbb{N}}\frac{1}{n^2}.$$

We have thus obtained the famous identity first discovered by L. Euler

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \zeta(2). \tag{20.2.4}$$

The value $\zeta(2)$ was determine by analyzing the Fourier series of $p_1(x)$. It turns out that the value $\zeta(2m)$ is obtained from Parseval's identity applied to the Fourier series of a polynomial of degree m, the Bernoulli polynomial $B_m(x)$. □

Example 20.2.3 (**Bernoulli polynomials and zeta functions**). Denote by $\mathbb{R}[x]$ the space of polynomials in one real variable x with real coefficients. Define the (shifted) *Bernoulli operator*

$$\mathcal{B}:\mathbb{R}[x]\to\mathbb{R}[x],\;\;\mathbb{R}[x]\ni p\mapsto Bp\in\mathbb{R}[x],\;\;\mathcal{B}p(x)=\frac{1}{2\pi}\int_x^{x+2\pi}p(s)ds.$$

Lemma 20.2.4. *The Bernoulli operator $\mathcal{B}$ is bijective.*

Proof. We set $p_n(x)=x^n$ and we denote by $\mathbb{R}[x]_n$ the space of polynomials of degree $\leqslant n$

$$\mathbb{R}[x]_n=\operatorname{span}\{p_0,\dots,p_n\}.$$

Note that

$$\mathcal{B}p_n(x)=\frac{1}{2\pi(n+1)}\Big((x+2\pi)^{n+1}-x^{n+1}\Big)=x^n+\text{lower order terms}.$$

Hence

$$\mathcal{B}\mathbb{R}[x]_n\subset\mathbb{R}[x]_n,\;\;\mathcal{B}\mu_n-\mu_n\in\mathbb{R}[x]_{n-1}.$$

Thus, the difference $J=\mathbb{1}-\mathcal{B}$ is nilpotent when restricted to $\mathbb{R}[x]_n$. More precisely $J^{n+1}=0$. Hence, on $\mathbb{R}[x]_n$ we have

$$\mathbb{1}=\mathbb{1}-J^{n+1}=(\mathbb{1}-J)(1+J+\cdots+J^n)=\mathcal{B}(1+J+\cdots+J^n).$$

Hence the map $\mathcal{B}:\mathbb{R}[x]_n\to\mathbb{R}[x]_n$ is bijective $\forall n$. □

The degree n *Bernoulli polynomial* B_n is the polynomial uniquely determined by the equality

$$\mathcal{B}B_n(x)=\frac{1}{2\pi}\int_x^{x+2\pi}B_n(s)ds=q_n(x):=\left(\frac{x+\pi}{2\pi}\right)^n,\;\;\forall n\geqslant 0,\;\;\forall x\in\mathbb{R}. \tag{20.2.5}$$

For example

$$B_0(x)=1,\;\;B_1(x)=\frac{x}{2\pi}. \tag{20.2.6}$$

Define

$$D,\Delta:\mathbb{R}[x]\to\mathbb{R}[x],\;\;Dp(x)=\frac{dp}{dx},\;\;\Delta p(x)=\frac{1}{2\pi}\big(p(x+2\pi)-p(x)\big).$$

Note that $\Delta=\mathcal{B}D$. Derivating (20.2.5) we deduce

$$\Delta B_n=\frac{n}{2\pi}q_{n-1}\Rightarrow\mathcal{B}DB_n=\frac{n}{2\pi}q_{n-1}\Rightarrow\mathcal{B}(DB_n)=\frac{n}{2\pi}\mathcal{B}(B_{n-1}).$$

Since $\mathcal{B}$ is injective we deduce

$$B_n'=\frac{n}{2\pi}B_{n-1},\;\;\forall n\in\mathbb{N}. \tag{20.2.7}$$

Observe that if we set $B_n^-(x):=B_n(-x)$, then

$$\mathcal{B}B_n^-(x)=\frac{1}{2\pi}\int_x^{x+2\pi}B_n(-s)ds=\frac{1}{2\pi}\int_{-x-2\pi}^{-x}B_n(t)dt$$

$$= q_n(-x-2\pi) = \left(\frac{-x-\pi}{2\pi}\right)^n = (-1)^n q_n(x) = (-1)^n \mathcal{B} B_n(x)$$

Hence

$$B_n(-x) = (-1)^n B_n(x), \quad \forall n \geqslant 0, \quad x \in \mathbb{R}.$$

Thus, the polynomials B_{2k} are even functions, while the polynomials B_{2k-1} are odd functions. Observe that $B_0(x) = 1$ and the polynomial B_1 of degree one is odd so it must have the form $B_1(x) = cx$. From the equality (20.2.7) we deduce

$$B_0(x) = 1, \;\; B_1(x) = \frac{1}{2\pi}x.$$

Form the equality $\Delta B_n = \frac{n}{2\pi} q_{n-1}$ we deduce that

$$\forall n > 1, \quad B_n(\pi) - B_n(-\pi) = n q_{n-1}(-\pi) = 0.$$

Define

$$a_{n,m} := a_n(B_m) = \int_{-\pi}^{\pi} B_m(x)\cos nx\, dx, \;\; b_{n,m} := b_n(B_m) = \int_{-\pi}^{\pi} B_m(x)\sin nx\, dx,$$

$$z_{n,m} := a_{n,m} + \boldsymbol{i} b_{n,m} = \int_{-\pi}^{\pi} B_m(x) e^{n\boldsymbol{i}x}\, dx.$$

Observe that

$$|a_{n,m}|^2 + |b_{n,m}|^2 = |z_{n,m}|^2.$$

We compute $z_{n,m}$ by induction on n. We have

$$z_{0,m} = \int_{-\pi}^{\pi} B_m(x)dx = 2\pi q_m(0) = \begin{cases} 2\pi, & m = 0, \\ 0, & m > 0. \end{cases}$$

$$z_{n,1} = \frac{1}{2\pi}\int_{-\pi}^{\pi} x e^{\boldsymbol{i}nx} dx = \frac{\boldsymbol{i}}{2\pi}\int_{-\pi}^{\pi} \pi x \sin nx dx \overset{(20.2.3)}{=} \frac{(-1)^{m+1}\boldsymbol{i}}{n}, \quad n > 0,$$

$$z_{n,0} = \int_{-\pi}^{\pi} e^{\boldsymbol{i}nx} dx = \begin{cases} 2\pi, & m = 0, \\ 0, & m > 0. \end{cases}, \quad n > 0.$$

For $m > 1$ we have

$$z_{n,m} = \frac{1}{m\boldsymbol{i}}\int_{-\pi}^{\pi} B_m(x)\frac{d}{dx}\big(\, e^{\boldsymbol{i}nx}\,\big)\, dx$$

$$= \frac{1}{n\boldsymbol{i}}\Big(e^{n\pi\boldsymbol{i}} B_m(\pi) - e^{-n\pi\boldsymbol{i}} B_m(-\pi)\Big) - \frac{1}{n\boldsymbol{i}}\int_{-\pi}^{\pi} B_m'(x) e^{\boldsymbol{i}nx} dx$$

$$= \frac{(-1)^n}{n\boldsymbol{i}} \underbrace{\Big(B_m(\pi) - B_m(-\pi)\Big)}_{=0} - \frac{1}{n\boldsymbol{i}}\int_{-\pi}^{\pi} B_m'(x) e^{\boldsymbol{i}nx} dx$$

$$= \frac{m\boldsymbol{i}}{2\pi n}\int_{-\pi}^{\pi} B_{m-1}(x) e^{\boldsymbol{i}nx}\, dx = \frac{m\boldsymbol{i}}{2\pi n} z_{n,m-1}.$$

We deduce inductively that

$$z_{n,m}=\left(\frac{\boldsymbol{i}}{2\pi n}\right)^{m-1}m(m-1)\cdots 2\cdot z_{n,1}=2\pi\frac{m!\boldsymbol{i}^n}{(2\pi n)^m}=2\pi m!\left(\frac{\boldsymbol{i}}{2\pi n}\right)^m$$

Parseval's formula implies that for any $n\in\mathbb{N}$ we have

$$\int_{-\pi}^{\pi}B_m(x)^2dx=\frac{1}{2\pi}|z_{n,0}|^2+\frac{1}{\pi}\sum_{m\in\mathbb{N}}|z_{n,m}|^2=4\pi(m!)^2\sum_{n\in\mathbb{N}}\frac{1}{(2\pi n)^{2m}}.$$

Hence

$$1+\frac{1}{2^{2m}}+\frac{1}{3^{2m}}+\cdots=\frac{(2\pi)^{2m}}{4\pi(m!)^2}\int_{-\pi}^{\pi}B_m(x)^2dx.$$

We can be even more precise. Set

$$\beta_n:=B_n(-\pi)\ \forall n\geqslant 0.$$

The numbers β_n also known as the *Bernoulli numbers*. Since $B_n(\pi)=B_n(-\pi)$ for $n>1$ we deduce that s $B_n(\pi)=B_n(-\pi)=\beta_n$, $\forall n\geqslant 2$. For $n\geqslant 1$ we have

$$\int_{-\pi}^{\pi}B_n(x)B_0(x)dx=\int_{-\pi}^{\pi}B_n(x)dx=2\pi q_n(-\pi)=0.$$

More generally, for $n\geqslant m>1$ we have

$$I_{n,m}:=\int_{-\pi}^{\pi}B_n(x)B_mdx=\frac{2\pi}{n+1}\int_{-\pi}^{\pi}B_{n+1}'B_mdx$$

$$=\frac{2\pi}{n+1}\underbrace{\big(B_{n+1}(\pi)B_m(\pi)-B_{n+1}(-\pi)B_m(-\pi)\big)}_{=0}-\frac{m}{n+1}\int_{-\pi}^{\pi}B_{n+1}(x)B_{m-1}(x)dx$$

$$=-\frac{m}{n+1}I_{n+1,m-1}.$$

Hence, for $n>1$ we have

$$\int_{-\pi}^{\pi}B_n(x)^2dx=I_{n,n}=-\frac{n}{n+1}I_{n+1,m-1}=\frac{n(n-1)}{(n+1)(n+2)}I_{n+2,n-2}=\cdots$$

$$=(-1)^{n-1}\frac{n(n-1)\cdots 2}{(n+1)(n+2)\cdots(2n-1)}I_{2n-1,1}.$$

We have

$$I_{2n-1,1}=\int_{-\pi}^{\pi}B_{2n-1}(x)B_1(x)dx=\frac{2\pi}{2n}\int_{-\pi}^{\pi}B_{2n}'(x)B_1(x)dx=\frac{1}{2n}\int_{-\pi}^{\pi}B_{2n}'(x)xdx$$

$$=\frac{\pi}{2n}\big(B_{2n}(\pi)+B_{2n}(-\pi)\big)-\frac{1}{2n}\underbrace{\int_{-\pi}^{\pi}B_{2n(x)}dx}_{=0}=\frac{\beta_{2n}\pi}{n}.$$

Hence

$$\int_{-\pi}^{\pi}B_m(x)^2=(-1)^{m-1}\frac{m(m-1)\cdots 2\pi\beta_{2m}}{(m+1)(m+2)\cdots(2m-1)m}=(-1)^{m-1}\frac{2\pi\beta_{2m}}{\binom{2m}{m}}. \tag{20.2.8}$$

We deduce

$$\boxed{1+\frac{1}{2^{2m}}+\frac{1}{3^{2m}}+\cdots=(-1)^{m-1}\frac{(2\pi)^{2m}\beta_{2m}}{2(2m)!}. \qquad (20.2.9)}$$

Recall (see page 82) Riemann's zeta function

$$\zeta:(1,\infty)\to\mathbb{R},\ \ \zeta(r)=\sum_{n=1}^{\infty}\frac{1}{n^r}.$$

The last equality can be restated succinctly

$$\boxed{\zeta(2m)=(-1)^{m-1}\frac{(2\pi)^{2m}\beta_{2m}}{2(2m)!},\ \ \forall n\in\mathbb{N} \qquad (20.2.10)}$$

There is a very convenient recurrence formula relating the Bernoulli numbers.

