

Höhere Mathematik

RĂZVAN GHEORGHE GURĂU

INSTITUT FÜR THEORETISCHE PHYSIK DER UNIVERSITÄT HEIDELBERG,
PHILOSOPHENWEG 19, 69120 HEIDELBERG, GERMANY

Contents

1	Introduction	1
I	HöMa I	3
2	Sets and numbers	5
2.1	Sets (Mengen)	5
2.1.1	Functions (Funktionen)	6
2.1.2	The axiom of choice (Das Auswahlaxiom)	6
2.1.3	Binary relations (Zweistellige Relationen)	7
2.1.4	Zorn's lemma (Das Lemma von Zorn)	8
2.2	Numbers (Zahlen)	8
2.2.1	Natural numbers $\mathbb{N}$ (Natürliche Zahlen)	9
2.2.2	Ordinals and cardinal numbers (Ordinal und Kardinal Zahlen)	9
2.2.3	Integer numbers $\mathbb{Z}$ (Ganze Zahlen)	11
2.2.4	Rational numbers $\mathbb{Q}$ (Rationale Zahlen)	12
2.2.5	Real numbers $\mathbb{R}$ (Reelle Zahlen)	13
2.2.6	Complex numbers $\mathbb{C}$ (Komplexe Zahlen)	16
3	Topological spaces	19
3.1	Topology	19
3.1.1	Compact sets (Kompakter Menge)	20
3.1.2	Topological boundary, closure and interior (Topologischer Rand)	21
3.2	Metric spaces and the standard topology	22
3.2.1	The $\mathbb{R}^D$ metric space	23
4	Sequences	27
4.1	Sequences in metric spaces	27
4.1.1	Cauchy sequences (Cauchy Folgen)	28
4.1.2	Banach fixed-point theorem (Fixpunktsatz von Banach)	29
4.2	Sequences in $\mathbb{R}$ and $\mathbb{R}^D$	30
4.2.1	Homogeneous linear recurrence (Homogener lineare Differenzengleichung)	30
4.2.2	Limits of real sequences, lim sup and lim inf	32
4.2.3	Monotone convergence theorem (Satz von der monotonen Konvergenz)	32
4.2.4	The Bolzano-Weierstrass theorem (Satz von Bolzano-Weierstrass)	33

4.2.5	$\mathbb{R}^D$ is Cauchy complete	34
4.2.6	Heine–Borel theorem (Satz von Heine–Borel)	34
4.3	Series	36
4.3.1	Absolute convergent series (absolut konvergente Reihen)	37
4.3.2	Power series (Potenzreihen)	39
4.3.3	The exponential function (Exponentialfunktion)	40
5	Vector spaces, first contact	41
5.1	Vector spaces	41
5.2	Hamel bases, dimension of a vector space	42
5.3	Finite dimensional vector spaces	43
5.4	Linear operators	45
5.4.1	The dual space	45
5.5	Affine spaces	46
6	Matrices	49
6.1	The algebra of matrices	50
6.2	Change of basis	52
6.3	Gaussian elimination	53
6.3.1	Finding bases in the kernel and the image of a matrix	54
6.3.2	Solving systems of linear equations with constant coefficients	55
6.3.3	Computing the matrix inverse	56
6.4	The determinant	57
6.4.1	The determinant and the inverse of a matrix	59
6.4.2	Block matrices	61
6.5	Minors of a matrix	62
6.5.1	The matrix tree theorem	63
7	Functions, first contact	67
7.1	Continuity	67
7.2	The intermediate value theorem (Zwischenwertsatz)	68
7.2.1	The exponential function (Exponentialfunktion)	69
7.3	The extreme value theorem	70
7.4	Differentiation in one variable	72
7.4.1	The case $f : \mathbb{R} \rightarrow \mathbb{R}$	73
II	HöMa II	77
8	Spectral theory in finite dimensions	79
8.1	Eigenvalues and eigenvectors	80
8.1.1	Diagonalizable matrices	81
8.1.2	Spectral decomposition of diagonalizable matrices	82
8.1.3	Decomposition in invariant subspaces for a generic matrix	83
8.1.4	Spectral decomposition for generic matrices	84
8.1.5	The Jordan normal form	85

8.2	Bilinear, quadratic and sesquilinear forms	88
8.3	Scalar product	89
8.3.1	The volume of a parallelepiped	92
8.4	Spectral decomposition for self adjoint matrices	93
8.4.1	Singular values	94
8.5	Matrix norms	95
9	Banach spaces	97
9.1	Banach spaces	97
9.2	Linear operators	100
10	Functions	103
10.1	Continuity	103
10.2	The uniform limit theorem	104
10.3	Differentiation in Banach spaces	105
10.4	The inverse function theorem	108
10.5	The case $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$	109
10.6	The implicit function theorem	110
10.7	The case $F : \mathbb{R}^N \rightarrow \mathbb{R}$	112
10.7.1	Critical points	112
11	The Riemann integral	115
11.1	Continuous functions of one variable and primitives	115
11.1.1	Indefinite integrals, integration by parts and change of variables	118
11.1.2	Integrals depending on a parameter	118
11.1.3	Limitations of the Riemann integral	120
11.2	The Riemann integral in $\mathbb{R}^n$	120
11.2.1	The Jordan measure	120
11.2.2	The multidimensional Riemann integral	124
11.2.3	Indefinite integrals, change of variables	126
12	Elements of differential geometry	127
12.1	Manifolds	127
12.2	The tangent and cotangent spaces	129
12.2.1	Local coordinates and bases	130
12.2.2	Curves	131
12.2.3	Coordinate transformations	132
12.3	(Pseudo-)Riemannian manifolds	133
12.4	Differential forms	136
12.4.1	Operations on differential forms	137
12.4.2	Maxwell's equations	139
12.5	Pushforward, pullback and the volume form	141
12.5.1	The normal(s) to a submanifold	142
12.6	Lagrange multipliers	144
12.7	The Lie derivative, the covariant derivative and geodesics	145

12.8 The Poincaré lemma	149
13 The Stokes theorem	151
13.1 Stokes's theorem for a box domain	151
13.2 Euclidean $\mathbb{R}^n$	156
13.3 Triangulated spaces	159
13.3.1 The boundary operator	159
13.3.2 Simplicial (pseudo-)manifolds	161
13.3.3 Stokes–Cartan theorem	162
III HöMa III	165
14 Complex analysis	167
14.1 Differentiation in $\mathbb{C}$	167
14.2 Analytic functions	169
14.2.1 Analytic continuation	170
14.2.2 The complex logarithm	172
14.3 Contour integrals	172
14.3.1 Cauchy's integral formula	175
14.3.2 Entire functions	176
14.4 Critical points of holomorphic functions and the steepest descent method	177
14.5 Mermorphic functions and the residue theorem	179
14.5.1 Isolated singularities of complex functions	180
14.6 Borel summation	182
15 Hilbert spaces (Hilbert Raum)	187
15.1 Properties of Hilbert spaces	187
15.2 Operators on Hilbert spaces	191
15.2.1 Linear forms and the dual space	192
15.2.2 Operators from V to V	193
15.3 Spectral theory	195
15.4 Quantum mechanics	199
16 The Lebesgue integral, first contact	203
16.1 The Lebesgue measure	203
16.1.1 Lebesgue measurable sets	206
16.1.2 Measure spaces	208
16.2 The Lebesgue integral	209
16.2.1 The L^p spaces	214
17 The Fourier transform	217
17.1 Fourier series	217
17.1.1 Canonical quantization and the spectrum of the $-i\frac{d}{dx}$ differential operator	220
17.1.2 Summing over the Fourier frequencies, Matsubara weighting function	221
17.1.3 Other bases in $L^2([0, T], \mathbb{C})$	222

17.2	The Schwartz space	224
17.3	The Fourier transform	226
18	Distributions	231
18.1	Tempered distributions	231
18.1.1	The Sokhotski–Plemelj formula	234
18.1.2	The Dirac comb and the Poisson summation formula	235
18.1.3	The Riemann zeta function	235
18.2	Differential operators and fundamental solutions	236
18.2.1	The Laplacian	238
18.2.2	The heat diffusion operator and the heat kernel	239
18.2.3	The d’Alembert operator	240
18.2.4	Green’s functions	242
18.3	Generalized eigenvectors	242
19	Differential equations	249
19.1	Initial value problems	249
19.1.1	Flows of vector fields	251
19.1.2	First integrals	252
19.1.3	The Kepler problem	252
19.2	Linear ordinary differential equations	253
19.2.1	Linear equations with constant coefficients	256
19.2.2	Generic linear first order ODE	257
19.3	Sturm–Liouville theory	258
19.3.1	The role of the boundary / initial conditions	261
19.3.2	Hermite functions	262
19.3.3	The Legendre polynomials	263
19.3.4	The multipole expansion	266
19.4	Power series solutions	267
19.4.1	The hypergeometric series	267
19.5	Partial differential equations	268
19.5.1	The Poisson equation	269
19.5.2	The spherical harmonics	271
19.5.3	Separation of variables	273
19.5.4	The Hydrogen atom	273
Appendices		
A	Lie groups and representations, a primer	279
A.1	Classical Lie groups	279
A.1.1	The $SU(N)$ Lie algebra	280
A.1.2	The adjoint representation	281
A.2	$SO(3)$ versus $SU(2)$	282
A.3	The Haar measure	283
A.4	Finite dimensional unitary representations of $SU(2)$	285

A.4.1	Classification of the irreducible finite dimension unitary representations of $SU(2)$	286
A.4.2	$SU(2)$ recoupling theory	287
A.4.3	Wigner D -matrices	288
A.5	$SU(N)$ Yang Mills theory	289
B	Exams	291
B.1	Sample exams HöMa II	291
B.1.1	Sample 1	291
B.1.2	Sample 2	293
B.1.3	Sample 3	294
B.1.4	Sample 4	297
B.2	Sample exams HöMa III	299
B.2.1	Sample 1	299
B.2.2	Sample 2	301
B.2.3	Sample 3	304
B.2.4	Sample 4	307

Chapter 1

Introduction

Axioms are statements which are accepted to be true. *Theorems* (lemmas, propositions, etc.) are statements that can be proven using logical reasoning starting from axioms. We know since the work of Gödel that no system of axioms is complete, that is for any system of axioms which includes the natural numbers it is possible to formulate statements that are undecidable, that is can neither be proved true nor be proven false. This is because, essentially, there are more possible statements that can be formulated than there are possible proofs.

Reductio ad absurdum is a strategy of proving a statement is true by first assuming the statement is false and proving that this implies a contradiction (i.e. some other false statement). As a true statement can not imply a false one, one concludes that the original statement is true.

Mathematical induction is a strategy of proving a statement indexed by a natural number n is true by first proving it is true for $n = 0$ and then proving that if the statement for n is true then the one for $n + 1$ is also true. Transfinite induction is indexed by ordinal numbers instead of natural ones.

The following logic symbols are useful $\exists, \forall, \Rightarrow, \Leftrightarrow, =$. We denote $\equiv$ when the left hand side and the right hand side are equal by definition.

Part I

HöMa I

Chapter 2

Sets and numbers

We review some elementary notions which will be useful for the sequel. In the beginning we will denote the multiplication of numbers by $\times$, but drop this notation later on.

2.1 Sets (Mengen)

The notion of *set* is a primitive notion in mathematics, therefore it can not be defined in terms of other more fundamental notions. Instead of providing a definition, one lists the axioms sets obey. Informally, a set is the mathematical model for a collection of different things.

A set X contains elements x , which we denote $x \in X$. The elements are mathematical objects of any kind. The only question one can ask about a set and an element is whether the element belongs to the set. This has the consequence that, when listing the elements of a set, every element is written only once: $\{2\} = \{2, 2\} = \{2, 2, 2\}$ and so on, as they all represent the set which contains only the element 2.¹

We follow the Zermelo–Fraenkel axioms of set theory plus the axiom of choice. The axiom of choice is sometimes contested, but not by us. Most axioms are intuitive.

Axiom 2.1 (*Zermelo–Fraenkel axioms (ZF-Axiome)*). We denote the empty set $\emptyset$ the set with no elements. We have

- (Extension) A set is determined by its elements, that is, if $x \in A \Leftrightarrow x \in B \Rightarrow$ then $A = B$. We denote $B \subset A$ if $x \in B \Rightarrow x \in A$.
- (Pairing) If A and B are sets, then $\{A, B\}$ is a set.
- (Subsets or specification) If A is a set and $P(x)$ is a property of some of the elements x , then $\{x \in A \mid P(x)\}$ is a set.
- (Union, or sum set) For any set of sets² A there exist a set $\bigcup A = \{x \mid x \in D \in A\}$ which is the union of all the elements of the elements of A . In particular this allows us to unpack the pairs of sets:

$$\bigcup \{A, B\} = \{x \mid x \in A \text{ or } x \in B\} \equiv A \cup B.$$

For instance $\bigcup \{\{1, 2\}, \{2, 3\}\} = \{1, 2, 3\}$.

- (Power set) For any set A there exists the power set $\mathcal{P}(A)$ consisting in all the subsets of A , $\mathcal{P}(A) = \{B \mid B \subset A\}$. As the empty set is a subset of any set, $\emptyset \in \mathcal{P}(A)$ for any A .

¹Multi sets which can contain the same element more than once can be defined.

²In ZF theory everything is a set, so any set is a set of sets.

- (Infinity). Let w be some set. Then, by the axiom of pairing $\{w\}$ is a set (obtained by pairing w with itself $\{w, w\} \equiv \{w\}$) and so is $S(w) = w \cup \{w\} = \bigcup\{w, \{w\}\}$. Note that $\bigcup\{w\} = w$, but $\{w\} \neq w$.

There exist a set A such that $\emptyset \in A$ and $w \in A \Rightarrow S(w) \in A$.

- (Replacement) The image of a set under a function is a set, that is if f is a function and A is a set, $f(A) = \{f(x) \mid x \in A\}$ is a set.
- (Foundation, or regularity) Every non empty set A contains an element that is disjoint from A .

We denote the intersection of the sets A and B the set consisting of the elements which belong to both A and B , that is $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$. Two sets are disjoint if their intersection is the empty set. The axiom of foundation states that for all $A \neq \emptyset$, there exist $x \in A$ such that $x \cap A = \emptyset$.

This implies that no set is an element of itself: for any set A , $\{A\}$ is a set by the axiom of pairing and by the axiom of regularity $\{A\}$ must contain an element which is disjoint from $\{A\}$. As $\{A\}$ contains only A , it follows that $A \cap \{A\} = \emptyset$. Reading this from right to left, as $A \in \{A\}$ it follows that $A \notin A$.

The set difference of X and Y consists in the element of X which do not belong to Y , $X \setminus Y = \{x \mid x \in X \text{ and } x \notin Y\}$. The Cartesian product of two non empty sets X and Y is the set of ordered pairs $X \times Y = \{(x, y) \mid x \in X \text{ and } y \in Y\}$.

2.1.1 Functions (Funktionen)

Definition 2.1 (**Functions**). A function from a set X to a set Y is an assignment of an unique element of Y to each element of X . The set X is called the domain of the function and the set Y is called the codomain of the function.

A function, its domain, and its codomain, are declared by the notation $f : X \rightarrow Y$, and the value of a function f at an element $x \in X$, denoted by $f(x)$, is called the image of x under f , or the value of f applied to the argument x . Functions are sometimes called maps or mappings. For example, the *identity function* $\text{id} : X \rightarrow X$ is the function which assigns to each element of X itself, that is $\forall x \in X, \text{id}_X(x) = x$.

Two functions f and g are equal if their domain and codomain sets are the same and their output values agree on the whole domain, that is for $f, g : X \rightarrow Y$ we denote $f = g$ if $f(x) = g(x), \forall x \in X$. The *composition* of two functions $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ is the function $g \circ f : X \rightarrow Z$ defined by $\forall x \in X, g \circ f(x) = g(f(x)) \in Z$. The *image* of a set $U \subset X$ under the function $f : X \rightarrow Y$ is the set $f(U) = \{y \in Y \mid \exists x \in U \text{ s.t. } f(x) = y\} \subset f(X)$. The *preimage* of a set $V \subset f(X)$ is the set $f^{-1}(V) = \{x \in X \mid f(x) \in V\} \subset X$.

A function $f : X \rightarrow Y$ is called *injective*, or *one to one*, if each element in Y is the image of at most one element in X , that is $\forall x_1, x_2 \in X, f(x_1) = f(x_2) \Leftrightarrow x_1 = x_2$. A function $f : X \rightarrow Y$ is called *surjective*, or *onto*, if $f(X) = Y$ that is each element in Y is the image of some element in X : $\forall y \in Y, \exists x \in X$ such that $f(x) = y$. A function which is both injective and surjective (one to one and onto) is called *bijective*. A bijection $f : X \rightarrow Y$ has a well defined *inverses*, that there exists a function $f^{-1} : Y \rightarrow X$ such that $f^{-1} \circ f = \text{id}_X$ and $f \circ f^{-1} = \text{id}_Y$. Indeed, as f is a bijection, then $\forall y \in Y, \exists x \in X$ (due to surjectivity) and x is unique (due to injectivity) such that $f(x) = y$. Let us denote this element x_y . The *inverse* of f is the function $f^{-1} : Y \rightarrow X$ defined by $f^{-1}(y) = x_y$.

2.1.2 The axiom of choice (Das Auswahlaxiom)

The most disputed axiom of set theory is the axiom of choice.

Axiom 2.2 (**The axiom of choice**). For any set $\mathcal{X}$ of nonempty sets, there exists a choice function $f : \mathcal{X} \rightarrow \bigcup \mathcal{X}$ that maps each set $X \in \mathcal{X}$ to an element $x \in X$.

Reformulated, this means that for any set of nonempty sets $\{X_\alpha \neq \emptyset \mid \alpha \in A\}$, there exists a set that contains exactly one element per X_α , that is there exists the set $\{x_\alpha \in X_\alpha \mid \alpha \in A\}$. What is non obvious about this axiom is that the index set A can be anything, for instance the real numbers, a set of ordinal numbers (see below) etc. In such cases it is often non trivial to exhibit a choice function and one needs to rely on the axiom of choice to conclude that such a function exists.

2.1.3 Binary relations (Zweistellige Relationen)

Elements of sets can be in binary relations of several types.

Definition 2.2 (**Binary relations**). *Several binary relations among the elements of a set X important for the sequel. We say that:*

- $\sim$ is an equivalence relation on X if:
 - (reflexive) $\forall x \in X$, then $x \sim x$,
 - (symmetric) $\forall x, y \in X$, if $x \sim y$ then $y \sim x$,
 - (transitive) $\forall x, y, z \in X$, if $x \sim y$ and $y \sim z$ then $x \sim z$.
- $\leq$ is a partial order relation on X if:
 - (reflexive) $\forall x \in X$, then $x \leq x$,
 - (antisymmetric) $\forall x, y \in X$, if $x \leq y$ and $y \leq x$ then $y = x$,
 - (transitive) $\forall x, y, z \in X$, if $x \leq y$ and $y \leq z$ then $x \leq z$.

A set with a partial order relation is called a partially ordered set, or a poset. A poset can contain pairs of elements that are not ordered, that is there can exist $x, y \in X$ such that $x \not\leq y$ and $y \not\leq x$. A partial order is a total order if any two elements are ordered:

- (connected) $\forall x, y \in X$, either $x \leq y$ or $y \leq x$.
- $<$ is a strict partial order relation on X if:
 - (irreflexive) $\forall x \in X$, then $x \not< x$,
 - (asymmetric) $\forall x, y \in X$, if $x < y$ then $y \not< x$,
 - (transitive) $\forall x, y, z \in X$, if $x < y$ and $y < z$ then $x < z$.

For every partial order $\leq$ there exists an associated strict partial order $<$ defined by $x < y$ if $x \leq y$ and $x \neq y$.

A strict partial order is a strict total order relation on X if:

- (connected) $\forall x, y \in X$, if $x \neq y$ then either $x < y$ or $y < x$,
- $\leq_w$ is a well-ordering relation on X if
 - (total order) $\leq_w$ is a total order on X ,
 - (least element) every non empty subset of X has a least element, that is $\forall Y \subset X, Y \neq \emptyset$, there exists $y \in Y$ such that $\forall z \in Y, y \leq_w z$.

A finite poset can be represented by a Hasse diagram with nodes the elements $x \in X$ and links between x and y if $x < y$ but there exists no z with $x < z < y$.

2.1.4 Zorn's lemma (Das Lemma von Zorn)

Lemma 2.1 (**Zorn's lemma**). A chain T in a partially order set X is a subset $T \subset X$ that is totally ordered, that is $\forall x, y \in T$ either $x \leq y$ or $y \leq x$. If any chain T in X has an upper bound, that is $\forall T$ chain, $\exists x_T \in X$ such that $\forall x \in T$, $x \leq x_T$, then X has at least one maximal element, that is an element that is not smaller than any other element in X . Formally, $\exists x_X \in X$ such that $\forall x \in X \setminus \{x_X\}$, $x_X \not\leq x$.

This lemma is in fact equivalent to the axiom of choice. In proving this lemma, the axiom of choice is used as follows. Assume there exists a partially ordered set X such that every chain has an upper bound and also for every element in X there is another element bigger than it. For every chain T we then define $b(T) \in X$, an element *strictly* larger than any element in T (this exists because T has an upper bound, and that upper bound, being an element in X , has a bigger element). In order to actually choose $b(T)$ we need to use the axiom of choice. The contradiction comes from building, using the function b , a strictly totally ordered sequence in X which is “too long”, that is it has more elements than there can be in any set as it is indexed by the ordinal numbers which are a proper class, not a set (see below).

Exercise 2.1 (The poset of partitions). A partition Π of a set X is a set of subsets of X , $\Pi = \{X_a \mid a \in A \neq \emptyset\}$ such that $X = \bigcup_{a \in A} X_a$ and $X_a \cap X_b = \emptyset$ if $a \neq b$. A partition Π_1 is finer than Π_2 (and Π_2 is coarser than Π_1) if $\forall X \in \Pi_1$, $\exists Y \in \Pi_2$ such that $X \subset Y$. We denote $\Pi_1 \leq \Pi_2$.

Prove that $\leq$ is a partial order on the set of partitions. Check that this poset verifies Zorn's lemma. Show that this poset is a lattice, that is a poset in which every pair of elements x, y has a unique supremum $x \vee y$ (smallest element larger than both, called the join) and a unique infimum $x \wedge y$ (largest element smaller than both, called the meet). Show that this lattice has a greatest lower bound as well as a smallest upper bound. For Zorn's lemma - the upper bound on any chain in the poset is the join of all the partitions in the chain; the poset has a maximal element, the one set partition.

Exercise 2.2 (Equivalence classes, quotient). Let X be a set with an equivalence relation. For any $x \in X$, we define the equivalence class $[x] = \{y \in X \mid y \sim x\}$. Prove that an equivalence relation partitions a set into the disjoint union of equivalence classes. First $X = \bigcup_{x \in X} [x]$, that is X is the union of equivalence classes of its elements. Second, two equivalence classes are either equal or disjoint: assume that two equivalence classes are not disjoint $[x] \cap [y] \neq \emptyset$. Then $\exists z \in [x] \cap [y]$ hence $z \sim x$ and $z \sim y$ and (by transitivity) $x \sim y$. But then, again by transitivity $\forall z' \in [x]$, $z' \sim x \sim y$ hence $z' \in [y]$ and $[x] \subset [y]$. Similarly, $[y] \subset [x]$ hence $[x] = [y]$.

We denote $\{[x] \mid x \in X\} \equiv X / \sim$ and we call it the quotient of X by $\sim$.

Exercise 2.3 (Equivalence classes of permutations.). The permutations of n elements are the bijections on the finite set with n elements $\sigma : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$. We denote the set of all the permutations over n elements by S_n . The composition of permutations is sometimes denoted as a product $\sigma_2 \circ \sigma_1 \equiv \sigma_2 \sigma_1$. For $\sigma_1, \sigma_2 \in S_n$, we denote $\sigma_1 \sim \sigma_2$ if there exists $\pi \in S_n$ such that $\sigma_1 = \pi \sigma_2 \pi^{-1}$.

How many elements are there in S_n ? Prove that $\sim$ is an equivalence relation. Characterize the quotient set $S_n / \sim$. How many elements are there in the equivalence class of a permutation? Hint: any permutation can be decomposed into a product of disjoint cycles.

2.2 Numbers (Zahlen)

We now discuss the various sets of numbers we will encounter in these notes. Different numbers are introduced in order to solve equations: solving $x + n = 0$ with n natural leads to integers, solving $mx + n = 0$ for m, n integers leads to rational numbers, $x^2 = r > 0$ with r rational leads to real numbers and finally $P[x] = 0$ for some polynomial P leads to complex numbers.

2.2.1 Natural numbers $\mathbb{N}$ (Natürliche Zahlen)

The natural numbers are a primitive notion. Informally they are obtained by starting from 0 and iteratively adding 1 an arbitrary number of times. One can construct the natural numbers starting from sets. This construction is then generalized to ordinal numbers.

Construction 2.1 (Natural numbers $\mathbb{N}$). *By the axiom of infinity, there exist sets which contain the empty set $\emptyset$ and are closed under the successor function (such sets are said to be inductive). The smallest such set is the set of natural numbers $\mathbb{N}$ and is constructed in detail as follows:*

- zero is the empty set $0 = \emptyset$.
- the successor $S(a)$ of any set a is $S(a) = a \cup \{a\}$ and we label:

$$\begin{aligned} 0 &= \emptyset, \\ 1 &= 0 \cup \{0\} = \{0\} = \emptyset \cup \{\emptyset\} = \{\emptyset\}, \\ 2 &= 1 \cup \{1\} = \{0, 1\} = \{\emptyset\} \cup \{\{\emptyset\}\} = \{\emptyset, \{\emptyset\}\}, \\ 3 &= 2 \cup \{2\} = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}\} \cup \left\{ \{\emptyset, \{\emptyset\}\} \right\} = \left\{ \emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\} \right\}. \end{aligned}$$

- $\mathbb{N}$ is the set consisting in $\emptyset$ and all its successors.

We endow $\mathbb{N}$ with further structure:

- (addition) the *addition* of natural numbers is the internal binary operation $+: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ defined as $\forall a, b \in \mathbb{N}, a + 0 = a$ and $a + S(b) = S(a + b)$,
- (multiplication) the *multiplication* is the internal binary operation $\times: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ defined as $\forall a, b \in \mathbb{N}, a \times 0 = 0$, $a \times S(b) = a + a \times b$.
- (total order) the natural numbers are *totally ordered* by the relation $a \leq b$ if $a \in b$ or $a = b$. Equivalently, there exists $c \in \mathbb{N}$ such that $a + c = b$. This total order is a well order.

The operation $+$ substitutes the successor function by denoting $S(0) = 1$, $S(a) = a + S(0) = a + 1$. We use the notation $1, 1+1=2, 2+1=3$ and so on and write the natural numbers as $\mathbb{N} = \{0, 1, 2, \dots, n, \dots\}$. We also denote $n \times n \equiv n^2$, $n \times n^2 = n^3$ and so on.

Exercise 2.4. *Prove that $+$ is commutative $a+b = b+a$ and associative $a+(b+c) = (a+b)+c$. We have $0+0=0$. For $a \neq 0$ we proceed by induction:*

we show that the statement for $S^{-1}(a)$ implies the statement for a . We first show that 0 commutes with every a . It is enough to show that $0+a = a$. Indeed, $0+a = S(0+S^{-1}(a)) = S(S^{-1}(a)) = a$. Now we want to show that $S(0)$ commutes with any a . First $a+S(0) = S(a+0) = S(a)$. Second, for $a \neq 0$ we have $a = S(S^{-1}(a)) = S(S^{-1}(a)+0) = S^{-1}(a)+S(0)$. Assuming $S(0)+S^{-1}(a) = S^{-1}(a)+S(0)$, then $S(0)+a = S(S(0)+S^{-1}(a)) = S(S^{-1}(a)+S(0)) = S(a)$. We then proceeded to $S(S(0))$ and so on.

Prove that $\times$ is commutative, associative and distributive with respect to the addition, that is $a \times (b+c) = a \times b + a \times c$. Prove that $\leq$ is a total order on $\mathbb{N}$ and that it is compatible with $+$ and $\times$, that is $\forall a, b, c \in \mathbb{N}, a \leq b \Rightarrow a+c \leq b+c$ and $a \times c \leq b \times c$. Finally, prove that $n \times p = n + n + \dots + n$ where the addition is performed p times.

2.2.2 Ordinals and cardinal numbers (Ordinal und Kardinal Zahlen)

Ordinal numbers generalise the notion of numerals (first, second, third and so on) to infinite sets. Roughly speaking, ordinals are labels which include the natural numbers, are totally ordered and are such that every set of ordinals is well ordered.

Definition 2.3 (von Neumann ordinals). An ordinal λ is the set having as elements precisely the ordinals strictly smaller than itself, which we denote by $\lambda = [0, \lambda)$. We stress that λ itself is not an element of λ .

More precisely, an ordinal λ is a transitive strictly well ordered set by set membership:

- $\in$ is a strict well order relation among the elements in the set λ :
 - that is for all $x \in \lambda$, $x \notin x$; for all $x, y \in \lambda$, if $x \neq y$ then either $x \in y$, or $y \in x$; for all $x, y, z \in \lambda$, if $x \in y$ and $y \in z$ then $x \in z$,
 - for all $A \subset \lambda$, $A \neq \emptyset$, $\exists x \in A$ such that $\forall y \in A$ either $x \in y$ or $x = y$.
- λ is transitive, that is for all $x \in \lambda$ then $x \subset \lambda$.

For every ordinal $\lambda = [0, \lambda)$, we denote $\lambda + 1 = \lambda \cup \{\lambda\}$ the successor of λ . It is immediate that $\lambda \cup \{\lambda\}$ is also an ordinal, and it is in fact the least element of all the ordinals strictly larger than λ . We always have $\bigcup \lambda \subset \lambda$. The ordinals themselves are strictly well ordered by set membership: for β and λ two ordinals, $\beta < \lambda$ if $\beta \in \lambda$. Two ordinals either coincide or one is strictly large than the other.

It can be proven that the union of any set of ordinals is an ordinal and every set of ordinals has a least upper bound (a supremum) obtained by taking the union of all the ordinals in the set. A *limit ordinal* is an ordinal that is neither 0 nor the successor of an ordinal³ and limit ordinals arise as least upper bounds on sets of all smaller ordinals:⁴ λ is a limit ordinal if $\forall \beta < \lambda$, $\exists \gamma$ s.t. $\beta < \gamma < \lambda$. Every ordinal is either $0 = \emptyset$, or a successor ordinal or a limit ordinal and in the latter case, $\lambda = \bigcup \lambda$.

By construction, the natural numbers are ordinals. The smallest ordinal greater than every natural number, denoted ω , is a limit ordinal: this is nothing but the well ordered infinite set of all the natural numbers. To check that ω is a limit ordinal we observe that, by definition, $n < \omega \Rightarrow n \in \mathbb{N}$ but $n < n + 1 \in \mathbb{N}$ and, because $n + 1$ is itself a natural number, $n + 1 < \omega$. The successor of ω is the ordinal $\omega + 1 = \omega \cup \{\omega\}$. A construction of ω is provided by the two letter words with alphabet $\mathbb{N}$ arranged in lexicographical order $(0, 0), (0, 1), (0, 2), \dots, (1, 0), (1, 1), \dots$. Then ω is the position of $(1, 0)$, $\omega + 1$ the one of $(1, 1)$, $\omega \cdot 2$ the one of $(2, 0)$ and so on. The smallest ordinal larger than any $\omega \cdot n$ is denoted ω^2 .

The class of all ordinals is *not a set*, but a proper class. If it were a set, this class would be an ordinal λ (obtained by taking the supremum of all the ordinals) hence a set. But, as λ contains all the ordinals, $\lambda \in \lambda$, which is a contradiction.

The axiom of choice implies that every set can be well ordered (Zermelo's theorem) by choosing the element given by the choice function on the subset of elements not yet chosen. In detail, take f a choice function for the non empty subsets of A and define for every ordinal α the element $a_\alpha = f(A \setminus \{a_\xi \mid \xi < \alpha\})$ if $A \setminus \{a_\xi \mid \xi < \alpha\} \neq \emptyset$. Then the listing of elements of A in the order $\{a_\alpha\}_\alpha$ is a well ordering of A . For every well ordered set S , there exists an order preserving bijection between S and exactly one ordinal.

Cardinal numbers are a generalization of the natural numbers used to measure the cardinality (size) of sets. Two sets have the same cardinality (the same cardinal number) if and only if there exist a bijection between them. The cardinal of a set X is the least ordinal number such that there exists a not necessarily order preserving bijection between X and that ordinal.

The cardinality of a finite set is the number of elements in that set hence a natural number. The “transfinite” cardinal numbers, often denoted using the Hebrew Aleph symbol $\aleph$ followed by a subscript, describe the sizes of infinite sets. There exists a transfinite sequence of cardinal numbers:

$$0, 1, 2, \dots, n, \dots \aleph_0, \aleph_1, \dots \aleph_\alpha \dots ,$$

³A limit ordinal is a limit point of the class of ordinal numbers, with respect to the order topology defined in the next chapter. The other ordinals are isolated points in the same topology.

⁴Of course, there are sets of ordinals whose supremum is *not* a limit ordinal.

where the transfinite cardinal numbers are indexed by ordinals. The cardinal of $\mathbb{N}$ is the first transfinite cardinal number, Aleph zero $\aleph_0$, which is the smallest ordinal greater than every natural number, ω . A set with cardinality $\aleph_0$ is called *countable*. Cardinals and ordinals are distinct objects: the ordinal $\omega + 1 = \omega \cup \{\omega\}$ is the successor of ω , hence a distinct ordinal, but the cardinal of $\omega + 1$ is still $\aleph_0$. The same holds for $\omega + 2, \omega \cdot 2, \omega^2, \omega^\omega$.

We denote the cardinal of a set X by $|X|$. The power set of the set $P(X) = \{Y \mid Y \subset X\}$ has cardinal $|P(X)| = 2^{|X|}$, and is a cardinal number strictly larger than $|X|$.

Theorem 2.1 (Cantor's theorem). *Let f be a map from a set X to its power set $\mathcal{P}(X)$. Then f is not surjective and $|X| < |\mathcal{P}(X)|$.*

Proof. Consider the set $B = \{x \in X \mid x \notin f(x)\} \subset X$ hence $B \in \mathcal{P}(X)$. Assume f is surjective. Then there exists $y \in X$ such that $f(y) = B$. But by construction $y \in B \Leftrightarrow y \notin f(y) = B$, which is a contradiction. The last assertion follows by noting that $f(x) = \{x\}$ is an injective function from X to $\mathcal{P}(X)$. $\square$

2.2.3 Integer numbers $\mathbb{Z}$ (Ganze Zahlen)

Informally, the integer numbers are constructed as $\mathbb{Z} = \mathbb{N} \cup \{-1, -2, \dots\}$, where $-n$ is the inverse of n with respect to the addition. Integers can be modeled as differences or natural numbers $a - b$.

Construction 2.2 (Integer numbers $\mathbb{Z}$). *The integer numbers can be constructed as the quotient set $\mathbb{Z} = \mathbb{N} \times \mathbb{N} / \sim$ with $\sim$ the equivalence relation $(a, b) \sim (c, d)$ if $a + d = b + c$. The addition, multiplication and total order on the integers are defined as:*

- $[(a, b)] + [(c, d)] = [(a + c, b + d)],$
- $[(a, b)] \times [(c, d)] = [(a \times c + b \times d, a \times d + b \times c)],$
- $[(a, b)] < [(c, d)]$ if $a + d < b + c$.

The natural numbers $\mathbb{N}$ are embedded in $\mathbb{Z}$ by $n \rightarrow [(n, 0)]$.

Note that $[(a, b)] + [(b, a)] = [(a + b, a + b)] = [(0, 0)]$ hence the additive inverse of $[(a, b)]$ is $[(b, a)]$ which we denote $[(b, a)] \equiv -[(a, b)]$. If $a \geq b$ then $a = b + c$ and $[(a, b)] = [(c, 0)] \equiv c$; if on the contrary $a < b$ then $b = a + c$ and $[(a, b)] = -[(b, a)] \equiv -c$ which leads to the usual notation $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$. We observe that $[(b, a)] = [(a, b)] \times [(0, 1)] = -[(a, b)]$, that is $(-1) \times c = -c$ and $(-1) \times (-c) = c$.

Exercise 2.5. *Prove that $\sim$ is an equivalence relation. Prove that the addition of integers is commutative, associative and compatible with the total order. Prove that the multiplication on integers is associative, distributive with respect to the addition and the multiplication by positive elements is compatible with the total order while the multiplication by negative elements reverses the order.*

Definition 2.4 (Group (Gruppe)). *A group $(G, \circ, e)$ is a set G with an internal binary operation $\circ : G \times G \rightarrow G$ such that:*

- (identity element) there exist $e \in G$ such that $e \circ g = g \circ e = g$,
- (associative) $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$, $\forall g_1, g_2, g_3 \in G$,
- (inverse) for any $g \in G$, there exists $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

A group is Abelian, or commutative, if $\forall g_1, g_2 \in G$, $g_1 \circ g_2 = g_2 \circ g_1$.

Definition 2.5 (Axiomatic definition of $\mathbb{Z}$). *$(\mathbb{Z}, +, 0)$ is the only non trivial totally ordered Abelian group whose positive elements (i.e. elements strictly greater than 0) are well-ordered.*

2.2.4 Rational numbers $\mathbb{Q}$ (Rationale Zahlen)

Informally, rational numbers are modeled as ratios p/q with p and q integers.

Construction 2.3 (**Rational numbers $\mathbb{Q}$**). *The rational numbers can be constructed as the quotient set $\mathbb{Q} = \mathbb{Z} \times \mathbb{Z}_{\neq 0} / \sim$ with $\sim$ the equivalence relation $(p_1, q_1) \sim (p_2, q_2)$ if $p_1 \times q_2 = q_1 \times p_2$. Clearly $[(p, q)] = [(a \times p, a \times q)]$ for any $a \in \mathbb{Z}_{\neq 0}$, in particular $[(p, q)] = [(-p, -q)]$. The addition, multiplication and total order on the rational numbers are:*

- $[(p_1, q_1)] + [(p_2, q_2)] = [(p_1 \times q_2 + q_1 \times p_2, q_1 \times q_2)]$,
- $[(p_1, q_1)] \times [(p_2, q_2)] = [(p_1 \times p_2, q_1 \times q_2)]$,
- $[(p_1, q_1)] < [(p_2, q_2)]$ if $q_1, q_2 > 0$ and $p_1 \times q_2 < q_1 \times p_2$.

