Refresher for Machine Learning

Maxime Sangnier

April 21, 2023

Contents

Introduction

These notes have been written after a careful reading of:

1. J. Garnier, S. Méléard, and N. Touzi. *Aléatoire.* École Polytechnique, 2021.
2. A. Guyader. *Statistique.* Sorbonne Université, 2021.
3. G.H. Golub and C.F. Van Loan. *Matrix Computations.* Baltimore, Maryland, The Johns Hopkins University Press, 2013.
4. L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* New York Inc., Springer-Verlag, 2004.

I claim no paternity of the content lying in this manuscript, which is mainly a reformulation of results exposed in the previous works. My poor contribution is only in selecting results important in my eyes, fixing a consistent notation and articulating arguments.

For the sake of conciseness, results are not proved but rigorous demonstrations can be found in the previous works.

I am particularly grateful to Arnaud Guyader, whose rigorous work is a daily source of learning and inspiration.

Chapter 1

Matrix and functional analysis

1.1 Matrices

1.1.1 Basics of linear algebra

Definition 1.1.1 (Linear independence and subspace)**.**

1. *A set of vectors* $\{x_1, \dots, x_n\} \subset \mathbb{R}^p$ *is linearly independent if*

$$\forall \alpha \in \mathbb{R}^n : \quad \sum_{i=1}^{n} \alpha_i x_i = 0 \implies \alpha = 0.$$

   In addition, $\{y_1, \dots, y_k\} \subset \{x_1, \dots, x_n\}$ *is a maximal linearly independent subset of* $\{x_1, \dots, x_n\}$ *if it is linearly independent and*

$$\forall S \subset \{x_1, \dots, x_n\}, S \text{ linearly independent and } S \neq \{y_1, \dots, y_n\} : \{y_1, \dots, y_k\} \not\subset S.$$

2. *A subspace of* $\mathbb{R}^p$ *is a subset that is also a vector space.*
3. *Let* $S_1, \dots, S_k \subset \mathbb{R}^p$ *be subspaces.* $S = S_1 + \dots + S_k$ *is a direct sum, and we note* $S = S_1 \oplus \dots \oplus S_k$*, if*

$$\forall x \in S, \exists!(x_1, \dots, x_k) \in S_1 \times \dots \times S_k : x = x_1 + \dots + x_k.$$

Property 1.1.1. *Let* $\{x_1, \dots, x_n\} \subset \mathbb{R}^p$*.*

1. $\operatorname{span}\{x_1, \dots, x_n\} = \{\sum_{i=1}^n \alpha_i x_i : \alpha \in \mathbb{R}^n\}$ *is a subspace of* $\mathbb{R}^p$*.*
2. *If* $\{y_1, \dots, y_k\} \subset \{x_1, \dots, x_n\}$ *is a maximal linearly independent subset of* $\{x_1, \dots, x_n\}$*, then* $\operatorname{span}\{y_1, \dots, y_k\} = \operatorname{span}\{x_1, \dots, x_n\}$ *and* $\{y_1, \dots, y_k\}$ *is a basis for* $\operatorname{span}\{y_1, \dots, y_k\}$*.*
3. *Let* S *be a subspace of* $\mathbb{R}^p$*. Then there exists* $\{y_1, \dots, y_k\}$ *linearly independent such that* $S = \operatorname{span}\{y_1, \dots, y_k\}$, i.e. $\{y_1, \dots, y_k\}$ *is a basis for* S*. In addition, all bases have the same number of elements, called the dimension of* S *and denoted* $\dim(S)$*.*

Definition 1.1.2. *Let $\boldsymbol{A} = [a_1| \dots |a_p] \in \mathbb{R}^{n\times p}$. The range and the kernel (or null space) of $\boldsymbol{A}$ are respectively*

$$\operatorname{range}(\boldsymbol{A}) = \{\boldsymbol{A}x, x \in \mathbb{R}^p\} = \operatorname{span}\{a_1, \dots, a_p\} \quad \textit{and} \quad \ker(\boldsymbol{A}) = \{x \in \mathbb{R}^p, \boldsymbol{A}x = 0\}\,.$$

Moreover, the rank of $\boldsymbol{A}$ is $\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{range}(\boldsymbol{A}))$. It is the maximal number of linearly independent columns (or rows).

At last, $\boldsymbol{A}$ is said rank-deficient if $\operatorname{rank}(\boldsymbol{A}) < \min(n, p)$.

Proposition 1.1.2 (Rank-nullity theorem)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. Then*

$$\operatorname{rank}(\boldsymbol{A}) + \dim(\ker(\boldsymbol{A})) = p.$$

The rank-nullity theorem says that the dimension of the subspace generated by $\boldsymbol{A}$ cannot be larger than the number of columns and if $\boldsymbol{A}$ is full rank, then $\ker(\boldsymbol{A}) = \{0\}$, meaning that $x \mapsto \boldsymbol{A}s$ is injective.

A direct application of the rank-nullity theorem is that for any basis $\{x_1, \dots, x_k\}$ of $\ker(\boldsymbol{A})$ and basis $\{\boldsymbol{A}x_{k+1}, \dots, \boldsymbol{A}x_p\}$ of $\operatorname{range}(\boldsymbol{A})$, $\{x_1, \dots, x_p\}$ is a basis of $\mathbb{R}^p$.

As another consequence, a square matrix is injective (or surjective) if and only if it is both injective and surjective, *i.e.* if and only if it is bijective.

Definition 1.1.3. *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is said non-singular (also non-degenerate or invertible) if it has an inverse,* i.e. *a matrix $\boldsymbol{A}^{-1} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$.*

Property 1.1.3. *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be two non-singular matrices. Then*

1. *$\boldsymbol{A}\boldsymbol{B}$ is non-singular and $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.*
2. *$\boldsymbol{A}^\top$ is non-singular and $(\boldsymbol{A}^\top)^{-1} = (\boldsymbol{A}^{-1})^\top$.*
3. *$\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1} = -\boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1}$*
4. *Sherman-Morrison-Woodbury formula (or matrix inversion lemma): let $\boldsymbol{U} \in \mathbb{R}^{n\times k}$, $\boldsymbol{C} \in \mathbb{R}^{k\times k}$ non-singular, $\boldsymbol{V} \in \mathbb{R}^{k\times p}$. Then $\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}$ is non-singular if and only if $\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}$ is invertible and in this case:*

   $$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}.$$

   We observe that a rank-k correction to a matrix results in a rank-k correction to the inverse.
5. *Sherman–Morrison formula: let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$. Then $\boldsymbol{A} + uv^\top$ is invertible if*

and only if $1 + v^\top \boldsymbol{A}^{-1} u \neq 0$ *and in this case:*

$$(\boldsymbol{A} + uv^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1} uv^\top \boldsymbol{A}^{-1}}{1 + v^\top \boldsymbol{A}^{-1} u}.$$

Definition 1.1.4 (Determinant)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *be a square matrix. Then*

$$\det(\boldsymbol{A}) = \sum_{\sigma \text{ permutation in } [\![1,n]\!]} \operatorname{sign}(\sigma) \prod_{i=1}^{n} \boldsymbol{a}_{\sigma(i),i},$$

where $\operatorname{sign}(\sigma)$ *is the sign of the permutation* σ*,* i.e. 1 *if* σ *is built with an even number of inversions and* -1 *otherwise.*

The determinant of a matrix $\boldsymbol{A}$ has a nice geometric meaning: its the signed volume of the unit hypercube transformed by the mapping $x \in \mathbb{R}^n \mapsto \boldsymbol{A}x$. The sign is negative if an odd number of axes are flipped by $\boldsymbol{A}$.

Property 1.1.4. *Let* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ *be two square matrices. Then*

1. *if* $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $\det(\boldsymbol{A}) = ad - bc$*;*
2. *if* $\boldsymbol{A}$ *is diagonal,* $\det(\boldsymbol{A}) = \prod_{i=1}^n \boldsymbol{a}_{ii}$*;*
3. $\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$*;*
4. $\det(\boldsymbol{A}^\top) = \det(\boldsymbol{A})$*;*
5. $\forall \alpha \in \mathbb{R} : \det(\alpha \boldsymbol{A}) = \alpha^n \det(\boldsymbol{A})$*;*
6. $\det(\boldsymbol{A}) \neq 0 \iff \boldsymbol{A}$ *is non-singular.*

Definition 1.1.5 (Orthogonality)**.** *A set of vectors* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ *is*

- ◇ *orthogonal if for every* $i, j \in [\![1, n]\!]$, $i \neq j \implies x_i^\top x_j = 0$*;*
- ◇ *orthonormal if for every* $i, j \in [\![1, n]\!]$, $x_i^\top x_j = \delta_{ij}$*.*

Let $S \subset \mathbb{R}^p$ *be a subspace.*

- ◇ *the orthogonal complement of* S *is:*

$$S^\perp = \left\{ y \in \mathbb{R}^p : y^\top x = 0, \forall x \in S \right\};$$

- ◇ *The vectors* $x_1, \ldots, x_k \subset S$ *form an orthonormal basis of* S *if they are orthonormal and if*

$$\operatorname{span}\{x_1, \ldots, x_k\} = S.$$

In particular, n *orthonormal vectors from* $\mathbb{R}^n$ *form an orthonormal basis of* $\mathbb{R}^n$*.*

A square matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is said orthogonal if

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_n,$$

or equivalently if its columns (respectively is rows)

A particular case is the set of rotation matrices: they are orthogonal matrices with determinant 1 (*i.e.* axes are rotated but not flipped).

Property 1.1.5. *Let $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ be a matrix. Then,* $\operatorname{range}(\boldsymbol{A})^\perp = \ker(\boldsymbol{A}^\top)$.

Property 1.1.6. *Let $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ be a square matrix. The following statements are equivalent:*

1. *$\boldsymbol{Q}$ is orthogonal;*
2. *$\boldsymbol{Q}$ is non-singular with $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$;*
3. *the columns of $\boldsymbol{Q}$ are orthonormal vectors (*i.e. *they form an orthonormal basis of $\mathbb{R}^n$);*
4. *the rows of $\boldsymbol{Q}$ are orthonormal vectors (*i.e. *they form an orthonormal basis of $\mathbb{R}^n$).*

Proposition 1.1.7 (Basis extension theorem)**.** *Let $\{x_1, \dots, x_k\} \subset \mathbb{R}^n$ (with $k < n$) be orthonormal vectors. Then, there exists $\{x_{k+1}, \dots, x_n\} \subset \mathbb{R}^n$ such that $\{x_1, \dots, x_n\}$ form an orthonormal basis of $\mathbb{R}^n$.*

As a consequence, if $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ is a matrix with orthonormal columns, then there exists $\boldsymbol{B} \in \mathbb{R}^{n \times (n-k)}$ such that the matrix $[\boldsymbol{A}|\boldsymbol{B}] \in \mathbb{R}^{n \times n}$ is orthogonal. In addition, $\operatorname{range}(\boldsymbol{B}) = \operatorname{range}(\boldsymbol{A})^\perp$.

1.1.2 Eigendecomposition

Definition 1.1.6 (Eigenvalue and eigenvector)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix. A vector $u \in \mathbb{R}^n \backslash \{0\}$ is an eigenvector of $\boldsymbol{A}$ if*

$$\exists \lambda \in \mathbb{R} : \quad \boldsymbol{A}u = \lambda u.$$

The scalar λ is called the eigenvalue of u.

Property 1.1.8. *Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix.*

1. *$\boldsymbol{A}$ has at most n different eigenvalues (they are the roots of a the degree-n characteristic polynomial of $\boldsymbol{A}$), with multiplicities summing to n;*
2. *the rank of $\boldsymbol{A}$ is the number of non-zero eigenvalues;*
3. *if $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ is similar to $\boldsymbol{A}$,* i.e. *$\exists \boldsymbol{P} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{P}$ is non-singular and*

$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{-1}$, *then* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *have the same eigenvalues.*

Definition 1.1.7 (Diagonalizable matrix)**.** *A matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is diagonalizable (or non-defective) if it is similar to a diagonal matrix,* i.e. *if*

$$\exists \boldsymbol{P}, \boldsymbol{D} \in \mathbb{R}^{n\times n}, \boldsymbol{P} \text{ non-singular and } \boldsymbol{D} \text{ diagonal, such that } \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}.$$

$\boldsymbol{D}$ *and* $\boldsymbol{P}$ *are respectively known as the spectral matrix and the modal matrix.*

Property 1.1.9 (Eigendecomposition)**.** *A matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is diagonalizable if and only if it has linearly independent eigenvectors* $\{v_1, \ldots, v_n\}$*. In this case, denoting* $\{\lambda_1, \ldots, \lambda_n\}$ *the corresponding eigenvalues, the eigendecomposition of* $\boldsymbol{A}$ *is:*

$$\boldsymbol{A} = [v_1|\ldots|v_n] \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix} [v_1|\ldots|v_n]^{-1}.$$

Property 1.1.10 (Inversion)**.** *A square and diagonalizable matrix* $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ *is non-singular if and only if all the entries of its spectral matrix* $\boldsymbol{D}$ *are non-zero. In this case,*

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}^{-1}.$$

Property 1.1.11 (Trace and determinant)**.** *Let* $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ *be a square and diagonalizable* $n \times n$ *matrix. Then* $\det(\boldsymbol{A}) = \prod_{i=1}^{n} \boldsymbol{d}_{ii}$ *and* $\operatorname{trace}(\boldsymbol{A}) = \sum_{i=1}^{n} \boldsymbol{d}_{ii}$.

Theorem 1.1.12 (Schur decomposition)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *be a symmetric matrix. Then,* $\boldsymbol{A}$ *is diagonalizable by an orthogonal matrix (also called change-of-basis or transition matrix):*

$$\begin{cases} \exists \boldsymbol{Q} \in \mathbb{R}^{n\times n}, \boldsymbol{Q} \text{ orthogonal,} \\ \exists \boldsymbol{D} \in \mathbb{R}^{n\times n}, \boldsymbol{D} \text{ diagonal} \end{cases} \quad : \quad \boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\top}.$$

In addition, the entries of $\boldsymbol{D}$ *are uniquely defined:*

$$\exists!(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n : \lambda_1 \leq \cdots \leq \lambda_n \text{ and } \boldsymbol{D} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix} \text{ up to a permutation.}$$

$\{\lambda_1, \ldots, \lambda_n\}$ *are the eigenvalues of* $\boldsymbol{A}$ *and the columns of* $\boldsymbol{Q}$ *are its normed eigenvectors.*

Let us remark that by denoting $\boldsymbol{Q} = [v_1|\dots|v_n]$, the eigendecomposition can be expressed:

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i v_i v_i^\top,$$

where $\{\lambda_1, \dots, \lambda_n\}$ are the eigenvalues and $\{v_1, \dots, v_n\}$ are the normed eigenvectors of $\boldsymbol{A}$.

Property 1.1.13. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvectors $\{v_1, \dots, v_n\}$ and eigenvalues $\lambda_1 \leq \dots \leq \lambda_n$. Then*

$$\lambda_1 = \min_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^\top \boldsymbol{A} x}{x^\top x} = \frac{v_1^\top \boldsymbol{A} v_1}{v_1^\top v_1},$$

and

$$\lambda_n = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^\top \boldsymbol{A} x}{x^\top x} = \frac{v_n^\top \boldsymbol{A} v_n}{v_n^\top v_n}.$$

Property 1.1.14 (Projection matrix)**.** *Let $S \subset \mathbb{R}^p$ be a subspace and $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ its orthogonal projection matrix. Then*

1. *$\boldsymbol{P}$ is unique;*
2. *$\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^\top$, where $\boldsymbol{V} = \left[v_1|\dots|v_{\dim(S)}\right]$ and $\{v_1|\dots|v_{\dim(S)}\}$ is an orthonormal basis for S;*
3. *$\mathrm{range}(\boldsymbol{P}) = S$;*
4. *$\mathrm{rank}(\boldsymbol{P}) = \dim(S)$;*
5. *$\boldsymbol{P}$ is symmetric and idempotent;*
6. *$\boldsymbol{P}$ has $\dim(S)$ eigenvalues equal to 1 and $p - \dim(S)$ equal to 0;*
7. *the orthogonal projection matrix onto $S^\perp$ is $\boldsymbol{I}_p - \boldsymbol{P}$.*

1.1.3 Norms

Definition 1.1.8 (Norm)**.** *Let $\mathcal{V}$ be a vector space. The function $x \in \mathcal{V} \mapsto \|x\| \in \mathbb{R}$ is a norm on $\mathcal{V}$ if it has the following property:*

1. *non-negativity: $\forall x \in \mathcal{V}, \|x\| \geq 0$;*
2. *definiteness: $\forall x \in \mathcal{V}, \|x\| = 0 \implies x = 0$;*
3. *triangle inequality: $\forall x, y \in \mathcal{V}, \|x + y\| \leq \|x\| + \|y\|$;*
4. *absolute homogeneity: $\forall x \in \mathcal{V}, \forall a \in \mathbb{R}, \|ax\| = |a|\|x\|$.*

Definition 1.1.9 (Vector norms)**.** *Let $\alpha \geq 1$. These are some norms on $\mathbb{R}^p$:*

1. *1-norm: $\forall x \in \mathbb{R}^p, \|x\|_1 = \sum_{i=1}^p |x_i|$;*
2. *2-norm: $\forall x \in \mathbb{R}^p, \|x\|_2 = \sqrt{\sum_{i=1}^p x_i^2}$;*

3. *α-norm:* $\forall x \in \mathbb{R}^p, \|x\|_\alpha = \sqrt[\alpha]{\sum_{i=1}^p x_i^\alpha}$;
4. *∞-norm:* $\forall x \in \mathbb{R}^p, \|x\|_\infty = \max_{1\leq i\leq n} |x_i|$;

Proposition 1.1.15 (Hölder and Cauchy-Schwarz inequalities)**.** *Let $\alpha, \beta \geq 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Then (Hölder inequality),*

$$\forall x, y \in \mathbb{R}^p : \quad |x^\top y| \leq \|x\|_\alpha \|y\|_\beta.$$

In particular (Cauchy-Schwarz inequality),

$$\forall x, y \in \mathbb{R}^p : \quad |x^\top y| \leq \|x\|_2 \|y\|_2 .$$

Property 1.1.16. *In $\mathbb{R}^p$, all norms are equivalent and in particular:*

1. $\|\cdot\|_2 \leq \|\cdot\|_1 \leq \sqrt{p}\,\|\cdot\|_2$;
2. $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \sqrt{p}\,\|\cdot\|_\infty$;
3. $\|\cdot\|_\infty \leq \|\cdot\|_1 \leq p\,\|\cdot\|_\infty$;
4. $\forall \alpha \geq 1 : \|\cdot\|_\infty \leq \|\cdot\|_\alpha \leq p^{1/\alpha}\,\|\cdot\|_\infty$;
5. $\forall x \in \mathbb{R}^p : \quad \lim_{\alpha\to\infty} \|x\|_\alpha = \|x\|_\infty$.