Observe that

$$D^kB_n=\frac{(n)_k}{(2\pi)^k}B_{n-k},\ \ (n)_k:=n(n-1)\cdots(n-k+1),\ \ \forall k,n\in\mathbb{N},\ \ n\geqslant k.$$

Using Taylor's expansion at $x_0=-\pi$ we deduce

$$\begin{aligned}B_n(x)=\sum_{k=0}^{n}\frac{1}{k!}D^kB_n(-\pi)(x+\pi)^k&=\sum_{k=0}^{n}\frac{(n)_k}{k!}B_{n-k}(-\pi)\left(\frac{x+\pi}{2\pi}\right)^k\\&=\sum_{k=1}^{n}\binom{n}{k}\beta_{n-k}\left(\frac{x+\pi}{2\pi}\right)^k.\end{aligned}\qquad (20.2.11)$$

Consider the formal power series

$$B_t(x)=\sum_{n\geqslant 0}B_n(x)\frac{t^n}{n!},\ \ \beta(t)=\sum_{n\geqslant 0}\frac{\beta_nt^n}{n!}.$$

Note that

$$B_t(-\pi)=\beta(t).$$

The equality (20.2.11) is equivalent to

$$e^{t\frac{(x+\pi)}{2\pi}}\beta(t)=B_t(x). \qquad (20.2.12)$$

From the equalities,

$$B_n(x+2\pi)-B_n(x)=2\pi\Delta B_n=n\left(\frac{x+\pi}{2\pi}\right)^{n-1},\ \ \forall n\geqslant 1,$$

we deduce

$$\frac{t^n}{n!}\big(B_n(x+2\pi)-B_n(x)\big)=\frac{t}{2\pi}\frac{t^{n-1}(x+\pi)^{n-1}}{(2\pi)^{n-1}(n-1)!}. \qquad (20.2.13)$$

Summing over n we deduce

$$B_t(x+2\pi)-B_t(x)=te^{\frac{t(x+\pi)}{2\pi}}.$$

On the other hand, the equality (20.2.12) implies

$$B_t(x+2\pi)-B_t(x)=e^{\frac{t(x+\pi)}{2\pi}}\big(e^t-1\big)\beta(t).$$

Hence

$$te^{\frac{t(x+\pi)}{2\pi}}=e^{\frac{t(x+\pi)}{2\pi}}\big(e^t-1\big)\beta(t),$$

i.e.,

$$t=\beta(t)(e^t-1).$$

In other words, $\beta(t)$ is the Taylor series of $\frac{t}{e^t-1}$. From a computational point of view, it is more convenient to interpret the equality $t=\beta(t)(e^t-1)$ as defining a recurrence relation on the Bernoulli numbers. More precisely, we have

$$\beta_0=1,\;\;\beta_1=B_1(-\pi)=-\frac{1}{2},$$

$$\sum_{k=1}^{n}\beta_{n-k}\binom{n}{k}=0,$$

or more explicitly

$$\binom{n}{1}\beta_{n-1}+\binom{n}{2}\beta_{n-2}+\cdots+\binom{n}{n}\beta_0=0,$$

so that

$$\boxed{\beta_{n-1}=-\frac{1}{n}\left(\binom{n}{2}\beta_{n-2}+\cdots+\binom{n}{n}\beta_0\right).}$$

Here are the first few values of the Bernoulli numbers.

n	0	1	2	4	6	8	10
β_n	1	$-\frac{1}{2}$	$\frac{1}{6}$	$-\frac{1}{30}$	$\frac{1}{42}$	$-\frac{1}{30}$	$\frac{5}{66}$

.

All the Bernoulli numbers β_{2k+1}, $k\geqslant 1$ are zero. To see this note that

$$\beta(t)-\beta_1 t=\frac{t}{e^t-1}+\frac{t}{2}=t\frac{e^t+1}{e^t-1}=t\frac{e^{t/2}+e^{-t/2}}{e^{t/2}-e^{-t/2}}=\frac{t\cosh t}{\sinh t}.$$

The last function is even so all its Taylor coefficients of odd degree are zero. □

Remark 20.2.5. The above definition of Bernoulli polynomials differs from the traditional one. The traditional Bernoulli polynomials $\overline{B}_n$ are related to the polynomials B_n we defined above via the change of variables $x=\pi(2y-1)$, or $y=\frac{x+\pi}{2\pi}$, so that

$$\overline{B}_n(y)=B_n\big(\pi(2y-1)\big).$$

Thus $\beta_n=B_n(-\pi)=\overline{B}_n(0)$

$$x^n=\int_x^{x+1}\overline{B}_n(t)dt,\;\;\overline{B}_n'(x)=n\overline{B}_{n-1}(x),\;\;\overline{B}_n(x+1)-\overline{B}_n(x)=nx^{n-1}\tag{20.2.14a}$$

$$\sum_{n\geqslant 0}\overline{B}_n(x)\frac{t^n}{n!}=e^{tx}\beta(t),\;\;\beta(t)=\frac{t}{e^t-1}.\tag{20.2.14b}$$

$$\overline{B}_n(y) = \sum_{k=1}^{n} \binom{n}{k} \beta_{n-k} y^k. \tag{20.2.14c}$$

In particular,

$$\frac{1}{k+1}\big(\overline{B}_{k+1}(n) - \overline{B}_{k+1}(0)\big) = 1^k + 2^k + \cdots + (n-1)^k, \ \ \forall k, n \in \mathbb{N}, \ \ n \geqslant 2.$$

Here are a few of these polynomials.

$$\begin{gathered} \overline{B}_1(x) = x - \frac{1}{2}, \ \ \overline{B}_2(x) = x^2 - x + \frac{1}{6}, \ \ \overline{B}_3(x) = x^3 - \frac{3\,x^2}{2} + \frac{x}{2}, \\ \overline{B}_4(x) = x^4 - 2\,x^3 + x^2 - \frac{1}{30}, \ \ \overline{B}_5(x) = x^5 - \frac{5\,x^4}{2} + \frac{5\,x^3}{3} - \frac{x}{6}. \end{gathered} \tag{20.2.15}$$

For example

$$\bar{B}_3(x) = \frac{x}{2}\big(2x^2 - 3x + 1\big) = \frac{1}{2}x(2x-1)(x-1),$$

$$\sum_{k=1}^{n} k^2 = \frac{1}{3}\bar{B}_3\big(n+1\big) = \frac{1}{6}n(n+1)(2n+1).$$

Similarly,

$$\bar{B}_4(x) - \bar{B}_4(0) = x^2(x-1)^2$$

so

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

The confirm the claims in Exercise 3.4. □

20.2.2. The Fourier series of an L^1 function. For our further developments it is convenient to allow complex valued functions in our considerations. Observe first that $\mathbb{T}$ can be identified with set of complex numbers of norm 1 and, as such, it has a group structure with respect to the multiplication of complex numbers. An element of $\mathbb{T}$ can be identified with the complex number $e^{\boldsymbol{i}x} = \cos x + \boldsymbol{i}\sin x$, $x \in (-\pi, \pi]$.

For every real valued function $f \in L^2(\mathbb{T})$ we defined its Fourier coefficients

$$a_n(f) = \int_{-\pi}^{\pi} f(x)\cos nx\,dx, \ \ b_m(f) = \int_{-\pi}^{\pi} f(x)\sin mx\,dx, \ \ m, n \in \mathbb{Z},$$

where

$$a_{-n}(f) = a_n(f), \ \ b_{-m}(f) = -b_m(f).$$

Note that

$$a_{-n}(f)\cos(-nx) = a_n(f)\cos nx, \ \ b_{-n}(f)\sin(-nx) = b_m(f)\sin nx, \ \ \forall n \in \mathbb{Z}.$$

The associated Fourier series is

$$\frac{1}{2\pi}a_0(f) + \frac{1}{\pi}\sum_{n\in\mathbb{N}}\big(a_n(f)\cos nx + b_n(f)\sin nx\big)$$

$$= \frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\big(\, a_n(f)\cos nx + b_n(f)\sin nx\,\big).$$

The partial sums of this series are

$$S_n\big[\, f\,\big] = \frac{1}{2\pi}a_0(f) + \frac{1}{\pi}\sum_{k=1}^{n}\big(\, a_k(f)\cos kx + b_k(f)\sin kx\,\big)$$

$$= \frac{1}{2\pi}\sum_{|k|\leqslant n}\big(\, a_k(f)\cos kx + b_k(f)\sin kx\,\big),$$

and they converge in L^2 to f.

For complex valued functions $f \in L^2(\mathbb{T},\mathbb{C})$ we define the Fourier coefficients

$$\widehat{f}(n) := \int_{-\pi}^{\pi} f(\theta)e^{-\boldsymbol{i}n\theta}\, d\theta, \;\; n\in\mathbb{Z}.$$

The complex Fourier series of $f = u + \boldsymbol{i}v \in L^2(\mathbb{T},\mathbb{C})$ is

$$\frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\widehat{f}(n)e^{\boldsymbol{i}nx}, \;\; \widehat{f}(n) = \int_{-\pi}^{\pi} f(\theta)e^{-\boldsymbol{i}n\theta}\, d\theta.$$

Its partial sums are

$$S_n^{\mathbb{C}}\big[\, f\,\big] = \frac{1}{2\pi}\sum_{|k|\leqslant n}\widehat{f}(k)e^{\boldsymbol{i}kx} = \frac{1}{2\pi}\sum_{|k|\leqslant n}\big(\, \widehat{u}(k) + \boldsymbol{i}\widehat{v}(k)\,\big)e^{\boldsymbol{i}kx}$$

Now observe that $\widehat{u}(0) = a_0(u)$ while for $k > 0$

$$\widehat{u}(k)e^{\boldsymbol{i}kx} + \widehat{u}(-k)e^{-\boldsymbol{i}kx} = 2\big(\, a_k(u)\cos kx + b_k(u)\sin kx\,\big),$$

so

$$S_n^{\mathbb{C}}(f) = S_n\big[\, u\,\big] + \boldsymbol{i}S_n\big[\, v\,\big]$$

proving that

$$\lim_{n\to\infty}\|S_n^{\mathbb{C}}(f) - f\|_{L^2} = 0.$$

Let us observe that

$$\widehat{f}(n) = a_n(f) - \boldsymbol{i}b_n(f) = a_n(u) + b_n(v) - \boldsymbol{i}\big(\, a_n(v) + b_n(u)\,\big).$$

Hence

$$\big|\,\widehat{f}(n)\,\big|^2 + \big|\,\widehat{f}(-n)\,\big|^2 = 2\big(\, |a_n(u)|^2 + |b_n(u)|^2 + |a_n(v)|^2 + |b_n(v)|^2\,\big),$$

and we deduce the complex Parseval formula

$$\int_{-\pi}^{\pi}|f(x)|^2\, dx = \frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\big|\,\widehat{f}(n)\,\big|^2. \tag{20.2.16}$$

If f is real valued we have $S_n^{\mathbb{C}}\big[\, f\,\big] = S_n\big[\, f\,\big]$. For this reason we will drop the more complicated notation $S_n^{\mathbb{C}}$ and we will stick with the simpler one S_n with the understanding that $S_n\big[\, f\,\big] = S_n^{\mathbb{C}}\big[\, f\,\big]$ when f is complex valued.

We have produced a bijective continuous map

$$\mathcal{F} : L^2(\mathbb{T}, \mathbb{C}) \to L^2(\mathbb{Z}, \mathbb{C}), \;\; f \mapsto \widehat{f}$$

called *the Fourier transform* for the Abelian group $\mathbb{T}$. Note that the rescaled map $\frac{1}{\sqrt{2\pi}}\mathcal{F}$ is an isometry.

Observe that the expression

$$\widehat{f}(n) = \int_{-\pi}^{\pi} f(x) e^{-\boldsymbol{i}nx} dx,$$

make sense for any $f \in L^1(\mathbb{T}, \mathbb{C})$ and

$$\big|\, \widehat{f}(n) \,\big| \leqslant \| f \|_{L^1}$$

so the Fourier transform extends to a continuous linear map

$$\mathcal{F} : L^1(\mathbb{T}, \mathbb{C}) \to L^\infty(\mathbb{Z}, \mathbb{C}).$$

A few natural questions come to mind.

(i) Suppose that $f \in C(\mathbb{T})$. In particular, $f \in L^2$ so

$$\lim_{n \to \infty} \| S_n[\, f \,] - f \|_{L^2} = 0.$$

Given $x \in \mathbb{T}$, is it true that $S_n[\, f \,](x)$ converges to $f(x)$ as $n \to \infty$?

(ii) We have seen that the complex Fourier coefficients $\widehat{f}(n)$ uniquely determine the function f if $f \in L^2$. Is this true also for $f \in L^1 \supsetneq L^2$?

20.2.3. Pointwise convergence of Fourier series. Let $f \in L^1(\mathbb{T}, \mathbb{C})$. Then for any $n \in \mathbb{N}$ we have

$$S_n[\, f \,] = \frac{1}{2\pi} \sum_{|k| \leqslant n} \widehat{f}(k) e^{\boldsymbol{i}kx} = \sum_{|k| \leqslant n} \frac{1}{2\pi} \left(\int_{-\pi}^{\pi} f(y) e^{-\boldsymbol{i}ky} dy \right) e^{\boldsymbol{i}kx}$$

$$= \int_{-\pi}^{\pi} \underbrace{\frac{1}{2\pi} \left(\sum_{|k| \leqslant n} e^{\boldsymbol{i}k(x-y)} \right)}_{=:D_n(x-y)} f(y) dy.$$

The function $D_n : \mathbb{T} \to \mathbb{C}$ is called the *Dirichlet kernel.* Despite its appearance, it is real valued. Indeed, observe

$$\sum_{|k| \leqslant n} e^{\boldsymbol{i}k(x-y)} = 1 + 2 \sum_{k=1}^{n} \cos k(x-y).$$

The sum

$$\sum_{k=1}^{n} \cos kt$$

can be expressed in more compact form. To see this, set $z = \cos t$ so

$$\sum_{k=1}^{n} \cos kt = \mathbf{Re}\,(z + \cdots + z^n).$$

On the other hand, we have $\bar{z} = \frac{1}{z}$ and

$$z + \cdots + z^n = z\frac{z^n - 1}{z - 1} = \frac{z^n - 1}{1 - \bar{z}} = \frac{\cos nt + \boldsymbol{i} \sin nt - 1}{1 - \cos t + \boldsymbol{i} \sin t}$$

$(1 - \cos\theta = 2\sin^2\theta/2,\ \sin\theta = 2\sin\theta/2\cos\theta/2)$

$$= \frac{-2\sin^2 nt/2 + 2\boldsymbol{i}\sin nt/2\cos nt/2}{2\sin^2 t/2 + 2\boldsymbol{i}\sin t/2\cos t/2} = \frac{\boldsymbol{i}\sin nt/2}{\boldsymbol{i}\sin t/2}\frac{(\cos nt/2 + \boldsymbol{i}\sin nt/2)}{(\cos t/2 - \boldsymbol{i}\sin t/2)}$$

$$= \frac{\sin nt/2}{\sin t/2}(\cos nt/2 + \boldsymbol{i}\sin nt/2)(\cos t/2 + \boldsymbol{i}\sin t/2)$$

$$= \frac{\sin nt/2}{\sin t/2}\big(\cos(n+1)t/2 + \boldsymbol{i}\sin(n+1)t/2\big).$$

Now observe that

$$2\sin(nt/2)\cos(n+1)t/2 = \sin(2n+1)t/2 - \sin t/2$$

$$2\sin(nt/2)\sin(n+1)t/2 = \cos t/2 - \cos(2n+1)t/2.$$

Hence, $\forall n \in \mathbb{N}$, we have,

$$\cos t + \cdots + \cos nt = \frac{1}{2}\frac{\sin(2n+1)t/2 - \sin t/2}{\sin t/2} \tag{20.2.17a}$$

$$\sin t + \cdots + \sin nt = \frac{1}{2}\frac{\cos t/2 - \cos(2n+1)t/2}{\sin t/2}. \tag{20.2.17b}$$

In particular

$$D_n(t) = 1 + \frac{\sin(2n+1)t/2 - \sin t/2}{\sin t/2} = \frac{\sin(2n+1)t/2}{\sin t/2}.$$

Note that $D_n(t)$ is even, $D_n(0) = 2n + 1$; see Figure 20.1.