The integer numbers $\mathbb{Z}$ are embedded in $\mathbb{Q}$ by $p \rightarrow [(p, 1)]$. We use the notation $[(p, q)] \equiv \frac{p}{q}$.

The additive and multiplicative inverses of $x = [(p, q)]$ are $-x = [(-p, q)]$ and $x^{-1} = [(q, p)]$ (where we need to assume $p \neq 0$ for the multiplicative inverse to exist). We can define the ratio of two rational numbers as the multiplication of the first by the multiplicative inverse of the second: $x/y = x \times y^{-1}$.

Theorem 2.2 (**$\mathbb{Q}$ is dense in itself**). *The set $\mathbb{Q}$ is dense in itself, that is for any $x, y \in \mathbb{Q}$ and $x < y$ there exists $z \in \mathbb{Q}$ such that $x < z < y$. More precisely (see Sec. 3) each point in $\mathbb{Q}$ is a limit point of $\mathbb{Q}$ in the standard topology.*

Proof. We have $x = [(p_1, q_1)]$ and $y = [(p_2, q_2)]$ with $q_1, q_2 > 0$ and $p_1 \times q_2 < q_1 \times p_2$. We denote $z = (x + y)/2 = [(p_1 \times q_2 + q_1 \times p_2, 2 \times q_1 \times q_2)]$, where $2 \times q_1 \times q_2 = q_1 \times q_2 + q_1 \times q_2$ as the multiplication is distributive. Taking into account that the multiplication is compatible with the total order on $\mathbb{Z}$ we have:

$$\begin{aligned} q_1 \times (p_1 \times q_2) &< q_1 \times (q_1 \times p_2) \Rightarrow 2 \times q_1 \times q_2 \times p_1 < q_1 \times (p_1 \times q_2 + q_1 \times p_2) \Rightarrow x < z, \\ (p_1 \times q_2) \times q_2 &< (q_1 \times p_2) \times q_2 \Rightarrow (p_1 \times q_2 + q_1 \times p_2) \times q_2 < 2 \times q_1 \times q_2 \times p_2 \Rightarrow z < y. \end{aligned}$$

□

Theorem 2.3 (**$\mathbb{Q}$ is countable**). *There exists a bijection between $\mathbb{N}$ and $\mathbb{Q}$. More generally, a countable union of countable sets is countable, that is there exists a bijection between $\mathbb{N}$ and the set $\bigcup_{n \in \mathbb{N}} S_n$, with $S_n = \{s_{np} | p \in \mathbb{N}\}$.*

Proof. Let us arrange the sets $(S_n = \{s_{np} | p \in \mathbb{N}\})_{n \in \mathbb{N}}$ on a square grid in the plane putting s_{np} at coordinates (n, p) . We list the sets S_n along the horizontal and pile them S_1, S_2 and so on along the vertical. The bijection is obtained by stepping along the axes and along the diagonals:

$$s_{11} \rightarrow s_{12} \rightarrow s_{21} \rightarrow s_{31} \rightarrow s_{22} \rightarrow s_{13} \rightarrow s_{14} \rightarrow s_{23} \rightarrow s_{32} \rightarrow s_{41} \rightarrow \dots$$

It follows that $\mathbb{Z} \subset \bigcup_{a \in \mathbb{N}} \{(a, b) | b \in \mathbb{N}\}$ is countable and $\mathbb{Q} \subset \bigcup_{p \in \mathbb{Z}} \{(p, q) | q \in \mathbb{Z}_{\neq 0}\}$ is also.

□

Exercise 2.6. *Prove that $\sim$ is an equivalence relation. Prove that the addition of rational numbers is commutative, associative and compatible with the total order. Prove that the multiplication of rational numbers is commutative, associative, distributive with respect to the addition and compatible with the total order.*

Exercise 2.7. *Prove that the decimal representation of any rational number either stops or ends up with an infinitely repeated (periodic) sequence of decimals. Consider $p < q$ and recall the algorithm of long division (we denote $(..0)$ a possibly empty sequences of the decimal zero):*

```

p.000000..... | q
p0(..0)         | 0.(.0)c1(..0)c2(..0)c3(..0) ****
-c1 x q
-----
r10(..0)
-c2 x q
-----
r20(..0)
-c3 x q
-----
r30(..0)
*****

```

The rests $r_1, r_2, \dots$ can only take a finite number of values $0, 1, \dots, q-1$. If one of the rests is 0 the decimal representation stops. If not, there is a first repeat of the rests $r_i = r_j$ and the decimal representation ends up in a periodically repeating sequence.

2.2.5 Real numbers $\mathbb{R}$ (Reelle Zahlen)

Perhaps the most useful numbers we will use are the real numbers. Informally the real numbers sit “in between” the rational ones, that is they are numbers that are not rational but can be arbitrarily well approximated by rational ones. An example of a number that is not rational is $\sqrt{2}$.⁵

Real numbers can be constructed using Dedekind cuts. Morally, a Dedekind cut represents the real number b by the set $(-\infty, b) \cap \mathbb{Q}$.

Construction 2.4 (Real numbers $\mathbb{R}$). *The real numbers $\mathbb{R}$ can be constructed as the set of Dedekind cuts of the rational numbers. A Dedekind cut (A, B) of the rational numbers is a partition of $\mathbb{Q}$ into two non empty disjoint sets $\mathbb{Q} = A \cup B$, $A, B \neq \emptyset$, $A \cap B = \emptyset$ such that:*

- (*A is closed downwards*) if $x, y \in \mathbb{Q}$ with $x < y$ and $y \in A$ then $x \in A$,
- (*A does not contain a greatest element*) if $x \in A$ then there exists $y \in A$ such that $x < y$.

As $B = \mathbb{Q} \setminus A$, we can identify the cut (A, B) by its first set only, namely A , which is a nonempty, downwards closed strict subset of $\mathbb{Q}$ (hence bounded from above) but with no greatest element. $\mathbb{Q}$ is embedded in the set of Dedekind cuts by $r \rightarrow A = \{a \in \mathbb{Q} \mid a < r\}$, which does not have a greatest element in $\mathbb{Q}$ since $\mathbb{Q}$ is dense in itself. For example the Dedekind cut corresponding to 0 is $\mathbb{Q}_{<0}$, the set of strictly negative rational numbers. The arithmetic operations and total order on the Dedekind cuts are defined as:

- (*addition*) $A_1 + A_2 = \{a_1 + a_2 \mid a_1 \in A_1 \text{ and } a_2 \in A_2\}$,
- (*subtraction*) $A_1 - A_2 = \{a_1 - b_2 \mid a_1 \in A_1 \text{ and } b_2 \in \mathbb{Q} \setminus A_2\}$,
- (*negation*) $-A_2 = 0 - A_2 = \{a_1 - b_2 \mid a_1 \in \mathbb{Q}_{<0} \text{ and } b_2 \in \mathbb{Q} \setminus A_2\}$,
- (*total order*) $A_1 < A_2$ if A_1 is a strict subset of A_2 , that is $A_1 \subset A_2$ and $A_2 \setminus A_1 \neq \emptyset$, also denoted $A_1 \subsetneq A_2$, and this is a total order as the Dedekind cuts are closed downwards,
- (*multiplication*) we have two cases:
 - for $A_1, A_2 \geq 0$ we define $A_1 \times A_2 = \{a_1 \times a_2 \mid 0 \leq a_1 \in A_1 \text{ and } 0 \leq a_2 \in A_2\} \cup \mathbb{Q}_{<0}$,
 - otherwise we use one of the equalities $A_1 \times A_2 = -(A_1 \times (-A_2)) = -((-A_1) \times A_2) = ((-A_1) \times (-A_2))$,
- (*division*) we again have two cases:

⁵Euclid, Elements (proposition 117, Book X) Alexandria, Ptolemaic Egypt, year ≈ 300 BCE. Assume $\sqrt{2} = p/q$ with p and q co primes. Then $p^2 = 2q^2$ hence p^2 divides by 2 hence p divides by 2. Then $q^2 = 2(p/2)^2$ hence q^2 divides by 2 hence q divides by 2 hence p and q are not co primes.

- if $A_1 \geq 0$ and $A_2 > 0$ then $A_1/A_2 = \{a_1/b_2 \mid 0 \leq a_1 \in A_1 \text{ and } b_2 \in \mathbb{Q} \setminus A_2\} \cup \mathbb{Q}_{<0}$,
- otherwise we use one of the equalities $A_1/A_2 = -(A_1/(-A_2)) = -((-A_1)/A_2) = (-A_1)/(-A_2)$.

The important remark is that any nonempty subset of the real numbers thus constructed which is bounded from above has a *supremum*, that is a least upper bound.

Theorem 2.4 (**Every nonempty, bounded subset of $\mathbb{R}$ has a supremum in $\mathbb{R}$**). *The supremum, that is the least upper bound, of a bounded above subset $S = \{A \mid A \in X \text{ and } A < B\}$ of a totally ordered set X is an element $I \in X$ such that $\forall A \in S, A \leq I$ and $\forall I' < I, \exists A \in S$ with $A \not\leq I'$.⁶ The supremum of the set of Dedekind cuts $S = \{A \mid A \in \mathbb{R} \text{ and } A < B\}$ is the Dedekind cut $I = \bigcup_{A \in S} A$, that is any nonempty subset of $\mathbb{R}$ bounded above has a supremum in $\mathbb{R}$.*

Proof. We consider $S = \{A \mid A \in \mathbb{R} \text{ and } A < B\}$. The set $I = \bigcup_{A \in S} A$ is a Dedekind cut:

- as $A \subset \mathbb{Q}$ then $I \subset \mathbb{Q}$; as $I \supset A \neq \emptyset$ then $I \neq \emptyset$; as $I < B$, then $I \neq \mathbb{Q}$,
- I is closed downwards, if $y \in I$ then $y \in A$ for some A and as A is closed downwards, $x < y \Rightarrow x \in A \subset I$.
- if I has a greatest element x , then $x \in A$ for some A and, as $A \subset I$, x is the greatest element of A , which is impossible.

Moreover, I is an upper bound for S because $\forall A \in S, A \subset I$ hence $A \leq I$. Finally, any $I' < I$ is not an upper bound for S because, if $I' < I$ then $\exists x \in I \setminus I'$. As $x \in I$, then $\exists A \in S$ such that $x \in A$ hence (as $x \notin I'$) $A \not\leq I'$. $\square$

Definition 2.6 (**Algebraic field (Körper)**). *An algebraic field $(K, +, 0, \times, 1)$ is a set such that:*

- $(K, +, 0)$ is an Abelian group, $(K \setminus \{0\}, \times, 1)$ is an Abelian group and $\times$ is distributive with respect to $+$ that is $a \times (b + c) = a \times b + a \times c$.

If moreover K is totally ordered by $\leq$ we say that $\leq$ is compatible with $+$ and $\times$ if for all $x, y, z \in K$, if $x \leq y$ then $x + z \leq y + z$, and if $0 \leq x, y$ then $0 \leq x \times y$. We say that $\leq$ is Dedekind complete if every non-empty subset of K bounded from above has a least upper bound, that is $\forall X \subset K, X \neq \emptyset$, if $B = \{b \in K \mid \forall x \in X, x \leq b\} \neq \emptyset$ then $\exists u \in B$ such that $\forall b \in B, u \leq b$. In this case u is called the supremum of X .

The rational numbers $\mathbb{Q}$ are an algebraic field with a compatible order $\leq$ which is not Dedekind complete: the set $A = \{x \in \mathbb{Q}, x^2 < 2\}$ has upper bounds in $\mathbb{Q}$, e.g. 1.416, but it has no least upper bound as $\sqrt{2} \notin \mathbb{Q}$.

Definition 2.7 (**Axiomatic definition of $\mathbb{R}$**). *$(\mathbb{R}, +, 0, \times, 1)$ is the unique Dedekind complete compatibly totally ordered algebraic field.*

The real numbers can be represented as points on an infinite straight line, called the real line, depicted in Fig. 2.1.

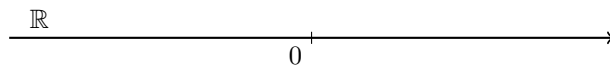


Figure 2.1: The real axis.

Theorem 2.5 (**$\mathbb{R}$ has the Archimedean property (Archimedisches Axiom)**). *The real numbers have the Archimedean property, that is for any $x, y \in \mathbb{R}$ with $0 < x < y$, there exists $n \in \mathbb{N}$ such that $y < nx$.*

Proof. Without loss of generality we can assume $y = 1$. Assume there exists $0 < x < 1$ such that $nx \leq 1$ for any $n \in \mathbb{N}$. Denote S the set of all such x and observe that S is bounded from above by 1. Then S has a supremum, which we denote c . As c is the supremum of S , $n(c/2) \leq 1$ for all $n \in \mathbb{N}$ ⁷ but there exists $p \in \mathbb{N}$ such that $1 < p(2c)$, which is impossible. $\square$

⁶Equivalently, if B is such that $\forall A \in S, A \leq B$ then $I \leq B$.

⁷As there exists $c/2 < x \leq c$ with $x \in S$, and as the order is compatible $n(c/2) < nx \leq 1$

Theorem 2.6 ($\mathbb{Q}$ is dense in $\mathbb{R}$). For any $x, y \in \mathbb{R}$ with $x < y$ there exists a rational number $q \in \mathbb{Q}$ such that $x < q < y$.

Proof. We denote $\tilde{x}$ the cut associated to the real number x , that is $\tilde{x}$ is a subset of $\mathbb{Q}$ that is closed downwards and is bounded from above but has no greatest element. The Dedekind cut of $q \in \mathbb{Q}$ is denoted $\tilde{q} = \{p \in \mathbb{Q} \mid p < q\}$. We denote $\tilde{x} \subsetneq \tilde{y}$ the relation of strict inclusion $\tilde{x} \subset \tilde{y}$ and $\tilde{y} \setminus \tilde{x} \neq \emptyset$. As $x < y$ we have $\tilde{x} \subsetneq \tilde{y}$ and:

If $x \in \mathbb{Q}$ then $x \in \tilde{y}$. This follows if $\exists s \in \tilde{y}, x < s$. Assume the contrary $\forall s \in \tilde{y}, s \leq x$. But then the inequality needs to be strict, otherwise, as $\tilde{y}$ has no greatest element, $\exists s' \in \tilde{y}, s' > s = x$. Finally, if $\forall s \in \tilde{y}, s < x$ then $\tilde{y} \subset \tilde{x}$ and $y \leq x$, contradicting the hypothesis $x < y$.

There exist $q \in \mathbb{Q}, q \in \tilde{y} \setminus \tilde{x}$ and if $x \in \mathbb{Q}$ we can always chose $q > x$. The first part follows from $\tilde{x} \subsetneq \tilde{y}$. If $x \in \mathbb{Q}$ then by the previous item $x \in \tilde{y}$ and as $\tilde{y}$ has no greatest element, there exists $q \in \mathbb{Q}, x < q$ and $q \in \tilde{y}$.

We have $\tilde{x} \subsetneq \tilde{q}$. Any $r \in \tilde{x}, r \neq q$. As $\mathbb{Q}$ is totally ordered then either $r < q$ or $q < r$. If $q < r$, as $\tilde{x}$ is downwards closed, $q \in \tilde{x}$ contradicting $q \in \tilde{y} \setminus \tilde{x}$. We conclude that $r < q$ hence $r \in \tilde{q}$ and $\tilde{x} \subset \tilde{q}$. If $x \notin \mathbb{Q}$ then $x \neq q$ and the inclusion is strict. If $x \in \mathbb{Q}$, as we can chose $q > x$, the inclusion is again strict.

We have $\tilde{q} \subsetneq \tilde{y}$. This is trivial as $q \in \tilde{y}$ and $\tilde{y}$ is closed downwards.

□

Theorem 2.7 ($\mathbb{R}$ is uncountable). There is no surjective function $f : \mathbb{N} \rightarrow (0, 1) \subset \mathbb{R}$, that is the set of real numbers is not countable.

Proof. Every $a \in (0, 1)$ admits a decimal representation as $0.a_1a_2\dots a_n\dots$ obtained by choosing inductively a_k to be the largest digit 0, 1, 2 up to 9 such that $\sum_{p=1}^k a_p 10^{-p} \leq x$ and $x = \sup_{k \geq 1} \{\sum_{p=1}^k a_p 10^{-p}\}$. This definition avoids the “problem of trailing 9s”, i.e. $0.49999\dots$ is not the decimal representation of any real number, as the decimal representation of $1/2$ is $0.50000\dots$. The price to pay is that not any list of digits is an admissible list of decimals of a real number: lists ending in trailing 9s, $a_k = 9$ for $k > n$ are not admissible and need to be converted in lists of decimals ending in trailing 0s, as in $0.a_1\dots a_n9999 \rightarrow 0.a_1\dots(a_n+1)0000\dots$. The gain is that there is a one to one correspondence between real numbers and lists of decimals that is $x = y$ if and only if all their decimals are equal.

Consider now any function $f : \mathbb{N} \rightarrow (0, 1)$ and consider for all $i \in \mathbb{N}$ the decimal representation $f(i) = 0.i_1i_2\dots i_n\dots$. We now build a real number $x \in (0, 1)$ such that $\forall i, f(i) \neq x$. We take x with decimal representation $x = 0.x_1x_2\dots x_n\dots$ and we ensure that $x \neq f(i)$ by choosing the i 'th decimal x_i of x to be different from the i 'th decimal i_i of $f(i)$ (if we make a table with rows the decimals of $f(i)$, we modify the diagonal decimals and collect them to form x). For instance we can chose $x_i = i_i + 1$ if $0 \leq i_i \leq 7$ and $x_i = 0$ for $i_i = 8, 9$. In particular x can not contain any decimal 9, so it is an admissible decimal representation of a real number. As $x \neq f(i)$ for any i , $x \notin f(\mathbb{N})$ and f is not surjective. □

It can be shown that $|\mathbb{R}| = 2^{\aleph_0}$, that is the cardinal of the real numbers is equal to the cardinal of the power set of the natural numbers.⁸ The continuum hypothesis states that there is no other cardinal number in between $\aleph_0$ and $2^{\aleph_0}$, that is $2^{\aleph_0} = \aleph_1$. This statement is undecidable using the standard Zermelo–Fraenkel set theory axioms.

Some simple functions of real numbers are:

- *Inverses.* For $x \in \mathbb{R}$ the additive inverse of x is $-x$ and for $x \in \mathbb{R}_{\neq 0}$ the multiplicative inverse is $1/x = x^{-1}$. The multiplicative inverses of 0 is undetermined.
- *Absolute value.* For $x \in \mathbb{R}$ the absolute value $|x|$ of x is x if $x \geq 0$ and $-x$ otherwise.

⁸Think of the decimal representation in base 2.

- *Integer powers, polynomials.* For $x \in \mathbb{R}_{\neq 0}$, $x^0 = 1$ and $x^{n+1} = x \times x^n$; $0^1 = 0$ but 0^0 is undermined. Polynomials are linear combinations of integer powers. A polynomial of degree n is $P(x) = a_n \times x^n + a_{n-1} \times x^{n-1} + \dots + a_0$, with $a_0, \dots, a_n \in \mathbb{R}$ and $a_n \neq 0$.
- *n -roots.* For $x \in \mathbb{R}_{\geq 0}$, the root of order n for $n \in \mathbb{Z}_{\neq 0}$ is:
 - $n > 0$, it is the Dedekind cut $x^{\frac{1}{n}} = \{0 \leq q \in \mathbb{Q} \mid q^n < x\} \cup \mathbb{Q}_{<0}$.
 - for $n < 0$ it is defined as $x^{\frac{1}{n}} = (x^{-\frac{1}{n}})^{-1}$.

For $n \neq 0$, $0^{\frac{1}{n}} = 1$ and $0^{\frac{1}{0}}$ is undetermined.

- *Rational powers.* For $x \in \mathbb{R}_{\geq 0}$, and $p, q \in \mathbb{Z}$, $x^{\frac{p}{q}} = (x^{\frac{1}{q}})^p$

Exercise 2.8. *Prove that the arithmetic operations defined for Dedekind cuts are commutative, associative, compatible with the ordering and that the multiplication is distributive with respect to the addition. Prove that if $0 < x, y$ then $0 < x \times y$. Prove that this implies that if $a > 0$ and $x < y$ then $a \times x < a \times y$.*

Exercise 2.9. *Prove that $A = \{a \in \mathbb{Q} \mid a^2 = a \times a < 2 \text{ or } a < 0\}$ is a valid Dedekind cut which does not correspond to a rational number. What real number does this cut represent?*

Exercise 2.10. *Prove, using Dedekind cuts, that the sequence $0.999\dots$ of trailing 9 is the real number 1. We denote the truncated sequence of n decimals $9, 0.(9)_n = 1 - 10^{-n} < 1$, and $0.999\dots$ the supremum of this subset of real numbers. The Dedekind cut of $0.999\dots$ consists in the rational numbers r such that there exists some n with $r < 0.(9)_n < 1$, hence r belongs to the cut of 1. Consider now any element in the cut of 1, $\frac{a}{b} < 1$ hence $0 < a < b$. Then $1 - \frac{a}{b} = \frac{b-a}{b} \geq \frac{1}{b} > \frac{1}{10^6}$ hence $0.(9)_6 > \frac{a}{b}$, therefore $\frac{a}{b}$ belongs to the cut of $0.(9)_6$ and consequently to the one of $0.999\dots$*

2.2.6 Complex numbers $\mathbb{C}$ (Komplexe Zahlen)

The last class of numbers we will be using are the complex numbers.

Construction 2.5 (**Complex numbers $\mathbb{C}$**). *The complex numbers $\mathbb{C}$ are the Cartesian product $\mathbb{R} \times \mathbb{R}$ endowed with the operations:*

- (addition) $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$,
- (multiplication) $(x_1, y_1) \times (x_2, y_2) = (x_1 \times x_2 - y_1 \times y_2, x_1 \times y_2 + y_1 \times x_2)$,

and form an algebraic field. The real numbers $\mathbb{R}$ are embedded in $\mathbb{C}$ by $x \rightarrow (x, 0)$

In order to simplify the notation, whenever no confusion can arise, we suppress the symbol $\times$ for multiplication.

As $\mathbb{C} = \mathbb{R} \times \mathbb{R}$, the complex numbers are represented as an infinite plane, called the complex plane. The real numbers $\mathbb{R}$ are embedded into $\mathbb{C}$ by the mapping $x \rightarrow (x, 0)$. In particular $(1, 0) \equiv 1$ and denoting $(0, 1) \equiv i$ we can write any complex number as:

$$z = (x, y) = (x, 0) + (0, 1) \times (y, 0) = x + i \times y, \quad i^2 = i \times i = -1.$$

Let $z = x + iy$ be a complex number. We call x the *real part* of z which we denote $\text{Re}(z)$, and y the *imaginary part* of z , which we denote by $\text{Im}(z)$. The *complex conjugated* of z , denoted $\bar{z}$, is obtained by keeping the real part fixed, but taking the additive inverse of the imaginary part $z = x + iy \rightarrow \bar{z} = x - iy$.

The complex numbers can be represented as points in a plane, called the complex plane spanned by two orthogonal real axes: $z = x + iy = (x, y)$ represents a point with coordinates x on the horizontal and y on the vertical, as depicted in Fig. 2.2.

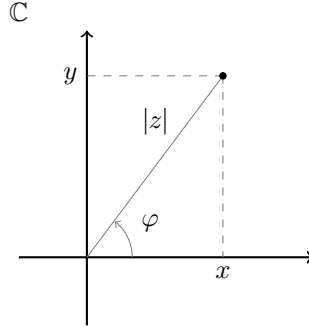


Figure 2.2: The complex plane.

The *absolute value* (or modulus) of z is the non negative real number⁹ $|z| = (x^2 + y^2)^{1/2} = (z\bar{z})^{1/2}$, $|z| > 0$ for $z \neq 0$. It represents the distance from the origin $0 = (0, 0)$ and $z = (x, y)$. The *argument* of z is the angle φ between the horizontal axis and the line $(0, 0), (x, y)$ and:

$$x = |z| \cos \varphi, \quad y = |z| \sin \varphi \quad \Rightarrow \quad z = |z|(\cos \varphi + \imath \sin \varphi).$$

A priori the angle φ belongs to the interval $(-\pi, \pi)$. This is the *principal determination* of the argument, denoted $\text{Arg}(z)$. However, the argument is not uniquely defined: it is a periodic function with period 2π , that is:

$$z = |z|(\cos \theta + \imath \sin \theta), \quad \theta = \text{Arg}(z) + 2k\pi, \quad k \in \mathbb{Z}.$$

Like for real numbers, simple functions of complete numbers are:

- *Inverses.* For $z \in \mathbb{C}$ the additive inverse of z is $-z$ and for $z \in \mathbb{C}_{\neq 0}$ the multiplicative inverse is $\frac{1}{z} = \frac{\bar{z}}{z \times \bar{z}}$. The multiplicative inverses of 0 is undetermined.
- *Integer powers, polynomials.* For $z \in \mathbb{R}_{\neq 0}$, $z^0 = 1$ and $z^{n+1} = z \times z^n$; $0^1 = 0$ but 0^0 is undermined. Polynomials are linear combinations of integer powers $P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0$, with $a_0, \dots, a_n \in \mathbb{R}$ and $a_n \neq 0$.
- *n-roots.* For $z = |z|(\cos \varphi + \imath \sin \varphi) \in \mathbb{C}_{\neq 0}$ and for any $n \in \mathbb{N}_{\neq 0}$ we have n distinct roots of order n of z :

$$\xi^n = z \Rightarrow \xi \in \{\xi_0, \dots, \xi_{n-1}\}, \quad \xi_k = |z|^{1/n} \left(\cos\left(\frac{\varphi + 2k\pi}{n}\right) + \imath \sin\left(\frac{\varphi + 2k\pi}{n}\right) \right).$$

- *Rational powers.* For $x \in \mathbb{C}_{\neq 0}$, and $p \in \mathbb{Z}$ and $q \in \mathbb{Z}_+$, $z^{\frac{p}{q}} = (z^{\frac{1}{q}})^p$

The most important property of the complex numbers, whose proof is postponed to Chapter 10, states that $\mathbb{C}$ is an algebraically closed field, that is any polynomial with complex coefficients admits a root in $\mathbb{C}$. In detail, for any polynomial with complex coefficients $P(z)$ there exist $r \in \mathbb{C}$ such that $P(r) = 0$. Denoting r a root of the polynomial $P(z)$, we note that $z^n - r^n = (z - r) \sum_{q=0}^{n-1} z^q r^{n-q}$ hence $P(z) = P(z) - P(r) = (z - r)Q(z)$ for Q some polynomial of degree $n - 1$, and iterating we get $P(z) = a_n (z - r_1) \times \dots (z - r_n)$ with $r_1, \dots, r_n$ the n (not necessarily distinct) roots of $P(z)$ and a_n the coefficient of the monomial of highest degree.

Exercise 2.11. Prove that the addition and multiplication of complex numbers are commutative and associative and the multiplication is distributive with respect to the addition, that is the complex numbers are an unordered algebraic field.

Exercise 2.12. Prove that $\imath^2 = \imath \times \imath = -1$

⁹Strictly speaking the non negative determination of the square root is constructed as a Dedekind cut.

Exercise 2.13. *Prove using the trigonometric circle that $\cos(\varphi_1 + \varphi_2) = \cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2)$ and $\sin(\varphi_1 + \varphi_2) = \sin(\varphi_1)\cos(\varphi_2) + \cos(\varphi_1)\sin(\varphi_2)$. Conclude that for $z_1 = |z_1|(\cos \varphi_1 + \imath \sin \varphi_1)$ and $z_2 = |z_2|(\cos \varphi_2 + \imath \sin \varphi_2)$ then $z_1 z_2 = |z_1||z_2|(\cos(\varphi_1 + \varphi_2) + \imath \sin(\varphi_1 + \varphi_2))$. Show that for any n -root of z , $(\xi_k)^n = z$. Prove that any complex number such that $\xi^n = z$ is one of the n -roots of z .*

Exercise 2.14. *What is the image of the line $\operatorname{Re}(z) = \text{const.}$ under the function $f(z) = 1/z$?*

Chapter 3

Topological spaces

Topology formalizes the notion of being close in the absence of a way to measure distances.

3.1 Topology

Definition 3.1 (**Topology (Topologie)**). A topological space is a set X endowed with a topology $\mathcal{T}$. A topology $\mathcal{T}$ on X is a set of subsets of X , $\mathcal{T} = \{U \mid U \subset X\} \subset \mathcal{P}(X)$ such that:

- $\emptyset, X \in \mathcal{T}$,
- (closed under arbitrary union) for any index set¹ A if $(U_\alpha)_{\alpha \in A} \in \mathcal{T}$ then $\bigcup_{\alpha \in A} U_\alpha \in \mathcal{T}$ (or, stated fancier, $\mathcal{A} \subset \mathcal{T} \Rightarrow \bigcup \mathcal{A} \subset \mathcal{T}$),
- (closed under finite intersection) for any finite set I , if $(U_i)_{i \in I} \in \mathcal{T}$ then $\bigcap_{i \in I} U_i \in \mathcal{T}$.

A subset Y of X is endowed with the subspace (or induced) topology $\mathcal{T}_Y = \{U \cap Y \mid U \in \mathcal{T}\}$.

The topologies of a set X form a poset, in fact a lattice, with the inclusion. For $\mathcal{T}_1$ and $\mathcal{T}_2$ two topologies on X , if $\mathcal{T}_1 \subset \mathcal{T}_2$ we say that $\mathcal{T}_2$ is finer than $\mathcal{T}_1$ (resp. $\mathcal{T}_1$ is coarser than $\mathcal{T}_2$).

Any set X has two trivial topologies: the *discrete topology* consisting in all the subsets of X (the finest topology on X) and the *indiscrete topology* consisting in only the empty set and the set itself (the coarsest topology in X). Topologies that sit in between these two are usually more interesting.

Definition 3.2 (**Open sets and neighborhoods (Offene Menge, Umgebungen)**). The elements U in a topology $U \in \mathcal{T}$ are called open sets. We reserve U as a notation for open sets. A set is called closed if its complement is open.² Sets that are both open and closed (for instance $\emptyset$ and X) are sometimes called clopen.

For any $x \in X$, a neighborhood of x in X is an set $V \subset X$ which includes an open set containing x , that is $x \in U \subset V$, with $U \in \mathcal{T}$. Without loss of generality we restrict to open neighborhoods, that is a neighborhood of a point is any open set containing that point.

A topology is called Hausdorff if for any two distinct points $x_1, x_2 \in X$, $x_1 \neq x_2$ one can find two disjoint neighborhoods, one containing x_1 and the other one x_2 , that is $x_1 \in V_1$, $x_2 \in V_2$ and $V_1 \cap V_2 = \emptyset$.

In a Hausdorff topology the points are not indistinguishably close to each other. The discrete topology is Hausdorff, while the indiscrete one is not.

¹Strictly speaking, any ordinal A , finite, countable or uncountable.

²As $\bigcap_\alpha (X \setminus U_\alpha) = X \setminus (\bigcup_\alpha U_\alpha)$ and $\bigcup_i (X \setminus U_i) = X \setminus (\bigcap_i U_i)$, arbitrary intersection and finite unions of closed sets are closed.

A *subbase* in a topology is a subset of open sets $U \in \mathcal{T}$ such that any other open set is obtained by taking arbitrary unions of finite intersections of them. A *base* is a subset of open sets such that any other open set can be obtained by taking only unions of them. For example, every totally ordered set X can be endowed with a topology, called the order topology, generated by the subbase $\{x|x < b\}$ and $\{x|a < x\}$ for all $a, b \in X$. A base in the order topology consists in these sets together with $(a, b) = \{x|a < x < b\}$. The ordinal numbers $\lambda = [0, \lambda)$, $\lambda + 1 = \lambda \cup \{\lambda\}$ are totally ordered and the space of ordinal numbers can be endowed with the order topology.

Definition 3.3 (**Continuous and open functions**). A map (function) $f : X \rightarrow Y$ from a topological space $(X, \mathcal{T}_1)$ to a topological space $(Y, \mathcal{T}_2)$ is called

- open if for all $U \in \mathcal{T}_1$, $f(U) \in \mathcal{T}_2$,
- continuous if for all $V \in \mathcal{T}_2$, $f^{-1}(V) \in \mathcal{T}_1$.

If $\mathcal{T}_1$ and $\mathcal{T}_2$ are topologies on X with $\mathcal{T}_1 \subset \mathcal{T}_2$ then the map $\text{id}_X : (X, \mathcal{T}_1) \rightarrow (X, \mathcal{T}_2)$ is open and $\text{id}_X : (X, \mathcal{T}_2) \rightarrow (X, \mathcal{T}_1)$ is continuous.

A topological space is called *connected* if it can not be divided into two disjoint nonempty open sets, $X \neq X_1 \cup X_2$ with $X_1 \cap X_2 = \emptyset$. For $x, y \in X$, a *path* from x to y is a continuous function $f : [0, 1] \rightarrow X$ with $f(0) = x$ and $f(1) = y$. A topological space is path connected if any two points in it are connected by a path.

3.1.1 Compact sets (Kompakter Menge)

A class of sets that will play an important role below is that of compact sets.

Definition 3.4 (**Compact sets**). A set $K \subset X$ is called compact if from any open cover of K , $\bigcup_{\alpha \in A} U_\alpha \supset K$ with $U_\alpha \in \mathcal{T}$, one can select a finite open cover of K , $\bigcup_{\alpha \in I \subset A} U_\alpha \supset K$ for some finite set I .

Theorem 3.1 (**Properties of compact sets**). We have:

- let K be a compact subset of X and $U \subset K$ an open set. Then $K \setminus U$ is compact.

Proof. Let any open cover $(U_\alpha)_{\alpha \in A}$ of $K \setminus U$, then $U, (U_\alpha)_{\alpha \in A}$ is an open cover of K and it possesses a finite subcover $(U_i)_{i \in I}$ (where one of U_i can be U). As U contains no point in $K \setminus U$, $(U_i)_{i \in I} \setminus U$ is a finite cover of $K \setminus U$. $\square$

- if K is a compact subset of $(X, \mathcal{T})$ then K is compact in the induced topology $(K, \mathcal{T}_K)$.

Proof. For any cover of K by open sets in the induced topology $(K \cap U_\alpha)_{\alpha \in A}$ with $U_\alpha \in \mathcal{T}$, then $(U_\alpha)_{\alpha \in A}$ is an open cover of K in $\mathcal{T}$ and admits a finite subcover $(U_i)_{i \in I}$ which is such that $(K \cap U_i)_{i \in I}$ covers K . $\square$

- the image of a compact set under a continuous function is a compact set.

Proof. Let $f : X \rightarrow Y$ a continuous function and consider $K \in X$ compact. For any open cover of $f(K)$, $f(K) \subset \bigcup_{\alpha \in A} V_\alpha \subset f(X)$, the sets $(f^{-1}(V_\alpha))_{\alpha \in A}$ are open (as f is continuous) and form a cover of K , $\bigcup_{\alpha \in A} f^{-1}(V_\alpha) \supset K$. Hence one can select a finite cover of K , $(f^{-1}(V_i))_{i \in I}$ leading to a finite open cover of $(V_i)_{i \in I}$ of $f(K)$. $\square$

3.1.2 Topological boundary, closure and interior (Topologischer Rand)

Definition 3.5 (**Topological boundary, closure and interior**). Consider a subset of a topological space $Y \subset X$. For any point in $x \in X$, we denote $U(x)$ an open neighborhood of x , that is an open set $U(x) \in \mathcal{T}$ containing x , $x \in U(x)$. The point x is:

- an accumulation (or limit) point of Y if every neighborhood of x contains at least one other point in Y , that is $\forall U(x), (U(x) \cap Y) \setminus \{x\} \neq \emptyset$. A limit point of Y does not necessarily belong to Y .
- an interior point of Y if it has a neighborhood entirely contained in Y , that is $\exists U(x) \subset Y$.
- a isolated point of Y if it belongs to Y but is not a limit point of Y , that is $x \in Y$ and $\exists U(x)$ such that $(U(x) \cap Y) \setminus \{x\} = \emptyset$.³

We stress that the isolated points of a set, while by definition are not limit points, can be interior points: if X is endowed with the discrete topology and $Y \subset X$, every point of $y \in Y$ is at the same time an isolated and an interior point of Y as $\{y\}$ is open.

Exercise 3.1. The interior Y° of a subset Y of a topological space X is the union of all the open sets contained in Y , that is $Y^\circ = \bigcup(\mathcal{P}(Y) \cap \mathcal{T})$. Prove that:

- $x \in Y^\circ$ if and only if x is an interior point of Y . If x is interior to Y then $\exists U(x) \subset Y$ hence $U(x) \in \mathcal{P}(Y) \cap \mathcal{T}$ thus $U(x) \subset Y^\circ$. Conversely, any point in Y° must belong to at least one of the open sets in $\mathcal{P}(Y) \cap \mathcal{T}$, all of which are entirely included in Y .
- the interior $Y^\circ \subset Y$ is an open set in $\mathcal{T}$. By definition.
- a set U is open if and only if $U = U^\circ$. If $U = U^\circ$ then U is open (as U° is open). Conversely, consider an open U ; first, $U^\circ \subset U$; second, as U is open then both $U \in \mathcal{P}(U)$ and $U \in \mathcal{T}$, hence $U \subset \bigcup(\mathcal{P}(U) \cap \mathcal{T}) = U^\circ$.