Definition 1.1.10 (Matrix norms)**.** *Let $\alpha \geq 1$ and $\|\cdot\|$ be a norm. These are some norms on $\mathbb{R}^{n\times p}$:*

1. *Frobenius norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_F = \sqrt{\sum_{1\leq i,j\leq n} \boldsymbol{a}_{ij}^2} = \sqrt{\operatorname{trace}(\boldsymbol{A}^\top \boldsymbol{A})}$;
2. *operator norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\| = \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|}{\|x\|}$;
3. *α-norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_\alpha = \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|_\alpha}{\|x\|_\alpha}$;
4. 1*-norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_1 = \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|_1}{\|x\|_1} = \max_{1\leq j\leq p} \sum_{i=1}^n |\boldsymbol{a}_{ij}|$;
5. *∞-norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_\infty = \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq n} \sum_{j=1}^p |\boldsymbol{a}_{ij}|$;
6. 2*- or spectral norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_2 = \sup_{x\neq 0} \frac{\|\boldsymbol{A}x\|_2}{\|x\|_2} = \sqrt{\lambda_{max}(\boldsymbol{A}^\top \boldsymbol{A})}$*, where λ_{max} is the largest eigenvalue. If $\boldsymbol{A}$ is square and diagonalizable, $\|\boldsymbol{A}\|_2 = \lambda_{max}(\boldsymbol{A})$;*
7. *max norm:* $\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \|\boldsymbol{A}\|_{max} = \max_{1\leq i,j\leq n} |\boldsymbol{a}_{ij}|$.

Let us remark that while 1- and ∞- norm can be computed in $O(np)$, this is more challenging for the spectral norm for it requires to diagonalize the matrix.

Property 1.1.17 (Compatibility)**.** *Operator norms are compatible with their vector norm. In particular,*

$$\forall \boldsymbol{A} \in \mathbb{R}^{n\times p}, \forall x \in \mathbb{R}^p; \|\boldsymbol{A}x\|_2 \leq \|\boldsymbol{A}\|_2 \|x\|_2 .$$

Property 1.1.18 (Mutual consistency)**.** *The operator norms for different sizes of matrices are mutually consistent:*

$$\forall \boldsymbol{A} \in \mathbb{R}^{n\times k}, \forall \boldsymbol{B} \in \mathbb{R}^{k\times p}, \|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|.$$

Property 1.1.19. *In $\mathbb{R}^{n\times p}$, all norms are equivalent and in particular, for every $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ with $r = \operatorname{rank}(\boldsymbol{A})$:*

1. $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \sqrt{r}\|\boldsymbol{A}\|_2 \leq \sqrt{\min(n,p)}\|\boldsymbol{A}\|_2$*;*
2. $\|\boldsymbol{A}\|_{max} \leq \|\boldsymbol{A}\|_2 \leq \sqrt{np}\|\boldsymbol{A}\|_{max}$*;*
3. $\frac{1}{\sqrt{p}}\|\boldsymbol{A}\|_\infty \leq \|\boldsymbol{A}\|_2 \leq \sqrt{n}\|\boldsymbol{A}\|_\infty$*;*
4. $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\|_1 \leq \|\boldsymbol{A}\|_2 \leq \sqrt{p}\|\boldsymbol{A}\|_1$*.*

Proposition 1.1.20 (Hölder inequality for matrices)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. Then,*

$$\|\boldsymbol{A}\|_2 \leq \sqrt{\|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty}.$$

Property 1.1.21 (Orthogonal invariance)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Q}_1 \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{p\times p}$ two orthogonal matrices. Then,*

1. $\|\boldsymbol{Q}_1\boldsymbol{A}\boldsymbol{Q}_2\|_F = \|\boldsymbol{A}\|_F$*;*
2. $\|\boldsymbol{Q}_1\boldsymbol{A}\boldsymbol{Q}_2\|_2 = \|\boldsymbol{A}\|_2$*.*

1.1.4 Singular value decomposition

Theorem 1.1.22 (The (thin) singular value decomposition)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}\backslash\{\boldsymbol{0}\}$ be a matrix and denote $r = \operatorname{rank}(\boldsymbol{A})$. Then,*

$$\begin{cases} \exists \boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_r, \\ \exists \boldsymbol{V} \in \mathbb{R}^{p\times r}, \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_r, \\ \exists \boldsymbol{D} \in \mathbb{R}^{r\times r}, \boldsymbol{D} \textit{ diagonal} \end{cases} \quad : \quad \boldsymbol{A} = \boldsymbol{UDV}^\top.$$

In addition, the entries of $\boldsymbol{D}$ are positive and uniquely defined:

$$\exists!(\sigma_1, \ldots, \sigma_r) \in (\mathbb{R}_+^*)^n : \sigma_1 \leq \cdots \leq \sigma_r \textit{ and } \boldsymbol{D} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_r \end{pmatrix} \textit{ up to a permutation.}$$

1. *$\{\sigma_1, \ldots, \sigma_r\}$ are the singular values of $\boldsymbol{A}$;*
2. *the columns of $\boldsymbol{U}$ are its normed left singular vectors (and the normed eigenvectors of $\boldsymbol{AA}^\top$ with eigenvalues $\{\sigma_1^2, \ldots, \sigma_r^2\}$);*

3. *the columns of $\boldsymbol{V}$ are its normed right singular vectors (and the normed eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ with eigenvalues $\{\sigma_1^2, \dots, \sigma_r^2\}$).*

Besides observing that we have the decomposition:

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

the previous theorem states that, for every $i \in [\![1, r]\!]$, the singular vectors u_i and v_i (respectively the i^{th} column of $\boldsymbol{U}$ and $\boldsymbol{V}$) verify:

$$\boldsymbol{A}^\top u_i = \sigma_i v_i \quad \text{and} \quad \boldsymbol{A} v_i = \sigma_i u_i.$$

Property 1.1.23. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}\backslash\{\mathbf{0}\}$ be a matrix with singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$. With the same notation as before, we have:*

1. $\|\boldsymbol{A}\|_2 = \sigma_r$*;*
2. $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$*;*
3. $\operatorname{range}(\boldsymbol{A}) = \operatorname{range}(\boldsymbol{U})$*;*
4. $\ker(\boldsymbol{A}) = \operatorname{range}(\boldsymbol{V})^\perp$*;*
5. $\boldsymbol{U}\boldsymbol{U}^\top$ *is the projection matrix onto* $\operatorname{range}(\boldsymbol{A})$*;*
6. $\boldsymbol{V}\boldsymbol{V}^\top$ *is the projection matrix onto* $\operatorname{range}(\boldsymbol{A}^\top) = \ker(\boldsymbol{A})^\perp$*;*
7. *if $\boldsymbol{V}$ is completed by $\boldsymbol{V}_0$ such that $[\boldsymbol{V}|\boldsymbol{V}_0]$ is orthogonal, $\boldsymbol{V}_0\boldsymbol{V}_0^\top$ is the projection matrix onto* $\ker(\boldsymbol{A})$*;*
8. *if $\boldsymbol{U}$ is completed by $\boldsymbol{U}_0$ such that $[\boldsymbol{U}|\boldsymbol{U}_0]$ is orthogonal, $\boldsymbol{U}_0\boldsymbol{U}_0^\top$ is the projection matrix onto* $\ker(\boldsymbol{A}^\top) = \operatorname{range}(\boldsymbol{A})^\perp$.

At a last very interesting property of the singular value decomposition, it makes it possible to define a pseudo-inverse for singular matrices.

Definition 1.1.11 (Pseudo-inverse (or Moore-Penrose inverse))**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}\backslash\{\mathbf{0}\}$ be a matrix with singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$. The pseudo-inverse of $\boldsymbol{A}$ is*

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\top.$$

Besides, $\mathbf{0}^+ = \mathbf{0}$.

Property 1.1.24. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ be a rectangular matrix. Then,*

1. *$\boldsymbol{A}^+$ is the unique minimal Frobenius norm solution in* $\arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times n}} \|\boldsymbol{A}\boldsymbol{B} - \boldsymbol{I}_n\|_F$*;*
2. *$\boldsymbol{A}^+$ is the unique $p \times n$-matrix that satisfies the Moore-Penrose conditions:*
   (a) $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$*;*
   (b) $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$*;*
   (c) $\left(\boldsymbol{A}\boldsymbol{A}^+\right)^\top = \boldsymbol{A}\boldsymbol{A}^+$*;*
   (d) $\left(\boldsymbol{A}^+\boldsymbol{A}\right)^\top = \boldsymbol{A}^+\boldsymbol{A}$.

3. $\boldsymbol{A}\boldsymbol{A}^{+}$ *is the projector onto* $\mathrm{range}(\boldsymbol{A})$*;*
4. $\boldsymbol{A}^{+}\boldsymbol{A}$ *is the projector onto* $\mathrm{range}(\boldsymbol{A}^\top)$*;*
5. $\boldsymbol{A}^{+} = \lim_{\lambda \to 0^+}(\boldsymbol{A}^\top\boldsymbol{A} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{A}^\top$*, and in particular, if* $\mathrm{rank}(\boldsymbol{A}) = p$*, then* $\boldsymbol{A}^{+}$ *is a left inverse of* $\boldsymbol{A}$ *and* $\boldsymbol{A}^{+} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$*;*
6. $\boldsymbol{A}^{+} = \lim_{\lambda \to 0^+}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top + \lambda \boldsymbol{I}_n)^{-1}$*, and in particular, if* $\mathrm{rank}(\boldsymbol{A}) = n$*, then* $\boldsymbol{A}^{+}$ *is a right inverse of* $\boldsymbol{A}$ *and* $\boldsymbol{A}^{+} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}$*;*
7. *in particular, if* $\boldsymbol{A}$ *is square and non-singular,* $\boldsymbol{A}^{+} = \boldsymbol{A}^{-1}$*;*
8. $(\boldsymbol{A}^{+})^{+} = \boldsymbol{A}$*;*
9. $\forall \alpha \neq 0; (\alpha\boldsymbol{A})^{+} = \alpha^{-1}\boldsymbol{A}^{+}$*;*
10. *for every* $x \in \mathbb{R}^n\backslash\{0\}$*,* $x^{+} = \frac{x^\top}{\|x\|_2^2}$*;*
11. *for* $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ *be two orthogonal matrices,* $(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V})^{+} = \boldsymbol{V}^{-1}\boldsymbol{A}^{+}\boldsymbol{U}^{-1} = \boldsymbol{V}^\top\boldsymbol{A}^{+}\boldsymbol{U}^\top$*;*
12. $\boldsymbol{A}^{+} = (\boldsymbol{A}^\top\boldsymbol{A})^{+}\boldsymbol{A}^\top = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{+}$*;*
13. $(\boldsymbol{A}^\top)^{+} = (\boldsymbol{A}^{+})^\top$*;*

The pseudo-inverse is particularly useful for expressing solutions to arbitrary linear systems (see the exercises).

1.1.5 Symmetric positive definite matrices

Definition 1.1.12. *A square matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is symmetric positive definite if:*

1. $\boldsymbol{A} = \boldsymbol{A}^\top$*;*
2. $\forall x \in \mathbb{R}^n\backslash\{0\}, x^\top\boldsymbol{A}x > 0$.

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is symmetric positive semi-definite if:*

1. $\boldsymbol{A} = \boldsymbol{A}^\top$*;*
2. $\forall x \in \mathbb{R}^n, x^\top\boldsymbol{A}x \geq 0$.

Property 1.1.25. *A square symmetric matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is:*

1. *positive definite if and only if its eigenvalues are positive;*
2. *positive semi-definite if and only if its eigenvalues are non-negative;*
3. *positive semi-definite if and only if* $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\top$ *for some* $\boldsymbol{B} \in \mathbb{R}^{n\times k}$ *(*$k \leq n$*).*

Theorem 1.1.26 (Cholesky decomposition)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *be a symmetric positive definite matrix. Then, there exists a unique lower triangular matrix* $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ *with positive diagonal entries such that:*

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top.$$

If $\boldsymbol{A}$ *is only symmetric positive semi-definite, then the diagonal entries of* $\boldsymbol{L}$ *are non-negative.*

As for non-negative scalar, it is possible to define the square root of a symmetric positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$:

1. the most general definition is $\boldsymbol{A}^{1/2} = \boldsymbol{Q}\boldsymbol{D}^{1/2}\boldsymbol{Q}^\top$, where $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ is the Schur decomposition of $\boldsymbol{A}$ and the square root of a diagonal matrix is computed entry-wise. This provides the unique symmetric positive semi-definite matrix $\boldsymbol{A}^{1/2}$ such that ${\boldsymbol{A}^{1/2}}^2 = \boldsymbol{A}^{1/2}{\boldsymbol{A}^{1/2}}^\top = \boldsymbol{I}_n$;
2. less frequently (but in a consistent idea with multivariate normal vectors), $\boldsymbol{A}^{1/2} = \boldsymbol{L}$, where $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ is the Cholesky decomposition of $\boldsymbol{A}$.

With both definitions, considering $X \sim \mathcal{N}(0, \boldsymbol{I}_n)$, we have $\mathbb{V}(\boldsymbol{A}^{1/2}X) = \boldsymbol{A}^{1/2}\,\mathbb{V}(X){\boldsymbol{A}^{1/2}}^\top = \boldsymbol{A}$. That is, $\boldsymbol{A}^{1/2}X \sim \mathcal{N}(0, \boldsymbol{A})$, which is what is of interest when simulating multivariate Gaussian vectors.

1.2 Functions

1.2.1 Differential calculus

Definition 1.2.1 (Fréchet-differentiability)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}^n$ is differentiable at $x \in \mathbb{R}^n$ if*

$$\exists \boldsymbol{J}_f(x) \in \mathbb{R}^{n\times p} : \quad \lim_{h\to 0} \frac{f(x+h) - f(x) - \boldsymbol{J}_f(x)h}{\|h\|_2} = 0.$$

$\boldsymbol{J}_f(x) \in \mathbb{R}^{n\times p}$ is called the Jacobian matrix of f at x.

f is said differentiable if it is differentiable at all $x \in \mathbb{R}^p$.

Property 1.2.1. *Let $f : \mathbb{R}^p \to \mathbb{R}^n$ be a differentiable function. Then, for every $x \in \mathbb{R}^p$*

1. *the Jacobian matrix $\boldsymbol{J}_f(x)$ is unique and for every $h \in \mathbb{R}^p$,*

$$\boldsymbol{J}_f(x)h = \lim_{t\to 0^+} \frac{f(x+th) - f(x)}{t};$$

2. *f is continuous at x;*
3. *the set of differentiable functions is a vector space and the operator $f \mapsto \boldsymbol{J}_f$ is linear on this space (with the notation introduced just below, this means that ∇ is a linear operator).*

In practice, we are mainly interested in scalar-valued functions $f : \mathbb{R}^p \to \mathbb{R}$. Then, f is differentiable at $x \in \mathbb{R}^p$ if there exists $L_f(x) \in \mathbb{R}^p$ such that

$$\lim_{h\to 0} \frac{f(x+h) - f(x) - L_f(x)^\top h}{\|h\|_2} = 0,$$

where $L_f(x)$ is called the gradient of f at x, denoted $\nabla f(x)$ (actually $\nabla f(x)^\top = \boldsymbol{J}_f(x)$). In other words, there exists a function $\varepsilon_x : \mathbb{R} \to \mathbb{R}$ such that $\lim_0 \varepsilon_x = 0$ and for every $h \in \mathbb{R}^p \backslash \{0\}$,

$$f(x+h) = f(x) + \nabla f(x)^\top h + \|h\|_2\, \varepsilon_x(\|h\|_2).$$

This leads to a very important interpretation of the gradient: it defines a hyperplan tangent to the graph of f at x or a first order approximation of f at x. We can even refine this approximation thanks to the descent lemma.

