Stochastic modeling
MACS201, MACS203a

François Roueff

September 9, 2021

Preliminary comments

Foreword

The main goal of these lecture notes is to provide a solid mathematical introduction to the most important principles of stochastic modeling.

We start with the foundations on Hilbert spaces, probability and mathematical statistics required for this goal. Then we introduce the general framework of random processes, with focuses on weakly stationary time series and martingales in discrete time. The final part is dedicated to Markov chains, which embed the vast majority of stochastic models used in applications.

These lecture notes have been written as a material for the courses MACS201 and MACS203a, which roughly corresponds to the first semester of second years students at Télécom ParisTech enrolled in the program *Modélisation aléatoire et calcul scientifique* (stochastic modeling and numerical analysis). They gather what we perceive as the most fundamental results that are useful to any scientists working with stochastic models. We tried to give a rigorous account which embraces these topics as completely as possible. The sufficient but also necessary prerequisites are limited to a good undergraduate level in probability and functional analysis (for which we recommend e.g. [6, 7, 8]). Starting from there, these notes are almost self-contained, as only a very few results rely on more advanced external references: the two existence Theorems 2.1.7 and 4.2.2, and Appendix A, which contains a brief account on the weak convergence in metric spaces and which is provided here for sake of completeness. Most of the material about Markov chains originates from a preliminary version of [3] that was courteously made available by their authors. We refer the reader to this book for a very complete view of the up-to-date knowledge on this topic.

Most of the mathematical theory presented here has been developed in the second half of XIXth century, building on the theory of probability and functional analysis completed in the previous half century. They are at the core of many recent advances in applied sciences, not restricted to those actually relying on random models (such as structured data analysis, mathematical finance, ecology or insurance). Indeed random approaches have been proved efficient to tackle many practical problems involving complex objects: to compute a high dimensional integral, evaluate the connectivity of a huge network, to find a minima of a (possibly rough) non-convex function...

To conclude this foreword, these lecture notes mainly contain an account of the fundamental results used in stochastic modeling and the efforts have been principally aimed at selecting, gathering and presenting these important results in a consistent, almost self-contained and concise fashion. The expected benefits are, hopefully, to give a reassuring view to the reader of topics that otherwise appear disconnected from one another.

Notation and conventions

Vectors of $\mathbb{C}^d$ are identified to $d \times 1$ matrices.

The Hermitian norm of $x \in \mathbb{C}^d$ is denoted by $|x|$.

The transpose of matrix A is denoted by A^T.

The conjugate transpose of matrix A is denoted by A^H.

The set $\mathbb{T}$ is the quotient space $\mathbb{R}/(2\pi\mathbb{Z})$ (or any interval congruent to $[0, 2\pi)$).

The variance of the random variable X is denoted by $\mathrm{Var}\,(X)$.

The variance-covariance matrix of the random vector $\mathbf{X}$ is denoted by $\mathrm{Cov}(\mathbf{X})$.

The covariance matrix between the random vectors $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})$.

The Gaussian distribution with mean μ and covariance Q is denoted by $\mathcal{N}(\mu, Q)$.

$X \sim P$ means that the random variable X has distribution P.

For a r.v. X on $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{P}^X$ denotes the probability distribution of X, $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$.

$(X_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$ means that $(X_t)_{t\in\mathbb{Z}}$ is a weak white noise with variance σ^2.

$(X_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0, \sigma^2)$ means that $(X_t)_{t\in\mathbb{Z}}$ is a strong white noise with (finite) variance σ^2.

$(X_t)_{t\in T} \overset{\text{iid}}{\sim} P$ means that $(X_t)_{t\in T}$ are independent variables with common distribution P.

Given $f : \mathbb{R}^p \to \mathbb{R}^q$ differentiable, $\partial f : \mathbb{R}^p \to \mathbb{R}^{p\times q}$ is the gradient of each component of f stacked columnwise.

Pursuing with the former example, if f is twice differentiable, $\partial\partial^T f : \mathbb{R}^p \to \times\mathbb{R}^p \times \mathbb{R}^{p\times q}$ are the Hessian matrices conveniently stacked depending of the context.

X_n converges a.s., in probability or weakly to X is denoted by $X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{P} X$ or $X_n \Longrightarrow X$, respectively.

The finite distributions of X_n converge weakly to that of X is denoted by $X_n \overset{\text{fidi}}{\Longrightarrow} X$.

$\sigma(X)$ denotes the smallest σ-field containing the reciprocal images of X.

$\mathcal{F} \vee \mathcal{G}$ denotes the smallest σ-field containing all the ellements of $\mathcal{F}$ and $\mathcal{G}$.

$\mathcal{F} \otimes \mathcal{G}$ denotes the smallest σ-field containing all the ellements $A \times B$ with $A \in \mathcal{F}$ and $B \in \mathcal{G}$.

Functions and measures spaces symbols

The set of all measurable functions defined on $(\mathsf{X}, \mathcal{X})$ and valued in $(\bar{\mathbb{R}}_+, \mathcal{B}(\bar{\mathbb{R}}_+))$ is denoted by $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$

The set of all bounded continuous functions defined on the metric space (X, d) and valued in $\mathbb{R}$ is denoted by $\mathrm{C}_b(\mathsf{X}, d)$

The set of all Lipschitz functions and valued in $\mathbb{R}$ defined on the metric space (X, d) is denoted by $\mathrm{Lip}(\mathsf{X}, d)$

The set of all bounded and Lipschitz functions defined on the metric space (X, d) and valued in $\mathbb{R}$ is denoted by $\mathrm{Lip}_b(\mathsf{X}, d)$

The set of all non-negative σ-finite measures defined on $(\mathsf{X}, \mathcal{X})$ is denoted by $\mathbb{M}_+(\mathsf{X}, \mathcal{X})$

The set of all probability measures defined on $(\mathsf{X}, \mathcal{X})$ is denoted by $\mathbb{M}_1(\mathsf{X}, \mathcal{X})$

In all above definitions, when no ambiguity arises, we will often omit the second arguments $\mathcal{X}$ or d *e.g.* simply writing $\mathrm{C}_b(\mathsf{X})$ or $\mathbb{M}_1(\mathsf{X})$.

Contents

Part I

Foundations

Chapter 1

Hilbert spaces

Basic knowledge of Hilbert spaces is quite useful for time series, and, more generally stochastic modeling. Here we gather some essential definitions and results on Hilbert spaces. Most results are elementary. A detailed account on this topic can be found in [8].

1.1 Definitions

Definition 1.1.1 (Inner-product spaces)**.** *Let $\mathcal{H}$ be a complex linear space. An* Inner-product *on $\mathcal{H}$ is a function*

$$\langle \cdot, \cdot \rangle : x, y \in \mathcal{H} \times \mathcal{H} \mapsto \langle x, y \rangle \in \mathbb{C}$$

which satisfies the following properties

(i) for all $(x, y) \in \mathcal{H} \times \mathcal{H}$, $\langle x, y \rangle = \overline{\langle y, x \rangle}$

(ii) for all $(x, y) \in \mathcal{H} \times \mathcal{H}$ and all $(\alpha, \beta) \in \mathbb{C} \times \mathbb{C}$, $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

(iii) for all $x \in \mathcal{H}$, $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$.

Then the application

$$\| \cdot \| : x \in \mathcal{H} \mapsto \sqrt{\langle x, x \rangle} \geq 0$$

defines a norm on $\mathcal{H}$.

Example 1.1.1 ($\mathbb{C}^n$)**.** *The space of column vectors $x = [x_1 \quad \cdots \quad x_n]^T$, where $x_k \in \mathbb{C}$ is a linear space on which the application*

$$\langle x, y \rangle = y^H x = \sum_{k=1}^{n} x_k \overline{y_k}$$

defines an inner product.

Example 1.1.2 (ℓ^2)**.** *The space of complex-valued sequences $\{x_k\}_{k \in \mathbb{N}}$ such that $\sum_{k=0}^{\infty} |x_k|^2 < \infty$ is a linear space. Define for all x et y in this space,*

$$\langle x, y \rangle = \sum_{k=0}^{\infty} x_k \overline{y_k} \, .$$

The sum is well defined and finite since $|x_k \overline{y_k}| \leq (|x_k|^2 + |y_k|^2)/2$. Properties (i-iii) of definition 1.1.1 are easily verified. We thus obtained an inner-product space, denoted as ℓ^2.

Example 1.1.3 (Squared integrable functions)**.** *The space $\mathcal{L}^2(T)$ of $\mathbb{C}$-valued Borel functions defined on an interval $T \subset \mathbb{R}$ whose modulus is squared integrable ($\int_T |f(t)|^2 dt < \infty$) is a linear space. Define*

$$(f,g) \in \mathcal{L}^2(T) \times \mathcal{L}^2(T) \mapsto \langle f,g \rangle = \int_T f(t)\overline{g(t)}\mathrm{d}t \ .$$

As for ℓ^2, Properties (i) and (ii) of Definition 1.1.1 hold. However Property (iii) fails to hold since :

$$\langle f,f \rangle = 0 \not\Rightarrow \forall t \in T \ f(t) = 0$$

Instead it implies $f = 0$ a.-e. (almost-everywhere). As a consequence, the space $\mathcal{L}^2(T)$ endowed with $\langle \cdot,\cdot \rangle$ is not an inner-product space. Nevertheless the space $L^2(T)$ of the equivalence classes of $\mathcal{L}^2(T)$ for the a.e. equality is an inner-product space.

Example 1.1.4 (Finite variance random variables)**.** *As in Example 1.1.3, for all probability space $(\Omega, \mathcal{F}, \mathbb{P})$, one defines $\mathcal{H} = \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ (denoted by $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ if no possible confusion occurs) as the space of all complex-valued random variables X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

$$\mathbb{E}\left[|X|^2\right] < \infty \ .$$

Define moreover

$$(X,Y) \in \mathcal{L}^2(\Omega) \times \mathcal{L}^2(\Omega) \mapsto \langle X,Y \rangle = \mathbb{E}\left[X\overline{Y}\right] \ .$$

For the same reasons as in Example 1.1.3, we define the inner-product space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ (or simply L^2 if no ambiguity occurs) as the space of the equivalence classes of $\mathcal{L}^2(\Omega)$ for the a.s. equality. This example and Example 1.1.3 can be extended to all measured space $(\Omega, \mathcal{F}, \mu)$ by setting

$$(f,g) \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu) \times \mathcal{L}^2(\Omega, \mathcal{F}, \mu) \mapsto \langle f,g \rangle = \int f\, \overline{g}\ \mathrm{d}\mu \ .$$

We have the following result.

Theorem 1.1.1. *For all $x, y \in \mathcal{H} \times \mathcal{H}$, we have :*

a*) Cauchy-Schwarz Inequality: $|\langle x,y \rangle| \leq \|x\|\|y\|$,*

b*) Triangular inequality: $|\|x\| - \|y\|| \leq \|x-y\| \leq \|x\| + \|y\|$,*

c*) Parallelogram inequality:*

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

Definition 1.1.2 (Convergence in $\mathcal{H}$)**.** *Let (x_n) be a sequence included in an inner-product space $\mathcal{H}$ and $x \in \mathcal{H}$. We say that (x_n) converges to x in $\mathcal{H}$ if $\|x_n - x\| \to 0$ as $n \to +\infty$. We will denote $x_n \to x$ if no confusion occurs with another convergence.*

It is easy to show that, for all $y \in \mathcal{H}$, the application $\langle \cdot, y \rangle : \mathcal{H} \to \mathbb{C}$, $x \mapsto \langle x,y \rangle$ is a continuous linear form. In fact we have the following continuity result.

Theorem 1.1.2 (Continuity of the inner product)**.** *If $x_n \to x$ and $y_n \to y$ in the inner-product space $\mathcal{H}$, then $\langle x_n, y_n \rangle \to \langle x,y \rangle$. In particular, $\|x_n\| \to \|x\|$.*

Proof. Using the triangle inequality and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\langle x, y\rangle - \langle x_n, y_n\rangle &= \langle (x - x_n) + x_n, (y - y_n) + y_n\rangle - \langle x_n, y_n\rangle \\ &= \langle x - x_n, y - y_n\rangle + \langle x - x_n, y_n\rangle + \langle x_n, y - y_n\rangle \\ &\leq \|x_n - x\|\|y_n - y\| + \|x_n - x\|\|y_n\| + \|y_n - x\|\|x_n\|\end{aligned}$$

This implies the result, since the sequences (x_n) are (y_n) bounded. □

Definition 1.1.3 (Hilbert space)**.** *An inner-product space $\mathcal{H}$ is called an* Hilbert space *if it is complete (that is, every Cauchy sequence converges).*

Recall that a normed space is complete if and only if every absolutely convergent series is convergent see [6, Proposition 5 in Chapter 6, Page 124].

Example 1.1.5 (ℓ^2)**.** *The space ℓ^2 is a Hilbert space. Let (a_n) be a Cauchy sequence in ℓ^2. Denote*

$$a_n = (a_{n,1}, a_{n,2}, \dots) ,$$

then, for all $\epsilon > 0$, there exists N such that, for all $n, m \geq N$,

$$\sum_{k=1}^{\infty} |a_{m,k} - a_{n,k}| \leq \epsilon^2 . \tag{1.1}$$

Let k be fixed. The previous displays shows that $(a_{n,k})_n$ is a Cauchy sequence in $\mathbb{C}$. Let α_k denote its limit. Further denote $a = (\alpha_k)$. It remains to show that $a \in \ell^2$ and $\lim_{n\to\infty} \|a_n - a\| = 0$. Using (1.1), we have for all $p \in \mathbb{N}$, and all $m, n \geq N$,

$$\sum_{k=1}^{p} |a_{m,k} - a_{n,k}|^2 \leq \sum_{k=1}^{\infty} |a_{m,k} - a_{n,k}|^2 \leq \epsilon^2.$$

Hence, for all $p \in \mathbb{N}$ and all $n \geq N$, $\lim_{m\to\infty} \sum_{k=1}^{p} |a_{m,k} - a_{n,k}|^2 = \sum_{k=1}^{p} |\alpha_k - a_{n,k}|^2 \leq \epsilon^2$. Taking the limit as $p \to \infty$, we thus get, for all $n \geq N$,

$$\|a - a_n\|^2 = \sum_{k=1}^{\infty} |\alpha_k - a_{n,k}|^2 \leq \epsilon^2,$$

which implies $(a - a_n) \in \ell^2$, thus $a \in \ell^2$. Since ϵ is arbitrary, we also get that $\lim_{n\to\infty} \|a - a_n\| = 0$.

Proposition 1.1.3 (L^2 spaces)**.** *For all measured space $(\Omega, \mathcal{F}, \mu)$, the space $L^2(\Omega, \mathcal{F}, \mu)$ (see Example 1.1.4) endowed with*

$$\langle f, g\rangle = \int f\bar{g} \, \mathrm{d}\mu$$

is a Hilbert space.

A more general result on L^p spaces is given in [6, Proposition 6 in Chapter 6, Page 126].

Example 1.1.6 (A non-complete inner-product space)**.** *Let $\mathcal{C}([-\pi,\pi])$ the space of continuous functions on $[-\pi,\pi]$. It is a subspace of the Hilbert space $L^2([-\pi,\pi])$. However it is not closed since $\mathbb{1}_{[-\pi/2,\pi/2]}$ can be approximated by continuous functions with arbitrarily small L^2 error. Hence $\mathcal{C}([-\pi,\pi])$ endowed with*

$$\langle f,g\rangle = \int_{-\pi}^{\pi} f\bar{g}$$

is not a complete space, although it is an inner product space.

Definition 1.1.4 (Generated subspace and its closure)**.** *Let $\mathcal{X}$ be a subspace of $\mathcal{H}$. We denote by* Span $(\mathcal{X})$ *the subspace of all finite linear combinations of vectors in $\mathcal{X}$ and by* $\overline{\text{Span}}\,(\mathcal{X})$ *the* closure *of* Span $(\mathcal{X})$ *in $\mathcal{H}$, that is the smallest closed subspace of $\mathcal{H}$ that contains* Span $(\mathcal{X})$*. In fact* $\overline{\text{Span}}\,(\mathcal{X})$ *contains and only contains all elements of $\mathcal{H}$ which are L^2 limits of sequences included in* Span $(\mathcal{X})$*.*

Definition 1.1.5 (Orthogonality)**.** *Two vectors $x,y\in\mathcal{H}$ are* orthogonal*, if $\langle x,y\rangle = 0$, which we denoted by $x\perp y$. If $\mathcal{S}$ is a subspace of $\mathcal{H}$, we write $x\perp\mathcal{S}$ if $x\perp s$ for all $s\in\mathcal{S}$. Also we write $\mathcal{S}\perp\mathcal{T}$ if all vectors in $\mathcal{S}$ are orthogonal to $\mathcal{T}$.*

Take two subspaces $\mathcal{A}$ et $\mathcal{B}$ such that $\mathcal{H}=\mathcal{A}+\mathcal{B}$, that is, for all $h\in\mathcal{H}$, there exist $a\in\mathcal{A}$ et $b\in\mathcal{B}$ such that $h=a+b$. If moreover $\mathcal{A}\perp\mathcal{B}$ we will denote $\mathcal{H}=\mathcal{A}\overset{\perp}{\oplus}\mathcal{B}$.

Definition 1.1.6 (Orthogonal set)**.** *Let $\mathcal{E}$ be a subset of an Hilbert space $\mathcal{H}$. The orthogonal set of $\mathcal{E}$ is defined as*

$$\mathcal{E}^{\perp} = \{x\in\mathcal{H} : \forall y\in\mathcal{E}\ \ \langle x,y\rangle = 0\}$$

We will need the following result, whose proofs are left to the reader as an exercise.

Theorem 1.1.4. *If $\mathcal{E}$ is a subset of an Hilbert space $\mathcal{H}$, then $\mathcal{E}^{\perp}$ is closed.*

1.2 Orthogonal and orthonormal bases

Definition 1.2.1 (Orthogonal and orthonormal sets)**.** *Let E be a subset of $\mathcal{H}$. It is an orthogonal set if for all $(x,y)\in E\times E$, $x\neq y$, $\langle x,y\rangle = 0$. If moreover $\|x\|=1$ for all $x\in E$, we say that E is orthonormal.*

Linear combinations of vectors in an orthogonal set have the following remarkable property. Let E be an orthogonal set and $x_1,\dots,x_n\in E$ distinct. Then for all $(\alpha_1,\dots,\alpha_n)\in\mathbb{C}^n$,

$$\left\|\sum_{k=1}^{n}\alpha_k x_k\right\|^2 = \sum_{k=1}^{n}|\alpha_k|^2\|x_k\|^2\ . \tag{1.2}$$

Thus the vectors of an orthogonal set are linearly independent. Relation (1.2) is well known in Euclidean geometry. In Hilbert spaces, we can extend this formula to infinite sums.

Theorem 1.2.1. *Let $(e_i)_{i\geq 1}$ be an orthonormal sequence of an Hilbert space $\mathcal{H}$ and let $(\alpha_i)_{i\geq 1}$ be a sequence of complex numbers. The series*

$$\sum_{i=1}^{\infty}\alpha_i e_i \tag{1.3}$$

converges in $\mathcal{H}$ if and only if $\sum_i |\alpha_i|^2 < \infty$, in which case

$$\left\| \sum_{i=1}^{\infty} \alpha_i e_i \right\|^2 = \sum_{i=1}^{\infty} |\alpha_i|^2 \, . \tag{1.4}$$

Proof. For all $m > k > 0$, as in (1.2), we have

$$\left\| \sum_{i=k}^{m} \alpha_i e_i \right\|^2 = \sum_{i=k}^{m} |\alpha_i|^2 \, .$$

since $\sum_{i=1}^{\infty} |\alpha_i|^2 < \infty$, the sequence $s_m = \sum_{i=1}^{m} \alpha_i e_i$ is Cauchy in $\mathcal{H}$. Since $\mathcal{H}$ is complete, it converges. Relation (1.4) is obtained by taking the limit.

Conversely, if $\sum_{i=1}^{\infty} \alpha_i e_i$ is convergent series, then (1.4) again holds, which implies the converse result. □

An Orthonormal series allows us to approximate any $x \in \mathcal{H}$ by a finite partial sum of the infinite sum (1.3).

Proposition 1.2.2. *Let x be a vector of the Hilbert space $\mathcal{H}$ and $E = \{e_1, \cdots, e_n\}$ a finite orthonormal set of vectors, then*

$$\left\| x - \sum_{k=1}^{n} \langle x, e_k \rangle \, e_k \right\|^2 = \|x\|^2 - \sum_{k=1}^{n} |\, \langle x, e_k \rangle \,|^2 \, . \tag{1.5}$$

In addition, $\sum_{k=1}^{n} \langle x, e_k \rangle \, e_k$ is the vector of $\mathrm{Span}\,(e_1, \ldots, e_n)$ *that is the closest of x. Hence the left-hand side of (1.5) equals*

$$\inf \left\{ \|x - y\|^2 \ : \ y \in \mathrm{Span}\,(e_1, \ldots, e_n) \right\} \, .$$

Proof. We have for all $j = 1, \ldots, n$,

$$\left\langle x - \sum_{k=1}^{n} \langle x, e_k \rangle \, e_k, e_j \right\rangle = \langle x, e_j \rangle - \langle x, e_j \rangle = 0 \, .$$

Hence, we may write

$$x = \left(x - \sum_{k=1}^{n} \langle x, e_k \rangle \, e_k \right) + \sum_{i=1}^{n} \langle x, e_k \rangle \, e_k \, ,$$

which is the sum of two orthogonal vectors. By Pythagore's Identity and (1.2) with $x_k = e_k$ and $\alpha_k = \langle x, e_k \rangle$, we get (1.5).

Similarly,, for all $(\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$,

$$\left\| x - \sum_{k=1}^{n} \alpha_k e_k \right\|^2 = \left\| x - \sum_{k=1}^{n} \langle x, e_k \rangle \, e_k \right\|^2 + \sum_{k=1}^{n} |\, \langle x, e_k \rangle - \alpha_k|^2 \, ,$$

and thus $\sum_{k=1}^{n} \langle x, e_k \rangle \, e_k$ achieves the best approximation of x by a linear combinations of $e_1, \ldots, e_n$. □

Example 1.2.1 (Gram-Schmidt algorithm)**.** *Let $(y_i)_{i\geq 1}$ be a sequence in a Hilbert space $\mathcal{H}$. The Gram-Schmidt algorithm is an iterative algorithm to construct an orthogonal sequence such that* $\mathrm{Span}\,(e_1,\ldots,e_n) = \mathrm{Span}\,(y_1,\ldots,y_n)$ *for all $n \geq 1$.*

Algorithm 1: Gram-Schmidt algorithm.

Data: A set of vectors $y_1,\ldots,y_n$
Result: An orthogonal sequence $e_1,\ldots,e_n$
Initialization: set $e_1 = y_1$.
for $t = 2,\ldots,n$ **do**
  Define
$$e_t = y_t - \sum_{k=1}^{t-1} \frac{\langle y_t, e_k\rangle}{\|e_k\|^2} e_k \ ,$$
  with the convention $0/0 = 0$.
end

Proposition 1.2.2 also yields the following result.

Corollary 1.2.3 (Bessel Inequality)**.** *Let $(e_i)_{i\geq 1}$ be an orthonormal sequence of a Hilbert space $\mathcal{H}$. Then, for all $x \in \mathcal{H}$,*

$$\sum_{i=1}^{\infty} |\,\langle x, e_i\rangle\,|^2 \leq \|x\|^2 \ .$$

The Bessel inequality implies that for all $x \in \mathcal{H}$, $\lim_{n\to\infty} \langle x, e_n\rangle = 0$ and also that $(\langle x, e_i\rangle)_{i\geq 1}$ is in ℓ^2. By Theorem 1.2.1, we get that

$$\sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i \tag{1.6}$$

is a convergent series. It is called the *Fourier expansion* of x; the coefficients $\langle x, e_i\rangle$ are called the *Fourier coefficients* with respect to the orthonormal sequence (e_i). Note however that, although $\sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i$ always converges, its limit is not always equal to x.

Example 1.2.2. *Let $\mathbb{T}$ denote the quotient space $\mathbb{R}/(2\pi\mathbb{Z})$ (or any interval congruent to $[0, 2\pi)$). Consider $\mathcal{H} = L^2(\mathbb{T})$ and define $e_n(t) = \pi^{-1/2}\sin(nt)$ pour $n = 1, 2, \cdots$. The sequence (e_n) is orthonormal in $\mathcal{H}$, but for $x(t) = \cos(t)$, we have*

$$\sum_{n=1}^{\infty} \langle x, e_n\rangle\, e_n(t) = \sum_{n=1}^{\infty} \left[\pi^{-1/2} \int_{\mathbb{T}} \cos(t)\sin(nt)\mathrm{d}t\right] \pi^{-1/2}\sin(nt)$$
$$= \sum_{n=}^{\infty} 0 \cdot \sin(nt) = 0 \neq \cos t \ .$$

In fact the limit is x, if an additional property is assumed.