The closure $\bar{Y}$ of Y is the union of Y with all its limit points. The boundary of Y is the set difference between its closure and its interior, $\partial Y = \bar{Y} \setminus Y^\circ$. We have that:

- $x \in \bar{Y}$ if and only if $\forall U(x), U(x) \cap Y \neq \emptyset$. If $x \in \bar{Y}$ then either $x \in Y$ or x is a limit point of Y and in both cases $\forall U(x), U(x) \cap Y \neq \emptyset$. Conversely, if $x \in Y$ then $x \in \bar{Y}$ and if $x \notin Y$ then $\forall U(x), x \notin U(x) \cap Y$ hence $U(x) \cap Y \setminus \{x\} = U(x) \cap Y \neq \emptyset$ and x is a limit point of Y .
- the complement of the closure is the interior of the complement $X \setminus \bar{Y} = (X \setminus Y)^\circ$. Indeed, if $x \in X \setminus \bar{Y}$ then on the one hand $\exists U(x)$ such that $U(x) \cap Y \setminus \{x\} = \emptyset$ and on the other $x \notin Y$ (as $Y \subset \bar{Y}$) hence $U(x) \cap Y = \emptyset$. It follows $U(x) \subset X \setminus Y$, that is $x \in (X \setminus Y)^\circ$. Conversely, if $x \in (X \setminus Y)^\circ$ then $\exists U(x), U(x) \subset X \setminus Y$, which implies $U(x) \cap Y = \emptyset$ hence x can not belong to $\bar{Y}$.
- C is closed if and only if $C = \bar{C}$. By the previous item, $\bar{C}$ is closed hence $C = \bar{C}$ implies C is closed. Conversely, if C is closed then $X \setminus C$ is open and $X \setminus C = (X \setminus C)^\circ = X \setminus \bar{C}$ by the previous item.
- X splits into the disjoint partition $X = \bar{Y} \cup (X \setminus \bar{Y}) = Y^\circ \cup \partial Y \cup (X \setminus Y)^\circ$. From the definition of $\bar{Y}$ and item 2.
- the boundaries of a set and of its complement are equal $\partial Y = \partial(X \setminus Y)$ and $\partial Y = \bar{Y} \cap \overline{X \setminus Y}$. By applying the previous item to Y and $Y \rightarrow X \setminus Y$.
- ∂Y is a closed set. As its complement is open.
- ∂Y consists of all the points $x \in X$ such that $\forall U(x), U(x) \cap Y \neq \emptyset$ and $U(x) \cap (X \setminus Y) \neq \emptyset$. Indeed, any point $x \in \partial Y$ is either a point in Y or a limit point of Y (or both) and any $U(x)$ will contain a point in Y ; the same goes for $X \setminus Y$. Conversely, by the first item above, $x \in \bar{Y}$ and $x \in \overline{X \setminus Y}$ hence $x \in \bar{Y} \cap \overline{X \setminus Y} = \partial Y$.

³ $\{x\}$ is an open set in the induced topology $\mathcal{T}_Y$.

- for any set Y , $\partial Y \supset \partial\partial Y$ and $\partial Y = \partial\partial Y$ if and only if ∂Y has no interior points. *Indeed, as ∂Y is closed, $\overline{\partial Y} = \partial Y$ and $\partial\partial Y = \partial Y \setminus (\partial Y)^\circ$.*
- the boundary ∂V of any open set V has empty interior. The same holds for closed sets. *As V is open, $V = V^\circ$ hence $V \cap \partial V = \emptyset$ as the boundary and the interior of any set are disjoint. Take any point $x \in \partial V$; for any $U(x)$, $U(x) \cap V \neq \emptyset$ hence $U(x)$ can not be a subset of ∂V . The same result follows for close sets as they have the same boundary as their complements, which are open.*
- every closed set with empty interior is the boundary of its complement. *Indeed, as $B = \bar{B}$ and $B^\circ = \emptyset$ then on the one hand $\partial B = \bar{B} = B$ and on the other $\partial B = \partial(X \setminus B) = \partial(X \setminus B)^\circ$.*
- we stress that:
 - * the isolated points of Y are not necessarily limit points of $X \setminus Y$, as in the discrete topology.
 - * for any Y the interior of Y° is Y° as Y° is open and the closure of $\bar{Y}$ is $\bar{Y}$ as $\bar{Y}$ is closed.
 - * a set Y is dense in X if $\bar{Y} = X$. If U is dense in X and open then $\partial U = X \setminus U$.
 - * in general $\partial\bar{Y} \neq \partial Y \neq \partial Y^\circ$ hence in general $\bar{Y}^\circ \neq \bar{Y}$ and $(\bar{Y})^\circ \neq Y^\circ$! The isolated points of Y belong to $\bar{Y}$ but are not accumulation points of Y . Such points usually do not belong to Y° (unless the topology is discrete and they belong as one element sets to the topology), hence usually they belong to ∂Y but not to ∂Y° . If U is open dense in X , then $\partial(\bar{U}) = \partial X = \emptyset$ as X is both closed and open but $\partial U = X \setminus U \neq \emptyset$.

3.2 Metric spaces and the standard topology

In many cases of interest, the topology of a space comes from a notion of metric. Formally, a topological space X is called *metrizable* if it is homeomorphic⁴ to a metric space. A metric space is a space in which one can measure distances.

We use ϵ as a notation for a (possibly small) strictly positive real number.

Definition 3.6 (**Metric space (Metrischer Raum)**). A metric space X is a set endowed with a distance function $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ (the non negative reals) such that:

- $d(x, x) = 0$,
- (symmetry and positivity) if $x \neq y$ then $d(x, y) = d(y, x) > 0$,
- (triangle inequality) for all x, y, z , $d(x, y) \leq d(x, z) + d(z, y)$.

The most interesting property of a distance is the triangle inequality. Morally it means that the distance between x and y is measured along the shortest path, and passing through another point z will always represent a detour which lengthens the distance.

A distance function of a metric space induces a topology.

Definition 3.7 (**The standard topology**). The standard topology on a metric space (X, d) is the topology generated by the subbase of all open balls $B_r(x) = \{y \mid d(x, y) < r\} \subset X$, for all $x \in X$, $r \in \mathbb{R}_{>0}$.

The closure of a ball in the standard topology is $\bar{B}_r(x) = \{y \mid d(x, y) \leq r\}$. Indeed, if y is an accumulation point of $B_r(x)$, then $\forall \epsilon, \exists z \in B_\epsilon(y) \cap B_r(x)$ hence $d(x, y) \leq d(x, z) + d(z, y) < r + \epsilon$. Assuming $d(x, y) = r' > r$, $\epsilon < r' - r$ leads to a contradiction. The boundary of the ball is $\partial B_r(x) = \{y \mid d(x, y) = r\}$.

⁴That is mapped by a *bicontinuous bijection*, a continuous bijection whose inverse is also continuous.

Theorem 3.2 (In the standard topology, every open set contains a ball around any of its points). For any open set U in the standard topology on some metric space (X, d) , and for any $y \in U$, there exists a small enough $\epsilon > 0$ such that $B_\epsilon(y) \subset U$. The converse (if $\forall y \in U, \exists B_\epsilon(y) \subset U$ then U is open) is trivially true.

A topological space X is called first countable if $\forall x \in X$, there exists a countable set of open neighborhoods of x , $\{U_n(x) | n \in \mathbb{N}\}$ such that for any open set U , $x \in U \Rightarrow U_i(x) \subset U(x)$ for some $i \in \mathbb{N}$. All metric spaces are first countable.

Proof. We first show that this property holds for any ball, then for an arbitrary union of balls and finally for a finite intersection of balls.

Consider a ball $B_r(x)$ centered at x for some x and some other point $y \in B_r(x)$. As $d(x, y) < r$, there exists ϵ such that $0 < \epsilon < r - d(x, y)$. By the triangle inequality, $\forall z \in B_\epsilon(y)$, $d(z, x) \leq d(z, y) + d(y, x) \leq \epsilon + d(y, x) < r$ therefore $B_\epsilon(y) \subset B_r(x)$.

Let $y \in \bigcup_{\alpha \in A} B_{r_\alpha}(x_\alpha)$ and denote $B_{r_1}(x_1)$ (one of) the ball(s) to which y belongs; then there exists ϵ such that $B_\epsilon(y) \subset B_{r_1}(x_1)$ and the same ball is contained in the union $B_\epsilon(y) \subset \bigcup_{\alpha \in A} B_{r_\alpha}(x_\alpha)$. Finally, if $y \in B_{\epsilon_1}(y) \subset B_{r_1}(x_1)$ and $y \in B_{\epsilon_2}(y) \subset B_{r_2}(x_2)$, then $y \in B_{\min(\epsilon_1, \epsilon_2)}(y) \subset B_{\epsilon_1}(y) \cap B_{\epsilon_2}(y)$. The lemma follows by noting that any open set can be obtained by taking arbitrary unions and finite intersection balls.

The countable set $B_{\frac{1}{n}}(x)$ fulfills the requirement of first countability at x . □

From this lemma it follows that the balls form in fact a base, not only a subbase in the standard topology, that is one can obtain any open set U by taking unions of balls, one for each of its points, which are fully contained in U .

An important result is that metrizable topologies are Hausdorff.

Theorem 3.3 (The standard topology is Hausdorff). The standard topology on a metric space is Hausdorff; the same holds for all metrizable topologies.

Proof. Consider any two points $x \neq y$, hence such that $d(x, y) > 0$. Then the balls $B_{d(x, y)/3}(x)$ and $B_{d(x, y)/3}(y)$ are totally disjoint. Indeed, assume there exists $z \in B_{d(x, y)/3}(x) \cap B_{d(x, y)/3}(y)$. Then $d(z, x), d(z, y) < d(x, y)/3$, and from the triangle inequality $d(x, y) \leq d(x, z) + d(z, y) < 2d(x, y)/3$ which is impossible. □

Sorgenfrey, or lower limit topology (Sorgenfrey-Gerade)

Although in many cases topologies are metrizable, there exist also non metrizable topologies. An example is provided for instance by the Sorgenfrey topology of $\mathbb{R}$ generated by the basis of all half open intervals $[a, b)$. This is a non metrizable topology on $\mathbb{R}$.

Exercise 3.2. Let us consider the real numbers $\mathbb{R}$. Prove that:

$$d : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, \quad d(x, y) = |x - y| = \begin{cases} x - y, & x - y \geq 0 \\ -x + y, & x - y < 0 \end{cases}.$$

is a distance function on $\mathbb{R}$. Any $x, y, z \in \mathbb{R}$ with $x < y$, we have $z < x < y$ or $x < z < y$ or $x < y < z$. Consider now the complex numbers $\mathbb{C}$. Prove that $d : \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{R}_{\geq 0}$, $d(z_1, z_2) = |z_1 - z_2|$ is a distance function. This follows from $|z_1 + z_2| \leq |z_1| + |z_2|$ which follows from $(x_1 + x_2)^2 + (y_1 + y_2)^2 \leq x_1^2 + y_1^2 + x_2^2 + y_2^2 + 2\sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2)} \Leftrightarrow x_1^2 x_2^2 + y_1^2 y_2^2 + 2x_1 x_2 y_1 y_2 \leq x_1^2 x_2^2 + x_1^2 y_2^2 + y_1^2 x_2^2 + y_1^2 y_2^2$ which is true as $0 \leq (x_1 y_2 - x_2 y_1)^2$.

3.2.1 The $\mathbb{R}^D$ metric space

The D dimensional Euclidean space is the set $\mathbb{R}^D = \{(x^1, \dots, x^D) | x^\mu \in \mathbb{R}\}$. We denote $0 = (0, \dots, 0)$ the origin in $\mathbb{R}^D$, that is the point with coordinates $x^\mu = 0$, $\mu = 1 \dots D$. In most cases $\mathbb{R}^D$ is regarded as a topological space endowed with the standard topology with respect to a distance function which we now discuss. Starting from the next chapter, unless otherwise specified, we will always assume that the $\mathbb{R}^D$ is endowed with this topology.

3. TOPOLOGICAL SPACES

The space $\mathbb{R}^D$ is a metric space. In fact several standard distance functions are usually defined on it:

$$d_p(x, y) = \|x - y\|_p = \left(\sum_{\mu=1}^D |x^\mu - y^\mu|^p \right)^{1/p}, \quad p \in [1, \infty),$$

and $d_\infty(x, y) = \|x - y\|_\infty = \max_{\mu=1, \dots, D} |x^\mu - y^\mu|$. Observe that for $D = 1$ all these distance functions are equal and reproduce the absolute value of the difference.

Lemma 3.1. *The distance functions d_p are equivalent on $\mathbb{R}^D$, that is for any $p, q \geq 1$ there exist two constants c_{pq} and C_{pq} such that*

$$c_{pq}d_p(x, y) \leq d_q(x, y) \leq C_{pq}d_p(x, y).$$

Proof. It is enough to prove that d_p is equivalent to d_∞ . We have $\max_\mu |x^\mu - y^\mu| \leq (\sum_{\mu=1}^D |x^\mu - y^\mu|^p)^{1/p} \leq D^{1/p} \max_\mu |x^\mu - y^\mu|$. $\square$

This lemma implies that the standard topology on the D dimensional Euclidean space $(\mathbb{R}^D, d_p)$ is independent on which distance function d_p is chosen (i.e. a set is open with respect to one of these distance functions if and only if it is open with respect to any other). For simplicity the common choice is to use d_2 :

$$d_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{\mu=1}^D |x^\mu - y^\mu|^2} \equiv \|x - y\| = d(x, y),$$

that is $\|\cdot\|$ and $d(\cdot, \cdot)$ with no index p are understood to correspond to $p = 2$. This is just the familiar Euclidean distance between points in dimension n . Note that, from the topological point of view, $\mathbb{R}^2$ endowed with the Euclidean $\|x - y\|$ is identical to the complex plane $\mathbb{C}$ with distance given by the absolute value $|z_1 - z_2|$ on complex numbers.

Theorem 3.4 (In the standard topology $\overline{\mathbb{Q}} = \mathbb{R}$). *In the standard topology the closure $\overline{\mathbb{Q}}$ of $\mathbb{Q}$ is $\mathbb{R}$.*

Proof. The balls in the standard topology of $\mathbb{R}$ (which incidentally is also the order topology) with distance function $d(x, y) = |x - y|$ have the form $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$. From Theorem 2.6 we have that $\mathbb{Q}$ is dense in $\mathbb{R}$, that is for any $x, y \in \mathbb{R}$ with $x < y$ there exists a rational number $q \in \mathbb{Q}$ such that $x < q < y$. Consider any $x \in \mathbb{R}$. From Theorem 3.2 any neighborhood of x contains a ball $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$ and as $\mathbb{Q} \cap B_\epsilon(x) \setminus \{x\} = \mathbb{Q} \cap ((x - \epsilon, x) \cup (x, x + \epsilon)) \neq \emptyset$ we conclude that x is a limit point of $\mathbb{Q}$. $\square$

Exercise 3.3. *Give an example of an open dense subset of $\mathbb{R}$ in the standard topology. $\mathbb{R} \setminus \{0\}$ or more generally $\mathbb{R} \setminus M$ for any finite set M .*

Exercise 3.4 (Hölder's inequality (Hölder-Ungleichung)). *Let $p, q \in \mathbb{R}_{\geq 1}$ such that $1/p + 1/q = 1$. Then, for any $(a^\mu, b^\mu)_{\mu=1, \dots, D} \in \mathbb{R}_{\geq 0}$ we have the inequality:*

$$\sum_{\mu=1}^D a^\mu b^\mu \leq \left[\sum_{\mu=1}^D (a^\mu)^p \right]^{1/p} \left[\sum_{\mu=1}^D (b^\mu)^q \right]^{1/q}.$$

We first note that, for any $p, q > 1$ with $1/p + 1/q = 1$ and any $x, y > 0$, as the logarithm (see Section 7.2.1 for a reminder on the definition and properties of the logarithm) is a concave function we have:

$$\ln \left(\frac{1}{p} x^p + \left(1 - \frac{1}{p}\right) y^q \right) \geq \frac{1}{p} \ln(x^p) + \frac{1}{q} \ln(y^q) = \ln(xy),$$

hence, as the logarithm is strictly increasing, $xy \leq \frac{x^p}{p} + \frac{y^q}{q}$, which is also known as the Young inequality. Thus: $\sum_{\mu=1}^D x^\mu y^\mu \leq \frac{\sum_{\mu=1}^D (x^\mu)^p}{p} + \frac{\sum_{\mu=1}^D (y^\mu)^q}{q}$, and the lemma follows by setting

$$x^\mu = \frac{a^\mu}{(\sum_{\mu=1}^D (a^\mu)^p)^{1/p}} \text{ and } y^\mu = \frac{b^\mu}{(\sum_{\mu=1}^D (b^\mu)^q)^{1/q}}.$$

Prove the Minkowski inequality $\|x + y\|_p \leq \|x\|_p + \|y\|_p$. Conclude that that all the functions $d_p(x, y) = \|x - y\|_p$ are distance functions. We have, with $x + y = (x^1 + y^1, \dots, x^D + y^D)$

$$\|x + y\|_p^p = \sum_{\mu} |x^\mu + y^\mu|^p = \sum_{\mu} |x^\mu + y^\mu| |x^\mu + y^\mu|^{p-1} \leq \sum_{\mu} (|x^\mu| + |y^\mu|) |x^\mu + y^\mu|^{p-1} = \sum_{\mu} |x^\mu| |x^\mu + y^\mu|^{p-1} + \sum_{\mu} |y^\mu| |x^\mu + y^\mu|^{p-1} \leq (\|x\|_p + \|y\|_p) (\|x + y\|_p)^{p-1} = (\|x\|_p + \|y\|_p) \|x + y\|_p^{p-1},$$

where we used Hölder's inequality. We conclude by observing that $\|(|x+y|^{p-1})\|_{1-\frac{1}{p}} = \left(\sum_{\mu} (|x^{\mu} + y^{\mu}|^{p-1})^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} = \|x+y\|_p^{p-1}$.

Exercise 3.5 (One point compactification (Alexandroff-Kompaktifizierung)). *To any topological space $(X, \mathcal{T})$ one can add a point, usually denoted $\{\infty\}$, and build the space $X^* = X \cup \{\infty\}$ endowed with the topology $\mathcal{T}^*$ consisting in all the open sets $U \in \mathcal{T}$ together with the sets $V = (X \setminus K) \cup \{\infty\}$ for K compact in X .*

Prove that $(X^*, \mathcal{T}^*)$ is compact. Any open cover of X^* will need to cover $\{\infty\}$, hence will contain a set of the form $(X \setminus K) \cup \{\infty\}$ whose complement in X is compact. What are the open sets in $\mathbb{N} \cup \{\infty\}$? In the discrete topology any subset of $\mathbb{N}$ and a set is compact if and only if it is finite. The neighborhoods of $\{\infty\}$ are $\{\infty\} \cup P$ for P any infinite subset of $\mathbb{N}$. Give a base in the topology of $\mathbb{R} \cup \{\infty\}$. Besides the balls $B_{\epsilon}(x) = (x - \epsilon, x + \epsilon)$ we add $B_{\epsilon}(\infty) = (-\infty, -1/\epsilon) \cup \{\infty\} \cup (1/\epsilon, \infty)$. Prove that $\mathbb{R} \cup \{\infty\}$ is homeomorphic to a circle. $\mathbb{R} \cup \{\infty\}$ is mapped on a circle by the inverse stereographic projection: draw a circle tangent at the south pole to $0 \in \mathbb{R}$ and map any point $x \in \mathbb{R}$ to the intersection of the line (x, N) (where N is the north pole of the circle) with the circle then identify $\{\infty\}$ with the north pole N on the circle. Prove that the one point compactification of $\mathbb{C}$ is homeomorphic to a 2 dimensional sphere, the Riemann sphere.

Chapter 4

Sequences

In this chapter we discuss sequences. One can consider sequences in generic topological spaces, metric spaces in $\mathbb{R}^D$ endowed with the standard topology. Sequences in non metrizable topological spaces are more complicated and will not be discussed here. However, it is important to discuss as much as possible general metric spaces: one does in fact encounter in physics such spaces, for instance in quantum mechanics where a central role is played by spaces of functions. Such spaces are metric spaces in their own right: one does talk about distances between functions, sequences of functions, and so on.

4.1 Sequences in metric spaces

From now on X is always understood to designate a metric space (X, d) endowed with the standard topology. Unless otherwise specified, when discussing the space $\mathbb{R}^D$, we always consider it endowed with some distance function $d(x, y) = \|x - y\|$.

A sequence in a metric space X is a list of elements $x_0, x_1, x_2, \dots, x_n, \dots$ in X in which repetitions are allowed and the order matters. Formally sequences are defined as follows.

Definition 4.1 (**Sequence (Folgen)**). A sequence in a metric space (X, d) is a function with domain $\mathbb{N}$ and co domain X , $x : \mathbb{N} \rightarrow X$. We denote $x(n) = x_n$ and we identify a sequence by its image $(x_0, x_1, x_2, \dots) \equiv (x_n)_{n \in \mathbb{N}}$. In order to simplify the notation we sometimes denote a sequence simply (x_n) .

A subsequence of a sequence is obtained by deleting some elements without changing the order of the remaining ones, that is a subsequence of $(x_n)_{n \in \mathbb{N}}$ is the sequence $(x_{n_k})_{k \in \mathbb{N}}$ with n_k a strictly increasing sequence of natural numbers $n_k > n_{k'}$ if $k > k'$. In particular, for any subsequence, $n_k \geq k$.

A sequence converges if x_n approaches some fixed value when $n \rightarrow \infty$.

Definition 4.2 (**Convergence (Konvergenz)**). A sequence $(x_n)_{n \in \mathbb{N}}$ in (X, d) is said to converge to x if for any $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that $n > N_\epsilon \Rightarrow d(x, x_n) < \epsilon$. Equivalently, for any ball $B_\epsilon(x)$, there exists N_ϵ such that $n > N_\epsilon \Rightarrow x_n \in B_\epsilon(x)$. We denote this by $\lim_{n \rightarrow \infty} x_n = x$, $\lim x_n = x$ or simply $x_n \rightarrow x$. The point x is called the limit of (x_n) .

This definition can be extended with zero effort to general topological spaces.

Lemma 4.1. A sequence in (X, d) converges if and only if all of its subsequences converge. A convergent sequence and its subsequences converge to the same limit.

Proof. As any sequence is one of its own subsequences, one implication is trivial. For the other, assume (x_n) converges to x , that is for any ϵ , $\exists N_\epsilon$ such that $k > N_\epsilon \Rightarrow d(x_k, x) < \epsilon$. As $n_k \geq k$ for any subsequence, $d(x_{n_k}, x) < \epsilon$ also, and the lemma follows. $\square$

Limits of sequences and limit points of subspaces of metric spaces are intimately related.

Theorem 4.1 (**Limit points in metric spaces are limits of sequences**). *A point x is an accumulation (limit) point of a subset Y of a metric space (X, d) if and only if there exists a sequence (y_n) with $y_n \in Y$, $y_n \neq x$ such that $\lim y_n = x$.*

Proof. If $\lim y_n = x$ then every ball $B_\epsilon(x)$ will contain all the elements $(y_n)_{n>N_\epsilon}$ which belong to Y and are different from x . Conversely, Take the sequence of balls of radius $1/n$ centered at x . As x is an accumulation point $B_{\frac{1}{n}}(x) \cap Y \setminus \{x\} \neq \emptyset$. From the axiom of choice (!) we can chose an element y_n per set $B_{\frac{1}{n}}(x) \cap Y \setminus \{x\} \neq \emptyset$ which yields a sequence such that $\forall N > 0, n > N \Rightarrow d(y_n, x) < 1/N$ and we conclude from the Archimedean property of $\mathbb{R}$. $\square$

In metric spaces, limit of sequences go through continuous functions. For X and Y topological spaces, a function $f : X \rightarrow Y$ is called *sequentially continuous* if for all $(x_n) \subset X$ such that $x_n \rightarrow x \in X$, the sequence $(f(x_n)) \subset Y$ is a sequence which converges to $f(x)$, $f(x_n) \rightarrow f(x)$. It can be shown that if X and Y are arbitrary topological spaces, a continuous function is sequentially continuous. The reciprocal does not hold for general topological spaces but holds if X is first countable.

Theorem 4.2 (**In metric spaces continuous is sequentially continuous**). *For X and Y metric spaces, a function $f : X \rightarrow Y$ is sequentially continuous if and only if it is continuous.*

Proof. Let f a continuous function. For any ϵ , let $B_\epsilon(f(x))$ be the open ball of radius ϵ around $f(x)$. Then $x \in f^{-1}(B_\epsilon(f(x)))$ (note that $f^{-1}(B_\epsilon(f(x)))$ includes other points in X if f is not injective) thus $f^{-1}(B_\epsilon(f(x)))$ is an open set in X (due to continuity) that includes x . It follows that there exists some δ such that $B_\delta(x) \subset f^{-1}(B_\epsilon(f(x)))$. As $x_n \rightarrow x$, there exists some N such that $x_n \in B_\delta(x)$ for $n > N$ and then $n > N \Rightarrow f(x_n) \in B_\epsilon(f(x))$.

Suppose now that f is sequentially continuous, but not continuous at some point $x \in X$, which means there exists an open neighborhood V of $f(x)$ such that $f^{-1}(V)$ is not an open neighborhood of x . As any open neighborhood of x contains some ball $B_\delta(x)$, it follows that for any δ , $B_\delta(x) \setminus f^{-1}(V) \neq \emptyset$. Choosing a sequence (x_n) with $x_n \in B_{\frac{1}{n}}(x) \setminus f^{-1}(V)$ we have that $x_n \rightarrow x$ but $f(x_n) \notin V$. But then $f(x_n) \not\rightarrow f(x)$, because V is open and contains a ball around $f(x)$. Hence f is not sequentially continuous, leading to a contradiction. $\square$

Theorem 4.3 (**Any distance function is continuous**). *Any distance function is continuous in either of its arguments.*

Proof. It is enough to prove sequential continuity. Fix a point z and consider $d(x, z)$ as a function of its first argument. For any x and x_n we have $d(x_n, z) \leq d(x_n, x) + d(x, z)$ and $-d(x_n, x) \leq d(x_n, z) - d(x, z)$ hence $|d(x_n, z) - d(x, z)| \leq d(x_n, x)$. We conclude that for any sequence $x_n \rightarrow x \Rightarrow d(x_n, z) \rightarrow d(x, z)$. $\square$

4.1.1 Cauchy sequences (Cauchy Folgen)

A Cauchy sequence is a sequence such that its terms become increasingly closer for $n \rightarrow \infty$.

Definition 4.3 (**Cauchy sequence**). *A sequence $(x_n)_{n \in \mathbb{N}}$ in (X, d) is called a Cauchy sequence if for any $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that $m, n > N_\epsilon \Rightarrow d(x_m, x_n) < \epsilon$.*

Every convergent sequence is Cauchy. The converse (reciprocal) of this statement is *not true*. For instance the real sequence defined recursively by $x_{n+1} = \frac{1}{2}(x_n + \frac{2}{x_n})$, $x_0 = 1$ is a Cauchy sequence in $\mathbb{Q}$ but does not converge in $\mathbb{Q}$.

Lemma 4.2. *A Cauchy sequence in (X, d) converges if and only if it has a convergent subsequence.*

Proof. One implication is trivial. Assume now we have a Cauchy sequence $(x_n)_{n \in \mathbb{N}}$ with a convergent subsequence $(x_{n_k})_{k \in \mathbb{N}}$. For any ϵ chose N_ϵ such that for $k, l > N_\epsilon$ both $d(x_k, x_l) < \epsilon/2$ and $d(x_{n_k}, x) < \epsilon/2$. as $n_k \geq k$, it follows that $d(x_k, x) \leq d(x_k, x_{n_k}) + d(x_{n_k}, x) < \epsilon$. $\square$

Exercise 4.1. *Prove that any convergent sequence is Cauchy. Use $d(x_m, x_n) < d(x_m, x) + d(x, x_n)$. Prove that $x_{n+1} = \frac{1}{2}(x_n + \frac{2}{x_n})$, $x_0 = 1$ is a Cauchy sequence in $\mathbb{Q}$. The fact that the sequence stays in $\mathbb{Q}$ is trivial. Then denote $x_n = \sqrt{2} - \epsilon_n$, the recursion becomes $\sqrt{2} - \epsilon_{n+1} = \frac{2+2\epsilon_n^2-2\sqrt{2}\epsilon_n}{2\sqrt{2}-2\epsilon_n} \Rightarrow \epsilon_{n+1} = \frac{4-2\sqrt{2}\epsilon_n-4-\epsilon_n^2+2\sqrt{2}\epsilon_n}{2\sqrt{2}-2\epsilon_n} = -\frac{\epsilon_n^2}{2(\sqrt{2}-\epsilon_n)}$. By induction we show that $|\epsilon_{n+1}| < \epsilon_n^2$ and $\epsilon_0 < 1/2$. Thus $|x_{n+p} - x_n| \leq |\epsilon_{n+1}| + |\epsilon_{n+2}| \dots |\epsilon_{n+p+1}| \leq \epsilon_n^2 + \epsilon_n^4 + \dots + \epsilon_n^{2^n} < \frac{\epsilon_n^2}{1-\epsilon_n^2} < \frac{1}{1-1/4} (\frac{1}{2})^n = 2^{2-n}$ which becomes arbitrary small with n .*

Theorem 4.4 (**In metric spaces compact is sequentially compact**). *A subset $K \subset X$ of a topological space X is called sequentially compact if every sequence in K admits a subsequence that converges to some point in K . If (X, d) is a metric space then $K \subset X$ is compact if and only if it is sequentially compact.*

Proof. We have:

Sequentially compact $\Rightarrow$ compact. Assume K is sequentially compact and take $\{U_\alpha\}_\alpha$ an open cover of K .

First, there exists $\delta > 0$ (independent of x) such that $\forall x \in K, B_\delta(x) \subset U_\alpha$ for some α . Assume no such δ exists. Then for any n there exists $x_n \in K$ such that $\forall \alpha, B_{1/n}(x_n) \not\subset U_\alpha$. Chose a convergent subsequence of $(x_n)_n$, say $x_{n_i} \rightarrow x \in K$. Now, $x \in U_{\alpha_1}$ for some α_1 and there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset U_{\alpha_1}$. Pick n_i such that $d(x, x_{n_i}) < \frac{\epsilon}{2}$ and $\frac{1}{n_i} < \frac{\epsilon}{2}$, thus $B_{1/n_i}(x_{n_i}) \subset U_{\alpha_1}$ which is a contradiction.

Second, there are finitely many balls of radius δ centered at points in K whose union covers K . Suppose this is false. Pick $x_1 \in K$ and $B_\delta(x_1) \not\subset K$, pick $x_2 \in K \setminus B_\delta(x_1)$ and $B_\delta(x_1) \cup B_\delta(x_2) \not\subset K$, iterate. We build a sequence x_n such that $j \neq j' \Rightarrow d(x_j, x_{j'}) > \delta$, hence x_n does not contain any Cauchy subsequence hence can not have any convergent subsequence.

As each $B_\delta(x)$ is a subset of some U_α , the finite cover of K by balls of radius δ selects a finite cover of K by sets U_α .

Compact $\Rightarrow$ sequentially compact. Assume K is compact and take a sequence $(x_n)_n$ in K . If x_n has a constant subsequence we conclude. If not, we can always pick a subsequence consisting in pairwise distinct elements which we label $(x_p)_{p \in \mathbb{N}}$.

Assume that $\forall y \in K, \exists \delta_y > 0$ such that $B_{\delta_y}(y)$ contains only finitely many terms in the sequence $(x_p)_p$. As K is compact, the open cover $\bigcup_y B_{\delta_y}(y) \supset K$ has a finite subcover $\bigcup_{i=1}^l B_{\delta_{y_i}}(y_i) \supset K$. But this finite union can only contains finitely many of the x_p 's, a contradiction.

Then $\exists y \in K$ such that $\forall \delta > 0, B_\delta(y)$ contains infinitely many terms in the sequence (x_p) . Then the subsequence obtained by picking one element $x_{p_k} \in \{x_p, p \in \mathbb{N}\} \cap B_{1/k}(y)$ for each $k \in \mathbb{N}$ is a convergent subsequence of x_p .

□

Definition 4.4 (**Cauchy complete spaces**). *A metric space such that every Cauchy sequence converges is called complete or Cauchy complete.*

Not all metric spaces are Cauchy complete: for instance $\mathbb{Q}$ is not, as can be seen by considering the Cauchy sequence consisting in the finite truncations of the decimal representation of $\sqrt{2}$ which does not converge in $\mathbb{Q}$. Starting from a Cauchy incomplete space X one can build a Cauchy complete one (called the completion of X) by adding to X the limits of all the Cauchy sequences¹ of elements in X . This procedure provides an alternative construction of the real numbers starting from the rational ones.

4.1.2 Banach fixed-point theorem (Fixpunktsatz von Banach)

A function (map) $T : X \rightarrow X$ is called a contracting map if there exists $q \in [0, 1)$ such that $d(T(x), T(y)) \leq q d(x, y)$ for all $x, y \in X$. The constant q is called a *Lipschitz constant* for T .

¹Strictly speaking one must take equivalence classes of Cauchy sequences with the same limit.

Theorem 4.5 (**Banach fixed-point theorem**). *A contracting map T on a (non empty) Cauchy complete metric space X admits a fixed point x^* , that is $T(x^*) = x^*$. Furthermore, for any x_0 the recursively defined sequence $x_{n+1} = T(x_n)$ for $n \geq 0$ is such that $d(x^*, x_n) < \frac{q^n}{1-q} d(x_1, x_0)$ hence converges and $\lim x_n = x^*$.*

The fixed point exist under the weaker condition $d(T(x), T(y)) \leq d(x, y)$ if X is compact.

Proof. Starting from x_0 , we build $x_{n+1} = T(x_n)$ for $n \geq 0$. Then $d(x_{n+1}, x_n) \leq q^n d(x_1, x_0)$ hence for $p \geq 1$:

$$d(x_{n+p}, x_n) \leq \sum_{i=0}^{p-1} d(x_{n+1+i}, x_{n+i}) \leq \sum_{i=0}^{p-1} q^{n+i} d(x_1, x_0) = \frac{1-q^p}{1-q} q^n d(x_1, x_0) < \frac{1}{1-q} q^n d(x_1, x_0),$$

hence, as $q^n \rightarrow 0$, x_n is Cauchy and, as X is complete, it converges $\lim x_n = x^* \in X$. Taking $p \rightarrow \infty$ above we find $d(x^*, x_n) \leq q^n/(1-q)d(x_1, x_0)$. We have that $\forall \epsilon$, there exists N_ϵ such that $d(x^*, x_k) < \epsilon/2$ for all $k > N_\epsilon$. Picking any $k > N_\epsilon$ we have:

$$d(x^*, T(x^*)) \leq d(x^*, x_{k+1}) + d(T(x_k), T(x^*)) < \frac{\epsilon}{2} + q \frac{\epsilon}{2} < \epsilon,$$

hence $x^* = T(x^*)$. Note that the fixed point is unique as T is contracting. $\square$

Note that in this proof we used the fact that $q^n \rightarrow 0$ for $0 \leq q < 1$. We will see below that this is a direct consequence of the monotone convergence theorem.

Exercise 4.2. *We will prove later that $\mathbb{R}$ is Cauchy complete. Consider the function $T(x) = \frac{x}{2} + \frac{1}{x}$. Prove that $T([1, \infty)) \subset [1, \infty)$. $\frac{x}{2} + \frac{1}{x} \geq 1 \Leftrightarrow x^2 + 1 \geq 2x \Leftrightarrow (x-1)^2 \geq 0$. Prove T is a contraction on $[1, \infty)$. $T(x) - T(y) = (x-y)(\frac{1}{2} - \frac{1}{xy})$ and for all $x, y > 1$, we have $0 < (xy)^{-1} < 1 \Rightarrow -2^{-1} < (xy)^{-1} - 2^{-1} < 2^{-1} \Rightarrow |2^{-1} - (xy)^{-1}| < 2^{-1}$. Find the fixed point of this contraction. $x^* = \sqrt{2}$.*

Propose an algorithm to compute $\sqrt{2}$ as the limit of a rational sequence. $x_{n+1} = T(x_n)$, $x_0 = 1$.

4.2 Sequences in $\mathbb{R}$ and $\mathbb{R}^D$

A particularly rich case is the one of real sequence $(x_i)_{i \in \mathbb{N}}$, $x_i \in \mathbb{R}$.

4.2.1 Homogeneous linear recurrence (Homogener lineare Differenzengleichung)

Often one encounters recursively defined sequences in which a term is fixed by a finite number of its predecessors. $x_{n+q} = F(x_n, x_{n+1}, \dots, x_{n+q-1})$. If F is continuous, then the limit of a recursively defined sequence, if it exists, is a solution of the equation $x^* = F(x^*, \dots, x^*)$.

Let us consider real or complex valued sequences. If a recursively defined sequence is not convergent, one is interested in finding a closed formula for the general term x_n (this is the equivalent in the world of sequences of solving differential equations). A particular case or recurrence that can always be solved is that of a homogeneous linear recursion with constant coefficients:²

$$x_{n+q} = \sum_{p=0}^{q-1} c_p x_{n+p}, \quad x_0, x_1, x_2, \dots, x_{q-1},$$

where c_p are fixed. In order to start the recursive computation of all the x_n s, the first q elements in the sequence $x_0, x_1, x_2, \dots, x_{q-1}$ need to be given. Linear recursions can always be translated in *finite difference equations* using the finite difference operator:

$$(\Delta x)_n = x_{n+1} - x_n, \quad (\Delta^{k+1} x)_n = \left(\Delta(\Delta^k x) \right)_n, \quad x_{n+p} = \sum_{k=0}^p \binom{p}{k} (\Delta^k x)_n.$$

²One can consider more general recursions with non constant, n dependent, coefficients $c_p(n)$ or inhomogeneous obtained by adding a known sequence b_n to the right hand side of the recurrence relation. If the homogeneous recurrence is solved, the inhomogeneous case can also be solved by the so called "variation of constants" method. We will discuss this later in the context of differential equations.

This provides a straightforward parallel between such recursions and differential equations which we will discuss later.

We note that if $(y_n^{(1)})$ and $(y_n^{(2)})$ are two sequences respecting a homogeneous recursion, so is any linear combination of them. In order to solve the recursion it is then enough to:

- identify q linearly independent solutions $y_n^{(i)}$ for $i = 1, \dots, q$,
- posit the general solution:

$$x_n = \sum_{i=1}^q d_i y_n^{(i)},$$

- fix the constants d_i by solving the boundary conditions $x_j = \sum_{i=1}^q d_i y_j^{(i)}$ for $j = 0, \dots, q-1$. Indeed, as long as $y_n^{(i)}$ are linearly independent, this linear system of equations will admit a unique solution.