We deduce that for any $f \in L^1(\mathbb{T}, \mathbb{C})$ we have

$$S_n[\,f\,](0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)f(t)\,dy = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sin(2n+1)t/2}{\sin t/2}f(t)dt. \tag{20.2.18}$$

The space of $C(\mathbb{T})$ of continuous functions $\mathbb{T} \to \mathbb{R}$ is a Banach space with respect to the sup-norm $\|-\|_\infty$.

Theorem 20.2.6. *Denote by $\mathfrak{X}_0$ the subset of $C(\mathbb{T})$ consisting of functions such that $S_n[\,f\,](0)$ converges to $f(0)$. Then $\mathfrak{X}_0$ is contained in a meagre set. In other words, the collection of continuous functions for which one can expect pointwise convergence of the associated Fourier series is extremely "skinny".*

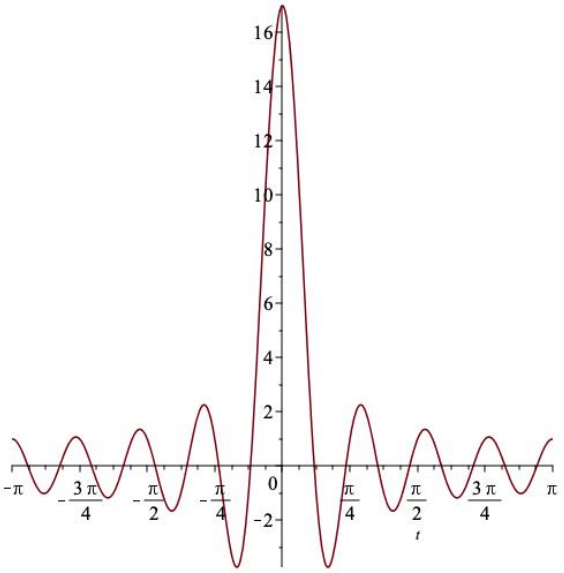

Figure 20.1. *The graph of $D_8(t)$.*

Proof. We have linear functionals

$$\lambda_n : C(\mathbb{T}) \to \mathbb{R},\ \ f \mapsto S_n\big[\, f\,\big](0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sin(2n+1)t/2}{\sin t/2} f(t)dt$$

and

$$\lambda_\infty : C(\mathbb{T}) \to \mathbb{R},\ \ \lambda_\infty(f) = f(0).$$

Thus

$$\mathcal{X}_0 := \big\{\, f \in C(\mathbb{T});\ \ \lim_{n\to\infty} \lambda_n(f) = \lambda_\infty(f)\,\big\}.$$

Observe that the above functionals are continuous since

$$|\lambda_n(f)| \leqslant \frac{1}{2\pi}\|D_n\|_{L^1} \cdot \|f\|_\infty.$$

Lemma 20.2.7. *For each n, denote by $\|\lambda_n\|_*$ the norm of the linear functional λ_n. Then*

$$\lim_{n\to\infty} \|\lambda_n\|_* = \infty.$$

Let us first show that the above Lemma implies the conclusion of Theorem 20.2.6. For $k \in \mathbb{N}$ we set

$$X_k := \big\{\, f \in C(\mathbb{T});\ \ |\lambda_n(f)| \leqslant k\|f\|_\infty,\ \ \forall n \in \mathbb{N}\,\big\}.$$

Observe that X_k is a closed set since it is an intersection of closed sets

$$X_k = \bigcap_{n\in\mathbb{N}} \big\{\, f \in C(\mathbb{T});\ \ |\lambda_n(f)| \leqslant k\|f\|_\infty\,\big\}.$$

On the other hand, X_k has empty interior, for any k. To see this we argue by contradiction. Suppose that $f_0 \in \boldsymbol{int}\, X_k$. Then there exists $r > 0$ such that $B_r(f_0) \subset X_k$. Thus,

$$\forall g \in C(\mathbb{T}),\ \ \|g\|_\infty < r \Rightarrow |\lambda_k(f_0 + g)| \leqslant k\|f_0 + g\|_\infty.$$

Now let $f \in C(\mathbb{T})\backslash\{0\}$. Define

$$\bar{f} = \frac{1}{\|f\|_\infty}.$$

Then

$$\|\bar{f}\|_\infty = 1,\ \ f_0 + r\bar{f}/2 \in X_k$$
$$\Rightarrow \forall n \in \mathbb{N},\ \ \frac{r}{2}|\lambda_n(\bar{f})| = |\lambda_n(r\bar{f})| \leqslant |\lambda_n(f_0 + r\bar{f}/2)| + |\lambda_n(f_0))|$$
$$\leqslant k\big(\ \|f_0 + r\bar{f}/2)\|_\infty + \|f_0\|_\infty\ \big) \leqslant k\big(\ (2\|f_0\|_\infty + r/2)$$

Hence, $\forall f \in C(\mathbb{T})\backslash\{0\}$ and $\forall n \in \mathbb{N}$ we have

$$\frac{|\lambda_n(f)|}{\|f\|_\infty} = |\lambda_n(\bar{f})| \leqslant \frac{2k\big(\ (2\|f_0\|_\infty + r/2)}{r}$$

In other words

$$\|\lambda_n\|_* \leqslant \frac{2k\big(\ (2\|f_0\|_\infty + r/2)}{r},\ \ \forall n \in \mathbb{N}.$$

This contradicts Lemma 20.2.7. This proves that the union

$$\mathcal{X} := \bigcup_{k \in \mathbb{N}} X_k$$

is a meagre set as a countable union of closed sets with empty interiors. On the other hand, observe that $\mathcal{X}_0 \subset \mathcal{X}$. Indeed, if $f \in \mathcal{X}_0\backslash\{0\}$, then

$$\frac{1}{\|f\|_\infty}\lambda_n(f) \to \frac{1}{\|f\|_\infty}f(0).$$

Hence, the sequence $\frac{1}{\|f\|_\infty}\lambda_n(f)$ is bounded so there exists $k \in \mathbb{N}$ such that

$$\frac{1}{\|f\|_\infty}|\lambda_n(f)| < k,\ \ \forall n,$$

i.e., $f \in X_k$. □

Proof of Lemma 20.2.7. For $n \in \mathbb{N}$ Consider the function

$$f_m : [-\pi, \pi] \to \mathbb{R},\ \ f_n(t) = \sin\frac{(2n+1)|t|}{2}$$

this is continuous and $f_n(\pi) = f_n(-\pi)$ and this defines a continuous function on the unit circle $\mathbb{T}$. Moreover

$$\lambda_n(f_n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sin(2n+1)t/2}{\sin t/2}\sin\frac{(2n+1)|t|}{2}\,dt = \frac{1}{\pi}\int_0^{\pi}\frac{\big(\sin\frac{(2n+1)t}{2}\big)^2}{\sin t/2}dt$$

Set $h := \frac{2\pi}{2n+1}$. We have

$$\int_0^\pi \frac{\left(\sin\frac{(2n+1)t}{2}\right)^2}{\sin t/2} dt = \sum_{k=1}^n \int_{(k-1)h}^{kh} \frac{\left(\sin\frac{(2n+1)t}{2}\right)^2}{\sin t/2} dt + \int_{\frac{2n\pi}{2n+1}}^\pi \frac{\left(\sin\frac{(2n+1)t}{2}\right)^2}{\sin t/2} dt$$

(use $\frac{1}{\sin t/2} \geqslant \frac{2}{t} \geqslant \frac{2}{kh}$ for $t \in [(k-1)h, kh]$)

$$\geqslant \sum_{k=1}^n \frac{2}{kh} \int_{(k-1)h}^{kh} \left(\sin\frac{(2n+1)t}{2}\right)^2 dt$$

$x := \frac{(2n+1)t}{2}$

$$= \sum_{k=1}^n \underbrace{\frac{4}{kh(2n+1)}}_{=\frac{2}{k\pi}} \underbrace{\int_{(k-1)\pi}^{k\pi} (\sin x)^2 dx}_{=\frac{\pi}{2}} = \sum_{k=1}^n \frac{1}{k}.$$

Observe that $\|f_n\|_\infty = 1$ so

$$\left\|\lambda_n(f_n)\right\|_* \geqslant \lambda_n(f_n) \geqslant \frac{1}{2\pi}\sum_{k=1}^n \frac{1}{k} \to \infty \text{ as } n \to \infty.$$

□

According to Theorem 20.2.6, the continuous functions $f : \mathbb{T} \to \mathbb{R}$ such that the partial Fourier sums $S_n\big[\,f\,\big]$ converge uniformly (or even pointwisely) to f are a rare species. However, they do exist. For example, if f is constant, $f(z) = 1$, $\forall z \in \mathbb{T}$, then

$$a_n(f) = b_n(f) = 0, \quad \forall n \in \mathbb{N},$$

so that

$$S_n\big[\,f\,\big] = \frac{1}{2\pi} a_0(f) = 1, \quad \forall n \in \mathbb{N}$$

so obviously $S_n\big[\,1\,\big] \to 1$ uniformly. Let us observe that this implies that

$$S_n\big[\,1\,\big] = \frac{1}{2\pi}\int_{-\pi}^\pi D_n(x-y)\,dy = 1, \quad \forall x \in [-\pi,\pi]. \tag{20.2.19}$$

However, this convergence phenomenon is not limited to this trivial case.

Let us first investigate when the Fourier series of a function $f \in L^1(\mathbb{T})$ converges. In the sequel we will think of functions $\mathbb{T} \to \mathbb{R}$ as 2π-periodic functions. Observe that if $f \in L^1(\mathbb{T})$, then

$$\int_{-\pi}^\pi f(y)dy = \int_x^{x+2\pi} f(y)dy, \quad \forall x \in \mathbb{R}.$$

Fix $x \in [-\pi,\pi]$ and $f \in L^1(\mathbb{T})$. We want to investigate when the partial sums $S_n\big[\,f\,\big](x)$ converge to a given real number s. We follow the excellent presentation in [**39**, Ch.13]. We have

$$S_n\big[\,f\,\big](x) = \frac{1}{2\pi}\int_{-\pi}^\pi D_n(x-y)f(y)dy \overset{y=x+t}{=} \frac{1}{2\pi}\int_{x-\pi}^{x+\pi} D_n(-t)f(x+t)dt$$

(use the 2π-periodicity of $t \mapsto D_n(-t)$ and $t \mapsto f(x+t)$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(-t)f(x+t)dt$$

(D_n is even)

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)f(x+t)dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)f(x-t)dt$$

On the other hand, using (20.2.19), we deduce

$$s = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)sdt.$$

We deduce

$$\lim_{n\to\infty} S_n\big[\,f\,\big](x) = s \Longleftrightarrow \lim_{n\to\infty}\int_{-\pi}^{\pi} D_n(t)\big(\,f(x+t)-s\,\big)dt = 0. \tag{20.2.20}$$

Thus

$$\lim_{n\to\infty} S_n\big[\,f\,\big](x) = f(x) \Longleftrightarrow \lim_{n\to\infty}\int_{-\pi}^{\pi} D_n(t)\big(\,f(x+t)-f(x)\,\big)dt = 0. \tag{20.2.21}$$

The next result will play a fundamental role in our investigation.