Proposition 1.2.2 (Descent lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with L-Lipschitz continuous gradient ($L > 0$):*

$$\forall x, y \in \mathbb{R}^n : \|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 .$$

Then

$$\forall x, y \in \mathbb{R}^n : f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2 .$$

Definition 1.2.2 (Twice differentiability)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is twice differentiable at $x \in \mathbb{R}^n$ if f is differentiable at x and ∇f is differentiable at x:*

$$\exists \boldsymbol{H}_f(x) \in \mathbb{R}^{p \times p} : \quad \lim_{h \to 0} \frac{\nabla f(x+h) - \nabla f(x) - \boldsymbol{H}_f(x)h}{\|h\|_2} = 0.$$

$\boldsymbol{H}_f(x)$ is called the Hessian matrix of f at x and is denoted $\nabla^2 f(x)$.

f is said twice differentiable if it is twice differentiable at all $x \in \mathbb{R}^p$.

It is essential for us to be able to compute the gradient and the Hessian of a function $f : \mathbb{R}^p \to \mathbb{R}$. For this purpose, we establish the link with partial derivatives.

Definition 1.2.3 (Partial derivative)**.** *The j^{th} partial derivative of $f : \mathbb{R}^n \to \mathbb{R}$ ($j \in [\![1, n]\!]$) at $x \in \mathbb{R}^n$ is:*

$$\frac{\partial f}{\partial x_j}(x) = \lim_{t \to 0} \frac{f(x + te_j) - f(x)}{t} = \boldsymbol{J}_f(x) e_j,$$

if it exists, where $e_j = (\mathbf{1}_{j=1}, \ldots, \mathbf{1}_{j=n})$.

The $(i, j)^{th}$ second partial derivative of f ($i, j \in [\![1, n]\!]$) is:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \lim_{t \to 0} \frac{\frac{\partial f}{\partial x_j} f(x + te_i) - f(x)}{t},$$

if it exists.

In practice, computing a partial derivative $\frac{\partial f}{\partial x_j}$ boils down to differentiate f while considering all variables other than x_j fixed. Now, the link between differentiability and the existence of partial derivatives is not as easy as it seems. This is enlightened in the next two results.

Proposition 1.2.3. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then it has all its partial derivatives and*

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_x}(x) \end{pmatrix}.$$

More generally, by denoting $(f_i)_{1 \leq i \leq p}$ the p components of $f : \mathbb{R}^n \to \mathbb{R}^p$,

$$\boldsymbol{J}_f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1}(x) & \dots & \frac{\partial f_p}{\partial x_n}(x) \end{pmatrix}.$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then it as all its second partial derivatives and

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}.$$

The converse is not true. For example, let us consider the function $f : x \in \mathbb{R}^2 \mapsto \frac{x_1^4 x_2}{x_1^6 + x_2^3}$ if $x \neq 0$ and $f(0) = 0$. Then, we get easily that $\frac{\partial f}{\partial x_1}(0) = 0$ and $\frac{\partial f}{\partial x_2}(0) = 0$. However, $\lim_{t \to 0} f((t, t^2)) = \lim_{t \to 0} \frac{t^4 t^2}{t^6 + t^6} = \frac{1}{2} \neq 0$ so f is not continuous at 0. Consequently, f cannot be differentiable at $x = 0$ while partial derivatives exist.

Theorem 1.2.4. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ has all its partial derivatives at $x \in \mathbb{R}^n$ and if they are continuous at x, then f is differentiable at x.*

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, if it has all its second partial derivatives at $x \in \mathbb{R}^n$ and if they are continuous at x, then f is twice differentiable at x.

Theorem 1.2.5 (Schwarz'z theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function at $x \in \mathbb{R}^n$. Then its Hessian matrix at x is symmetric.*

Schwarz's theorem says that the order of differentiation does not matter: $\frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial^2}{\partial x_j \partial x_i}$. Let us remark that, in a broader view, gradients and Hessian matrices can be defined without requiring differentiability. In this case, we are not guaranteed to have the good properties expected when a function is differentiable:

1. the continuity of the function;
2. the tangent space;
3. the symmetry of the Hessian;
4. the Taylor expansion (see thereafter).

Since the Hessian is symmetric, it is diagonalizable and we have the following characterization in order to check if the descent lemma hold.

Proposition 1.2.6. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Then ∇f is L-Lipschitz continuous (for some $L > 0$) if and only if for every $x \in \mathbb{R}^n$, $\lambda_{max}(\nabla^2 f(x)) \leq L$, where λ_{max} denotes the largest eigenvalue of a diagonalizable matrix.*

Proposition 1.2.7 (Chain rule)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}^n$ be two functions. If g is differentiable at $x \in \mathbb{R}^p$ and f is differentiable at $g(x)$, then $f \circ g$ is differentiable at x and*

$$\nabla(f \circ g)(x) = \boldsymbol{J}_g(x)^\top \nabla f(g(x)),$$

where $\boldsymbol{J}_g(x)$ is the Jacobian matrix of g at x.

For $n = 1$,

$$\nabla(f \circ g)(x) = f'(g(x))\nabla g(x).$$

More generally, if $f : \mathbb{R}^n \to \mathbb{R}^m$,

$$\boldsymbol{J}_{f\circ g}(x) = \boldsymbol{J}_f(g(x))\boldsymbol{J}_g(x),$$

with self-evident notation.

Example 1.2.1 (Gradients of classical functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. The following examples are differentiable everywhere the gradient is given.*

1. *If f is constant, f is differentiable and $\nabla f(x) = 0, \forall x \in \mathbb{R}^n$.*
2. *If $f(x) = \|x\|_2$, $\nabla f(x) = \frac{x}{\|x\|_2}, \forall x \in \mathbb{R}^n\backslash\{0\}$.*
3. *If $f(x) = \|x\|_2^2$, $\nabla f(x) = 2x, \forall x \in \mathbb{R}^n$.*
4. *If $f(x) = x^\top \boldsymbol{A}x + b^\top x + c$, $\nabla f(x) = (\boldsymbol{A} + \boldsymbol{A}^\top)x + b, \forall x \in \mathbb{R}^n$.*
5. *If $f(x) = \|\boldsymbol{A}x + b\|_2^2$, $\nabla f(x) = 2\boldsymbol{A}^\top(\boldsymbol{A}x + b), \forall x \in \mathbb{R}^n$.*

Theorem 1.2.8 (Taylor's theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function at $x \in \mathbb{R}^n$. Then, there exists a function $\varepsilon_x : \mathbb{R} \to \mathbb{R}$ such that $\lim_0 \varepsilon_x = 0$ and for every $h \in \mathbb{R}^n\backslash\{0\}$:*

$$f(x + h) = f(x) + \nabla f(x)^\top h + \frac{1}{2}h^\top \nabla^2 f(x)h + \|h\|_2^2\,\varepsilon_x(\|h\|_2^2).$$

We end this section with a very useful result for probability.

Definition 1.2.4 (C^1-diffeomorphism)**.** *Let U and V be two open sets from $\mathbb{R}^n$. A mapping $f : U \to V$ is a C^1-diffeomorphism if:*

1. *f is bijective;*
2. *f is differentiable on U with continuous partial derivatives;*
3. *the inverse of f is differentiable on V with continuous partial derivatives.*

Property 1.2.9. *Let U and V be two open sets from $\mathbb{R}^n$. A mapping $\varphi : U \to V$ is a C^1-diffeomorphism if and only if:*

1. *φ is bijective;*
2. *φ has all its partial derivatives and they are continuous;*
3. *for every $x \in U$, the Jacobian matrix of φ at x has a non-zero determinant: $\det(\boldsymbol{J}_\varphi(x)) \neq 0$.*

In addition, if φ is a C^1-diffeomorphism, then for every $x \in U$, denoting $y = f(x) \in V$:

$$\det(\boldsymbol{J}_\varphi(x)) = \det(\boldsymbol{J}_{\varphi^{-1}}(y))^{-1}.$$

Theorem 1.2.10 (Integration by substitution (or change of variables))**.** *Let U and V be two open sets from $\mathbb{R}^n$, $\varphi : U \to V$ is a C^1-diffeomorphism and $f : V \to \mathbb{R}$ a measurable function. Then*

$$\int_V f = \int_U (f \circ \varphi) |\det(\boldsymbol{J}_\varphi)|,$$

or saying it in a different manner with the C^1-diffeomorphism $\psi = \varphi^{-1}$,

$$\int_V f(x)\,dx = \int_U (f \circ \psi^{-1})(y) |\det(\boldsymbol{J}_{\psi^{-1}}(y))|\,dy = \int_U (f \circ \psi^{-1})(y)\,\frac{dy}{|\det(\boldsymbol{J}_\psi(\psi^{-1}(y)))|}.$$

The last formula says that if one sets the change of variables $y = \psi(x)$, then

$$dy = |\det(\boldsymbol{J}_\psi(x))|\,dx, \quad \text{so} \quad dx = \frac{dy}{|\det(\boldsymbol{J}_\psi(\psi^{-1}(y)))|},$$

since $x = \psi^{-1}(y)$. Starting from this, we also have $dx = |\det(\boldsymbol{J}_{\psi^{-1}}(y))|\,dy$.

1.2.2 Convex functions

Definition 1.2.5 (Convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in (0,1)$,*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y).$$

In addition, f is strictly convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in (0,1)$,

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

Proposition 1.2.11 (Elementary properties)**.** *We consider two convex functions $f_1 \colon \mathbb{R}^n \to \mathbb{R}$ and $f_2 \colon \mathbb{R}^n \to \mathbb{R}$.*

1. *For any non-negative α and β, $\alpha f_1 + \beta f_2$ is convex.*
2. $x \mapsto \max(f_1(x), f_2(x))$ *is convex.*
3. *Let $(A, b) \in \mathbb{R}^{d \times d'} \times \mathbb{R}^n$, then $x \in \mathbb{R}^{d'} \mapsto f_1(Ax + b)$ is convex.*
4. *For every $x, y \in \mathbb{R}^n$ and $t \geq 1$, denoting $z_t = x + t(y - x)$, $f_1(z_t) \geq f_1(x) + t(f_1(y) - f_1(x))$.*
5. *Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing and $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function, then $\varphi \circ f$ is convex.*
6. *The perspective function of f_1:*

$$g\colon (x,t) \in \mathbb{R}^n \times \mathbb{R} \mapsto \begin{cases} t f_1(\frac{1}{t}x) & \text{if } t > 0 \\ \infty & \text{otherwise,} \end{cases}$$

is convex.

Theorem 1.2.12 (Jensen's inequality)**.** *A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex if and only if: $\forall n \geq 2, \forall (x_i)_{1 \leq i \leq n} \in (\mathbb{R}^n)^n, \forall (t_i)_{1 \leq i \leq n}$ such that $t_i \geq 0, \forall i \in [n]$ and $\sum_{i=1}^n t_i = 1$,*

$$f\left(\sum_{i=1}^n t_i x_i\right) \leq \sum_{i=1}^n t_i f(x_i).$$

Example 1.2.2 (Convex functions)**.**

1. *Every norm $\|\cdot\|$ on $\mathbb{R}^n$ is convex (this comes from the triangle inequality and homogeneity).*
2. *p-norms for $1 < p < \infty$ are strictly convex and only convex for $p = 1$ and $p = \infty$.*
3. *For $\varphi\colon \mathbb{R} \to \mathbb{R}$ convex nondecreasing and every norm $\|\cdot\|$ on $\mathbb{R}^n$, $\varphi(\|\cdot\|)$ is convex. In particular, $\|\cdot\|^p$ is convex provided that $p \geq 1$.*
4. *For a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$, $f\colon x \in \mathbb{R}^n \to x^\top A x$ is convex. If A is positive definite, f is strictly convex.*

Proposition 1.2.13 (Coordinate supremum)**.** *Let $\mathcal{Y} \subset \mathbb{R}^p$ be a non-empty set and $F\colon (x,y) \in \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}$ be a function convex in x (that is, $\forall y \in \mathcal{Y}, F(\cdot, y)$ is convex). Then $f\colon x \in \mathbb{R}^n \mapsto \sup_{y \in \mathcal{Y}} F(x,y)$ is convex.*

Proposition 1.2.14 (Coordinate infimum)**.** *Let $F\colon (x,y) \in \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be a (jointly) convex function. Then $f\colon x \in \mathbb{R}^n \mapsto \inf_{y \in \mathbb{R}^{d'}} F(x,y)$ is convex.*

Definition 1.2.6 (Strong convexity)**.** *Let $\mu \in \mathbb{R}_+^*$. A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex.*

Proposition 1.2.15 (Characterization of a strongly convex function)**.** *Let $\mu \in \mathbb{R}_+^*$. A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if and only if*

$$\forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n, \forall t \in (0,1)\colon \quad f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)-\frac{\mu}{2}t(1-t)\|x-y\|_2^2.$$

Proposition 1.2.16 (Relation between convexities)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function.*

$$f \text{ strongly convex} \implies f \text{ strictly convex} \implies f \text{ convex.}$$

Proposition 1.2.17 (First-order conditions of convexity)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function.*

1. *f is convex if and only if:*

$$\forall x, y \in \mathbb{R}^n\colon \quad f(y) \geq f(x) + \nabla f(x)^\top (y-x).$$

2. *f is convex if and only if:*

$$\forall x, y \in \mathbb{R}^n\colon \quad (\nabla f(y) - \nabla f(x))^\top (y-x) \geq 0.$$

3. *f is strictly convex if and only if:*

$$\forall x, y \in \mathbb{R}^n, x \neq y\colon \quad f(y) > f(x) + \nabla f(x)^\top (y-x).$$

4. *Let $\mu \in \mathbb{R}_+^*$. f is μ-strongly convex if and only if:*

$$\forall x, y \in \mathbb{R}^n\colon \quad f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|_2^2.$$

Remark 1.2.1. *For convex functions,*

$$\forall x, y \in \mathbb{R}^n\colon \quad f(y) \geq f(x) + \nabla f(x)^\top (y-x).$$

This is perhaps the most important property of convex functions since it shows that from a local information ($\nabla f(x)$), we can derive a global information concerning f (we have a global underestimator). In particular, if $\nabla f(x) = 0$, then x is a global minimizer.

Proposition 1.2.18 (Second-order conditions of convexity)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. We denote λ_{min} the smallest eigenvalue of a diagonalizable matrix.*

1. *f is convex if and only if:*

$$\forall x \in \mathbb{R}^n\colon \quad \lambda_{min}(\nabla^2 f(x)) \geq 0.$$

2. *f is strictly convex if and only if:*

$$\forall x \in \mathbb{R}^n : \quad \lambda_{min}(\nabla^2 f(x)) > 0.$$

3. *Let $\mu \in \mathbb{R}_+^*$. f is μ-strongly convex if and only if:*

$$\forall x \in \mathbb{R}^n : \quad \lambda_{min}(\nabla^2 f(x)) \geq \mu.$$

1.2.3 Minimization

Let us consider a function $f : \mathbb{R}^n \to \mathbb{R}$ along with the optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x). \tag{P1}$$

Definition 1.2.7 (Minimizers)**.** *A point $x^\star \in \mathbb{R}^n$ is a global minimizer of Problem (P1) if*

$$\forall x \in \mathbb{R}^n \colon f(x^\star) \leq f(x).$$

A point $x^\star \in \mathbb{R}^n$ is a local minimizer of Problem (P1) if there exists $\epsilon > 0$, $N = \{x \in \mathbb{R}^n : \|x^\star - x\|_2 \leq \epsilon\}$ such that

$$\forall x \in N \colon f(x^\star) \leq f(x).$$

Remark 1.2.2. *Global minimizers may not exist, as we can see for:*

- ◇ $f(x) = \frac{1}{x^2}$ *(agreeing that $f(0) = \infty$);*
- ◇ $f(x) = \exp(-x^2)$*;*
- ◇ $f(x) = x + \chi_{\mathbb{R}_+^*}(x)$.

As we can see, the existence of minimizers of Problem (P1) is not guaranteed, even though f is continuous. Consequently, the remaining of this section is devoted to characterizing the existence of minimizers and their properties. We do so, first, for constrained optimization problems.