The q linearly independent solutions $y_n^{(i)}$ are found by solving the *characteristic equation* of the recurrence. Posit that $x_n = t^n$ with some fixed t , then the homogeneous linear recursion becomes the characteristic equation:

$$P(t) = t^q - \sum_{p=0}^{q-1} c_p t^p = 0.$$

Being a polynomial equation of degree q , this equation has q not necessarily distinct complex roots. Clearly any root r of this polynomial is such that $x_n = r^n$ will respect the recurrence, hence if all the roots r_i are distinct the general solution of the recursion relation is:

$$x_n = \sum_{i=1}^q d_i r_i^n.$$

In particular, as advertised, the linear system of equations determining the coefficients d_i admits a unique solution as:

$$\det(y_j^{(i)}) = \begin{vmatrix} 1 & \dots & 1 \\ r_1 & \dots & r_q \\ \vdots & & \vdots \\ r_1^{q-1} & \dots & r_q^{q-1} \end{vmatrix} = \prod_{1 \leq i < j < q} (r_i - r_j) \neq 0.$$

If the root r has multiplicity ν , then the linearly independent solutions corresponding to it are $r^n, n r^n, n^2 r^n$ and so on up to $n^{\nu-1} r^n$.

Exercise 4.3. Solve for the Fibonacci numbers $x_0 = 0, x_1 = 1, x_{n+2} = x_{n+1} + x_n$. The characteristic equation is $t^2 = t + 1$ with roots $t = \frac{1 \pm \sqrt{1+4}}{2}$ hence $x_n = d_+ \left(\frac{1+\sqrt{5}}{2}\right)^n + d_- \left(\frac{1-\sqrt{5}}{2}\right)^n$ with $d_+ + d_- = 0, (d_+ - d_-)\sqrt{5} = 2$ hence $x_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

Exercise 4.4. Solve and compare the general term of the sequences $a_0 = 0$ and $a_1 = 1$ and recursive definitions:

$$a_{n+2} - 2a_{n+1} + a_n + \omega^2 a_n = 0, \quad a_{n+2} - 2a_{n+1} + a_n - \lambda^2 a_n = 0$$

Exercise 4.5. First order non-homogeneous linear recursions can always be solved. Given two known sequences (f_n) with $f_n \neq 0$ and (g_n) prove that the general term in the recursive sequence $a_0, a_{n+1} = f_n a_n + g_n$ is:

$$a_n = \left(\prod_{k=0}^{n-1} f_k \right) \left(a_0 + \sum_{m=0}^{n-1} \frac{g_m}{\prod_{k=0}^m f_k} \right).$$

We verify that $a_1 = f_0(a_0 + \frac{g_0}{f_0})$ and $f_n a_n + g_n = f_n \left(\prod_{k=0}^{n-1} f_k \right) \left(a_0 + \sum_{m=0}^{n-1} \frac{g_m}{\prod_{k=0}^m f_k} \right) + g_n = \left(\prod_{k=0}^n f_k \right) \left(a_0 + \sum_{m=0}^{n-1} \frac{g_m}{\prod_{k=0}^m f_k} + \frac{g_n}{\prod_{k=0}^n f_k} \right).$

4.2.2 Limits of real sequences, $\limsup$ and $\liminf$

The distance on $\mathbb{R}$ is the absolute value $d(x, y) = |x - y|$. Convergent real sequence can be manipulated much like numbers: they can be added, subtracted, multiplied, divided ordered and so on.

Theorem 4.6 (Properties of limits). *If (a_n) and (b_n) are two convergent sequences in $\mathbb{R}$, $\lim a_n = a$ and $\lim b_n = b$, then:*

- $\lim |a_n| = |a|$.
- $\lim(a_n \pm b_n) = a \pm b$.
- $\forall c \in \mathbb{R}, \lim(ca_n) = ca$.
- $\lim(a_n b_n) = ab$.
- if $b \neq 0$ then $\lim \frac{a_n}{b_n} = \frac{a}{b}$.
- if $\forall n > N, a_n \leq b_n$ then $a \leq b$. This implies the so called sandwich theorem (Einschnürungssatz): if c_n is a sequence such that $a_n \leq c_n \leq b_n$ and $\lim a_n = \lim b_n = L$ then $\lim c_n = L$.

Proof. The proof is straightforward. For instance, as $|a| \leq |a - a_n| + |a_n|$ and $|a_n| \leq |a_n - a| + |a|$ we conclude that $||a| - |a_n|| \leq |a - a_n|$ and $\lim |a_n| = |a|$; also:

$$\left| \frac{a_n}{b_n} - \frac{a}{b} \right| = \left| \frac{a_n}{b_n} - \frac{a_n}{b} + \frac{a_n}{b} - \frac{a}{b} \right| \leq |a_n| \frac{|b - b_n|}{|b_n| |b|} + \frac{|a_n - a|}{|b|},$$

and for n large enough $|b| - \epsilon < |b_n|$, $|a_n| < |a| + \epsilon$.

For the last point, assuming that $b < a$ the ball $B_{\frac{|b-a|}{3}}(a)$ contains all the a_n , $n > N_1$ and the ball $B_{\frac{|b-a|}{3}}(b)$ contains all the b_n , $n > N_2$, hence for $N > \max(N_1, N_2)$, $b_n \leq a_n$ contradicting the hypothesis. $\square$

Any real sequence $(x_n)_n \in \mathbb{R}$ admits an asymptotic upper (lower) bound:

$$\limsup x_n = \lim_{n \rightarrow \infty} \left(\sup_{m \geq n} x_m \right), \quad \liminf x_n = \lim_{n \rightarrow \infty} \left(\inf_{m \geq n} x_m \right),$$

where we allow $\limsup x_n = \pm\infty$, $\liminf x_n = \pm\infty$. We have that:

- we denote $\limsup x_n = \infty$ if for any $y \in \mathbb{R}$, there exist infinitely many terms in (x_n) such that $x_n > y$, that is $\forall y, \forall N, \exists n > N$ such that $x_n > y$.
- if $\limsup x_n = L < \infty$, then for any $\epsilon > 0$ only a finite number of terms in (x_n) are such that $x_n > L + \epsilon$, but infinitely many terms are such that $x_n > L - \epsilon$, that is $\forall \epsilon, \exists N, \forall n > N, L + \epsilon \geq x_n$ but $\forall \epsilon, \forall N, \exists n > N$ such that $x_n > L - \epsilon$.

Clearly, the sequence (x_n) converges to x if and only if $\limsup x_n = \liminf x_n = x \neq \pm\infty$.³

4.2.3 Monotone convergence theorem (Satz von der monotonen Konvergenz)

An interesting case of sequences in $\mathbb{R}$ are the monotonic, bounded ones.

Definition 4.5 (Monotonicity). *A sequence in $\mathbb{R}$ is called:*

- *monotonically increasing* if $a_n \geq a_p$ for $n > p$.
- *monotonically decreasing* if $a_n \leq a_p$ for $n > p$.

³One sometimes encounters the notation $\lim x_n = \infty$ if $\forall y, \exists N_y$ such that $n > N_y \Rightarrow x_n > y$. This is however a dangerous notations because such a sequence x_n does not converge.

- bounded if $|a_n| \leq M$ form some fixed M .

Theorem 4.7 (**Monotone convergence theorem**). A sequence (x_n) , $x_n \in \mathbb{R}$ that is monotonic and bounded is convergent. If x_n is monotonically increasing (resp. decreasing) then it converges to its supremum (resp. infimum).

Proof. Let (x_n) monotonically increasing and bounded. $\{x_n\}$ is a bounded subset of $\mathbb{R}$ hence has a supremum (least upper bound) which we denote x . Now for all ϵ , there exists N such that $x_N > x - \epsilon$ (otherwise $x - \epsilon$ would be an upper bound on (x_n)) and $\forall n > N$, $x - \epsilon < x_N \leq x_n \leq x$. $\square$

Exercise 4.6 (Limits of classical sequences.). Prove Bernoulli's inequality $(1+x)^n \geq 1+nx$ for x real, $x > -1$ and $n \in \mathbb{N}$.

By induction, $(1+x)^{n+1} = (1+x)^n(1+x) \geq (1+nx)(1+x) = 1+(n+1)x+nx^2 \geq 1+(n+1)x$. Prove that:

- $\forall s \in \mathbb{Q}_{>0} \lim_{n \rightarrow \infty} \frac{1}{n^s} = 0$. Chose $N_\epsilon > 1/(\epsilon^{1/s})$.
- $\forall a > 0 \lim_{n \rightarrow \infty} a^{1/n} = 1$. It is enough to prove this for $a > 1$. We have $a = (1+a^{1/n}-1)^n \geq 1+n(a^{1/n}-1) \Rightarrow 0 < a^{1/n}-1 \leq \frac{a-1}{n}$ hence by the sandwich criterion $\lim_{n \rightarrow \infty} a^{1/n} = 1$.
- $\lim_{n \rightarrow \infty} n^{1/n} = 1$. Note that if $n > 1$ then $n^{1/q} > 1$ for any q . Now $n^{1/2} = [1+(n^{1/(2n)}-1)]^n \geq 1+n(n^{1/(2n)}-1) \Rightarrow 0 < n^{1/(2n)}-1 < n^{-1/2}-n^{-1} \rightarrow 0$ and $\lim_{n \rightarrow \infty} (n^{1/(2n)}) = 1$.
- $0 < q < 1$, $\lim_{n \rightarrow \infty} q^n = 0$. As $0 < q^{n+1} < q^n$ this is a monotonic bounded sequence hence converges at its infimum. Assume this infimum is $c > 0$. Then $c < q^n$ for all n , but $q^p \leq \frac{1}{q}c$ for some p , as $\frac{1}{q}c$ is not the infimum, which is a contradiction.
- for $n \in \mathbb{N}$, $x > 1$, $\lim_{n \rightarrow \infty} \frac{n^k}{x^n} = 0$. We have $x^n \geq 1+n(x-1) \geq n^{1/2}n^{1/2}(x-1) \Rightarrow \frac{n^{1/2}}{x^n} < \frac{1}{n^{1/2}(x-1)} \rightarrow 0$. Then $\frac{n^k}{x^n} = \left(\frac{n^{1/2}}{x^{1/2}n}\right)^{2k} \rightarrow 0$.
- denoting $n! = 1 \cdot 2 \cdot \dots \cdot n$, prove that $\lim_{n \rightarrow \infty} \frac{x^n}{n!} = 0$ for any real x . It is enough to check for $x > 0$. We denote $a_n = \frac{x^n}{n!} > 0$ and note that $a_{n+1} = \frac{x}{n+1}a_n$ hence a_n is monotonic decreasing for $n > x$ hence converges to its infimum. Assuming this infimum is $c > 0$, then $c < \frac{x^n}{n!}$ for all n , but there exists some $n > x$ such that $\frac{x^n}{n!} < c \frac{n+1}{x}$, as $c \frac{n+1}{x}$ is not an upper bound on a_n for $n > x$.
- prove that $x_n = (1+n^{-1})^n$ converges. We denote $\lim_{n \rightarrow \infty} (1+n^{-1})^n = e$ and we call it Euler's number. $(1+n^{-1})^n = \sum_{p=0}^n \binom{n}{p} \frac{1}{n^p} = 1 + \sum_{p=1}^n \frac{1}{p!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{p-1}{n}\right)$. Hence $x_{n+1} - x_n = \sum_{p=1}^n \frac{1}{p!} \left[\left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) \dots \left(1 - \frac{p-1}{n+1}\right) - \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{p-1}{n}\right)\right] + \frac{1}{(n+1)!} \left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) \dots \left(1 - \frac{n}{n+1}\right)$, and as $1 - \frac{1}{n+1} > 1 - \frac{1}{n} > 0$ it follows that $x_{n+1} \geq x_n$ hence x_n is monotonically increasing. Furthermore, as $x_n \leq \sum_{p \geq 0} \frac{1}{p!}$ and $p! > \frac{1}{8} 2^p$ for $p = 0, 1, \dots$ then $x_n \leq 8 \sum_{p \geq 0} \frac{1}{2^p} = 8 \frac{1 - \lim_{N \rightarrow \infty} (1/2)^N}{1 - 1/2} < 16$ hence it is bounded from above.

4.2.4 The Bolzano-Weierstrass theorem (Satz von Bolzano-Weierstrass)

We have seen that sequences that are monotonic and bounded converge. In fact, boundedness suffices to ensure convergence of a subsequence.

Theorem 4.8 (**Nested intervals theorem (Intervallschachtelung)**). Let $[a_n, b_n] \in \mathbb{R}$ be a family of nested closed intervals $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ indexed by $n \in \mathbb{N}$. If $\lim_{n \rightarrow \infty} (b_n - a_n) = 0$ then there exists exactly one real number $\{x\} = \bigcap_{n \in \mathbb{N}} [a_n, b_n]$.

Proof. Due to nesting, both (a_n) and (b_n) are monotonic and bounded, hence convergent. If they converge to different limits then $\lim_{n \rightarrow \infty} (b_n - a_n) \neq 0$ contradicting the hypothesis. It follows that $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = x$ and, due to monotonicity, $a_n \leq x \leq b_n$ for all n hence $\{x\} \in \bigcap_{n \in \mathbb{N}} [a_n, b_n]$. Finally, assuming there exist some $x' \neq x$ with $\{x'\} \in \bigcap_{n \in \mathbb{N}} [a_n, b_n]$, x' can not belong to the intervals with diameter $b_n - a_n < |x - x'|$, which leads to a contradiction. $\square$

Theorem 4.9 (**Bolzano-Weierstrass theorem**). A sequence in $\mathbb{R}^D$ is called bounded if $\|x_n\| = d(x_n, 0) \leq M$. Any bounded sequence in $\mathbb{R}^D$ has a convergent subsequence.

Proof. It is enough to prove this for $D = 1$, as the general D case can be obtained by going component by component.⁴ A sequence in $\mathbb{R}^D$ is bounded if and only if the sequence of any of its components $(x_n^\mu)_{n \in \mathbb{N}}$ are bounded. We first select a convergent subsequence for the first component, then we select a subsequence of the subsequence such that the second component converges also and so on.

We denote a_0 and b_0 an upper and a lower bound for (x_n) and we declare the interval $I_0 = [a_0, b_0]$. We chose the midpoint $\frac{a_0+b_0}{2}$ and note that at least one of the intervals $[a_0, \frac{a_0+b_0}{2}]$ and $[\frac{a_0+b_0}{2}, b_0]$ has an infinity of elements of (x_n) . We declare this interval as $I_1 = [a_1, b_1]$ and we iterate. As the diameter of the interval halves at every step we have built a family of nested intervals. We denote $\{x\} = \bigcap_{k \in \mathbb{N}} [a_k, b_k]$.

We pick a subsequence (x_{n_k}) of (x_n) as follows. We first pick x_{n_0} one element in the sequence (x_n) . At each step, among the successors of $x_{n_{k-1}}$, that is x_n with $n > n_{k-1}$, an infinity belongs to I_k (as an infinity of elements in the sequence belongs to I_k and $x_{n_{k-1}}$ has only finitely many predecessors). We pick one of them and declare it x_{n_k} . Then (x_{n_k}) is a subsequence of (x_n) and $d(x_{n_k}, x) < \frac{|a_0-b_0|}{2^k}$. We have $\lim_k x_{n_k} = x$. Indeed, for any ϵ , there exists k_ϵ such that $\frac{|a_0-b_0|}{2^{k_\epsilon}} < \epsilon$ hence for any $k > k_\epsilon$, as $I_k \subset I_{k_\epsilon}$, we have that $d(x_{n_k}, x) < \epsilon$.⁵ $\square$

Whenever the splitting in half of an interval I_n leads to two intervals containing an infinity of elements of (x_n) , each of these subsets of (x_n) has its own convergent subsequence: a bounded sequence can have subsequences which converge to different limits.

4.2.5 $\mathbb{R}^D$ is Cauchy complete

An example of a Cauchy complete metric space is provided by the real numbers.

Theorem 4.10 (*$\mathbb{R}^D$ is Cauchy complete*). *The real numbers $\mathbb{R}$ endowed with the standard topology are a Cauchy complete metric space. The same holds for $\mathbb{R}^D$.*

Proof. Any Cauchy sequence is bounded. Indeed, pick some ϵ , and for $n > N_\epsilon$, $|x_{N_\epsilon} - x_n| < \epsilon \Rightarrow |x_n| < |x_{N_\epsilon}| + \epsilon$. Then, by the Bolzano-Weierstrass theorem any Cauchy sequence in $\mathbb{R}$ has a convergent subsequence we conclude by Lemma 4.2.

This generalizes to $\mathbb{R}^D$ by noting that the sequence corresponding to the first component x_n^1 of any Cauchy sequence $x_n = (x_n^1, \dots, x_n^D) \in \mathbb{R}^D$ is Cauchy. $\square$

Corollary 4.1. *$\mathbb{R}$ is the Cauchy completion of $\mathbb{Q}$.*

Proof. As $\mathbb{Q} \subset \mathbb{R}$ and $\mathbb{R}$ is Cauchy complete, any Cauchy sequence of rational numbers converges to some number in $\mathbb{R}$.

In order to prove that $\mathbb{R}$ is the Cauchy completion of $\mathbb{Q}$ we need to prove that $\forall x \in \mathbb{R}$, there exists a Cauchy sequence of rational numbers converging to x . From Theorem 2.6, for any real number x and for any n there exists a rational number q_n such that $x - n^{-1} < q_n < x$, hence the sequence $(q_n)_{n \in \mathbb{N}}$ is a sequence of rational numbers converging to x . As any sequence in $\mathbb{R}$ is convergent if and only if it is Cauchy, $(q_n)_{n \in \mathbb{N}}$ is Cauchy. $\square$

4.2.6 Heine–Borel theorem (Satz von Heine–Borel)

An important result is the characterisation of compact sets in the standard topology on $\mathbb{R}^D$. A set in $(\mathbb{R}^D, d)$ is called *bounded* if it can be entirely covered by one (big enough) ball centered at the origin.

Theorem 4.11 (*The Heine–Borel theorem*). *In the standard topology $(\mathbb{R}^D, d)$, a set is compact if and only if it is closed and bounded.*

⁴We note that for all the distances we have defined $\|x - y\| < \sum_{i=1}^D |x^i - y^i|$.

⁵Alternatively, if (x_n) has a constant subsequence $(x_{n_k})_{k \in \mathbb{N}}$, we conclude. If not, then x is an accumulation point of the set $\{x_n | n \in \mathbb{N}\}$ and we conclude using Lemma 4.1. Indeed, for any ϵ consider the bubble $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$. As $\lim(a_n - b_n) \rightarrow 0$ there exist N_ϵ such that $n \geq N_\epsilon$ $|a_n - b_n| < \epsilon$ hence, as $a_n \leq x \leq b_n$, $I_n \subset B_\epsilon(x)$. In particular $I_{N_\epsilon} \subset B_\epsilon(x)$ and I_{N_ϵ} contains an infinity of elements of (x_n) , at least one of which is different from x as (x_n) does not have a constant subsequence.

Proof. The theorem follows from the following assertions:

If K is compact then K is closed. This holds in any Hausdorff topology. The standard topology on $\mathbb{R}^D$ Hausdorff by Theorem 3.3.

Take x a limit point of K . Any finite collection of open sets $U_i, i \in I$ such that $\forall i, \exists V_i(x)$ some open neighborhood of x such that $U_i \cap V_i(x) = \emptyset$ can not cover K . Indeed, $\bigcap_{i \in I} V_i(x)$ is an open set, thus a neighborhood of x hence it contains a point in $x' \in K$ and $x' \notin \bigcup_{i \in I} U_i$.

If K is compact but not close, then $\exists x$ limit point of K such that $x \notin K$. For every $y \in K$, pick an open set $U(y)$ containing y such that it is disjoint from some neighborhood $U_y(x)$ of x . This exists as the topology is Hausdorff. The family $U(y), y \in K$ is an open cover of K , $\bigcup_{y \in K} U(y) \supset K$, but any finite subfamily will be of the type discussed previously, hence can not cover K .

Note that if $x \in K$, such $U_y(x)$ exist only for $y \in K, y \neq x$ and $\bigcup_{y \in K, y \neq x} U(y)$ does not cover K .

If K is compact then K is bounded. This hold for any metric space in the standard topology.

Consider a compact K . The family $\{U_1(x) | x \in K\}$ of bubbles of radius 1 is an open cover of K hence admits an open subcover $\{U_1(x_i) | x_i \in I\}$ for some finite set I . The set of pairwise distances between the centers of the bubbles $\{d(x_i, x_j) | x_i, x_j \in I\}$ is a finite set hence its maximum $M = \max_{x_i, x_j \in I} d(x_i, x_j)$ is a finite number. As any point belongs to one of the bubbles in this family, by the triangle inequality, any two points are at a distance of at most $d(x, y) \leq 1 + M + 1$, hence K has finite diameter.

A closed subset of a compact set is compact. This holds for any topological space.

Let $S \subset K$ a closed subset of the compact set K , and let $\{U_\alpha | \alpha \in A\}$ an open cover of S . Then $U = \mathbb{R}^D \setminus S$ is open, $\{U\} \cup \{U_\alpha | \alpha \in A\}$ is an open cover of K , hence contains a finite cover $\{U_i | i \in I\}$ (where one of the U_i can be U) of K which also covers S . As U does not contain any element in S , $\{U_i | i \in I\} \setminus \{U\}$ is an open cover of S .

If a set is closed and bounded, then it is compact. This holds for finite dimensional Cauchy complete metric space (we will define later the dimension of a space).

By the property above, it is enough to prove that the set $T_0 = [-a, a]^D$ with $a > 0$ is compact. Assume T_0 is not compact. Then there exists an infinite open cover $\bigcup_{\alpha \in A} U_\alpha \supset K$ which admits no finite subcover for K . Similarly to the nested intervals construction, we subdivide $T_0 = \times_{i=1}^D ([-a, 0] \cup [0, a])$ into 2^n closed boxes of the type $[0, a]^p \times [-a, 0]^{n-p}$. At least one of these sub boxed requires an infinite subcover of $(U_\alpha)_{\alpha \in A}$. Call this box T_1 and iterate. We obtain a sequence of nested closed boxes $T_0 \supset T_1 \supset \dots$, where the diameter of T_k is $(2a)2^{-k}$, each of which requires an infinite subcover of $(U_\alpha)_{\alpha \in A}$. We pick a sequence $x_k \in T_k$ and x_k is Cauchy as $\forall p \geq k, x_p \in T_k$.

As $\mathbb{R}^D$ is Cauchy complete, x_k converges and $x = \lim x_k \in T_k$ for all k as T_k are closed. There exists U , one of the U_α such that $x \in U$, and since U is open there exists some ϵ such that $B_\epsilon(x) \subset U$, hence for k larger than some N_ϵ , $T_k \subset B_\epsilon(x) \subset U$, thus only one set in $(U_\alpha)_{\alpha \in A}$ suffices to cover all these T_k , leading to a contradiction.

Hence T_0 is compact.

□

Note that this property uses the completeness of $\mathbb{R}^D$, hence fails for incomplete metric spaces like $\mathbb{Q}$. It also fails in complete infinite dimensional spaces (more on this later) in which the subdivision of boxes into boxes fails.

Corollary 4.2. *The inverse of a continuous bijection between two compact Hausdorff spaces is continuous.*

Proof. Consider $f : X \rightarrow Y$ a continuous bijection with X compact, hence $Y = f(X)$ is compact. Take any $U \subset X$ open set, then $X \setminus U$ is compact (always), $f(X \setminus U)$ is compact (f continuous), $Y \setminus f(X \setminus U)$ is open (Y Hausdorff plus the first item in the proof of Heine-Borel), and using $f(U) \cap f(X \setminus U) = \emptyset$ (f bijection) we conclude that $f(U) = Y \setminus f(X \setminus U)$ is open. We thus established that f is an open map. Denote $g : Y \rightarrow X$ the inverse of f and take any open set $U \subset X$. Then $g^{-1}(U) = f(U)$ which is open. $\square$

4.3 Series

Series are infinite sums over sequences.

Definition 4.6 (**Series (Reihen)**). Starting from any sequence (a_n) valued in $\mathbb{R}$ or $\mathbb{C}$ one can build a series, which is a formal infinite sum denoted:

$$\sum_{k=0}^{\infty} a_k = a_0 + a_1 + a_2 + \dots$$

The elements a_n of the sequence (a_n) are called the general term of the series. A series is said to be:

- convergent if the sequence of partial sums $s_n = \sum_{k=0}^n a_k$ converges $s_n \rightarrow s$. In this case we say that $\sum_{k \geq 0} a_k$ converges to s .
- absolutely convergent if the series of the absolute values $\sum_{k=0}^{\infty} |a_k|$ is convergent.
- divergent if it is not convergent.

Note that if a series is convergent then the general term $a_n = s_n - s_{n-1}$ must converge to 0, $a_n \rightarrow 0$ (otherwise the sequence of partial sums is not Cauchy).⁶ Not only this, but it must in fact converge fast enough. Some examples of series are:

$$\sum_{k \geq 1} \frac{1}{k^2}, \quad \sum_{k \geq 1} \frac{(-1)^k}{k}, \quad \sum_{k \geq 1} \frac{1}{k}.$$

Also, if a series is absolutely convergent then it is convergent: indeed, $|\sum_{k=n}^m a_k| \leq \sum_{k=n}^m |a_k|$ and one concludes that if the sequence of partial sums of $\sum_k |a_k|$ is Cauchy, then the sequence of partial sums of $\sum_k a_k$ is also Cauchy.

Exercise 4.7. Prove that the three series above are respectively absolutely convergent, convergent and divergent. The first series is absolutely convergent as $\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$ hence $s_n < 1 + 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \dots + \frac{1}{n-1} - \frac{1}{n} = 2 - \frac{1}{n}$ thus s_n is monotonic and bounded. The second series is convergent as $|s_m - s_n| \leq \frac{1}{n+1}$ for $m > n$ hence s_n is Cauchy, but it is not absolutely convergent. To see this, consider as $s_n = \sum_{k \geq 1} \frac{1}{k}$ and grouping the terms we have:

$$s_{2n} = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \dots > 1 + 2^0 \times \frac{1}{2} + 2^1 \times \frac{1}{4} + 2^2 \times \frac{1}{8} + \dots 2^{n-1} \frac{1}{2^n} = 1 + \frac{n}{2},$$

hence s_{2n} is a subsequence of s_n which is not convergent.

The main question one faces when encountering a series is to determine whether the series is absolutely convergent, convergent, or divergent. Two important series are encountered often.

- *The geometric series.* The geometric series $\sum_{k \geq 0} q^k$ is absolutely convergent and converges to $(1 - q)^{-1}$ for $0 < q < 1$; it is divergent for $q \geq 1$. This follows by noting that $s_n = \frac{1 - q^{n+1}}{1 - q}$ for $q \neq 1$ and $s_n = n + 1$ for $q = 1$.
- *The generalized harmonic series.* The series $\zeta(\alpha) = \sum_{k \geq 1} \frac{1}{k^\alpha}$ is absolutely convergent for $\alpha > 1$ and divergent for $\alpha \leq 1$. By the Cauchy condensation test (see below) the series has the same nature as:

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

⁶If s_n converges then it is Cauchy then $\forall \epsilon > 0, \exists N_\epsilon$ such that $p > N_\epsilon \Rightarrow |s_{p+1} - s_p| = |a_{p+1}| < \epsilon$ hence $a_p \rightarrow 0$.

Remark that if a series with positive terms converges then it converges absolutely.

Theorem 4.12 (**Convergence tests (Konvergenzkriterium)**). *Given a generic series $\sum_{k \geq 0} a_k$:*

- Cauchy condensation test. *If $a_k \geq 0$ and a_k is monotonically decreasing then $\sum_{k \geq 1} a_k$ converges if and only if $\sum_{k \geq 1} 2^k a_{2^k}$ converges and moreover $\sum_{k \geq 1} a_k \leq \sum_{k \geq 0} 2^k a_{2^k} \leq 2 \sum_{k \geq 1} a_k$. This follows by noting that:*

$$\begin{aligned} \sum_{k \geq 1} a_k &= a_1 + (a_2 + a_3) + (a_4 + a_5 + a_6 + a_7) + \cdots \leq \sum_{k \geq 0} 2^k a_{2^k}, \\ \sum_{k \geq 0} 2^k a_{2^k} &= (a_1 + a_2) + (a_2 + a_4) + (a_4 + a_4) + (a_4 + \cdots) \leq 2 \sum_{k \geq 1} a_k. \end{aligned}$$

- Direct comparison test. *If $|a_k| \leq b_k$ (hence $b_k \geq 0$) and $\sum_{k \geq 0} b_k$ is convergent then $\sum_{k \geq 0} a_k$ is absolutely convergent and $|\sum_{k \geq 0} a_k| \leq \sum_{k \geq 0} b_k$. Indeed, as $\sum_{k \geq 0} b_k$ is convergent, the sequence $u_n = \sum_{k \geq 0}^n b_k$ is convergent hence Cauchy; but then $|s_m - s_n| \leq \sum_{k=n+1}^m |a_k| \leq \sum_{k=n+1}^m b_k \leq u_m - u_n$, hence s_m is Cauchy thus convergent and $|s_n| \leq u_n$.*
- Alternating series. *If a_k is monotonically decreasing and $a_k \rightarrow 0$ then $\sum_{k \geq 0} (-1)^k a_k$ is convergent. This follows by noting that for $m > n$, $|s_m - s_n| \leq |a_{n+1}|$.⁷*
- Cauchy-Hadamard. *Denote $\lim_{n \rightarrow \infty} (\sup_{m \geq n} |a_m|^{1/m}) = r$. If $r < 1$ then the series is absolutely convergent. If $r > 1$ it is divergent.*
- if $\lim \left| \frac{a_{n+1}}{a_n} \right| = r$ then, if $r < 1$ the series is absolutely convergent whereas if $r > 1$ it is divergent.
- Integral criterion. *This criterion can be revisited after the reader familiarizes themselves with integrals, discussed in Chapter 11. Assume a_k is monotone decreasing and $a_k \rightarrow 0$. If there exists monotonically decreasing function $f: [1, \infty) \rightarrow \mathbb{R}_+$ such that $f(n) = a_n$ then the series $\sum_{n \geq 1} a_n$ is convergent if and only if $\int_1^\infty f(t) dt < \infty$.*
- Abel's test. *If $\sum_{k \geq 1} a_k$ is convergent and (b_k) is a monotonic bounded sequence then $\sum_{k \geq 1} a_k b_k$ is convergent. The importance of this test is that it applies when $\sum_{k \geq 1} a_k$ is convergent, but not absolutely convergent. Call $A_n = \sum_{k=1}^n a_k$ and note⁸ that $\sum_{k=1}^n a_k b_k = A_n b_{n+1} - \sum_{k=1}^n A_k (b_{k+1} - b_k)$. Both (b_n) and (A_n) converge, then on the one hand $A_n b_{n+1}$ converges and on the other $\sum_{k=n}^m A_k (b_{k+1} - b_k) \leq (\sup_k |A_k|) \sum_{k=n}^m |b_{k+1} - b_k|$. As (b_n) is monotonic $\sum_{k=n}^m |b_{k+1} - b_k| = |b_m - b_n|$ and, as b_m is convergent hence it is Cauchy, $\sum_{k=n}^m A_k (b_{k+1} - b_k)$ is also Cauchy and consequently convergent.*

Exercise 4.8. *Prove the unproven statements above.*

Exercise 4.9. *We have defined $e = \lim(1 + n^{-1})^n$. Prove that $e = \sum_{k \geq 0} \frac{1}{k!}$.*

We have $s_n = \sum_{k=1}^n \frac{1}{k!}$ and $t_n = (1 + n^{-1})^n = \sum_{p=0}^n \binom{n}{p} \frac{1}{n^p} = \sum_{p=1}^n \frac{1}{p!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{p-1}{n}\right) \leq s_n$. We know that $t_n \rightarrow e$ and it is trivial to check that s_n converges. We have $\lim t_n = e \leq \lim s_n$. Now define $v_m = \sum_{p \geq 0}^m \frac{1}{p!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{p-1}{n}\right)$ for $m < n$. We have $v_m \leq t_n$ hence $\lim_{m \rightarrow \infty} v_m \leq \lim t_n = e$, but $\lim_{m \rightarrow \infty} v_m = s_m$ hence $s_m \leq e$ and $\lim s_m \leq e$.

4.3.1 Absolute convergent series (absolut konvergente Reihen)

Absolute convergent series have some additional, more constraining properties.

Lemma 4.3. *The sum of an absolutely convergent series is invariant under rearranging the terms in the series. That is, let an absolutely convergent series $\sum_k |a_k| < \infty$, $\sum_k a_k = A$. For any permutation $\sigma: \mathbb{N} \rightarrow \mathbb{N}$ (bijection), $\sum_k a_{\sigma(k)} = A$.*

⁷Alternatively $[s_{2m+1}, s_{2m+2}] \subset [s_{2m-1}, s_{2m}]$ and $s_{2m} - s_{2m-1} = a_{2m} \rightarrow 0$ hence by the nested interval theorem $\bigcap_m [s_{2m-1}, s_{2m}] = \lim s_{2m} = \lim s_{2m-1} = \sum_{k \geq 0} (-1)^k a_k$.

⁸More generally we have $\sum_{k=m}^n b_k (A_k - A_{k-1}) = b_{n+1} A_n - b_m A_{m-1} - \sum_{k=m}^n (b_{k+1} - b_k) A_k$.

4. SEQUENCES

Proof. For any ϵ let N_ϵ such that for $N > N_\epsilon$ both $\sum_{k=N}^\infty |a_k| < \epsilon/2$ and $|\sum_{k=1}^N a_k - A| < \epsilon/2$. Fix some σ and take N also larger than $\max\{\sigma^{-1}(1) \dots \sigma^{-1}(N_\epsilon)\} = \max \sigma^{-1}(\{1, \dots, N_\epsilon\})$, that is the list $a_{\sigma(0)}, a_{\sigma(1)}, \dots, a_{\sigma(N)}$ includes all the terms $a_0, a_1, \dots, a_{N_\epsilon}$, and possibly others. Then:

$$\left| \sum_{i=1}^N a_{\sigma(i)} - A \right| \leq \left| \sum_{j=1}^{N_\epsilon} a_j - A \right| + \left| \sum_{i=1, \sigma(i) > N_\epsilon}^N a_{\sigma(i)} \right| \leq \left| \sum_{j=1}^{N_\epsilon} a_j - A \right| + \sum_{j=N_\epsilon}^\infty |a_j| \leq \epsilon.$$

□

Theorem 4.13 (Cauchy product). *The Cauchy product of the two series $\sum_{k \geq 0} a_k$ and $\sum_{k \geq 0} b_k$ is the series $\sum_{p \geq 0} c_p$ with $c_p = \sum_{k=0}^p a_k b_{p-k}$. If both $\sum_{k \geq 0} a_k$ and $\sum_{k \geq 0} b_k$ converge and at least one of them converges absolutely then $\sum_{p \geq 0} c_p$ converges to $(\sum_{k \geq 0} a_k)(\sum_{k \geq 0} b_k)$. If both series are absolutely convergent then their Cauchy product is also absolutely convergent.*

Proof. Denote the partial sums $A_n = \sum_{k=0}^n a_k$, $B_n = \sum_{k=0}^n b_k$ and $C_n = \sum_{k=0}^n c_k$ and assume $\sum_k a_k$ converges absolutely. We have:

$$C_n = \sum_{i=0}^n \sum_{k=0}^i a_k b_{i-k} = \sum_{k=0}^n \sum_{i=k}^n a_k b_{i-k} = \sum_{k=0}^n a_k B_{n-k} = \sum_{k=0}^n a_{n-k} B_k = \sum_{k=0}^n a_{n-k} (B_k - B) + A_n B.$$

Since $B_n \rightarrow B$ and $\sum_k |a_k| < \infty$ then there exists some N_ϵ such that for $n > N_\epsilon$, $|B_n - B| \leq \frac{\epsilon/3}{\sum_k |a_k| + 1}$. On the other hand as A_n converges then $a_k \rightarrow 0$ hence there exists an M_ϵ such that for $n > M_\epsilon$, $|a_n| \leq \frac{\epsilon/3}{N_\epsilon(\max_{i \leq N_\epsilon-1} |B_i - B| + 1)}$. Finally, there exists an L_ϵ such that $n > L_\epsilon$, $|A - A_n| \leq \frac{\epsilon/3}{|B| + 1}$, thus for $n \geq \{L_\epsilon, M_\epsilon + N_\epsilon\}$:

$$\begin{aligned} |C_n - AB| &= \left| \sum_{k=0}^n a_{n-k} (B_k - B) + (A_n - A) B \right| \leq \sum_{k=0}^{N_\epsilon-1} \underbrace{|a_{n-k}|}_{\geq M_\epsilon} |B_k - B| + \sum_{k=N_\epsilon}^n |a_{n-k}| |B_k - B| + |A_n - A| |B| \\ &\leq \frac{\epsilon/3}{N_\epsilon(\max_{i \leq N_\epsilon-1} |B_i - B| + 1)} N_\epsilon \max_{i \leq N_\epsilon-1} |B_i - B| + \left(\sum_k |a_k| \right) \frac{\epsilon/3}{\sum_k |a_k| + 1} + \frac{\epsilon/3}{|B| + 1} |B| \leq \epsilon. \end{aligned}$$

Now, if both $\sum_k a_k$ and $\sum_k b_k$ converge absolutely, then:

$$\sum_{i=0}^n |c_i| \leq \sum_{i=0}^n \sum_{k=0}^i |a_k| |b_{i-k}| \leq \sum_{k=0}^n |a_k| \sum_{i=k}^n |b_{i-k}| \leq \left(\sum_k |a_k| \right) \left(\sum_i |b_i| \right),$$

hence $\sum_{i \geq 0} c_i$ is absolutely convergent.