Theorem 20.2.8 (Riemann-Lebesgue Lemma)**.** *Let $a, b \in \mathbb{R}$, $a < b$ and $f \in L^1([a,b], \boldsymbol{\lambda})$. Then*

$$\lim_{\lambda\to\infty}\int_a^b f(x)\cos\lambda x\,dx = \lim_{\lambda\to\infty}\int_a^b f(x)\sin\lambda x\,dx = 0.$$

Proof. Suppose first that f is a polynomial. Integrating by parts we deduce

$$\int_a^b f(x)\cos\lambda x dx = \frac{f(x)\sin\lambda x}{\lambda}\Big|_{x=a}^{x=b} - \frac{1}{\lambda}\int_a^b f'(x)\sin\lambda x dx$$

Both terms above go to zero as $\lambda\to\infty$ so

$$\lim_{\lambda\to\infty}\int_a^b f(x)\cos\lambda x = 0$$

The other equality is proved in a similar fashion. This proves the Riemann-Lebesgue Lemma when f is a polynomial.

Suppose that $f \in L^1([a,b])$. Let $\varepsilon > 0$. From the Weierstrass approximation theorem (Corollary 17.4.11) and Corollary 19.4.15 we deduce that exists a polynomial p such that

$$\int_a^b |f(x)-p(x)|dx < \frac{\varepsilon}{2}.$$

From the Riemann-Lebesgue Lemma for polynomials we deduce that there exists $\Lambda(\varepsilon) > 0$ such that

$$\forall\lambda > \Lambda(\varepsilon),\ \left|\int_a^b p(x)\cos\lambda x dx\right| < \frac{\varepsilon}{2}.$$

Hence, for all $\lambda > \Lambda(\varepsilon)$ we have

$$\left|\int_a^b f(x)\cos\lambda x dx\right| \leqslant \left|\int_a^b \big(\,f(x)-p(x)\,\big)\cos\lambda x dx\right| + \left|\int_a^b p(x)\cos\lambda x dx\right|$$

$$< \int_a^b \big|\,\big(\,f(x)-p(x)\,\big)\cos\lambda x\,\big|dx + \frac{\varepsilon}{2} \leqslant \int_a^b \big|\,f(x)-p(x)\,\big|\,dx + \frac{\varepsilon}{2} < \varepsilon.$$

Hence $\forall f \in L^1([a,b])$ we have

$$\lim_{\lambda\to\infty}\int_a^b f(x)\cos\lambda x = 0$$

The other equality is proved in a similar fashion. □

Corollary 20.2.9. *For any $f \in L^1(\mathbb{T})$ we have*

$$\lim_{n\to\infty} a_n(f) = \lim_{n\to\infty} b_n(f) = 0.$$ □

Remark 20.2.10. The Riemann-Lebesgue Lemma is a rather miraculous result because the function $f(x)\cos\lambda x$ does not converge to 0 as $\lambda \to \infty$. The only reason why the integral of this function is small for λ large has to be a mysterious balancing act: the area of the graph below the x-axis is close to the area above this axis. The proof we gave hides this miraculous cancellation.

Intuitively, the main reason behind this cancellation is the highly oscillatory behavior of $\cos\lambda x$ with no particular bias for positive or negative values. In Figure 20.2 we have depicted the graph of $f(x)\cos(50x)$ where $f(x) = (x-2)^2-1$ where this "unbiased" highly oscillatory behavior is very visible. □

Definition 20.2.11. Let $f \in \mathbb{T} \to \mathbb{R}$ be a Lebesgue integrable function. We say that f satisfies the *Dini condition* at $x \in [-\pi,\pi]$ if there exists $\delta > 0$ such that

$$\int_{-\delta}^{\delta} \frac{|f(x+t)-f(x)|}{|t|}dt < \infty.$$ □

Remark 20.2.12. Let $f \in L^1(\mathbb{T})$ Observe first that

$$\int_{-\delta}^{\delta} \frac{|f(x+t)-f(x)|}{|t|}dt = \int_{-\delta}^{\delta} \frac{|f(x-t)-f(x)|}{|t|}dt$$

so f satisfies the Dini condition at x if and only if

$$\int_{-\delta}^{\delta} \frac{|f(x+t)-f(x)|}{|t|}dt + \int_{-\delta}^{\delta} \frac{|f(x-t)-f(x)|}{|t|}dt < \infty.$$

Observe next that the Dini condition is really a constraint on the behavior of the function f near x. For the function $t \mapsto \frac{f(x+t)-f(x)}{t}$ to be integrable the numerator of the fraction should be a counterweight to the explosive behavior of $1/t$ as $t \to 0$. For example, if f is Lipschitz, then the function $t \mapsto \frac{f(x+t)-f(x)}{t}$ is bounded hence integrable. Similarly, if f is differentiable at x, then it satisfies the Dini condition at x. □

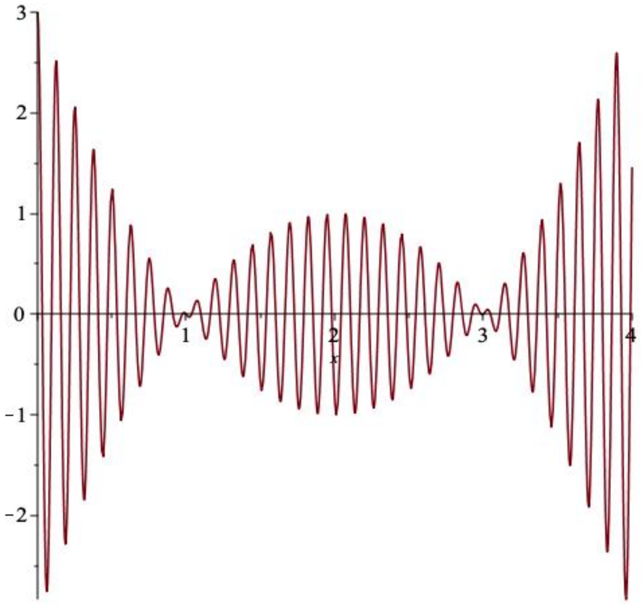

Figure 20.2. *The graph of* $(x^2-4x+3)\cos(50x)$ *on the interval* $[0,4]$.

Theorem 20.2.13. *Suppose that* $f \in L^1(\mathbb{T})$ *satisfies the Dini condition at* $x \in [-\pi,\pi]$. *Then*

$$\lim_{n\to\infty} S_n\big[\,f\,\big](x) = f(x).$$

Proof. Let $\delta > 0$ be as in the Dini condition. We have

$$S_n\big[\,f\,\big] - f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)\big(\,f(x+t)-f(x)\,\big)dt$$

$$= \underbrace{\frac{1}{2\pi}\int_{|t|\leqslant\delta} D_n(t)\big(\,f(x+t)-f(x)\,\big)dt}_{X_n} + \underbrace{\frac{1}{2\pi}\int_{\delta<|t|<\pi} D_n(t)\big(\,f(x+t)-f(x)\,\big)dt}_{Y_n}.$$

We write

$$D_n(t)\big(\,f(x+t)-f(x)\,\big) = \underbrace{\frac{f(x+t)-f(x)}{\sin t/2}}_{=g(t)}\sin\frac{2n+1}{2}t.$$

The function $g(t)$ is integrable on $\{\delta < |t| \leqslant \pi\} = [-\pi,-\delta) \cup (\delta,\pi]$ since the function $\frac{1}{|\sin t/2|}$ is bounded above by $\sin\delta/2$ in this region. The Riemann-Lebesgue Lemma implies that

$$\lim_{n\to\infty} X_n = \lim_{n\to\infty}\int_{\delta<|t\leqslant\pi} g(t)\sin\frac{2n+1}{2}t = 0.$$

In the region $|t| \leqslant \delta$ we write

$$D_n(t)\big(f(x+t-f(x)\big)) = \underbrace{\frac{f(x+t)-f(x)}{t}\frac{t}{\sin t/2}}_{=:h)t)}\sin\frac{2n+1}{2}t.$$

The Dini condition implies that the function $t \mapsto \frac{f(x+t)-f(x)}{t}$ is integrable on $[-\delta,\delta]$ while the function $t \mapsto \frac{t}{\sin t/2}$ is bounded over this interval. Hence the function $h(t)$ is integrable over $[-\delta,\delta]$ and Riemann-Lebesgue Lemma implies that

$$\lim_{n\to\infty} Y_n = \lim_{n\to\infty}\int_{|t|\leqslant\delta} h(t)\sin\frac{2n+1}{2}t = 0.$$

□

Example 20.2.14. Consider the function $f:[-\pi,\pi]\to\mathbb{R}$, $f(x)=|x|$. Since

$$f(-\pi)=f(\pi)=\pi$$

this function extends to a periodic Lipschitz function $\mathbb{R}\to\mathbb{R}$. Thus, for every $x\in[-\pi,\pi]$ the series

$$\frac{1}{2\pi}a_0(f)+\frac{1}{\pi}\sum_{n\in\mathbb{N}}\big(a_n(f)\cos nx+b_n(f)\sin nx\big)$$

converges to $f(x)$. In particular,

$$0=f(0)=\frac{1}{2\pi}a_0(f)+\frac{1}{\pi}\sum_{n\in\mathbb{N}}a_n(f).$$

We have

$$a_0(f)=\int_{-\pi}^{\pi}|x|\,dx = 2\int_0^{\pi}x\,dx=\pi^2.$$

If $n>0$,

$$a_n(f)=\int_{-\pi}^{\pi}|x|\cos nx\,dx = 2\int_0^{\pi}x\cos nx\,dx=\frac{2}{n}\int_0^{\pi}xd(\sin nx)$$

$$=\underbrace{\Big(\frac{2x}{n}\sin nx\Big)\Big|_{x=0}^{x=\pi}}_{=0}-\frac{2}{n}\int_0^{\pi}\sin nxdx=\frac{2}{n^2}\cos nx\Big|_{x=0}^{x=\pi}=\begin{cases}0, & n \text{ even},\\ -\frac{4}{n^2}, & n \text{ odd}.\end{cases}$$

Hence

$$\frac{\pi}{2}=\frac{4}{\pi}\sum_{k=0}^{\infty}\frac{1}{(2k+1)^2},$$

i.e.,

$$\sum_{k=0}^{\infty}\frac{1}{(2k+1)^2}=\frac{\pi^2}{8}.$$

If we write

$$S := 1+\frac{1}{2^2}+\frac{1}{3^2}+\cdots,$$

then we deduce

$$S = \sum_{k=1}^{\infty} \frac{1}{(2k)^2} + \sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} = \frac{S}{4} + \frac{\pi^2}{8},$$

so that

$$\frac{3}{4} S = \frac{\pi^2}{8},$$

i.e.,

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

This is Euler's identity (20.2.4). □

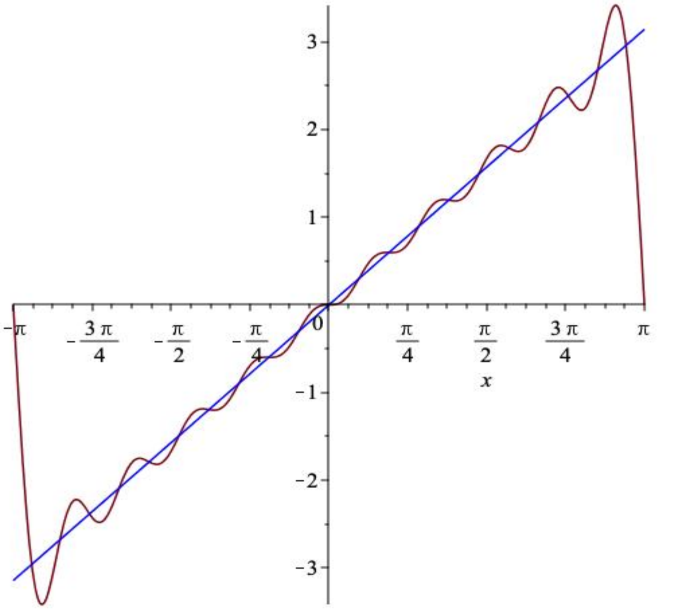

Figure 20.3. *The function $f(x) = x$ and its Fourier approximation $S_{10}[\,f\,](x)$.*

Example 20.2.15. Consider the function $f \in L^2(\mathbb{T})$, $f(x) = x$, $x \in (-\pi, \pi]$. According to the computations in Example 4.6.7 the Fourier series of f is

$$\sum_{n\in\mathbb{N}} \frac{2(-1)^{n+1}}{n} \sin nx = 2\Big(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \cdots \Big).$$

The function f satisfies the Dini condition at any $x \in (-\pi, \pi)$ so its Fourier series converges for any $x \in (-\pi, \pi)$.

For example for $x = \frac{\pi}{2}$ we deduce

$$\frac{\pi}{2} = 2\Big(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} - \cdots \Big).$$

This is a formula first obtained by Leibniz by using the Taylor series of $\arctan x$.

In Figure 20.3 we have depicted in the same coordinate system the function f and the partial Fourier sum $S_{10}\big[\, f\,\big](x)$. □

The two results we proved so far shows that the pointwise convergence of a Fourier series is a rather subtle issue and we have barely scratched the surface. For more details we refer to [**26, 39, 46**]. In the 1920's A. N. Kolmogorov, in his first mathematical contribution, constructed examples of functions $f \in L^1(\mathbb{T})$ such that their Fourier series diverge almost everywhere. These functions are however discontinuous. Lusin, Kolmogorov's adviser, if such a thing is possible if f is continuous. More than four decades later, L. Carleson showed in a notoriously difficult paper that the Fourier series of an L^2-function converges almost everywhere.