Definition 1.2.2 (Dense sets, Hilbert Bases)**.** *A subset E of a Hilbert space $\mathcal{H}$ is said* dense *if $\overline{\mathrm{Span}}\,(E) = \mathcal{H}$. An* orthonormal dense *sequence is called a* Hilbert basis.

Let us give an example of a dense set for measured spaces.

Proposition 1.2.4. *Consider the measured space $(\Omega, \mathcal{F}, \mu)$ and the Hilbert space $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mu)$.*

$$\overline{\mathrm{Span}}\,(\mathbb{1}_A, A \in \mathcal{F}) = L^2(\Omega, \mathcal{F}, \mu)\ ,$$

Proof. For any nonnegative square integrable function f defined on $(\Omega, \mathcal{F}, \mu)$, denote

$$f_n = \sum_{k=0}^{n2^n} k2^{-n}\mathbb{1}_{f^{-1}([k2^{-n},(k+1)2^{-n}))} \in \overline{\mathrm{Span}}\,(\mathbb{1}_A, A \in \mathcal{F})\ .$$

Since $0 \le f_n \le f$, by dominated convergence, we get that $\int |f_n - f|^2 \mathrm{d}\mu \to 0$. Since any $g \in L^2(\Omega, \mathcal{F}, \mu)$ is a linear combination of at most 4 nonnegative functions (the positive and negative part of the real and complex parts), we get the result. □

A Hilbert basis allows us to "reach" any point in $\mathcal{H}$.

Theorem 1.2.5. *Let $(e_i)_{i\ge1}$ be a Hilbert basis of the Hilbert space $\mathcal{H}$. Then for all $x \in \mathcal{H}$,*

$$x = \sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i\ . \tag{1.7}$$

Proof. We already known the series in (1.6) converges. On the other hand, since (e_i) is dense, there exists $(\alpha_{p,n})_{1\le i\le n}$ such that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \alpha_{i,n} e_i = x\ .$$

Now from Proposition 1.2.2, we have

$$\left\| x - \sum_{i=1}^{n} \langle x, e_i\rangle\, e_i \right\| \le \left\| x - \sum_{i=1}^{n} \alpha_{i,n} e_i \right\|\ .$$

Hence the result. □

Theorem 1.2.5 implies that an orthonormal sequence (e_i) is a Hilbert basis if and only if Relation (1.7) holds for all $x \in \mathcal{H}$. The proof of the following result is left as an exercise.

Theorem 1.2.6. *Let $(e_i)_{i\ge1}$ be an orthonormal sequence of the Hilbert space $\mathcal{H}$. The following assertions are equivalent.*

(i) $(e_i)_{i\ge1}$ *is an Hilbert basis.*

(ii) If some $x \in \mathcal{H}$ satisfies

$$\langle x, e_i\rangle = 0 \quad \textit{for all} \quad i \ge 1\ ,$$

then $x = 0$.

(iii) For all $x \in \mathcal{H}$,

$$\|x\|^2 = \sum_{i=1}^{\infty} |\langle x, e_i\rangle|^2\ . \tag{1.8}$$

Example 1.2.3 (Fourier basis)**.** *Define*

$$e_n(x) = (2\pi)^{-1/2}\mathrm{e}^{\mathrm{i}nx}\ , n \in \mathbb{Z}\ .$$

Then (e_n) *is an Hilbert basis of* $L^2(\mathbb{T})$*, see e.g. [8].*

A Hilbert space is called separable if it contains a countable dense subset.

Theorem 1.2.7. *A Hilbert space* $\mathcal{H}$ *is separable if and only if it admits a Hilbert basis.*

Proof. Let (e_i) be a Hilbert basis of $\mathcal{H}$. The set $S = \bigcup_{n=1}^{\infty} S_n$, where, for $n \in \mathbb{N}$,

$$S_n \stackrel{\text{def}}{=} \left\{ \sum_{l=1}^{n} (\alpha_k + \mathrm{i}\beta_k)e_k\ , (\alpha_k, \beta_k) \in \mathbb{Q} \times \mathbb{Q}, k = 1, \ldots, n \right\}$$

is countable. Since for $x \in \mathcal{H}$,

$$\lim_{n\to\infty} \left\| \sum_{k=1}^{n} \langle x, e_k \rangle\, e_k - x \right\| = 0\ ,$$

the set S is dense in $\mathcal{H}$.

If $\mathcal{H}$ is separable then there exists a dense sequence $(y_i)_{i\geq 1}$. The Gram-Schmidt algorithm of Example 1.2.1 provides an orthogonal sequence $(e_i)_{i\geq 1}$ such that $\mathrm{Span}\,(e_1, \ldots, e_n) = \mathrm{Span}\,(y_1, \ldots, y_n)$ for all n. Removing the null vectors in this sequence and normalizing the others by the square root of their norms, we get a Hilbert basis. □

1.3 Fourier series

Define the sequence of complex exponential functions

$$\phi_n(x) = (2\pi)^{-1/2}\mathrm{e}^{\mathrm{i}nx}, \quad n \in \mathbb{Z}\ . \tag{1.9}$$

We shall see that (ϕ_n) is a dense set of $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ for any finite measure μ on the Borel sets of $\mathbb{T}$. If moreover μ est the Lebesgue mesure, it is a Hilbert basis.

Let $L^1(\mathbb{T})$ denote the set of 2π-periodic locally integrable (with respect to the Lebesgue measure) functions. For $f \in L^1(\mathbb{T})$, set

$$f_n = \sum_{k=-n}^{n} \left(\int_{\mathbb{T}} f\bar{\phi}_k \right) \phi_k\ , \quad n = 0, 1, 2, \ldots$$

Then

$$f_n(x) = \sum_{k=-n}^{n} \frac{1}{2\pi} \int_{\mathbb{T}} f(t)\mathrm{e}^{\mathrm{i}k(x-t)}\mathrm{d}t\ . \tag{1.10}$$

The following result can be found in [8].

Theorem 1.3.1. *Suppose that* f *is a continuous* 2π*-periodic function. Then the Cesaro sequence*

$$\left(\frac{1}{n} \sum_{k=0}^{n-1} f_k \right)_{n\in\mathbb{N}^*}$$

converges uniformly to f*.*

An interesting consequence for us is the following result (see Exercise 1.2).

Corollary 1.3.2. *Let μ be a finite measure on the Borel sets of $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$. The sequence $(\phi_n)_{n\in\mathbb{Z}}$ defined in (1.9) is linearly dense in the Hilbert space $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$, that is, $\overline{\mathrm{Span}}\,(\phi_n,\, n \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$.*

In the case of the Lebesgue measure, we get the following.

Corollary 1.3.3. *The sequence $(\phi_n)_{n\in\mathbb{Z}}$ defined in (1.9) is a Hilbert basis in $L^2(\mathbb{T})$. In particular, for all $f \in L^2(\mathbb{T})$,*

$$f = \sum_{k=-\infty}^{\infty} \alpha_k \phi_k \quad \textit{with} \quad \alpha_k = (2\pi)^{-1/2} \int_{\mathbb{T}} f(x) \mathrm{e}^{-\mathrm{i}kx} \mathrm{d}x \,,$$

where the infinite sum converges in $L^2(\mathbb{T})$. The Parseval identity then reads

$$\int_{\mathbb{T}} |f(x)|^2 \mathrm{d}x = \sum_{k=-\infty}^{\infty} |\alpha_k|^2 \,.$$

1.4 Projection and orthogonality principle

The following theorem allows us to define the orthogonal projection onto a closed subspace of a Hilbert space.

Theorem 1.4.1 (Projection theorem)**.** *Let $\mathcal{E}$ be a closed convex subset of a Hilbert space $\mathcal{H}$ and let $x \in \mathcal{H}$. Then the following assertions hold.*

(i) There exists a unique vector $\mathrm{proj}\,(x|\,\mathcal{E}) \in \mathcal{E}$ *such that*

$$\|x - \mathrm{proj}\,(x|\,\mathcal{E})\,\| = \inf_{w\in\mathcal{E}} \|x - w\|$$

(ii) If moreover $\mathcal{E}$ is a linear subspace, $\mathrm{proj}\,(x|\,\mathcal{E})$ *is the unique $\hat{x} \in \mathcal{E}$ such that $x - \hat{x} \in \mathcal{E}^\perp$.*

We call $\mathrm{proj}\,(x|\,\mathcal{E})$ *the orthogonal projection of x onto $\mathcal{E}$.*

Proof. Let $x \in \mathcal{H}$. Set $h = \inf_{w\in\mathcal{E}} \|x - w\| \geq 0$. Let $w_1, w_2, \cdots$, be in $\mathcal{E}$ such that:

$$\lim_{m\to+\infty} \|x - w_m\|^2 = h^2 \geq 0 \tag{1.11}$$

The parallelogram identity $\|a-b\|^2+\|a+b\|^2 = 2\|a\|^2+2\|b\|^2$ with $a = w_m - x$ and $b = w_n - x$ gives that

$$\|w_m - w_n\|^2 + \|w_m + w_n - 2x\|^2 = 2\|w_m - x\|^2 + 2\|w_n - x\|^2$$

Since $(w_m + w_n)/2 \in \mathcal{E}$, we have $\|w_m + w_n - 2x\|^2 = 4\|(w_m + w_n)/2 - x\|^2 \geq 4h^2$. By (1.11), for all $\epsilon > 0$,there exists N such that $\forall m, n > N$:

$$\|w_m - w_n\|^2 \leq 2(h^2 + \epsilon) + 2(h^2 + \epsilon) - 4h^2 = 4\epsilon \,,$$

which shows that $\{w_n, n \in \mathbb{N}\}$ is a Cauchy sequence and thus converges to some limit y in $\mathcal{E}$, since $\mathcal{E}$ is closed. By continuity of the norm, we have $\|y - x\| = h$.

It remains to show the uniqueness. Let $z \in \mathcal{E}$ such that $\|x-z\|^2 = \|x-y\|^2 = h^2$. Then the parallelogram identity implies

$$0 \leq \|y-z\|^2 = -4\|(y+z)/2 - x\|^2 + 2\|x-y\|^2 + 2\|x-z\|^2 \\ \leq -4h^2 + 2h^2 + 2h^2 = 0 \,,$$

where we used that $(y+z)/2 \in \mathcal{E}$ with the convexity assumption and thus $\|(y+z)/2 - x\|^2 \geq h^2$. We get $y = z$, which conclude the proof of this assertion.

We now prove the second assertion. Let $\hat{x}$ be the orthogonal projection of x onto $\mathcal{E}$. If there exists $u \in \mathcal{E}$ such that $x - u \perp \mathcal{E}$, we have

$$\begin{aligned} \|x-\hat{x}\|^2 &= \langle x-u+u-\hat{x}, x-u+u-\hat{x}\rangle \\ &= \|x-u\|^2 + \|u-\hat{x}\|^2 + 2\mathrm{Re}(\langle u-\hat{x}, x-u\rangle) \\ &= \|x-u\|^2 + \|u-\hat{x}\|^2 + 0 \geq \|x-u\|^2 \,, \end{aligned}$$

and thus $u = \hat{x}$ by the previous assertion.

Conversely suppose that $x - \hat{x} \not\perp \mathcal{E}$ and let us find a contradiction. Then there exists $y \in \mathcal{E}$ such that $\|y\| = 1$ and $c = \langle x-\hat{x}, y\rangle \neq 0$. Set $\tilde{x} = \hat{x} + cy \in \mathcal{E}$. We have

$$\begin{aligned} \|x-\tilde{x}\|^2 &= \langle x-\hat{x}+\hat{x}-\tilde{x}, x-\hat{x}+\hat{x}-\tilde{x}\rangle \\ &= \|x-\hat{x}\|^2 + \|\hat{x}-\tilde{x}\|^2 + 2\mathrm{Re}(\langle \hat{x}-\tilde{x}, x-\hat{x}\rangle) \\ &= \|x-\hat{x}\|^2 - |c|^2 < \|x-\hat{x}\|^2 \,. \end{aligned}$$

Thus we get a contradiction with the definition of $\hat{x}$. □

Assertion (ii) provides a quite practical way to determine the projection, since it replaces a minimization problem by a system of linear equations to solve.

Example 1.4.1 (Projection onto a one dimension space)**.** *Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{C} = \mathrm{Span}\,(v)$ with $v \in \mathcal{H}$. For any $x \in \mathcal{H}$, we have $\mathrm{proj}\,(x|\,\mathcal{C}) = \alpha v$ with $\alpha = \langle x, v\rangle / \|v\|^2$. Denoting $\epsilon = x - \mathrm{proj}\,(x|\,\mathcal{C})$, we get*

$$\|\epsilon\|^2 = \|x\|^2\left(1 - \|\rho\|^2\right) \quad \textit{where} \quad \rho = \frac{\langle x, v\rangle}{\|x\|\|v\|} \quad \textit{with} \quad |\rho| \leq 1$$

The projection operator defined by Theorem 1.4.1 has the following interesting properties, whose proofs are left to the reader as an exercise.

Proposition 1.4.2. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{E}$ a closed subspace of $\mathcal{H}$. Then the following assertions hold.*

(i) Suppose that $\mathcal{E} = \overline{\mathrm{Span}}\,(e_k, k \in \mathbb{N})$ with (e_k) being an orthonormal sequence. Then

$$\mathrm{proj}\,(h|\,\mathcal{E}) = \sum_{k=0}^{\infty} \langle h, e_k\rangle\, e_k \,.$$

(ii) The function $\mathrm{proj}\,(\cdot|\,\mathcal{E}) : \mathcal{H} \to \mathcal{H}$, $x \mapsto \mathrm{proj}\,(x|\,\mathcal{E})$ is linear and continuous on $\mathcal{H}$.

(iii) $\|x\|^2 = \|\,\mathrm{proj}\,(x|\,\mathcal{E})\,\|^2 + \|x - \mathrm{proj}\,(x|\,\mathcal{E})\,\|^2$,

(iv) $x \in \mathcal{E}$ *if and only if* $\mathrm{proj}\,(x|\,\mathcal{E}) = x$.

(v) $x \in \mathcal{E}^{\perp}$ *if and only if* $\mathrm{proj}\,(x|\,\mathcal{E}) = 0$.

(vi) Let $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two closed subspace of* $\mathcal{H}$*, such that* $\mathcal{E}_1 \subset \mathcal{E}_2$*. Then*

$$\forall x \in \mathcal{H}, \qquad \mathrm{proj}\,(\mathrm{proj}\,(x|\,\mathcal{E}_2)|\,\mathcal{E}_1) = \mathrm{proj}\,(x|\,\mathcal{E}_1)\ .$$

(vii) Let $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two closed subspace of* $\mathcal{H}$*, such that* $\mathcal{E}_1 \perp \mathcal{E}_2$*. Then*

$$\forall x \in \mathcal{H}, \qquad \mathrm{proj}\left(x|\,\mathcal{E}_1 \overset{\perp}{\oplus} \mathcal{E}_2\right) = \mathrm{proj}\,(x|\,\mathcal{E}_1) + \mathrm{proj}\,(x|\,\mathcal{E}_2)\ .$$

The following result will be useful.

Theorem 1.4.3. *Let* $(\mathcal{M}_n)_{n\in\mathbb{Z}}$ *be an increasing sequence of closed subspaces of an Hilbert space* $\mathcal{H}$*.*

(i) Denote $\mathcal{M}_{-\infty} = \bigcap_n \mathcal{M}_n$*. Then for all* $h \in \mathcal{H}$*, we have*

$$\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty}) = \lim_{n\to-\infty} \mathrm{proj}\,(h|\,\mathcal{M}_n)$$

(ii) Let $\mathcal{M}_\infty = \overline{\bigcup_{n\in\mathbb{Z}} \mathcal{M}_n}$*. Then, for all* $h \in \mathcal{H}$*,*

$$\mathrm{proj}\,(h|\,\mathcal{M}_\infty) = \lim_{n\to\infty} \mathrm{proj}\,(h|\,\mathcal{M}_n)\ .$$

Proof. We first note that (ii) can be deduced from (i). Indeed, we have

$$\mathcal{M}_\infty^{\perp} = \bigcap_n \mathcal{M}_n^{\perp}\ ,$$

and thus, since $\mathcal{M}_\infty$ and the $\mathcal{M}_n$'s are closed, Assertion (vii) of Proposition 1.4.2 yields $\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty}) = h - \mathrm{proj}\left(h|\,\mathcal{M}_\infty^{\perp}\right)$ and the same holds for $\mathcal{M}_n$. Now, since $\mathcal{M}_n^{\perp}$ are closed by Theorem 1.1.4, we can apply (i).

It remains to show (i). Since $\mathcal{M}_n$ is a closed subspace of $\mathcal{H}$, $\mathcal{M}_{-\infty}$ is a closed subspace of $\mathcal{H}$. The projection theorem, Theorem 1.4.1, shows that $\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty})$ exists. For $m < n$, define $\mathcal{M}_n \ominus \mathcal{M}_m$ as the orthogonal complement of $\mathcal{M}_m$ in $\mathcal{M}_n$, that is $\mathcal{M}_m^{\perp} \cap \mathcal{M}_n$.This is a closed subset of $\mathcal{H}$ by Theorem 1.1.4. Using Assertion (vii) of Proposition 1.4.2,

$$\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_m) = \mathrm{proj}\,(h|\,\mathcal{M}_n) - \mathrm{proj}\,(h|\,\mathcal{M}_m)\ .$$

It follows that, for all $m \geq 1$,

$$\sum_{n=-m+1}^{0} \|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2 = \|\,\mathrm{proj}\,(h|\,\mathcal{M}_0 \ominus \mathcal{M}_{-m})\,\|^2 \leq \|h\|^2 < \infty\ .$$

We obtain that the series $(\|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2)_{n\leq 0}$ is convergent and since for all $m \leq p \leq 0$,

$$\|\,\mathrm{proj}\,(h|\,\mathcal{M}_p) - \mathrm{proj}\,(h|\,\mathcal{M}_n)\,\|^2 = \sum_{n=-m+1}^{p} \|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2\ ,$$

the sequence $\{\mathrm{proj}\,(h|\,\mathcal{M}_n)\,, n = 0, -1, -2, \dots\}$ is a Cauchy sequence. Since $\mathcal{H}$ is complete, $\mathrm{proj}\,(h|\,\mathcal{M}_n)$ converges in $\mathcal{H}$, say to z. We have to show that $z = \mathrm{proj}\,(h|\,\mathcal{M}_{-\infty})$. By the projection theorem, this is equivalent to $z \in \mathcal{M}_{-\infty}$ and $h - z \perp \mathcal{M}_{-\infty}$. Since $\mathrm{proj}\,(h|\,\mathcal{M}_n) \in \mathcal{M}_p$ for all $n \leq p$, we have $z \in \mathcal{M}_p$ for all p and thus $z \in \mathcal{M}_{-\infty}$. Take now $p \in \mathcal{M}_{-\infty}$. Then $p \in \mathcal{M}_n$ for all $n \in \mathbb{Z}$, and, for all $n \in \mathbb{Z}$, $\langle h - \mathrm{proj}\,(h|\,\mathcal{M}_n)\,, p\rangle = 0$ and $\langle h - z, p\rangle = 0$ by taking the limit, which achieves the proof. □

1.5 Riesz representation theorem

We start with some simple results on the orthogonal set.

Proposition 1.5.1. *Let $\mathcal{E}$ and $\mathcal{F}$ be two subspaces of a Hilbert space $\mathcal{H}$. If $\mathcal{E} \overset{\perp}{\oplus} \mathcal{F} = \mathcal{H}$, then $\mathcal{F} = \mathcal{E}^{\perp}$.*

Proof. Any $x \in \mathcal{H}$ can be written as $x = y + z$ with $y \in \mathcal{E}$ and $z \in \mathcal{F} \subseteq \mathcal{E}^{\perp}$. Hence $x \in \mathcal{E}^{\perp}$ if and only if $y \in \mathcal{E}^{\perp}$, and so $y = 0$, and thus $x = z \in \mathcal{F}$. □

However, one needs an additional assumption on $\mathcal{E}$ in order to have that $\mathcal{E} \overset{\perp}{\oplus} \mathcal{F} = \mathcal{H}$ and $\mathcal{F} = \mathcal{E}^{\perp}$ are two equivalent assertions.

Theorem 1.5.2. *If $\mathcal{E}$ is a closed subspace of a Hilbert space $\mathcal{H}$, then $\mathcal{E} \overset{\perp}{\oplus} \mathcal{E}^{\perp} = \mathcal{H}$. Moreover $(\mathcal{E}^{\perp})^{\perp} = \mathcal{E}$.*

Proof. Let $x \in \mathcal{H}$ and set $y = \mathrm{proj}\,(x|\,\mathcal{E})$. Then $z = x - y \in \mathcal{E}^{\perp}$ by characterization of the orthogonal projection. Hence $x = y + z$ with $y \in \mathcal{E}$ and $z \in \mathcal{E}^{\perp}$, which shows the first assertion of the theorem.

The second assertion is a consequence of the first one and of Proposition 1.5.1. □

We can now state the main result of this section.

Theorem 1.5.3 (Riesz representation theorem)**.** *Let $\mathcal{H}$ be a Hilbert space. Then $F : \mathcal{H} \to \mathbb{C}$ is a non-zero continuous linear form if and only if there exists $x \in \mathcal{H} \backslash \{0\}$ such that $F(y) = \langle y, x\rangle$ for all $y \in \mathcal{H}$.*

Proof. Let $x \in \mathcal{H} \setminus \{0\}$ and $F : \mathcal{H} \to \mathbb{C}$ defined by $F(y) = \langle y, x\rangle$ for all $y \in \mathcal{H}$. Then F is a continuous linear form by linearity of the scalar product with respect to the first argument and by the Cauchy-Schwarz inequality. Moreover F is non-zero since $F(x) > 0$.

Let us now show the direct implication. Let $F : \mathcal{H} \to \mathbb{C}$ be a non-zero continuous linear form. Denote by $\mathcal{E}$ the null space of F. Then $\mathcal{E}$ is a closed subspace of $\mathcal{H}$. By Theorem 1.5.2, $\mathcal{E}^{\perp}$ is a supplementary set of $\mathcal{E}$ in $\mathcal{H}$. Since $\mathcal{E}$ has codimension 1, we conclude that $\mathcal{E}^{\perp}$ has dimension 1. Let $z \in \mathcal{E}^{\perp}$ such that $\|z\| = 1$, hence $\mathcal{E}^{\perp} = \mathrm{Span}\,(z)$. Then for all $y \in \mathcal{H}$, we have $\mathrm{proj}\,(y|\,\mathrm{Span}\,(z)) = \langle y, z\rangle\, z$ and, since $y - \mathrm{proj}\,(y|\,\mathrm{Span}\,(z)) \in \mathcal{E} = (\mathcal{E}^{\perp})^{\perp}$, we get $F(y) = \langle y, z\rangle\, F(z)$. We conclude the proof by setting $x = \overline{F(z)}z$. □

1.6 Unitary operators

Definition 1.6.1 (Unitary operators)**.** *Let $\mathcal{H}$ and $\mathcal{I}$ be two Hilbert spaces. An* isometric operator *S from $\mathcal{H}$ to $\mathcal{I}$ is a linear application $S : \mathcal{H} \to \mathcal{I}$ such that $\langle Sv, Sw\rangle_{\mathcal{I}} = \langle v, w\rangle_{\mathcal{H}}$ for*

all $(v,w) \in \mathcal{H}$. *If it is moreover bijective, we say that it is a* unitary operator. *In this case we also say that* $\mathcal{H}$ *and* $\mathcal{I}$ *are isomorphic.*

Observe that an isometric operator is always continuous.

Theorem 1.6.1. *Let* $\mathcal{H}$ *be a separable Hilbert space.*

(i) If $\mathcal{H}$ *has infinite dimension, it is isomorphic to* ℓ^2.

(ii) If $\mathcal{H}$ *has dimension* n, *it is isomorphic to* $\mathbb{C}^n$.

Proof. It is a direct application of Theorem 1.2.7 and Theorem 1.2.1. □

The following result is very convenient to construct isometric operators.

Theorem 1.6.2. *Let* $(\mathcal{H}, \langle,\rangle_{\mathcal{H}})$ *and* $(\mathcal{I}, \langle,\rangle_{\mathcal{I}})$ *be two Hilbert spaces. Let* $\mathcal{G}$ *be a subspace of* $\mathcal{H}$.

(i) Let $S : \mathcal{G} \to \mathcal{I}$ *be isometric on* $\mathcal{G}$. *Then* S *admits a unique isometric extension* $\bar{S} : \overline{\mathcal{G}} \to \mathcal{I}$ *and* $\bar{S}(\overline{\mathcal{G}})$ *is the closure of* $S(\mathcal{G})$ *in* $\mathcal{I}$.

(ii) Let $(v_t, t \in \mathrm{T})$ *and* $(w_t, t \in \mathrm{T})$ *be two sets of vectors in* $\mathcal{H}$ *and* $\mathcal{I}$ *indexed by an arbitrary index set* T. *Suppose that for all* $(s,t) \in \mathrm{T} \times \mathrm{T}$, $\langle v_t, v_s\rangle_{\mathcal{H}} = \langle w_t, w_s\rangle_{\mathcal{I}}$. *Then, there exists a unique isometric operator* $S : \overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T}) \to \overline{\mathrm{Span}}\,(w_t, t \in \mathrm{T})$ *such that for all* $t \in \mathrm{T}$, $Sv_t = w_t$. *Moreover,* $S\left(\overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T})\right) = \overline{\mathrm{Span}}\,(w_t, t \in \mathrm{T})$.