Theorem 1.2.19 (Weierstrass extreme value theorem)**.** *Let $C \subset \mathbb{R}^n$ be a non-empty compact set and assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous. Then $\arg\min_{x \in C} f(x)$ and $\arg\max_{x \in C} f(x)$ are non-empty.*

Now, we state an existence result for unconstrained optimization problems.

Definition 1.2.8 (Coercivity)**.** *A function $f \colon \mathbb{R}^n \to \mathbb{R}$ is coercive if for every sequence*

$(x_n)_{n\in\mathbb{N}}$ *such that* $\lim_{n\to\infty}\|x_n\| = \infty$,

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

Theorem 1.2.20 (Existence of a solution for unconstrained problems)**.** *Assume that* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuous and coercive. Then* $\arg\min_{x\in C} f(x)$ *is non-empty,* i.e. *Problem (P1) admits a global minimizer.*

Knowing that continuous[1] functions can attain their minima, let us go back to convex functions.

Proposition 1.2.21 (Minimizers of convex functions)**.** *Assume that* $f : \mathbb{R}^n \to \mathbb{R}$ *is convex. Then*

1. *a local minimizer of* f *is a global one;*
2. *the set of minimizers of* f *is convex;*
3. *if* f *is strictly convex, then* f *has a unique minimizer.*

Remark 1.2.3. *When an estimator is built as a minimizer of an optimization problem, we are interested in a global minimizer. However, in order to verify that a point* $x^\star$ *is a global minimizer, one would have to compare* $f(x^\star)$ *to every other value* $f(x)$*, no matter how far from* $x^\star$*, the point* x *is. The fact that for convex functions, local minimizers are also global minimizers essentially explains our interest in convex optimization and the availability of efficient numerical methods. Indeed, local minimizers can be found by greedy approaches (such as gradient descent).*

Differentiability plays a key role in optimization. First because it helps characterizing convexity (see Proposition 1.2.17), second (this is a consequence) because it is inherent in the mainly used optimality condition: Fermat's rule.

Theorem 1.2.22 (Fermat's rule)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a convex and differentiable function.* $x^\star \in \mathbb{R}^n$ *is a global minimizer of* f *if and only if*

$$\nabla f(x^\star) = 0.$$

[1] A refined theory makes use of the notion of lower semi-continuous functions.

1.3 Exercises

1.3.1 Matrices

Exercise 1.1 (Range and null space (proof or Property 1.1.5))**.** Let $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ be a matrix. Prove that $\text{range}(\boldsymbol{A})^\perp = \ker(\boldsymbol{A}^\top)$.

Exercise 1.2 (Projection matrix (proof of Property 1.1.14))**.**

1. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Show that the projection of $x \in \mathbb{R}^n$ onto $\text{range}(\boldsymbol{X})$ is $\boldsymbol{X}\alpha$, with $\alpha \in \mathbb{R}^p$ such that $\boldsymbol{X}^\top \boldsymbol{X}\alpha = \boldsymbol{X}^\top x$.
2. Let $S \subset \mathbb{R}^p$ be a subspace and $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ its orthogonal projection matrix. Prove that $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^\top$, where $\boldsymbol{V} = \left[v_1 | \dots | v_{\dim(S)}\right]$ and $\left\{v_1 | \dots | v_{\dim(S)}\right\}$ is an orthonormal basis for S.
3. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be a matrix and its singular value decomposition. Show that $\boldsymbol{U}\boldsymbol{U}^\top$ is the projector onto $\text{range}(\boldsymbol{A})$.

Exercise 1.3 (Equivalence of $\|\cdot\|_2$ and $\|\cdot\|_1$)**.**

1. Prove that $\forall x \geq 0, \sqrt{1+x^2} \leq 1 + x$.
2. Deduce that $\|\cdot\|_2 \leq \|\cdot\|_1$.
3. Show that $\|\cdot\|_1 \leq \sqrt{p}\,\|\cdot\|_2$ in $\mathbb{R}^p$.

Exercise 1.4 (Solution to linear systems)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n \times p} \backslash \{0\}$ and $y \in \mathbb{R}^n$.

1. Show that $\boldsymbol{A}^+ y \in \arg\min_{x \in \mathbb{R}^p} \|\boldsymbol{A}x - y\|_2$.
2. Simplify this solution when $\text{rank}(\boldsymbol{A}) = p$ and $\text{rank}(\boldsymbol{A}) = n$.
3. Prove that $y \in \text{range}(\boldsymbol{A}) \iff \boldsymbol{A}\boldsymbol{A}^+ y = y$.
4. Characterize a solution, denoted x_0, to the linear system $\boldsymbol{A}x = y$, where the unknown quantity is x.
5. Show that for arbitrary $z \in \mathbb{R}^p$, $x_0 + (\boldsymbol{I}_p - \boldsymbol{A}^+\boldsymbol{A})z$ is also solution.

Exercise 1.5 (Symmetric positive semi-definite matrices (proof of Property 1.1.25))**.** Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square symmetric matrix.

1. Show that $\boldsymbol{A}$ is positive semi-definite if and only if its eigenvalues are non-negative.
2. Prove that $\boldsymbol{A}$ is positive semi-definite if and only if $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\top$ for some $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ $(k \leq n)$.

1.3.2 Functions and matrices

Exercise 1.6 (On the chain rule)**.**

1. Compute the gradients of $\|\cdot\|_2^2$ and $f : x \in \mathbb{R}^n \mapsto x^\top \boldsymbol{A} x + b^\top x + c$, where $\boldsymbol{A}$ is a square symmetric matrix.
2. Show that $\forall x \in \mathbb{R}^n, f(x) = x^\top \boldsymbol{A} x \geq 0 \iff \boldsymbol{A}$ is semi-definite positive.
3. For a C^1-diffeomorphism $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, show that for every $x \in \mathbb{R}^n$, $\det(J_\varphi(x)) = \det(J_{\varphi^{-1}}(y))^{-1}$, where $y = f(x)$.

Exercise 1.7. Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, $b \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. We denote $f : x \in \mathbb{R}^p \mapsto \|\boldsymbol{A}x - b\|_2^2 + \lambda \|x\|_2^2$.

1. Give a condition for f to be convex, or μ-strongly convex ($\mu > 0$).
2. Show that if $\boldsymbol{A} \neq \boldsymbol{0}$, ∇f is Lipschitz continuous.
3. What about $g : x \in \mathbb{R}^p \mapsto \mathrm{e}^{-b^\top \boldsymbol{A} x}$?

Exercise 1.8 (The descent lemma (proof of Proposition 1.8))**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with L-Lipschitz continuous gradient ($L > 0$). Let $x, y \in \mathbb{R}^n$ and $g : t \in [0, 1] \mapsto f(x + t(y - x))$.

1. Show that $f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^\top (y - x)\, dt$.
2. Deduce that $f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2$.

Exercise 1.9 (Fermat's rule (proof of Theorem 1.2.22))**.** Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function and $x^\star \in \mathbb{R}^n$.

1. Prove that if $\nabla f(x^\star) = 0$, then $x^\star \in \mathbb{R}^n$ is a global minimizer of f.
2. By remarking that $\|\nabla f(x^\star)\|_2^2 = \lim_{t\to 0} \frac{f(x^\star + t\nabla f(x^\star)) - f(x^\star)}{t}$, show that if $x^\star \in \mathbb{R}^n$ is a global minimizer of f, then $\nabla f(x^\star) = 0$.

Chapter 2

Probability

2.1 Probability space

2.1.1 First definitions

Definition 2.1.1 (Measurable space)**.** *$(\Omega, \mathcal{A})$ is a measurable space if $\mathcal{A}$ is a σ-algebra on the set Ω:*

- ◇ $\Omega \in \mathcal{A}$*;*
- ◇ $A \in \mathcal{A} \implies A^c \in \mathcal{A}$*;*
- ◇ $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \implies \cup_{n\in\mathbb{N}} A_n \in \mathcal{A}$*.*

Definition 2.1.2 (Probability measure)**.** *A positive measure μ on a measurable space $(\Omega, \mathcal{A})$ is a function from $\mathcal{A}$ to $\mathbb{R}_+$.*

In addition, μ is a probability measure if:

- ◇ $\mu : \mathcal{A} \to [0, 1]$*;*
- ◇ $\mu(\Omega) = 1$*;*
- ◇ *σ-additivity: $\mu(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$ for any disjoint events $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$.*

The triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space when μ is a positive measure and a probability space when μ is a probability measure.

From now on, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Property 2.1.1. *One has:*

- ◇ $\mathbb{P}(\emptyset) = 0$*;*
- ◇ $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$ *for any* $A \in \mathcal{A}$*;*
- ◇ $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$*;*
- ◇ $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$ *for any* $(A, B) \in \mathcal{A}^2$*.*

Definition 2.1.3 (σ-algebra generated by an event)**.** *Let $\mathcal{P}(\Omega)$ be the set of all subsets of Ω and $C \subset \mathcal{P}(\Omega)$ be a set of subsets of Ω. The σ-algebra generated by C, denoted $\sigma(C)$, is the smallest σ-algebra that contains C:*

$$\sigma(C) = \bigcap_{\substack{\mathcal{B}\subset\mathcal{P}(\Omega):\\ \mathcal{B}\ \sigma\text{-algebra and } C\subset\mathcal{B}}} \mathcal{B}.$$

Two important measures that are not probabilities are:

- the counting measure on $(\mathbb{M}, \mathcal{P}(\mathbb{M}))$, where $\mathbb{M} \subset \mathbb{R}$ is a countable subset (often $\mathbb{M} = \mathbb{N}$) is $\mu(A) = |A|$ for any $A \in \mathcal{P}(\mathbb{M})$ ($\mu(A) = \infty$ if A is infinite);
- the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the σ-algebra generated by open sets of $\mathbb{R}$ (also called Borelian σ-algebra), is

$$\mu(A) = \inf\left\{\sum_{n\in\mathbb{N}}(b_n - a_n) : (a_n)_{n\in\mathbb{N}} \subset \mathbb{R}, (b_n)_{n\in\mathbb{N}} \subset \mathbb{R}, A \subset \bigcup_{n\in\mathbb{N}}(a_n, b_n)\right\},$$

for any $A \in \mathcal{B}$.

2.1.2 Conditioning and independence

Definition 2.1.4 (Conditional probability)**.** *Let $(A, B) \in \mathcal{A}^2$ with $\mathbb{P}(B) > 0$. The conditional probability of A given B is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$.*

Proposition 2.1.2. *Let $B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$. Then $A \in \mathcal{A} \mapsto \mathbb{P}(A|B)$ is a probability measure on $(\Omega, \mathcal{A})$, called conditional probability.*

In addition, for $A \in \mathcal{A}$ with $\mathbb{P}(A) > 0$, we have $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A\cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$.

Proposition 2.1.3 (Law of total probability and Bayes' Theorem)**.** *Let $(B_i)_{i\in I} \subset \mathcal{A}$ be a finite or countable set of events such that $(B_i)_{i\in I}$ is a partition of Ω and $\mathbb{P}(B_i) > 0$ for all $i \in I$. Then, for all $A \in \mathcal{A}$,*

$$\mathbb{P}(A) = \sum_{i\in I}\mathbb{P}(A\cap B_i) = \sum_{i\in I}\mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

In addition, if $\mathbb{P}(A) > 0$,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A\cap B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{i\in I}\mathbb{P}(A|B_i)\mathbb{P}(B_i)}, \quad \forall i \in I.$$

Definition 2.1.5 (Independent events)**.** *Two events A and B from $\mathcal{A}$ are independent if $\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

A sequence of events $(A_i)_{i\in I} \subset \mathcal{A}$ *is independent if* $\mathbb{P}(\cap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i)$ *for every finite subset* $J \subset I$*.*

Proposition 2.1.4. *Let* A *and* B *be two independent events from* $\mathcal{A}$*. Then* A *and* B^c*,* A^c *and* B*,* A^c *and* B^c *are independent.*

In addition, if $\mathbb{P}(B) > 0$*, then* $\mathbb{P}(A|B) = \mathbb{P}(A)$*.*

2.2 Random variables

2.2.1 Univariate distributions

Let $(\Omega, \mathcal{A}, \mathbb{P}) = ([0,1], \mathcal{B}_{[0,1]}, \lambda_{[0,1]})$ (where $\mathcal{B}_{[0,1]}$ is the Borelian σ-algebra of $[0,1]$ and $\lambda_{[0,1]}$ is the Lebesque measure restricted to $[0,1]$) and $(E, \mathcal{E})$ be a measurable space (either $(\mathbb{M}, \mathcal{P}(\mathbb{M}))$, where $\mathbb{M}$ is countable, or $(\mathbb{R}, \mathcal{B})$).

Definition 2.2.1 (Random variable)**.** *A real-valued random variable is a measurable mapping* $X : \Omega \to E$*, that is:* $\forall B \in \mathcal{E}, X^{-1}(B) \in \mathcal{A}$*.*

The distribution P *of* X *is the unique probability measure on* $(E, \mathcal{E})$ *such that* $\forall B \in \mathcal{E}, P(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B)$*.*

From now on, assertions on a random variable X are understood *almost surely*. For instance, $X \in \mathbb{N}$, could be written $X \in \mathbb{N}$ *a.s.*, and means $\mathbb{P}(X \in \mathbb{N}) = 1$.

Proposition 2.2.1. *Let* P *be a probability measure on* $(E, \mathcal{E})$*. Then, there exists a random variable* X *with distribution* P*.*

Definition 2.2.2 (Cumulative distribution function)**.** *The cumulative distribution function of a random variable* X *is:* $F_X : x \in \mathbb{R} \mapsto \mathbb{P}(X \leq x)$*.*

Property 2.2.2. *A cumulative distribution function* F *is:*

- *non-decreasing:* $x \leq y \implies F(x) \leq F(y)$*;*
- *right-continuous:* $\lim_{x\to a+} F(x) = F(a)$ *for any* $a \in \mathbb{R}$*;*
- *normalized:* $\lim_{x\to -\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$*.*

Proposition 2.2.3. *Let* $F : \mathbb{R} \to [0,1]$ *be a non-decreasing right-continuous and normalized function. Then, there exists a unique probability measure* P *on* $(\mathbb{R}, \mathcal{B})$ *such that* $F(x) = P((-\infty, x])$ *for all* $x \in \mathbb{R}$*.*

Theorem 2.2.4 (Radon–Nikodym theorem)**.** *Let ν and μ be two positive and σ-finite measures on the measurable space $(E, \mathcal{E})$*[a] *(here μ is either the counting or the Lebesgue measure). If ν is absolutely continuous with respect to μ,* i.e.

$$\mu(B) = 0 \implies \nu(B) = 0 \qquad \forall B \in \mathcal{E},$$

then there exists a measurable function $f : E \to \mathbb{R}_+$ such that for any $B \in \mathcal{E}$,

$$\nu(B) = \int_B f \, d\mu = \int_{x \in B} f(x) \, d\mu(x) = \begin{cases} \int_{x\in B} f(x)\,dx & \text{if } \mu \text{ is the Lebesgue measure;} \\ \sum_{x \in B \cap \mathbb{M}} f(x) & \text{if } \mu \text{ is the counting measure.} \end{cases}$$

The function f is called the density of ν with respect to μ.

[a] A positive measure ν is σ-finite on $(E, \mathcal{E})$ if there exists $(B_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ such that $\nu(B_n) < \infty$ for all $n \in \mathbb{N}$ and $E = \bigcup_{n\in\mathbb{N}} B_n$. This is the case for the counting and the Lebesgue measures.

There are two particular families of random variables:

- ◇ X is a discrete random variable if it takes countably many values (supposed in $\mathbb{N}$ for simplicity), equivalently if its cumulative distribution function has countably many discontinuities or if its distribution P is absolutely continuous with respect to the counting measure. In this case, the density f of P can be defined by $f(x) = \mathbb{P}(X = x)$ for all $x \in \mathbb{R}$ and is called the probability mass function of X.
- ◇ X is a continuous random variable if it takes uncountably many values and if its cumulative distribution function is absolutely continuous, equivalently if its distribution P is absolutely continuous with respect to the Lebesgue measure. In this case, the density f of P is called the probability density function of X and we have, for all $x \in \mathbb{R}$: $F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$. Conversely, $f(x) = F'(x)$ whenever F is differentiable.

In both cases, the density of a distribution P characterizes it completely.

Property 2.2.5. *Let X be a random variable on the measure space $(E, \mathcal{E}, \mu)$ (here μ is either the counting or the Lebesgue measure) and assume it has density $f : E \to \mathbb{R}_+$. Then:*

- ◇ $\int_E f \, d\mu = 1$*;*
- ◇ $\forall B \in \mathcal{E}, \mathbb{P}(X \in B) = \int_{x\in B} f(x)\,d\mu(x) = \int_{x \in E} f(x) \mathbf{1}_{x\in B} \, d\mu(x)$

2.2.2 Bivariate distributions and independence

Let $(E_1, \mathcal{E}_1, \mu_1)$ and $(E_2, \mathcal{E}_2, \mu_2)$ be two measure spaces and $X : \Omega \to E_1$ and $Y : \Omega \to E_2$ be two random variables.