20.2.4. Uniqueness. We want to address the second of the questions we formulated above: do the Fourier coefficients of a function $f \in L^1(\mathbb{T})$ uniquely determine the function?

More concretely, given that $f, g \in L^1(\mathbb{T})$ such that $a_n(f) = a_n(g)$ and $b_n(f) = b_n(g)$, $\forall n \in \mathbb{Z}$ can we conclude that $f = g$ in L^1? Equivalently, if $S_n\big[\, f\,\big] = S_n\big[\, g\,\big]$, $\forall n \in \mathbb{N}$, can we deduce that $f(x) = g(x)$ a.e.?

We can phrase this in a more conceptual way. Observe that for any $f \in L^1(\mathbb{T}, \mathbb{C})$ and any $n \in \mathbb{Z}$ we have

$$|\widehat{f}(n)| = \left| \int_{-\pi}^{\pi} f(x) e^{-\boldsymbol{i}nx} dx \right| \leqslant \int_{-\pi}^{\pi} \big|\, f(x) e^{-\boldsymbol{i}nx}\,\big|\, dx = \|f\|_{L^1}$$

The Fourier transform is then the map

$$L^1(\mathbb{T}, \mathbb{C}) \ni f \mapsto \widehat{f} = \big(\, \widehat{f}(n)\,\big)_{n \in \mathbb{Z}} \in L^\infty(\mathbb{Z}, \mathbb{C}) = \text{bounded functions } \mathbb{Z} \to \mathbb{C}.$$

Note that

$$\|\widehat{f}\|_{L^\infty} \leqslant \|f\|_{L^1}$$

Thus the Fourier transform is a bounded linear map $L^1(\mathbb{T}, \mathbb{C}) \to L^\infty(\mathbb{C})$ and we want to investigate its injectivity or, equivalently, its kernel.

We will do this in a rather roundabout way that will afford us a side trip in some beautiful parts of classical analysis. First some terminology.

Definition 20.2.16. Consider a Banach space $(X, \| - \|)$. The sequence $(x_n)_{n \in \mathbb{N}}$ in X is said to *Cesàro converge* or *C-converge* to $x \in X$ if the sequence of running averages

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$$

converges to x. We indicate this using the notation

$$C - \lim_{n \to \infty} x_n = x. \qquad \square$$

.

Lemma 20.2.17 (Cesàro). *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence a Banach space. X. Then*

$$\lim_{n\to\infty} x_n = x \Rightarrow C - \lim_{n\to\infty} x_n = x.$$

Proof. Set

$$y_n = x_n = x,\ \bar{y}_n = \frac{1}{n}\sum_{k=1}^{n} y_n = \frac{1}{n}\sum_{k=1}^{n} x_n - x.$$

Thus it suffices to prove that $\bar{y}_n \to 0$ given that $y_n \to 0$. Set

$$C := \sup_n \|y_n\|.$$

Fix $\varepsilon > 0$. There exists $n_0 = n_0(\varepsilon)$ such that $\|y_n\| < \varepsilon/2$, $\forall n > n_0$. Next, choose $n_1 = n_1(\varepsilon) > n_0$ such that

$$\frac{n_0 C}{n} < \frac{\varepsilon}{2},\ \ \forall n > n_1.$$

Let $n > n_1$. We have

$$\|\bar{y}_n\| \leqslant \frac{\|y_1\| + \cdots + \|y_n\|}{n}$$

$$= \underbrace{\frac{\|y_1\| + \cdots + \|y_{n_0}\|}{n}}_{<\frac{n_0 C}{n}} + \underbrace{\frac{\|y_{n_0+1} + \cdots + \|y_n\|}{n}}_{<\frac{n-n_0}{n}\frac{\varepsilon}{2}} - < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.$$

□

Remark 20.2.18. The converse is not true. There are Cesàro convergent sequences that are not convergent. For example, the sequence

$$1, -1, 1, -1, \ldots$$

Cesàro converges to 0 but it is obviously not convergent. □

We want to investigate the Cesàro convergence of the partial sums $S_n[\,f\,]$ of an integrable function $f : \mathbb{T} \to \mathbb{R}$. We have

$$\bar{S}_n[\,f\,](x) = \frac{1}{n}\big(\,S_1[\,f\,](x) + \cdots + S_n[\,f\,](x)\,\big) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(x-y)f(y)\,dy,$$

where $F_n(t)$ is the *Fejér kernel*

$$F_n(t) = \frac{1}{n}\big(\,D_1(t) + \cdots + D_n(t)\,\big) = \frac{1}{n\sin t/2}\sum_{k=1} \sin(2k+1)t/2.$$

The above sum can be simplified substantially. We write $\zeta = \cos t/2 + \boldsymbol{i}\sin t/2$, $z = \zeta^2$. Then

$$A_n(t) := \sum_{k=1}^{n} \sin(2k+1)t/2 = \mathbf{Im}\big(\,\zeta + \zeta^3 + \cdots + \zeta^{2n+1}\,\big) = \mathbf{Im}\,\zeta\big(\,1 + z + \cdots + z^n\,\big).$$

We have

$$
\begin{aligned}
\zeta(1+z+\cdots+z^n) &= \zeta\frac{1-\cos(n+1)t-\boldsymbol{i}\sin(n+1)t}{1-\cos t-\boldsymbol{i}\sin t}\\
&= \zeta\frac{2\sin^2(n+1)t/2-2\sin(n+1)t/2\cos(n+1)t/2}{2\sin^2 t/2-2\boldsymbol{i}\sin t/2\cos t/2}\\
&= (\cos t/2+\boldsymbol{i}\sin t/2)\frac{2\sin(n+1)t/2\big(\sin(n+1)t/2-\cos(n+1)t/2\big)}{-2\boldsymbol{i}\sin t/2(\cos t/2+\boldsymbol{i}\sin t/2)}\\
&= \frac{\cos(n+1)t/2\sin(n+1)t/2+\boldsymbol{i}\sin^2(n+1)t/2}{\sin t/2}.
\end{aligned}
$$

Thus

$$
A_n(t)=\frac{\sin^2(n+1)t/2}{\sin t/2},\quad F_n(t)=\frac{1}{n}\left(\frac{\sin(n+1)t/2}{\sin t/2}\right)^2.
$$

Note that $F_n(t)$ is nonnegative, even and 2π-periodic. The graph of F_9 is depicted in Figure 20.4. As in the previous subsection we deduce that for any $f\in L^1(\mathbb{T})$ we have

$$
\bar{S}_n\big[\,f\,\big](x)=\frac{1}{2\pi}\int_{-\pi}^{\pi}F_n(t)f(x+t)\,dt.
$$

We deduce

$$
1=\bar{S}_n\big[\,1\,\big](0)=\frac{1}{2\pi}\int_{-\pi}^{\pi}F_n(t)dt. \tag{20.2.22}
$$

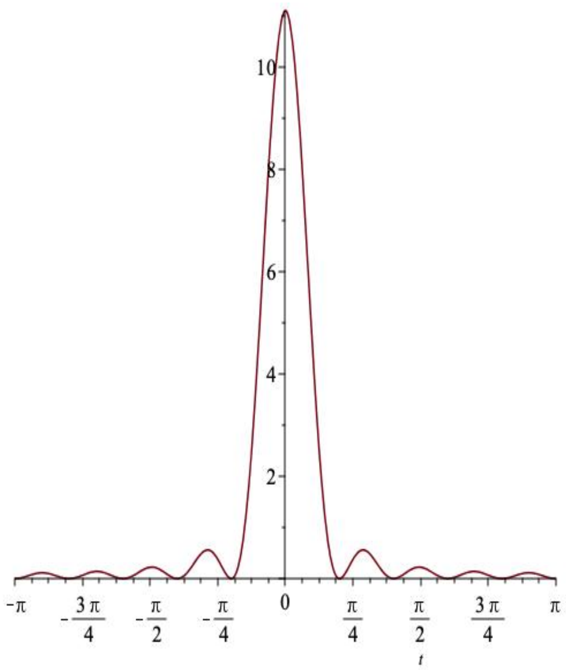

Figure 20.4. *The graph of $F_9(t)$.*

As Figure 20.4 suggests, most of the area under the graph of F_n seems to be concentrated near the origin. More precisely,

$$\forall \delta \in (0,\pi), \quad \lim_{n\to\infty} \int_{\delta<|t|\leqslant\pi} F_n(t)\,dt = 0. \tag{20.2.23}$$

Indeed

$$0 \leqslant F_n(t) = \frac{\sin^2(n+1)t/2}{n\sin^2 t/2} \leqslant \frac{1}{n\sin^2\delta/2}, \quad \forall \delta < |t| \leqslant \pi.$$

Proposition 20.2.19. *Let $p \in [1,\infty)$. Then for any $f \in L^p(\mathbb{T})$ and any $n \in \mathbb{N}$ we have $S_n[\,f\,] \in L^p(\mathbb{T})$ and*

$$\big\|\bar{S}_n[\,f\,]\big\|_{L^p} \leqslant \|f\|_{L^p}. \tag{20.2.24}$$

Proof. By Theorem 19.4.14, the space $C(\mathbb{T})$ is dense in $L^p(\mathbb{T})$ so it suffices to prove (20.2.24) for $f \in C(\mathbb{T})$.

Fix $n \in \mathbb{N}$ and consider the Borel measure μ_n on $[-\pi,\pi]$ defined by

$$\mu_n[\,dt\,] := \frac{F_n(t)}{2\pi}\,dt$$

We deduce from (20.2.22) that μ_n is a probability measure, i.e.,

$$\mu_n\big[\,[-\pi,\pi]\,\big] = 1.$$

Let $f \in C(\mathbb{T})$. We view f as a continuous 2π-periodic function on $\mathbb{R}$. We have

$$\big|\bar{S}_n[\,f\,](x)\,\big|^p = \left|\int_{-\pi}^{\pi} f(x+t)\frac{F_n(t)}{2\pi}\,dt\right|^p \leqslant \left|\int_{-\pi}^{\pi} |f(x+t)|\mu_n[\,dt\,]\right|^p.$$

Hólder's inequality implies

$$\begin{aligned}\int_{-\pi}^{\pi} |f(x+t)|\mu_n[\,dt\,] &\leqslant \left(\int_{-\pi}^{\pi} |f(x+t)|^p\,\mu_n[\,dt\,]\right)^{1/p} \cdot \left(\int_{-\pi}^{\pi} 1^q\,\mu_n[\,dt\,]\right)^{1/q} \\ &= \left(\int_{-\pi}^{\pi} |f(x+t)|^p\,\mu_n[\,dt\,]\right)^{1/p}.\end{aligned}$$

Hence

$$\big|\bar{S}_n[\,f\,](x)\,\big|^p \leqslant \int_{-\pi}^{\pi} |f(x+t)|^p\,\mu_n[\,dt\,].$$

We deduce

$$\big\|\bar{S}_n[\,f\,]\big\|_{L^p}^p = \int_{-\pi}^{\pi} \big|\bar{S}_n[\,f\,](x)\,\big|^p\,dx \leqslant \int_{-\pi}^{\pi}\left(\int_{-\pi}^{\pi} |f(x+t)|^p\,\mu_n[\,dt\,]\right)dx$$

(use Fubini)

$$= \int_{-\pi}^{\pi}\left(\int_{-\pi}^{\pi} \boxed{|f(x+t)|^p}\,dx\right)\mu_n[\,dt\,] = \int_{-\pi}^{\pi}\underbrace{\left(\int_{-\pi}^{\pi} |f(x)|^p\,dx\right)}_{\|f\|_{L^p}^p}\mu_n[\,dt\,]$$

$$= \int_{-\pi}^{\pi} \|f\|_{L^p}^p \mu_n\big[\,dt\,\big] = \|f\|_{L^p}^p.$$

This proves (20.2.24) □

Theorem 20.2.20 (Fejér)**.** *Let $f \in L^1(\mathbb{T})$ and set*

$$\bar{S}_n\big[\,f\,\big] = \frac{1}{n}\big(\,S_1\big[\,f\,\big] + \cdots + S_n\big[\,f\,\big]\,\big),\ \ n \in \mathbb{N}.$$

(i) *If f is continuous then $\bar{S}_n\big[\,f\,\big]$ converges uniformly to f on $\mathbb{T}$.*

(ii) *If $f \in L^p(\mathbb{T})$, $p \in [1, \infty)$, then*

$$\lim_{n\to\infty} \big\|\,\bar{S}_n\big[\,f\,\big] - f\,\big\|_{L^p} = 0.$$

Proof. (i) Suppose f is continuous. Since $\mathbb{T}$ is compact, the function f is uniformly continuous. As usual, we identify it with a 2π-periodic continuous function $f : \mathbb{R} \to \mathbb{R}$. Set

$$\omega(\delta) := \sup_{|x-y|\leqslant\delta} \big|\,f(x) - f(y)\,\big|$$

The uniform continuity implies that

$$\lim_{\delta\searrow 0} \omega(\delta) = 0.$$

Denote by $\|-\|_\infty$ the sup-norm on $C(\mathbb{T})$. For any $x \in [-\pi, \pi]$ we have

$$\big|\,\bar{S}_n\big[\,f\,\big](x) - f(x)\,\big| = \frac{1}{2\pi}\left|\int_{-\pi}^{\pi} F_n(t)\big(\,f(x+t) - f(x)\,\big)\,dt\right|$$

$$\leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(t)\big|\,f(x+t) - f(x)\,\big|dt$$

$$= \frac{1}{2\pi}\int_{|t|\leqslant\delta} F_n(t)\big|\,f(x+t) - f(x)\,\big|dt + \frac{1}{2\pi}\int_{|t|>\delta} F_n(t)\big|\,f(x+t) - f(x)\,\big|dt$$

$$\leqslant \omega(\delta)\int_{|t|\leqslant\delta} F_n(t)\omega(\delta)dt + \frac{\|f\|_\infty}{\pi}\int_{|t|>\delta} F_n(t)\,dt \overset{(20.2.22)}{\leqslant} \omega(\delta) + \frac{\|f\|_\infty}{\pi}\int_{|t|>\delta} F_n(t)\,dt$$