□

Theorem 4.14 (Fubini-Tonelli theorem for series). *Consider the double series $\sum_{i \geq 0} \sum_{j \geq 0} a_{ij}$ with a_{ij} complex numbers. If $\sum_{i \geq 0} \sum_{j \geq 0} |a_{ij}| < \infty$ then the series is convergent and the two sums commute $\sum_{i \geq 0} \sum_{j \geq 0} a_{ij} = \sum_{j \geq 0} \sum_{i \geq 0} a_{ij}$.*

Proof. The hypothesis $\sum_{i \geq 0} \sum_{j \geq 0} |a_{ij}| < \infty$ implies $\forall i, M_i = \sum_{j=0}^\infty |a_{ij}| < \infty$ and $\sum_{i=0}^\infty M_i < \infty$.

Let us denote $f_n(m) = \sum_{i=0}^n \sum_{j=0}^m a_{ij}$ and $f(m) = \sum_{i=0}^\infty \sum_{j=0}^m a_{ij}$ and this series converges as $|\sum_{j=0}^m a_{ij}| < M_i$. Observe that for any m , $|f_n(m) - f(m)| \leq \sum_{i=k+1}^\infty |\sum_{j=0}^m a_{ij}| \leq \sum_{i=k+1}^\infty M_i$ and taking into account that $\sum_{i=0}^\infty M_i$ converges it follows that $f_n \rightarrow f$ uniformly in m , that is

$$\forall \epsilon, \forall m \quad \exists N_\epsilon \text{ such that } n > N_\epsilon \Rightarrow |f_n(m) - f(m)| < \epsilon.$$

As $|f(k) - f(m)| \leq |f(k) - f_n(k)| + |f_n(k) - f_n(m)| + |f_n(m) - f(m)|$, this uniform convergence and the convergence of $\sum M_i$ imply that $f(k)$ is a Cauchy sequence hence convergent, $\lim_{k \rightarrow \infty} f(k) = f$,

On the other hand, for any i , $\sum_{j=0}^\infty a_{ij}$ converges absolutely, therefore $\lim_{m \rightarrow \infty} f_n(m) = \sum_{i=0}^n \lim_{m \rightarrow \infty} \sum_{j=0}^m a_{ij} = A_n$ exists and, as $|A_n - A_k| \leq \sum_{i=k+1}^n M_i$, A_n is a Cauchy sequence, hence convergent $\lim_{n \rightarrow \infty} A_n = A$.

Finally:

$$|f - A| \leq |f - f(k)| + |f(k) - f_n(k)| + |f_n(k) - A_n| + |A_n - A|,$$

which is arbitrarily small for k and n large enough, therefore $f = A$. $\square$

4.3.2 Power series (Potenzreihen)

An important subclass of series are the power series, generalizing polynomials, of the form $P(z) = \sum_{n \geq 0} c_n (z - z_0)^n$, where z_0 is some fixed complex number (called the center of the disk of convergence) and z is a complex variable.

Lemma 4.4. *Let $P(z)$ a power series. If $P(z)$ converges at some z , then $P(z')$ converges absolutely for all z' such that $|z' - z_0| < |z - z_0|$.*

Proof. As $P(z)$ converges then $c_n (z - z_0)^n \rightarrow 0$ hence there exists some constant such that $\forall n, |c_n (z - z_0)^n| < S$. For any z' with $|z' - z_0| < |z - z_0|$, pick u such that $|z' - z_0| < u < |z - z_0|$. Consider the partial sum:

$$S_N = \sum_{n=0}^N |c_n (z' - z_0)^n| \leq \sup_n \{|c_n (z - z_0)^n|\} \sum_{n=0}^N \left| \frac{z' - z_0}{z - z_0} \right|^n \leq S \frac{1}{1 - \frac{u}{|z - z_0|}}.$$

We have that S_N is monotonic and bounded hence convergent. $\square$

Definition 4.7 (**Radius of converge**). *The radius of convergence of the power series $P(z) = \sum_{n \geq 0} c_n (z - z_0)^n$ is the nonnegative real number $r \in \mathbb{R}_{\geq 0}$ such that the series converges (absolutely) if $|z - z_0| < r$ and diverges if $|z - z_0| > r$, that is:*

$$r = \sup \left\{ |z - z_0| \mid |P(z)| < \infty \right\}.$$

The domain (disk) of convergence of the series is the open disk $B_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| < r\}$, and the power series converges absolutely inside its domain of convergence. The radius of convergence of a power series is:

$$\frac{1}{r} = \limsup |c_n|^{1/n}.$$

The radius of convergence of a power series can also be computed as $r = \lim_{n \rightarrow \infty} \left| \frac{c_n}{c_{n+1}} \right|$ if the limit exists. For instance, the radius of convergence of the geometric series $\sum_{n \geq 0} z^n$ is 1.

Lemma 4.5. *Consider a power series and pick z such that $|z - z_0| \leq \rho < r$. Then there exists a constant C such that $|P(z) - \sum_{n=0}^N c_n (z - z_0)^n| \leq C |z - z_0|^{N+1}$.*

Proof. There exists u with $\rho < u < r$, thus $|c_n u^n|$ is bounded by some S hence:

$$\left| \sum_{n=N+1}^P c_n (z - z_0)^n \right| \leq S \sum_{n=N+1}^P \frac{|z - z_0|^n}{u^n} \leq |z - z_0|^{N+1} \frac{S}{u^{N+1}} \frac{1}{1 - \frac{\rho}{u}}.$$

$\square$

Exercise 4.10 (Cauchy–Hadamard theorem). *Prove that the series $P(z) = \sum_n c_n z^n$ converges absolutely for $|z| < r$ and diverges for $|z| > r$ where $r^{-1} = \limsup |c_n|^{1/n}$. Denote $t = \frac{1}{r} = \limsup |c_n|^{1/n}$. Then for any ϵ only a finite number of terms are such that $|c_n| > (t + \epsilon)^n$ hence*

$\sum_n |c_n z^n| \leq \sum_n |z|^n (t + \epsilon)^n$ which converges for $|z| < \frac{1}{t + \epsilon}$. For any ϵ , $|c_n| > (t - \epsilon)^n$ for an infinity of terms, hence for $|z| = \frac{1}{t - \epsilon}$, $|c_n z^n| \not\rightarrow 0$ and the series can not converge.

4.3.3 The exponential function (Exponentialfunktion)

Many simple functions are defined as the sum of power series. We now discuss some examples. The *exponential function* is defined as the sum of the power series:

$$\exp(z) \equiv e^z = \sum_{n \geq 0} \frac{1}{n!} z^n \quad r = \lim_{n \rightarrow \infty} \frac{(n+1)!}{n!} = \infty ,$$

where we stress that the radius of convergence of this series is ∞ , that is the series converges for any $z \in \mathbb{C}$. The exponential has a number of interesting properties.

Proposition 4.1. *We have $e^{z_1} e^{z_2} = e^{z_1+z_2}$, in particular $e^z \neq 0$. Furthermore, for any $z \in \mathbb{C}$ we have:*

$$e^z = \lim_{n \rightarrow \infty} \left(1 + \frac{z}{n}\right)^n .$$

Proof. The Cauchy product of e^{z_1} and e^{z_2} is absolutely convergent to:

$$e^{z_1} e^{z_2} = \sum_{n \geq 0} \frac{1}{n!} \sum_{p=0}^n \frac{n!}{p!(n-p)!} z_1^p z_2^{n-p} = \sum_{n \geq 0} \frac{1}{n!} (z_1 + z_2)^n = e^{z_1+z_2} ,$$

and we note that $e^z e^{-z} = e^0 = 1$.

In order to prove the last equality, for any ϵ , let N such that $\sum_{k \geq N+1} \frac{|z|^k}{k!} < \frac{\epsilon}{4}$, which is possible as the exponential has infinite radius of convergence). As $|\binom{n}{k} \frac{z^k}{n^k}| \leq \frac{1}{k!} |z|^k$, for any $n > N$ we have that:

$$\begin{aligned} \left| \left(1 + \frac{z}{n}\right)^n - \sum_{k \geq 0} \frac{1}{k!} z^k \right| &\leq \sum_{k=1}^N \frac{|z|^k}{k!} \left| \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) - 1 \right| + \sum_{k=N+1}^n \binom{n}{k} \frac{|z|^k}{n^k} + \sum_{k \geq N+1} \frac{|z|^k}{k!} \\ &\leq \sum_{k=1}^N \frac{|z|^k}{k!} \left| \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) - 1 \right| + \frac{\epsilon}{2} < \epsilon , \end{aligned}$$

for n large enough. □

Exercise 4.11. *We call $f(z) = \sum_{k \geq 0} a_k z^k$ the generating function of the coefficients a_k . It might happen that the series is divergent, in which this is just a formal equality, or it is convergent. Compute the generating function of the coefficients defined recursively as $a_{n+1} = \sum_{k=0}^n a_n a_{n-k}$, $a_0 = 1$. What is the radius of converge of this generaing function?*

Exercise 4.12 (Ramanujan's trick). *One can sometimes attribute a finite value to the sum of a divergent series. For instance:*

$$\sum_{k=1}^{\infty} k = 1 + 2 + \dots + n + \dots = -\frac{1}{12} .$$

Denote $1 + 2 + \dots + n + \dots = c$. Show that $c - 4c = \sum_{p \geq 1} (-1)^{p+1} p$. We have by subtracting 4 from the second term, 8 from the fourth, 12 from the sixth and so on that $c - 4c = 1 + (2-4) + 3 + (4-8) + 5 + (6-12) + \dots = 1 - 2 + 3 - 4 + 5 - 6 + \dots = \sum_{p \geq 1} (-1)^{p+1} p$. Use $\frac{1}{1+x} = \sum_{k \geq 0} (-1)^k x^k$ to conclude that $\sum_{p \geq 1} (-1)^{p+1} p = \frac{1}{4}$. We have $\frac{1}{(1+x)^2} = \left[\sum_{k \geq 0} (-1)^k x^k \right] \left[\sum_{k' \geq 0} (-1)^{k'} x^{k'} \right] = \sum_{p \geq 0} \sum_{k=0}^p (-1)^{p-k} x^{p-k} (-1)^k x^k = \sum_{p \geq 0} (-1)^p (p+1) x^p \Rightarrow \sum_{p \geq 0} (-1)^p (p+1) = \frac{1}{4}$. We will see later that this resummation can be more rigorously justified.

Chapter 5

Vector spaces, first contact

Vector spaces are ubiquitous in mathematics.

5.1 Vector spaces

The simplest examples of vector spaces are $\mathbb{R}^D$ and $\mathbb{C}^D$, but many other vector spaces are relevant.

Definition 5.1 (**Vector space (Vektorraum)**). A vector space V over the algebraic field $\mathbb{F} = \mathbb{R}, \mathbb{C}$ is a set endowed with two operations:

- the addition $+: V \times V \rightarrow V$, $(x, y) \rightarrow x + y$ such that $(V, +, 0)$ is an Abelian group (that is $+$ is associative and commutative, 0 is the identity element for $+$ and every element $x \in V$ has an additive inverse, or negation, $-x$).
- multiplication by scalars $\cdot: \mathbb{F} \times V \rightarrow V$, $(\alpha, x) \rightarrow \alpha \cdot x$ such that:
 - $\cdot$ is compatible with the multiplication $\times$ in $\mathbb{F}$:

$$\alpha \cdot (\beta \cdot x) = (\alpha \times \beta) \cdot x .$$

- the identity of $\cdot$ is the multiplicative identity in $\mathbb{F}$, denoted 1 :

$$1 \cdot x = x .$$

- $\cdot$ is distributive with respect to both the addition in $\mathbb{F}$ and the addition in V :

$$\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y , \quad (\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x .$$

$\mathbb{F}$ is called the field of scalars. When no confusion can arise, we suppress the $\cdot$ and $\times$ notation for the multiplication.

The D dimensional Euclidean space $\mathbb{R}^D = \{x = (x^1, \dots, x^D) \mid x^\mu \in \mathbb{R}\}$ is an example of vector space over $\mathbb{R}$ with:

- the addition of $x = (x^1, \dots, x^D)$ and $y = (y^1, \dots, y^D)$ is $x + y = (x^1 + y^1, \dots, x^D + y^D)$.
- the multiplication of x by $\alpha \in \mathbb{R}$ is $\alpha x = (\alpha x^1, \dots, \alpha x^D)$.

Similarly, $\mathbb{C}^D$ is a vector space over $\mathbb{C}$. We typically represent vectors as column vectors in these cases.

Exercise 5.1. Prove that $0 \cdot x = 0$. $\alpha \cdot x = (\alpha + 0) \cdot x = \alpha \cdot x + 0 \cdot x$.

We often encounter vector spaces with more structure.

Definition 5.2 (**Norm**). A norm on the vector space $(V, \mathbb{F})$ is a function $\|\cdot\| : V \rightarrow \mathbb{R}_{\geq 0}$ such that:

- $\|v\| \geq 0$ and $\|v\| = 0$ if and only if $v = 0$.¹
- $\|\alpha x\| = |\alpha| \|x\|$, $\forall \alpha \in \mathbb{F}$ and $\forall x \in V$.
- $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in V$.

Any normed vector space is a metric space with distance function $d(x, y) = \|x - y\|$. A Banach space is a Cauchy complete normed vector space.

For instance $\mathbb{R}^D$ and $\mathbb{C}^D$ are normed spaces (and in fact Banach spaces) with norm any of the p -norms $\|x\|_p = (\sum_{\mu=1}^D |x^\mu|^p)^{1/p}$. Remark that $\mathbb{Q}^D$ with $\|\cdot\|_p$ is a finite dimensional normed vector space over $\mathbb{Q}$ which is not Banach as it is not Cauchy complete. In fact all the finite dimensional vector spaces (see below) over Cauchy complete fields of scalars are Banach spaces.

Note that if V is a vector space and $Y \subset V$ is a vector subspace then $V/Y = \{[x] \mid x_1 \sim x_2 \text{ if } x_1 - x_2 \in Y\}$ is a vector space. An important lemma is the following.

Lemma 5.1 (**Direct sum decomposition**). Consider Y_1 and Y_2 are subspaces of X such that $Y_1 + Y_2 = \{y_1 + y_2 \mid y_1 \in Y_1, y_2 \in Y_2\} = X$, that is any $x \in X$ can be decomposed as a sum $x = y_1 + y_2$ with $y_1 \in Y_1$ and $y_2 \in Y_2$. If $Y_1 \cap Y_2 = \{0\}$ (or equivalently Y_1 and Y_2 are linearly independent)², then the decomposition is unique. In this case we denote $X = Y_1 \oplus Y_2$ and say that X is the direct sum of Y_1 and Y_2 .

Proof. If x admits two decompositions $x = y_1 + y_2 = y'_1 + y'_2$, then $y_1 - y'_1 = y'_2 - y_2$ hence, as $y_1 - y'_1 \in Y_1$ and $y'_2 - y_2 \in Y_2$ we conclude that $y_1 = y'_1$ and $y_2 = y'_2$. $\square$

Given two vector spaces one can build new vector spaces.

Definition 5.3 (**Direct sum of vector spaces**). Let Y_1 and Y_2 be two vector spaces over $\mathbb{F}$. The direct sum $Y_1 \oplus Y_2$ is the set of pairs (y_1, y_2) with $y_1 \in Y_1$ and $y_2 \in Y_2$ endowed with the vector space operations $\alpha(y_1, y_2) = (\alpha y_1, \alpha y_2)$ and $(y_1, y_2) + (y'_1, y'_2) = (y_1 + y'_1, y_2 + y'_2)$.

5.2 Hamel bases, dimension of a vector space

Vector spaces are closed under taking linear combinations. Thus given a vector space, an important question is to identify a subset of vectors such that any other vector can be obtained as a linear combination of them.

Definition 5.4 (**Hamel basis**). A Hamel basis in a vector space $(V, \mathbb{F})$ is a set of vectors $B \subset V$ such that:

- Linearly independent. For every finite subset $\{e_1, \dots, e_n\} \subset B$, $\alpha_1 e_1 + \dots + \alpha_n e_n = 0 \Rightarrow \alpha_1 = \alpha_2 = \dots = \alpha_n = 0$.
- Spanning. Every vector writes as a finite linear combination of vectors in the basis, that is $\forall v \in V$ there exist $\beta_1, \beta_2, \dots, \beta_m \in \mathbb{F}$ and $\{e_1, \dots, e_m\} \subset B$ such that $v = \beta_1 e_1 + \dots + \beta_m e_m$.

The scalars $\beta_1, \dots, \beta_m$ are called the components of v in the basis B .

¹It is in fact sufficient to require $\|v\| > 0$ if $v \neq 0$.

²For any $y_1 \in Y_1$ and $y_2 \in Y_2$, $\alpha_1 y_1 + \alpha_2 y_2 = 0$ if and only if $\alpha_1 = \alpha_2 = 0$.

The span of the vectors S is the set of all finite linear combinations of these vectors $\text{Span } S = \{\alpha_1 e_1 + \dots + \alpha_n e_n \mid \alpha_i \in \mathbb{F}, e_i \in S\}$. If a vector $w \neq 0$ is not in the span of the linearly independent vectors S then $\{w\} \cup S$ are linearly independent. Indeed, assume

$$\alpha w + \alpha_1 e_1 + \dots + \alpha_n e_n = 0,$$

such that not all $\alpha, \alpha_1, \dots, \alpha_n$ are zero. Then $\alpha \neq 0$ (otherwise e_i would not be linearly independent) and

$$w = -\frac{\alpha_1}{\alpha} e_1 - \dots - \frac{\alpha_n}{\alpha} e_n,$$

and at least one of the α_i s is not zero (as $w \neq 0$) hence $w \in \text{Span } S$. The span of any set of vectors is a vector spaces.

We stress that even though a set B might be infinite, both linear independence and span involve only a *finite* number of vectors in B . On the positive side, every vector space has a Hamel basis.

Theorem 5.1 (**Every vector space has a Hamel basis**). *Every vector space has a Hamel basis. All Hamel basis of a vector space have the same cardinality. The cardinal number of a basis is called the dimension of the vector space.*

Proof. Let $(V, \mathbb{F})$ a vector space and let $\mathcal{X}$ the set of linearly independent subsets of V . $\mathcal{X}$ is non empty (as it contains the empty set) and is partially ordered by inclusion $\subset$.

For any subset $\mathcal{Y} \subset \mathcal{X}$ which is totally ordered (i.e. $\forall Y_1, Y_2 \in \mathcal{Y}$, either $Y_1 \subset Y_2$ or $Y_2 \subset Y_1$) denote $S_{\mathcal{Y}} = \bigcup_{Y \in \mathcal{Y}} Y$. $S_{\mathcal{Y}}$ is linearly independent: every finite subset of $S_{\mathcal{Y}}$ is a subset of some element in $\mathcal{Y}$. Indeed, consider a finite subset $\{s_1, \dots, s_n\} \in S_{\mathcal{Y}}$. Then there exist $Y_1 \ni s_1, \dots, Y_n \ni s_n$, $Y_1, \dots, Y_n \in \mathcal{Y}$ and, as $\mathcal{Y}$ is totally ordered, one of $Y_1, \dots, Y_n$ is the largest of them and contains $s_1, \dots, s_n$. It follows that $S_{\mathcal{Y}} \in \mathcal{X}$ and is an upper bound for $\mathcal{Y}$ with respect to the order relation $\subset$.

By Zorn's lemma, $\mathcal{X}$ has a maximal element, i.e. an element that is not smaller than any other element. Denote $S_{\max} \in \mathcal{X}$ a maximal element. Then $S_{\max}$ is linearly independent, as it belongs to $\mathcal{X}$. Moreover, $S_{\max}$ spans V . Indeed, assume it does not, then $\exists w \in V$ such that w is not in the span of $S_{\max}$ (i.e. no finite linear combination of elements in $S_{\max}$ equals w). Then $S_{\max} \cup \{w\}$ is linearly independent (as any finite subset of $S_{\max} \cup \{w\}$ is linearly independent) and strictly larger than $S_{\max}$, contradicting the fact that $S_{\max}$ is maximal.

The fact that all basiss have the same cardinality can be proven abstractly (using again Zorn's lemma). As we comment below, the infinite dimensional case is not of practical use and mostly academic. In finite dimensions it follows from lemma 5.2 below. $\square$

The drawback of Hamel basis is that they are useful only if the dimension of the space is finite, that is the Hamel bases have a finite number of elements. In this case a Hamel basis is called simply a basis. If the cardinal of the Hamel basis is infinite, things are more complicated.

5.3 Finite dimensional vector spaces

Theorem 5.2 (**Steinitz exchange lemma**). *Let U and W be finite subsets of a vector space V . If U is a set of linearly independent vectors and W spans V , then:*

- $|U| \leq |W|$,
- there is a set $W' \subset W$ with $|W'| = |W| - |U|$ such that $U \cup W'$ spans V .

Proof. We denote $U = \{u_1, \dots, u_m\}$ and $W = \{w_1, \dots, w_n\}$. We will show that “we can exchange k elements in W for elements in U ”, that is $\forall k, 0 \leq k \leq m$ the set $\{u_1, \dots, u_k, w_{k+1}, \dots, w_n\}$, where the w 's have been reordered in a k dependent manner, spans V .

We proceed by induction. $k = 0$ is trivial. Assume that for $k < m$, $\{u_1, \dots, u_k, w_{k+1}, \dots, w_n\}$ spans V then $\exists \mu_1, \dots, \mu_n$ such that:

$$u_{k+1} = \sum_{j=1}^k \mu_j u_j + \sum_{j=k+1}^n \mu_j w_j,$$

and there must exist $j \geq k+1$ with $\mu_j \neq 0$ (as the u s are linearly independent). In particular, this implies that $k+1 \leq n$. We relabel this $j = k+1$ and write:

$$w_{k+1} = \frac{1}{\mu_{k+1}} \left(u_{k+1} - \sum_{j=1}^k \mu_j u_j - \sum_{j=k+2}^n \mu_j w_j \right).$$

Thus $w_{k+1} \in \text{Span}\{u_1, \dots, u_k, u_{k+1}, w_{k+2}, \dots, w_n\}$. As, in addition, $\text{Span}\{u_1, \dots, u_k, u_{k+1}, w_{k+2}, \dots, w_n\}$ contains all $u_1, \dots, u_k, w_{k+2}, \dots, w_n$, then $\text{Span}\{u_1, \dots, u_k, u_{k+1}, w_{k+2}, \dots, w_n\} \supset \text{Span}\{u_1, \dots, u_k, w_{k+1}, w_{k+2}, \dots, w_n\} = V$. $\square$

Although an analog of this lemma holds for V an infinite dimensional vector space, its true force comes into play in the finite dimensional case. In particular, in finite dimensions, we can build a basis in a vector space by completing a basis in a subspace.

Corollary 5.1 (**Vector subspace**). *Let X be a vector space with a finite Hamel basis. Then:*

- *if a basis in X has a finite number of elements, any other basis in X will have the same number of elements. We denote $\dim X$ this common number and call it the dimension of X .*

Proof. As any two basis B_1 and B_2 are both spanning and linearly independent, we have both $|B_1| \leq |B_2|$ and $|B_2| \leq |B_1|$. $\square$

- *any set of $\dim X$ linearly independent vectors in X spans X (hence is a basis in X); any set of $\dim X$ vectors that spans X is linearly independent (hence a basis in X).*

Proof. We denote $n = \dim X$. Take $U = \{u_1 \dots u_n\} \in X$ a set of n linearly independent vectors. Choosing a basis $W = \{w_1, \dots, w_n\}$ in X and exchanging its elements with the elements of U , as $|U| = |W|$ we conclude that U spans X . For the second assertion, assume that the $\dim X$ vectors in W span X but are not linearly independent. Then one can choose a subset $W' \subset W$ with $\dim X - 1$ elements which spans X . Exchanging W' for $\dim X - 1$ vectors in a basis we obtain $\dim X - 1$ vectors which are both linearly independent and span X , which is impossible. $\square$

Let $Y \subset X$ be a vector subspace of X , then:

- *any basis in Y can be completed to a basis in X ; if $\dim Y = \dim X$ then $Y = X$ and $Y \subsetneq X$ if and only if $\dim Y < \dim X$.*

Proof. We denote $\dim Y = m$. The assertion follows by exchanging the elements of a basis $\{w_1, \dots, w_n\}$ in X for the elements of a basis $\{u_1, \dots, u_m\}$ in Y to obtain the set of $n = \dim X$ vectors $\{u_1, \dots, u_m, w_{m+1}, \dots, w_n\}$ that span X , hence are a basis in X . $\square$

If $Y_1, Y_2 \subset X$ are vector subspaces of X , then:

- *$Y_1 \cap Y_2$ is a vector subspace of X and $\dim X + \dim(Y_1 \cap Y_2) \geq \dim Y_1 + \dim Y_2$.*

Proof. This follows by choosing a basis $\{e_1, \dots, e_{\dim(Y_1 \cap Y_2)}\}$ in $Y_1 \cap Y_2$ and completing it to a basis in Y_1 by adding $\{f_{\dim(Y_1 \cap Y_2)+1}, \dots, f_{\dim Y_1}\}$ and respectively to a basis in Y_2 by adding $\{h_{\dim(Y_1 \cap Y_2)+1}, \dots, h_{\dim Y_2}\}$. As the set of all vectors e, f, h is linearly independent³ in X , we have $\dim X \geq \dim Y_1 + \dim Y_2 - \dim(Y_1 \cap Y_2)$. $\square$

³As $\alpha e + \beta f \in Y_1$ and $h \in Y_2 \setminus Y_1$, if $\alpha e + \beta f = -\gamma h$ then $\gamma = 0$ and $\alpha e + \beta f$ implies $\alpha = \beta = 0$.

Definition 5.5 (**Tensor product of vector spaces**). Let X and Y be two vector spaces over $\mathbb{F}$, with $(x_i)_{i \in I}$ and $(y_j)_{j \in J}$ bases in X and respectively Y . The tensor product space $X \otimes Y$ is the linear span of all the pairs $(x_i \otimes y_j)_{i \in I, j \in J}$, that is:

$$X \otimes Y = \text{Span}\{x_i \otimes y_j, \ i \in I, j \in J\} = \left\{ \sum_{i,j=1}^{\text{finite}} \alpha_{ij} x_i \otimes y_j, \ \alpha_{ij} \in \mathbb{F}, i \in I, j \in J \right\},$$

where the tensor product $\otimes$ is linear in each of its arguments. We have $\dim(X \otimes Y) = \dim(X) \dim(Y)$.

We stress that $X \otimes Y$ contains the tensor products $v \otimes w$ with $v \in X$ and $w \in Y$, that is $v \otimes w = \sum_{i,j} v_i w_j x_i \otimes y_j$, but contains also more general linear combinations $\sum_{i,j} \alpha_{ij} x_i \otimes y_j$.

5.4 Linear operators

Definition 5.6 (**Linear operators**). An operator $\mathcal{A} : X \rightarrow Y$ from a vector space to a vector space is called linear if $\mathcal{A}(\alpha x + \beta y) = \alpha \mathcal{A}(x) + \beta \mathcal{A}(y)$. We sometimes denote $\mathcal{A}(x) = Ax$ in order to simplify the notation. The operators $\mathcal{A}$ also form a vector space over $\mathbb{F}$, with $(\alpha \mathcal{A} + \beta \mathcal{B})(x) = \alpha \mathcal{A}(x) + \beta \mathcal{B}(x)$. The kernel of $\mathcal{A}$ is the set $\ker \mathcal{A} = \{x \in X \mid Ax = 0\}$, and the image of $\mathcal{A}$ is the set $\text{Im } \mathcal{A} = \{y \in Y \mid \exists x \in X, Ax = y\}$. Both $\ker \mathcal{A}$ and $\text{Im } \mathcal{A}$ are vector spaces.

A linear operator is also called a linear homomorphism. If $\mathcal{A} : X \rightarrow X$, then $\mathcal{A}$ is called a linear endomorphism.

Theorem 5.3 (**The rank nullity theorem**). Let $\mathcal{A} : X \rightarrow Y$ a linear operator on a finite dimensional vector space X . Then $\dim(\ker \mathcal{A}) + \dim(\text{Im } \mathcal{A}) = \dim(X)$.

Proof. We denote $\dim X = n$. The kernel $\ker \mathcal{A}$ is a vector subspace of X . We denote $k = \dim(\ker \mathcal{A})$, we chose a basis $\{v_1, \dots, v_k\}$ in the kernel and we extend it to a basis $= \{v_1, \dots, v_k, w_{k+1}, \dots, w_n\}$ in X . Then:

$$\text{Im } \mathcal{A} = \text{Span}\{\mathcal{A}v_1, \dots, \mathcal{A}v_k, \mathcal{A}w_{k+1}, \dots, \mathcal{A}w_n\} = \text{Span}\{\mathcal{A}w_{k+1}, \dots, \mathcal{A}w_n\},$$

that is the vectors $\mathcal{A}w_{k+1}, \dots, \mathcal{A}w_n$ span $\text{Im } \mathcal{A}$. These vectors are linearly independent as:

$$\sum_{i=k+1}^n \alpha_i \mathcal{A}w_i = 0 \Rightarrow \mathcal{A}\left(\sum_{i=k+1}^n \alpha_i w_i\right) = 0 \Rightarrow \sum_{i=k+1}^n \alpha_i w_i \in \ker \mathcal{A} \Rightarrow \sum_{i=k+1}^n \alpha_i w_i + \sum_{i=1}^k \alpha_i v_i = 0 \Rightarrow \alpha_i = 0, \ i = 1, \dots, n,$$

hence form a basis in $\text{Im } \mathcal{A}$. □

Definition 5.7 (**Operator norm**). The operator norm of $\mathcal{A}$ is:

$$\|\mathcal{A}\|_{\text{op.}} = \inf \left\{ c \geq 0 \mid \|\mathcal{A}v\| \leq c\|v\|, \forall v \in V \right\} = \sup_{v \in V} \frac{\|\mathcal{A}v\|}{\|v\|} = \sup_{v \in V, \|v\|=1} \|\mathcal{A}v\|.$$

We will prove later that any linear operator on a finite dimensional vector space is bounded, that is $\|\mathcal{A}\|_{\text{op.}} \leq \infty$.

Exercise 5.2. Prove that $\|\cdot\|_{\text{op.}}$ is a norm. Prove it is submultiplicative $\|AB\|_{\text{op.}} \leq \|A\|_{\text{op.}} \|B\|_{\text{op.}}$.

The linear operators $\mathcal{A} : X \rightarrow Y$, with point wise addition and multiplication by scalars form a vector space, and in fact a Banach space with respect to $\|\cdot\|_{\text{op.}}$, which we denote $B_{\text{op.}}(X, Y)$.

5.4.1 The dual space

Definition 5.8 (**The dual**). A linear form or linear functional on the vector space V over $\mathbb{F}$ is a linear function $\varphi : V \rightarrow \mathbb{F}$, $\varphi(\alpha x + \beta y) = \alpha \varphi(x) + \beta \varphi(y)$, for $x, y \in V$, $\alpha, \beta \in \mathbb{F}$

The dual space V^* is the vector space of linear functionals on V endowed with the vector space operations:

- $(\varphi_1 + \varphi_2)(x) = \varphi_1(x) + \varphi_2(x)$, $\varphi_1, \varphi_2 \in V^*$, $x \in V$,
- $(\alpha \cdot \varphi)(x) = \alpha \cdot \varphi(x)$, $\varphi \in V^*$, $\alpha \in \mathbb{F}$, $x \in V$.

The elements of the dual are also called covectors. There exists a natural homomorphic injection of V into $V^{**} = \{\Phi : V^* \rightarrow \mathbb{F} \mid \Phi \text{ linear}\}$ (which can be promoted to an isomorphic bijection in finite dimension) $v \rightarrow \Phi(v)$ defined by $(\Phi(v))(\varphi) = \varphi(v)$, $\forall v \in V$ and $\forall \varphi \in V^*$.

For every basis in a finite dimensional vector space, one can build a basis in the dual space.

Construction 5.1 (Dual basis). If the dimension of V is finite, V^* has the same dimension. Given a basis $\{e_1, \dots, e_N\}$ in V one can build a basis in V^* called the dual basis by specifying the linear functionals θ^j defined by $\theta^j(e_i) = \delta_i^j$, that is $\theta^j(\sum_{i=1}^N v^i e_i) = v^j$. By definition θ^j are linear functionals and they are linearly independent as:

$$\sum_{i=1}^N \alpha_i \theta^i(v) = 0, \quad \forall v \Rightarrow \sum_{i=1}^N \alpha_i \theta^i(e_k) = \alpha_k = 0,$$

and span V^* as any functional φ writes as:

$$\forall v, \quad \varphi(v) = \varphi\left(\sum_{i=1}^N v^i e_i\right) = \sum_{i=1}^N \varphi(e_i) v^i = \sum_{i=1}^N \varphi(e_i) \theta^i(v) \Rightarrow \varphi = \sum_{i=1}^N \varphi(e_i) \theta^i.$$

Lemma 5.2 (Basis in $B_{\text{op.}}(X, Y)$). If X and Y are finite dimensional vector spaces, then the vector space of the linear operators is the tensor product $B_{\text{op.}}(X, Y) = Y \otimes X^*$. Given $\{e_j, \theta^j\}_{j \in J}$ a basis in X and its dual basis in X^* , respectively $\{f_i, \mu^i\}_{i \in I}$ dual bases in Y and Y^* , any linear operator $\mathcal{A} : X \rightarrow Y$ decomposes in the tensor product basis $f_i \otimes \theta^j$ as $\mathcal{A} = \sum_{i \in I, j \in J} [\mu^i(\mathcal{A}e_j)] f_i \otimes \theta^j$. For instance, the identity operator on $\mathcal{I} : V \rightarrow V$, $\mathcal{I}(v) = v$ writes as $\mathcal{I} = \sum_{i=1}^{\dim V} e_i \otimes \theta^i$.

Proof. It is enough to note that any vector $y \in Y$ decomposes as $y = \sum_{i \in I} y^i f_i = \sum_{i \in I} f_i \mu^i(y)$ hence:

$$\mathcal{A}(x) = \sum_{i \in I} f_i \mu^i(\mathcal{A}x) = \sum_{i \in I, j \in J} f_i [\mu^i(\mathcal{A}e_j)] x^j = \sum_{i \in I, j \in J} [\mu^i(\mathcal{A}e_j)] f_i \theta^j(x) = \sum_{i \in I, j \in J} [\mu^i(\mathcal{A}e_j)] (f_i \otimes \theta^j)(x).$$

We denote $\mu^i(\mathcal{A}e_j) = A_{ij}^i$ and call $(A_{ij}^i)_{i \in I, j \in J}$ the matrix elements of $\mathcal{A}$ in the given bases. □

5.5 Affine spaces

An affine space is morally what is left from a vector space when one forgets all the information that depends on the origin. For instance, while the sum of two vector depends on the origin and is not captured in the affine structure, the difference (or any affine combination $\lambda \vec{v} + (1 - \lambda) \vec{w}$) does not and is encoded in the affine structure.

Definition 5.9 (Affine space). An affine space is a set V and a vector space $\vec{V}$ and an operation $+: V \times \vec{V} \rightarrow V$, $(a, \vec{v}) \mapsto a + \vec{v}$ such that (denoting with an arrow the elements in $\vec{V}$):

- (right identity) $a + \vec{0} = a$, $\forall a \in V$, where $\vec{0}$ is the vector zero in $\vec{V}$,
- (associativity) $\forall \vec{v}, \vec{w} \in \vec{V}$ and $\forall a \in V$ we have $(a + \vec{v}) + \vec{w} = a + (\vec{v} + \vec{w})$, where $\vec{v} + \vec{w}$ is the addition in $\vec{V}$,
- (free and transitive action) $\forall a \in V$, the mapping from $\vec{V}$ to V , $\vec{v} \mapsto a + \vec{v}$ is a bijection.

The dimension of an affine space is the dimension of the associated vector space.

For every $\vec{v} \in \vec{V}$, the translation by $\vec{v}$, that is the map $T_{\vec{v}} : V \rightarrow V$, $T_{\vec{v}}(a) = a + \vec{v}$, is a bijection. Note that $T_{\vec{0}}$ is the identity and $T_{\vec{v}+\vec{w}} = T_{\vec{w}}T_{\vec{v}}$. If $T_{\vec{v}}(a) = T_{\vec{v}}(b)$ then $T_{-\vec{v}}T_{\vec{v}}(a) = T_{-\vec{v}}T_{\vec{v}}(b)$ hence $a = b$ and for any b , $T_{\vec{v}}(T_{-\vec{v}}(b)) = b$.

The third property in the definition is equivalent to saying that for every $a, b \in V$, there exist a unique $\vec{v}$ such that $b = a + \vec{v}$.⁴ We sometimes denote this vector $\vec{ab}$ or $b - a$. The subtraction is such that:

- $\forall a \in V, \forall \vec{v} \in \vec{V}$, there exists a unique $b \in V$ such that $b - a = \vec{v}$, namely $a + \vec{v}$,
- $\forall a, b, c \in V, (c - b) + (b - a) = c - a$, as $(a + (b - a)) + (c - b) = c = a + ((b - a) + (c - b))$.

An *affine subspace* of V is a subset of the form $a + \vec{W} = \{a + \vec{w}, \vec{w} \in \vec{W}\}$ with a a point in V and $\vec{W}$ a vector subspace of $\vec{V}$. The vector subspace $\vec{W}$ is called the *direction* of the affine subspace and two affine subspaces with the same direction are called *parallel*. Given any direction $\vec{W}$ for any point $a \in V$, there is only one affine subspace of direction $\vec{W}$, namely $a + \vec{W}$, which is known as Euclid's fifth postulate. Any translation of the base point maps an affine subspace to a parallel subspace.

Exercise 5.3. Consider a linear operator on a finite dimensional vector space $\mathcal{H} : X \rightarrow Y$. Prove that for any fixed $y \in \text{Im } \mathcal{H}$, the space of solutions of the equation $\mathcal{H}x = y$ is an affine space over $\ker \mathcal{H}$. We denote $W_y = \{x \in X \mid \mathcal{H}x = y\}$ and define $+$: $W_y \times \ker \mathcal{H} \rightarrow W_y$ by $(x, v) \rightarrow x + \vec{v}$, with $+$ the addition in X . Clearly, for any $x \in W_y$, $\mathcal{H}(x + \vec{v}) = \mathcal{H}(x) + \mathcal{H}(\vec{v}) = y$ hence $x + \vec{v} \in W_y$ and the associativity is trivial. It remains to check that for any x , the function $\vec{v} \rightarrow x + \vec{v}$ is a bijection. As x is fixed, this is clearly an injection. For any $x' \in W_y$ we have $\mathcal{H}(x' - x) = y - y = 0$ hence $x' - x \in \ker \mathcal{H}$ and $x' = x + (x' - x)$.