The pair $(X, Y) : \Omega \to E_1 \times E_2$ is a bivariate random variable with the output measure space $(E_1 \times E_2, \mathcal{E}_1 \times \mathcal{E}_2, \mu_1 \times \mu_2)$[1] and distribution P defined for every $(A, B) \in \mathcal{E}_1 \times \mathcal{E}_2$ by $P(A, B) = \mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B))$.

Property 2.2.6. *If X has density f_X with respect to μ_1 and Y has density f_Y with respect to μ_2, then (X, Y) has a density, denoted $f_{X,Y}$ (not necessarily $f_X f_Y$), with respect to $\mu_1 \times \mu_2$.*

In addition, for every $(A, B) \in \mathcal{E}_1 \times \mathcal{E}_2$, $P(A, B) = \iint_{(x,y) \in A \times B} f_{X,Y}(x, y)\, d\mu_1(x) d\mu_2(y)$.

Proposition 2.2.7 (Marginal and conditional distributions)**.** *If (X, Y) has density $f_{X,Y}$ with respect to $\mu_1 \times \mu_2$, then X and Y have densities f_X and f_Y respectively with respect to μ_1 and μ_2, that are defined by: $f_X : x \in E_1 \mapsto \int_{y \in E_2} f_{X,Y}(x, y)\, d\mu_2(y)$ and $f_Y : y \in E_2 \mapsto \int_{x \in E_1} f_{X,Y}(x, y)\, d\mu_1(x)$. These densities define the marginal distributions of (X, Y) respectively for the random variables X and Y.*

In addition, let $y \in E_2$ with $f_Y(y) > 0$. Then the function $f_{X|Y}(\cdot, y) : x \in E_1 \mapsto \frac{f_{X,Y}(\cdot, y)}{f_Y(y)}$ is a density with respect to μ_1 and defines the distribution of the random variable $X|Y = y$ (called the conditional distribution of X given $Y = y$).

Proposition 2.2.8. *The distribution of the random pair (X, Y) is characterized by the marginal distribution for Y and the conditional distribution of $X|Y = y$ for all $y \in E_2$ such that $f_Y(y) > 0$.*

Definition 2.2.3 (Independence)**.** *Two random variables X and Y (respectively defined on $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$) are independent, denoted $X \perp\!\!\!\perp Y$, if for every $A \in \mathcal{E}_1$ and $B \in \mathcal{E}_2$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.*

A sequence $(X_n)_{n \geq 1}$ of random variables (X_n being defined on $\mathcal{E}_n$) is said independent if for every positive integer n, $\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{1 \leq i \leq n} \mathbb{P}(X_i \in A_i)$ for any events $A_1 \in \mathcal{E}_1$, …, $A_n \in \mathcal{E}_n$.

Proposition 2.2.9 (Coalitions lemma)**.** *Let $(X_i)_{1 \leq i \leq n}$ be independent random variables, $p \in [\![1, n-1]\!]$, $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^{n-p} \to \mathbb{R}$ be two measurable functions. Then $f(X_1, \ldots, X_p) \perp\!\!\!\perp g(X_{p+1}, \ldots, X_n)$.*

In particular, two arbitrary random variables in the sequence are independent.

Proposition 2.2.10. *Let X and Y be two random variables (respectively defined on $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$) with joint density $f_{X,Y}$. $X \perp\!\!\!\perp Y$ if and only if $f_{X,Y} = f_X f_Y$ almost surely or if there exist two functions $g : E_1 \to \mathbb{R}_+$ and $h : E_2 \to \mathbb{R}_+$ such that $f_{X,Y}(x, y) = g(x)h(y)$ for almost all $(x, y) \in E_1 \times E_2$.*

[1] The product mesure is defined by $(\mu_1 \times \mu_2)(A, B) = \mu_1(A)\mu_2(B)$ for every $(A, B) \in \mathcal{E}_1 \times \mathcal{E}_2$.

If $X \perp\!\!\!\perp Y$, then for all $y \in E_2$ such that $f_Y(y) > 0$, the distribution of $X|Y = y$ is that of X.

Definitions of joint, marginal and conditional distributions can be extended to random vectors $\mathbf{X} = (X_1, \dots, X_n)$ ($\mathbf{X}$ is said to have a multivariate distribution). This makes it possible to define a sample or independent and identically distributed (*iid*) random variables.

Definition 2.2.4 (*iid* random variables and samples)**.** *Random variables $X_1, \dots, X_n$ are said* iid *if they are independent and if they have the same marginal distribution.*

In addition, a random vector $\mathbf{X} = (X_1, \dots, X_n)$ is called a sample if $X_1, \dots, X_n$ are iid*.*

2.2.3 Expectation

Definition 2.2.5 (Expectation)**.** *Let X be a random variable on the measure space $(E, \mathcal{E}, \mu)$ with density f. If*

$$\int_{x\in E} |x| f(x)\, d\mu(x) < \infty,$$

X is said to be integrable and we note L^1 the set of all integrable random variables (supposed on the measure space $(E, \mathcal{E}, \mu)$).

In this case, we define

$$\mathbb{E}(X) = \int_{x\in E} x f(x)\, d\mu(x).$$

The expectation $\mathbb{E}(X)$ of a random variable X is also called expected value or first order moment.

Proposition 2.2.11. *Let $X \in L^1$, $Y \in L^1$ and Z be a random variable. Then, one has:*

- ⋄ *L^1 is a vector space;*
- ⋄ *$\mathbb{E}$ is a linear function on L^1: $\forall (a,b) \in \mathbb{R}^2, \mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y)$;*
- ⋄ *if $\exists C \in \mathbb{R} : |Z| \leq C$ a.s., then $Z \in L^1$;*
- ⋄ *$|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$ and $Z \in L^1 \iff |Z| \in L^1$;*
- ⋄ *$X \leq Y$ a.s. $\implies \mathbb{E}(X) \leq \mathbb{E}(Y)$.*

Theorem 2.2.12 (Transfert theorem)**.** *Let $X \in L^1$, h be a measurable function and $Y = h(X)$. If $Y \in L^1$, then:*

$$\mathbb{E}(Y) = \mathbb{E}(h(X)) = \int_{x\in\mathcal{E}} h(x) f(x)\, d\mu(x).$$

Proposition 2.2.13. *Let X and Y be two independent random variables, f and g two measurable functions such that $f(X) \in L^1$ and $g(Y) \in L^1$. Then $f(X)g(Y) \in L^1$ and*

$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$.

Definition 2.2.6 (Square-integrable random variable)**.** *A random variable X is square-integrable if $X^2 \in L^1$ and we note L^2 the set of all square-integrable random variables.*

Proposition 2.2.14. ◇ *L^2 is a vector space;*
- ◇ *$L^2 \subset L^1$, that is $X \in L^2 \implies X \in L^1$;*
- ◇ *$\forall X \in L^2, |\mathbb{E}(X)| \leq \mathbb{E}(|X|) \leq \sqrt{\mathbb{E}(X^2)}$;*

Definition 2.2.7 (Variance and standard deviation)**.** *Let $X \in L^2$. The variance of X is $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2)] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ and the standard deviation is $\mathrm{sd}(X) = \sqrt{\mathbb{V}(X)}$.*

Property 2.2.15. *For all $X \in L^2$, $Y \in L^2$ and $(a, b) \in \mathbb{R}^2$:*

- ◇ *$aX + b \in L^2$ and $\mathbb{V}(aX + b) = a^2\,\mathbb{V}(X)$;*
- ◇ *$X \perp\!\!\!\perp Y \implies \mathbb{V}(aX + bY) = a^2\,\mathbb{V}(X) + b^2\,\mathbb{V}(Y)$.*

2.2.4 Conditional expectation

Proposition 2.2.16. *Let X and Y be two random variables. If $X \in L^1$, then $X|Y = y \in L^1$ for every $y \in E_2$ such that $f_Y(y) > 0$ and $\mathbb{E}(X|Y = y) = \int_{x\in E_1} x f_{X|Y}(x, y)\, d\mu_1(x)$.*

In addition, if $r : E_1 \times E_2 \to \mathbb{R}$ is a measurable function such that $r(X, Y) \in L^1$, then $r(X, Y)|Y = y \in L^1$ for every $y \in E_2$ such that $f_Y(y) > 0$ and $\mathbb{E}[r(X, Y)|Y = y] = \int_{x\in E_1} r(x, y) f_{X|Y}(x, y)\, d\mu_1(x)$.

The last part of the proposition says that given $Y = y$, Y should be considered "as a constant". For instance, $\mathbb{E}(XY|Y = y) = \mathbb{E}(X|Y = y)y$.

Definition 2.2.8 (Conditional expectation)**.** *Let X and Y be two random variables. If $X \in L^1$, the conditional expectation of X given Y is the random variable $\mathbb{E}(X|Y) = h(Y)$, where $h : y \in E_2 \mapsto \mathbb{E}(X|Y = y)\mathbf{1}_{f_Y(y)>0}$.*

In addition, if $r : E_1 \times E_2 \to \mathbb{R}$ is a measurable function such that $r(X, Y) \in L^1$, $r(X, Y)|Y = h(Y)$, where $h : y \in E_2 \mapsto \mathbb{E}(r(X, Y)|Y = y)\mathbf{1}_{f_Y(y)>0}$.

Property 2.2.17. ◇ *$\mathbb{E}(X|Y)$ has a density with respect to μ_2;*
- ◇ *for any $Z \in L^1$ and $(a, b) \in \mathbb{R}^2$, $\mathbb{E}(aX + bZ|Y) = a\,\mathbb{E}(X|Y) + b\,\mathbb{E}(Z|Y)$;*
- ◇ *for a measurable function h, $\mathbb{E}(Xh(Y)|Y) = \mathbb{E}(X|Y)h(Y)$;*
- ◇ *$\mathbb{E}(X|Y) \geq 0$ a.s. if $X \geq 0$ a.s.;*

- $\mathbb{E}(1|Y) = 1$ *a.s.*.

Theorem 2.2.18 (The rule of iterated expectations)**.** *Let X and Y be two random variables, such that $X \in L^1$. Then $\mathbb{E}(X|Y) \in L^1$ and $\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X)$.*

Proposition 2.2.19. *Let X and Y be two independent random variables such that $X \in L^1$. Then for every $y \in E_2$ such that $f_Y(y) > 0$, $\mathbb{E}(X|Y = y) = \mathbb{E}(X)$. In addition, $\mathbb{E}(X|Y) = \mathbb{E}(X)$ a.s..*

2.2.5 Transformations of random variables

Proposition 2.2.20. *One has:*

- *any continuous transformation of two random variables X and Y is a random variable;*
- $\inf_{n\in\mathbb{N}} X_n$ *and* $\sup_{n\in\mathbb{N}} X_n$ *are two random variables, for any sequence of random variables* $(X_n)_{n\in\mathbb{N}}$.

Let X be a random variable and $Y = h(X)$, where h is a measurable function. We give here some tools in order to determine the distribution of Y. The first method is to make the cumulative distribution function F_Y explicit. If F_Y is piecewise differentiable, Y is a continuous random variable and it is possible to exhibit its probability density function as F_Y' almost everywhere.

Property 2.2.21 (The test function method)**.** *If there exists a function g such that for all continuous and bounded functions φ, $\mathbb{E}(\varphi(Y)) = \int_{y\in\mathbb{R}} \varphi(y)g(y)\,dy$, then Y is continuous and has g for probability density function.*

In particular, if X has density f with respect to the Lebesgue measure, the transfert theorem says that $\mathbb{E}(\varphi(Y)) = \mathbb{E}[\varphi(h(X))] = \int_{x\in\mathbb{R}} (\varphi \circ h)(x) f(x)\,dx$, which involves a change of variable in order to exhibit the probability density function g.

A last method, particularly useful for sum of random variables is using the characteristic function, that characterizes the distribution of a random variable.

Definition 2.2.9 (Characteristic function)**.** *The characteristic function of a random variable X is*

$$\phi_X(t) = \mathbb{E}(\mathrm{e}^{\imath tX}), \quad \forall t \in \mathbb{R}.$$

Proposition 2.2.22. *Let X and Y be two random variables with characteristic functions ϕ_X and ϕ_Y. If $\phi_X = \phi_Y$, then X and Y have the same distribution.*

Moreover, if $X \perp\!\!\!\perp Y$, then the characteristic function of $X + Y$ is $\phi_{X+Y} = \phi_X \phi_Y$.

Name	**Support**	**Notation**	**Parameters**	**Density** $f(x), \forall x \in \mathbb{R}$	**Expectation**	**Variance**
Point mass						
Discrete uniform	$[\![1, m]\!]$					
Bernoulli	$\{0, 1\}$	$\mathcal{B}(p)$	$p \in (0, 1)$	$p^x(1-p)^{1-x}$	p	$p(1-p)$
Binomial						
Geometric						
Poisson						
Uniform	$[0, 1]$					
Normal						
Laplace						
Student						
Exponential						
Chi-square						
Fisher						
Multinomial						
Multivariate normal						

Table 2.1: Classical distributions

Theorem 2.2.23 (Universality of the uniform distribution). *Let X be a random variable with cumulative distribution function F_X and $U \sim \mathcal{U}([0, 1])$. Let F_X^{-1} be the generalized inverse of F_X, defined by:*

$$F_X^{-1} : u \in [0, 1] \mapsto \inf\{x \in \mathbb{R} : F_X(x) \geq u\} \in \mathbb{R} \cup \{\pm\infty\}.$$

Then

1. *X and $F_X^{-1}(U)$ have same distribution;*
2. *if F_X is continuous, then $F_X(X)$ and U have same distribution.*

2.3 Famous inequalities

2.3.1 Probability inequalities

Proposition 2.3.1 (Markov's inequality). *Let $X \in L^1$ be a non-negative random variable. For any $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Proposition 2.3.2 (Bienaymé-Tchebychev's inequality). *Let $X \in L^2$. For any $t > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2}.$$

As a particular case of Bienaymé-Tchebychev's inequality, consider $X \in L^2$ such that $\mathbb{E}(X) = 0$ and $\mathbb{V}(X) = 1$. Then for any $k \in \mathbb{N}^*$, $\mathbb{P}(|X| \geq k) \leq \frac{1}{k}$.

Proposition 2.3.3 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables and assume that for each $i \in [\![1, n]\!]$, $\exists (a_i, b_i) \in \mathbb{R}^2 : a_i \leq X_i \leq b_i$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

In particular, if $X_1, \dots, X_n$ are iid*, denoting $c = b_1 - a_1$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have:*

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}(X_1)\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{c^2}\right).$$

Proposition 2.3.4 (Mill's inequality)**.** *Let $Z \sim \mathcal{N}(0, 1)$. Then, for any $t > 0$,*

$$\mathbb{P}(Z > t) = \mathbb{P}(Z < -t) \leq \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-t^2/2}}{t}.$$

A fairly general recipe for obtaining concentration inequalities is Chernoff's method.

Proposition 2.3.5 (Chernoff's method)**.** *Let $X_1, \dots, X_n$ be independent random variables and assume that for each $i \in [\![1, n]\!]$, there exists a function $\phi_i : \mathbb{R}_+^* \to \mathbb{R}$ such that*

$$\forall \lambda > 0, \quad \mathbb{E}\left(\mathrm{e}^{\lambda(X_i - \mathbb{E}(X_i))}\right) \leq \mathrm{e}^{\phi_i(\lambda)}.$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq t\right) \leq \inf_{\lambda > 0} \exp\left(\sum_{i=1}^{n} \phi_i(\lambda) - \lambda t\right).$$

2.3.2 Expectation inequalities

Proposition 2.3.6 (Cauchy-Schwartz inequality)**.** *Let $X \in L^2$ and $Y \in L^2$. Then $XY \in L^1$ and*

$$|\,\mathbb{E}(XY)| \leq \mathbb{E}\left(|XY|\right) \leq \sqrt{\mathbb{E}(X)^2}\sqrt{\mathbb{E}(Y)^2}.$$

In addition, equalities hold if and only if $Y = 0$ a.s. or $\exists \alpha \in \mathbb{R} : X = \alpha Y$ a.s..

Proposition 2.3.7 (Jensen's inequality)**.** *Let $X \in L^1$ and f be a convex function. If $f(X) \in L^1$,*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

In addition, if f is strictly convex,

$$f(\mathbb{E}(X)) < \mathbb{E}(f(X)) \iff X \text{ is not constant } a.s..$$

From the last inequality, we obtain that for $X \in L^1$ such that $X > 0$ *a.s.*, $\frac{1}{\mathbb{E}(X)} \leq \mathbb{E}\left(\frac{1}{X}\right)$ and $\mathbb{E}(\log(X)) \leq \log(\mathbb{E}(X))$ (if $1/X \in L^1$ and $\log(X) \in L^1$).

Markov's inequality explains how to obtain a probability inequality from an expectation inequality. Here is how to go in the other side.