Let $\varepsilon > 0$. There exists $\delta_0 = \delta_0(\varepsilon)$ such that $\omega(\delta_0) < \frac{\varepsilon}{2}$. From (20.2.23) we deduce that there exists $N = N(\varepsilon)$ such that

$$\frac{\|f\|_\infty}{\pi}\int_{|t|>\delta_0} F_n(t)\,dt < \frac{\varepsilon}{2},\ \ \forall n > N_\varepsilon.$$

Hence $\forall\varepsilon > 0$, $\exists N = N(\varepsilon) > 0$ such that $\forall n > N(\varepsilon)$ and $\forall x \in [-\pi, \pi]$,

$$\big|\,\bar{S}_n\big[\,f\,\big](x) - f(x)\,\big| < \varepsilon.$$

(ii) Observe first that for any $h \in C(\mathbb{T})$ we have

$$\|h\|_{L^p} = \left(\int_{-\pi}^{\pi} |h(x)|^p\,dx\right)^{1/p} \leqslant \left(\int_{-\pi}^{\pi} \|h\|_\infty^p\,dx\right)^{1/p} (1\pi)^{1/p}\|h\|_\infty.$$

Let $f \in L^p(\mathbb{T})$. Fix $\varepsilon > 0$. According to Theorem 19.4.14, the space $C(\mathbb{T})$ is dense in $L^p(\mathbb{T})$ so there exists $g \in C(\mathbb{T})$ such that

$$\|f - g\|_{L^p} < \frac{\varepsilon}{3}.$$

From (ii) we deduce

$$\lim_{n\to\infty} \big\|\bar{S}_n\big[\,g\,\big] - g\,\big\|_\infty = 0.$$

Hence there exists $N = N(\varepsilon) > 0$ such that

$$(2\pi)^{1/p}\big\|\bar{S}_n\big[\,g\,\big] - g\,\big\|_\infty < \frac{\varepsilon}{3},\;\;\forall n > N(\varepsilon).$$

Thus for all $n > N(\varepsilon)$ we have

$$\begin{aligned}\big\|\bar{S}_n\big[\,f\,\big] - f\,\big\|_{L^p} &\leqslant \big\|\bar{S}_n\big[\,f\,\big] - \bar{S}_n\big[\,g\,\big]\big\|_{L^p} + \big\|\bar{S}_n\big[\,g\,\big] - g\,\big\|_{L^p} + \|g - f\|_{L^p}\\ &\overset{(20.2.24)}{\leqslant} 2\|g - f\|_{L^p} + (2\pi)^{1/p}\big\|\bar{S}_n\big[\,g\,\big] - g\,\big\|_\infty < \varepsilon.\end{aligned}$$

□

Corollary 20.2.21. *Suppose $f, g \in L^1(\mathbb{T})$ satisfy*

$$a_n(f) = a_n(g),\;\; b_m(f) = b_m(g),\;\;\forall m, n \in \mathbb{Z}.$$

Then $f = g$ a.e..

Proof. Set $h = f - g$. Then $a_n(h) = b_n(h) = 0$, $\forall n \in \mathbb{Z}$ so $S_n\big[\,h\,\big] = 0$, $\forall n \in \mathbb{N}$ and we deduce $\bar{S}_n\big[\,h\,\big] = 0$, $\forall n \in \mathbb{N}$. We deduce

$$\|\,h\,|_{L^1} = \lim_{n\to\infty}\big\|\bar{S}_n\big\|_{L^1} = 0.$$

□

Remark 20.2.22. Let $f \in C(\mathbb{T})$. Observe that for any $n \in \mathbb{N}$ and any $x \in [-\pi, \pi]$ we have

$$\bar{S}_n\big[\,f\,\big] = \frac{1}{2\pi}a_0(f) + \sum_{k=1}^n \frac{n+1-k}{n}\big(\,a_k(f)\cos kx + b_k(f)\sin kx\,\big).$$

This sequence of trigonometric polynomials, determined *explicitly* from f, converges uniformly to f on $[-\pi, \pi]$. □

20.3. Elements of point set topology

20.4. Fundamental results about Banach spaces

20.5. Exercises

Exercise 20.1. Suppose that H is a real Hilbert space with inner product $\langle -,-\rangle$, and $x,y\in H\backslash\{0\}$. Prove that the following are equivalent.

(i) $\|x+y\| = \|x\|+\|y\|$.

(ii) $\exists t\geqslant 0$ such that $y = tx$.

□

Exercise 20.2. Suppose that H is a real Hilbert space with inner product $\langle -,-\rangle$ and $\mathcal{X}\subset H$. Prove that the following are equivalent.

(i) $\operatorname{span}(\mathcal{X})$ is not dense in H.

(ii) There exists $h\in H\backslash\{0\}$ such that $\langle h,x\rangle = 0$, $\forall x\in\mathcal{X}$.

Hint. Use Theorem 20.1.15. □

Exercise 20.3. Consider the sequence of functions $H_n : [0,1]\to\mathbb{R}$ defined by

$$H_{-1}(x) = 1,\ \ H_{0,0} = \boldsymbol{I}_{[0,1/2)} - \boldsymbol{I}_{[1/2,1)}$$

and, generally,

$$H_{n,k}(x) = 2^{n/2}H_{0,0}\big(2^n x - k\big),\ \ 0\leqslant k < 2^n.$$

(i) Draw the graphs of $H_0, H_{0,0}, H_{1,0}, H_{1,1}$.

(ii) Prove that

$$\int_0^1 H_{-1}(x)H_{n,k}(x)dx = 0,$$

$$\int_0^1 H_{m,j}(x)H_{n,k}(x)dx = \begin{cases} 1, & (m,j) = (n,k),\\ 0, & (m,j)\neq(n,k).\end{cases}$$

(iii) For $n\geqslant 0$ we set

$$\mathcal{H}_n = \operatorname{span}\big\{\,H_{-1},\ \ H_{m,k},\ \ 0\leqslant m\leqslant n,\ \ 0\leqslant k<2^m\,\big\}\subset L^2\big(\,[0,1],\boldsymbol{\lambda}\big).$$

Prove that for any $n\geqslant 1$ and any $0\leqslant k<2^n$ we have $\exists h\in\mathcal{H}_{n-1}$ such that $h = \boldsymbol{I}_{[k/2^n,(k+1)/2^n]}$ a. e..

(iv) Set

$$\mathcal{H}_\infty = \bigcup_{n\geqslant 0}\mathcal{H}_n.$$

Prove that $\mathcal{H}_\infty$ is dense in $L^2([0,1])$.

□

Exercise 20.4. For $n = 0,1,\dots$ denote by $P_n(x)$ the degree n Legendre polynomial defined in Exercise 9.22,

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\big(\,1-x^2\,\big)^n.$$

(i) Compute

$$c_n := \|P_n\|_{L^2([-1,1])}.$$

(ii) Prove that span $\{ P_n(x);\ \ n = 0, 1, 2, \dots \}$ is dense in $L^2([-1, 1])$. **Hint.** Prove that span $\{ P_n(x);\ \ n = 0, 1, 2, \dots \}$ = span $\{ x^n;\ \ n = 0, 1, 2, \dots \}$.

(iii) Prove that the collection

$$\{ c_n^{-1} P_n;\ \ n = 0, 1, \dots \}$$

is a Hilbert basis of $L^2\big([-1, 1] \big)$.

(iv) Compute

$$a_n = \int_{-1}^{1} \cos(\pi x) P_n(x) dx,\ \ n = 0, 1, \dots,$$

and then show

$$\sum_{n\geqslant 0} \frac{a_n^2}{c_n^2} = 1.$$

Hint. Use integration by parts to compute a_n.

□

Exercise 20.5. Let $(\Omega, \mathcal{S}, \mu)$ be a finite measured space and $f \in H := L^2(\Omega, \mathcal{S}, \mu)$. Fix a sigma-subalgebra $\mathcal{A} \subset \mathcal{S}$.

(i) Show that $U := L^2(\Omega, \mathcal{A}, \mu)$ is a closed subspace of H.

(ii) Denote by $\bar{f}$ the orthogonal projection of f on U. Prove that (compare with Exercise 19.72)

$$\int_A f(x)\mu\big[\, dx \,\big] = \int_A \bar{f}(x)\mu\big[\, dx \,\big],\ \ \forall A \in \mathcal{A}.$$

□

Exercise 20.6. Let H be a real Hilbert space with inner product $\langle -, - \rangle$. Given $n \in \mathbb{N}$ and $u_1, \dots, u_n \in H$ we define the *Gram determinant* of $u_1, \dots, u_n$ to be the determinant of the *Gramian* matrix

$$G(u_1, \dots, u_n) = \big[\langle u_i, u_j \rangle \big]_{1\leqslant i,j\leqslant n}$$

(i) Let $u_1, \dots, u_n$. Fix any *orthonormal* basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_m\}$ of span$\{u_1, \dots, u_n\}$. Denote by A the $m \times n$ matrix with entries $a_{ij} = \langle \boldsymbol{e}_i, \boldsymbol{u}_j \rangle$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$. Show that

$$G(u_1, \dots, u_n) = A^\top A,$$

where $A^\top$ is the transpose of A.

(ii) Prove that $\det G(u_1, \dots, u_n) \geqslant 0$ with equality if and only if the vectors $u_1, \dots, u_n$ are linearly dependent. **Hint** Prove that $\ker A = \ker G$.

(iii) Suppose that $u_1, \dots, u_n$ are linearly independent and set $U := \operatorname{span}\{u_1, \dots, u_n\}$. The subspace U is closed since it is finite dimensional. Let $y \in H$ and denote by y_0 the orthogonal projection of y on U. Prove that

$$\|y - y_0\|^2 = \frac{\det G(y - y_0, u_1, \dots, u_n)}{\det G(u_1, \dots, u_n)} = \frac{\det G(y, u_1, \dots, u_n)}{\det G(u_1, \dots, u_n)}.$$

Hint. Write $y_\perp := y - y_0$. Prove that

$$\|y_\perp\|^2 \det G(u_1, \dots, u_n) = \det G(y_\perp, u_1, \dots, u_n) = \det G(y_\perp + y_0, u_1, \dots, u_n).$$

You will need to use (ii) at some point.

(iv) Let $H = L^2\big([0,1], \boldsymbol{\lambda}\big)$, $a_1, \dots, a_n \geqslant 0$. Define $u_1, \dots, u_n \in H$ by $u_k(x) = x^{a_k}$. Prove that

$$\det G(u_1, \dots, u_n) = \frac{\prod_{1 \leqslant j < k \leqslant n} (a_j - a_k)^2}{\prod_{j,k=1}^{n} (a_j + a_k + 1)}.$$

□

Exercise 20.7. Let $H = L^2\big([0,1], \boldsymbol{\lambda}\big)$. Suppose that $(u_n)_{n \geqslant 1}$ is a linearly independent sequence of functions in H. For each nonnegative integer k we denote by m_k the monomial $m_k(x) = x^k$, $x \in [0,1]$. Set

$$U = \operatorname{span}\big\{ u_n;\ \ n \in \boldsymbol{N} \big\}$$

Prove that the following are equivalent.

(i) The subspace U is dense in H.

(ii) For any $f \in H$

$$\lim_{n \to \infty} \frac{\det G(f, u_1, \dots, u_n)}{\det G(u_1, \dots, u_n)} = 0,$$

where $G(u_1, \dots, u_n)$ is the Gram determinant of $u_1, \dots, u_n \in H$.

(iii) For any $k \geqslant 0$

$$\lim_{n \to \infty} \frac{\det G(m_k, u_1, \dots, u_n)}{\det G(u_1, \dots, u_n)} = 0.$$

Hint. Use Exercise 20.6. □

Exercise 20.8. Let $H = L^2\big([0,1], \boldsymbol{\lambda}\big)$. Suppose that $(a_n)_{n \geqslant 1}$ is a strictly increasing sequence of positive numbers. For $n \geqslant 1$ define $u_n \in H$, $u_n(x) = x^{a_n}$. Prove that the following are equivalent

(i) The functions $(u_n)_{n \in \mathbb{N}}$ span a dense subspace of H.

(ii) $\sum_{n \geqslant 1} \frac{1}{a_n} = \infty$.

Hint. Use Exercises 20.6 and 20.7. □

Exercise 20.9. Fix a finite measured space $(\Omega, \mathcal{S}, \mu)$ and $f \in \mathcal{L}^0_+(\Omega, \mathcal{S})$. Prove that the following are equivalent.

(i) $f \in L^2(\Omega, \mathcal{S}, \mu)$.