One often uses the same notation for S and its extension $\bar{S}$.

Proof. We first show Assertion (i). Let $v \in \bar{\mathcal{G}}$. For all sequence $(v_n) \subset \mathcal{G}$ converging to v, the sequence (Sv_n) is a Cauchy sequence in $\mathcal{I}$ (since (v_n) is Cauchy in $\mathcal{G}$ and S is isometric). Thus there exists $w \in \mathcal{I}$ such that $w = \lim_{n\to\infty} Sv_n$. If (v'_n) is also converging to v, we have $\|v'_n - v'_n\|_{\mathcal{H}} \to 0$ and thus $\|Sv_n - Sv'_n\|_{\mathcal{I}} \to 0$, which shows that w only depends on v. Set $\bar{S}v = w$. Linearity and isometric properties are preserved by taking the limit and $\bar{S} : \bar{\mathcal{G}} \to \mathcal{I}$ is thus an isometric extension of S. The uniqueness of this extension is obvious.

By definition $\bar{S}(\bar{\mathcal{G}})$ is included in the closure of $S(\mathcal{G})$. Conversely, let $w \in \overline{S(\mathcal{G})}$. there exists a sequence $(v_n) \in \mathcal{G}$ such that $w = \lim_{n\to\infty} Sv_n$. The sequence (Sv_n) is Cauchy and thus so is (v_n) in $\mathcal{G}$. Let $v \in \bar{\mathcal{G}}$ its limit. We have $\bar{S}v = \lim_{n\to\infty} Sv_n$ and thus $\bar{S}v = w$, which shows that $\overline{S(\mathcal{G})} \subseteq \bar{S}(\bar{\mathcal{G}})$. The first assertion is proved.

We next show the second assertion. For all finite subset J of T and all complex numbers $(a_t)_{t\in J}$ and $(b_t)_{t\in J}$, we have

$$\sum_{t\in J} a_t v_t = \sum_{t\in J} b_t v_t \Rightarrow \sum_{t\in J} a_t w_t = \sum_{t\in J} b_t w_t$$

since by setting $c_t = a_t - b_t$,

$$\left\|\sum_{t\in J} c_t v_t\right\|^2_{\mathcal{H}} = \sum_{t\in J}\sum_{t'\in J} c_t\overline{c_{t'}}\left\langle v_t, v'_t\right\rangle_{\mathcal{H}} = \sum_{t\in J}\sum_{t'\in J} c_t\overline{c_{t'}}\left\langle w_t, w'_t\right\rangle_{\mathcal{I}} = \left\|\sum_{t\in I} c_t w_t\right\|_{\mathcal{I}},$$

using the linearity and isometric properties. This allows us to define $Sf = \sum_{t\in I} a_t w_t$ for all f such that $f = \sum_{t\in I} a_t v_t$ with I finite subset of T. We just defined S on $\mathcal{G} = \mathrm{Span}\,(v_t, t \in \mathrm{T})$ and it is an isometric operator. Applying (i), it admits a unique isometric extension $\bar{S} : \bar{\mathcal{G}} \to \mathcal{I}$ such that $\bar{S}(\bar{\mathcal{G}}) = \overline{S(\mathcal{G})}$. By definition, $\bar{\mathcal{G}} = \overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T})$ and $S(\mathcal{G}) = \mathrm{Span}\,(w_t, t \in \mathrm{T})$. □

1.7 Exercises

Exercise 1.1. Let X and Y be two complex valued random variables in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, for some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Determine the constant $m = \mathrm{proj}\,(X | \,\mathrm{Span}\,(1))$.
2. Determine the random variable $Z = \mathrm{proj}\,(X | \,\mathrm{Span}\,(1, Y))$.

Exercise 1.2 (The Fourier basis is dense). The first questions of this exercise are dedicated to the proof of Theorem 1.3.1. Let f be a continuous 2π-periodic function and f_n be defined as in (1.10).

1. Determine the Fejér kernel J_n, which satisfies
$$\frac{1}{n}\sum_{k=0}^{n-1} f_k = \int_{\mathbb{T}} J_n(x-t)\, f(t)\, \mathrm{d}t\ .$$
2. Show that we can write, for all $t \in \mathbb{R}$,
$$J_n(t) = \frac{1}{2\pi}\sum_{k=-n+1}^{n-1} (1-|k|/n)\mathrm{e}^{\mathrm{i}kt} = \frac{1}{2\pi n}\left|\sum_{j=0}^{n-1} \mathrm{e}^{\mathrm{i}jt}\right|^2\ .$$
3. Deduce that $J_n \geq 0$, $\int_{\mathbb{T}} J_n = 1$ and that for any $\epsilon \in (0, \pi]$,
$$\sup_{n\geq 1} n \sup_{\epsilon \leq |t| \leq \pi} J_n(t) < \infty\ .$$
4. Conclude the proof of Theorem 1.3.1.

Let now μ be a finite measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. Let F be a closed set in $\mathbb{T}$. Define $f_n(x) = (1 - n\, d(F, x))_+$, where $d(F, x) = \inf\{|y - x| \ :\ y \in F\}$.

5. Show that $f_n \to \mathbb{1}_F$ in $L^1(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$.

By Proposition A.1.3, we know that μ is regular that is, for all $A \in \mathcal{B}(\mathbb{T})$,
$$\mu(A) = \inf\{\mu(U) \ :\ U \text{ open set } \supset A\} = \sup\{\mu(F) \ :\ F \text{ closed set} \subset A\}\ .$$

6. Deduce that for all $A \in \mathcal{B}(\mathbb{T})$ and all $\epsilon > 0$, there exists a continuous 2π-periodic function g_ϵ such that
$$\int |\mathbb{1}_A - g_\epsilon| \mathrm{d}\mu \leq \epsilon\ .$$
7. Deduce that the set of continuous 2π-periodic functions is dense in $L^1(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ endowed with the L^1 norm.
8. Deduce that the set of continuous 2π-periodic functions is also dense in $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ endowed with the L^2 norm.
9. Conclude the proof of Corollary 1.3.2.

Chapter 2

Probability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, a non empty space Ω endowed with a σ-field $\mathcal{F}$ and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. Here we focus on conditional distribution calculus and we do not recall the basic results in probability such as the construction of measures or of the integrals, see for instance the classical reference [6]. However, characterizing a probability measure from its application over a rich enough subclass of sets is repeatedly used in these lecture notes and it is worthwhile to remind the reader of the arguments leading to this conclusion. Recall that a π-system $\mathcal{C}$ on Ω is a non-empty class of subsets of Ω which is stable by finite intersection, for all $A, B \in \mathcal{C}$, $A \cap B \in \mathcal{C}$. Also recall that λ-system $\mathcal{C}$ on Ω is a class of subsets of Ω which contains Ω and is stable by taking the complementary set or a countable union of disjoint sets finite, for all $A \in \mathcal{C}$, $A^c \in \mathcal{C}$, and for all $(A_n)_{n\in\mathbb{N}} \in \mathcal{C}^{\mathbb{N}}$ such that $A_n \cap A_p = \emptyset$ for $n < p$, $\cup_n A_n \in \mathcal{C}$. The following theorem is very useful for extending a result from a π-system to its generating σ-field.

Theorem 2.0.1 ($\pi - \lambda$-theorem)**.** *If $\mathcal{A} \subset \mathcal{C}$ with $\mathcal{A}$ a π-system and $\mathcal{C}$ a λ-system, then $\sigma(\mathcal{A}) \subset \mathcal{C}$.*

The classical characterization theorem for finite measures easily follows from this result. For instance, in the case of a probability measure:

Theorem 2.0.2 (Characterization of probability measures)**.** *Let $\mathcal{C}$ be a π-system on Ω and $\mathcal{F} = \sigma(\mathcal{C})$ be the smallest σ-field containing $\mathcal{C}$. Then a probability measure μ on $(\Omega, \mathcal{F})$ is uniquely characterized by $\mu(A)$ on $A \in \mathcal{C}$.*

2.1 Conditional calculus

2.1.1 Conditional Expectation

Recall that, for $p > 0$ we denote by $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ the space of random variables X such that $\mathbb{E}\left[|X|^p\right] < \infty$ and by $L^p(\Omega, \mathcal{F}, \mathbb{P})$ the one obtain by identifying random variables that are equal $\mathbb{P}$-a.s..

Recall also that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space and observe that for any sub-σ-field $\mathcal{G}$ of $\mathcal{F}$, the space $L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Thus, by Proposition 1.4.2, for any real valued $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, $Y = \text{proj}\left(X | L^2(\Omega, \mathcal{G}, \mathbb{P})\right)$ is well defined and satifies

(i) $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$,

(ii) $\mathbb{E}\left[|X-Y|^2\right] = \inf_{Z\in L^2(\Omega,\mathcal{G},\mathbb{P})} \mathbb{E}\left[|X-Z|^2\right]$.

(iii) For all $Z \in L^2(\Omega,\mathcal{G},\mathbb{P})$, $\mathbb{E}\left[(X-Y)Z\right] = 0$.

Moreover (i) and (ii) are sufficient to characterize Y (up to $\mathbb{P}$-a.s.equality) as well as (i) and (iii). Looking at (iii) we see that it can be written as $\mathbb{E}\left[XZ\right] = \mathbb{E}\left[YZ\right]$ and that this identity continues to make sense if one relaxes the condition $X \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$ into $X \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$ and impose Z to be bounded. In fact, it turns out that taking Z of the form of an indicator function $\mathbb{1}_A$ is sufficient for constructing and defining Y as stated in the following result.

Lemma 2.1.1. *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $X \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Then there exists $Y \in \mathcal{L}^1(\Omega,\mathcal{G},\mathbb{P})$ such that*

$$\mathbb{E}\left[X\mathbb{1}_A\right] = \mathbb{E}\left[Y\mathbb{1}_A\right] \quad \text{for all } A \in \mathcal{G}. \tag{2.1}$$

Moreover the following assertions hold.

(i) If $Y' \in \mathcal{L}^1(\Omega,\mathcal{G},\mathbb{P})$ also satisfies (2.1), then $Y = Y'$ $\mathbb{P}$-a.s.

(ii) If $X \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$, then $Y = \mathrm{proj}\left(X|\, L^2(\Omega,\mathcal{G},\mathbb{P})\right)$.

Proof. We first observe that if $X \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$. Then $Y = \mathrm{proj}\left(X|\, L^2(\Omega,\mathcal{G},\mathbb{P})\right)$ satisfies (2.1). In particular (ii) follows from this fact and (i).

Now let $X \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$ and let $X_n = n \wedge (X \vee (-n))$ so that $X_n \to X$ in L^1 with $X_n \in L^2(\Omega,\mathcal{F},\mathbb{P})$ for all $n \geq 1$. Define $Y_n = \mathrm{proj}\left(X_n|\, L^2(\Omega,\mathcal{G},\mathbb{P})\right)$. Then for all $n \geq 1$ and $A \in \mathcal{G}$, we have

$$\mathbb{E}\left[X_n\mathbb{1}_A\right] = \mathbb{E}\left[Y_n\mathbb{1}_A\right] . \tag{2.2}$$

Moreover for all $p \geq n \geq 1$, we have $\mathrm{proj}\left(X_n - X_p|\, L^2(\Omega,\mathcal{G},\mathbb{P})\right) = Y_n - Y_p$ and so

$$\mathbb{E}\left[|Y_n - Y_p|\right] = \mathbb{E}\left[(Y_n - Y_p)\mathrm{sgn}(Y_n - Y_p)\right] = \mathbb{E}\left[(X_n - X_p)\mathrm{sgn}(Y_n - Y_p)\right] ,$$

since $\mathrm{sgn}(Y_n - Y_p) \in L^2(\Omega,\mathcal{G},\mathbb{P})$. Hence we get that

$$\mathbb{E}\left[|Y_n - Y_p|\right] = \left|\mathbb{E}\left[(X_n - X_p)\mathrm{sgn}(Y_n - Y_p)\right]\right| \leq \mathbb{E}\left[|X_n - X_p|\right] .$$

We conclude that $(Y)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^1(\Omega,\mathcal{G},\mathbb{P})$. Its limit Y then satisfies (2.1) by letting $n \to \infty$ in (2.2). Hence we have proven the main assertion of the lemma.

We now prove (i). Let $Z = Y - Y' \in L^1(\Omega,\mathcal{G},\mathbb{P})$ so that $\mathbb{E}\left[Z\mathbb{1}_A\right] = 0$ for all $A \in \mathcal{G}$. Then taking A successively equal to $\{Z > 0\}$ and $\{Z < 0\}$ we get $Z = 0$ $\mathbb{P}$-a.s.. □

Lemma 2.1.1 allows us to introduce the following definition.

Definition 2.1.1 (Conditional expectation)**.** *Let $X \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The unique $Y \in L^1(\Omega,\mathcal{F},\mathbb{P})$ defined by (2.1) is called the* conditional expectation *of X* given *$\mathcal{G}$, and denoted by $Y = \mathbb{E}\left[X|\,\mathcal{G}\right]$.*

Conditional expectations are defined as elements of L^1, thus they are random variables defined up to $\mathbb{P}$-almost sure equality. Hence, when writing $\mathbb{E}\left[X|\,\mathcal{G}\right] = Y$ for instance, we always mean that this relations holds $\mathbb{P}$-a.s., that is, Y is a version of the conditional expectation.

We now have a series of simple lemmas.

Lemma 2.1.2. *Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The mapping $X \mapsto \mathbb{E}[X|\mathcal{G}]$ is linear continuous from $L^1(\Omega, \mathcal{F}, \mathbb{P})$ to itself. Moreover we have*

$$\mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|] \leq \mathbb{E}[|X|] \ . \tag{2.3}$$

Proof. The linearity of conditional expectation is left as an exercise (see Proposition 2.1.5(a) and Exercise 2.3). Continuity is then a byproduct of (2.3), which we now prove. Let Y be a $\mathcal{G}$-measurable version of $\mathbb{E}[X|\mathcal{G}]$ and $A = \{Y \geq 0\}$. Then $A, A^c \in \mathcal{G}$ and thus

$$\mathbb{E}[|Y|] = \mathbb{E}[\mathbb{1}_A Y] - \mathbb{E}[\mathbb{1}_{A^c} Y] = \mathbb{E}[\mathbb{1}_A X] - \mathbb{E}[\mathbb{1}_{A^c} X] = \mathbb{E}[(\mathbb{1}_A - \mathbb{1}_{A^c})X] \leq \mathbb{E}[|X|] \ ,$$

where we used that $|\mathbb{1}_A - \mathbb{1}_{A^c}| = 1$. Hence we conclude (2.3). □

We now state an intermediary lemma, which will be extended to more general assumptions afterwards.

Lemma 2.1.3. *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and let Y be a version of $\mathbb{E}[X|\mathcal{G}]$. The following assertions hold.*

(i) Equality (2.1) continues to hold with $\mathbb{1}_A$ extended as follows, we have

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all } Z \in \mathcal{L}^\infty(\Omega, \mathcal{G}, \mathbb{P}).$$

(ii) For all $Z \in \mathcal{L}^\infty(\Omega, \mathcal{G}, \mathbb{P})$, we have

$$\mathbb{E}[XZ|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}] \ .$$

Proof. By linearity of the expectation, (2.1) continues to holds when replacing $\mathbb{1}_A$ by any simple $\mathcal{G}$-measurable random variable Z. Since $X, Y \in L^1$, we also get by dominated convergence that (2.1) continues to holds when replacing $\mathbb{1}_A$ by any bounded $\mathcal{G}$-measurable random variable Z, since we can find Z_n converging pointwise to Z with $|Z_n| \leq |Z|$ for all n. Hence we obtain (i).

Take any $A \in \mathcal{G}$. Then $Z\mathbb{1}_A \in \mathcal{L}^\infty(\Omega, \mathcal{G}, \mathbb{P})$ and by applying (i) we get $\mathbb{E}X(Z\mathbb{1}_A) = \mathbb{E}Y(Z\mathbb{1}_A)$. But since $YZ \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P})$, we obtain that YZ is a version of $\mathbb{E}[XZ|\mathcal{G}]$ and the proof is concluded. □

Finally we have the following further extension of identity (2.1).

Lemma 2.1.4. *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and let Y be a version of $\mathbb{E}[X|\mathcal{G}]$. Equality (2.1) continues to hold with $\mathbb{1}_A$ extended as follows,*

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all } \mathcal{G}\text{-measurable r.v. } Z \text{ such that } \mathbb{E}[|XZ|] < \infty.$$

Proof. Take first a non-negative $\mathcal{G}$-measurable r.v. Z such that $\mathbb{E}[|XZ|] < \infty$. Then we can find a non-decreasing sequence $(Z_n)_{n\in\mathbb{N}}$ of simple non-negative $\mathcal{G}$-measurable r.v. Z such that $(Z_n)_{n\in\mathbb{N}}$ converges to Z pointwise. Thus $|XZ_n| = |X|\,Z_n \leq |X|\,Z = |XZ|$, and so by dominated convergence, XZ_n converges to XZ in L^1. Using Lemma 2.1.3 (ii), we have that YZ_n is a version of $\mathbb{E}[XZ_n|\mathcal{G}]$ and by Lemma 2.1.2, we deduce that $(YZ_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in L^1. Since it converges to YZ pointwise, we conclude that $(YZ_n)_{n\in\mathbb{N}}$ converges to YZ in L^1 by Fatou's lemma. Thus $\mathbb{E}[XZ_n] = \mathbb{E}[YZ_n]$ implies $\mathbb{E}[XZ] = \mathbb{E}[YZ]$ by letting $n \to \infty$. It remains to consider the case of a signed $\mathcal{G}$-measurable r.v. Z such that $\mathbb{E}[|XZ|] < \infty$. In this case, we observe that Z_+, Z_- are non-negative $\mathcal{G}$-measurable r.v.'s such that $\mathbb{E}[|XZ_+|], \mathbb{E}[|XZ_-|] < \infty$. Since $Z = Z_+ - Z_-$ we get the result. □

Many of the useful properties of expectations extend to conditional expectations. We state below some of these useful properties. In the following statements, all equalities and inequalities between random variables, and convergence of such, should be understood to hold $\mathbb{P}$-a.s. The proofs are left as an exercise, (see Exercise 2.3).

Proposition 2.1.5 (Elementary Properties of Conditional Expectation)**.** *Suppose that* $X, Y, Z, X_n \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ *for all* $n \geq 1$.

*(a) (*linearity*) For all* $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + bY | \mathcal{G}] = a\mathbb{E}[X | \mathcal{G}] + b\mathbb{E}[Y | \mathcal{G}] \ .$$

(b) If X *is* $\mathcal{G}$*-measurable,* $\mathbb{E}[X | \mathcal{G}] = X$.

(c) If $\mathcal{G} = \{\emptyset, \Omega\}$ *is the trivial* σ*-field, then* $\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X]$.

(d) If X *is independent of* $\mathcal{G}$*, then*

$$\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X] \ . \tag{2.4}$$

*(e) (*positivity*) If* $X \leq Y$*, then* $\mathbb{E}[X | \mathcal{G}] \leq \mathbb{E}[Y | \mathcal{G}]$.

(f) $\mathbb{E}[X | \mathcal{G}] \vee \mathbb{E}[Y | \mathcal{G}] \leq \mathbb{E}[X \vee Y | \mathcal{G}]$, $\mathbb{E}[X | \mathcal{G}]_+ \leq \mathbb{E}[X_+ | \mathcal{G}]$ *and* $|\mathbb{E}[X | \mathcal{G}]| \leq \mathbb{E}[|X| | \mathcal{G}]$.

*(g) (*tower property*) If* $\mathcal{H}$ *is a sub-*σ*-field of* $\mathcal{F}$ *such that* $\mathcal{G} \subseteq \mathcal{H}$*, then*

$$\mathbb{E}[\mathbb{E}[X | \mathcal{H}] | \mathcal{G}] = \mathbb{E}[X | \mathcal{G}] \ .$$

(h) The expectation is not modified by conditional expectation,

$$\mathbb{E}[\mathbb{E}[X | \mathcal{G}]] = \mathbb{E}[X] \ .$$

(i) If X *is* $\mathcal{G}$*-measurable and* $XY \in L^1(\Omega, \mathcal{F}, \mathbb{P})$*, then*

$$\mathbb{E}[XY | \mathcal{G}] = X\mathbb{E}[Y | \mathcal{G}] \ . \tag{2.5}$$

We conclude this section with a special case of the conditional expectation.

Definition 2.1.2 (Conditional expectation given a random element)**.** *Let* Y *be a random variable and let* $\sigma(X)$ *be the sub-*σ*-field generated by a random variable* X*. If* $\mathbb{E}[Y | \sigma(X)]$ *is well-defined, it is written as* $\mathbb{E}[Y | X]$ *and is called the conditional expectation of* Y *given* X.

By construction, $\mathbb{E}[Y | X]$ is a $\sigma(X)$-measurable random variable. Thus, there exists a Borel function g on X such that $\mathbb{E}[Y | X] = g(X)$. The choice of g is unambiguous in the sense that any two functions g and $\tilde{g}$ satisfying this equality must be equal $\mathbb{P}^X$-a.e. It is usual (although not recommended) to write $\mathbb{E}[Y | X = x]$ for such a $g(x)$.

2.1.2 Conditional Distribution

Definition 2.1.3 (Version of Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. We denote*

$$\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A | \, \mathcal{G}] \quad \text{for any event } A \in \mathcal{F}.$$

A mapping Q on $\Omega \times \mathcal{F}$ valued in $[0,1]$ is called a version of the conditional probability given *$\mathcal{G}$ if, for all $A \in \mathcal{F}$, $\omega \mapsto Q(\omega, A)$ is a version of $A \mapsto \mathbb{P}[A \,|\, \mathcal{G}]$.*

Since $\mathbb{P}[1 \,|\, \mathcal{G}] = 1$ and $\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A | \, \mathcal{G}]$ and the conditional expectation $X \mapsto \mathbb{E}[X | \, \mathcal{G}]$ satisfies the usual properties of an expectation (positivity, linearity), we might expect a version Q of the conditional probability given $\mathcal{G}$ to be a (random!) probability on $\mathcal{F}$, in the sense that for $\mathbb{P}$-almost every ω, $A \mapsto Q(\omega, A)$ would be a probability. More precisely we would like to exhibit a *regular conditional probability* as defined in the following.

Definition 2.1.4 (Regular Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A* regular version of the conditional probability of $\mathbb{P}$ given *$\mathcal{G}$ is a function*

$$\mathbb{P}^{\mathcal{G}} : \Omega \times \mathcal{F} \to [0,1]$$

such that

(i) For all $A \in \mathcal{F}$, $\mathbb{P}^{\mathcal{G}}(A) : \omega \mapsto \mathbb{P}^{\mathcal{G}}(\omega, A)$ is $\mathcal{G}$-measurable and is a version of the conditional probability of A given $\mathcal{G}$, $\mathbb{P}^{\mathcal{G}}(A) = \mathbb{P}[A \,|\, \mathcal{G}]$;

(ii) For all $\omega \in \Omega$, the mapping $A \mapsto \mathbb{P}^{\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{F}$.

When dealing with a regular conditional probability $\mathbb{P}^{\mathcal{G}}$, one can use all the usual properties of the measure for $\mathbb{P}^{\mathcal{G}}(\omega, \cdot)$. For instance, for any $Y \geq 0$, one can define the random variable $\mathbb{E}^{\mathcal{G}}[Y]$ as

$$\mathbb{E}^{\mathcal{G}}[Y](\omega) = \int Y(\omega') \, \mathbb{P}^{\mathcal{G}}(\omega, \mathrm{d}\omega') \; .$$

As usual this definition extends to a signed Y by setting

$$\mathbb{E}^{\mathcal{G}}[Y] = \mathbb{E}^{\mathcal{G}}[Y_+] - \mathbb{E}^{\mathcal{G}}[Y_-] \; ,$$

provided that this difference is well defined $\mathbb{P}$-a.s. (that is, for $\mathbb{P}$-a.e. ω, at least one the terms $\mathbb{E}^{\mathcal{G}}[Y_+]$ or $\mathbb{E}^{\mathcal{G}}[Y_-]$ is finite). All the usual properties (monotone convergence, dominated convergence, Fubini,...) can then be applied for all ω. In particular we have the following result.