Definition 5.10 (Affine map). An affine map (affine homomorphism) between two affine spaces $(X, \vec{X})$ and $(Y, \vec{Y})$ is a map $\mathcal{A} : X \rightarrow Y$ such that the map $\vec{\mathcal{A}} : \vec{X} \rightarrow \vec{Y}$, $\vec{\mathcal{A}}(b - a) = \mathcal{A}(b) - \mathcal{A}(a)$ is a well defined linear map, that is if $b - a = d - c$ then $\mathcal{A}(b) - \mathcal{A}(a) = \mathcal{A}(d) - \mathcal{A}(c)$. It follows that $\mathcal{A}(a + \vec{v}) = \mathcal{A}(a) + \vec{\mathcal{A}}(\vec{v})$, hence $\mathcal{A}$ is defined by its value at a point and the associated linear map $\vec{\mathcal{A}}$.

An affine transformation, or endomorphisms is an affine map from $\mathcal{A} : V \rightarrow V$. The translation $T_{\vec{v}}$ is an endomorphisms. Given a point b and a linear map $\vec{\mathcal{A}} : \vec{V} \rightarrow \vec{V}$ we can build the affine transformation $L_{b, \vec{\mathcal{A}}} : V \rightarrow V$, $L_{b, \vec{\mathcal{A}}}(a) = b + \vec{\mathcal{A}}(b - a)$, that is the linear map $\vec{\mathcal{A}}$ centered at “the origin” b .

For two choices of an origin o, o' , if $\lambda_1 + \dots + \lambda_n = 0$, then for any $a_1, \dots, a_n$:

$$\lambda_1(a_1 - o) + \dots + \lambda_n(a_n - o) = \lambda_1(a_1 - o') + \dots + \lambda_n(a_n - o').$$

Indeed, using $\lambda_n = -\lambda_1 - \dots - \lambda_{n-1}$, both sides equal the vector $\lambda_1(a_1 - a_n) + \dots + \lambda_{n-1}(a_{n-1} - a_n)$. As this vector is independent of the choice of origin o , we denote it $\lambda_1 a_1 + \dots + \lambda_n a_n$. Note that the subtraction is just the case with two terms.

Exercise 5.4. Prove that $g = o + \lambda_1(a_1 - o) + \dots + \lambda_n(a_n - o)$ is independent of o if $\lambda_1 + \dots + \lambda_n = 1$. If $g = o + \vec{v}$ and $g' = o' + \vec{v}'$ then $g - o = \vec{v}$, $g' - o' = \vec{v}'$ hence $(g - o) - (g' - o') = \vec{v} - \vec{v}'$, but the right hand side is also $(g - o) + (o - g') - [(o - g') + (g' - o')] = (g - g') - (o - o')$, therefore $g - g' = o - o' + \vec{v} - \vec{v}'$. Denoting $g = o + \lambda_1(a_1 - o) + \dots + \lambda_n(a_n - o)$ and $g' = o' + \lambda_1(a_1 - o') + \dots + \lambda_n(a_n - o')$, and taking into account that $a - b = -(b - a)$ (if $a = b + \vec{v}$ then $b = a + (-\vec{v})$), we get $g - g' = (o - o') - \lambda_1(o - o') - \dots - \lambda_n(o - o') = \vec{0}$.

For $\lambda_1 + \dots + \lambda_n = 1$ and for some choice of the origin o , we denote $g = o + \lambda_1(a_1 - o) + \dots + \lambda_n(a_n - o)$. This is independent of o and we simplify the notation to $g = \lambda_1 a_1 + \dots + \lambda_n a_n$ and say that g is the affine combination (or the barycenter) of the a_i s with weights λ_i . The affine span of $X \subset A$ is the set of all finite affine combinations of elements in X and its direction is the linear span of $x - y$ with $x, y \in X$. Fixing $x_0 \in X$, the direction of the affine span is the linear span of $x - x_0$ for all $x \in X$.⁵

⁴As $\vec{v} \mapsto a + \vec{v}$ is surjective, there exist some $\vec{v}$ such that $a + \vec{v} = b$. As $\vec{v} \mapsto a + \vec{v}$ is injective, if $a + \vec{v}_1 = a + \vec{v}_2$ then $a + (\vec{v}_1 - \vec{v}_2) = a$ but at the same time $a + \vec{0} = a$ hence $\vec{v}_1 - \vec{v}_2 = \vec{0}$.

⁵Equivalently, the affine span of X is the smallest affine space containing X and it is the intersection of all the affine subspaces containing X . Its direction is the intersection of the directions of the affine subspaces containing it.

Definition 5.11 (**Affine basis, affine and barycentric coordinates**). An affine basis in a affine space $V, \vec{V}$ of dimension n is a set of $n+1$ elements in $\{e_0, \dots, e_n\}$ such that $\{e_1 - e_0, \dots, e_n - e_0\}$ is a basis in $\vec{V}$. Every $x \in V$ has a unique decomposition in an affine basis $x = e_0 + \lambda_1(e_1 - e_0) + \dots + \lambda_n(e_n - e_0)$, or equivalently:

$$x = \lambda_0 e_0 + \lambda_1 e_1 + \dots + \lambda_n e_n, \quad \lambda_0 + \dots + \lambda_n = 1.$$

The coordinates $\lambda_0, \dots, \lambda_n$ are called the barycenter coordinates of x .

An affine coordinate frame in an affine space $V, \vec{V}$ of dimension n is a point $o \in V$, called the origin and a basis $\{\vec{v}_1, \dots, \vec{v}_n\}$ in $\vec{V}$. For each point $x \in V$, there exists a unique n -tuple $\lambda_1, \dots, \lambda_n$ such that $x = o + \lambda_1 \vec{v}_1 + \dots + \lambda_n \vec{v}_n$. The numbers λ_i are called the affine coordinates of x in the frame $\{o, \vec{v}_1, \dots, \vec{v}_n\}$.

The point with barycentric coordinates $\lambda_0, \dots, \lambda_n, \lambda_0 + \dots + \lambda_n = 1$ in the affine basis $\{e_0, \dots, e_n\}$ corresponds to the point with affine coordinates $\lambda_1, \dots, \lambda_n$ in the affine frame $\{o = e_0, \vec{v}_1 = e_1 - e_0, \dots, \vec{v}_n = e_n - e_0\}$. Conversely, the point with affine coordinates $\lambda_1, \dots, \lambda_n$ in the affine frame $\{o, \vec{v}_1, \dots, \vec{v}_n\}$ corresponds to the point with barycentric coordinates $1 - \lambda_1 - \dots - \lambda_n, \lambda_1, \dots, \lambda_n$ in the affine basis $\{e_0 = o, e_1 = o + \vec{v}_1, \dots, e_n = o + \vec{v}_n\}$.

Exercise 5.5. Prove that $x = e_0 + \lambda_1(e_1 - e_0) + \dots + \lambda_n(e_n - e_0)$ if and only if $x = (1 - \lambda_1 - \dots - \lambda_n)e_0 + \lambda_1 e_1 + \dots + \lambda_n e_n$.

If $x = (1 - \lambda_1 - \dots - \lambda_n)e_0 + \lambda_1 e_1 + \dots + \lambda_n e_n$ then, choosing e_0 as origin, we have $x = e_0 + (1 - \lambda_1 - \dots - \lambda_n)(e_0 - e_0) + \lambda_1(e_1 - e_0) + \dots + \lambda_n(e_n - e_0)$ and we conclude. Conversely, if $x = e_0 + \lambda_1(e_1 - e_0) + \dots + \lambda_n(e_n - e_0)$ then $x - e_0 = (1 - \lambda_1 - \dots - \lambda_n)(e_0 - e_0) + \lambda_1(e_1 - e_0) + \dots + \lambda_n(e_n - e_0)$ and we conclude.

Exercise 5.6. Any vector space is an affine space over itself. We denote the vectors in $\mathbb{R}^3$ as (x, y, z) . Prove that the planes defined as $P_a = \{(x, y, z) | x + y + z = a, a \in \mathbb{R}\}$ are parallel affine subsets of $\mathbb{R}^3$. Identify barycentric and affine coordinates on P_a . We denote $e_0 = (1, 0, 0)$, $e_1 = (0, 1, 0)$ and $e_2 = (0, 0, 1)$. The planes P_a can be parametrized as $P_a = \{(a - y - z, y, z) | y, z \in \mathbb{R}\}$. Denoting $a = (a, 0, 0)$ and $\vec{v}_1 = e_1 - e_0$ and $\vec{v}_2 = e_2 - e_0$ we have $P_a = a + y\vec{v}_1 + z\vec{v}_2$, with + the addition in $\mathbb{R}^3$, hence they are all parallel affine subspaces in $\mathbb{R}^3$. An affine basis in P_a is $e_0^a = (a, 0, 0)$, $e_1^a = (0, a, 0)$ and $e_2^a = (0, 0, a)$: any element in P_a decomposes in this basis as $(1 - \lambda_1 - \lambda_2)e_0^a + \lambda_1 e_1^a + \lambda_2 e_2^a = (1 - a\lambda_1 - a\lambda_2, a\lambda_1, a\lambda_2)$ and an affine frame is $(o = e_0^a, \vec{v}_1 = e_1^a - e_0^a, \vec{v}_2 = e_2^a - e_0^a)$.

Chapter 6

Matrices

We will now be more precise with the notation and distinguish between vectors, forms and operators and their expressions in various bases. We will denote $\mathbf{x}$ the vectors in the (finite dimensional) vector space X , $\varphi \in X^*$ the linear forms and $\mathcal{A}: X \rightarrow Y$ the linear operators from X to Y .

Once we pick a basis $\{\mathbf{e}_i\}_{i=1,\dots,N}$, any vector will decompose uniquely as $\mathbf{x} = \sum_{i=1}^N x^i \mathbf{e}_i$, with x^i the components of the vector in the basis. These components can be organized in a *column vector*:

$$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^N \end{pmatrix} = x^1 e_1 + \dots x^N e_N, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots \quad e_N = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

where the entry i in the column vector x corresponding to $\mathbf{x} \in X$ is the component x^i of the vector in the direction $\mathbf{e}_i$. In particular, the basis vectors are columns with only one entry 1 and the other 0. In a basis, any vector space X is identified with the vector space $\mathbb{R}^N$ or $\mathbb{C}^N$ of column vectors.

Choosing a basis in the N dimensional vector space X and a basis in the M dimensional vector space Y allows us to represent the linear operators $\mathcal{A}: X \rightarrow Y$ as $M \times N$ *matrices*, that is a table of numbers with M rows and N columns. In detail, let $\{\mathbf{e}_j\}_{j=1\dots N}$ be a basis in X and $\{\theta^j\}_{j=1\dots N}$ its dual basis in X^* , that is $\theta^j(\mathbf{e}_k) = \delta_k^j$, $\theta^j(\sum_{k=1}^N x^k \mathbf{e}_k) = x^j$ and respectively let $\{\mathbf{f}_i, \mu^i\}_{i=1,\dots,M}$ be a pair of dual bases in Y and Y^* . For any vector in $\mathbf{y} \in Y$ we have $\mathbf{y} = \sum_{i=1}^M y^i \mathbf{f}_i = \sum_{i=1}^M \mu^i(\mathbf{y}) \mathbf{f}_i$, and the linear operator $\mathcal{A}: X \rightarrow Y$ decomposes in the tensor product basis $\mathbf{f}_i \otimes \theta^j$ as :

$$\begin{aligned} \mathcal{A}\mathbf{x} &= \sum_{i=1}^N \mathbf{f}_i \mu^i(\mathcal{A}(\mathbf{x})) = \sum_{i=1}^N \mathbf{f}_i \mu^i \left(\mathcal{A} \left(\sum_{j=1}^M x^j \mathbf{e}_j \right) \right) = \sum_{i=1}^N \sum_{j=1}^M \mathbf{f}_i [\mu^i(\mathcal{A}\mathbf{e}_j)] x^j = \sum_{i=1}^N \sum_{j=1}^M [\mu^i(\mathcal{A}\mathbf{e}_j)] \mathbf{f}_i \theta^j(\mathbf{x}) \\ \Rightarrow \mathcal{A} &= \sum_{i=1}^N \sum_{j=1}^M [\mu^i(\mathcal{A}\mathbf{e}_j)] \mathbf{f}_i \otimes \theta^j. \end{aligned}$$

We denote $\mu^i(\mathcal{A}\mathbf{e}_j) = A_{ij}^i = a_{ij}$ and call $(a_{ij})_{i=1\dots N, j=1\dots M}$ the *matrix elements* of $\mathcal{A}$ in the given bases. For instance, the identity operator on a vector space V , $\mathcal{I}: V \rightarrow V$, $\mathcal{I}(v) = v$ writes as $\mathcal{I} = \sum_{i=1}^N \mathbf{e}_i \otimes \theta^i$ and has matrix elements $I_{ij}^i = \delta_{ij}$.

As a point of notation, we use calligraphic letters $\mathcal{A}, \mathcal{B}$ and so on for operators, capital letters A, B and so on for the matrices of these operators in some bases, capital letters A_{ij}^i, B_{ij}^i for the matrix elements if we track the position of the indices and lower case letters $A_{ij}^i = a_{ij}, B_{ij}^i = b_{ij}$ if we write all the indices in subscript.

The numbers $(a_{ij})_{i=1\dots N, j=1\dots M}$ form a matrix, that is a table with M rows and N columns, having thus $M \times N$ entries a_{ij} , which are just (real or complex) numbers. We denote $\mathfrak{M}_{M \times N}(\mathbb{F})$ the space of $M \times N$ matrices whose entries a_{ij} belong

to $\mathbb{F} = \mathbb{R}, \mathbb{C}$:

$$A \in \mathfrak{M}_{M \times N}(\mathbb{F}), \quad A = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & & \\ a_{M1} & \dots & a_{MN} \end{pmatrix}.$$

The action of the operator $\mathcal{A} : X \rightarrow Y$ on a vector $\mathbf{x}$ is the *row-column multiplication* of the corresponding matrix $A \in \mathfrak{M}_{M \times N}(\mathbb{F})$ with column vector $x \in \mathbb{F}^N$, that is the component i in the basis $\mathbf{f}_i$ of the vector $\mathcal{A}\mathbf{x}$ is:

$$\mu^i(\mathcal{A}\mathbf{x}) \equiv (Ax)^i = \sum_{j=1}^N A_{ij}^i x^j = \sum_{j=1}^N a_{ij} x^j,$$

which writes in terms of tables of numbers:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & & \\ a_{M1} & \dots & a_{MN} \end{pmatrix}, \quad x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^N \end{pmatrix}, \quad Ax = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & & \\ a_{M1} & \dots & a_{MN} \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^N \end{pmatrix} = \begin{pmatrix} a_{11}x^1 + \dots + a_{1N}x^N \\ \vdots \\ a_{M1}x^1 + \dots + a_{MN}x^N \end{pmatrix}.$$

In words, one multiplies term by term the entries in the rows in A with the entries in the column vector and adds up the results.

The linear forms $\varphi : X \rightarrow \mathbb{F}$ are represented as $1 \times N$ matrices, that is line vectors. The form $\varphi = \sum_{i=1}^N \varphi_i \theta^i$ acts on vectors as $\varphi(\mathbf{x}) = \sum_{j=1}^N \varphi_j \theta^j(\mathbf{x}) = \sum_{j=1}^N \varphi_j x^j$ and it is represented by the line vector:

$$\varphi = (\varphi_1 \quad \varphi_2 \quad \dots \quad \varphi_N), \quad \theta^1 = (1 \quad 0 \quad \dots \quad 0), \quad \theta^2 = (0 \quad 1 \quad \dots \quad 0), \quad \theta^N = (0 \quad 0 \quad \dots \quad 1).$$

Note that, while the components $x^j = \theta^j(\mathbf{x})$ of the vector $\mathbf{x}$ and the matrix elements $A_{ij}^i = \mu^i(\mathcal{A}\mathbf{e}_j)$ of the operator $\mathcal{A}$ depend the bases we have chosen, the vectors $\mathbf{x}$ and $\mathcal{A}\mathbf{x}$ are basis independent. We will see later that, in order to reflect this, the components of vectors and the matrix elements of operators change in a precise way when changing basis. The identity operator is special: its matrix elements in any basis are $I_j^i = \delta_{ij}$.

6.1 The algebra of matrices

Once we fixed bases, the set of operators is $B_{\text{op.}}(X, Y)$ with $\dim X = N$ and $\dim Y = M$ is represented by the set of matrices $\mathfrak{M}_{M \times N}(\mathbb{F})$ and consequently the matrices form a vector space with $(A+B)_{ij}^i = a_{ij} + b_{ij}$ and $(\alpha A)_{ij}^i = \alpha a_{ij}$, that is entry-wise addition and multiplication by scalars. The dimension of this vector space is $M \times N$, and a basis consists in the matrices associated to $\mathbf{f}_i \otimes \theta^j$:

$$(\mathbf{f}_i \otimes \theta^j)_l^k \equiv (E_{(ij)})_l^k = \delta_i^k \delta_l^j \quad E_{(ij)} = \begin{matrix} & & j \\ & & \begin{pmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix} \end{matrix}.$$

In order to simplify the notation we often suppress the tensor product symbol.

Linearity operators can be composed, that is if $\mathcal{B} : X \rightarrow Y$ and $\mathcal{A} : Y \rightarrow Z$ with $\dim(X) = N$, $\dim Y = M$, and $\dim(Z) = P$ are linear operators, then $\mathcal{A} \circ \mathcal{B} : X \rightarrow Z$ is a linear operator. Fixing bases $\{\mathbf{e}_j, \theta^j\}$, $\{\mathbf{f}_i, \mu^i\}$, $\{\mathbf{h}_k, \rho^k\}$ in X, Y respectively

Z , and taking into account that the identity in Y is $\sum_i \mathbf{f}_i \theta^i$, the matrix element of the composition is:

$$\rho^k[\mathcal{A} \circ \mathcal{B}(\mathbf{e}_j)] \equiv (A \cdot B)^k_j = \sum_{i=1}^{M=\dim Y} \rho^k[\mathcal{A}(\mathbf{f}_i)] \mu^i[\mathcal{B}(\mathbf{e}_j)],$$

corresponding to the the *row-column product* of then two matrices:

$$A \in \mathfrak{M}_{P \times M}(\mathbb{F}), \quad B \in \mathfrak{M}_{M \times N}(\mathbb{F}), \quad A \cdot B \in \mathfrak{M}_{P \times N}(\mathbb{F}), \quad (A \cdot B)^k_j = \sum_{i=1}^M A^k_i B^i_j = \sum_{i=1}^M a_{ki} b_{ij}.$$

In terms of tables of numbers one multiplies term by term the rows in A with columns in B and adds up the results:

$$\begin{pmatrix} a_{11} & \dots & a_{1M} \\ \vdots & & \\ a_{P1} & \dots & a_{PM} \end{pmatrix} \begin{pmatrix} b_{11} & \dots & b_{1N} \\ \vdots & & \\ b_{M1} & \dots & b_{MN} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + \dots + a_{1M}b_{M1} & \dots & a_{11}b_{1N} + \dots + a_{1M}b_{MN} \\ \vdots & & \\ a_{P1}b_{11} + \dots + a_{PM}b_{M1} & \dots & a_{P1}b_{1N} + \dots + a_{PM}b_{MN} \end{pmatrix}.$$

In particular the basis matrices are matrix products $E_{(ij)} = \mathbf{f}_i \cdot \theta^j$. Note that the identity matrix with matrix elements $I^i_j = \delta_{ij}$ is an identity for the matrix product $A \cdot I = I \cdot A = A$.

Definition 6.1 (**Algebra over a field**). *An algebra $\mathfrak{M}$ over a field of scalars $\mathbb{F}$ is a vector space over $\mathbb{F}$ endowed with a bilinear product $\cdot : \mathfrak{M} \times \mathfrak{M} \rightarrow \mathfrak{M}$. The bilinearity of the product can be states equivalently as:*

- (right distributive) $(A + B) \cdot C = A \cdot C + B \cdot C$ for any $A, B, C \in \mathfrak{M}$,
- (left distributive) $A \cdot (B + C) = A \cdot B + A \cdot C$ for any $A, B, C \in \mathfrak{M}$,
- (compatible with scalars) $(\alpha A) \cdot (\beta B) = (\alpha\beta)(A \cdot B)$, for any $\alpha, \beta \in \mathbb{F}$ and $A, B \in \mathfrak{M}$.

The algebra is associative if $\cdot$ is associative $A \cdot (B \cdot C) = (A \cdot B) \cdot C$, commutative if $A \cdot B = B \cdot A$, and unital if $\cdot$ has a unit element $I \cdot A = A \cdot I = A$.

The set of $N \times N$ matrices, $\mathfrak{M}_{N \times N}(\mathbb{F})$ with the matrix product is an associative unital but non commutative algebra over $\mathbb{F}$. The same holds for the set of operators $B_{\text{op.}}(X, X)$ with the composition. For convenience, we supress from now on the $\cdot$ in the matrix product.

Theorem 6.1 (**The left and right inverses of a matrix are equal**). *Let A be a $N \times N$ matrix. The left inverse of A , if it exist, is the matrix L such that $LA = I$, and the right inverse is the matrix R such that $AR = I$. If the left inverse of a matrix exists then the right inverse of the matrix exists and the two are equal.*

In any associative unital algebra, if both the left and the right inverses exist then they are equal.

Proof. Consider a finite dimensional vector space V , and $\Phi : V \rightarrow V$ injective ($\Phi(v_1) = \Phi(v_2) \Rightarrow v_1 = v_2 \Rightarrow \ker(\Phi) = 0$). By the rank nullity theorem, $\dim(\text{Im}(\Phi)) = \dim V$ hence $\text{Im}(\Phi) = V$ and Φ is surjective.

Consider A a matrix with left inverse L . We define the linear operator $\Phi : \mathfrak{M}_{N \times N}(\mathbb{F}) \rightarrow \mathfrak{M}_{N \times N}(\mathbb{F})$, $\Phi(X) = AX$. Then $\Phi(X) = 0 \Leftrightarrow AX = 0 \Leftrightarrow LAX = X = 0$, hence $\dim(\ker(\Phi)) = 0$ and Φ is surjective and there exist R such that $AR = I$.

If both the left and right inverses exists then $L = L(AR) = (LA)R = R$. $\square$

Note that this is trivial when recalling that a matrix represents an operator in some basis. By the rank nullity theorem an operator $\mathcal{A} : X \rightarrow X$ is surjective if and only if it is injective (it has zero kernel) and in this case it is invertible.

6.2 Change of basis

Vectors and linear operators are intrinsic notions and do not depend on a particular basis, but their representations (components of vectors, matrix elements and so on) do. For instance the matrix associated to an operator $\mathcal{A} : X \rightarrow Y$ in some bases has entries $a_{ij} = \boldsymbol{\mu}^i(\mathcal{A}\mathbf{e}_j)$ hence depends on the basis vector $\mathbf{e}_i$ and forms $\boldsymbol{\mu}^i$.

Let us assume we have a basis $\mathbf{e}_i$ in X and we select a new basis $\tilde{\mathbf{e}}_i$. As both $\mathbf{e}_i$ and $\tilde{\mathbf{e}}_i$ are bases, we can decompose the vectors in the new basis as linear combinations of the vectors in the old one:

$$\tilde{\mathbf{e}}_i = \sum_{j=1}^N \mathbf{e}_j S_{ij}^j, \quad \tilde{\mathbf{e}}_i = \begin{pmatrix} S_{i1}^1 \\ \vdots \\ S_{iN}^N \end{pmatrix}, \quad S = \begin{pmatrix} \tilde{\mathbf{e}}_1 & \dots & \tilde{\mathbf{e}}_N \end{pmatrix},$$

where the last notation signifies that the matrix S is just the matrix obtained by juxtaposing the column vectors $\tilde{\mathbf{e}}_j$ corresponding to the vector $\tilde{\mathbf{e}}_j$ decomposed in the basis $\mathbf{e}_i$. We can of course expand $\mathbf{e}_i$ on the basis $\tilde{\mathbf{e}}_i$, $\mathbf{e}_i = \sum_j \tilde{\mathbf{e}}_j Q_{ji}^j = \sum_{j,l} \mathbf{e}_l S_{jl}^l Q_{ji}^j$ and, as the $\mathbf{e}_i$ s are linearly independent, we conclude that $Q = S^{-1}$, hence in particular S is invertible.

The dual basis changes accordingly (where all the sums are understood to run from 1 to N):

$$\tilde{\boldsymbol{\theta}}^i = \sum_j (S^{-1})^i_j \boldsymbol{\theta}^j, \quad \tilde{\boldsymbol{\theta}}^i(\tilde{\mathbf{e}}_k) = \sum_j (S^{-1})^i_j \boldsymbol{\theta}^j \left(\sum_l \mathbf{e}_l S_{lk}^l \right) = \sum_{j,l} (S^{-1})^i_j \delta_l^j S_{lk}^l = \delta_k^i.$$

The components of a vector $\sum_i v^i \mathbf{e}_i$ and of a form $\sum_i \varphi_i \boldsymbol{\theta}^i$, with $v^i = \boldsymbol{\theta}^i(v)$ and $\varphi_i = \varphi(\mathbf{e}_i)$ transform as:

$$\tilde{v}^i = \tilde{\boldsymbol{\theta}}^i(v) = \sum_j (S^{-1})^i_j \boldsymbol{\theta}^j(v) = \sum_j (S^{-1})^i_j v^j, \quad \tilde{\varphi}_i = \varphi(\tilde{\mathbf{e}}_i) = \sum_j \varphi(\mathbf{e}_j) S_{ij}^j = \sum_j \varphi_j S_{ij}^j,$$

and we say that φ_i transforms *covariantly*, as it transforms like the basis elements $\mathbf{e}_i$, while $\boldsymbol{\theta}^i$ and v^i transform *contravariantly*. The matrix elements of an operator $\mathcal{A} : X \rightarrow X$ transform under a change of basis as:

$$\tilde{A}_{ij}^i = \tilde{\boldsymbol{\theta}}^i(\mathcal{A}\tilde{\mathbf{e}}_j) = \sum_l (S^{-1})^i_l \boldsymbol{\theta}^l \left(\mathcal{A} \sum_q \mathbf{e}_q S_{jq}^q \right) = \sum_{l,a} (S^{-1})^i_l A_{la}^l S_{jq}^q = [S^{-1}AS]^i_j.$$

Note that, consistently, $\tilde{A}\tilde{v} = (S^{-1}AS)S^{-1}v = S^{-1}(Av)$.

Two matrices related by a transformation $\tilde{A} = S^{-1}AS$, for S some invertible matrix are called similar, and similar matrices correspond to the same operator expressed in different bases. The matrix of an operator $\mathcal{A} : X \rightarrow Y$ changes under a change of bases in both X and Y as $\tilde{A} = T^{-1}AS$.

Definition 6.2 (The trace). The trace of a square matrix is the sum of its diagonal elements, $\text{Tr}(A) = \sum_i A_{ii}^i = \sum_i a_{ii}$. The trace is cyclical $\text{Tr}(AB) = \text{Tr}(BA)$ and it is invariant under a change of basis, $\text{Tr}(S^{-1}AS) = \text{Tr}(A)$.

As it is common in the literature, when no risk of confusion arises, we will conflate the operators with their matrix representation in some basis. We will then talk about the kernel of a matrix consisting in vectors in the kernel of the corresponding operator expressed in the appropriate basis, and so on, of course keeping in mind that the matrix is basis dependent.

In order to simplify the notation, from now on we write the matrix indices in subscript.

6.3 Gaussian elimination

Matrices can be brought into a standard form by linear manipulations. The algorithm we present below is used to determine the kernel and image of a matrix, to solve system of linear equations with constant coefficients and to compute the inverse of a matrix (if it exists).

Consider $A \in \mathfrak{M}_{M \times N}(\mathbb{F})$ a $M \times N$ matrix. The process of Gaussian elimination simplifies the matrix thorough a sequence of the following steps:

- swapping two rows, say i and j of the matrix A , that is $A \rightarrow SA$ with $S_{nn} = 1$ if $n \neq i, j$, $S_{ij} = S_{ji} = 1$ and all the other entries set to 0.
- multiplying one row by a non zero constant, $a_{i\bullet} \rightarrow \alpha a_{i\bullet}$ (the dot signifies that we apply the operations along the entire row), that is $A \rightarrow D_{\alpha}^i A$ with $(D_{\alpha}^i)_{kl} = \delta_{kl} + (\alpha - 1)\delta_{ik}\delta_{il}$:

$$(D_{\alpha}^i A)_{kq} = \sum_l (\delta_{kl} + (\alpha - 1)\delta_{ik}\delta_{il})a_{lq} = a_{kq} + (\alpha - 1)\delta_{ik}a_{iq}.$$

- taking a linear combination of two rows, $a_{i\bullet} \rightarrow a_{i\bullet} + \alpha a_{j\bullet}$, that is $A \rightarrow L_{\alpha}^{i,j} A$ with $(L_{\alpha}^{i,j})_{kl} = \delta_{kl} + \alpha\delta_{ik}\delta_{jl}$:

$$(L_{\alpha}^{i,j} A)_{kq} = \sum_l (\delta_{kl} + \alpha\delta_{ik}\delta_{jl})a_{lq} = a_{kq} + \alpha\delta_{ik}a_{jq}.$$

Exercise 6.1. Prove that $S^2 = D_{\alpha}^i D_{-\alpha}^i = L_{\alpha}^{i,j} L_{-\alpha}^{i,j} = I$.

As all the elementary operations are invertible, the matrix obtain after any sequence of such operations can be written as $B = FA$ for F some invertible $M \times M$ matrix.

Construction 6.1 (**Gaussian elimination**). Using a combination of these operations one reduces the matrix as follows:

- if an entry on the first column is non zero, by permuting the rows we can always bring it in the position a_{11} . For all the rows with $i > 1$, we take the row linear combination $a_{i\bullet} \rightarrow a_{i\bullet} - \frac{a_{i1}}{a_{11}}a_{1\bullet}$. This sets all the entries on the first column a_{i1} with $i > 1$ to 0 and we pass to the next column.
- when we reached the column k , we only look at the rows with $i \geq k$. If any of these rows has a non zero entry in the column k , by row permutations we can bring it in the position a_{kk} . For all the rows below, that is with $i > k$, we take the linear combination $a_{i\bullet} \rightarrow a_{i\bullet} - \frac{a_{ik}}{a_{kk}}a_{k\bullet}$, which sets all $a_{ik} = 0$ for $i > k$ and we pass to the next column.
- at the end of these sequence of steps we obtain the row echelon form of the matrix, which is such that:
 - all the rows with only zero entries are at the bottom.
 - the left-most nonzero entry in every nonzero row (called the leading entry, or the pivot) is on the right of the leading entry of every row above.
- we use rescalings to set all the pivot entries to 1, and linear combinations of the rows to set to zero all the elements above the pivots. This leads to the reduced row echelon form (RREF) of the matrix, which is unique.

We call dominant the columns in the RREF which contain the pivots, and non dominant the other ones.

An example of Gaussian elimination applied to a matrix goes as follows:

$$\begin{pmatrix} 1 & 3 & 1 & 9 \\ 1 & 1 & -1 & 1 \\ 3 & 11 & 5 & 35 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 3 & 1 & 9 \\ 0 & -2 & -2 & -8 \\ 0 & 2 & 2 & 8 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 3 & 1 & 9 \\ 0 & -2 & -2 & -8 \\ 0 & 0 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -2 & -3 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & -2f_1 + f_2 & -3f_1 + 4f_2 \end{pmatrix},$$

where we denoted f_1, f_2, f_3 the standard basis in $\mathbb{R}^3$.

Theorem 6.2 (**The RREF of a matrix is unique**). *The RREF of a matrix A is unique.*

Proof. The dominant columns in a RREF are the elements in the standard basis $f_1, \dots, f_M$ in $\mathbb{R}^M$ encountered in this order. The non dominant columns are linear combinations of the dominant columns which situated at their left, with coefficients exactly the entries in the column. It follows that a RREF writes on columns like:

$$\begin{pmatrix} f_1 & \dots & f_k & \sum_{i=1}^k \alpha_{ik+1} f_i & \sum_{i=1}^k \alpha_{ik+2} f_i & \dots & f_l & \dots \end{pmatrix}.$$

If a column is identically zero, we see it as a linear combination with zero coefficients of the columns at its left. In particular, any matrix which can be obtain from B as a product FB with F some invertible matrix will have the form:

$$\begin{pmatrix} Ff_1 & \dots & Ff_k & \sum_{i=1}^k \alpha_{ik+1} Ff_i & \sum_{i=1}^k \alpha_{ik+2} Ff_i & \dots & Ff_l & \dots \end{pmatrix}.$$

If B' is a RREF of A , then $B' = FB$ and its dominant columns must reproduce $f_1, \dots, f_M$ in $\mathbb{R}^M$ in this order. Thus the first k dominant columns must respect $Ff_1 = f_1, Ff_2 = f_2$, up to $f_k = Ff_k$. Then the first non dominant column is $\sum_{i=1}^k \alpha_{ik+1} Ff_i$ which equals $\sum_{i=1}^k \alpha_{ik+1} f_i$ and iterating we conclude that $B' = B$. $\square$

6.3.1 Finding bases in the kernel and the image of a matrix

The columns of the matrix A can also be interpreted as the column vectors Ae_i , with $e_1, \dots, e_N$ the standard basis in $\mathbb{R}^N$. As $A = FB$ with F invertible and B the RREF of A , it follows that:

$$\begin{pmatrix} Ae_1 & \dots & Ae_N \end{pmatrix} = \begin{pmatrix} Ff_1 & \dots & Ff_k & \sum_{i=1}^k \alpha_{ik+1} Ff_i & \sum_{i=1}^k \alpha_{ik+2} Ff_i & \dots & Ff_l & \dots \end{pmatrix},$$

and we conclude that:

- a basis in the image of A , which is the span of its columns $Ae_1, \dots, Ae_N$, is $Ff_1, Ff_2, \dots, Ff_l, \dots$ (from the RREF it is clear these vector are spanning, and they are linear independent as F is invertible). A basis in the image is given by the dominant columns of A , that is the columns which contain the pivots after reduction to the RREF.
- each non dominant column of A yields a relation $Ae_j = \sum_{i < j} \alpha_{ij} Ae_i$, hence a non trivial vector $v_j = e_j - \sum_{i < j} \alpha_{ij} e_i \in \ker(A)$. All such vectors are linearly independent. Indeed, denoting J the set of indices of the non dominant columns, let us assume that:

$$\sum_{j \in J} \gamma_j v_j = \sum_{j \in J} \gamma_j \left(e_j - \sum_{i < j} \alpha_{ij} e_i \right) = 0.$$

Denoting $j_{\max}$ the maximal index in J , $e_{j_{\max}}$ is multiplied only by $\gamma_{j_{\max}}$ in this linear combination, hence $\gamma_{j_{\max}} = 0$ as the e_j s are linearly independent. Iterating we conclude that $\gamma_j = 0$. As we have built $N - \dim(\text{Im } A)$ linearly independent vectors in $\ker(A)$, they form a basis by the rank-nullity theorem.

Consider for instance the matrix A with RREF B as before:

$$A = \begin{pmatrix} 1 & 3 & 1 & 9 \\ 1 & 1 & -1 & 1 \\ 3 & 11 & 5 & 35 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & -2 & -3 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & -2f_1 + f_2 & -3f_1 + 4f_2 \end{pmatrix},$$

where we denoted f_1, f_2, f_3 the standard basis in $\mathbb{C}^3$. We have the bases:

$$\text{Im}(A): \quad Ae_1 = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \quad Ae_2 = \begin{pmatrix} 3 \\ 1 \\ 11 \end{pmatrix}, \quad \ker(A): \quad e_3 + 2e_1 - e_2 = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \quad e_4 + 3e_1 - 4e_2 = \begin{pmatrix} 3 \\ -4 \\ 0 \\ 1 \end{pmatrix}.$$

Indeed, $e_3 + 2e_1 - e_2$ and $e_4 + 3e_1 - 4e_2$ are linearly independent and:

$$\begin{pmatrix} 1 & 3 & 1 & 9 \\ 1 & 1 & -1 & 1 \\ 3 & 11 & 5 & 35 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 - 3 + 1 \\ 2 - 1 - 1 \\ 6 - 11 + 5 \end{pmatrix}, \quad \begin{pmatrix} 1 & 3 & 1 & 9 \\ 1 & 1 & -1 & 1 \\ 3 & 11 & 5 & 35 \end{pmatrix} \begin{pmatrix} 3 \\ -4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 - 12 + 9 \\ 3 - 4 + 1 \\ 9 - 44 + 35 \end{pmatrix}.$$

One can reconstruct $B = L_{-3}^{1;2} D_{-\frac{1}{2}}^2 L_1^{3;2} L_{-3}^{3;1} L_{-1}^{2;1} A$.

6.3.2 Solving systems of linear equations with constant coefficients

Any system of M linear equations with N unknowns and constant coefficients:

$$a_{11}x^1 + \cdots + a_{1N}x^N = y^1, \quad a_{21}x^1 + \cdots + a_{2N}x^N = y^2, \quad \dots, \quad a_{M1}x^1 + \cdots + a_{MN}x^N = y^M,$$

can be written as a matrix equation $Ax = y$ with:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \\ a_{M1} & \cdots & a_{MN} \end{pmatrix}, \quad x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^N \end{pmatrix}, \quad y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^M \end{pmatrix}.$$

There are three cases:

- $y \notin \text{Im}(A)$. In this case the system has no solutions (the system is incompatible).
- $y \in \text{Im}(A)$ and $\ker(A) = 0$. In this case the system has a unique solution (the system is compatible determined).
- $y \in \text{Im}(A)$ and $\ker(A) \neq 0$. In this case the system has an infinity of solutions (the system is compatible undermined).
The set of solutions $\{x, Ax = y\}$ is an affine space over $\ker(A)$, that is given x_0 a particular solution of $Ax_0 = y$, $\{x, Ax = y\}$ is the affine subspace $x_0 + \ker(A)$ of $\mathbb{C}^N$.