Property 2.3.8. *Let X be a random variable with $X^k \geq 0$ a.s. and $k > 0$ such that $X^k \in L^1$. Then $\mathbb{E}(X^k) = \int_0^\infty \mathbb{P}(X^k \geq t)\, dt$.*

Proposition 2.3.9. *Let $X \in L^2$ such that $X \geq 0$ a.s. and $\mathbb{P}(X \geq t) \leq C\,\mathrm{e}^{-\alpha t^2}$ for some $C > 1$ and $\alpha > 0$ and all $t \geq 0$. Then,*

$$\mathbb{E}(X) \leq \sqrt{\frac{\log(C\,\mathrm{e}\,)}{\alpha}}.$$

2.4 Limit theorems

Definition 2.4.1 (Types of convergence)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables (with cumulative distribution functions F_n) and X be a random variable (with cumulative disitrubtion function F).*

1. *Almost sure convergence: $X_n \xrightarrow[n\to\infty]{a.s.} X$ if $\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$.*
2. *Convergence in probability: $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ if for every $\epsilon > 0$, $\mathbb{P}\left(|X_n - X| \geq \epsilon\right) \xrightarrow[n\to\infty]{} 0$.*
3. *Convergence in mean (if $X_n \in L^1$ and $X \in L^1$): $X_n \xrightarrow[n\to\infty]{L^1} X$ if $\mathbb{E}(|X_n - X|) \xrightarrow[n\to\infty]{} 0$.*
4. *Convergence in quadratic mean (if $X_n \in L^2$ and $X \in L^2$): $X_n \xrightarrow[n\to\infty]{L^2} X$ if $\mathbb{E}((X_n - X)^2) \xrightarrow[n\to\infty]{} 0$.*
5. *Convergence in distribution: $X_n \xrightarrow[n\to\infty]{d} X$ if $F_n(x) \xrightarrow[n\to\infty]{} F(x)$ for all x for which F is continuous.*

Proposition 2.4.1. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable. Then $X_n \xrightarrow[n\to\infty]{d} X$ if and only if for every continuous and bounded function $h : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}(h(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(h(X))$.*

Theorem 2.4.2 (Lévy's theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable. Then $X_n \xrightarrow[n\to\infty]{d} X$ if and only if*

$$\forall t \in \mathbb{R}: \quad \phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi_X(t).$$

In addition, if

$$\forall t \in \mathbb{R}: \quad \phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi(t),$$

for some function $\phi : \mathbb{R} \to \mathbb{C}$, which is continuous at 0, then ϕ is the characteristic function of a random variable Y and $X_n \xrightarrow[n\to\infty]{d} Y$.

Proposition 2.4.3. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable.*

- ◇ $X_n \xrightarrow[n\to\infty]{a.s.} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{d} X$*;*
- ◇ *if $X_n \in L^1$ and $X \in L^1$, $X_n \xrightarrow[n\to\infty]{L^1} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$;*
- ◇ *if $X_n \in L^2$ and $X \in L^2$, $X_n \xrightarrow[n\to\infty]{L^2} X \implies X_n \xrightarrow[n\to\infty]{L^1} X$;*
- ◇ *if $X_n \in L^1$ and $X \in L^1$, $X_n \xrightarrow[n\to\infty]{L^1} X \implies \mathbb{E}(X_n) \xrightarrow[n\to\infty]{} \mathbb{E}(X)$.*

Proposition 2.4.4. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable.*

- ◇ *if X is constant almost surely, $X_n \xrightarrow[n\to\infty]{d} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$;*
- ◇ *if $(X_n)_{n\geq 1}$ is non-decreasing almost surely, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{a.s.} X$;*
- ◇ *if $\exists C \in \mathbb{R}_+ : |X_n| \leq C$ a.s. and $X \in L^1$, then $X_n \in L^1$ and $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{L^1} X$.*

Theorem 2.4.5 (Corollary of Borel-Cantelli's lemma)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be a random variable. If for each $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| \geq \epsilon) < \infty,$$

or if there exists a sequence $(\epsilon_n)_{n\geq 1}$ such that $\epsilon_n \xrightarrow[n\to\infty]{} 0$ and

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| \geq \epsilon_n) < \infty,$$

then $X_n \xrightarrow[n\to\infty]{a.s.} X$.

Theorem 2.4.6 (The dominated convergence theorem)**.** *Let* $(X_n)_{n\geq 1}$ *be a sequence of random variables and* X *be a random variable. If there exists* $Z \in L^1$ *such that* $|X_n| \leq Z$ *and if* $X_n \xrightarrow[n\to\infty]{a.s.} X$, *then* $X_n \in L^1$, $X \in L^1$ *and* $X_n \xrightarrow[n\to\infty]{L^1} X$.

Theorem 2.4.7 (The continuous mapping theorem)**.** *Let* $(X_n)_{n\geq 1}$ *be a sequence of random variables and* X *be a random variable. For every continuous function* $g : \mathbb{R} \to \mathbb{R}$*:*

- ◇ $X_n \xrightarrow[n\to\infty]{a.s.} X \implies g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$;
- ◇ $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies g(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} g(X)$;
- ◇ $X_n \xrightarrow[n\to\infty]{d} X \implies g(X_n) \xrightarrow[n\to\infty]{d} g(X)$.

In addition, for every Lipschitz continuous (or continuous and bounded) function $g : \mathbb{R} \to \mathbb{R}$*:*

- ◇ $X_n \xrightarrow[n\to\infty]{L^1} X \implies g(X_n) \xrightarrow[n\to\infty]{L^1} g(X)$;
- ◇ $X_n \xrightarrow[n\to\infty]{L^2} X \implies g(X_n) \xrightarrow[n\to\infty]{L^2} g(X)$.

Example 2.4.1 (A counterexample by Arnaud Guyader)**.** *For every positive integer* n, *let* X_n *be a discrete random variable taking values in* $\{0, n\}$, *such that* $\mathbb{P}(X_n = n) = \frac{1}{n^3}$ *and* $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n^3}$. *We have* $\mathbb{E}(X_n^2) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$, *so* $X_n \xrightarrow[n\to\infty]{L^2} 0$. *However, taking now the continuous (but not Lipschitz continuous) function* $f(x) = x^2$, $\mathbb{E}[f(X)^2] = n \not\xrightarrow[n\to\infty]{} 0$, *so* $g(X_n) \not\xrightarrow[n\to\infty]{L^2} 0$.

Proposition 2.4.8 (Convergence of sums and products)**.** *Let* $(X_n)_{n\geq 1}$ *and* $(Y_n)_{n\geq 1}$ *be two sequences of random variables and* X *and* Y *be two random variables. If* X_n *converges to* X *and* Y_n *converges to* Y *almost surely (respectively in probability, in mean or in quadratic mean), then* $X_n + Y_n$ *converges to* $X + Y$ *and* $X_n Y_n$ *converges to* XY *almost surely (respectively in probability, in mean or in quadratic mean).*

Theorem 2.4.9 (Slutsky's theorem)**.** *Let* $(X_n)_{n\geq 1}$ *and* $(Y_n)_{n\geq 1}$ *be two sequences of random variables,* X *a random variable and* $c \in \mathbb{R}$. *If* $X_n \xrightarrow[n\to\infty]{d} X$ *and* $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} c$, *then* $X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$ *and* $X_n Y_n \xrightarrow[n\to\infty]{d} cX$.

Theorem 2.4.10 (The law of large numbers)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of* iid *random variables such that $X_n \in L^1$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then,*

$$\bar{X}_n \xrightarrow[n\to\infty]{a.s.\ \&\ L^1} \mathbb{E}(X_1).$$

This result is often called the strong law of large numbers. In particular,

$$\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}(X_1),$$

which is referred to as the weak law of large numbers.

Theorem 2.4.11 (The central limit theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of* iid *random variables such that $X_n \in L^2$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\sqrt{n}\left(\bar{X}_n - \mathbb{E}(X_1)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mathbb{V}(X_1)).$$

Proposition 2.4.12 (The delta method)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables, $(a_n)_{n\geq 1}$ a sequence of real numbers such that $a_n \xrightarrow[n\to\infty]{} \infty$, $c \in \mathbb{R}$ and Y a random variable. Suppose that $a_n\left(X_n - c\right) \xrightarrow[n\to\infty]{d} Y$. Then, for every function $g : \mathbb{R} \to \mathbb{R}$ differentiable in c, $a_n\left(g(X_n) - g(c)\right) \xrightarrow[n\to\infty]{d} g(c)Y$.*

Theorem 2.4.13 (The delta method applied to the central limit theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of* iid *random variables such that $X_n \in L^2$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for every function $g : \mathbb{R} \to \mathbb{R}$ differentiable in $\mathbb{E}(X_1)$ such that $g'(\mathbb{E}(X_1)) \neq 0$, $\sqrt{n}\left(g(\bar{X}_n) - g(\mathbb{E}(X_1))\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, g'(\mathbb{E}(X_1))^2\, \mathbb{V}(X_1)\right)$.*

2.5 The multivariate normal vector

Definition 2.5.1 (Multivariate normal vector)**.** *Let $\mathbf{X}$ be a d-dimensional random vector. $\mathbf{X}$ is a multivariate normal vector (or Gaussian vector) if for every $a \in \mathbb{R}^d$,*

$$\exists \mu \in \mathbb{R} : \quad a^\top \mathbf{X} \sim \mathcal{N}(\mu, \sigma^2) \textit{ for some } \sigma^2 > 0 \textit{ or } a^\top \mathbf{X} = \mu \textit{ a.s.},$$

i.e. *$a^\top \mathbf{X}$ follows a (potentially degenerated) normal distribution.*

Proposition 2.5.1. *Let $\mathbf{X}$ be a d-dimensional Gaussian vector. Then $\mathbf{X} \in L^2$ and we note $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = \mathbb{E}(\mathbf{X})$ and $\Sigma = \mathbb{V}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top\right]$.*

In addition, for every $a \in \mathbb{R}^d$, $a^\top \mathbf{X}$ is not constant almost surely if and only if Σ is non-singular. This is equivalent to $\mathbf{X}$ having a density with respect to the Lebesgue measure, which can be expressed:

$$\forall x \in \mathbb{R}^d : \quad f_{\mathbf{X}}(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \, \mathrm{e}^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \, .$$

Proposition 2.5.2. *Let $\mathbf{X}$ be a d-dimensional Gaussian vector with $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. Let I and J be two disjoint sets of indexes from $[\![1, d]\!]$ and denote $\mathbf{X}_I$ and $\mathbf{X}_J$ the corresponding subvectors. Then*

$$\mathbf{X}_I \perp\!\!\!\perp \mathbf{X}_J \iff \mathrm{Cov}(\mathbf{X}_I, \mathbf{X}_J) = \mathbb{E}\left[(\mathbf{X}_I - \mathbb{E}(\mathbf{X}_I))(\mathbf{X}_J - \mathbb{E}(\mathbf{X}_J))^\top\right] = 0,$$

in other words if the items of Σ at the intersection of I and J are 0 outside the diagonal.

In particular, components of $\mathbf{X}$ are independent if and only if Σ is diagonal.

Proposition 2.5.3. *Let $\mathbf{X}$ be a d-dimensional Gaussian vector with $\mathbf{X} \sim (\mu, \Sigma)$. Let also $A \in \mathbb{R}^{k \times d}$ for some positive integer k and I be a set of indexes from $[\![1, d]\!]$. Then,*

- ◇ *the marginal distribution of $\mathbf{X}_I$ is $\mathcal{N}(\mu_I, \Sigma_{I,I})$;*
- ◇ *$A\mathbf{X} \sim \mathcal{N}(A\mu, A\Sigma A^\top)$;*
- ◇ *$\mathbf{X}$ has same distribution as $\mu + \Sigma^{1/2}\mathbf{Y}$, where $\mathbf{Y} \sim \mathcal{N}(0, I_d)$;*
- ◇ *if Σ is non-sigular, $(\mathbf{X} - \mu)^\top \Sigma^{-1}(\mathbf{X} - \mu) \sim \chi^2_d$;*

Theorem 2.5.4 (Cochran's theorem)**.** *Let $V \subset \mathbb{R}^d$ be a vector subspace of $\mathbb{R}^d$, $V^\perp$ its orthogonal space ($\mathbb{R}^d = V \oplus V^\perp$) and $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2 I_d)$ a d-dimensional multivariate normal vector. By denoting P and $P_\perp = I_d - P$ the orthogonal projectors respectively on V and $V^\perp$, we have:*

1. *$P\mathbf{X} \sim \mathcal{N}(P\mu, \sigma^2 P)$ and $P_\perp \mathbf{X} \sim \mathcal{N}(P_\perp \mu, \sigma^2 P_\perp)$;*
2. *$P\mathbf{X} \perp\!\!\!\perp P_\perp \mathbf{X}$;*
3. *$\frac{1}{\sigma^2}\|P\mathbf{X} - P\mu\|^2 \sim \chi^2_{\dim(V)}$ and $\frac{1}{\sigma^2}\|P_\perp \mathbf{X} - P_\perp \mu\|^2 \sim \chi^2_{\dim(V^\perp)}$.*

Theorem 2.5.5 (The multivariate central limit theorem)**.** *Let $(\mathbf{X}_n)_{n \geq 1}$ be a sequence of* iid *random vectors such that $\mathbf{X}_n \in L^2$ and let $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Then*

$$\sqrt{n}\left(\bar{\mathbf{X}}_n - \mathbb{E}(\mathbf{X}_1)\right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, \mathbb{V}(\mathbf{X}_1)).$$

2.6 Exercises

Exercise 2.1 (Integration by parts)**.** What are the expectation and the variance of the following distribution:

1. $\mathcal{U}([a,b])$, where $a < b$?
2. $\mathcal{E}(\lambda)$, where $\lambda > 0$? Deduce that for every positive integer k, $\mathbb{E}(X^k) = \frac{k!}{\lambda^k}$.
3. $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$?

Exercise 2.2 (The test function method)**.**

1. Let $X \sim \mathcal{E}(\lambda)$, where $\lambda > 0$. What is the distribution of $\sqrt{X}$?
2. Let X be a random variable distributed according to a Cauchy distribution, *i.e.* with density $x \in \mathbb{R} \mapsto \frac{1}{\pi(1+x^2)}$ with respect to the Lebesgue measure. What is the distribution of $1/X$?
3. Let $X \sim \mathcal{U}([0,\pi])$ and $Y = \sin(X)$. What is the distribution of Y?

Exercise 2.3 (Uniform distribution)**.** Let $U \sim \mathcal{U}([0,1])$ and $X = \max(U, 1-U)$.

1. What is the distribution of X?
2. What is the distribution of $Y = \min(U, 1-U)$?
3. What is the distribution of $Z = -\log(U)$?
4. Let $\lambda > 0$. What is the distribution of Z/λ?

Exercise 2.4 (Marginal distributions)**.** Let (X,Y) be a pair of random variables with joint density $f : (x,y) \in \mathbb{R}^2 \mapsto \frac{2}{\pi}\,\mathrm{e}^{-x(1+y^2)}\,\mathbf{1}_{x,y\geq 0}$ with respect to the Lebesgue measure.

1. Show that f is a probability density function.
2. What are the distributions of X and Y.

Exercise 2.5 (Marginal distributions of two pairs)**.** Let (X,Y) and (X',Y') be two pairs of random variables with respective densities (with respect to the Lebesgue measure):

$$f(x,y) = \frac{1}{4}(1+xy)\mathbf{1}_{[0,1]^2}(x,y) \quad \text{and} \quad g(x,y) = \frac{1}{4}\mathbf{1}_{[0,1]^2}(x,y).$$

1. Show that f and g are probability density functions.
2. How can we assure that (X,Y) and (X',Y') do not have the same distribution?
3. Show that the marginal distributions of the two pairs are the same.

Exercise 2.6 (Convergence in distribution)**.** Let $(X_n)_{n\geq 1}$ be a sequence of random variables and $f : \mathbb{R} \to \mathbb{R}$ a continuous function. Show that if $X_n \xrightarrow[n\to\infty]{d} X$ for some random variable

X, then $f(X_n) \xrightarrow[n\to\infty]{d} f(X)$.

Exercise 2.7 (Law of large numbers)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables, each with probability density function (with respect to the Lebesgue measure):

$$x \in \mathbb{R} \mapsto 2x\,\mathrm{e}^{-x^2}\,\mathbf{1}_{\mathbb{R}_+}(x).$$

1. Compute $\mathbb{E}(X_1^2)$.
2. Show that $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is convergent almost surely.
3. Same question for $Z_n = \frac{n}{\sum_{i=1}^{n} X_i^2}$.

Exercise 2.8 (Asymptotic normality)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables such that $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1)$ exist (with $\sigma^2 > 0$). Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

1. Show that $\bar{X}_n$ is convergent almost surely.
2. Same question for $\hat{\sigma}_n^2$.
3. Show that $\sqrt{n}\frac{\bar{X}_n-\mu}{\sqrt{\hat{\sigma}_n^2}}$ is convergent in distribution and exhibit the limit.