Lemma 2.1.6. *Let $\mathbb{P}^{\mathcal{G}}$ be a regular version of the conditional probability of $\mathbb{P}$ given $\mathcal{G}$ and let $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\mathbb{E}[Y | \, \mathcal{G}] = \mathbb{E}^{\mathcal{G}}[Y] \quad \mathbb{P}\text{-a.s.}$$

Proof. We already have this property for $Y = \mathbb{1}_B$ with $B \in \mathcal{F}$ by Definition 2.1.4(i). By linearity of the conditional probability, it remains true for all simple random variables Y. Also by linearity, it is now sufficient to prove the lemma in the case where $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ is non-negative, in which case we can build a non-decreasing sequence of simple random variables $(Y_n)_{n \in \mathbb{N}}$ converging to Y pointwise and so also in L^1. By Lemma 2.1.2, it follows that $\mathbb{E}[Y_n | \, \mathcal{G}] = \mathbb{E}^{\mathcal{G}}[Y_n]$ converges in L^1 to $\mathbb{E}[Y | \, \mathcal{G}]$. We also know by monotone convergence that $\mathbb{E}^{\mathcal{G}}[Y_n](\omega)$ converges to $\mathbb{E}^{\mathcal{G}}[Y](\omega)$ for every ω, hence $\mathbb{E}^{\mathcal{G}}[Y]$ coincides $\mathbb{P}$-a.s. with the L^1 limit $\mathbb{E}[Y | \, \mathcal{G}]$ of $\big(\mathbb{E}^{\mathcal{G}}[Y_n]\big)_{n \in \mathbb{N}}$. □

Also, if now Y is a $(\mathsf{Y}, \mathcal{Y})$-valued random variable, for all $A \in \mathcal{Y}$, we have

$$\mathbb{P}^{\mathcal{G}}(Y \in A) = \mathbb{P}\left[Y \in A \,|\, \mathcal{G}\right] .$$

For each ω the image probability of Y under $\mathbb{P}^{\mathcal{G}}(\omega, \cdot)$ is of course a probability and provides what we call a *regular version of the conditional distribution* of Y given $\mathcal{G}$.

Definition 2.1.5 (Regular conditional distribution of Y given $\mathcal{G}$)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space and let Y be an Y-valued random variable. A* regular version of the conditional distribution of Y given $\mathcal{G}$ *is a function*

$$\mathbb{P}^{Y|\mathcal{G}} : \Omega \times \mathcal{Y} \to [0, 1]$$

such that

(i) For all $A \in \mathcal{Y}$, $\omega \mapsto \mathbb{P}^{Y|\mathcal{G}}(\omega, A)$ is $\mathcal{G}$-measurable and is a version of conditional distribution of Y given $\mathcal{G}$, $\mathbb{P}^{Y|\mathcal{G}}(\cdot, A) = \mathbb{P}\left[Y \in A \,|\, \mathcal{G}\right]$ $\mathbb{P}$-a.s.

(ii) For every ω, $A \mapsto \mathbb{P}^{Y|\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{Y}$.

Finally, in the case where $\mathcal{G} = \sigma(X)$, the two following definitions are more convenient.

Definition 2.1.6. *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A* kernel *Q on $\mathsf{X} \times \mathcal{Y}$ is a mapping $Q : \mathsf{X} \times \mathcal{Y} \to [0, \infty]$ satisfying the following conditions:*

(i) for every $A \in \mathcal{Y}$, the mapping $Q(\cdot, A) : x \to Q(x, A)$ is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $[0, \infty]$.

(ii) for every $x \in \mathsf{X}$, the mapping $Q(x, \cdot) : A \mapsto Q(x,\ A)$ is a measure on $\mathcal{Y}$,

- *Q is called a probability kernel if $Q(x, \mathsf{Y}) = 1$, for all $x \in \mathsf{X}$.*
- *Q is called a Markov kernel if it is a probability kernel on $\mathsf{X} \times \mathcal{X}$, that is, in the case where $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ are taken to be the same measurable space.*

Example 2.1.1 (Measure seen as a *constant* kernel)**.** *A σ-finite positive measure ν on a space $(\mathsf{Y}, \mathcal{Y})$ can be seen as a kernel on $\mathsf{X} \times \mathcal{Y}$ by defining $N(x, A) = \nu(A)$ for all $x \in \mathsf{X}$ and $A \in \mathcal{Y}$. It is a probability kernel if and only if ν is a probability measure.*

Example 2.1.2 (Discrete state-space kernel)**.** *Assume that X and Y are countable sets. Each element $x \in \mathsf{X}$ is then called a state. A kernel N on $\mathsf{X} \times \mathcal{P}(\mathsf{Y})$, where $\mathcal{P}(\mathsf{Y})$ is the set of all parts of Y, is specified by a (possibly doubly infinite) matrix $N = [N(x, y)]_{x,y \in \mathsf{X} \times \mathsf{Y}}$ with nonnegative entries. Each row $[N(x, y)]_{y \in \mathsf{Y}}$ defines a measure on $(\mathsf{Y}, \mathcal{P}(\mathsf{Y}))$ by setting (using the same notation N as for the matrix symbol)*

$$N(x, A) = \sum_{y \in A} N(x, y) ,$$

for $A \subset \mathsf{Y}$. The obtained kernel N is a probability kernel if every row $[N(x, y)]_{y \in \mathsf{Y}}$ defines a probability on $(\mathsf{Y}, \mathcal{P}(\mathsf{Y}))$, that is,

$$\sum_{y \in \mathsf{Y}} N(x, y) = 1$$

for all $x \in \mathsf{X}$. Note that in the discrete case, the kernel is characterized by setting $N(x, \{y\}) = N(x, y)$ for all $x, y \in \mathsf{X}$.

Example 2.1.3 (Kernel density)**.** *Let λ be a positive σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$ and $n : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ be a nonnegative function, measurable with respect to the product σ-field $\mathcal{X} \otimes \mathcal{Y}$. Then, the application N defined on $\mathsf{X} \times \mathcal{Y}$ by*

$$N(x, A) = \int_A n(x, y)\lambda(\mathrm{d}y) ,$$

is a kernel. The function n is called the density of the kernel N with respect to the measure λ. The kernel N is a probability kernel if and only if $\int_{\mathsf{Y}} n(x, y)\lambda(\mathrm{d}y) = 1$ for all $x \in \mathsf{X}$.

Definition 2.1.7 (Regular conditional distribution of Y given X)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X and Y be random variables with values in the measurable spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. A* regular version of the conditional distribution of Y given X *is a probability kernel*

$$\mathbb{P}^{Y|X} : \mathsf{X} \times \mathcal{Y} \to [0, 1]$$

such that for all $A \in \mathcal{Y}$,

$$\mathbb{P}^{Y|X}(X, A) = \mathbb{P}\left[Y \in A \,|\, X\right] \quad \mathbb{P}\text{-a.s.} \tag{2.6}$$

Going back to the conditional probability $F \to \mathbb{P}\left[F \,|\, \mathcal{G}\right]$, it is not always possible to find a regular version. The difficulty follows from the fact that for each F, $\mathbb{P}\left[F \,|\, \mathcal{G}\right]$ is defined up to a $\mathbb{P}$-null set and this null set may change from one F to another. Because unless in very specific cases the σ-field $\mathcal{F}$ is not countable, there is no guarantee in general that one can choose a particular version such that $F \mapsto \mathbb{P}\left[F \,|\, \mathcal{G}\right]$ is a probability.

However in the context of these lecture notes we can always rely on the following result which says that a regular conditional distribution of Y always exists if Y takes its values in a "nice" space. (We admit this result here, which is actually a very specific case which can be extended to more general topological spaces, in particular infinite dimensional ones, see [4, Theorem 10.2.2]).

Theorem 2.1.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let $d \geq 1$ and Y be an $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued random variable. Then there exists a regular version of the conditional distribution of Y given $\mathcal{G}$, $\mathbb{P}^{Y|\mathcal{G}}$, and this version is unique in the sense that for any other regular version $\bar{\mathbb{P}}^{Y|\mathcal{G}}$ of this distribution, for $\mathbb{P}$-almost every ω it holds that*

$$\mathbb{P}^{Y|\mathcal{G}}(\omega, F) = \bar{\mathbb{P}}^{Y|\mathcal{G}}(\omega, F) \quad \text{for all } F \in \mathcal{F} .$$

Moreover, if $\mathcal{G} = \sigma(X)$ for some r.v. X with values in a measurable spaces $(X, \mathcal{X})$, there also exists a unique regular version (hence a probability kernel) $\mathbb{P}^{Y|X}$ of the conditional distribution of Y given X.

When a regular version of a conditional distribution of Y given $\mathcal{G}$ exists, conditional expectations can be written as integrals for each ω.

Finally, it is of interest to define the regular conditional distribution of a random variable Y given another random variable X. Exactly as in Lemma 2.1.6, we have the following useful result.

Lemma 2.1.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X and Y be random variables with values in the measurable spaces $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{X}, \mathcal{X})$, respectively, and let $\mathbb{P}^{Y|X}$ be a regular version of the conditional expectation of Y given X. Then for any real-valued measurable function g on Y such that $\mathbb{E}\left[|g(Y)|\right] < \infty$, we have*

$$\mathbb{E}\left[g(Y)|\, X\right] = \int g(y)\, \mathbb{P}^{Y|X}(X, \mathrm{d}y) \quad \mathbb{P}\text{-a.s.}$$

2.1.3 Disintegration of a measure on a product space

Recall that, given two σ-finite measures $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ and $\nu \in \mathbb{M}_+(\mathsf{Y}, \mathcal{Y})$, the product measure $\mu \otimes \nu$ is uniquely defined on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ by

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B)\,, \qquad A \in \mathcal{X},\, B \in \mathcal{Y}\,.$$

Moreover, in the case of probability measures, (X, Y) has distribution $\mu \otimes \nu$ exactly means that X has distribution μ, Y has distribution ν and X and Y are independent. This can be summarized as

$$\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes \mathbb{P}^Y \quad \text{with } \mathbb{P}^X = \mu \text{ and } \mathbb{P}^Y = \nu\,. \tag{2.7}$$

Now consider any two random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. The existence of a regular version $\mathbb{P}^{Y|X}$ of the conditional distribution of Y given X allows us to extend the above product, which only holds if X and Y are independent to the following one, which will always hold :

$$\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}\,. \tag{2.8}$$

This is called a disintegration of $\mathbb{P}^{(X,Y)}$. The goal of this section is to define the $\otimes$ that appears in this formula, which can be seen as a shortcut for the identity

$$\mathbb{P}^{(X,Y)}(A \times B) = \mathbb{E}\left[\mathbb{1}_A(X)\mathbb{E}\left[\mathbb{1}_B(Y)|\, X\right]\right] = \mathbb{E}\left[\mathbb{1}_A(X)\mathbb{P}^{Y|X}(X, B)\right]\,. \tag{2.9}$$

Let us first explain how general kernels also act on measures. Let N be a kernel on $\mathsf{X} \times \mathcal{Y}$ and μ be a positive measure on $(\mathsf{X}, \mathcal{X})$ and define the measure μN by

$$\mu N(A) = \int_{\mathsf{Y}} \mu(\mathrm{d}x)\; N(x,\; A)\,, \qquad A \in \mathcal{Y}\,.$$

Proposition 2.1.9. *Let N be a kernel on $\mathsf{X} \times \mathcal{Y}$ and $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Then $\mu N \in \mathbb{M}_+(\mathsf{Y}, \mathcal{Y})$. If N is a probability kernel, then $\mu(\mathsf{X}) = \mu N(\mathsf{X})$.*

Proof. Note first that $\mu N(A) \geq 0$ for all $A \in \mathcal{Y}$ and $\mu N(\emptyset) = 0$ since $N(x, \emptyset) = 0$ for all $x \in \mathsf{X}$. Therefore, it suffices to establish the countable additivity of μN. Let $(A_i)_{i\in\mathbb{N}} \subset \mathcal{X}$ be a sequence of pairwise disjoints sets. Since $N(x, \cdot)$ is a measure for all $x \in \mathsf{X}$, the countable additivity implies that $N(x, \bigcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty N(x, A_i)$. Moreover, the function $x \mapsto N(x, A_i)$ is nonnegative and measurable for all $i \in \mathbb{N}$, thus the monotone convergence theorem yields

$$\mu N\left(\bigcup_{i=1}^\infty A_i\right) = \int \mu(\mathrm{d}x) N\left(x, \bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \int \mu(\mathrm{d}x) N(x, A_i) = \sum_{i=1}^\infty \mu N(A_i)\,.$$

□

Next we introduce the notation $\mu \otimes N$. If μ is a σ-finite positive measure on $(\mathsf{X}, \mathcal{X})$ and N is a kernel on $\mathsf{X} \times \mathcal{Y}$, then we can also define the tensor product of μ and N, denoted by $\mu \otimes N$, which is the measure on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ defined by

$$\mu \otimes N(C) = \int \left(\int \mathbb{1}_C(x, y)\, N(x, \mathrm{d}y)\right) \mu(\mathrm{d}x)\,.$$

Note in particular that if $C = A \times B$ with $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, we have

$$\mu \otimes N(A \times B) = \int \mathbb{1}_A(x)\, N(x, B)\, \nu(\mathrm{d}x)\ , \tag{2.10}$$

which, in the special case $A = \mathsf{X}$ gives that

$$\mu N = \mu \otimes N(\mathsf{X} \times \cdot)\ . \tag{2.11}$$

Moreover, as usual if $\mu \otimes N$ is a probability measure, Identity (2.10) on all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$ entirely characterizes $\mu \otimes N$ (see Theorem 2.0.2). This is the case if μ is a probability and N is a probability kernel since then one easily gets that $\mu \otimes N$ is also a probability. Hence we immediately see that, with these definitions, we indeed have that (2.9) is equivalent to (2.8). In fact the converse is true: if one is able to "disintegrate" the distribution of (X, Y) conveniently, it means that one has exhibited the marginal $\mathbb{P}^X$ and (a regular version of) the conditional distribution $\mathbb{P}^{Y|X}$. We state this as follows.

Theorem 2.1.10. *Let X and Y be two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. Let μ be a probability on $(\mathsf{X}, \mathcal{X})$ and N be a probability kernel on $\mathsf{X} \times \mathcal{Y}$. Then we have $\mathbb{P}^{(X,Y)} = \mu \otimes N$ if and only if $\mathbb{P}^X = \mu$ and N is a regular version of the conditional distribution of Y given X. Moreover it then follows that $\mathbb{P}^Y = \mu N$.*

Proof. As explained above, the "if" part follows from observing that (2.8) holds as soon as a regular version of the conditional distribution of Y given X exists, as a consequence of (2.9). Now, the "only if" part in turn follows from the converse implication: if $\mathbb{P}^{(X,Y)} = \mu \otimes N$ then, for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

$$\mathbb{P}^{(X,Y)}(A \times B) = \int \mathbb{1}_A(x) N(x, B)\ \mu(\mathrm{d}x)\ .$$

In particular taking $B = \mathsf{Y}$ yields $\mu = \mathbb{P}^X$ but then the latter equation can be written as

$$\mathbb{P}^{(X,Y)}(A \times B) = \mathbb{E}\left[\mathbb{1}_A(X) N(X, B)\right]\ ,$$

which yields $N(X, B) = \mathbb{E}\left[\mathbb{1}_B(Y) \middle| X\right]$, which concludes the converse implication. The last assertion is a direct application of (2.11). □

2.1.4 Conditional distribution for Gaussian vectors

An important application of the projection theorem in Hilbert spaces is the computation of the conditional mean for L^2 random variables, see Lemma 2.1.1(ii). It also provides an easy way to compute the conditional distribution in a Gaussian context, where the following result holds. The poof is left as an exercise (see Exercise 2.4).

Proposition 2.1.11. *Let $p, q \geq 1$. Let $\mathbf{X}$ and $\mathbf{Y}$ be two jointly Gaussian vectors, respectively valued in $\mathbb{R}^p$ and $\mathbb{R}^q$. Define*

$$\widehat{\mathbf{X}} := \mathrm{proj}\left(\mathbf{X} \middle| \left\{a + B\mathbf{Y}\ :\ a \in \mathbb{R}^p\,,\ B \in \mathbb{R}^{p \times q}\right\}\right)\ .$$

Then the following assertions hold.

(i) We have

$$\mathbb{E}\left[\mathbf{X}|\,\mathbf{Y}\right] = \widehat{\mathbf{X}}\,.$$

(ii) We have

$$\mathrm{Cov}(\mathbf{X}-\widehat{\mathbf{X}}) = \mathbb{E}\left[\mathbf{X}(\mathbf{X}-\widehat{\mathbf{X}})^T\right] = \mathbb{E}\left[(\mathbf{X}-\widehat{\mathbf{X}})\mathbf{X}^T\right]$$

and the conditional distribution of $\mathbf{X}$ *given* $\mathbf{Y}$ *is given by*

$$\mathbb{P}^{\mathbf{X}|\mathbf{Y}}(\mathbf{Y},\cdot) = \mathcal{N}(\widehat{\mathbf{X}}, \mathrm{Cov}(\mathbf{X}-\widehat{\mathbf{X}}))\,.$$

(iii) If $\mathrm{Cov}(\mathbf{Y})$ *is invertible, then* $\widehat{\mathbf{X}}$ *is given by*

$$\widehat{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\right] + \mathrm{Cov}\left(\mathbf{X},\mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\left(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}\right]\right),$$

and

$$\mathrm{Cov}(\mathbf{X}-\widehat{\mathbf{X}}) = \mathrm{Cov}(\mathbf{X}) - \mathrm{Cov}\left(\mathbf{X},\mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\mathrm{Cov}\left(\mathbf{Y},\mathbf{X}\right)\,.$$

2.2 Radon-Nikodym derivative

2.2.1 Domination (absolute continuity)

Let μ, λ be σ-finite measures on $(\Omega, \mathcal{F})$ such that for all $A \in \mathcal{F}$ we have

$$\mu(A) = \int_A \phi \,\mathrm{d}\lambda\,,$$

where ϕ is some Borel function. Then we say that μ admits a density ϕ with respect to λ. Note that $\phi \geq 0$ λ-a.s. and the property $\mu(A) = \int_A \phi\,\mathrm{d}\lambda$ for all $A \in \mathcal{F}$ defines ϕ uniquely up to the equality λ-a.e. (See Exercise 2.5) Hence the following definition.

Definition 2.2.1 (Radon-Nikodym derivative)**.** *If* $\mu(A) = \int_A \phi\,\mathrm{d}\lambda$ *for all* $A \in \mathcal{F}$*, we say that the* λ-a.e. *equivalent class of* ϕ *is the Radon-Nikodym derivative of* μ *with respect to* λ *and write* $\phi = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$*.*

We note that if μ admits a density ϕ with respect to λ, then $\lambda(A) = 0$ implies $\mu(A) = 0$. It turns out that this property is sufficient to have that μ admits a density with respect to λ.

Definition 2.2.2 (Absolute continuity of measures)**.** *Let* λ *be a measure on* $(\Omega, \mathcal{F})$*. We say that a* σ*-finite measure* μ *is* absolutely continuous with respect to λ *or that* λ dominates μ *and we write* $\mu \ll \lambda$ *if for all* $A \in \mathcal{F}$*,* $\lambda(A) = 0$ *implies* $\mu(A) = 0$*.*

The following result holds.

Theorem 2.2.1 (Radon-Nikodym theorem)**.** *Let* $\lambda, \mu \in \mathbb{M}_+(\Omega, \mathcal{F})$ *be* σ*-finite measures on* $(\Omega, \mathcal{F})$ *such that* $\mu \ll \lambda$*. Then there exists a non-negative Borel function* ϕ *such that for all* $A \in \mathcal{F}$*,* $\mu(A) = \int_A \phi\,\mathrm{d}\lambda$*.*

Proof. We assume that μ and λ are finite with $\mu \ll \lambda$ (it then easily extends to the case where μ is σ-finite case by partitioning the space with finite μ and λ-measures subsets). Let $\rho = \lambda + \mu$. Let us define, for all $f \in L^2(\Omega, \mathcal{F}, \rho)$,

$$I(f) = \int f \,\mathrm{d}\mu\,.$$

Then we have

$$|I(f)| \leq \int |f| \, \mathrm{d}\mu \leq (\mu(\Omega))^{1/2} \left(\int |f|^2 \, \mathrm{d}\mu \right)^{1/2} \leq (\mu(\Omega))^{1/2} \left(\int |f|^2 \, \mathrm{d}\rho \right)^{1/2} .$$

Hence $f \mapsto I(f)$ is a continuous linear form on $L^2(\Omega, \mathcal{F}, \rho)$. By the Riesz representation theorem (see Theorem 1.5.3), there exists $g \in L^2(\Omega, \mathcal{F}, \rho)$ such that for all $f \in L^2(\Omega, \mathcal{F}, \rho)$, $I(f) = \langle f, \bar{g} \rangle$, that is

$$\int f \, \mathrm{d}\mu = \int f \, g \, \mathrm{d}\rho \, . \tag{2.12}$$

Remark moreover that for all $f \geq 0$, $\langle f, \bar{g} \rangle \geq 0$, so one easily gets that the $L^2(\Omega, \mathcal{F}, \rho)$-norm of $\Im(g)$ is zero, hence g is real and the $L^2(\Omega, \mathcal{F}, \rho)$-norm of g_- is zero, hence $g \geq 0$ ρ-a.e.

Let $A \in \mathcal{F}$. We have $\mathbb{1}_A \in L^2(\Omega, \mathcal{F}, \rho)$ and applying (2.12) with $f = \mathbb{1}_A$, we get

$$\mu(A) = \int_A g \, \mathrm{d}\rho \, . \tag{2.13}$$

Take $A = \{g \geq 1\}$. Since $g \mathbb{1}_A \geq \mathbb{1}_A$, we get

$$\mu(A) = \int_A g \, \mathrm{d}\rho \geq \rho(A) = \mu(A) + \lambda(A) \, .$$

Hence $\lambda(A) = 0$ and since $\mu \ll \lambda$ we also have $\mu(A) = 0$ and then $\rho(A) = 0$. Hence finally $0 \leq g < 1$ ρ-a.e. Modifying g by setting it to 0 on the set $\{0 \leq g < 1\}^c$, we thus finally have (2.13) for all $A \in \mathcal{F}$ and $0 \leq g < 1$. Hence since $\rho = \lambda + \mu$, we get that, for all $A \in \mathcal{F}$,

$$\int \mathbb{1}_A (1 - g) \, \mathrm{d}\mu = \int \mathbb{1}_A g \, \mathrm{d}\lambda.$$

This identity extends to any non-negative Borel function f in place of $\mathbb{1}_A$ by monotone convergence and in particular to $f = \mathbb{1}_A/(1-g)$ and obtain

$$\mu(A) = \int_A \phi \, \mathrm{d}\lambda \, ,$$

where $\phi = g/(1-g)$. □

2.2.2 Conditional density

The following definition is classically seen in elementary probability courses in the discrete case (ξ and ξ' are compting measures on countable sets) or in the Lebesgue case (ξ and ξ' are Lebesgue measures of given dimensions).

Definition 2.2.3 (Conditional density)**.** *Let (X, Y) be two random elements admitting a density f with respect to $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then the function $(x, y) \mapsto f(y|x)$ defined for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ by*

$$f(y|x) = \frac{f(x, y)}{\int f(x, y') \, \mathrm{d}\xi'(y')}$$

is called the conditional density of Y given X.

Observe that the denominator is the density of X with respect to ξ applied to x. Hence (see Exercise 2.5) it satisfies

$$0 < \int f(x, y') \, \mathrm{d}\xi'(y') < \infty \quad \text{for } \mathbb{P}^X\text{-a.e. } x.$$

For x's such that this is not true, we can set $y \mapsto f(y|x)$ to be any arbitrary density such as the density of Y, in which case

$$y \mapsto f(y|x) \text{ is a density with respect to } \xi' \text{ for all } x \in \mathsf{X}.$$

In the case where (X, Y) satisfies the assumptions of Definition 2.2.3, the conditional distribution of Y given X is given by the following result, whose proof is left as an exercise (Exercise 2.6).

Theorem 2.2.2. *Let (X, Y) be two random elements admitting a density $f : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ with respect to $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then the conditional distribution of Y given X is given by*

$$\mathbb{P}^{Y|X}(x, A) = \int_A f(y|x) \, \xi'(\mathrm{d}y) \quad \textit{for all } x \in \mathsf{X} \textit{ and } A \in \mathcal{Y} \, ,$$

where $(x, y) \mapsto f(y|x)$ is the conditional density of Y given X.

2.2.3 Kullback-Leibler divergence

There are many ways to define a metric on the set of probability measures defined on the same measurable set. We here introduce the Kullback-Leibler divergence which is a classical way to quantify how close two given probability are. It is not a metric however since it is not symmetric but it is a very useful notion in Information theory and also in statistics. We start with an important lemma.

Lemma 2.2.3. *Let P and Q be two probabilities on the measurable space $(\Omega, \mathcal{F})$ and let $\nu \in \mathbb{M}_+(\Omega, \mathcal{F})$ dominate both P and Q, for instance, take $\nu = P + Q$. Let f_P and f_Q denote the densities of P and Q with respect to ν. Then the quantity*

$$\mathrm{KL}\left(P \,\|\, Q\right) = \int \ln \frac{f_P}{f_Q} \, \mathrm{d}P \tag{2.14}$$

is always well defined and takes values in $[0, \infty]$. Moreover we have the following properties.

(i) If Q does not dominate P then $\mathrm{KL}\left(P \,\|\, Q\right) = \infty$.

(ii) If $P \ll Q$ then

$$\mathrm{KL}\left(P \,\|\, Q\right) = \int \ln \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P \, .$$

(and the latter may be finite or infinite).