We define the *augmented matrix* of the system $[A|y]$ as the matrix consisting in the matrix A with an additional column y . In order to determine in which of the three cases we are (and solve the system if possible), we observe that x is a solution of $Ax = y$ if and only if x is a solution of $RAx = Ry$ for any invertible R , that is x is a solution of the system with augmented matrix $[RA|Ay]$. The system becomes trivial if $[A|y]$ is brought in RREF, $[B|Y]$ namely:

- if the last column, Y , is dominant, then in the system of equations corresponding to the RREF one of the equations reads $0 = 1$, hence the system is incompatible.
- if the last column, Y , is non dominant, then:
 - either $M = N$ and the RREF of A is the identity, $B = I$. In this case the system has a unique solution $x = Y$.

- or the non dominant columns in B lead to a basis in $\ker(A)$ and a particular solution is $x_0^{j(i)} = Y^i$, where $j(i)$ is the label of the i 'th dominant column, and we complete x_0 with 0 on all other entries.

As examples, we consider the systems of equations with augmented matrices (we eliminate the vertical line separating y in order to simplify the notation) below:

$$\begin{pmatrix} 1 & 1 & -3 & 4 \\ 2 & 1 & -1 & 2 \\ 3 & 2 & -4 & 7 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & -3 & 4 \\ 0 & -1 & 5 & -6 \\ 0 & -1 & 5 & -5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & -3 & 4 \\ 0 & -1 & 5 & -6 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The last column is dominant and the system incompatible (no solution).

$$\begin{pmatrix} 1 & 1 & -3 & 4 \\ 2 & 1 & -1 & 2 \\ 3 & 2 & -4 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & -3 & 4 \\ 0 & -1 & 5 & -6 \\ 0 & -1 & 5 & -6 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & -3 & 4 \\ 0 & -1 & 5 & -6 \\ 0 & 0 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 2 & -2 \\ 0 & 1 & -5 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

A basis in $\ker(A)$ is $e_3 - 2e_1 + 5e_2$, and the general solution writes:

$$x = \begin{pmatrix} -2 \\ 6 \\ 0 \end{pmatrix} + t \begin{pmatrix} -2 \\ 5 \\ 1 \end{pmatrix}.$$

$$\begin{pmatrix} 1 & 2 & 4 & 23 \\ 1 & 1 & 1 & 7 \\ 4 & 2 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 4 & 23 \\ 0 & -1 & -3 & -16 \\ 0 & -6 & -15 & -90 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 4 & 23 \\ 0 & 1 & 3 & 16 \\ 0 & 0 & 3 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 4 & 23 \\ 0 & 1 & 3 & 16 \\ 0 & 0 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -2 & -9 \\ 0 & 1 & 3 & 16 \\ 0 & 0 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & -5 \\ 0 & 1 & 0 & 10 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

The solution of the system is unique:

$$x = \begin{pmatrix} -5 \\ 10 \\ 2 \end{pmatrix}.$$

$$\begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 2 & 1 & -3 & 0 & 2 \\ 5 & -2 & 0 & -3 & 5 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 3 & -5 & 2 & 0 \\ 0 & 3 & -5 & 2 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -\frac{5}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -\frac{2}{3} & -\frac{1}{3} & 1 \\ 0 & 1 & -\frac{5}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

The solution of the system is:

$$x = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + t_1 \begin{pmatrix} \frac{2}{3} \\ \frac{5}{3} \\ 1 \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 0 \\ 1 \end{pmatrix}.$$

6.3.3 Computing the matrix inverse

Let A be a $N \times N$ matrix. The inverse of A is such that $AC = I$ or, looking at the columns, the column vectors of the inverse solve the equations $AC_i = e_i$, with e_i the i 'th basis vector in the standard basis. The augmented matrix of these N

systems of N equations is $[A|I]$, where the i 'th column of the identity matrix is e_i . All these systems can be simultaneously solved if and only if the RREF of A is I and in this case the RREF of $[A|I]$ is $[I|C]$. For instance:

$$\begin{pmatrix} 1 & 2 & 4 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 4 & 2 & 1 & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 4 & 1 & 0 & 0 \\ 0 & -1 & -3 & -1 & 1 & 0 \\ 0 & -6 & -15 & -4 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -2 & -1 & 2 & 0 \\ 0 & 1 & 3 & 1 & -1 & 0 \\ 0 & 0 & 3 & 2 & -6 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & -2 & -1 & 2 & 0 \\ 0 & 1 & 3 & 1 & -1 & 0 \\ 0 & 0 & 1 & \frac{2}{3} & -2 & \frac{1}{3} \end{pmatrix} \\ \rightarrow \begin{pmatrix} 1 & 0 & 0 & \frac{1}{3} & -2 & \frac{2}{3} \\ 0 & 1 & 0 & -1 & 5 & -1 \\ 0 & 0 & 1 & \frac{2}{3} & -2 & \frac{1}{3} \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 4 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & -2 & \frac{2}{3} \\ -1 & 5 & -1 \\ \frac{2}{3} & -2 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Note that A is invertible if and only if $\ker(A) = 0$, that is the only solution of $Ax = 0$ is $x = 0$. If A is not invertible, then $Ax = 0$ admits a non trivial solution.

6.4 The determinant

The determinant of a $N \times N$ matrix is a degree N homogeneous polynomial in the matrix entries with some special properties.

We denote S_n the symmetric group, that is the group of permutations of n elements with group operation the composition:

$$S_n = \{ \sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}, \sigma \text{ bijection} \}, \quad |S_n| = n!.$$

The composition of permutations is also called the product of permutations. A permutation is sometimes denoted in a two line notation:

$$\sigma = \begin{pmatrix} 1 & 2 & \dots & N \\ \sigma(1) & \sigma(2) & \dots & \sigma(N) \end{pmatrix},$$

and a rearrangement (permutation) of the columns denotes the same permutation.

Any permutation can be decomposed into disjoint cycles $\sigma = (p, \sigma(p), \sigma^2(p) \dots)(q, \sigma(q) \dots) \dots$, where each cycle collects all the iterated images $p, \sigma(p), \dots$ in the order in which they are encountered. A fixed point $\sigma(p) = p$ is denoted (p) . A cyclical permutation of the elements in a cycle leads to the same cycle $(p, \sigma(p) \dots \sigma^{-1}(p)) = (\sigma^{-1}(p)p\sigma(p) \dots)$. For instance

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} = (1)(2)(34), \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = (12)(34), \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 4 \end{pmatrix} = (123)(4), \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} = (1234).$$

We denote d_i the number of cycles of length i in the cycle decomposition of σ , and the cycle type of σ is the collection $\{d_i\}_{i \geq 1}$. Two permutations have the same cycle type $\{d_i\}_{i \geq 1}$ if and only if they are conjugated $\sigma_1 = \pi \sigma_2 \pi^{-1}$ for some $\pi \in S_n$.

Exercise 6.2. Prove that two permutations have the same cycle type if and only if they are conjugated. If $\sigma_2(i) = j$ and $\sigma_1 = \pi \sigma_2 \pi^{-1}$

then $\sigma_1(\pi(i)) = \pi(j)$. Listing the cycles of σ_2 and the replacing all the a by $\pi(a)$ we obtain the cycles of σ_1 . Conversely, we list the cycles of σ_2 on a line and the cycles of σ_1 immediately below, making sure that the cycles match in length top and bottom, interpreting this as a two line notation for a permutation π , $\sigma_1 = \pi \sigma_2 \pi^{-1}$.

Exercise 6.3. Prove that there are $\frac{n!}{\prod_i d_i! i^{d_i}}$ distinct permutations with a cycle type $\{d_i\}_{i \geq 1}$. Prove that $\sum_{d_i \geq 0}^{\sum i d_i = n} \frac{n!}{\prod_i d_i! i^{d_i}} = n!$.

We list the permutation τ in cycle notation. The $n!$ permutations π of the labels (balls) while keeping the cycles of τ (boxes) fixed generates all the permutations of the same cycle type. However, the permutations are over counted: a permutation π whose only effect is to permute circularly the labels in a cycle of τ lead to the same τ , and a permutation π whose only effect is to permute all the labels in a cycle with all the labels in another cycle of the same length preserving their order also leads to the same τ , hence τ is generated $\prod_i d_i! i^{d_i}$ times by the $n!$ permutations of labels π . For the second part:

$$\sum_{d_i \geq 0}^{\sum i d_i = n} \frac{n!}{\prod_i d_i! i^{d_i}} = n! \left[\sum_{d_i \geq 0} \prod_{i \geq 1} \frac{1}{d_i! i^{d_i}} x^{d_i} \right]_{x^n} = n! \left[e^{-\ln(1-x)} \right]_{x^n} = n!,$$

where $[\dots]_{x^n}$ denotes the coefficient of x^n in a formal power series.

Exercise 6.4. Prove that there are $\frac{n!}{\prod_i d_i!(i!)^{d_i}}$ ways to distribute N labelled balls into d_1 boxes of size 1, d_2 boxes of size 2 and so on. As before, with the exception that any permutation of the labels of all the balls in the same box leads to the same assignment of balls to boxes.

A transposition is a permutation with only one cycle of length two, and $n-2$ cycles of length 1. Any permutation can be written as a product, that is a composition, of transpositions. Indeed, any permutation is a product of its cycles completed by fixed points and any cycle writes as a product of transpositions (where the product is read left to right and we do not write the fixed points in the transpositions):

$$(a_1 \ a_2 \ a_3 \ \dots \ a_i) = (a_1 \ a_2) \cdot (a_2 \ a_3) \cdot \dots \cdot (a_{i-1} \ a_i).$$

A cycle of length i needs $i-1$ transpositions, and a permutation of a give cycle type needs $\sum_i d_i(i-1)$. The signature of the permutation $\epsilon(\sigma) = (-1)^{\sum_i d_i(i-1)}$ tracks the parity of the number of transpositions in such a decomposition. Equivalently, a permutation is odd (signature -1) if the number of cycles of even length of the permutation is odd and it is even (signature $+1$) otherwise. The same permutation can be decomposed in transpositions in many different ways, but the parity of the number of transposition is the same in all such decompositions.¹

The signature of a product of permutation is the product of signatures of the permutations (as we can first decompose both permutations in transpositions and then multiply to obtain a product of transpositions).

Exercise 6.5. Prove that $\epsilon(\tau) = \epsilon(\tau^{-1})$. τ and τ^{-1} have the same cycle type.

Exercise 6.6. Prove that a permutation can be written as the product of its cycles, completed by fixed points, for instance $(12)(34) = (12)(3)(4) \cdot (1)(2)(34)$.

Definition 6.3 (**The determinant**). The determinant of a square $N \times N$ matrix $A \in \mathfrak{M}_{N \times N}(\mathbb{F})$ with matrix elements a_{ij} is the polynomial:

$$\det(A) = \begin{vmatrix} a_{11} & \dots & a_{1N} \\ \vdots & & \\ a_{N1} & \dots & a_{NN} \end{vmatrix} = \sum_{\sigma \in S_N} \epsilon(\sigma) \prod_{i=1}^n a_{i\sigma(i)}.$$

The determinant is the unique function $\det : \mathfrak{M}_{N \times N}(\mathbb{F}) \rightarrow \mathbb{F}$ such that $\det(I) = 1$, is antisymmetric under permutations of the rows of A , linear in each row and zero if two rows coincide.² In detail, denoting $r_1, \dots, r_N$ the rows of the matrix:

$$\begin{vmatrix} r_1 + \tilde{r}_1 \\ r_2 \\ \vdots \\ r_N \end{vmatrix} = \begin{vmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{vmatrix} + \begin{vmatrix} \tilde{r}_1 \\ r_2 \\ \vdots \\ r_N \end{vmatrix}, \quad \begin{vmatrix} \alpha r_1 \\ \alpha r_2 \\ \vdots \\ \alpha r_N \end{vmatrix} = \alpha \begin{vmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{vmatrix}, \quad \begin{vmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_N \end{vmatrix} = (-1)^{i+1} \begin{vmatrix} r_i \\ r_1 \\ \vdots \\ r_N \end{vmatrix}.$$

The same properties hold for columns, as $\det(A) = \det(A^t)$ with the transposed matrix defined as $(A^t)_{ij} = a_{ji}$.³

The minor $\det(A_{\hat{K}\hat{J}})$ of A is the determinant of the matrix obtained from A by erasing the rows with indices in K and columns with indices in J (this is defined only for $|K| = |J|$). Minors can also be defined for $M \times N$ matrices and correspond to determinants of square sub matrices.

¹The group S_n is the group of words with generators $\tau_i = (i, i+1)$ and relations $\tau_i^2 = 1$, $\tau_i \tau_{i+1} \tau_i = \tau_{i+1} \tau_i \tau_{i+1}$ for $i < n$ and $\tau_i \tau_j = \tau_j \tau_i$ if $|i-j| \geq 2$. All the relations preserve the parity of the number of letters of the words.

²Antisymmetry follows from noting that for any τ fixed permutation, $\sum_{\sigma \in S_N} \epsilon(\sigma) \prod_{i=1}^n a_{\tau(i)\sigma(i)} = \epsilon(\tau) \sum_{\sigma \in S_N} \epsilon(\sigma) \prod_{i=1}^n a_{i\sigma(i)}$ and trivially implies the determinant is zero if two rows coincide. Linearity is obvious as the determinant is a homogeneous polynomial of degree 1 in the entries in each row.

³We have $\sum_{\sigma \in S_N} \epsilon(\sigma) \prod_{i=1}^n a_{\sigma(i)i} = \sum_{\sigma \in S_N} \epsilon(\sigma) \prod_{j=1}^n a_{\sigma\sigma^{-1}(j)\sigma^{-1}(j)}$, and we change summation variable to σ^{-1} .

The determinant of a diagonal matrix is just the product of the diagonal entries:

$$\begin{vmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{vmatrix} = \lambda_1 \lambda_2 \dots \lambda_N ,$$

as in this case only the identity permutation contributes. The same holds if the matrix is upper (or lower) triangular:

$$a_{ij} = 0, \quad i > j \Rightarrow \det(A) = \prod_{i=1}^N a_{ii} .$$

Exercise 6.7. Prove that $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$ and $\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{31}a_{22} .$

An important property of the determinant is that it respects the matrix product.

Proposition 6.1 (**The determinant respects the matrix product**). For any two $N \times N$ matrices A and B we have $\det(AB) = \det(A) \det(B)$.

Proof. By linearity of the determinant in the columns we have:

$$\det(AB) = \begin{vmatrix} \sum_{j_1=1}^N a_{1j_1} b_{j_1 1} & \dots & \sum_{j_N=1}^N a_{1j_N} b_{j_N N} \\ \vdots & \ddots & \vdots \\ \sum_{j_1=1}^N a_{Nj_1} b_{j_1 1} & \dots & \sum_{j_N=1}^N a_{Nj_N} b_{j_N N} \end{vmatrix} = \sum_{j_1=1}^N \dots \sum_{j_N=1}^N b_{j_1 1} \dots b_{j_N N} \begin{vmatrix} a_{1j_1} & \dots & a_{1j_N} \\ \vdots & \ddots & \vdots \\ a_{Nj_1} & \dots & a_{Nj_N} \end{vmatrix} .$$

If two of the j s are equal, the last determinant has two identical columns and is zero. Ordering the j 's and taking into account that the determinant is antisymmetric by permutation of the columns we get:

$$\det(AB) = \det(A) \sum_{\sigma \in S_N} \epsilon(\sigma) b_{j_{\sigma(1)} 1} \dots b_{j_{\sigma(N)} N} = \det(A) \det(B) .$$

□

This leads to the following crucial remark.

Corollary 6.1 (**The determinant is basis independent**). Under a change of basis a matrix changes like $\tilde{A} = S^{-1}AS$, and $\det(\tilde{A}) = \det(S^{-1}) \det(A) \det(S) = \det(A) \det(SS^{-1}) = \det(A)$, that is the determinant is invariant under a change of basis.

This might seem surprising at first sight as the matrix elements change under a change of basis, and the determinant is a complicated polynomial in the matrix entries.

6.4.1 The determinant and the inverse of a matrix

Lemma 6.1 (**A is invertible if and only if $\det(A) \neq 0$**). Let us denote B the row echelon form of A . Then

$$\pm \det(A) = \det(B) = \prod_{i=1}^N b_{ii} .$$

We have seen that a matrix is invertible if and only if its RREF is I , or equivalently all the diagonal entries in its row echelon form are non zero $b_{ii} \neq 0$ for $i = 1, \dots, N$, therefore A is invertible if and only if $\det(A) \neq 0$.

Proof. Up to a sign, the determinant is invariant under the linear manipulations which bring A to its row echelon form. This can be checked by observing that the matrices S which encode row permutations and $L_\alpha^{i,j}$ which take linear combinations of rows have determinant ± 1 ($L_\alpha^{i,j}$ is either upper or lower triangular). $\square$

Lemma 6.2 ($\det(A) = 0$ if and only if the columns of A are linearly dependent). Let us denote B the row echelon form of A . $\det(A) = 0$ if and only if B has at least a row of 0, hence at least one non dominant column and a linear relation between the columns.

The determinant gives a criterion of invertibility for matrices. A recursive way to compute the determinant is given by the following theorem.

Theorem 6.3 (Cofactor expansion of determinant). The determinant of a matrix can be computed by expanding along the rows:

$$\det(A) = \sum_{j=1}^N (-1)^{i+j} a_{ij} \det(A_{i\hat{j}}), \quad \forall i = 1, \dots, N,$$

where the cofactor $\det(A_{i\hat{j}})$ is the determinant of the matrix obtained by deleting the row i and the column j in A . The same holds when expanding along columns, as the determinant is invariant under taking the transposed.

Proof. Without loss of generality we fix $i = 1$ and we classify the permutations σ according to $\sigma(1)$:

$$\det(A) = \sum_{j=1}^N \sum_{\sigma \in S_n}^{\sigma(1)=j} \epsilon(\sigma) a_{1j} \prod_{i' \neq 1} a_{i'\sigma(i')}.$$

Each term above is the determinant of a matrix with only one non zero entry on the first, row a_{1j} and all the others set to 0 (that is the determinant is linear in its first row):

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ \vdots & & & \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{vmatrix} = \begin{vmatrix} a_{11} & 0 & \dots & 0 \\ \vdots & & & \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & \dots & 0 \\ \vdots & & & \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{vmatrix} + \dots + \begin{vmatrix} 0 & 0 & \dots & a_{1N} \\ \vdots & & & \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{vmatrix}$$

and by permuting the columns, the j 'th term is:

$$(-1)^{j+1} \begin{vmatrix} a_{1j} & 0 \\ \vdots & A_{1\hat{j}} \end{vmatrix} = (-1)^{j+1} \sum_{\sigma \in S_n} \epsilon(\sigma) \prod_i b_{i\sigma(i)},$$

where we labelled the elements of the new matrix generically by b_{ij} . Now, the only non trivial element on the first row is $b_{11} = a_{1j}$, hence the only permutations contributing must have $\sigma(1) = 1$. Such a permutation consists in a cycle (1) and a permutation τ on $\{2, \dots, n-2\}$, with $\epsilon(\tau) = \epsilon(\sigma)$, hence

$$(-1)^{j+1} \sum_{\sigma \in S_n} \epsilon(\sigma) \prod_i b_{i\sigma(i)} = (-1)^{j+1} a_{1j} \sum_{\tau \in S_{n-1}} \epsilon(\tau) \prod_{i=2}^n b_{i\tau(i)} = (-1)^{j+1} a_{1j} \det(A_{1\hat{j}}).$$

$\square$

This cofactor expansion allows us to give an explicit formula for the inverse A^{-1} of any matrix A with $\det A \neq 0$.

Corollary 6.2 (**Inverse of a matrix**). If A is a $N \times N$ matrix with $\det(A) \neq 0$, then A is invertible, that is there exist a matrix A^{-1} such that $(AA^{-1})_{ik} = (A^{-1}A)_{ik} = \delta_{ik}$. The inverse matrix has matrix elements:

$$(A^{-1})_{ij} = \frac{1}{\det(A)} (-1)^{i+j} \det(A_{\hat{j}\hat{i}}),$$

where we stress that the indices of the cofactor are permuted. If A represents the operator $\mathcal{A}$ in some basis, the inverse matrix A^{-1} represents the inverse operator $\mathcal{A}^{-1}$ in the same basis.

Proof. We have:

$$(AA^{-1})_{ik} = \sum_{j=1}^N a_{ij} (A^{-1})_{jk} = \frac{1}{\det(A)} \sum_{j=1}^N a_{ij} (-1)^{k+j} \det(A_{\hat{k}\hat{j}}).$$

By the cofactor expansion, if $j = k$, then the right hand side is 1. If $j \neq k$, then

$$\sum_{j=1}^N a_{ij} (-1)^{k+j} \det(A_{\hat{k}\hat{j}}),$$

is nothing but the determinant of the matrix A , where the k 'th row has been replaced by a copy of the i 'th row. As this matrix has two identical rows, its determinant is zero. $\square$

A matrix is invertible if and only if it has zero kernel, that is $\ker A = 0 \Leftrightarrow \det A \neq 0$.

Corollary 6.3 (**Cramer's rule**). Let A be an invertible $N \times N$ matrix. Then the system of equations $Ax = y$ has a unique solution $x = A^{-1}y$ and, using the formula for the inverse and the cofactor expansion of the determinant we have:

$$x^i = \frac{1}{\det A} \sum_{j=1}^N (-1)^{i+j} \det(A_{\hat{j}\hat{i}}) y^j = \frac{\det(A^i)}{\det A},$$

where the matrix A^i is obtained from A by replacing the column i with the column vector y .

6.4.2 Block matrices

Often one deals which are partitioned by some criterion into blocks. Consider four matrices: A of sizes $M \times M$, B of size $M \times N$, C of size $N \times M$ and D of size $N \times N$ which are organized in a $(M+N) \times (M+N)$ matrix:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Proposition 6.2 (**Determinant and inverse of a block matrix**). If $B = 0$ then the determinant of the block matrix is:

$$\begin{vmatrix} A & 0 \\ C & D \end{vmatrix} = \det(A) \det(D),$$

and similarly if $C = 0$. If A is invertible then:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \det(A) \det(D - CA^{-1}B),$$

and if D is invertible it is $\det(D)\det(A - BD^{-1}C)$. If both A and D are invertible, then the inverse of the block matrix is:

$$\begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix} \\ \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

Proof. Assume the block $B = 0$. Then, as the entries i, j for $i \in \{1, \dots, M\}$ and $j \in \{M + 1, \dots, M + N\}$ in the block matrix are zero, the permutations which contribute to the evaluation of the determinant of the block matrix must respect $\sigma(\{1, \dots, M\}) \subset \{1, \dots, M\}$ and consequently $\sigma(\{M + 1, \dots, M + N\}) \subset \{M + 1, \dots, M + N\}$. It follows that σ factors into two permutations each acting on one of the subsets of indices leading to:

$$\begin{vmatrix} A & 0 \\ C & D \end{vmatrix} = \det(A)\det(D),$$

and similarly if $C = 0$.

If A is invertible, then, denoting I_N and I_M the $N \times N$ respectively $M \times M$ identity matrix we have:

$$\begin{aligned} \begin{vmatrix} A & B \\ C & D \end{vmatrix} &= \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \det \begin{pmatrix} I_M & -B \\ 0 & I_N \end{pmatrix} = \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \det \begin{pmatrix} A & 0 \\ 0 & I_N \end{pmatrix} \det \begin{pmatrix} A^{-1} & -A^{-1}B \\ 0 & I_N \end{pmatrix} \\ &= \det(A)\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \det \begin{pmatrix} A^{-1} & -A^{-1}B \\ 0 & I_N \end{pmatrix} = \det(A)\det \begin{pmatrix} I_M & 0 \\ CA^{-1} & D - CA^{-1}B \end{pmatrix} = \det(A)\det(D - CA^{-1}B) \end{aligned},$$

and the corresponding manipulation works if D invertible. The last part is just a direct check. $\square$

6.5 Minors of a matrix

Let us consider a $M \times N$ matrix A , where M can equal N or not. We denote $A_{IJ} = (a_{ij})_{i \in I, j \in J}$, where I is a set of row indices and J a set of column indices the sub-matrix consisting in the rows with indices in I and the columns with indices in J . A_{IJ} is a square matrix if $|I| = |J|$, and in this case the determinant $\det(A_{IJ}) \equiv [A_{IJ}]$ is called a minor of the matrix A . If A is a square matrix, A_{II} are called the *principal minors* of A . A matrix has many minors and some of them can be zero.

Proposition 6.3 (Minors and rank). *The rank of A is the dimension of the image of A , $\text{rank}(A) = \dim(\text{Im}(A))$. The matrix A has rank r if and only if it has a nonzero $r \times r$ minor.*

Proof. If r columns of A are linearly dependent then all the $r \times r$ minors of A containing these columns are zero (as their columns are linearly dependent). Conversely (up to a permutation of the columns) if all the $r \times r$ minors of A containing the first r columns are zero, bringing A in its RREF:

$$(Ae_1 \quad \dots \quad Ae_N) = (Ff_1 \quad \dots \quad Ff_k \quad \sum_{i=1}^k \alpha_{ik+1}Ff_i \quad \sum_{i=1}^k \alpha_{ik+2}Ff_i \quad \dots \quad Ff_l \quad \dots),$$

with F is invertible we conclude that $r > k$ hence the r columns are linearly dependent. $\square$

The determinant of a product of rectangular matrices can be computed in terms of the minors.

Theorem 6.4 (Cauchy-Binet formula). *Let A be an $M \times N$ matrix and B be a $N \times M$ matrix with $M \leq N$. Let $1 \leq j_1 < \dots < j_M \leq N$ and denote $A_{j_1 \dots j_M}$ the $M \times M$ sub-matrix of A with column indices $j_1, \dots, j_M$ and $B_{j_1 \dots j_M}$ the $M \times M$*

sub-matrix of B with row indices $j_1 \dots j_M$. Then:

$$\det(AB) = \sum_{1 \leq j_1 < \dots < j_M \leq N} \det(A_{j_1 \dots j_M}) \det(B_{j_1 \dots j_M}).$$

If $M = N$ this reduces to $\det(AB) = \det(A) \det(B)$.

Proof. The proof parallels the case $N = M$. By linearity of the determinant in the columns we have:

$$\det(AB) = \begin{vmatrix} \sum_{j_1=1}^N a_{1j_1} b_{j_1 1} & \dots & \sum_{j_M=1}^N a_{1j_M} b_{j_M M} \\ \vdots & \ddots & \vdots \\ \sum_{j_1=1}^N a_{Mj_1} b_{j_1 1} & \dots & \sum_{j_M=1}^N a_{Mj_M} b_{j_M M} \end{vmatrix} = \sum_{j_1=1}^N \dots \sum_{j_M=1}^N b_{j_1 1} \dots b_{j_M M} \begin{vmatrix} a_{1j_1} & \dots & a_{1j_M} \\ \vdots & \ddots & \vdots \\ a_{Mj_1} & \dots & a_{Mj_M} \end{vmatrix}.$$

If two of the j s are equal, the last determinant has two identical columns and is zero. Ordering the j 's and taking into account that the determinant is antisymmetric by permutation of the columns we get:

$$\det(AB) = \sum_{1 \leq j_1 < \dots < j_M \leq N} \det(A_{j_1 \dots j_M}) \sum_{\sigma \in S_M} \epsilon(\sigma) b_{j_{\sigma(1)} 1} \dots b_{j_{\sigma(M)} M} = \sum_{1 \leq j_1 < \dots < j_M \leq N} \det(A_{j_1 \dots j_M}) \det(B_{j_1 \dots j_M}).$$

□

As a direct consequence, the minors of a product of rectangular matrices can also be computed.

Corollary 6.4. Let A be an $M \times N$ matrix and B be a $N \times P$ matrix with $M \leq N$ and let $I \subset \{1, \dots, M\}$ respectively $J \subset \{1, \dots, P\}$ with $|I| = |J|$. Then:

$$\det[(AB)_{IJ}] = \sum_K \det(A_{IK}) \det(B_{KJ}),$$

where K runs over all the subsets of $\{1, \dots, N\}$ with $|K| = |I| = |J|$.

Classifying the permutations by their fixed points one can write a diagonal expansion of the determinant.

Lemma 6.3 (**Diagonal expansion of the determinant**). Let A be a $N \times N$ matrix and let $I \subset \{1, \dots, N\}$. We denote $A_{\hat{I}\hat{I}}$ the sub matrix of A obtained from A by deleting the rows and columns with indices in the set I . Then:

$$\det(A) = \sum_{I \subset \{1, \dots, N\}} \det(A_{\hat{I}\hat{I}}) \prod_{i \in I} a_{ii}.$$

6.5.1 The matrix tree theorem

Definition 6.4 (**Graph**). A graph $G = (V, E)$ is a finite set of vertices $V = \{1, \dots, N\}$ and a set $E = \{(i, j) | i, j \in V\}$ of edges, that is pairs of vertices.⁴ A subgraph of G is a graph having the same vertices as G , but only some of the edges. A root of the graph is a marked vertex, say the vertex 1.

Examples of graphs are:

$$\begin{aligned} G_1 : N = 3, \quad E &= \{(1, 2), (1, 3), (2, 3)\} \\ G_2 : N = 4, \quad E &= \{(1, 2), (1, 3), (1, 4), (2, 3), (3, 4)\}. \end{aligned}$$

A *path* in a graph is a sequence of edges $e_1, \dots, e_n$ such that any two consecutive edges share a vertex $e_l \cap e_{l+1} \neq \emptyset$. A path can also be identified by the ordered list of vertices $i_0, i_1, \dots, i_n$ it visits, $e_0 = (i_0, i_1), e_1 = (i_1, i_2), \dots, e_n = (i_{n-1}, i_n)$. A path is *simple* if all the intermediate vertices $v_1, \dots, v_n$ are different and a simple path is a *loop* (or *cycle*) if $i_0 = i_n$. Any path can be decomposed into simple paths and cycles.

⁴Self loops and multiple edges are excluded.

Definition 6.5 (**Connected graph, tree**). A graph is connected if any two vertices of the graph can be connected by a path. A graph is a tree if it is connected and has no loops. In a tree any two vertices are connected by only one simple path (if two different simple paths exist, their union contains a loop). We denote $\kappa(G)$ the number of subgraphs of the connected graph G which are trees.

Lemma 6.4. A graph over N vertices is a tree if and only if has no cycle and has exactly $N - 1$ edges.

Proof. Deleting any edge in a tree disconnects the graph, that is it increases the number of connected components of the graph by 1. Indeed, any edge (i, j) partitions the vertices in the tree into two sets

- the vertices in the subset I which are connected to the vertex i by simple paths which do not pass through (i, j) . They are connected to j by adding (i, j) to the paths connecting them to i .
- the vertices in the subset J which are connected to the vertex j by simple paths which do not pass through (i, j) . They are connected to i by adding (i, j) to the paths connecting them to j .

All the simple paths connecting vertices in I to vertices in J pass through (i, j) . Deleting (i, j) cuts all the paths that pass through (i, j) , hence splits the graph into two connected components containing the vertices i and I respectively j and J .

Iterating the edge deletions, the process stops when all the edges in the tree have been deleted and the graph has one vertex per connected component, thus at the end one has exactly N connected components. As T had one connected component, it follows that exactly $N - 1$ edges deletions have been performed.

Conversely, if T has N vertices, $N - 1$ edges and no cycle then T must be a tree. Indeed, assume T is not connected and denote $|E_i|$ and $|V_i|$ the number of edges and vertices in the connected component i , and $\sum_i |E_i| = N - 1$, $\sum_i |V_i| = N$. Then for at least one i , $|E_i| \geq |V_i|$. One can always chose a tree in a connected graph by a greedy algorithm: one iteratively tests at step l one of the untested edges, say e_l . If e_l does not form a loop with the edges already chosen at previous steps, then one ads e_l to the set of chosen edges and iterates. If e_l does form a loop with the edges already chosen at previous steps, one does not add it to the set of chosen edges, marks it as tested and iterates. A tree over the $|V_i|$ vertices uses only $|V_i| - 1$ edges and adding edges to a tree leads to a cycles (the two end vertices of the excess edge are connected by the tree and also by the edge). $\square$

There are several matrices associated to a graph:

- the $N \times N$ *adjacency matrix* A with entries $A_{ij} = 1$ if $(i, j) \in E$ and 0 if not. For the two graphs examples above we have:

$$A_1 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

- the $N \times |E|$ *incidence matrix* B . We orient the edges, say from the smaller to the larger label, and we define:

$$B_{ie} = \begin{cases} 1, & \text{if } e \text{ enters } i \\ -1, & \text{if } e \text{ exits } i \\ 0, & \text{otherwise} \end{cases}.$$

For the two examples before, ordering the edges lexicographically, we have

$$B_1 = \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

- the $N \times N$ diagonal *degree* matrix D with entries $D_{ii} = d_i$, with d_i the number of edges incident to the vertex i (the number of edges to which i belongs). For the examples before:

$$D_1 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

- the $N \times N$ *Laplacian* matrix $L = D - A$, which is for our two examples:

$$L_1 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

Lemma 6.5. *The product of the incidence matrix and its transposed is the Laplacian matrix $BB^t = L$.*

Proof. For our second example for instance we have:

$$\begin{pmatrix} -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

In general, $(BB^t)_{ij} = \sum_e B_{ie}B_{je}$. If i and $j \neq i$ are connected by an edge (ij) then the sum reduces to one term and $B_{i(ij)}B_{j(ij)} = -1$, as one of the entries is 1 and the other -1 ; if i and $j \neq i$ are not connected then the sum is zero; if $i = j$ then the sum is $\sum_e B_{ie}B_{ie}$ and each of the edges incident to i contribute a 1. $\square$

Note that the columns of L are linearly dependent, hence $\det(L) = 0$. We consider the matrix $\tilde{L}$ obtained from L by erasing the line and column corresponding to one vertex, chosen as root. We have $\tilde{L} = \tilde{B}\tilde{B}^t$, with $\tilde{B}$ the incidence matrix with the row corresponding to the root vertex erased.

Lemma 6.6. *The $(N-1) \times (N-1)$ minors of $\tilde{B}$ are the determinants of the matrices $\tilde{B}_I$, where I is a subsets of $N-1$ edges $I \subset E$. The minors are:*

$$\det(\tilde{B}_I) = \begin{cases} \pm 1 & \text{if } I \text{ is a tree} \\ 0 & \text{if not} \end{cases}.$$

Proof. Assume I has a simple cycle⁵ formed by the edges $e_1 = (i_0, i_1) \dots e_k = (i_{k-1}, i_0)$. By factoring a some global sign we can reorient all the edges along the cycle such that e_1 is oriented from i_0 to i_1 , e_2 from i_1 to i_2 and so on up to e_k oriented from i_{k-1} to i_0 . Then $B_{ie_1} + B_{ie_2} + \dots B_{ie_k} = 0$ for any i as:

- either i is not a vertex in the cycle and the all the terms are 0,
- or i is a vertex in the cycle, but then the consecutive terms B_{ie_s} and $B_{ie_{s+1}}$ to which i belongs are respectively $+1$ and -1 and cancel.

Thus in order for $\det(\tilde{B}_I) \neq 0$, I must be a graph with N vertices, $N - 1$ edges and no cycles, therefore I must be a tree. We consider one of the edges incident to the root, say (r, i) . This column has only one non zero entry, and expanding the minor along this column we obtain a minor corresponding to a graph in which the edge e has been deleted, and vertex i marked as root in the connected component generated by the deletion of e (that is the row corresponding to i is eliminated in B). Iterating, the process stops when all the edges have been exhausted. As at each step we multiply the result by ± 1 , the result follows. $\square$

Theorem 6.5 (**The matrix tree theorem**). *The determinant of $\tilde{L}$ equals the number of trees of G , $\det(\tilde{L}) = \kappa(G)$.*

Proof. As $\tilde{L} = \tilde{B}\tilde{B}^t$, from Cauchy-Binet $\det(\tilde{L}) = \sum_I |B_I|^2$ where the sum runs over all the subsets $I \subset E$ with $|I| = N - 1$. The only non zero contributions correspond to sets I which are trees, and each tree contributes exactly 1. $\square$

For the two examples we discussed:

$$\det(\tilde{L}_1) = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3, \quad \det(\tilde{L}_2) = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 8.$$

For the second graph we have the trees depicted in Fig. 6.1 below.

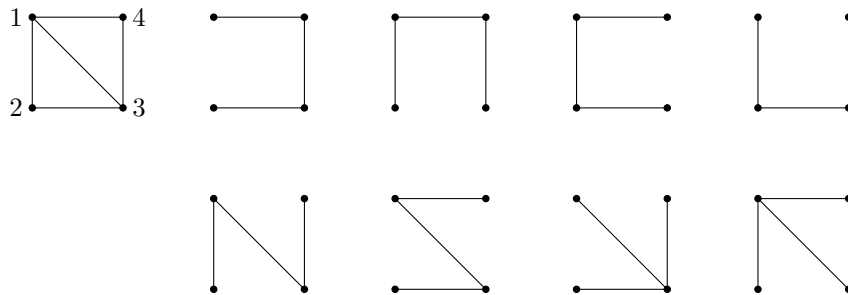


Figure 6.1: Trees in a graph.

⁵If I has a cycle it has a simple cycle.

Chapter 7

Functions, first contact

An important role in physics and mathematics is played by the continuous functions.

7.1 Continuity

Definition 7.1 (**Pointwise continuity**). For X and Y metric spaces, a function $f : X \rightarrow Y$ is called pointwise continuous if $\forall x, \forall \epsilon, \exists \delta_{\epsilon, x}$ such that $d(x, x') < \delta \Rightarrow d(f(x), f(x')) < \epsilon$. A function can be continued at a point x if there exists a function $\bar{f}$ whose domain includes x , is continuous at x and is such that $f = \bar{f}$ on the domain of f .

As we have seen in Chapter 3, a function is continuous if the preimage of any open set is open. A function between metric spaces is sequentially continuous (see Chapter 4) if for any $x_n \rightarrow x$, $f(x_n) \rightarrow f(x)$ and in metric spaces a function is continuous if and only if it is sequentially continuous.