Exercise 2.9 (Delta method)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables such that $X_1 \sim \mathcal{U}([0,1])$ and let $Y_n = \frac{1}{(\prod_{i=1}^{n} X_i)^{1/n}}$.

1. What is the distribution of $Z_1 = -\log(X_1)$?
2. Show that $\sqrt{n}(Y_n - \mathrm{e}\,)$ is convergent in distribution and exhibit the limit.

Chapter 3

Statistics

3.1 Inference

3.1.1 Model

Let $(E, \mathcal{E})$ be a measurable space.

Definition 3.1.1 (Statistical model)**.** *A statistical model $\mathcal{P}$ is a family of distributions (*i.e.* of probability measures) on the measurable space $(E, \mathcal{E})$. The model is parametric if there exists an integer p and $\Theta \subset \mathbb{R}^p$ such that $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where for every $\theta \in \Theta$, P_θ is a probability measure on $(E, \mathcal{E})$.*

In addition, $\mathcal{P}$ is a sampling model if there exists a positive integer n such that for every $\theta \in \Theta$, $P_\theta = Q_\theta^{\otimes n}$ for some distribution Q_θ (in other words, any random vector $\mathbf{X} \sim P_\theta$ is a n-sample).

Here, we will focus on parametric models. Then, in many cases, the parameter θ will reflects the natural degrees of liberty of the distribution P_θ (think for instance to $P_\theta = \mathcal{B}(\theta)^{\otimes n}$ or to $P_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)^{\otimes n}$). In other cases, θ will be any parameter helping to characterize the distribution P_θ, thanks to its cumulative distribution function or its density (if there is one). Thus, since we consider here distributions with densities with respect to the counting or the Lebesgue measure, a parametric model can also be extended to a family $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ of densities.

Assume that we are provided with some measurements $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}$. The principle of statistical modeling is to consider that these measurements are the realizations of some random variables $\underline{X}_1, \dots, \underline{X}_n$ and to determine in the best way the distribution of $\underline{\mathbf{X}} = (\underline{X}_1, \dots, \underline{X}_n)$ based on the model $\mathcal{P}$. In other words, it is about determining $\theta \in \Theta$ such that P_θ approximates the best as possible the distribution of $\underline{\mathbf{X}}$. In order to work only with parameters (and not also with distributions), we assume that the distribution of $\underline{\mathbf{X}}$ is in the model, *i.e.* $\exists \theta_0 \in \Theta : \underline{\mathbf{X}} \sim P_{\theta_0}$. Now, to emphasize that the value θ_0 is unknown, we will work in broad generality and try to approximate, for every $\theta \in \Theta$, the value of

θ based on the knowledge of the observation $\mathbf{X}_\theta = (X_1, \dots, X_n) \sim P_\theta$. In the forthcoming discussion, the phrase "of θ" means "of the generic parameter θ indexing the parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$" (but nowhere θ is a fixed value).

For ease of explanation, we will also assume that there is a single parameter of interest: $\Theta \subset \mathbb{R}$. In addition, we will sometimes only deal with sampling models (which are often the ones we use in practice), *i.e.* we will assume $X_1, \dots, X_n$ are *iid*. In this case, with a slight abuse, P_θ will be the distribution of X_1 instead of $\mathbf{X}_\theta = (X_1, \dots, X_n)$.

3.1.2 Estimation

Definition 3.1.2 (Statistic, estimator and estimation)**.** *A statistic is a random variable $T = f(\mathbf{X}_\theta)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a measurable function independent of θ.*

An estimator $\hat{\theta}_n$ of the parameter θ is a statistic aimed at approximating θ and with $\mathbb{P}(\hat{\theta}_n \in \Theta) \xrightarrow[n\to\infty]{} 1$.

An estimation of the value θ_0 is the realization $f((\underline{x}_1, \dots, \underline{x}_n))$ of an estimator $\hat{\theta}_n = f(\mathbf{X}_\theta)$ of θ.

The notation $\hat{\theta}_n$ for an estimator of θ is clear:

1. "$\hat{}$" indicates that $\hat{\theta}_n$ is aimed at approximating θ;
2. "n" points out that the approximation depends on the size of the observation;
3. "θ" makes it clear that $\hat{\theta}_n$ depends on θ (more precisely its distribution is ruled by θ).

Definition 3.1.3 (Bias and mean squared error)**.** *Let $\hat{\theta}_n$ be an estimator of θ.*

1. *The bias of $\hat{\theta}_n$ is: $B(\theta) = \mathbb{E}[\hat{\theta}_n] - \theta, \quad \forall \theta \in \Theta$.*
2. *The mean squared error of $\hat{\theta}_n$ is:*

$$R(\theta) = \mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] = B(\theta)^2 + \mathbb{V}(\hat{\theta}_n), \quad \forall \theta \in \Theta.$$

We should remember here that when θ varies, $\mathbb{E}[\hat{\theta}_n]$ changes too.

Definition 3.1.4. *An estimator $\hat{\theta}_n$ of θ is:*

1. *consistent if $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta, \quad \forall \theta \in \Theta$;*
2. *strongly consistent if $\hat{\theta}_n \xrightarrow[n\to\infty]{a.s.} \theta, \quad \forall \theta \in \Theta$;*
3. *asymptotically normal if for every $\theta \in \Theta$, there exists $(c_\theta, \sigma_\theta^2) \in \mathbb{R} \times \mathbb{R}_+^*$ and a sequence of real numbers $(a_n)_{n\geq 1}$ with $a_n \xrightarrow[n\to\infty]{} \infty$, such that:*

$$a_n(\hat{\theta}_n - c_\theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma_\theta^2).$$

We say that $\hat{\theta}_n$ has asymptotic variance σ_θ^2 and rate of convergence $1/a_n$.

Property 3.1.1. *Let $\hat{\theta}_n$ be an asymptotically normal estimator of θ with*

$$a_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2), \quad \forall \theta \in \Theta.$$

Then $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta, \quad \forall \theta \in \Theta$.

Property 3.1.2. *Let $\hat{\theta}_n$ be an estimator of θ with mean squared error R. If for every $\theta \in \Theta$, $R(\theta) \xrightarrow[n\to\infty]{} 0$, then $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta, \quad \forall \theta \in \Theta$.*

3.1.3 Confidence intervals

Definition 3.1.5 (Confidence interval)**.** *Let $\alpha \in (0,1)$. A confidence interval with level $1-\alpha$ for the parameter θ is an interval $I_{1-\alpha} = [T_1, T_2]$, where T_1 and T_2 are two statistics and*

$$\mathbb{P}(\theta \in I_{1-\alpha}) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

A confidence interval with asymptotic level $1-\alpha$ for the parameter θ is an interval $I_{1-\alpha,n} = [T_{1,n}, T_{2,n}]$, where $T_{1,n}$ and $T_{2,n}$ are two statistics such that for every $\theta \in \Theta$, $\lim_{n\to\infty} \mathbb{P}(\theta \in I_{1-\alpha})$ exists and

$$\lim_{n\to\infty} \mathbb{P}(\theta \in I_{1-\alpha}) \geq 1 - \alpha.$$

There are mainly two manners to build confidence intervals:

1. knowing an estimator $\hat{\theta}_n$ and its exact or asymptotic distribution (through its asymptotic normality for instance);
2. using concentration bounds such as Bienaymé-Tchebychev's or Hoeffding's inequality.

3.1.4 Hypothesis testing

Definition 3.1.6 (Hypothesis test)**.** *Let $\{\Theta_0, \Theta_1\}$ be partition of Θ and $\alpha \in (0,1)$. A test of level α for the hypotheses*

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1,$$

is a statistic $T = \mathbf{1}_{\mathbf{X}_\theta \in \mathcal{R}_\alpha}$ (where the rejection region $R_\alpha \subset E$ has to be defined), such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}(T = 1) \leq \alpha.$$

A test with asymptotic level α is a statistic $T_n = \mathbf{1}_{\mathbf{X}_\theta \in \mathcal{R}_\alpha}$, such that the limit exists and

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_0} \mathbb{P}(T_n = 1) \leq \alpha.$$

We can see a test as an estimator of the unknown quantity $\mathbf{1}_{\theta\in\Theta_1}$. The interpretation is the following : when a realization t of a test T equals 1, we say that we reject the null hypothesis (*i.e.* we conclude that $\theta \in \Theta_1$), while when $t = 0$, we say that we accept the null hypothesis (*i.e.* we conclude that $\theta \in \Theta_0$).

Often, a statistical test has the form $T = \mathbf{1}_{S>c_\alpha}$, where S is a test statistic and c_α is a critical value.

Definition 3.1.7 (Power and size)**.** *Let T be a hypothesis test for the hypotheses*

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1.$$

1. *The power function is $\beta : \theta \in \Theta \mapsto \mathbb{P}(T = 1)$.*
2. *The size is $\sup_{\theta\in\Theta_0} \beta(\theta) = \sup_{\theta\in\Theta_0} \mathbb{P}(T = 1)$.*
3. *H_0 (respectively H_1) is said simple if Θ_0 (respectively Θ_1) is a singleton, and composite otherwise.*
4. *The test T is one-sided if Θ_1 has the form $(-\infty, A) \cap \Theta$ or $\Theta \cap (A, \infty)$ for some $A \in \mathbb{R}$ and two-sided otherwise.*

Proposition 3.1.3 (Interval based test)**.** *Let $\{\Theta_0, \Theta_1\}$ be partition of Θ, $\alpha \in (0, 1)$ and an interval $J_{1-\alpha} = [T_1, T_2]$, where T_1 and T_2 are two statistics such that:*

$$\mathbb{P}(\theta \in J_{1-\alpha}) \geq 1 - \alpha, \quad \forall \theta \in \Theta_0.$$

Then, $T = \mathbf{1}_{\Theta_0 \cap J_{1-\alpha} = \emptyset}$ is a test of level α for the hypotheses

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1.$$

Let $I_{1-\alpha}(\theta)$ be a confidence interval for θ. In practice, we take $J_{1-\alpha} = I_{1-\alpha}$ when H_0 is composite and $J_{1-\alpha}$ to be "a confidence interval only true on H_0" when H_0 is simple. Let us remark that choosing such a test is not always a good idea since the construction of the interval $J_{1-\alpha}$ is *a priori* independent of Θ_0 but...there are often many confidence intervals for θ, all are not equivalent and all provide different tests...for the same hypotheses.

Definition 3.1.8 (p-value)**.** *Let $(T_\alpha)_{\alpha\in(0,1)}$ be a family of tests of level α for the same hypotheses. We call p-value*

$$\alpha_0 = \inf \{\alpha \in (0, 1) : T_\alpha = 1\}.$$

Proposition 3.1.4 (Interpretation of the p-values)**.** *Let $(T_\alpha)_{\alpha\in(0,1)}$ be a family of tests* *of size α for the hypotheses*

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1,$$

$(\alpha_0(\mathbf{X}_\theta))$ be their p-values and $x \in \mathbb{R}^n$ be a realization of $\mathbf{X}_\theta$ such that $\alpha_0(x) \in (0,1)$.

If, for every $\alpha \in (0,1)$, $T_\alpha = \mathbf{1}_{S(\mathbf{X}_\theta) > c_\alpha}$, where $S(\mathbf{X}_\theta)$ is a statistic and $\alpha \in (0,1) \mapsto c_\alpha$ is continuous, then

$$\alpha_0(x) = \sup_{\theta \in \Theta_0} \mathbb{P}(S(\mathbf{X}_\theta) > S(x)).$$

Besides being a measure of evidence against H_0, the p-value provides an alternative way to express the test T_α. Indeed, by definition of the p-value, for every $\theta \in \Theta$:

$$\begin{cases} T_\alpha = 1 & \implies \alpha_0 \leq \alpha; \\ T_\alpha = 0 & \implies \alpha_0 \geq \alpha; \\ \alpha_0 < \alpha & \implies T_\alpha = 1; \\ \alpha_0 > \alpha & \implies T_\alpha = 0. \end{cases}$$

Thus, on the event $\{\alpha_0 \neq \alpha\}$, $T_\alpha = \mathbf{1}_{\alpha_0 > \alpha}$. The next result exhibit a situation where the equivalence holds.

Proposition 3.1.5 (Usage of the p-value)**.** *Let $(T_\alpha)_{\alpha \in (0,1)}$ be a family of tests of size α for the hypotheses*

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1,$$

and $\alpha_0(\mathbf{X}_\theta)$ be their p-values.

If $\mathbb{P}(\alpha_0(\mathbf{X}_\theta) \in (0,1)) = 1$ and for every $\alpha \in (0,1)$, $T_\alpha = \mathbf{1}_{S(\mathbf{X}_\theta) > c_\alpha}$, where $S(\mathbf{X}_\theta)$ is a statistic and $\alpha \in (0,1) \mapsto c_\alpha$ is continuous, then $T_\alpha = \mathbf{1}_{\alpha_0(\mathbf{X}_{\theta_0}) < \alpha}$ a.s..

Proposition 3.1.6 (Uniformity of the p-value)**.** *Let $(T_\alpha)_{\alpha \in (0,1)}$ be a family of tests of size α for the hypotheses*

$$H_0 : \theta \in \Theta_0 = \{\theta_0\} \quad \textit{versus} \quad H_1 : \theta \in \Theta_1,$$

where $\theta_0 \in \Theta$ and $\alpha_0(\mathbf{X}_\theta)$ be their p-values.

If $\mathbb{P}(\alpha_0(\mathbf{X}_\theta) \in (0,1)) = 1$ and for every $\alpha \in (0,1)$, $T_\alpha = \mathbf{1}_{S(\mathbf{X}_\theta) > c_\alpha}$, where $S(\mathbf{X}_\theta)$ is a statistic such that $S(\mathbf{X}_{\theta_0})$ has a continuous cumulative distribution function $F_{S(\mathbf{X}_{\theta_0})}$ and $\alpha \in (0,1) \mapsto c_\alpha$ is continuous, then $\alpha_0(\mathbf{X}_{\theta_0}) \sim \mathcal{U}([0,1])$.

3.2 Parametric estimation

3.2.1 Method of moments

Definition 3.2.1 (Method of moments estimator)**.** *Let k be a positive integer. In a sampling model with $X_1 \in L^k$, a method of moments estimator (MME) of order k is any estimator*

$$\hat{\theta}_n = \varphi\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \dots, \frac{1}{n}\sum_{i=1}^n X_i^k\right),$$

where φ is a function such that:

$$\theta = \varphi(\mathbb{E}(X_1), \mathbb{E}(X_1^2), \dots, \mathbb{E}(X_1^k)), \quad \forall \theta \in \Theta.$$

With this definition, the MME may not be defined (assume for instance that $\phi(x) = \frac{1}{x}$, $\mathbb{E}(X_1) = \frac{1}{\theta}$ and $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i = 0) > 0$). The next proposition says that this is not a fatality.