(iii) We have $\mathrm{KL}\left(P \,\|\, Q\right) = 0$ if and only if $P = Q$.

As usual when a fraction appear one has to wonder what happens when the denominator appearing in (2.14) is zero. It is always clear that $a/0 = \infty$ for $a > 0$ and classical in probability to set $0/0 = 0$. Note however that in (2.14) taking the integral only on $\{f_P > 0\}$ does not modify its value since $f_P > 0$ P-a.s., so how one defines $0/0$ is in fact unimportant in (2.14).

Proof. First observe that the function $\ln : [0,\infty] \to [-\infty,\infty]$ satisfies $\ln x \leq x - 1$ and thus

$$\int \ln \frac{f_Q}{f_P} \, \mathrm{d}P \leq \int \frac{f_Q}{f_P} \, \mathrm{d}P - 1 = \int_{f_P>0} f_Q \, \mathrm{d}\nu - 1 \leq 0 \, . \tag{2.15}$$

This shows that $\mathrm{KL}\,(P \,\|\, Q)$ is well defined and takes values in $[0,\infty]$.

Let us now prove (i). Let $A \in \mathcal{F}$ be such that $Q(A) = 0$ and $P(A) > 0$. Then $f_Q(x) = 0$ for ν-a.e. $x \in A$, implying $\ln f_P(x)/f_Q(x) = \infty$ whenever $f_P(x) > 0$. Hence we get that

$$\int_A \ln \frac{f_P}{f_Q} \, \mathrm{d}P = \int_{A\cap\{f_P>0\}} \ln \frac{f_P}{f_Q} \, \mathrm{d}P = \int_{A\cap\{f_P>0\}} \infty \mathrm{d}P = \infty \, ,$$

since $P(A) > 0$. Hence $\mathrm{KL}\,(P \,\|\, Q) = \infty$.

Let us now prove (ii). Let $g = \mathrm{d}P/\mathrm{d}Q$. Then we have $f_P = g\, f_Q$ ν-a.e., hence $f_P/f_Q = g$ P-a.s., which yields (ii).

We conclude with the proof of (iii). If $P = Q$ then $f_P = f_Q$ ν-a.e. and we get $\mathrm{KL}\,(P \,\|\, Q) = 0$. Suppose now that $\mathrm{KL}\,(P \,\|\, Q) = 0$. Then all the inequalities in (2.15) are equalities and so

$$\int \left(\frac{f_Q}{f_P} - 1 - \ln \frac{f_Q}{f_P} \right) \mathrm{d}P = 0 \, .$$

Since the function between parentheses is non-negative and vanishes only on the set $\{f_P = f_Q\}$, this implies that $f_P = f_Q$ P-a.s. In particular for any set $A \in \mathcal{F}$, we have

$$P(A) = P(A \cap \{f_P = f_Q\}) = \int_{A\cap\{f_P=f_Q\}} f_P \, \mathrm{d}\nu = \int_{A\cap\{f_P=f_Q\}} f_Q \, \mathrm{d}\nu \leq Q(A) \, .$$

Applying this to A^c, we get $1 - P(A) \leq 1 - Q(A)$ so, finally, $P(A) = Q(A)$, and since this is true for all $A \in \mathcal{F}$, it concludes the proof. □

We see from Lemma 2.2.3 that the quantity $\mathrm{KL}\,(P \,\|\, Q)$ does not depend on the choice of the dominating measure ν. Hence the following definition.

Definition 2.2.4 (Kullback-Leibler divergence)**.** *Let P and Q be two probabilities on the measurable space $(\Omega, \mathcal{F})$. The quantity $\mathrm{KL}\,(P \,\|\, Q)$ defined by (2.14) is called the* Kullback-Leibler divergence *between P and Q.*

We conclude this section with the following useful theorem.

Theorem 2.2.4. *Let P and Q be two probabilities on the measurable space $(\Omega, \mathcal{F})$ and X a measurable mapping from $(\Omega, \mathcal{F})$ to $(\mathsf{X}, \mathcal{X})$. Then we have*

$$\mathrm{KL}\left(P^X \,\|\, Q^X\right) \leq \mathrm{KL}\,(P \,\|\, Q) \, .$$

Proof. We only have something to prove if $P \ll Q$. Let $g = \mathrm{d}P/\mathrm{d}Q$. Then for all $A \in \mathcal{X}$ we have

$$P^X(A) = \int_{X^{-1}(A)} \mathrm{d}P = \int_{X^{-1}(A)} g \, \mathrm{d}Q = \mathbb{E}_Q[\mathbb{1}_{\{X^{-1}(A)\}} g] \, ,$$

where $\mathbb{E}_Q$ denotes the expectation with respect to Q. Conditioning on X, we get that

$$P^X(A) = \mathbb{E}_Q[\mathbb{E}_Q[\mathbb{1}_{\{X^{-1}(A)\}} g | X]] = \mathbb{E}_Q[\mathbb{1}_{\{X^{-1}(A)\}} \mathbb{E}_Q[g|X]] \, .$$

Since $\mathbb{E}_Q[g|X]$ is $\sigma(X)$-measurable, it can be written as

$$\mathbb{E}_Q[g|X] = h \circ X \ ,$$

and thus we obtain

$$P^X(A) = \int \mathbb{1}_A \, h \, \mathrm{d}Q^X \ ,$$

which means that P^X admits h as a density with respect to Q^X. Hence we have

$$\mathrm{KL}\left(P^X \,\|\, Q^X\right) = \int \ln h \, \mathrm{d}P^X = \int h \, \ln h \, \mathrm{d}Q^X = \int h \circ X \, \ln h \circ X \, \mathrm{d}Q = \mathbb{E}_Q[H] \ ,$$

where

$$H = \mathbb{E}_Q[g|X] \, \ln \mathbb{E}_Q[g|X] = \phi\left(\mathbb{E}_Q[g|X]\right) \ .$$

where $\phi : x \mapsto x \ln x$ is defined from $[0, \infty) \to \mathbb{R}$. With the same notation, we can write

$$\mathrm{KL}\left(P \,\|\, Q\right) = \int \ln g \, \mathrm{d}P = \int \phi \circ g \, \mathrm{d}Q \ .$$

Hence to conclude the proof, we only need to show that

$$\mathbb{E}_Q[H] \leq \mathbb{E}_Q[\phi \circ g] \ . \tag{2.16}$$

and we only need to do that in the case where $\mathbb{E}_Q[|\phi \circ g|] < \infty$ (otherwise $\mathbb{E}_Q[\phi \circ g] = \infty$). Note that ϕ is convex and proceeding as in the Jensen inequality, we can write, if $\mathbb{E}_Q[g|X] > 0$,

$$\phi \circ g \geq \phi\left(\mathbb{E}_Q[g|X]\right) + \phi'\left(\mathbb{E}_Q[g|X]\right) \times \left(g - \mathbb{E}_Q[g|X]\right) \ .$$

Define, for any $n \geq 1$, $A_n = \{n^{-1} \leq \mathbb{E}_Q[g|X] \leq n\}$, so that multiplying by $\mathbb{1}_{A_n}$ on both sides and using that $\mathbb{E}_Q[|\phi \circ g|], \mathbb{E}_Q[|g|] < \infty$, we can apply $\mathbb{E}_Q[\cdot|X]$ on both sides, obtaining that

$$\mathbb{1}_{A_n} \mathbb{E}_Q[\phi \circ g|X] \geq \mathbb{1}_{A_n} \phi\left(\mathbb{E}_Q[g|X]\right) \quad Q\text{-a.s.}$$

Letting $n \to \infty$, we get that Q-a.e. on the set $\{\mathbb{E}_Q[g|X] > 0\}$,

$$\mathbb{E}_Q[\phi \circ g|X] \geq \phi\left(\mathbb{E}_Q[g|X]\right) = H \quad Q\text{-a.s.} \tag{2.17}$$

Let $B = \{\mathbb{E}_Q[g|X] = 0\}$. Then

$$\mathbb{E}_Q\left[g \mathbb{1}_B\right] = \mathbb{E}_Q\left[\mathbb{E}_Q\left[g|\, X\right] \mathbb{1}_B\right] = 0 \ ,$$

so $g\mathbb{1}_B = 0$ Q-a.s. and, since $\phi(0) = 0$, $\phi \circ g\mathbb{1}_B = 0$ Q-a.s. as well, from which we get that (2.17) holds also on B. Hence (2.16) follows by appluing $\mathbb{E}_Q$ on both sides. □

2.3 Exercises

Exercise 2.1 (Pseudo-inverse d'une fonction de répartition, ordre stochastique). Let F be the distribution function of a probability on $\mathbb{R}$. Define (as usual) its *pseudo-inverse* by

$$F^{-1}(t) = \inf\{x : \ F(x) \geq t\}, \qquad t \in (0,1) .$$

1. Show that for all $t \in (0,1)$, $F \circ F^{-1}(t) \geq t$, with equality if F is continuous.
2. Show that for all $x \in \mathbb{R}$, $F^{-1} \circ F(x) \leq x$, , with equality if F is (strictly) increasing.
3. Let X be a r.v. with distribution function F; show that $X = F^{-1}(F(X))$ a.s.
4. Show that if F is continuous, then $F(X)$ has a uniform distribution on $[0,1]$. Compute $\mathbb{E}[F^n(X)]$ for all $n \in \mathbb{N}$.
5. Show that if U is a uniform r.v. on $[0,1]$, then $F^{-1}(U)$ has distribution function F.

Let X and Y two r.v. with distribution functions F and G. Suppose that for all $s \in \mathbb{R}$, $F(s) \leq G(s)$. We say that F is stochastically larger than or equal to G and we denote $G \leq_{sto} F$.

6. Provide examples of F and G such that $G \leq_{sto} F$.
7. Show that $G \leq_{sto} F$ if and only if there exists a random bidimensional vector (X,Y) such that $X \sim F$, $Y \sim G$ (this is called a *coupling* with marginals (F,G)) and $Y \leq X$ a.s.
8. Provide an example of F and G such that $G \leq_{sto} F$ and a coupling (X,Y) with marginals (F,G) for which $\mathbb{P}(Y \leq X) < 1$.
9. Show that if $G \leq_{sto} F$, then for any non-decreasing $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[f(Y)] \leq \mathbb{E}[f(X)]$ for any coupling (X,Y) with marginals (F,G).

Exercise 2.2. Let Y be an L^2 real valued r.v. defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\mathcal{B}$ be a sub-σ-field of $\mathcal{A}$. Define the conditional variance of Y given $\mathcal{B}$ by

$$\sigma^2(Y|\mathcal{B}) = \mathbb{E}[Y^2 \mid \mathcal{B}] - (\mathbb{E}[Y \mid \mathcal{B}])^2.$$

Denoting by $\sigma^2(Z)$ the variance of Z, show that

$$\sigma^2(Y) = \sigma^2(\mathbb{E}[Y \mid \mathcal{B}]) + \mathbb{E}[\sigma^2(Y|\mathcal{B})].$$

What is the result when Y is independent of $\mathcal{B}$?

Exercise 2.3. Show all the properties of Proposition 2.1.5 and also the following ones.

(j) If $\sigma(X) \vee \mathcal{H} = \sigma(\sigma(X) \cup \mathcal{H})$ (the smallest σ-field containing $\sigma(X)$ and $\mathcal{H}$) is independent of $\mathcal{G}$, $\mathbb{E}[X | \mathcal{H} \vee \mathcal{G}] = \mathbb{E}[X | \mathcal{H}]$. [Hint : first take an element of $\mathcal{H} \vee \mathcal{G}$ of the form $A \cap B$. Use the $\pi - \lambda$-theorem to conclude.]

(k) Let $X = F(Y, Z)$ where Y and Z are two random vectors valued in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, and F measurable from $(\mathbb{R}^{p+q}, \mathcal{B}(\mathbb{R}^{p+q}))$ to $(\mathsf{X}, \mathcal{X})$. Suppose moreover that Y is $\mathcal{G}$-measurable and Z is independent of $\mathcal{G}$, then the conditional distribution of X given $\mathcal{G}$ is given by

$$\mathbb{P}^{X|\mathcal{G}}(\omega, A) = \mathbb{P}(F(Y(\omega), Z) \in A) \quad \text{for all } \omega \in \Omega \text{ and } A \in \mathcal{X} .$$

[Hint : first compute $\mathbb{P}^{(Y,Z)|\mathcal{G}}(\cdot, B \times C)$ for $B \in \mathcal{B}(\mathbb{R}^p)$ and $C \in \mathcal{B}(\mathbb{R}^q)$ and deduce $\mathbb{P}^{(Y,Z)|\mathcal{G}}(\cdot, D)$ for $D \in \mathcal{B}(\mathbb{R}^{p+q})$.]

Exercise 2.4. Let $\mathbf{X}$ and $\mathbf{Y}$ be as in Proposition 2.1.11.

1. In order to prove Proposition 2.1.11(i) and (ii), use Property (k) above.
2. Use the characterization of the orthogonal projection to prove Proposition 2.1.11(iii).

Exercise 2.5. Let μ and λ be two σ-finite measures on $(\Omega, \mathcal{F})$ and let a Borel function $\phi : \Omega \to \bar{\mathbb{R}}_+$ satisfy, for all $A \in \mathcal{F}$,

$$\mu(A) = \int_A \phi \, \mathrm{d}\lambda . \tag{2.18}$$

1. Show that if a Borel function $\psi : \Omega \to \bar{\mathbb{R}}$ satisfies, for all $A \in \mathcal{F}$,

$$\int_A \psi \, \mathrm{d}\lambda = 0 ,$$

then $\psi = 0$ λ-a.e.

2. Show that , for all $A \in \mathcal{F}$ such that $\lambda(A) = 0$, we have $\mu(A) = 0$.
3. Show that $\phi > 0$ μ-a.e.
4. Give an example where we do not have that $\phi > 0$ λ-a.e.
5. Show that $\phi < \infty$ λ-a.e. [Hint : first consider the case where μ is finite]
6. Show that (2.18) uniquely defines ϕ up to a λ-null set.

Exercise 2.6. Prove Theorem 2.2.2.

Exercise 2.7. Let $(\mathbf{X}, \mathbf{Y})$ be a random vector valued in $\mathbb{R}^{p+q}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Show that the conditional density of $\mathbf{X}$ given $\mathbf{Y}$ is always well defined when

1. $\mathbf{X}$ and $\mathbf{Y}$ are discrete random variables (i.e. take values in countable sets).
2. $(\mathbf{X}, \mathbf{Y})$ admit a density with respect to the Lebesgue measure.
3. $\mathbf{Y}$ is a discrete random variable [Hint : show that $(\mathbf{X}, \mathbf{Y})$ admits a density with respect to $\xi \otimes \xi'$ with ξ' a counting measure and $\xi = \mathbb{P}^{\mathbf{X}}$.].
4. Deduce a formula for $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ in all the previous cases.
5. Give a new proof of Proposition 2.1.11(i) and (ii) in the case where $(\mathbf{X}, \mathbf{Y})$ has an invertible covariance matrix.

Exercise 2.8. Let X be a real valued random variable admitting a symmetric density f (with respect to the Lebesgue measure), $f(-x) = f(x)$ for all $x \in \mathbb{R}$.

1. Compute the conditional distribution of X given $|X|$.
2. Let Y be any other random variable defined on the same probability space as X and let $\mathbb{P}^{X|Y}$ denote the conditional distribution of X given Y. Show that for all Borel set A with null Lebesgue measure, we have that, for $\mathbb{P}^Y$-a.e. y, $\mathbb{P}^{X|Y}(y, A) = 0$.
3. Do we have that $\mathbb{P}^{X|Y}(y, \cdot)$ admits a density (with respect to the Lebesgue measure) for $\mathbb{P}^Y$-a.e. y?

Exercise 2.9. Let (X, Y) be an $\mathbb{R}^2$-valued r.v. Determine the conditional expectation and distribution of X given Y in the following choices for the distribution of (X, Y) :

1. the uniform distribution on the triangle $(0, 0), (1, 0), (0, 1)$
2. the uniform distribution on the square $(0, 0), (1, 0), (1, 1), (0, 1)$

Exercise 2.10. Let X and Y be two r.v.s defined on the same probability space. Suppose that X has its values in $\mathbb{N}$ and Y follows an exponential distribution with unit mean. Suppose also that the conditional distribution of X given Y is Poisson with mean Y. Determine the distribution of (X, Y) and that of X. Compute the conditional distribution of Y given X.

Exercise 2.11. Let $X_1, \ldots, X_p$ be independent r.v.'s following Poisson distributions with parameters $\lambda_1, \ldots, \lambda_p$.

1. Determine the conditional distribution of $(X_1, \ldots, X_{p-1})$ given $X_1 + \ldots + X_p$.
2. Compute $\mathbb{E}[X_1 \mid X_1 + X_2]$.

Exercise 2.12. Let $X_1, \ldots, X_n$ be i.i.d. random variables with density f, assumed to be continuous on $\mathbb{R}$.

1. Recall that the order statistic $(X_{(1)}, \ldots, X_{(n)})$ obtained by ordering$X_1, \ldots, X_n$ in an increasing order admits a density.
2. Determine the conditional distribution of $\min_{1 \le i \le n} X_i$ given $\max_{1 \le i \le n} X_i$.
3. Assuming that $\mathbb{E}[|X_i|] < \infty$ for all $i = 1, \ldots, n$, deduce an expression of $\mathbb{E}[\min_{1 \le i \le n} X_i \mid \max_{1 \le i \le n} X_i]$.

Exercise 2.13 (Statistiques d'ordre). Let $X = (X_1, \ldots, X_n)$ be a random vector with density $f(x)$, $x \in \mathbb{R}^n$. Let $R = (R(1), \ldots, R(n))$ be *rank statistic* of X, that is, for all $i \in \{1, \ldots, n\}$,

$$R(i) = \sum_{j=1}^{n} \mathbb{1}_{\{X_i \ge X_j\}} .$$

1. Show that, with probability 1, there exists a permutation σ of $\{1, \ldots, n\}$ such that $X_{\sigma(1)} < \ldots < X_{\sigma(n)}$. What is the relationship between σ and R ?
2. Show that, for any permutation r of $\{1, \ldots, n\}$ and all borel function $g : \mathbb{R}^n \to \mathbb{R}_+$

$$\mathbb{E}\left[g(X_{\sigma(1)}, \ldots, X_{\sigma(n)})\mathbb{1}_{\{R=r\}}\right] = \int g(x_1, \ldots, x_n)\mathbb{1}_{\{x_1 < \cdots < x_n\}} f(x_{r(1)}, \ldots, x_{r(n)})\, \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

3. Deduce the conditional distribution of R given $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$.

4. What do you obtain when the X_i's are iid ?

Exercise 2.14. Let P and Q be two probabilities on $(\Omega, \mathcal{F})$.

1. Let X and Y be two measurable functions defined on $(\Omega, \mathcal{F})$ valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively, such that, under P and Q, X and Y are independent. Express $\mathrm{KL}\left(P^{(X,Y)} \,\|\, Q^{(X,Y)}\right)$ with $\mathrm{KL}\left(P^X \,\|\, Q^X\right)$ and $\mathrm{KL}\left(P^Y \,\|\, Q^Y\right)$.

2. Let $X_1, \dots, X_n$ be measurable functions defined on $(\Omega, \mathcal{F})$ such that the X_i's are iid under P and Q. Deduce $\mathrm{KL}\left(P^{X_{1:n}} \,\|\, Q^{X_{1:n}}\right)$ from $\mathrm{KL}\left(P^{X_1} \,\|\, Q^{X_1}\right)$.

Chapter 3

Mathematical statistics

Conditional calculus and domination are key tools in mathematical statistics. We here apply them to obtain some fundamental results in a quite general fashion.

3.1 Statistical modeling

Definition 3.1.1 (Statistical model)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{P}$ a collection of probabilities on this space. Let X be a measurable function from $(\Omega, \mathcal{F})$ to the* observation space *$(\mathsf{X}, \mathcal{X})$. We say that $\mathcal{P}$ is a* statistical model *for the* observation variable *X. We denote by $\mathcal{P}^X$ the corresponding collection of probability distributions, $\mathcal{P}^X = \left(P^X\right)_{P\in\mathcal{P}}$.*

It is usual in statistics to consider $\Omega = \mathsf{X}$ and $\mathcal{F} = \mathcal{X}$ and to set $X(\omega) = \omega$, in which case any $P \in \mathcal{P}$ directly corresponds to the distribution of X, $P = P^X$.

For any countable collection $\left(P_n^X\right)_{n\geq 1}$ of distributions, one can exhibit a common dominating probability measure by setting

$$\mu = \sum_{n\geq 1} 2^{-n} P_n^X$$

Having a common dominating measure is very useful to derived simple calculation based on densities. For non-countable models we rely on the following definition.

Definition 3.1.2 (Dominated model)**.** *Let $\nu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ and $\mathcal{P}$ be a statistical model for the observation variable X. We say that $\mathcal{P}$ is a ν-*dominated model *for the* observation variable *X or, in a shorter way, that $\mathcal{P}^X$ is ν-*dominated*, if for all $P \in \mathcal{P}$, $P^X \ll \nu$.*

Hence in a dominated model the distribution of X is characterized by a collection of densities with respect to a common σ-finite measure ν. This simplifies the computations a lot, a nice and widely used example of which is given by Theorem 3.3.2 given below.

Clearly if ν dominates a model $\mathcal{P}$, many other measures do. Surprisingly, we an always pick a dominating measure made up as a countable sum of probality measures taken in $\mathcal{C}$.

Lemma 3.1.1 (Halmos and Savage (1949))**.** *Let $\nu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Consider a ν-dominated model $\mathcal{P}$ for the variable X. Then there exists a countable collection $(P_n)_{n\geq 1}$ of probabilities in $\mathcal{P}$ such that $\mathcal{P}^X$ is also dominated by the probability measure*

$$\mu = \sum_{n\geq 1} 2^{-n} P_n^X \ .$$

Proof. First observe that since ν is σ-finite we can construct a probability measure which dominates ν and thus also $\mathcal{P}$. So in the following we can assume that ν is a probability measure.

Let $\mathcal{Q}$ be the set of probability measures defined by

$$\mathcal{Q} = \left\{ \sum_{n\in\mathbb{N}} c_n P_n \ : \ (P_n)_{n\in\mathbb{N}} \in \mathcal{P}^{\mathbb{N}}, (c_n)_{n\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}, \sum_{n\in\mathbb{N}} c_n = 1 \right\} .$$

Clearly $\mathcal{Q}^X$ is ν-dominated and for all $Q \in \mathcal{Q}$ we denote by f_Q a density of Q^X with respect to ν. To conclude the proof, we only need to exhibit some $\tilde{Q}^X \in \mathcal{Q}^X$ which dominates all $Q^X \in \mathcal{Q}^X$.

Let $\mathcal{C}$ be the class of all sets $A \in \mathcal{X}$ such that there exists $Q \in \mathcal{Q}$ for which $f_Q > 0$ ν-a.s. on A and $Q^X(A) > 0$. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of elements in $\mathcal{C}$ such that

$$\lim_{n\to\infty} \nu(A_n) = \sup_{A\in\mathcal{C}} \nu(A) .$$

For any $n \in \mathbb{N}$, let $Q_n \in \mathcal{Q}$ such that $f_{Q_n} > 0$ ν-a.s. on A_n and $Q_n^X(A_n) > 0$. We set

$$\tilde{Q} = \sum_{n\in\mathbb{N}} 2^{-n-1} Q_n \quad \text{and} \quad \tilde{A} = \bigcup_{n\in\mathbb{N}} A_n .$$

Then $\tilde{Q} \in \mathcal{Q}$ and $\tilde{Q}^X$ admits the density

$$f_{\tilde{Q}} = \sum_{n\in\mathbb{N}} 2^{-n-1} f_{Q_n}$$

with respect to ν. Moreover we have $f_{\tilde{Q}} > 0$ ν-a.s. on $\tilde{A}$ so $\tilde{A} \in \mathcal{C}$ and

$$\nu(\tilde{A}) = \sup_{A\in\mathcal{C}} \nu(A) . \tag{3.1}$$

Let $A \in \mathcal{X}$ such that $\tilde{Q}^X(A) = 0$ and let $Q \in \mathcal{Q}$. It only remains to show that $Q^X(A) = 0$. First note that $\tilde{Q}^X(A \cap \tilde{A}) = 0$ implies $\nu(A \cap \tilde{A}) = 0$ which implies

$$Q^X(A \cap \tilde{A}) = 0 .$$

Let $B = \{f_Q > 0\}$, hence

$$Q^X(A \cap \tilde{A}^c \cap B^c) = 0 .$$

We finally end up with the task to show that $Q^X(A \cap \tilde{A}^c \cap B) = 0$. We do this by contradiction. Suppose that

$$Q^X(A \cap \tilde{A}^c \cap B) > 0 . \tag{3.2}$$

Then we would have $A \cap \tilde{A}^c \cap B \subseteq B \in \mathcal{C}$, hence $\tilde{A} \cup \left(A \cap \tilde{A}^c \cap B\right) \in \mathcal{C}$ and from (3.1) we would obtain that

$$\nu\left(\tilde{A}\right) \leq \nu\left(\tilde{A} \cup \left(A \cap \tilde{A}^c \cap B\right)\right) \leq \nu\left(\tilde{A}\right) .$$

But this implies $\nu(A \cap \tilde{A}^c \cap B) = 0$ which contradicts (3.2), and the proof is concluded. □

3.2 Estimation of a parameter

It is convenient to index a model $\mathcal{P}$ by a finite dimensional parameter.