(ii) There exists $C > 0$ such that for any $g \in \mathcal{L}^2_+(\Omega, \mathcal{S}, \mu)$ we have

$$\int_\Omega fgd\mu \leqslant C \left(\int_\Omega g^2 d\mu \right)^{1/2}.$$

Hint. (ii) ⇒ (i) Use Theorem 19.6.13. □

Exercise 20.10. Fix finite measured spaces $(\Omega_i, \mathcal{S}_i, \mu_i)$, $i = 0, 1$ and

$$K \in L^2(\Omega_1 \times \Omega_0, \mathcal{S}_1 \otimes \mathcal{S}_0, \mu_1 \otimes \mu_0).$$

(i) Prove that if $f_i \in \mathcal{L}^2(\Omega_i, \mathcal{S}_i, \mu_i)$, $i = 0, 1$, then the function

$$f_1 \otimes f_0 : \Omega_1 \times \Omega_0 \to \mathbb{R}, \quad f_1 \otimes f_0(\omega_1, \omega_0) = f_0(\omega_0) f_1(\omega_1)$$

belongs to $\mathcal{L}^2(\Omega_1 \times \Omega_0, \mathcal{S}_1 \otimes \mathcal{S}_0, \mu_1 \otimes \mu_0)$ and

$$\| f_1 \otimes f_0 \|_{L^2} = \| f_0 \|_{L^2} \| f_1 \|_{L^2}$$

(ii) Prove that for any $f_i \in \mathcal{L}^2(\Omega_0, \mathcal{S}_i, \mu_i)$, $i = 0, 1$ the function

$$(\omega_0, \omega_1) \mapsto K(\omega_1, \omega_0) f_0(\omega_0) f_1(\omega_1)$$

is integrable with respect to the measure $\mu \otimes \mu$ and

$$\int_{\Omega \times \Omega} \left| K(\omega_1, \omega_0) f_0(\omega_0) f_1(\omega_1) \right| \mu_1 \otimes \mu_0 \big[d\omega_1 d\omega_0 \big] \leqslant \| K \|_{L^2} \| f_0 \|_{L^2} \| f_1 \|_{L^2}.$$

(iii) Deduce that for any $f \in \mathcal{L}^2(\Omega_0, \mathcal{S}_0, \mu_0)$ the function

$$K[f](\omega_1) = \int_{\Omega_0} K(\omega_1, \omega_0) f(\omega_0) \mu_0 \big[d\omega_0 \big]$$

is well defined and finite for ω_1 outside a μ_1-negligible set and

$$\| K[f] \|_{L^2(\Omega_1)} \leqslant \| K \|_{L^2(\Omega_1 \times \Omega_0)} \| f \|_{L^2(\Omega_0)}.$$

Hint. (i) Use Fubini. (ii) Use (i) and the Cauchy inequality (19.4.2). (iii) Use Fubini, (ii) and Exercise 20.9. □

Exercise 20.11. Let (K_i, d_i), $i = 1, 2$ be compact metric spaces and μ_i finite Borel measure on K_0. Suppose that $\big(f^i_n \big)_{n \in \mathbb{N}}$ is a Hilbert basis of $L^2\big(K^i, \mu \big)$. Prove that the collection of functions

$$f_{m,n} : K_1 \times K_2 \to \mathbb{R}, \quad f_{m,n}(x_1, x_2) = f^1_m(x_1) f^2_n(x_2), \quad m.n \in \mathbb{N}$$

is a Hilbert basis of $L^2\big(K_1 \times K_2, \mu_1 \otimes \mu_2 \big)$. **Hint.** Prove that for any Borel sets $B_i \subset K_i$, the indicator $\boldsymbol{I}_{B_1 \times B_2}$ is in the $L^2\big(K_1 \times K_2, \mu_1 \otimes \mu_2 \big)$-closure of $\operatorname{span}\{ f_{m,n}, \ m, n \in \mathbb{N} \}$. □

Exercise 20.12. Let H be the Hilbert space $L^2([-1, 1], \boldsymbol{\lambda})$. For $n \geqslant 0$ we define $\mu_n : [-1, 1] \to \mathbb{R}$, $\mu_n(x) = x^n$.

(i) Show that $\operatorname{span}\big\{ \mu_0, \mu_1, \dots \big\}$ is dense in H.

(ii) Denote by $L_n(x)$ the n-th *Legendre polynomial*

$$L_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}\big(x^2-1\big)^n.$$

Set $\bar{L}_n(x) = \sqrt{n+1/2}L_n(x)$. Prove that

$$\{\,\bar{L}_0, \bar{L}_1, \dots\,\}$$

is the Hilbert basis of H obtained from $\{\mu_0, \mu_1, \dots\}$ via the Gram-Schmidt procedure (see Example 20.1.19).

Hint. (ii) Have a look at Exercise 9.22. □

Exercise 20.13. The unit circle $\mathbb{T}$ can be identified with the set of complex numbers of length 1 and as such, it becomes an Abelian group with respect to multiplication of complex numbers. Suppose that $\chi : \mathbb{T} \to \mathbb{T}$ is a *continuous* group morphism, i.e., $\chi(z_0z_1) = \chi(z_0)\chi(z_1)$, $\forall z_0, z_1 \in \mathbb{T}$. Prove that there exists $n \in \mathbb{Z}$ such that $\chi(z) = z^n$, $\forall z \in \mathbb{T}$. □

Appendix A

A bit more set theory

We survey a few more advanced facts of set theory that are needed in the second part of the text. For proofs and more details we refer to most texts on set theory, e.g., [**25, 41**].

A.1. Order relations

Suppose that X is a set. A *binary relation* on X is a subset $\mathcal{R} \subset X \times X$.

Given a binary relation $\mathcal{R}$ on X, we say two elements $x_0, x_1 \in X$ are $\mathcal{R}$-related, and we write this $x_0 \,\mathcal{R}\, x_1$, if $(x_0, x_1) \in \mathcal{R}$.

Example A.1.1. (a) Let $X = \mathbb{R}$. The set

$$\left\{\, (x, y) \in \mathbb{R} \times \mathbb{R};\;\; y - x \geqslant 0 \,\right\}$$

describes the usual order relation on the set of real numbers.

(b) Let $X = \mathbb{Z}$. Consider the binary relation

$$\mathscr{C} := \left\{\, (x, y) \in \mathbb{Z} \times \mathbb{Z};\;\; y - x \in 2\mathbb{Z} \,\right\}.$$

Two integers are $\mathscr{C}$-related iff they have the same remainders when divided by 2; see Theorem 3.3.7. □

Definition A.1.2. Let $\mathcal{R}$ be a binary relation on a set X.

(i) The relation $\mathcal{R}$ is called *reflexive* if

$$\forall x \in X, \;\; x \,\mathcal{R}\, x.$$

(ii) The relation $\mathcal{R}$ is called *symmetric* if

$$\forall x_0, x_1 \in X \;\; x_0 \,\mathcal{R}\, x_1 \Longleftrightarrow x_1 \,\mathcal{R}\, x_0.$$

(iii) The relation $\mathcal{R}$ is called *antisymmetric* if

$$\forall x_0, x_1 \in X, \ \ x_0 \,\mathcal{R}\, x_1 \text{ and } x_1 \,\mathcal{R}\, x_0 \Rightarrow x_0 = x_1.$$

(iv) The relation $\mathcal{R}$ is called *transitive* if

$$\forall x_0, x_1, x_2 \in X, \ \ x_0 \,\mathcal{R}\, x_1 \text{ and } x_1 \,\mathcal{R}\, x_2 \Rightarrow x_0 \,\mathcal{R}\, x_2.$$

(v) A *partial order* on X is a binary relation that is *reflexive, antisymmetric and transitive.* A *partially ordered set* or *poset* is a set together with a choice of partial order on it.

(vi) An *equivalence relation* on X is a binary relation that is *reflexive, symmetric and transitive.*

□

The relation in Example A.1.1(a) is an order relation. If S is a set and $X = 2^S$ is the collection of all subsets of S, then X is partially ordered by the inclusion relation $\subset$.

Definition A.1.3. Suppose that $(X, \preceq)$ is a poset.

(i) We say that two elements $x_0, x_1 \in X$ are *comparable* if either $x_0 \preceq x_1$ or $x_1 \preceq x_0$.

(ii) The partial order "$\preceq$" is called a *total* or *linear* order if any two elements of X are comparable.

(iii) A *chain* in a poset $(X, \preceq)$ is a subset $C \subset X$ such that any two elements in C are comparable

(iv) A *maximal element* of a partial order $\preceq$ on X is an element $x^* \in X$ such that for any $x \in X$ either x and x^* are not comparable, or $x \preceq x^*$.

(v) Suppose that S is a subset of a poset $(X, \preceq)$. An *upper bound* for S is an element $\bar{x} \in X$ such that

$$\forall s \in S \ \ s \prec \bar{x}.$$

In particular $\bar{x}$ is comparable with all the elements in S.

□

Example A.1.4. Suppose that Y is a set with at least two elements and $\mathcal{X}$ is the collection of proper subsets of Y, i.e., subsets $S \neq Y$. The set $\mathcal{X}$ is naturally ordered by the inclusion of subset. Then for any $y \in Y$ the subset $S_y = Y \backslash \{y\}$ is maximal with respect to the inclusion relation. The poset $(\mathcal{X}, \subset)$ has no upper bound.

On the other hand, observe that the set of real numbers with the usual order relation $\leqslant$ has no upper bound or maximal elements. □

The next famous result is one of the most powerful tools we have at our disposal when dealing with infinite sets and, in particular, with infinite dimensions. For a proof and more details on its important role in set theory we refer to [25, Chap. 8, Thm. 1.13].

Theorem A.1.5 (Zorn's Lemma)**.** *Suppose that $(X, \preceq)$ is a poset such that every chain has an upper bound. Then X itself has a maximal element.* □

Let us emphasize that Zorn's Lemma is an *existence* result. It is non-constructive in the sense that it gives no generally applicable method of finding the claimed maximal. The next result is a typical application of Zorn's Lemma.

Theorem A.1.6. *Suppose that V is a, possibly infinite dimensional, real vector space and $S \subset V$ is a linearly independent collection of vectors. Then S is contained in some basis B of V, i.e., a linearly independent collection spanning V.*

Proof. . Denote by $\mathcal{X}_S$ the family of linearly independent collections $X \subset V$ such that $S \subset X$.

Let us first show that any chain $\mathscr{C} \subset \mathcal{X}_S$ has an upper bound. Denote by C^* the union of all the sets in the chain $\mathscr{C}$. Clearly C^* is an upper bound of $\mathscr{C}$. We will show that C^* is also a linearly independent collection containing S. Suppose that a linear combination of vectors on C^* is trivial

$$\sum_{k=1}^{n} \lambda_k v_k$$

where $\lambda_k \in \mathbb{R}$, $v_k \in C_k \in \mathscr{C}$, $\forall k = 1, \dots, n$. We have to show that

$$\lambda_1 = \cdots = \lambda_n = 0.$$

Since $\mathscr{C}$ is a chain of subsets, one of the collections $C_1, \dots, C_n$ contains all the others, say

$$C_k \subset C_n, \;\; \forall k \leqslant n.$$

Thus

$$v_k \in C_n, \;\; \forall k = 1, \dots, n.$$

Since the collection C_n is linearly independent we deduce $\lambda_1 = \cdots = \lambda_n = 0$ so that $C^* \in \mathcal{X}_S$. From Zorn's Lemma we deduce that $\mathcal{X}_S$ contains a *maximal* element B. We claim that B is a basis of V.

Note first that since $B \in \mathcal{X}_S$ the collection B is linearly independent and contains S. To prove that it spans V we argue by contradiction. Suppose that there exists an element $v \in V \backslash \operatorname{span}(B)$ and therefore the collection $\widehat{B} = B \cup \{v\}$ is linearly independent and contains S. This contradicts the maximality of B. □

Zorn's Lemma is equivalent to an innocent looking yet very debated axiom of set theory. Loosely speaking this axiom postulates that given any collection of sets there exists a procedure of extracting an element from each set of the collection. Here is the precise statement.

The Axiom of Choice. *For any collection nonempty sets $(S_i)_{i\in I}$ there exists a choice function, i.e., a function*

$$f : I \to \bigcup_{i\in I} S_i,$$

such that

$$f(i) \in S_i,\ \ \forall i \in I.$$

For more information about the special role this axiom plays in modern set theory we refer [25, Chap. 8].

The axiom of choice is a nonconstructive statement since it postulates the existence of an object without any indication on how one could effective find it. The proofs that are based on statements equivalent to the axiom of choice are called *nonconstructive*.

A.2. Equivalence relations

Let X be a set. An *equivalence relation* on X is a binary relation on X that is *reflexive*, *symmetric* and *transitive*.

Example A.2.1. Fix an integer $d > 1$. We define a binary relation on $\mathbb{Z}$ by declaring $x, y \in \mathbb{Z}$ related if their difference $x - y$ is a multiple of d. We use the notation

$$x \equiv y \bmod d$$

to indicated this. This relation is reflexive since $x - x = 0$ is clearly a multiple of d. It is symmetric because $x - y$ is a multiple of d iff $y - x = -(x - y)$ is a multiple of d. Finally, it is transitive because if $x - y$ and $y - z$ are multiples of d then so is their sum $x - z = (x - y) + (y - z)$. □

Proposition A.2.2. *Let X be a set and $\sim$ and equivalence relation on X. For each $x \in X$ we set*

$$C_x := \big\{\, y \in X;\ x \sim y \,\big\}. \tag{A.2.1}$$

The set C_x is called the equivalence class of x.