Proposition 3.2.1. *Let $\mathcal{P}$ be a sampling model, k be a positive integer and assume that $X_1 \in L^k$. Let φ be a function such that:*

$$\theta = \varphi(\mathbb{E}(X_1^k)), \quad \forall \theta \in \Theta,$$

and $\hat{\theta}_n = \varphi\left(\frac{1}{n}\sum_{i=1}^n X_i^k\right)$. Then,

1. *if φ is continuous in $\mathbb{E}(X_1^k)$ for every $\theta \in \Theta$, $\hat{\theta}_n$ is well defined with probability tending to 1 and is consistent;*
2. *if φ is differentiable in $\mathbb{E}(X_1^k)$ with $\varphi'(\mathbb{E}(X_1^k)) \neq 0$ for every $\theta \in \Theta$, $\hat{\theta}_n$ is asymptotically normal with rate $1/\sqrt{n}$ and asymptotic variance $\varphi'(\mathbb{E}(X_1^k))^2\,\mathbb{V}(X_1^k)$.*

3.2.2 Empirical quantiles

Definition 3.2.2 (Empirical cumulative distribution function)**.** *Assume that $\mathbf{X}_\theta$ is a sample. Its empirical cumulative distribution function is:*

$$F_n : x \in \mathbb{R} \mapsto \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \le x}.$$

Proposition 3.2.2. *Assume that $\mathbf{X}_\theta$ is a sample and let F_{X_1} be the cumulative distribution function of X_1. For every $x \in E$:*

1. *strong consistency: $F_n(x) \xrightarrow[n\to\infty]{a.s.} F_{X_1}(x), \quad \forall \theta \in \Theta$;*
2. *asymptotic normality: $\sqrt{n}(F_n(x) - F_{X_1}(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, F_{X_1}(x)(1 - F_{X_1}(x))\right), \quad \forall \theta \in \Theta$.*

Definition 3.2.3 (Quantile)**.** *Let $\alpha \in [0,1]$ and assume that $\mathbf{X}_\theta$ is a sample. The*

α-quantile of X_1 is

$$q_\alpha = F_{X_1}^{-1}(\alpha).$$

The α-empirical quantile of the sample $\mathbf{X}_\theta$ is:

$$q_{n,\alpha} = F_n^{-1}(\alpha).$$

Proposition 3.2.3. *Let $\alpha \in (0,1)$ and assume that $\mathbf{X}_\theta$ is a sample.*

1. *If F_{X_1} is strictly increasing at q_α, then*

$$q_{n,\alpha} \xrightarrow[n\to\infty]{a.s.} q_\alpha.$$

2. *If F is differentiable at q_α with $F'(q_\alpha) > 0$, then*

$$\sqrt{n}(q_{n,\alpha} - q_\alpha) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{F'(q_\alpha)^2}\right).$$

3.2.3 Maximum likelihood

Definition 3.2.4 (Maximum likelihood estimator)**.** *Assume that the model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is dominated by a measure μ and denote, for every $\theta \in \Theta$, f_θ the density of P_θ with respect to μ. For every $\theta \in \Theta$, we call likelihood function*

$$L_n(\theta') = f_{\theta'}(\mathbf{X}_\theta), \quad \forall \theta' \in \Theta.$$

Then, a maximum likelihood estimator (MLE) $\hat{\theta}_n$ is any random variable verifying:

$$\hat{\theta}_n \in \arg\max_{\hat{\theta}:estimator} L_n(\hat{\theta}) \ a.s., \quad \forall \theta \in \Theta.$$

Theorem 3.2.4. *Let us assume that:*

1. *$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a sampling model dominated by a measure μ and let, for every $\theta \in \Theta$, f_θ be the density of P_θ with respect to μ;*
2. *the model $\mathcal{P}$ is identifiable,* i.e. *the mapping $\theta \in \Theta \to P_\theta$ is injective;*
3. *the log-likelihood converges uniformly to the opposite of the Kullback-Leibler divergence:*

$$\sup_{\theta' \in \Theta} |M_n(\theta') - \mathbb{E}(M_n(\theta'))| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \quad \forall \theta \in \Theta,$$

where $M_n(\theta') = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_{\theta'}(X_i)}{f_\theta(X_i)}\right), \quad \forall \theta' \in \Theta.$

Then any MLE $\hat{\theta}_n$ of θ is consistent, i.e. $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta, \quad \forall \theta \in \Theta.$

Theorem 3.2.5. *Let us assume that:*

1. $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ *is a sampling model dominated by a measure* μ *and let, for every* $\theta \in \Theta$, f_θ *be the density of* P_θ *with respect to* μ*;*
2. *the model* $\mathcal{P}$ *is regular,* i.e.*:*
   (a) *for every* $\theta \in \Theta$*, the support* $\mathcal{S} = \{x \in E : f_\theta(x) > 0\}$ *of* P_θ*, is independent of* θ*;*
   (b) *for* μ*-almost all* $x \in E$*, the function* $\theta \in \Theta \mapsto f_\theta(x)$ *is continuously differentiable;*
   (c) *for every* $\theta \in \Theta$, $S_\theta \in L^2$ *and* $\theta \in \Theta \mapsto \mathbb{E}(S_\theta^2)$ *is continuous, where we have defined* $\ell : \theta' \mapsto \log f_{\theta'}(X_1)$ *and* $S_\theta = \ell'(\theta)$ *(*S_θ *is called the score function).*

Then, any MLE $\hat{\theta}_n$ *such that, for a given* $\theta \in \Theta$, $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta$ *and the Fisher information* $I(\theta) = \mathbb{E}(S_\theta^2) = \mathbb{V}(S_\theta) > 0$*, we have:*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right),$$

and

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

In a nutshell, the previous theorem says that if the model $\mathcal{P}$ is regular with a Fisher information positive for every $\theta \in \Theta$ and if an MLE $\hat{\theta}_n$ is consistent, then it is asymptotically normal. The last statement makes it possible to design confidence intervals.

3.3 Linear model

3.3.1 Multiple linear model

Let $(\underline{x}_1, \underline{y}_1), \ldots, (\underline{x}_n, \underline{y}_n) \in \mathbb{R}^p \times \mathbb{R}$ be some measurements. We would like to study the relationship between x_i and y_i (for ever $i \in [\![1, n]\!]$) having in mind that it is easier to measure x_i than y_i in the future. For this reason, we assume that measurements are the realization of n *iid* random pairs $(\underline{X}_1, \underline{Y}_1), \ldots, (\underline{X}_n, \underline{Y}_n)$, where for every $i \in [\![1, n]\!]$, $\underline{X}_i$ is a random vector of covariates (or features) with values in $\mathbb{R}^p$ and the response $\underline{Y}_i$ is a real-valued random variable.

Our goal here (that of regression actually) is to approximate the distribution of $\underline{Y}_i|\underline{X}_i = x$ (this is the same for every $i \in [\![1, n]\!]$), denoted P_x, for all $x \in \mathbb{R}^p$. For the sake of simplicity, least squares regression (a particular case of regression) focuses only on the regression function $x \in \mathbb{R}^p \mapsto \mathbb{E}(\underline{Y}_1|\underline{X}_1 = x)$ as a characteristic of P_x (having in mind that we would like to "predict" y_i based on the measurement x_i for any $i \in [\![1, n]\!]$). In addition, since we are interested in conditional distributions $(P_x)_{x\in\mathbb{R}^p}$, it is enough to work given that $\underline{X}_1 = x_1, \ldots, \underline{X}_n = x_n$ for any values $(x_1, \ldots, x_n) \subset \mathbb{R}^p$, *i.e.* with the random vector $(\underline{Y}_1|\underline{X}_1 = x_1, \ldots, \underline{Y}_n|\underline{X}_n = x_n)$, which is the same as $(\underline{Y}_1, \ldots, \underline{Y}_n)|(\underline{X}_1, \ldots, \underline{X}_n) = (x_1, \ldots, x_n)$. In other words, we work like if $(\underline{X}_1, \ldots, \underline{X}_n)$ is fixed to $(x_1, \ldots, x_n)$ and $(\underline{Y}_1, \ldots, \underline{Y}_n)$ only is random (what is called the fixed design, in opposition to the random design in which we work with the random vector $(\underline{X}_1, \ldots, \underline{X}_n)$).

Thus, let $(x_1, \ldots, x_n) \subset \mathbb{R}^p$ and $\mathcal{P}$ be the statistical model:

$$\mathcal{P} = \left\{ P_{(\beta,\sigma^2)}, (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^* \right\},$$

where for every $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$, if $(Y_1, \ldots, Y_n) \sim P_{(\beta,\sigma^2)}$, then:

1. $Y_1, \ldots, Y_n$ are independent;
2. for every $i \in [\![1, n]\!]$, $\mathbb{E}(Y_i) = \beta^\top x_i$;
3. for every $i \in [\![1, n]\!]$, $\mathbb{V}(Y_i) = \sigma^2$.

This statistical model (which is not identifiable) actually assumes that the regression function $r(x) = \mathbb{E}(Y) = \beta^\top x$ is linear.

Now, let us assume that the true distribution is in the model $\mathcal{P}$:

$$\exists (\beta_0, \sigma_0^2) \in \mathbb{R}^p \times \mathbb{R}_+^* : (\underline{Y}_1 | \underline{X}_1 = x_1, \ldots, \underline{Y}_n | \underline{X}_n = x_n) \sim P_{(\beta_0, \sigma_0^2)},$$

and let for every $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$,

$$\mathbf{Y}_{(\beta,\sigma^2)} = (Y_1, \ldots, Y_n) \sim P_{(\beta,\sigma^2)}.$$

The random vector $\mathbf{Y}_{(\beta,\sigma^2)}$ is our observation, based on which we have to estimate the generic parameters β (of the regression function) and σ^2 (variance of the noise). Let us remark that in this modeling, if the random variables $Y_1, \ldots, Y_n$ are independent, they are not *iid* (the notation $\mathbf{Y}$ different from $\mathbf{X}$ is aimed at underlying that $\mathbf{Y}$ is not a sample).

Let now $\mathbb{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times p}$ be the desing matrix. Then, we can characterize the distribution $P_{(\beta,\sigma^2)}$ (and so the model $\mathcal{P}$) in the following manner:

$$\mathbf{Y}_{(\beta,\sigma^2)} = \mathbb{X}\beta + \epsilon,$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a random vector such that:

1. $\epsilon_1, \ldots, \epsilon_n$ are independent;
2. for every $i \in [\![1, n]\!]$, $\mathbb{E}(\epsilon_i) = 0$;
3. for every $i \in [\![1, n]\!]$, $\mathbb{V}(\epsilon_i) = \sigma^2$.

Definition 3.3.1 (Least squares estimator)**.** *A least squares estimator of β is a random vector $\hat{\beta}$ such that*

$$\hat{\beta} \in \arg\min_{\beta' \in \mathbb{R}^p} \left\| \mathbf{Y}_{(\beta,\sigma^2)} - \mathbb{X}\beta' \right\|^2, \quad \forall (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*.$$

From now on, we shall assume that $\operatorname{rank}(\mathbb{X}) = p$, otherwise, the forthcoming estimator is not unique.

Property 3.3.1. *The least squares estimator of β is unique and can be expressed as:*

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbf{Y}_{(\beta,\sigma^2)}, \quad \forall(\beta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*.$$

In addition, $\mathbb{E}(\hat{\beta}) = \beta$ *and* $\mathbb{V}(\hat{\beta}) = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}$.

Proposition 3.3.2. *An unbiased estimator of σ^2 is*

$$\widehat{\sigma}^2 = \frac{\|\mathbf{Y}_{(\beta,\sigma^2)} - \mathbb{X}\hat{\beta}\|^2}{n-p}, \quad \forall(\beta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*.$$

3.3.2 Gaussian linear model

Let us now consider the statistical model

$$\mathcal{P} = \left\{ P_{(\beta,\sigma^2)} = \mathcal{N}(\mathbb{X}\beta, \sigma^2 \boldsymbol{I}_n), (\beta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^* \right\},$$

where $\boldsymbol{I}_n$ is the identity matrix of size n. We still assume that $\text{rank}(\mathbb{X}) = p$, making now the model $\mathcal{P}$ identifiable. Actually, considering this new model boils down to make stronger assumptions with respect to the previous section, since we now have:

$$\mathbf{Y}_{(\beta,\sigma^2)} = \mathbb{X}\beta + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. This, however, makes it possible to make the link with maximum likelihood estimation and to exhibit the exact distributions of the estimators introduced just before.

Proposition 3.3.3. *The MLE of (β, σ^2) is*

$$\tilde{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top \mathbf{Y}_{(\beta,\sigma^2)} = \hat{\beta} \quad \textit{and} \quad \widetilde{\sigma}^2 = \frac{\|\mathbf{Y}_{(\beta,\sigma^2)} - \mathbb{X}\tilde{\beta}\|^2}{n} = \frac{n-p}{n}\widehat{\sigma}^2,$$

for every $\forall(\beta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$.

The next result is an application of Cochran's theorem.

Property 3.3.4. *For every* $\forall(\beta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$, *we have:*

1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1})$;
2. $\hat{\beta} \perp\!\!\!\perp \widehat{\sigma}^2$;
3. $\frac{n-p}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-p}$.

In addition:

1. *for every* $j \in [\![1,p]\!]$,
$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbb{X}^\top \mathbb{X})^{-1}]_{jj}}} \sim \mathcal{T}_{n-p};$$
2. $$\frac{(\hat{\beta}-\beta)^\top \mathbb{X}^\top \mathbb{X}(\hat{\beta}-\beta)}{p\widehat{\sigma}^2} \sim \mathcal{F}^p_{n-p}.$$

This last property makes it possible to derive confidence intervals for σ^2 and the components of β. We also have access to a confidence hyperellipsoid for β.

We now focus on our initial wish, which is to "predict" an unseen response. To be a bit more formal, let $(\underline{x}_{n+1}, \underline{y}_{n+1})$ be a novel measurement (assumed to be a realization of some random pair $(\underline{X}_{n+1}, \underline{Y}_{n+1})$ independent from and identically distributed to $(\underline{X}_1, \underline{Y}_1), \dots, (\underline{X}_n, \underline{Y}_n)$), and suppose that we only know $\underline{x}_{n+1}$. We would like to "predict" $\underline{y}_{n+1}$ and a natural way to do it is to compute $\hat{\beta}_0^\top \underline{x}_{n+1}$ (where $\hat{\beta}_0$ is the realization of $\hat{\beta}$ based on $(\underline{x}_1, \underline{y}_1), \dots, (\underline{x}_n, \underline{y}_n)$).

From the beginning, the word "prediction" is quoted because its meaning is unclear. Indeed, it is not an estimation of $\underline{y}_{n+1}$ since this quantity is a realization of a random variable $\underline{Y}_{n+1}$ and not a fixed parameter. With our modeling (based on random variables, which are objects characterized by their distribution rather than their mapping input→output), the best we can do is to characterize the distribution of $\underline{Y}_{n+1}$ and without the Bayesian theory, there no best way[1] than providing a prediction interval, *i.e.* an interval containing $\underline{Y}_{n+1}$ with high probability (let us note that this is not a confidence interval since the quantity $\underline{Y}_{n+1}$ is random). This leads to the forthcoming definition of the prediction error and results.

Before proceeding, let us go back to our modeling: let $x_{n+1} \in \mathbb{R}$ be any real value and for every $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}^*_+$, let $Y_{n+1} \sim \mathcal{N}(\beta^\top x_{n+1}, \sigma^2) \perp\!\!\!\perp \mathbf{Y}_{(\beta,\sigma^2)}$.

Proposition 3.3.5. *For every* $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}^*_+$,

1. *the distribution of the prediction is:* $\hat{\beta}^\top x_{n+1} \sim \mathcal{N}(\beta^\top x_{n+1}, \sigma^2)$*;*
2. *the distribution of the prediction error is:*
$$Y_{n+1} - \hat{\beta}^\top x_{n+1} \sim \mathcal{N}\left(0, \sigma^2(1 + x_{n+1}^\top (\mathbb{X}^\top \mathbb{X})^{-1} x_{n+1})\right);$$
3. *for every* $\alpha \in (0,1)$*, denoting* t_α *the* $\left(1 - \frac{\alpha}{2}\right)$*-quantile of* $\mathcal{T}_{n-p}$*,*
$$\mathbb{P}\left(Y_{n+1} \in \left[\hat{\beta}^\top x_{n+1} \pm \hat{\sigma}\sqrt{1 + x_{n+1}^\top (\mathbb{X}^\top \mathbb{X})^{-1} x_{n+1}} t_\alpha\right]\right) = 1 - \alpha.$$

[1] Actually, we could think of "prediction" as estimating $\mathbb{E}(\underline{Y}_{n+1} | \underline{X}_{n+1} = \underline{x}_{n+1}) = \beta_0^\top \underline{x}_{n+1}$ but this quantity is of limited interest because there is no reason for $\underline{y}_{n+1}$ or $\underline{Y}_{n+1}$ to be close to it.

3.4 Exercises

Exercise 3.1 (Mean squared error)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables, such that $m = \mathbb{E}[X_1]$ and $\sigma^2 = \mathbb{V}(X_1)$. We assume that σ^2 is known but that m is unknown.

1. Show that the estimators $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $Z_n = \frac{X_{n-1}+X_n}{2}$ are unbiased.
2. Compute their mean squared error. Which estimtator is better?
3. What do you think of $W_n = 0$?

Exercise 3.2 (Method of moments)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables. For the following two distributions, propose an estimator by the method of moments and show that it is asymptotically normal.

1. $X_1 \sim \mathcal{U}([0,\theta])$, for $\theta > 0$.
2. $X_1 \sim \mathcal{E}(\lambda)$, for $\lambda > 0$.

Exercise 3.3 (Maximum likelihood)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables. For the following distributions, compute the maximum likelihood estimator and show that it is asymptotically normal.

1. $X_1 \sim \mathcal{N}(\mu, 1)$, for $\mu \in \mathbb{R}$.
2. $X_1 \sim \mathcal{E}(\lambda)$, for $\lambda > 0$.
3. $X_1 \sim \mathcal{P}(\lambda)$, for $\lambda > 0$.

Let $\alpha \in (0,1)$. Give confidence intervals of asymptotic level $1-\alpha$ for unknown parameters and propose tests to asses the null hypothesis that the unknown parameter equals 1.

Exercise 3.4 (Empirical mean and median)**.** Let $(X_n)_{n\geq 1}$ be a sequence of *iid* random variables. We consider two estimators: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\tilde{\mu} = q_{n,1/2}$ (the empirical median). For the following distributions, compare the asymptotic variances of $\hat{\mu}$ and $\tilde{\mu}$.

1. $X_1 \sim \mathcal{N}(\mu, 1)$, for $\mu \in \mathbb{R}$.
2. X_1 has density $f : x \in \mathbb{R} \mapsto \frac{1}{2}\,\mathrm{e}^{-|x-\mu|}$, for $\mu \in \mathbb{R}$. Let us remark that we have $\mathbb{E}[X_1] = \mu$ and $\mathbb{V}(X_1) = 2$.