Definition 3.2.1 (Parametric model)**.** *Let $\mathcal{P}$ be a statistical model for the* observation variable X*. We say that $\mathcal{P}$ is a* parametric model *for the* observation variable X *if there exists a finite dimensional set Θ such that $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$.*

In a parametric model, each probability P of the model is entirely determined by a finite-dimensional *parameter* θ. If in turn this parameter characterizes the probability of X, that is $P_\theta^X = P_{\theta'}^X$ implies $\theta = \theta'$ then we say that θ is an *identifiable parameter*. More generally, we rely on the following definition.

Definition 3.2.2 (Identifiable parameter)**.** *Let $\mathcal{P}$ be a statistical model for the* observation variable X*. Any finite dimensional quantity $t(P^X)$ only depending on P^X as $P \in \mathcal{P}$ is called an* identifiable parameter*.*

In statistics, we assume a model $\mathcal{P}$ for the observation X, which means that we assume that the observed data is the realization of the r.v. X drawn according to some P^X with $P \in \mathcal{P}$ but we do not know *which* P. The main idea is that an identifiable parameter $t(P^X)$ can be "guessed" (with some quantifiable error) from one realization of X. The vector quantities that one can get from the observation are called *statistics*.

Definition 3.2.3 (Statistic)**.** *Let $\mathcal{P}$ be a statistical model for the* observation variable X*. A* statistic *in this context is any random variable T valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with $d \geq 1$, defined by $T = g(X)$ where g is a Borel function not depending on $P \in \mathcal{P}$.*

If a statistic is used as a guess for a parameter $t(P^X) \in \mathbb{R}^d$, it is called an *estimator* of the *parameter* $t(P^X)$. In this case, the *bias* of T for estimating $t(P^X)$ is defined as

$$\mathrm{Bias}(T, P) = \int T \mathrm{d}P - t(P^X) \ ,$$

whenever $\int |T| \mathrm{d}P < \infty$. (Otherwise we say that the bias is infinite). We say that T is an *unbiased* estimator of $t(P^X)$ if

$$\int T \mathrm{d}P = t(P^X) \quad \text{for all } P \in \mathcal{P}.$$

It is very important to understand that in general the bias depends on P, so having an unbiased estimator is a strong constraint but achievable in many standard examples. In contrast, the *quadratic risk* or *mean squared error* defined by (in the case $d = 1$)

$$\mathrm{MSE}(T, P) = \int (T - t(P^X))^2 \mathrm{d}P = \mathrm{Var}\,(T) + (\mathrm{Bias}(T, P))^2 \ ,$$

usually depends on P except in very particular cases.

3.3 Sufficient statistics

Next, we introduce a very important notion in statistics, which relies on conditional probabilities.

Definition 3.3.1 (Sufficient statistic)**.** *Consider a statistical model $\mathcal{P}$ for the* observation variable *X. Let T be a statistic valued in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with $d \geq 1$. We say that T is a* sufficient statistic *for the model $\mathcal{P}$ if for all $P \in \mathcal{P}$, the conditional distribution of X given T does not depend on P, that is, there exists a probability kernel Q on $\mathbb{R}^d \times \mathcal{X}$ such that, for all $P \in \mathcal{P}$, Q is a regular version of $P^{X|T}$.*

Intuitively, a sufficient statistic has the property to contain all the information available in X about P. More precisely it allows one to build unbiased estimators with reduced variances from any unbiased estimator. This is called the *Rao-blackwellization* of an estimator. Here is how it works.

Lemma 3.3.1. *Let S be a sufficient statistic associated to the Markov kernel Q and let $T = g(X)$ be an unbiased estimator of the parameter $t(P^X)$ (both real valued). Define*

$$T^R = \int g(x)\, Q(S, \mathrm{d}x) \, .$$

Then T^R is an unbiased estimator of the parameter t and its variance is smaller than that of T. As a consequence we have, for all $P \in \mathcal{P}$,

$$\mathrm{MSE}(T^R, P) \leq \mathrm{MSE}(T, P) \, .$$

Except in the case where X takes its values in a countable set, computing the conditional distribution of X given a function of X can be pretty intricate since we cannot rely on a joint density with respect to a product probability measure and thus conditional densities cannot be used. Hopefully, for dominated models checking that a statistic is sufficient turns out to be a trivial thing thanks to the following powerful theorem.

Theorem 3.3.2 (Fisher Factorization theorem)**.** *Suppose that the observation space is given by $\mathsf{X} = \mathbb{R}^n$ and $\mathcal{X} = \mathcal{B}(\mathbb{R}^n)$. Let $\nu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Consider a ν-dominated model $\mathcal{P}$ for the observation variable X and let $S = g(X)$ be a d-dimensional statistic. Then S is a sufficient statistic for the model $\mathcal{P}$ if and only if there exists a non-negative Borel function h on the observation space X such that for all $P \in \mathcal{P}$, there exists a Borel function $f_P : \mathbb{R}^d \to \mathbb{R}_+$ such that*

$$\frac{\mathrm{d}P^X}{\mathrm{d}\nu} = h \times f_P \circ g \, . \tag{3.3}$$

Proof. Suppose that the factorization (3.3) holds for all $P \in \mathcal{P}$. We first prove that $g(X)$ is a sufficient statistic in the case where

$$p := \int h \, \mathrm{d}\nu < \infty \, . \tag{3.4}$$

In this case we can define $Q \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ by

$$Q(A) = p^{-1} \int_A h \, \mathrm{d}\nu \quad \text{for all } A \in \mathcal{X} \, .$$

Let now $B \in \mathcal{X}$ and $A \in \mathcal{B}(\mathbb{R}^d)$. We have

$$\begin{aligned} P[X \in B, S \in A] &= \int_{g^{-1}(A)\cap B} h \times f_P \circ g \, \mathrm{d}\nu \\ &= p\mathbb{E}_Q \left[\mathbb{1}_{g^{-1}(A)\cap B} f_P \circ g \right] , \end{aligned}$$

where $\mathbb{E}_Q$ denotes the expectation under Q. Then, since $\mathbb{1}_{g^{-1}(A)} f_P \circ g$ is $\sigma(g)$-measurable and $\mathbb{E}_Q \left[\mathbb{1}_{g^{-1}(A)\cap B} f_P \circ g \right] \leq p^{-1} < \infty$, we can write

$$\mathbb{E}_Q \left[\mathbb{1}_{g^{-1}(A)\cap B} \times f_P \circ g \right] = \mathbb{E}_Q \left[\mathbb{1}_{g^{-1}(A)} \times f_P \circ g \times \mathbb{E}_Q[\mathbb{1}_B|g] \right] .$$

With the previous display, we thus get that

$$P[X \in B, S \in A] = p\mathbb{E}_Q \left[\mathbb{1}_{g^{-1}(A)} \times f_P \circ g \times \mathbb{E}_Q[\mathbb{1}_B|g] \right] .$$

Let $K : \mathbb{R}^d \times \mathcal{X} \to [0,1]$ be a Markov kernel providing a regular version of the conditional probability of the identity variable given g so that

$$\mathbb{E}_Q[\mathbb{1}_B|g] = K(g, B) \quad Q\text{-a.s.} .$$

Hence we get that

$$\begin{aligned} P[X \in B, S \in A] &= \int_{g^{-1}(A)} h \times f_P \circ g K[g, B] \, \mathrm{d}\nu \\ &= \mathbb{E}_P \left[\mathbb{1}_A(S) \times K[S, B] \right] , \end{aligned}$$

Since this holds for all $A \in \mathcal{B}(\mathbb{R}^d)$, it shows that

$$P(X \in B \,|\, S) = K[S, B] .$$

And since in turn this holds for all $B \in \mathcal{X}$, it shows that the conditional distribution of X given S is given by the Markov kernel K. Observe that we defined the later without using P, since Q is defined from h and ν only. Hence S is a sufficient statistic.

We now need to consider the case where (3.4) does not hold. It is sufficient to show that the conditional distribution of X given S is the same for any two arbitrary probabilities in $\mathcal{P}$. We may thus assume that $\mathcal{P}$ only contains 2 probabilities, $\mathcal{P} = \{P_0, P_1\}$. Then we can write for $i = 0, 1$,

$$h \times f_{P_i} \circ g = \tilde{h} \times \tilde{f}_{P_i} \circ g ,$$

where $\tilde{h} = h \times (f_{P_0} \vee f_{P_1}) \circ g$ and $\tilde{f}_{P_i} = f_{P_i}/(f_{P_0} \vee f_{P_1})$ with the convention $0/0 = 0$. Now we have

$$\int \tilde{h} \, \mathrm{d}\nu \leq \int h \times f_{P_0} \circ g \, \mathrm{d}\nu + \int h \times f_{P_i} \circ g \, \mathrm{d}\nu = 2 .$$

Thus, we can apply the previous case to conclude that the conditional distribution of X given S is the same for both probabilities in $\mathcal{P}$.

Let us now prove the necessity of the given condition. By Lemma 3.1.1, we can set, without loss of generality,

$$\nu = \sum_{n \geq 1} 2^{-n} P_n^X , \tag{3.5}$$

for some countable collection $(P_n)_{n\geq 1}$ of probabilities in $\mathcal{P}$. Suppose now that S is a sufficient statistic. So there exists a Markov kernel K providing a regular conditional distribution of X given S under any $P \in \mathcal{P}$. It is then easy to show that the Markov kernel K also provides a regular conditional distribution of the identity variable given g under ν defined by (3.5). Now, for all $B \in \mathcal{X}$, we have

$$P(X \in B) = \mathbb{E}_P[P(X \in B \,|\, S)] = \mathbb{E}_P[K(S, B)] = \int K(s, B)\, P^S(\mathrm{d}s) \ .$$

Let f_P denote the density of P^S with respect to ν^g. The previous display then reads

$$\begin{aligned} P(X \in B) &= \int K(s, B)\, f_P(s)\, \nu^g(\mathrm{d}s) \\ &= \mathbb{E}_\nu \left[K(g, B)\, f_P(g)\right] \ . \end{aligned} \tag{3.6}$$

Since the Markov kernel K is a regular version of the conditional distribution of the identity given g under ν, we can write

$$\mathbb{E}_\nu \left[K(g, B)\, f_P(g)\right] = \mathbb{E}_\nu \left[\mathbb{E}_\nu[\mathbb{1}_B|g]\, f_P(g)\right] = \mathbb{E}_\nu \left[\mathbb{1}_B f_P(g)\right] \ .$$

So, finally, inserting this in (3.6), we get that, for all $B \in \mathcal{X}$,

$$P(X \in B) = \int_B f_P \circ g(x)\ \mathrm{d}\nu \ ,$$

hence P^X admits the density $f_P \circ g$ with respect to ν. □

3.4 Likelihood function

Many methods in statistics rely on the following notion.

Definition 3.4.1 (Likelihood function)**.** *Consider a ν-dominated model $\mathcal{P}$ for the observation variable X valued in $(\mathsf{X}, \mathcal{X})$. For all $P \in \mathcal{P}$, let us denote by f_P the density of P^X with respect to ν. The likelihood function is defined as $P \mapsto f_P \circ X$ on $P \in \mathcal{P}$.*

The likelihood function can alternatively be seen as a collection of *statistics* $(f_P \circ X)_{P\in\mathcal{P}}$. It is indeed very important not to be misled by the use of P here and to understand that the likelihood function does not rely on the distribution of X, which is often called the *true distribution* and denoted by P_* in order to distinguish it from the P of $P \mapsto f_P \circ X$. Defining a likelihood function does not even require that $X \sim P_*^X$ for some $P_* \in \mathcal{P}$, that is, one can define a likelihood function associated to a model $\mathcal{P}$ without assuming that X indeed is drawn according to the model $\mathcal{P}$. When the distribution P_*^X of X is known to be out of $\{P^X, P \in \mathcal{P}\}$, we often call the likelihood function associated to $\mathcal{P}$ a *quasi-likelihood* function.

The likelihood function plays a central role in parametric estimation or hypothesis testing and can also be used in nonparametric statistics. It provides a way to control how close a given P is from the true distribution P_* of X. Suppose indeed that $X \sim P_*^X$ and that P_*^X admits a density f_* with respect to ν. Then we have, for all $P_1, P_2 \in \mathcal{P}$,

$$f_{P_1}(X) \geq f_{P_2}(X) \Leftrightarrow \ln f_{P_1}(X) \geq \ln f_{P_2}(X) \Leftrightarrow \ln \frac{f_*(X)}{f_{P_1}(X)} - \ln \frac{f_*(X)}{f_{P_2}(X)} \leq 0 \ ,$$

where the last equivalence hods P_*-a.s. since it holds whenever $f_*(X) > 0$. The expectation of the log difference turns out to be equal to the difference of the Kullback-Leibler divergences

$$\mathrm{KL}\left(P_*^X \,\|\, P_1^X\right) - \mathrm{KL}\left(P_*^X \,\|\, P_2^X\right) .$$

The basic idea for using likelihood function in statistics is that, for any P_1, $P_2 \in \mathcal{P}$,

(L-KL) $\ln(f_*(X)/f_{P_1}(X)) - \ln(f_*(X)/f_{P_2}(X))$ has the same sign as its expectation $\mathrm{KL}(P_* \,\|\, P_1) - \mathrm{KL}(P_* \,\|\, P_2)$ with high probability.

So if one believes in this principle, comparing $f_{P_1}(X)$ and $f_{P_2}(X)$ provides an indication about which P^X among P_1^X, P_2^X is closer to P_*^X in the sense of the Kullback-Leibler divergence. More precisely, $f_{P_1}(X) \geq f_{P_2}(X)$ is an *indication* that $\mathrm{KL}\left(P_*^X \,\|\, P_1^X\right) \leq \mathrm{KL}\left(P_*^X \,\|\, P_2^X\right)$.

Remark 3.4.1. *Interestingly, we note that if one has a sufficient statistic $S = g(X)$, by the Fisher factorization theorem (Theorem 3.3.2), to compare $f_{P_1}(X)$ and $f_{P_2}(X)$, we only need to observe S.*

Let us now consider the case where $\mathcal{P}$ is a parametric model,

$$\mathcal{P} = (P_\theta)_{\theta\in\Theta} \quad \text{with } \Theta \subseteq \mathbb{R}^p,$$

for some dimension $p \geq 1$. In this case we define the likelihood function directly on Θ, $\theta \mapsto f_\theta \circ X$, where f_θ denotes the density of P_θ with respect to ν. The likelihood is now defined on a subset of $\mathbb{R}^p$, rather than a subset of probability measures, which is more convenient for performing simple tasks such as trying to find its maximum by differentiating. In this context, the following estimator is of primary importance in statistics.

Definition 3.4.2 (Maximum likelihood estimator (MLE))**.** *A statistic $\hat{\theta}_n$ valued in Θ such that*

$$f_{\hat{\theta}_n} \circ X = \max_{\theta\in\Theta} f_\theta \circ X$$

is called a maximum likelihood estimator *(MLE). One equivalently writes*

$$\hat{\theta}_n \in \operatorname*{argmax}_{\theta\in\Theta} f_\theta \circ X .$$

If Θ is closed and the likelihood function is continuous then the existence of a MLE is guarantied, but not its uniqueness. The motivation for using an MLE clearly comes from Property (L-KL) which indicates that an MLE provides the parameter $\hat{\theta}_n$ such that Kullback-Leibler divergence between the true distribution of X and $P_{\hat{\theta}_n}$ should be approximately the smallest among all P_θ, $\theta \in \Theta$.

Let us conclude by briefly explaining why we expect Property (L-KL) to hold. To make this property unambiguous, one should precise what *with high probability* means. Adopting an asymptotic perspective, it could mean *with probability tending to 1* as the *number of observations* tend to infinity. The number of observations is a natural quantity in the case where $X = (X_1, \ldots, X_n)$ with n denoting the number of observations, $\nu = \mu^{\otimes n}$, $f_*(x_1, \ldots, x_n) = p_*(x_1) \ldots p_*(x_n)$ and, for all $P \in \mathcal{P}$, $f_P(x_1, \ldots, x_n) = p_P(x_1) \ldots p_P(x_n)$, that is $X_1, \ldots, X_n$ are i.i.d. under P with density p_P and under P_* with density p_*. In this setting $\mathcal{P}$ and X depend on n, but neither p_P nor p_* does. We can moreover write

$$\Delta_n := \frac{1}{n}\left(\ln\frac{f_*(X)}{f_{P_1}(X)} - \ln\frac{f_*(X)}{f_{P_2}(X)}\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\ln\frac{p_*(X_k)}{p_{P_1}(X_k)} - \ln\frac{p_*(X_k)}{p_{P_2}(X_k)}\right) .$$

As $n \to \infty$, that is, for a large number of observations, the law of large number shows that

$$\Delta_n \xrightarrow{\text{a.s.}} \mathrm{KL}\left(P_*^{X_1} \,\|\, P_1^{X_1}\right) - \mathrm{KL}\left(P_*^{X_1} \,\|\, P_1^{X_1}\right) .$$

On the other hand, we know from Exercise 2.14 that, for any n,

$$\frac{1}{n}\left(\mathrm{KL}\left(P_* \,\|\, P_1\right) - \mathrm{KL}\left(P_* \,\|\, P_2\right)\right) = \mathrm{KL}\left(P_*^{X_1} \,\|\, P_1^{X_1}\right) - \mathrm{KL}\left(P_*^{X_1} \,\|\, P_1^{X_1}\right) .$$

Hence, Property (L-KL) holds in the sense that, provided that $\mathrm{KL}\left(P_* \,\|\, P_1\right) - \mathrm{KL}\left(P_* \,\|\, P_2\right)$ is non-zero, its sign is the same as that of $\ln(f_*(X)/f_{P_1}(X)) - \ln(f_*(X)/f_{P_2}(X))$ with probability tending to one.

3.5 Statistical testing

3.5.1 General definition

Consider a partition $\{\mathcal{P}_0, \mathcal{P}_1\}$ of a statistical model $\mathcal{P}$ for the observation variable X. This allows us to define two *hypotheses* respectively called the null hypothesis and the alternative hypothesis.

($\mathbf{H}_0$) The observation variable X has distribution P^X with $P \in \mathcal{P}_0$.

($\mathbf{H}_1$) The observation variable X has distribution P^X with $P \in \mathcal{P}_1$.

We say that an hypothesis is *simple* if the corresponding set $\mathcal{P}_i$ reduces to one point. In this context one wishes to decide from the data which one of the two hypotheses is *preferable*.

Definition 3.5.1 (Statistical test)**.** *A statistical test is a statistic δ with values in $\{0, 1\}$. If $\delta = 0$ we say that we* accept *the null hypothesis ($\mathbf{H}_0$). Otherwise we say that we* reject *the null hypothesis ($\mathbf{H}_0$).*

To evaluate the performance of a test δ, two type of risks are considered.

1. The *first type* risk is defined as $P \mapsto P(\delta = 1)$ as $P \in \mathcal{P}_0$.
2. The *second type* risk is defined as $P \mapsto P(\delta = 0)$ as $P \in \mathcal{P}_1$.

Finally we call the *power* of δ the application $P \mapsto P(\delta = 1)$ as $P \in \mathcal{P}_1$.

It seems difficult to compare two statistical tests in general, say δ and δ', first because their risks vary with P and it is difficult to compare them for all P and second, even in the case of two simple hypotheses (only one P in each $\mathcal{P}_0$ and $\mathcal{P}_1$) because a test may have a small first type risk but a large second type risk (that is a small power) or *vice versa*.

The Neyman-Pearson approach for comparing tests is to restrict the comparison to the powers under a constraint on the first type risk.

Definition 3.5.2 (Level of a test)**.** *Let $\alpha \in [0, 1]$ We say that a test δ is of level α if*

$$\sup_{P \in \mathcal{P}_0} P(\delta = 1) \leq \alpha .$$

We say that δ is uniformly more powerful *than δ' for level α if both δ and δ' are of level α and for all $P \in \mathcal{P}_1$, we have*

$$P(\delta = 1) \geq P(\delta' = 1) .$$

3.5.2 Simple hypotheses

We here consider the case where $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1\}$, so that $\mathcal{P} = \{P_0, P_1\}$, that is, we only wish to point out one distribution among two possible ones.

Let f_0 and f_1 denote the densities of P_0^X and P_1^X with respect to a common dominating measure ν, say $\nu = P_0 + P_1$.

Definition 3.5.3 (Likelihood ratio test)**.** *The statistic*

$$T = \frac{f_1(X)}{f_0(X)}$$

is called the likelihood ratio *statistic. Let $t \in [0, \infty]$. The test defined by*

$$\delta = \begin{cases} 1 & \text{if } T \geq t \\ 0 & \text{otherwise} \end{cases}$$

is called the likelihood ratio test *with threshold t.*

It turns out that for two simple hypotheses, it is possible to determine *the* uniformly more powerful test for any given level α and this test relies on the likelihood ratio.

Theorem 3.5.1. *Consider two simple hypotheses $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1\}$ with P_0^X and P_1^X both dominated by ν. Denote by T the corresponding likelihood ratio. Let $t \in [0, \infty]$ and set*

$$\alpha_t = P_0(T \geq t) \, .$$

Then the likelihood ratio test with threshold t is uniformly more powerful than any other test δ' for the level α_t. Moreover if δ' is of level α_t and as powerful as δ then they coïncide on the set $\{T \neq t\}$ P_i-a.s. for $i = 0, 1$.

To have this result in a general fashion one needs to rely on *random tests* which we avoid here for simplicity. Instead we only restrict ourselves to levels α of the form α_t. Such restriction is inadequate only if the distribution function of T under P_0 only takes a very limited set of values, which does not happen often in usual examples.

Proof of Theorem 3.5.1. Note that $f_0(X) > 0$ P_0-a.s. hence T is well defined ($f_1(X)$ and $f_0(X)$ do not vanish at the same time) and valued in $[0, \infty)$ P_0-a.s. Similarly T is well defined and valued in $(0, \infty]$ P_1-a.s. In this proof, we let $\delta' = g(X)$ be a statistical test of level α_t.

Consider first the case $t = \infty$. Then δ is of level $\alpha_t = P_0(T = \infty) = 0$ and of power $P_1(T = \infty) = P_1(f_0(X) = 0)$. A test $\delta' = g \circ X$ is of level $\alpha_\infty = 0$ means that $\delta' = 0$ P_0-a.s., that is,

$$\int \mathbb{1}_{\{g=0\}} f_0 \, \mathrm{d}\nu = 1 = \int f_0 \, \mathrm{d}\nu \, .$$

Hence we have, since g is valued in $\{0, 1\}$, $\mathbb{1}_{\{g=1\}} f_0 = 0$ ν-a.s., which implies $\mathbb{1}_{\{g=1\}} \leq \mathbb{1}_{\{f_0=0\}}$ ν-a.s. It follows that

$$P_1(\delta' = 1) = \int \mathbb{1}_{\{g=1\}} f_1 \, \mathrm{d}\nu \leq \int \mathbb{1}_{\{f_0=0\}} f_1 \, \mathrm{d}\nu = P_1(\delta = 1) \, .$$

Moreover equality holds only if $\mathbb{1}_{\{g=1\}} = \mathbb{1}_{\{f_0=0\}}$ P_1^X-a.s., that is, $\delta' = \delta$ P_1-a.s. We already knew that $\delta' = \delta$ P_0-a.s. as a consequence of the level constraint, so the case $t = \infty$ is completed.

Now, it remains to consider the case $t \in [0, \infty)$. In this case we have

$$T \geq t \iff f_1(X) \geq t\, f_0(X) \quad P_0\text{-a.s. and } P_1\text{-a.s.}$$

Hence if we set, for all $x \in \mathsf{X}$,

$$g_0(x) = \begin{cases} 1 & \text{if } t\, f_0(x) \leq f_1(x) \\ 0 & \text{otherwise,} \end{cases}$$

then we have

$$\delta = g_0(X) \quad P_0\text{-a.s. and } P_1\text{-a.s.}$$

Let $\delta' = g \circ X$ be statistical test. For all $t \geq 0$, we have that

$$\begin{aligned} t\, P_0(\delta' = 1) + P_1(\delta' = 0) &= \int \left(t \mathbb{1}_{\{g(x)=1\}}\, f_0(x) + \mathbb{1}_{\{g(x)=0\}}\, f_1(x) \right) \nu(\mathrm{d}x) \\ &\stackrel{\text{def}}{=} K(g)\ . \end{aligned}$$

Since g takes its values in $\{0, 1\}$ we see that

$$K(g) \geq K(g_0)\ ,$$

and moreover we have

$$K(g) = K(g_0) \iff \nu\left(\{g \neq g_0\} \cap \{t\, f_0 \neq f_1\}\right) = 0\ .$$

We deduce that

$$\begin{aligned} P_1(\delta' = 0) &= K(g) - t\, P_0(\delta' = 1) \\ &\geq K(g_0) - t\, P_0(\delta' = 1) \\ &\geq K(g_0) - t\, \alpha_t \\ &= t\, P_0(\delta = 1) + P_1(\delta = 0) - t\, \alpha_t \\ &= P_1(\delta = 0)\ . \end{aligned}$$

Hence we get that δ is more powerful than δ' and if it is as powerful as δ' then necessarily $K(g) = K(g_0)$ and thus

$$\nu\left(\{g \neq g_0\} \cap \{t\, f_0 \neq f_1\}\right) = 0\ .$$

This implies that δ and δ' coincide on $\{T \neq t\}$ P_i-a.s. for $i = 0, 1$ and the proof is concluded. □

3.5.3 Monotone hypotheses

(...)