(i) *For any $x_0, x_1 \in X$, $C_{x_0} = C_{x_1} \Longleftrightarrow x_0 \sim x_1$.*

(ii) *For any $x_0, x_1 \in X$, $C_{x_0} \cap C_{x_1} \neq \varnothing \Longleftrightarrow C_{x_0} = C_{x_1}$.*

Proof. (i) Note that $x_0 \sim x_1$, then $y \sim x_0$ if and only if $y \sim x_0 \sim x_1$, i.e., $y \in C_{x_0}$ if and only $y \in C_{x_0}$. Thus $C_{x_0} = C_{x_1}$. Conversely, if $C_{x_0} = C_{x_1}$ then $x_1 \in C_{x_0}$ so $x_0 \sim x_1$.

(ii)

□

Thus, an equivalence relation "$\sim$" classes partitions the set X into equivalence classes. The equivalence classes are sometime referred to as the *chambers* or *cells* of equivalence class. We denote by $X/\sim$ the collection of equivalence classes. Note that we have a natural surjection

$$\pi : X \to X/\sim, \;\; x \mapsto C_x.$$

The set $X/\sim$ is called the *quotient* of X with respect to the equivalence relation $\sim$.

A.3. Cardinals

We survey, mostly without proofs a few classical facts about *cardinality* theory. For more details we refer to [**25, 41**].

Two sets A, B are said to be *equipotent* or *have the same cardinality* when there is a bijection between the two sets. We will indicate this using the notation $A \sim B$.

One can prove[1] that there exists a set $\mathscr{Card}$ called *set of cardinals* and a "correspondence"[2] that associates to each set A its cardinality $\operatorname{card} A \in \mathscr{Card}$ so that

$$\operatorname{card} A = \operatorname{card} B \Longleftrightarrow A \sim B.$$

Given two sets A, B we write $\operatorname{card} A \leqslant \operatorname{card} B$ if there exists an injection $A \hookrightarrow B$. We write $\operatorname{card} A < \operatorname{card} B$ if $\operatorname{card} A \leqslant \operatorname{card} B$ yet $\operatorname{card} A \neq \operatorname{card} B$.

The set $\mathscr{Card}$ of cardinals is partially ordered by $\leqslant$. One can show that this is a total order. More precisely we have the following result.

Theorem A.3.1. *Given two sets we have either* $\operatorname{card} A \leqslant \operatorname{card} B$ *or* $\operatorname{card} B \leqslant \operatorname{card} A$. □

Example A.3.2. (a) For any natural number n we denote by $\mathbb{I}_n$ the "interval" $\mathbb{N} \cap [1, n]$ and we write $n := \operatorname{card} \mathbb{I}_n$. A set A is called *finite* if $\operatorname{card} A = n$ for some $n \in \mathbb{N}$. Otherwise the set is called *infinite*.

(b) We write $\operatorname{card} A = \aleph_0$ if $\operatorname{card} A = \operatorname{card} \mathbb{N}$. If this is the case we say that A is countable.[3]

(c) We set $\aleph_c := \operatorname{card} \mathbb{R}$ and we say that $\aleph_c$ is the cardinality of the continuum.

(d) Given two cardinals κ_0, κ_1 we define their sum $\kappa_0 + \kappa_1$ to be the cardinality of the union of two disjoint sets A_i, $\operatorname{card} A_i = \kappa_i$, $i = 0, 1$. Their product $\kappa_0 \times \kappa_1$ is the cardinality of $A_0 \times A_1$.

(e) For any set A we denote by 2^A the set of all the subsets of S. For any cardinal κ we denote by 2^κ the cardinality of 2^A, where $\operatorname{card} A = \kappa$. □

Theorem A.3.3. *Let A be a set. Then the following statements are equivalent.*

(i) *The set is infinite.*

(ii) $\operatorname{card} A \geqslant n$, $\forall n \in \mathbb{N}$.

[1] This is highly nontrivial!

[2] Take the word correspondence with a grain of salt. There is no set of all sets.

[3] The symbol $\aleph$ is the first letter of the Hebrew alphabet and it is pronounced aleph.

(iii) $\operatorname{card} A \geqslant \aleph_0$.

(iv) *There exists a proper subset* $S \subsetneq A$ *such that* $\operatorname{card} S = \operatorname{card} A$.

□

Theorem A.3.4 (Cantor-Bernstein)**.** *Let* A, B *be two sets. Then*

$$\operatorname{card} A = \operatorname{card} B \Longleftrightarrow \operatorname{card} A \leqslant \operatorname{card} B \text{ and } \operatorname{card} B \leqslant \operatorname{card} A.$$

□

Theorem A.3.5 (Cantor)**.** *For any cardinal* κ *we have* $\kappa < 2^\kappa$.

Proof. We present the clever short proof. Fix a set A such that $\operatorname{card} A = \kappa$ Clearly $\operatorname{card} A \leqslant \operatorname{card} 2^A$ since the map

$$A \ni a \mapsto \{a\} \in 2^A$$

is an injection. We argue by contradiction and we assume that there exists a bijection $F : A \to 2^A$. Define

$$X := \bigl\{ a \in A;\ a \notin F(a) \bigr\}.$$

Since F is surjective, we deduce that there exists $a_0 \in A$ such that $X = F(a_0)$.

There are two options: either $a_0 \in X$ or $a_0 \notin X$. We will show that each of them leads to contradictions.

Indeed, if $a_0 \in X$, then this means $a_0 \notin X = F(a_0)$ meaning $a_0 \in X$! If $a_0 \notin X = F(a_0)$, this means $a_0 \in X$! □

The next result takes much more effort to prove and requires a precise definition of $\mathcal{C}ard$. For a proof we refer to [41, Sec. 4.6]

Theorem A.3.6. *For any infinite cardinal* $\kappa \in \mathcal{C}ard$ *we have*

$$\kappa + \kappa = \kappa,\ \ \kappa \times \kappa = \kappa.$$

□

Theorem A.3.7. $\aleph_c = 2^{\aleph_0}$.

Proof. Observe first that $\operatorname{card}(0,1) = \operatorname{card} \mathbb{R}$. Indeed we have a bijection

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

and a bijection

$$(0,1) \to (-\pi/2, \pi/2),\ \ t \mapsto -\pi/2 + \pi t.$$

We identify $2^{\mathbb{N}}$ with the set of maps $\mathbb{N} \to \{0,1\}$ by associating to a subset $S \subset \mathbb{N}$ its indicator function $\boldsymbol{I}_S : \mathbb{N} \to \{0,1\}$.

To every $x \in (0,1)$ we associate its binary description

$$x = 0.\epsilon_1\epsilon_2\ldots\epsilon_k\ldots := \sum_{k\geqslant 1}\frac{\epsilon_k}{2^k}, \;\; \epsilon_k = 0,1$$

where infinitely many of the ϵ_k's are equal to 0. In other words we do not allow all ϵ's to be 1 after a while. For example

$$0.01111\ldots = \sum_{k\geqslant 2}\frac{1}{2^k} = \frac{1}{2} = 0.10000\ldots.$$

Such a binary sequence can be identified with the indicator function of a set $S \subset \mathbb{N}$ such that its complement is infinite. This proves

$$\aleph_c \leqslant 2^{\aleph_0}.$$

Thus we have identified $(0,1)$ with the family of subsets of $\mathbb{N}$ with infinite complement. This identification misses the subsets with finite complements. There are $\aleph_0$ of them. Hence

$$2^{\aleph_0} = \aleph_c + \aleph_0 \leqslant \aleph_c + \aleph_c = \aleph_c.$$

□

Bibliography

[1] V. I. Arnold: *Mathematical Methods of Classical Mechanics*, 2nd Edition, Springer Verlag, 1989.

[2] E. Artin: *The Gamma Function*, Holt, Rinehart & Winston, 1964.

[3] F. Sunyer i Balaguer, E. Corominas: *Sur des conditions pour qu'une fonction infiniment dérivable soit un polynôme*. Comptes Rendues Acad. Sci. Paris, 238 (1954), 558-559.

[4] V. Barbu: *Differential Equations*, Springer Undergraduate Mathematics Series, Springer Verlag, 2016.

[5] V. I. Bogachev: *Measure Theory. Vol 2*, Springer Verlag, 2007.

[6] E. Çinlar: *Probability and Stochastics*, Graduate Texts in Math., vol. 261, Springer Verlag, 2011.

[7] H. L. Cohn: *Measure Theory*, 2nd Edition, Birkhäuser, 2013.

[8] W.A. Coppel: *Number Theory. An Introduction to Mathematics*, 2nd Edition, Universitext, Springer Verlag, 2009.

[9] R. Courant, D. Hilbert: *Methods of Mathematical Physics*, vol. 1, Wiley Classics Library, John Wiley&Sons, 1989.

[10] P. J. Daniell: *Integrals in an infinite number of dimensions*, Ann. of Math. **20**(1919), 281-288.

[11] H. Davenport: *The Higher Arithmetic*, 7th Edition, Cambridge University Press, 1999.

[12] J. Dieudonné: *Foundations of Modern Analysis*, Academic Press, 1969.

[13] J.J. Duistermaat, J.A. Kolk: *Mutidimensional Real Analysis I. Differentiation*, Cambridge University Press, 2004.

[14] J.J. Duistermaat, J.A. Kolk: *Mutidimensional Real Analysis II. Integration*, Cambridge University Press, 2004.

[15] P. Duren: *Invitation to Classical Analysis*, Pure and Applied Undergraduate Texts, Amer. Math. Soc., 2012.

[16] W. Feller: *An Introduction to Probability Theory and Its applications*, vol. 1, 3rd Edition, John Wiley & Sons, 1968.

[17] G. B. Folland: *Real Analysis. Modern Techniques and Their Applications*, John Wiley & Sons, 1999.

[18] A. Friedman: *Advanced Calculus*, Dover, 2007.

[19] B. R. Gelbaum, J.M.H. Olmstead: *Counterexamples in Analysis*, Dover, 2003.

[20] T. Hales: *The Jordan curve theorem, formally and informally*, Amer. Math. Monthly, **114**(2007), 882-894

[21] G. H. Hardy:*Weierstrass's non-differentiable function*, Trans. A.M.S., **17**(1916), 301-325.

[22] G. H. Hardy: *Divergent Series*, Oxford University Press, 1949.

[23] G.H. Hardy, J.E. Littlewood, G. Polya: *Inequalities*, 2nd Edition, Cambridge University Press, 1952.

[24] D. Hilbert, P. Bernays: *Foundations of Mathematics I*, C.P.Wirth, J.Siekmann, M.Gabbay, Editors, College Publications, 2011.

[25] K. Hrbacek, T. Jech: *Introduction to Set Theory*, 3rd Edition, Marcel Dekker, 1999.

[26] Y. Katznelson: *An Introduction to Harmonic Analysis*, 3rd Edition, Cambridge University Press, 2011.

[27] A. Knapp: *Basic Real Analysis*, 2nd Digital Edition, `https://www.math.stonybrook.edu/~aknapp/download/b2-realanal-clickable.pdf`

[28] K. Knopp: *Theory and Application of Infinite Series*, Dover, 1990.

[29] S. Lang: *Undergraduate Analysis*, 2nd Edition, Springer Verlag, 1997.

[30] L. H. Loomis: *Introduction to Abstract Harmonic Analysis*, Dover, 2011.

[31] J. Munkres: *Topology*, 2nd Edition, Pearson Eduction Ltd., 2014.

[32] P. J. Nahin: *The Probability Integral. Its Origin, Its Importance and Its Calculation*, Springer Verlag, 2023.

[33] E. Noether: *Invariante Variationsprobleme*, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse (1918): 235-257.

[34] J.H. Silverman: *A Friendly Introduction to Number Theory*, Prentice Hall, 1997.

[35] M. Spivak: *Calculus on Manifolds*, Perseus Books, Cambridge MA, 1965.

[36] J.M. Steele: *The Cauchy-Schwarz Master Class*, Cambridge University Press, 2004.

[37] A. Tarski: *Introduction to Logic and to the Methodology of Deductive Sciences*, Dover Publications, 1996.

[38] M. E. Taylor: *Measure Theory and Integration*, Grad. Stud. Math., vol. 76, Amer. Math. Soc., 2006.

[39] E. C. Titchmarsh: *The Theory of Functions*, 2nd Edition, Oxford University Press, 1952.

[40] S. Treil: *Linear Algebra Done Wrong*, `https://www.math.brown.edu/~treil/papers/LADW/LADW.html`

[41] R. L. Vaught: *Set Theory. An Introduction*, 2nd Edition, Brikhäuser, 1995

[42] E.T. Whittaker, G.N. Watson: *A Course of Modern Analysis*, 4th Edition, Cambridge University Press, 1927.

[43] K. Yosida: *Functional Analysis*, Springer Verlag, any edition.

[44] V.A. Zorich: *Mathematical Analysis I*, Springer Verlag, 2004.

[45] V.A. Zorich: *Mathematical Analysis II*, Springer Verlag, 2004.

[46] A. Zygmund: *Trigonometric Series*, 3rd Edition, (volumes I and II combined), Cambridge University Press, 2002.

,

Index