Theorem 7.1 (**Sequential and pointwise continuity are equivalent**). A function $f : X \rightarrow Y$ with X and Y metric spaces is continuous if and only if it is sequentially continuous.

Proof. Consider a pointwise continuous function $f : X \rightarrow Y$ and $x_n \rightarrow x$. For any ϵ , there exist δ such that $d(x_n, x) < \delta \Rightarrow d(f(x_n), f(x)) < \epsilon$ and (for this δ but also for any other δ) there exist N_δ such that $n > N_\delta \Rightarrow d(x_n, x) < \delta$, hence $f(x_n) \rightarrow f(x)$.

Conversely, if f is not sequentially continuous at x , then there exist a sequence $x_n \rightarrow x$ such that $f(x_n)$ does not converge to $f(x)$. Then there exist some ϵ such that for any N_ϵ , there exist some $n > N_\epsilon$ with $d(f(x_n), f(x)) > \epsilon$. Consequently, for this ϵ , for any δ there exists $n > N_\delta$ with $d(x_n, x) < \delta$ but such that $d(f(x_n), f(x)) > \epsilon$ and f is not pointwise continuous at x . $\square$

It follows that a function is continuous if and only if it is pointwise continuous. A direct proof of this fact goes as follows.

Theorem 7.2 (**Continuous is pointwise continuous**). A function between two metric spaces is continuous if and only if it is pointwise continuous.

Proof. The proof follows the one establishing that, in metric spaces, sequential continuity and continuity are equivalent.

Let f a continuous function. For any ϵ , let $B_\epsilon(f(x))$ be the open ball of radius ϵ around $f(x)$. Then $x \in f^{-1}(B_\epsilon(f(x)))$ (note that $f^{-1}(f(x))$ includes other points in X if f is not injective) thus $f^{-1}(B_\epsilon(f(x)))$ is an open set in X (due to continuity) that includes x . It follows that there exists some $\delta_{\epsilon, x}$ such that $B_{\delta_{\epsilon, x}}(x) \subset f^{-1}(B_\epsilon(f(x)))$.

Suppose now that f is pointwise continuous but not continuous. This means that there exists an open set V such that $f^{-1}(V)$ is not open. Then there exist some $x \in f^{-1}(V)$ such that $\forall \delta, B_\delta(x) \setminus f^{-1}(V) \neq \emptyset$. As $f(x) \in V$ and V is

open, $\exists \epsilon$ with $B_\epsilon(f(x)) \subset V$. It follows that $\exists \epsilon$ such that $\forall \delta$, picking $x' \in B_\delta(x) \setminus f^{-1}(V)$ we have both $d(x, x') < \delta$ and $d(f(x), f(y)) \geq \epsilon$. $\square$

Exercise 7.1. Prove that the composition of continuous functions is continuous.

Exercise 7.2. Prove sums, products and ratios (when defined) of continuous functions are continuous functions. Conclude that any polynomial is a continuous function. $|z^n - w^n| = |z - w| \sum_{p=0}^{n-1} |z|^p |w|^{n-p} \leq |z - w| (s+1) (\max\{|z|, |w|\})^n$.

Exercise 7.3. Prove that the function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined as

$$f(x) = \begin{cases} \frac{1}{q}, & \text{for } x = \frac{p}{q} \text{ with } p, q \text{ co primes and } q > 0 \\ 1, & \text{for } x = 0 \\ 0, & \text{for } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases},$$

is continuous at all $x \in \mathbb{R} \setminus \mathbb{Q}$ but discontinuous at all $x \in \mathbb{Q}$. For any $x \in \mathbb{Q}$, $x + \frac{\sqrt{2}}{N}$ is irrational and arbitrarily close to x . For any $x \in \mathbb{R} \setminus \mathbb{Q}$ and for any q there exists an integer such that $\frac{p}{q} < x < \frac{p+1}{q}$ chose $\delta < \min\{|x - \frac{p}{q}|, |x - \frac{p+1}{q}|\} < \frac{1}{q}$. All the rational numbers in $(x - \delta, x + \delta)$ are of the form $\frac{p'}{q'}$ with denominator $q' > \frac{1}{\delta}$, hence $f(\frac{p'}{q'}) < \epsilon$.

Exercise 7.4. Prove that a power series defines a continuous function at all the points inside its convergence domain. Without loss of generality, let us consider a power series centered at zero, $P(z) = \sum_{n \geq 0} a_n z^n$. Note that $P(z)$ and $P'(z) = \sum_{n \geq 0} a_n (n+1) z^n$ have the same radius of convergence $r^{-1} = \limsup |a_n|^{1/n} = \limsup (n+1)^{1/n} |a_n|^{1/n}$ as $(n+1)^{1/n} \rightarrow 1$. Now for any two points in the convergence disk, denote $\rho = \max\{|z|, |w|\} < r$ then:

$$|P(z) - P(w)| \leq |z - w| \sum_{n \geq 0} |a_n| \sum_{p=0}^n z^p w^{n-p} \leq |z - w| \sum_{n \geq 0} a_n (n+1) \rho^n.$$

We will now focus on the case of $f: \mathbb{R} \rightarrow \mathbb{R}$. Note that a function f is continuous on a closed interval $[a, b]$ if it is continuous at $x \in (a, b)$ and the one sided limits exist at the end points $\lim_{x \searrow a} f(x) = f(a)$, $\lim_{x \nearrow b} f(x) = f(b)$.

7.2 The intermediate value theorem (Zwischenwertsatz)

Continuous functions do not jump between values.

Theorem 7.3 (The intermediate value theorem). Let $f: [a, b] \rightarrow \mathbb{R}$ a continuous function. Then $f([a, b])$ is an interval and $f([a, b]) \supset [\min\{f(a), f(b)\}, \max\{f(a), f(b)\}]$. That is any intermediate value between $f(a)$ and $f(b)$ is attained by some point in $[a, b]$. Equivalently, for any $t \in [0, 1]$ there exists $c \in [a, b]$ such that $f(c) = tf(a) + (1-t)f(b)$.

Proof. We assume $f(a) < u < f(b)$, as the other case is similar. Let $S = \{x \in [a, b] \mid f(x) \leq u\}$. S is nonempty as $a \in S$ and bounded from above (by b) hence it has a supremum $c = \sup S$. As f is continuous at c , for any ϵ there exists δ such that $|x - c| < \delta \Rightarrow |f(x) - f(c)| < \epsilon$, that is $f(x) - \epsilon < f(c) < f(x) + \epsilon$. As c is the supremum of S , there exists $x \in S \cap (c - \delta, c]$, hence such that $f(x) \leq u$ and $f(c) < u + \epsilon$. On the other hand $\forall x' \in (c, c + \delta)$, $x' \notin S$ hence $u < f(x')$ and $u - \epsilon < f(c)$. Thus $f(c) = u$.

Finally, $f([a, b])$ is an interval because, for any $c \in [a, b]$, all the values between $f(c)$ and $f(a)$ are attained. $\square$

Corollary 7.1. Let $I \subset \mathbb{R}$ an interval (could be open or unbounded), if $f: I \rightarrow J = f(I) \subset \mathbb{R}$ is strictly monotonic and continuous on I , then f is a bijection between I and J , J is an interval and f^{-1} is strictly monotonic and continuous on J .

Proof. Let us assume f is strictly increasing, that is $x_1 > x_2 \Leftrightarrow f(x_1) > f(x_2)$. Then f is injective (as $f(x_1) = f(x_2) \Leftrightarrow x_1 = x_2$) and it is also surjective on J , hence f is bijective. By the intermediate value theorem, J is an interval. As f is strictly increases, f^{-1} is also strictly increasing (as $y_1 > y_2 \Rightarrow f^{-1}(y_1) > f^{-1}(y_2)$).

It remains to prove that f^{-1} is continuous. Combining the fact that f is strictly monotonic with the intermediate value theorem we conclude that the image of any open interval (x_1, x_2) contains the open interval $(f(x_1), f(x_2))$, hence

is an open set.¹ As the standard topology is generated by the open intervals, f is an open map and it follows that f^{-1} is continuous. $\square$

For $f : \mathbb{R} \rightarrow \mathbb{R}$ we denote:

- $\lim_{x \rightarrow \infty} f(x) = L$ if $\forall \epsilon, \exists y$ such that $x > y \Rightarrow |f(x) - L| < \epsilon$
- $\lim_{x \rightarrow x_0} f(x) = \infty$ if $\forall L, \exists \delta$ such that $|x - x_0| < \delta \Rightarrow f(x) > L$.
- $\lim_{x \rightarrow \infty} f(x) = \infty$ if $\forall L, \exists y$ such that $x > y \Rightarrow f(x) > L$. In the same spirit, for a sequence $(x_n) \in \mathbb{R}$ we denote $\lim x_n = \infty$ if $\forall L, \exists N$ such that $n > N \Rightarrow x_n > L$.

7.2.1 The exponential function (Exponentialfunktion)

We revisited the exponential function restricted to real arguments, $\exp : \mathbb{R} \rightarrow \mathbb{R}_{>0}$, $\exp(x) \equiv e^x = \sum_{n \geq 0} \frac{1}{n!} x^n$.

Proposition 7.1. *Restricted to real arguments the exponential is monotonic, continuous and surjective on $\mathbb{R}_{>0} = (0, \infty)$.*

Proof. We have $e^0 = 1$ and for any positive x , $e^x > 1$. As $e^{x+y} = e^x e^y$, $\exp(x)$ is strictly increasing and $e^{-x} = (e^x)^{-1} > 0$, that is the exponential is always positive.

As $e^{\delta x} = 1 + \delta x \sum_{k \geq 1} \frac{1}{k!} \delta x^{k-1}$, for $0 < \delta x < 1$ we have $|e^{\delta x} - 1| \leq \frac{\delta x}{1 - \delta x}$, hence $|e^x - e^y| \leq e^x \frac{|x-y|}{1-|x-y|}$ for $|x-y| < 1$ that is e^x is continuous at any $x \in \mathbb{R}$ (by a similar argument we obtain continuity at any z in $\mathbb{C}$). For $x > 0$, $e^x > 1 + x$ hence $\lim_{x \rightarrow \infty} e^x = \infty$ and $\lim_{x \rightarrow -\infty} e^x = 0$ and by the intermediate value Theorem e^x takes all the values in the interval $(0, \infty)$. $\square$

It follows that the exponential has an inverse which is continuous, strictly increasing and surjective. We call this inverse the *natural logarithm* and we denote it:

$$\ln : \mathbb{R}_{>} \rightarrow \mathbb{R}, \quad \ln(x) = y, \text{ with } y \text{ such that } e^y = x.$$

Note that, because $e^{y_1} e^{y_2} = e^{y_1+y_2}$ we have $\ln(x_1 x_2) = \ln(x_1) + \ln(x_2)$. We are finally in the position to define irrational powers of real numbers. For all $\alpha \in \mathbb{R}$, $x \in \mathbb{R}_{>}$:

$$x^\alpha \equiv e^{\alpha \ln(x)}.$$

Exercise 7.5 (Convex and concave functions (Konvexe und konkave Funktionen)). *The graph of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is the set of points $(x, f(x)) \in \mathbb{R}^2$. We say that $f : \mathbb{R} \rightarrow \mathbb{R}$ is convex in the interval $[a, b]$ if $\forall t \in [0, 1]$:*

$$f(ta + (1-t)b) \leq tf(a) + (1-t)f(b),$$

and we say it is concave if $f(ta + (1-t)b) \geq tf(a) + (1-t)f(b)$. This has a graphical interpretation: a function is convex if at an intermediate point $ta + (1-t)b$ it is below the straight line from $f(a)$ to $f(b)$.

Prove that:

- for any $n \in \mathbb{N}$, $f : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}$ with $f(x) = x^n$ is convex.
- e^x is convex on $\mathbb{R}$.
- $\ln(x)$ is concave on $\mathbb{R}_{>0}$.

¹Any $y \in f((x_1, x_2))$, $y = f(x)$ for some $x \in (x_1, x_2)$ and there exists a δ such that $(x-\delta, x+\delta) \in (x_1, x_2)$, hence $f((x_1, x_2)) \supset (f(x-\delta), f(x+\delta)) \ni f(x) = y$.

7. FUNCTIONS, FIRST CONTACT

$f(x) = x^n$ is convex. The cases $n = 0, 1$ are trivial. For $n \geq 2$ we proceed by induction:

$$[ta + (1-t)b]^{n+1} = [ta + (1-t)b]^n [ta + (1-t)b] \leq [ta^n + (1-t)b^n][ta + (1-t)b] = t^2 a^{n+1} + (1-t)^2 b^{n+1} + t(1-t)[a^n b + ab^n],$$

where we used the induction hypothesis. A short manipulation shows that:

$$t^2 a^{n+1} + (1-t)^2 b^{n+1} + t(1-t)[a^n b + ab^n] = ta^{n+1} + (1-t)b^{n+1} - t(1-t)(a-b)^2 \left[\sum_{p=0}^n a^{n-p} b^p \right],$$

and convexity follows.

e^x is convex. Let us first prove it is convex on $\mathbb{R}_{>0}$. We have, using the previous point, that:

$$e^{ta+(1-t)b} = \sum_{n \geq 0} \frac{1}{n!} [ta + (1-t)b]^n \leq \sum_{n \geq 0} \frac{1}{n!} [ta^n + (1-t)b^n] = te^a + (1-t)e^b.$$

In general, we note that for any $x, y \in \mathbb{R}$, there exists a z such that $x+z > 0$ and $y+x > 0$ hence:

$$e^{tx+(1-t)y} = e^{-z} \left[e^{t(x+z)+(1-t)(y+z)} \right] \leq e^{-z} \left[te^{x+z} + (1-t)e^{y+z} \right] = te^x + (1-t)e^y.$$

$\ln(x)$ is concave. This follows from the previous point, combined with the observation that the inverse of a convex strictly increases function is concave. Indeed, assume $f : X \rightarrow Y$ is convex and strictly increasing. Then f^{-1} is also strictly increasing and

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \Rightarrow f^{-1}\left(f(tx_1 + (1-t)x_2)\right) \leq f^{-1}\left(tf(x_1) + (1-t)f(x_2)\right),$$

which, denoting $f(x_1) = y_1$ and $f(x_2) = y_2$ yields:

$$tf^{-1}(y_1) + (1-t)f^{-1}(y_2) \leq f^{-1}(ty_1 + (1-t)y_2).$$

7.3 The extreme value theorem

Continuous functions on compact sets attain their extrema. Note that the supremum c of any subset $S \subset \mathbb{R}$ can not be an interior point of S because for any ϵ , $B_\epsilon(c) = (c-\epsilon, c+\epsilon) \not\subset S$, as no element of S is larger than c . In fact the supremum:

- either is an isolated point of S (and necessarily belongs to S),
- or is a limit point of S (belonging or not to S).

Indeed, any point of S is either isolated or a limit point of S , hence if $c \in S$ we conclude. If $c \notin S$, then for any ϵ , $c-\epsilon$ is not the supremum of S hence $B_\epsilon(c) \cap S = B_\epsilon(c) \cap S \setminus \{c\} \neq \emptyset$, and c is a limit point of S .

Theorem 7.4 (**The extreme value theorem**). *Real valued continuous functions on compact domains attain their extrema, that is for any continuous function $f : K \rightarrow \mathbb{R}$ with compact domain K , there exists $x_m, x_M \in K$ such that $f(x_m) = \inf f(K)$ and $f(x_M) = \sup f(K)$.*

Proof. By definition, for any point $y \in f(K)$ there exists $x \in K$ such that $f(x) = y$. The continuous image $f(K)$ of the compact set K is compact and by Heine Borel a compact set in $\mathbb{R}$ is bounded and closed. As it is closed, $f(K)$ contains all the isolated points of $f(K)$ (if any) and all the limit points of $f(K)$, hence contains its supremum (and infimum). $\square$

We are now in the position to prove the fundamental theorem of algebra.

Theorem 7.5 (**The fundamental theorem of algebra (Fundamentalsatz der Algebra)**). *Any polynomial $P(z) = z^n + a_{n-1}z^{n-1} + \dots + a_0$ with complex coefficients $a_0, \dots, a_{n-1} \in \mathbb{C}$ admits a root in $\mathbb{C}$.*

Proof. We define $f(z) = |P(z)|$. It is easy to check that $f(z)$ is a continuous function on $\mathbb{C}$. The proof is divided into two steps:

$f(z)$ has a global minimum in $\mathbb{C}$. First we show that outside a big enough disk $B_R(0)$, the polynomial $P(z)$ has a large absolute value. Indeed, let R be a constant such that $R > \max\{1, 2 \sum_{i=0}^{n-1} |a_i|\}$. Then for $|z| > R$ we have:

$$|a_{n-1}z^{n-1} + \dots + a_0| \leq \sum_{i=0}^{n-1} |z|^i |a_i| \leq |z|^{n-1} \sum_{i=0}^{n-1} |a_i| < |z|^{n-1} \frac{R}{2} < \frac{|z|^n}{2}.$$

It follows that for $|z| > R$ we have:

$$f(z) = |P(z)| \geq |z|^n - |a_{n-1}z^{n-1} + \dots + a_0| > |z|^n - \frac{|z|^n}{2} = \frac{|z|^n}{2} > \frac{R^n}{2}.$$

Second we show that $f(z)$ has a minimum inside $\overline{B_R(0)}$. Indeed, the closed disk $\overline{B_R(0)}$ is closed and bounded, hence compact in $\mathbb{C}$. By the extreme value Theorem, the function $f(z)$ has a minimum in it, that is there exists $z_m \in \overline{B_R(0)}$ such that $f(z_m) \leq f(z)$ for all $z \in \overline{B_R(0)}$.

Third, we show that z_m is a global minimum in $\mathbb{C}$. Indeed, $f(z_m) \leq f(0) = |a_0| < \frac{R}{2} < \frac{R^n}{2}$, and outside the disk $f(z) > \frac{R^n}{2}$, therefore:

$$f(z_m) \leq f(z), \quad \forall z \in \mathbb{C}.$$

At its minimum, $f(z_m) = 0$. Observing that $z = z_m + (z - z_m)$, re expand the polynomial in terms of $z - z_m$:²

$$P(z) = P(z_m) + c_k(z - z_m)^k + c_{k+1}(z - z_m)^{k+1} + \dots + c_n(z - z_m)^n,$$

where $c_n = 1$ and we denoted k the lowest degree such that $c_k \neq 0$. Note that k is at most n , as $c_n = 1$, or lower.

We assume that $P(z_m) \neq 0$. Then, as $c_k \neq 0$, there exists $a \neq 0$ one of the roots of order k of $-\frac{P(z_m)}{c_k}$, that is $a^k = -\frac{P(z_m)}{c_k}$. We now evaluate P at the point $z_m + ta$:

$$P(z_m + ta) = P(z_m) + c_k t^k \left(-\frac{P(z_m)}{c_k} \right) + \sum_{p=k+1}^n c_p (ta)^p = (1 - t^k)P(z_m) + t^{k+1} \sum_{p=k+1}^n c_p a^p t^{p-k-1}.$$

For $t < 1$ we have $|\sum_{p=k+1}^n c_p a^p t^{p-k-1}| \leq \sum_{p=k+1}^n |c_p a^p| < C$ for some constant C and

$$|P(z_m + ta)| \leq (1 - t^k)|P(z_m)| + t^{k+1}C = 1 - t^k|P(z_m)| \left(1 - t \frac{C}{|P(z_m)|} \right),$$

which is strictly smaller than $|P(z_m)|$ for $t < \frac{|P(z_m)|}{C}$ leading to a contradiction. We conclude that $P(z_m) = 0$.

□

Theorem 7.6 (In finite dimensions all norms are equivalent). For any two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a finite dimensional Banach space V ($\mathbb{R}^N$ or $\mathbb{C}^N$), there exist two constants C_1 and C_2 such that:

$$C_1 \|x\|_b \leq \|x\|_a \leq C_2 \|x\|_b.$$

Proof. We will show that all norms are equivalent to $\|x\|_1 = \sum_{i=1}^N |x^i|$.

²Anticipating, this is just the pedestrian derivation of the fact that close to z_m the Taylor expansion at first non trivial order of $P(z)$ is $P(z_m + \delta z) \approx P(z_m) + \frac{1}{k!} P^{(k)}(z_m) (\delta z)^k$ for some $1 \leq k \leq n$ and $|P(z_m + \delta z)|$ decreases for $(\delta z)^k = -P(z_m) \frac{k!}{P^{(k)}(z_m)} t^k$.

Any norm $\|\cdot\|_a$ is continuous under $\|\cdot\|_1$. By the triangle inequality $|\|x_1\|_a - \|x_2\|_a| \leq \|x_1 - x_2\|_a$. On the other hand, choosing a basis e_i , we have:

$$\|x_1 - x_2\|_a \leq \sum_{i=1}^N |x_1^i - x_2^i| \|e_i\|_a \leq \|x_1 - x_2\|_1 \max_{i=1, \dots, N} \|e_i\|_a,$$

hence for any ϵ , choosing $\delta = \epsilon / \max_{i=1, \dots, N} \|e_i\|_a$ we have $\|x - x_0\|_1 < \delta \Rightarrow \|\|x_1\|_a - \|x_2\|_a\| < \|x_1 - x_2\|_a < \epsilon$.

The closed ball of radius 1 in $\|\cdot\|_1$, $B_1^{\|\cdot\|_1}(0) = \{x \in V \mid \|x\|_1 \leq 1\}$ is a bounded closed set in V , hence by the Heine-Borel theorem it is compact: it is here that we use the fact that V is finite dimensional. By the extreme value theorem, the continuous function $f(x) = \|x\|_a$ will attain a maximum and a minimum on this closed ball, which we denote $C_1 = \inf_{\|u\|_1=1} \|u\|_a$ and $C_2 = \sup_{\|u\|_1=1} \|u\|_a$ hence for any x :

$$C_1 \leq \left\| \frac{1}{\|x\|_1} x \right\|_a \leq C_2.$$

□

As a consequence we have the following result.

Theorem 7.7 (**In finite dimensions all the linear operators are bounded**). *If X is a finite dimensional normed vector space and Y is a normed vector space then any linear operator $T : X \rightarrow Y$ is bounded.*

Proof. Let us denote $e_1, \dots, e_n$ a Hamel basis in X . Then $\forall x \in X$, $x = \sum_{k=1}^n a_k e_k$:

$$\|T(x)\|_Y \leq \sum_{k=1}^n |a_k| \|T(e_k)\| \leq \left(\max_{k=1, \dots, n} \|T(e_k)\| \right) \sum_{k=1}^n |a_k| = \left(\max_{k=1, \dots, n} \|T(e_k)\| \right) \|x\|_1,$$

and we conclude as any norm on X is equivalent with $\|\cdot\|_1$.

□

7.4 Differentiation in one variable

The derivative of a function captures how the function behaves when one of the arguments is varied. For a function h continuous at x_0 we denote $\lim_{x \rightarrow x_0} h(x) = h(x_0)$.

Definition 7.2 (**Derivative**). *Consider $f : \mathbb{R} \rightarrow V$ for $(V, \|\cdot\|)$ some Banach space. We say that f is differentiable at x_0 if the function: $\frac{f(x) - f(x_0)}{x - x_0}$ is a continuous function of x at $x = x_0$ ³ and in this case we denote:*

$$f'(x_0) \equiv \frac{df}{dx}(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}, \quad \left(\forall \epsilon, \exists \delta, \text{ such that } |x - x_0| < \delta \Rightarrow \left\| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right\| < \epsilon \right)$$

and we call $f'(x_0)$ the derivative of f at x_0 . f is called differentiable on an interval I if it is differentiable at all $x_0 \in I$ and it is called continuously differentiable if $f'(x_0)$ is a continuous function. We denote the vector space of continuously differentiable functions $\mathcal{C}^1(\mathbb{R}, V)$.

Note that $f(x) - f(x_0)$ is a vector in V and $(x - x_0)^{-1}$ a scalar, hence $f'(x)$ is a vector in V . Denoting e_k a basis in V , $f(x) = \sum_k f_k(x) e_k$ and $f'(x) = \sum_k f'_k(x) e_k$ ⁴. Remark that a priori $\frac{f(x) - f(x_0)}{x - x_0}$ is defined for $x \neq x_0$ and needs to be extended by continuity to x_0 . If f is differentiable at x_0 then f is continuous at x_0 , as for any ϵ , $\|f(x) - f(x_0)\| \leq (\epsilon + \|f'(x_0)\|)|x - x_0|$ if $|x - x_0|$ is small enough.

³Strictly speaking, $\frac{f(x) - f(x_0)}{x - x_0}$ can be continuously continued at x_0 .

⁴This holds for Hamel basis in finite dimensions, but also for Schauder basis in infinite dimensions. Indeed, it can be shown that in any Banach space with a Schauder basis $(e_n)_n$, the coordinate functionals $F_n : V \rightarrow \mathbb{R}$, $F_n(\sum_{p=1}^{\infty} \alpha_p e_p) = \alpha_n$ and the projections $P_m : V \rightarrow V$, $P_m(\sum_{p=1}^{\infty} \alpha_p e_p) = \sum_{p=1}^m \alpha_p e_p$ are continuous functions.

Theorem 7.8 (Mean value inequality). Let $f : [x_1, x_2] \rightarrow V$ a function continuous on $[x_1, x_2]$, differentiable on (x_1, x_2) . Then:

$$\|f(x_2) - f(x_1)\| \leq |x_2 - x_1| \sup_{x \in (x_1, x_2)} \|f'(x)\|.$$

Proof. Assume that there exists some $\epsilon > 0$ such that $\|f(x_2) - f(x_1)\| - (x_2 - x_1) \left(\sup_{x \in (x_1, x_2)} \|f'(x)\| + \epsilon \right) > 0$. Consider the function:

$$g_\epsilon(x) = \|f(x) - f(x_1)\| - (x - x_1) \left(\sup_{x \in (x_1, x_2)} \|f'(x)\| + \epsilon \right), \quad g_\epsilon(x_1) = 0, \quad g_\epsilon(x_2) > 0,$$

and denote $x_0 = \sup\{x \in [x_1, x_2], g_\epsilon(x) = 0\}$. As g_ϵ is continuous it is sequentially continuous and considering a sequence⁵ $x_n \in S$, $x_n \rightarrow x_0$ we conclude that $g_\epsilon(x_0) = 0$. Furthermore, as $g_\epsilon(x_2) > 0$, we have $x_0 < x_2$ and $g_\epsilon(x) > 0$ for $x_0 < x < x_2$ (were $g_\epsilon(x) < 0$ for some such x , by the intermediate value theorem there would exist x' with $x < x' < x_2$ with $g_\epsilon(x') = 0$ and x_0 would not be the supremum). For all x with $x_0 < x < x_2$ we have:

$$\begin{aligned} 0 &\leq g_\epsilon(x) - g_\epsilon(x_0) = \|f(x) - f(x_1)\| - \|f(x_0) - f(x_1)\| - (x - x_0) \left(\sup_{x \in (x_1, x_2)} \|f'(x)\| + \epsilon \right) \\ &\leq \|f(x) - f(x_0)\| - (x - x_0) \left(\sup_{x \in (x_1, x_2)} \|f'(x)\| + \epsilon \right), \end{aligned}$$

hence

$$\sup_{x \in (x_1, x_2)} \|f'(x)\| + \epsilon \leq \left\| \frac{f(x) - f(x_0)}{x - x_0} \right\|,$$

and sending $x \rightarrow x_0$ we obtain a contradiction. □

7.4.1 The case $f : \mathbb{R} \rightarrow \mathbb{R}$

An important particular class of functions are the real functions of one variable.

Theorem 7.9 (Extremum criterion, Rolle's theorem and mean value theorem). Let $f : [x_1, x_2] \rightarrow \mathbb{R}$ continuous on $[x_1, x_2]$ and differentiable on (x_1, x_2) . Then:

- Extremum criterion if f has a local extremum (maximum or minimum) at $x_0 \in (x_1, x_2)$ then $f'(x_0) = 0$.
- Rolle's theorem if $f(x_1) = f(x_2)$ then there exists $x \in (x_1, x_2)$ such that $f'(x) = 0$.
- Mean value theorem $\exists x \in (x_1, x_2)$ such that $\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(x)$.

Proof. Assume that f attains a maximum at $x_0 \in (x_1, x_2)$. Then:

$$\lim_{x \nearrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \geq 0, \quad \lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \leq 0,$$

and as f is differentiable at x_0 , the two limits must coincide which is possible only if $f'(x_0) = 0$.

For the second item, by the extreme value theorem, f must attain a maximum and a minimum in the interval $[x_1, x_2]$. If $f(x_1)$ is both minimum and maximum, then the function is constant and its derivative is identically zero; if not we invoke the first item.

⁵If x_0 is the supremum of S there exists a sequence $x_n \in S$ with $x_n \rightarrow x_0$. If $x_0 \in S$ then the sequence $x_n = x_0$ belongs to S and converges to x_0 . If $x_0 \notin S$, as $x_0 - \epsilon$ is not the supremum of S there exist $x \in S$ with $x_0 - \epsilon < x < x_0$ hence $B_\epsilon(x_0) \cap S \setminus \{x_0\} \neq \emptyset$ and x_0 is a limit point of S hence the limit of some sequence in S from Theorem 4.1.

For the last item, we apply the second one to:

$$g(x) = f(x) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) .$$

□

Theorem 7.10 (**Convexity criterion**). Consider $f : [x_1, x_2] \rightarrow \mathbb{R}$ continuous on $[x_1, x_2]$ and differentiable on (x_1, x_2) . Then:

- f monotonically increasing if and only if $f' \geq 0$.
- recall that f is called convex if $\forall t \in [0, 1], f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$. If f is twice continuously differentiable on (x_1, x_2) then f is convex if and only if $f''(x) \geq 0$ for all $x \in (x_1, x_2)$.

Proof. If f is monotonically increasing, then using the definition we have $f'(x) \geq 0$. Conversely, if $f' \geq 0$ then f is increasing on any interval by the mean value theorem.

Setting $t = \frac{x_2 - x}{x_2 - x_1}$, convexity is equivalent to:

$$\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x} , \quad x_1 < x < x_2 .$$

Assume that $f''(x) < 0$ for some $x \in (x_1, x_2)$. Then, as f'' is continuous, $f''(x) < 0$ for x in some interval, $x \in [y_1, y_2] \subset (x_1, x_2)$, and f' is strictly decreasing on this interval. But if f is convex on $[y_1, y_2]$, then:

$$\frac{f(x) - f(y_1)}{x - y_1} \leq \frac{f(y_2) - f(x)}{y_2 - x} , \quad y_1 < x < y_2 ,$$

and by the mean value theorem, there exist $z_1 \in (y_1, x)$ and $z_2 \in (x, y_2)$, hence $z_1 < z_2$ such that:

$$f'(z_1) = \frac{f(x) - f(y_1)}{x - y_1} \leq \frac{f(y_2) - f(x)}{y_2 - x} = f'(z_2) ,$$

contradicting the fact that f' is strictly decreasing on $[y_1, y_2]$. □

Theorem 7.11 (**Properties of derivatives**). Consider $f, g : \mathbb{R} \rightarrow \mathbb{R}$. We have:

- Leibnitz rule if f and g are differentiable at x_0 , then so is their product and $(f \cdot g)'(x_0) = f'(x_0) \cdot g(x_0) + f(x_0) \cdot g'(x_0)$.
- chain rule if g is differentiable at x_0 and f is differentiable at $g(x_0)$, then the composition $(f \circ g)(x) = f(g(x))$ is differentiable at x_0 and $(f \circ g)'(x_0) = (f' \circ g)(x_0) \cdot g'(x_0)$.
- power series if $f = \sum_{n \geq 0} c_n x^n$ is a power series with radius of convergence $r > 0$ then $\sum_{n \geq 1} n c_n x^{n-1}$ is a power series with radius of convergence r and its sum equals $f'(x)$.

Proof. We have:

$$(f \cdot g)'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} = \lim_{x \rightarrow x_0} \frac{1}{x - x_0} \left(f(x)g(x) - f(x_0)g(x) + f(x_0)g(x) - f(x_0)g(x_0) \right) .$$

The composition is straightforward.

For the last part, it is easy to check that the radius of converge of $\sum_{n \geq 1} n c_n x^{n-1}$ is r . As a power series converges absolutely inside its domain of convergence we have:

$$h(x) = \frac{f(x) - f(x_0)}{x - x_0} = \sum_{n=1} c_n \sum_{p=0}^{n-1} x^p x_0^{n-p},$$

and the series converges absolutely for $|x|, |x_0| < \rho < r$. As, except at most a finite number of terms, $|c_n| \leq r^{-n}$, we have:

$$|h(x) - h(x_0)| \leq \sum_{n=2} |c_n| \sum_{p=0}^{n-1} |x^p - x_0^p| |x_0|^{n-p} \leq \sum_{n=2} \frac{1}{r^n} |x - x_0| \sum_{p=1}^{n-1} \sum_{q=1}^p |x|^q |x_0|^{p-q} |x_0|^{n-p} \leq |x - x_0| \sum_{n \geq 2} \frac{n(n-1)}{r^n} \rho^{n-2} = \frac{|x - x_0|}{(1 - \rho/r)^3}.$$

□

Theorem 7.12 (L'Hôpital's rule). Let c and L be real numbers or $\pm\infty$, and I an interval containing c if $c \in \mathbb{R}$ or open at c if $c = \pm\infty$. Let two functions $f, g : I \rightarrow \mathbb{R}$ differentiable on I (except possibly at c) such that $g'(x) \neq 0$ (except possibly at c). If $\lim_{x \rightarrow c} \frac{f'(x)}{g'(x)} = L$ and either $\lim_{x \rightarrow c} f(x) = \lim_{x \rightarrow c} g(x) = 0$ or $\lim_{x \rightarrow c} |g(x)| = \infty$ then $\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = L$.

Proof. We take the case $\lim_{x \nearrow c} f(x) = \lim_{x \nearrow c} g(x) = 0$ (the other cases can be deduced from this one by simple tricks). If needed, we continue f and g at c to $f(c) = g(c) = 0$. For any $a < c$, apply Rolle's theorem to the function $h : [a, c] \rightarrow \mathbb{R}$:

$$h(y) = (f(y) - f(a))(g(c) - g(a)) - (f(c) - f(a))(g(y) - g(a)) \Rightarrow \exists y \in (a, c), \quad \frac{f'(y)}{g'(y)} = \frac{f(c) - f(a)}{g(c) - g(a)} = \frac{f(a)}{g(a)}.$$

As $\lim_{y \nearrow c} \frac{f'(y)}{g'(y)} = L$, for any ϵ , $\exists \delta$ such that $c - y < \delta \Rightarrow \left| \frac{f'(y)}{g'(y)} - L \right| < \epsilon$. But then, for any a with $c - a < \delta$, $\left| \frac{f'(a)}{g'(a)} - L \right| = \left| \frac{f'(y)}{g'(y)} - L \right|$ for some $y \in (a, c)$ hence $\left| \frac{f'(a)}{g'(a)} - L \right| < \epsilon$. □

Theorem 7.13 (Taylor series). Let the function $f : \mathbb{R} \rightarrow \mathbb{R}$, continuous on the closed interval and $k+1$ times differentiable on the open interval between x and a . Then there exists ξ between a and x such that

$$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2!}(x-a)^2 f''(a) + \cdots + \frac{1}{k!}(x-a)^k f^{(k)}(a) + R_k(x), \quad R_k(x) = \frac{1}{(k+1)!}(x-a)^{k+1} f^{(k+1)}(\xi),$$

where $f^{(k)}(a) = \frac{d^k f}{dx^k}(a)$ is the k 'th derivative of f at the point a .

Proof. Define $F(t) = f(t) + (x-t)f'(t) + \frac{1}{2!}(x-t)^2 f''(t) + \cdots + \frac{1}{k!}(x-t)^k f^{(k)}(t)$ for t in the interval comprised between x and a and observe that $F'(t) = \frac{1}{k!}(x-t)^k f^{(k+1)}(t)$. By the mean value theorem, for any G , there exists ξ between x and a such that

$$\frac{F'(\xi)}{G'(\xi)} = \frac{F(x) - F(a)}{G(x) - G(a)},$$

and using $G(t) = (x-t)^{k+1}$ this reads $R_k(x) = \frac{F'(\xi)}{G'(\xi)}(G(x) - G(a)) = \frac{1}{k!}(x-\xi)^k f^{(k+1)}(\xi) \frac{(x-a)^{k+1}}{(k+1)(x-\xi)^k}$. □

Exercise 7.6 (Differentiable but not continuously differentiable). Prove that the function:

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x}), & x \neq 0 \\ 0, & x = 0 \end{cases},$$

is differentiable but not continuously differentiable at $x = 0$.

Exercise 7.7 (Derivatives of simple functions). Let $f : I \rightarrow \mathbb{R}$ for $I \in \mathbb{R}$ some suitable interval. Prove that:

- $f(x) = c$, $f'(x) = 0$. Prove that if $f'(x) = 0$ on an interval, then $f(x) = c$ for some constant c .

- $f(x) = x$, $f'(x) = 1$.
- $f(x) = x^n$, $f'(x) = nx^{n-1}$.
- $f(x) = e^x$, $f'(x) = e^x$.
- $f(x) = \ln(x)$, $f'(x) = \frac{1}{x}$.
- recalling that $x < \sin x < \tan x$ prove that $\sin'(x) = \cos(x)$ and $\cos'(x) = -\sin(x)$.
- prove that $e^{-ix}[\cos(x) + i\sin(x)] = 1$. The complex sinus and cosinus functions are defined as

$$e^{\pm iz} = \cos(z) \pm i\sin(z) \Rightarrow \begin{aligned} \sin(z) &= \frac{e^{iz} - e^{-iz}}{2i} = \sum_{k \geq 0} \frac{1}{(2k+1)!} (-1)^k z^{2k+1}, \\ \cos(z) &= \frac{e^{iz} + e^{-iz}}{2} = \sum_{k \geq 0} \frac{1}{(2k)!} (-1)^k z^{2k}. \end{aligned}$$

Exercise 7.8 (Generalized binomial theorem). Apply Taylor's theorem to $\sin(x)$, $\cos(x)$, e^x , $\ln(1+x)$ and $(1+x)^\alpha$.