3.6 EM algorithm

Let us consider a parametric model, $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$, with $\Theta \subseteq \mathbb{R}^p$ for some $p \geq 1$ and suppose that one wishes to compute the associated MLE $\hat{\theta}_n$. Even in the case where the likelihood function $f_\theta \circ X$ can be efficiently computed at each θ, it can be difficult to maximize it over $\theta \in \Theta$. Usual algorithms are based on gradient descent algorithm. Such a procedure is iterative and without strong convexity assumptions, each step of a gradient descent does not guaranty an increase of the likelihood. The EM algorithm, introduced in [2], does have this nice property. We here briefly describe this algorithm because it can be quite interesting and practical for a large class of models.

The EM (*Expectation-Minimization*) algorithm applies in the context of *hidden variables* or *partly observed data*. In such models, we suppose that we can introduce an additional variable as follows.

(H-EM-1) Let U be a measurable mapping defined on $(\Omega, \mathcal{F})$ and valued in $(\mathrm{U}, \mathcal{U})$ such that under any $P_\theta \in \mathcal{P}$, the variable $Z = (X, U)$ admits a density $\bar{f}_\theta$ with respect to the dominating measure $\nu \otimes \mu$, where $\mu \in \mathbb{M}_+(\mathrm{U}, \mathcal{U})$ does not depend on P.

Thus, we can defined a *joint likelihood* function $\theta \mapsto \bar{f}_\theta \circ Z$ defined on the parameter space Θ. The advantage in adding the variable U only occurs when this joint likelihood is easier to compute and to maximize.

However this variable is *unobserved*, one also say that U is a *hidden variable*. Hence the joint likelihood cannot be computed from the data and appears to be useless in practice. We do not indeed use the joint likelihood directly and instead rely on the following quantity. For all $\theta, \theta' \in \Theta$, define

$$\mathcal{Q}(x; \theta\,; \theta') \stackrel{\text{def}}{=} -\int \ln \tilde{f}_\theta(x, u)\; g_{\theta'}(u|x)\; \mu(\mathrm{d}u) \tag{3.7}$$

where

$$g_{\theta'}(u|x) = \frac{\tilde{f}_{\theta'}(x, u)}{f_{\theta'}(x)} \quad x \in \mathsf{X},\ u \in \mathrm{U}\ .$$

Recall that by (H-EM-1), $g_{\theta'}(u|x)$ defines a conditional density of U given X under $P_{\theta'} \in \mathcal{P}$, which, as a function of u, is a density with respect to μ for all $x \in \mathsf{X}$. Moreover these densities define the conditional probability of U given X. (See Theorem 2.2.2). Hence if

(H-EM-2) for all $\theta, \theta' \in \Theta$, $\mathbb{E}_{\theta'}\left[\left|\ln \tilde{f}_\theta \circ Z\right|\right] < \infty$,

we have

$$\mathcal{Q}(X; \theta\,; \theta') = -\mathbb{E}_{\theta'}\left[\ln \tilde{f}_\theta(X, U)\,\middle|\, X\right] \quad P_{\theta'}\text{-a.s.}$$

Here $\mathbb{E}_{\theta'}$ denotes the expectation or conditional expectation under $P_{\theta'}$.

Since $\mathcal{Q}(X; \theta\,; \theta')$ is by definition $\sigma(X)$-measurable, it is a statistic which can be computed from the data.

The EM algorithm is then given as follows, assuming that the argmin below is well defined

and easy to compute.

Algorithm 2: EM algorithm.

Data: Observation X, initial estimate θ_0.
Result: Numerical approximation of the MLE $\widehat{\theta}_n$.
Initialization: Set $k = 0$.
repeat
  Set

$$\theta_{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{Q}_n(X; \theta\,; \theta_k)\ . \tag{3.8}$$

  Increment k ($k \leftarrow k+1$).
until $\ln f_{\theta_k}(X)$ *and* $\ln f_{\theta_{k-1}}(X)$ *are "relatively close"*;

Remark 3.6.1. *The definition of $\mathcal{Q}(x;\theta\,;\theta')$ in (3.7) requires some care since $g_{\theta'}(u|x)$ is only well defined if $f_{\theta'}(x) > 0$ and even if $f_{\theta'}(x) > 0$ the integral may not be defined. To deal with the latter point, one can alternatively define $\mathcal{Q}(x;\theta\,;\theta')$ as*

$$\mathcal{Q}(x;\theta\,;\theta') \stackrel{\text{def}}{=} \int_{\{g_{\theta'}(u|x)>0\}} \ln \frac{\tilde{f}_{\theta'}(x,u)}{\tilde{f}_{\theta}(x,u)}\ g_{\theta'}(u|x)\ \mu(\mathrm{d}u)\ . \tag{3.9}$$

Then, whenever $\mathcal{Q}(x;\theta\,;\theta')$ defined by (3.7) is finite, then so is the one defined by (3.9) and the output of Algorithm 2 remains unchanged. Moreover we prove hereafter that this $\mathcal{Q}(x;\theta\,;\theta')$ is well defined whenever $f_{\theta'}(x), f_{\theta}(x) > 0$.

We have the following result which shows that the likelihood associated to the sequence $(\theta_n)_{n\in\mathbb{N}}$ defined iteratively by (3.8) is non-decreasing.

Lemma 3.6.1. *Suppose that Assumption (H-EM-1) holds and define $\mathcal{Q}(x;\theta\,;\theta')$ by (3.9). Let $\theta, \theta' \in \Theta$ and $x \in \mathsf{X}$ such that $f_{\theta'}(x), f_{\theta}(x) > 0$. Then $\mathcal{Q}(x;\theta\,;\theta')$ is well defined and if*

$$\mathcal{Q}(x;\theta\,;\theta') \geq \mathcal{Q}(x;\theta'\,;\theta')\ ,$$

then we have

$$f_{\theta}(x) \geq f_{\theta'}(x)\ .$$

Proof. We have $\mathcal{Q}(x;\theta'\,;\theta') = 0$ by definition (3.9) and since $f_{\theta'}(x), f_{\theta}(x) > 0$,

$$\mathcal{Q}(x;\theta\,;\theta') = \ln \frac{f_{\theta'}(x)}{f_{\theta}(x)} + \int_{\{g_{\theta'}(u|x)>0\}} \ln \frac{g_{\theta'}(u|x)}{g_{\theta}(u|x)}\ g_{\theta'}(u|x)\ \mu(\mathrm{d}u)\ .$$

Since $f_{\theta'}(x) > 0$, we have that $u \mapsto g(u|x)$ is a density with respect to μ. Hence the above integral can be interpreted as a Kullback-Leibler divergence and is thus non-negative. Hence the result. □

Of course, in practice, the EM algorithm is used if both the right-hand sides of (3.7) (the *Expectation step*) and (3.8) (the *Minimization step*) are easy to compute numerically.

3.7 Fisher information matrix

In this section we consider a parametric ν-dominated model $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$ for the observation variable X valued in $(\mathsf{X}, \mathcal{X})$, and denote by f_θ the density of P_θ with respect to ν. We assume in the following that Θ is an open subset of $\mathbb{R}^d$ and denote by

$$\|f\| := \left(\int_{\mathsf{X}} |f(x)|^2 \, \nu(\mathrm{d}x)\right)^{\frac{1}{2}},$$

the norm of the Hilbert space $L^2(\mathsf{X}, \mathcal{X}, \nu)$. Observe that $\xi_\theta := \sqrt{f_\theta}$ is in $L^2(\mathsf{X}, \mathcal{X}, \nu)$ for all $\theta \in \Theta$.

Definition 3.7.1 (Hellinger differentiability)**.** *We say that $\mathcal{P}$ is Hellinger differentiable at θ if the mapping $\theta' \mapsto \xi_\theta$ defined from $\Theta \to L^2(\mathsf{X}, \mathcal{X}, \nu)$ admits a derivative at θ, that is, there exists a (unique) $\dot\xi_\theta \in (L^2(\mathsf{X}, \mathcal{X}, \nu))^d$ such that*

$$\lim_{\theta'\to\theta} \frac{1}{|\theta' - \theta|} \left\|\xi_{\theta'} - \xi_\theta - \dot\xi_\theta^T(\theta' - \theta)\right\| = 0.$$

We now give two lemmas allowing us to verify Hellinger differentiability.

Lemma 3.7.1. *Let $\theta \in \Theta$ and $V \subset \Theta$ be a neighborhood of θ. Suppose that for ν-a.e. x and all $\theta' \in V$, we can write*

$$\xi_{\theta'}(x) = \xi_\theta(x) + \int_{t=0}^1 g_{t\theta'+(1-t)\theta}^T(x)(\theta' - \theta)\, dt\,, \tag{3.10}$$

where, for all $x \in \mathsf{X}$, g satisfies one of the two following assertions.

(i) We have $\lim_{\epsilon\downarrow 0} \left\| \sup_{|\theta'-\theta|\le\epsilon} |g_{\theta'} - g_\theta| \right\| = 0\,.$

(ii) For ν-a.e. x, $\theta' \mapsto g_{\theta'}(x)$ is continuous at θ and there exists $\epsilon > 0$ such that

$$\left\| \sup_{|\theta'-\theta|\le\epsilon} |g_{\theta'}| \right\| < \infty\,.$$

Then $\mathcal{P}$ is Hellinger differentiable at θ with derivative g_θ,

$$\lim_{\theta'\to\theta} \frac{1}{|\theta' - \theta|} \left\|\xi_{\theta'} - \xi_\theta - g_\theta^T(\theta' - \theta)\right\| = 0.$$

Proof. Note that, by dominated convergence, we have that (ii) implies (i).

Now suppose that (i) holds. By (3.10), we have, for ν-a.e. x and all $\theta' \in V$,

$$\begin{aligned}\left|\xi_{\theta'}(x) - \xi_\theta(x) - g_\theta^T(x)(\theta' - \theta)\right| &= \left|\int_{t=0}^1 \{g_{t\theta'+(1-t)\theta}(x) - g_\theta(x)\}^T(\theta' - \theta)\, dt\right| \\ &\le |\theta' - \theta| \sup_{|\theta''-\theta|\le\epsilon} |g_{\theta''} - g_\theta|\,,\end{aligned}$$

where we set $\epsilon = |\theta' - \theta|$. Hence the result by using (i). □

A second lemma is useful to relate the Hellinger derivative to the more usual *score* function defined as the derivative of $\theta \mapsto \ln f_\theta(X)$.

Lemma 3.7.2. *Suppose that the set*

$$A := \{f_\theta > 0\}$$

does not depend on θ and for all $x \in A$, $\theta \mapsto \ln f_\theta(x)$ is continuously differentiable on Θ with derivative $\theta \mapsto \dot{\ell}_\theta(x)$ and. Suppose moreover that for all $\theta \in \Theta$ there exists a neighborhood V of θ such that

$$\int \sup_{\theta' \in V} \left(|\dot{\ell}_\theta(x)|^2 \, f_\theta(x) \right) \nu(\mathrm{d}x) < \infty \, .$$

Then $\mathcal{P}$ is Hellinger differentiable with Hellinger derivative given by

$$\dot{\xi}_\theta(x) = \frac{1}{2} \dot{\ell}_\theta(x) \, \xi_\theta(x) \, \mathbb{1}_A(x), \quad x \in \mathsf{X} \, . \tag{3.11}$$

Proof. For all $x \in A$, $\theta \mapsto f_\theta(x)$ is continuously differentiable on Θ and so is $\theta \mapsto \xi_\theta(x)$ and denoting the derivatives by $\dot{f}_\theta(x)$ and $\dot{\xi}_\theta(x)$, respectively, we have

$$\dot{\xi}_\theta(x) = \frac{1}{2} \frac{\dot{f}_\theta(x)}{f_\theta^{1/2}(x)} = \frac{1}{2} \frac{\dot{f}_\theta(x)}{f_\theta(x)} \xi_\theta(x) \quad \text{and} \quad \dot{\ell}_\theta(x) = \frac{\dot{f}_\theta(x)}{f_\theta(x)} \, .$$

We then define $\dot{\xi}_\theta(x) = 0$ for $x \notin A$ and the result follows by applying Lemma 3.7.1 with Condition (ii). □

Under the assumptions of Lemma 3.7.2, the function $\theta \mapsto \dot{\ell}_\theta(X)$ is called the *score function* (derivative of the log-likelihood function). Under these (rather heavy) additional assumptions one can show that the score function has zero expectation under P_θ, which reads as

$$\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = \int_A \dot{\ell}_\theta(x) \, f_\theta(x) \, \nu(\mathrm{d}x) = 0 \, .$$

Interestingly, using (3.11), this is equivalent to

$$\int_{\mathsf{X}} \dot{\xi}_\theta(x) \, \xi_\theta(x) \, \nu(\mathrm{d}x) = 0 \tag{3.12}$$

which is always true as soon as the weaker Hellinger differentiability condition holds, since it follows by taking the derivative in the following equation

$$\|\xi_\theta\|^2 = 1 \, .$$

Definition 3.7.2 (Fisher information matrix)**.** *Let $\mathcal{P}$ be Hellinger differentiable with Hellinger derivative $\dot{\xi}_\theta$. The Fisher information matrix is defined as*

$$\mathcal{I}(\theta) \stackrel{\text{def}}{=} 4 \int_{\mathsf{X}} \dot{\xi}_\theta(x) \dot{\xi}_\theta(x)^T \, \nu(\mathrm{d}x) \, .$$

When the conditions of Lemma 3.7.2 hold, we have

$$\mathcal{I}(\theta) = \int_A \left(\dot{\ell}_\theta(x)\right)^2 f_\theta(x)\,\nu(\mathrm{d}x) = \mathbb{E}_\theta[\left(\dot{\ell}_\theta(X)\right)^2],$$

which is the more widespread definition of the Fisher information matrix (but requires more assumptions).

Without additional assumptions, we can now state a lower bound on the variance of a statistic based on the Fisher information matrix.

Theorem 3.7.3. *Let $\mathcal{P}$ be Hellinger differentiable with Hellinger derivative $\dot{\xi}_\theta$. Let $T = g(X)$ be a scalar statistic such that, for some $\epsilon > 0$,*

$$\sup_{|\theta'-\theta|\leq\epsilon} \mathbb{E}_\theta[T^2] < \infty.$$

Define $\psi : \theta \mapsto \mathbb{E}_\theta[T]$. Then ψ is differentiable at θ and, if $\mathcal{I}(\theta)$ is positive definite, we have

$$\mathrm{Var}_\theta\,(T) \geq \dot{\psi}(\theta)^T \mathcal{I}(\theta)^{-1} \dot{\psi}(\theta), \tag{3.13}$$

where $\dot{\psi}(\theta)$ denotes the derivative of ψ at θ.

Proof. We first prove that ψ is differentiable at θ with derivative

$$\dot{\psi}(\theta) = 2\int_{\mathsf{X}} g(x)\,\dot{\xi}_\theta(x)\,\xi_\theta(x)\,\nu(\mathrm{d}x). \tag{3.14}$$

Note that this quantity is well defined since $\dot{\xi}_\theta \in (L^2(\mathsf{X},\mathcal{X},\nu))^d$ and $T\xi_\theta \in L^2(\mathsf{X},\mathcal{X},\nu)$. We need to prove that

$$\lim_{\theta'\to\theta} \frac{1}{|\theta'-\theta|} \left| \int_{\mathsf{X}} g(x)\,\left(\xi_{\theta'}^2(x) - \xi_\theta^2(x) - 2\,\xi_\theta(x)\,\dot{\xi}_\theta(x)^T(\theta'-\theta)\right)\,\nu(\mathrm{d}x)\right| = 0. \tag{3.15}$$

Define

$$r(x;\theta') = \xi_{\theta'}(x) - \xi_\theta(x) - \dot{\xi}_\theta(x)^T(\theta'-\theta), \quad x \in \mathsf{X}.$$

The previous integral reads as

$$\int_{\mathsf{X}} g(x)\,\left(\dot{\xi}_\theta(x)^T(\theta'-\theta)(\xi_{\theta'}(x)-\xi_\theta(x)) + r(x;\theta')(\xi_{\theta'}(x)+\xi_\theta(x))\right)\,\nu(\mathrm{d}x)$$

This integral is divided as the sum of three integrals appearing as follows, for all $m > 0$,

$$\left|\int_{\{g(x)\leq m\}} g(x)\,\left(\dot{\xi}_\theta(x)^T(\theta'-\theta)(\xi_{\theta'}(x)-\xi_\theta(x))\right)\,\nu(\mathrm{d}x)\right| \leq m\,|\theta'-\theta|\,\|\dot{\xi}_\theta\|\,\|\xi_{\theta'}-\xi_\theta\| \tag{3.16}$$

$$\left|\int_{\{g(x)>m\}} g(x)\,\left(\dot{\xi}_\theta(x)^T(\theta'-\theta)(\xi_{\theta'}(x)-\xi_\theta(x))\right)\,\nu(\mathrm{d}x)\right|$$

$$\leq |\theta'-\theta|\,(\|T\,\xi_{\theta'}\| + \|T\,\xi_\theta\|)\left(\int_{\{g(x)>m\}} |\dot{\xi}_\theta(x)|^2\,\nu(\mathrm{d}x)\right)^{\frac{1}{2}}, \tag{3.17}$$

$$\left|\int_{\mathsf{X}} g(x)\;r(x;\theta')(\xi_{\theta'}(x)+\xi_\theta(x))\,\nu(\mathrm{d}x)\right| \leq (\|T\,\xi_{\theta'}\| + \|T\,\xi_\theta\|)\,\|r(\cdot;\theta')\|, \tag{3.18}$$

where we used the Cauchy-Schwarz inequality. Since $\|\xi_{\theta'} - \xi_\theta\| = O(|\theta' - \theta|)$ as $\theta' \to \theta$, the line (3.16) is $o(|\theta' - \theta|)$ for $m = o(|\theta' - \theta|^{-1})$. The line (3.17) is $o(|\theta' - \theta|)$ as $m \to \infty$ since $|\dot{\xi}_\theta(x)|^2$ is integrable and $\|g\,\xi_{\theta'}\| = \mathbb{E}_{\theta'}[T^2] < \infty$ for θ' close enough to θ. Now the line (3.18) is asymptotically equivalent to $\|r(\cdot;\theta')\|$ in a neighborhood of θ, and thus is $o(|\theta' - \theta|)$ as $\theta' \to \theta$. We thus set $m(\theta') \to \infty$ with $m(\theta') = o(|\theta' - \theta|^{-1})$ as $\theta' \to \theta$ and get (3.15) and thus (3.14).

Using (3.12), the derivative in (3.14) can be written as

$$\dot{\psi}(\theta) = 2 \int_{\mathsf{X}} (g(x) - \mathbb{E}_\theta[T])\, \dot{\xi}_\theta(x)\, \xi_\theta(x)\, \nu(\mathrm{d}x) \,.$$

Then, by Cauchy-Schwarz Inequality, for all $\lambda \in \mathbb{R}^d$,

$$\begin{aligned} \left|\lambda^T \dot{\psi}(\theta)\right| &= 2 \left| \int_{\mathsf{X}} g(x)\, \lambda^T \dot{\xi}_\theta(x)\, \xi_\theta(x)\, \nu(\mathrm{d}x) \right| \\ &\le 2 \left\| \lambda^T \dot{\xi}_\theta \right\| \, \|(T - \mathbb{E}_\theta[T])\xi_\theta\| \\ &= \sqrt{\lambda^T \mathcal{I}(\theta) \lambda \ \mathrm{Var}_\theta\,(T)}. \end{aligned}$$

If $\mathcal{I}(\theta)$ is positive definite, we get the result by optimizing the bound into

$$\mathrm{Var}_\theta\,(T) \ge \sup_{\lambda \in \mathbb{R}^d,\, \lambda \neq 0} \frac{\left|\lambda^T \dot{\psi}(\theta)\right|^2}{\lambda^T \mathcal{I}(\theta) \lambda} = \sup \left\{ (\lambda^T \mathcal{I}(\theta) \lambda)^{-1} \ : \ \lambda^T \dot{\psi}(\theta) = 1 \right\}.$$

The λ that minimizes $\lambda^T \mathcal{I}(\theta)\lambda$ under the constraint $\lambda^T \dot{\psi}(\theta) = 1$ is proportional to $\mathcal{I}(\theta)^{-1} \dot{\psi}(\theta)$ and we easily derive (3.13). □

In Theorem 3.7.3, when T serves as an estimator of $\psi(\theta)$, then it is unbiased. Then since the lower bound given by (3.13) applies to any unbiased estimator of $\psi(\theta)$, if the bound is reached by T then T has the minimal variance among all possible unbiased estimator of $\psi(\theta)$. This is why the following definition is useful.

Definition 3.7.3 (Efficient estimator)**.** *Let T be as in Theorem 3.7.3. If an equality holds in (3.13) for all $\theta \in \Theta$, we say that T is an efficient estimator of $\psi(\theta)$.*

It is not always possible to find an efficient estimator.

Example 3.7.1 (Translated doubly exponential distribution)**.** *Consider the Lebesgue-dominated model defined by the collection of densities*

$$f_\theta(x) = \frac{1}{2} \exp(-|x - \theta|), \quad \theta \in \mathbb{R} \,.$$

Then for all x, $\theta \mapsto f_\theta(x)$ is not differentiable at $\theta = x$. However it is Hellinger differentiable by Lemma 3.7.1. Here we have

$$\xi_\theta(x) = \frac{1}{\sqrt{2}} \exp(-\frac{1}{2}|x - \theta|) \,,$$

which satisfies (3.10) by setting

$$g_\theta(x) = \frac{\mathrm{sgn}(x - \theta)}{2\sqrt{2}} \exp(-|x - \theta|/2) \,.$$

Since $\theta \mapsto g_\theta(x)$ is continuous on $\mathbb{R}\backslash\{x\}$ and $|g_\theta(x)| \leq \exp(-|x|/2)\exp(|\theta|/2)/2\sqrt{2}$, we obtain that Condition (ii) of Lemma 3.7.1 is satified. Hence the model is Hellinger differentiable at every $\theta \in \mathbb{R}$ and Theorem 3.7.3 applies with

$$\mathcal{I}(\theta) = \frac{1}{2}\int_{\mathbb{R}} \exp(-|x-\theta|)\,\mathrm{d}x = 1$$

We conclude with Exercise 3.7 that for all unbiased estimator $\hat{\theta}_n$ of θ from n i.i.d.observations we have

$$\mathrm{Var}_\theta\left(\hat{\theta}_n\right) \geq n^{-1}\ .$$

3.8 Exercises

Exercise 3.1. Let $X_1, \dots, X_n$ be n i.i.d. r.v.s with common density

$$p_\theta(x) = \exp(\theta - x)\mathbb{1}_{\{x \geq \theta\}}\ , \theta \in \mathbb{R}\ .$$

We want to estimate the translation parameter θ.

1. Propose an unbiased estimator $\hat{\theta}_n$ based on the empirical mean.
2. Compute the quadratic risk of $\hat{\theta}_n$.

Define the estimator $\tilde{\theta}_n$ by $\tilde{\theta}_n = \inf_{1 \leq i \leq n} X_i$.

3. What is the distribution of $\tilde{\theta}_n$?
4. Is the estimator $\tilde{\theta}_n$ unbiased ?
5. Compute the quadratic risk of $\tilde{\theta}_n$.
6. What is the best estimator according to these results?

Exercise 3.2. Let $X_1, \dots, X_n$ be n i.i.d. r.v.'s with distribution $\mathbf{Ber}(\theta)$ (Bernouilli with parameter θ).

1. Show that $S_n = X_1 + \cdots + X_n$ is a sufficient statistic of the model.
2. Show that one can choose $\alpha > 0$ such that $\alpha(X_1 - X_2)^2$ is an unbiased estimator of the variance.
3. Compute the Rao Blackwellized estimator

$$Z = \mathbb{E}_\theta[\alpha(X_1 - X_2)^2 | S_n].$$

Exercise 3.3. Let $X_1, \dots, X_n$ n i.i.d. r.v.s with distribution $\mathbf{Pn}(\theta)$ (Poisson distribution with parameter $\theta > 0$) defined by

$$\boldsymbol{P}_\theta\{X_1 = k\} = \frac{\theta^k}{k!} e^{-\theta}, \quad k \in \mathbb{N}.$$

Define $S_n = \sum_{i=1}^n X_i$.

1. Show that if $Y_1 \sim \mathbf{Pn}(\theta_1)$, $Y_2 \sim \mathbf{Pn}(\theta_2)$ and Y_1, Y_2 are independent, then $Y_1 + Y_2 \sim \mathbf{Pn}(\theta_1 + \theta_2)$.
2. Show that S_n is a sufficient statistic.

Let $x_0 \in \mathbb{N}$ and $X \sim \mathbf{Pn}(\theta)$. We want to estimate $\boldsymbol{P}_\theta(X \geq x_0)$ from $X_1, \dots, X_n$. We consider the Rao-Blackwellized version of $\mathbb{1}(X_1 \geq x_0)$, that is

$$\widehat{P}_{x_0} := \mathbb{E}_\theta[\mathbb{1}(X_1 \geq x_0) | S_n].$$

3. Compute $\widehat{P}_{x_0}$ explicitly by using 1.

4. What is the result when $x_0 = 0$ and when $x_0 = 1$? Comment the result.

Exercise 3.4 (Neyman-Pearson test: Gaussian variable with known mean). 1. Let Y be a centered Gaussian vector of length n. We want to test the hypothesis H_1: $Y \sim \mathcal{N}(0, \Sigma_1)$ versus H_0: $Y \sim \mathcal{N}(0, \Sigma_0)$ where Σ_0, Σ_1 are invertible covariance matrices. Show that the Neyman-Pearson test consists in comparing $Y^T(\Sigma_1^{-1} - \Sigma_0^{-1})Y$ to a threshold.

2. Let X and V be two centered Gaussian r.v.'s with respective variances σ_X^2 and σ_V^2. The variable X is a true signal of interest and V is a measure noise. We observe $Y = X + V$ in the form of n independent observations.

   Propose a statistical test at level α for detecting the presence of the signal X.

3. For the previous test, give the value of the threshold as a function of the quantiles of the χ^2 distributions.

Exercise 3.5 (Uniform distribution). Let $X_1, \ldots, X_n$ be n i.i.d. r.v.s with uniform distribution on $[0, \theta]$, $\theta > 0$.

1. Let $\theta_0 < \theta_1$. What is the form of the Neyman-Pearson test of H_0 : "$\theta = \theta_0$" against H_1 : "$\theta = \theta_1$".

2. Same question for $\theta_0 > \theta_1$.

3. Let $\theta_0 > 0$. Build a test at level α of H_0 : "$\theta = \theta_0$" against H_1 : "$\theta \neq \theta_0$" from the test statistic $\hat{\theta}_n = \sup_{1 \leq i \leq n} X_i$.

4. Compute the power function of the previous test.

Exercise 3.6 (Mixture model and EM algorithm). Let $\{g_\theta,\, \theta \in \Theta\}$ be a parametric collection of density functions with respect to a measure ν on $(\mathsf{X}, \mathcal{X})$. Suppose that for all θ, θ',

$$\int |\ln g_{\theta'}|\ g_\theta \, \mathrm{d}\nu < \infty \, .$$

Let K be a positive integer. The parametric model

$$\left\{ \sum_{k=1}^{K} \alpha_k g_{\theta_k}(x)\, \nu(dx),\ (\theta_1, \ldots, \theta_K) \in \Theta^K,\ (\alpha_1, \ldots, \alpha_K) \in [0,1]^K,\ \sum_{k=1}^{K} \alpha_k = 1 \right\} \tag{3.19}$$

is called a *micture model*. We will denote $\eta = (\theta_1, \ldots, \theta_K, \alpha_1, \ldots, \alpha_K)$ the corresponding parameter and $\mathbb{P}_\eta$ the associated distribution. In the following we observe $X_1, \ldots, X_n$ i.i.d. with distribution $\mathbb{P}_\eta$.

1. Show that one can define i.i.d. *hidden variables* $Y_1, \ldots, Y_n$ valued in $\{1, \ldots, K\}$ such that for all $i = 1, \ldots, n$ and all $A \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}_\eta(X_i \in A \,|\, Y_i = k) = \int_A g_{\theta_k}(x)\, \nu(dx) \, .$$

2. Compute $\mathbb{P}_\eta(Y_1 = k)$ for all $k = 1, \ldots, K$.

3. Compute the joint density f_η of (X_1, Y_1) and the distribution of Y_1 given X_1 under $\mathbb{P}_\eta$.

The maximum likelihood estimator for the model (3.19) is denoted by $\hat{\eta} = (\hat{\theta_1}, \dots, \hat{\theta_K}, \hat{\alpha_1}, \dots, \hat{\alpha_K})$ and maximizes

$$\eta = (\theta_1, \dots, \theta_K, \alpha_1, \dots, \alpha_K) \mapsto \sum_{k=1}^{K} \alpha_k g_{\theta_k}(x).$$

Even in the cases where $\{g_\theta,\, \theta \in \Theta\}$ is a well known simple collection of density functions (e.g. Gaussian), the max is difficult to compute. We can rely on the hidden variables abiove in order to implement the EM algorithm. Define

$$\begin{aligned}\mathcal{Q}(X_{1:n}; \eta'\,; \eta) &= \mathbb{E}_\eta \left[\ln f_{\eta'}(X_1, \dots, X_n, Y_1, \dots, Y_n) \,|\, X_{1:n} \right] \\ &= \sum_{i=1}^{n} \mathbb{E}_\eta \left[\ln f_{\eta'}(X_i, Y_i) \,|\, X_i \right].\end{aligned}$$

On notera

$$\begin{aligned}P_{i,k}(\eta) &:= \frac{\alpha_k \, g_{\theta_k}(X_i)}{\sum_{j=1}^{K} \alpha_j \, g_{\theta_j}(X_i)}, \\ T_{n,k}(\eta) &:= \sum_{i=1}^{n} P_{i,k}(\eta), \\ S_{n,k}(\eta, \theta') &= \sum_{i=1}^{n} P_{i,k}(\eta) \, \ln(g_{\theta'_k}(X_i)).\end{aligned}$$

4. (**Etape E**) Compute $\mathcal{Q}(X_{1:n}; \eta'\,; \eta)$ explicitly.

5. (**Etape M**)

   (a) Compute $(\alpha_1^\star, \dots, \alpha_K^\star)$ that maximizes

   $$(\alpha'_1, \dots, \alpha'_K) \mapsto \mathcal{Q}(X_{1:n}; \eta' = (\theta'_1, \dots, \theta'_K, \alpha'_1, \dots, \alpha'_K)\,; \eta)$$

   for $\eta, (\theta'_1, \dots, \theta'_K)$ given.

   (b) Take $\Theta = (0, \infty)$ and g_θ as the Gaussian density with mean zero and variance θ. Compute $\eta^\star$ that maximizes

   $$\eta' \mapsto \mathcal{Q}(X_{1:n}; \eta'\,; \eta)\,.$$

Exercise 3.7. Let $\mathcal{P} = \{f_\theta(x)\, \nu(\mathrm{d}x),\, \theta \in \Theta\}$ and $\mathcal{Q} = \{g_\theta(x)\, \mu(\mathrm{d}x),\, \theta \in \Theta\}$ be Hellinger differentiable at θ with Fisher information matrices $\mathcal{I}(\theta)$ et $\mathcal{J}(\theta)$.

1. Show that $\{f_\theta(x) \otimes g_\theta(x)\ \nu(\mathrm{d}x) \otimes \mu(\mathrm{d}x),\, \theta \in \Theta\}$ is Hellinger differentiable at θ with Fisher information matrix $\mathcal{I}(\theta) + \mathcal{I}'(\theta)$.

2. Show that, for all $n \geq 1$, $\{\prod_{i=1}^{n} p_\theta(x_i)\, \mu^{\otimes n}(\mathrm{d}x),\, \theta \in \Theta\}$ is Hellinger differentiable at θ with Fisher information matrix $n \times \mathcal{I}(\theta)$.

3. Compare with Exercise 2.14.

Part II

Introduction to random processes

Chapter 4

Random processes: basic definitions

In this chapter, we introduce the basic foundations for stochastic modelling of time series such as random processes, stationary processes, Gaussian processes and finite distributions. We also provide some basic examples of real life time series.

4.1 Introduction

A time series is a sequence of observations x_t, each of them recorded at a time t. The time index can be discrete, in which case we will take $t \in \mathbb{N}$ or $\mathbb{Z}$ or can be continuous, $t \in \mathbb{R}$, $\mathbb{R}_+$ or $[0,1]$... Time series are encountered in various domains of application such as medical measurements, telecommunications, ecological data and econometrics. In some of these applications, spatial indexing of the data may also be of interest. Although we shall not consider this case in general, many aspects of the theory and tools introduced here can be adapted to spatial data.

In this course, we consider the observations as the realized values of a random process $(X_t)_{t\in T}$ as defined in Section 4.2. In other words, we will use a *stochastic modeling* approach of the data. Here are some examples which illustrate the various situations in which stochastic modelling of time series are of primary interest.

Example 4.1.1 (Heartbeats)**.** *Figure 4.1 displays the heart rate of a resting person over a period of* 900 *seconds. This rate is defined as the number of heartbeats per unit of time. Here the unit is the minute and is evaluated every* 0.5 *seconds.*

Example 4.1.2 (Internet traffic)**.** *Figure 4.2 displays the inter-arrival times of TCP packets, expressed in seconds, on the main link of Lawrence Livermore laboratory. This trace is obtained from a 2 hours record of the traffic going through this link. Over this period around 1.3 millions of packets have been recorded. Many traces are available on* The Internet Traffic Archive, *http://ita.ee.lbl.gov/*.

Example 4.1.3 (Speech audio data)**.** *Figure 4.3 displays a speech audio signal with a sampling frequency equal to 8000 Hz. This signal is a record of the unvoiced fricative phoneme* sh *(as in* sh*arp).*

Example 4.1.4 (Meteorological data)**.** *Figure 4.4 displays the daily record of the wind speed at the Kilkenny meteorological station.*

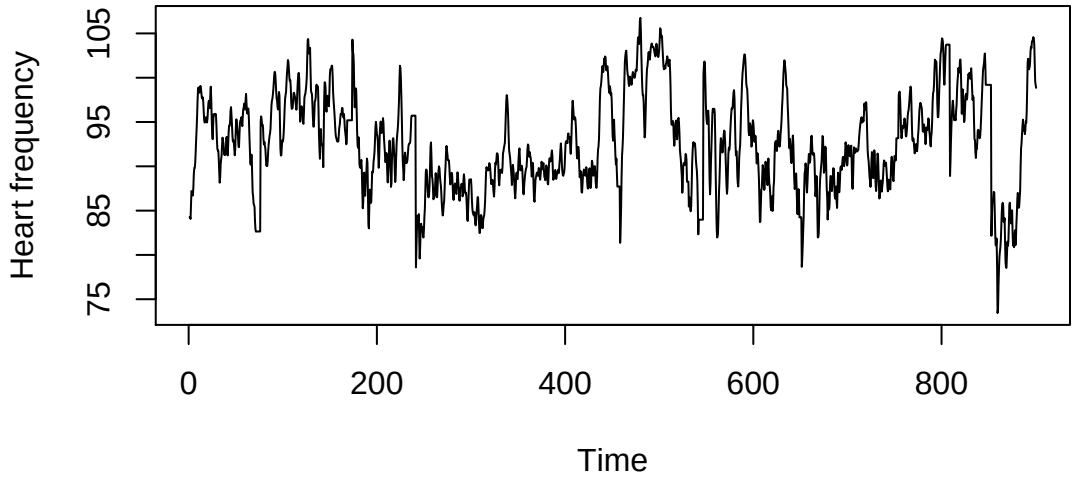

Figure 4.1: *Heartbeats: time evolution of the heart rate.*

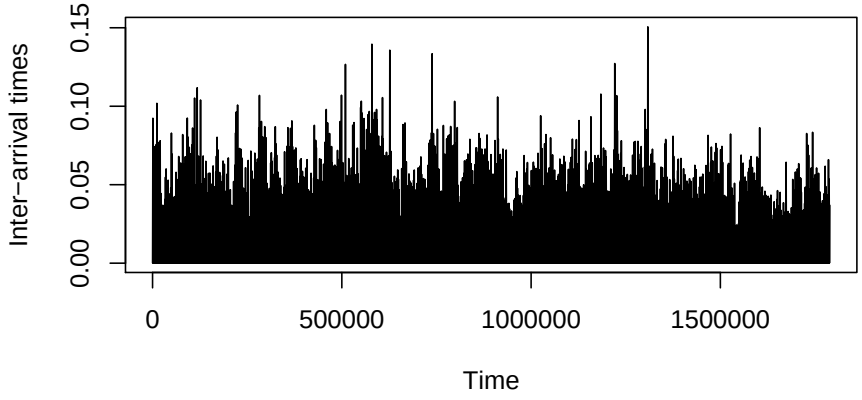

Figure 4.2: *Internet traffic trace : inter-arrival times of TCP packets.*

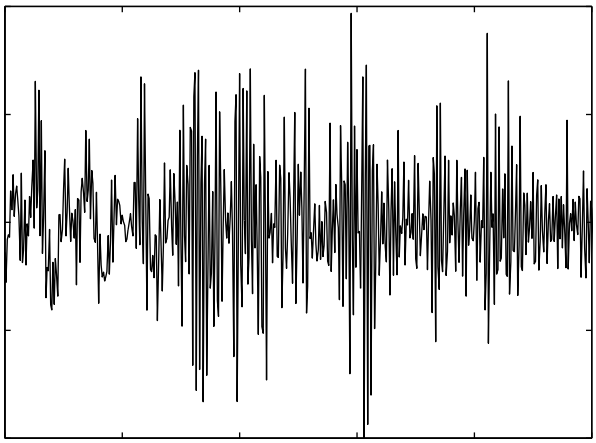

Figure 4.3: *A record of the unvoiced fricative phoneme* sh.

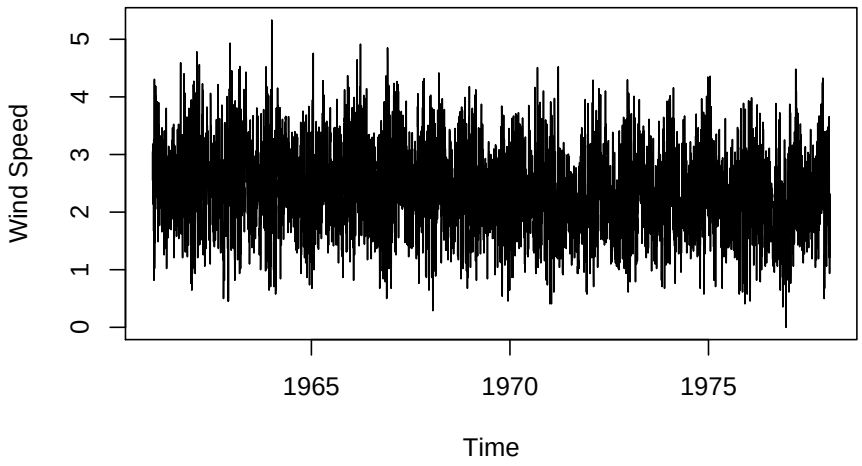

Figure 4.4: *Daily record of the wind speed at Kilkenny (Ireland).*

Example 4.1.5 (Financial index)**.** *Figure 4.5 displays the daily open value of the Standard and Poor 500 index. This index is computed as a weighted average of the stock prices of 500 companies traded at the New York Stock Exchange (NYSE) or NASDAQ. It is a widely used benchmark index which provides a good summary of the U.S. economy.*

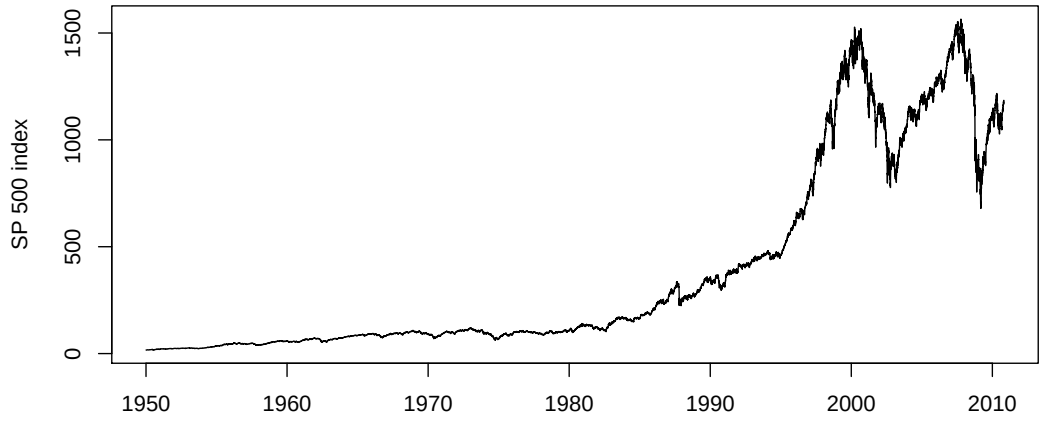

Figure 4.5: SP-500 stock index time series

4.2 Random processes

4.2.1 Definitions

In this section we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an index set T and a measurable space $(\mathsf{X}, \mathcal{X})$, called the *observation space*.

Definition 4.2.1 (Random process)**.** *A random process defined on $(\Omega, \mathcal{F}, \mathbb{P})$, indexed on T and valued in $(\mathsf{X}, \mathcal{X})$ is a collection $(X_t)_{t\in T}$ of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking there values in $(\mathsf{X}, \mathcal{X})$.*

The index t can for instance correspond to a time index, in which case $(X_t)_{t\in T}$ is a time series. When moreover $T = \mathbb{Z}$ or $\mathbb{N}$, we say that it is a *discrete time* process and when $T = \mathbb{R}$ or $\mathbb{R}_+$, it is a *continuous time* process. In the following, we shall mainly focus on discrete time processes with $T = \mathbb{Z}$. Concerning the space $(\mathsf{X}, \mathcal{X})$, we shall usually consider $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-field of $\mathbb{R}$), in which case we have a *real-valued process*, or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, in which case we have a *vector-valued process*, and in particular $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$, in which case we have a *complex-valued process*.

It is important to note that a random process can be seen as an application $X : \Omega \times T \to \mathsf{X}$, $(\omega, t) \mapsto X_t(\omega)$ such that, for each index $t \in T$, the function $\omega \mapsto X_t(\omega)$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathsf{X}, \mathcal{X})$.

Definition 4.2.2 (Path)**.** *For each $\omega \in \Omega$, the $T \to \mathsf{X}$ application $t \mapsto X_t(\omega)$ is called the* path *associated to the experiment ω.*

When $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$, it can be useful to associate a *filtration* to the process.

Definition 4.2.3 (Filtration)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$.*

(i) A filtration of a measurable space $(\Omega, \mathcal{F})$ is an increasing sequence $(\mathcal{F}_t)_{t\in T}$ of sub-σ-fields of $\mathcal{F}$. A filtered probability space $(\Omega,\ \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbb{P})$ is a probability space endowed with a filtration.

(ii) A random process $(X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be adapted to the filtration $(\mathcal{F}_t)_{t\in T}$ if for each $t \in T$, X_t is $\mathcal{F}_t$-measurable.

We will also directly say that $((X_t, \mathcal{F}_t))_{t\in T}$ is an adapted process to indicate that the process $(X_t)_{t\in T}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in T}$. The σ-field $\mathcal{F}_t$ can be thought of as the information available up to time t. Requiring the process to be adapted means that X_t can be computed using this available information.

Definition 4.2.4 (Natural filtration)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$. Let $(X_t)_{t\in T}$ be a random process. The* natural filtration *of the process $(X_t)_{t\in T}$ is the the smallest filtration with respect to which $(X)_{t\in T}$ is adapted,*

$$\mathcal{F}_t^X = \sigma(X_s :\ s\ \leq t)\ ,\quad t \in T\ .$$

By definition, a stochastic process is adapted to its natural filtration.

4.2.2 Finite dimensional distributions

Given two measurable spaces $(\mathsf{X}_1, \mathcal{X}_1)$ et $(\mathsf{X}_2, \mathcal{X}_2)$, one defines the product measurable space $(\mathsf{X}_1 \times \mathsf{X}_2, \mathcal{X}_1 \otimes \mathcal{X}_2)$ where $\times$ denotes the Cartesian product of sets and $\otimes$ the corresponding product for σ-field: $\mathcal{X}_1 \otimes \mathcal{X}_2$ is the smallest σ-field containing the set class $\{A_1 \times A_2, A_1 \in \mathcal{X}_1 : A_2 \in \mathcal{X}_2\}$, which will be written

$$\mathcal{X}_1 \otimes \mathcal{X}_2 = \sigma\left(A_1 \times A_2 : A_1 \in \mathcal{X}_1, A_2 \in \mathcal{X}_2\right)\ .$$

Since the set class $\{A_1 \times A_2 : A_1 \in \mathcal{X}_1, A_2 \in \mathcal{X}_2\}$ is stable under finite intersections, a probability measure on $\mathcal{X}_1 \otimes \mathcal{X}_2$ is uniquely defined by its restriction to this class by Theorem 2.0.2.

Similarly one defines a finite product measurable space $(\mathsf{X}_1 \times \cdots \times \mathsf{X}_n, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$ from n measurable spaces $(\mathsf{X}_t, \mathcal{X}_t)$, $t \in T$. We will also write $(\prod_{t\in T} \mathsf{X}_t, \bigotimes_{t\in T} \mathcal{X})$.

Let now $(X_t)_{t\in T}$ be random process $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$ and $I \in \mathcal{I}$, where $\mathcal{I}$ denotes the set of finite subsets of T. Let $\mathbb{P}^{X_I}$ denotes the probability distribution of the random vector $\{X_t, t \in I\}$, that is, the image measure of $\mathbb{P}$ defined on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ by

$$\mathbb{P}^{X_I}\left(\prod_{t\in I} A_t\right) = \mathbb{P}\left(X_t \in A_t,\, t \in I\right)\ , \tag{4.1}$$

where A_t, $t \in T$ are any sets of the σ-field $\mathcal{X}$. The probability measure $\mathbb{P}^{X_I}$ is a *finite dimensional* distribution.

Definition 4.2.5. *We call* finite dimensional distributions *or* fidi distributions *of the process X the collection of probability measures $\left(\mathbb{P}^{X_I}\right)_{I\in\mathcal{I}}$.*

If T is infinite, the definition of product σ-fields above is extended by considering the σ-field generated by the *cylinders* on the Cartesian product $\prod_{t\in T} \mathsf{X}_t$ defined as the set of T-indexed sequences $(x_t)_{t\in T}$ such that $x_t \in \mathsf{X}_t$ for all $t \in T$. Let us focus on the case where

$(\mathsf{X}_t, \mathcal{X}_t) = (\mathsf{X}, \mathcal{X})$ for all $t \in T$. Then $\mathsf{X}^T = \prod_{t\in T} \mathsf{X}$ is the set of sequences $(x_t)_{t\in T}$ such that $x_t \in \mathsf{X}$ for all $t \in T$ and

$$\mathcal{X}^{\otimes T} = \sigma\left(\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I} : I \in \mathcal{I}, \forall t \in I,\ A_t \in \mathcal{X}\right) . \tag{4.2}$$

Then we have the following theorem.

Theorem 4.2.1. *Let $(X_t)_{t\in T}$ be random process $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$. Then the mapping $X : \omega \mapsto (X_t(\omega))_{t\in T}$ is measurable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$. Moreover the push forward measure $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$ is entirely characterized by the collection of finite distributions $\left(\mathbb{P}^{X_I}\right)_{I\in\mathcal{I}}$.*

Proof. Let us denote

$$\mathcal{C} = \left\{\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I} : I \in \mathcal{I}, \forall t \in I,\ A_t \in \mathcal{X}\right\} .$$

This class of sets is a π-system and $\mathcal{X}^{\otimes T} = \sigma(\mathcal{C})$. Moreover it is easy to show that, for all $A \in \mathcal{C}$, $X^{-1}(A) \in \mathcal{F}$ since it is a finite intersection of sets of the form $X_t^{-1}(A_t)$ which are in $\mathcal{F}$. And for the same reason, $\mathbb{P}(X^{-1}(A))$ is determined by $\mathbb{P}^{X_I}$ for a well chosen finite set I. Hence by Theorem 2.0.2, $\mathbb{P}^X$ is uniquely determined by the collection of fidi distributions. □

This theorem thus shows that a process $(X_t)_{t\in T}$ can be seen as random variable $X = (X_t)_{t\in T}$ valued $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, whose *law* is determined by the fidi distributions.

Conversely, the fidi distributions can be obtain form $\mathbb{P}^X$ since the canonical projection Π_I of X^T on X^I defined by

$$\Pi_I(x) = (x_t)_{t\in I} \quad \text{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T , \tag{4.3}$$

is easily shown to be measurable from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ for all $I \subset T$. Hence, for all $I \subset T$,

$$\mathbb{P}^{(X_t)_{t\in I}} = \mathbb{P}^X \circ \Pi_I^{-1} .$$

Now, given a distribution $\mathbb{P}$ directly on the space of paths $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, it is also convenient to define a process on this measurable space, whose distribution will be given by $\mathbb{P}$. This construction is known as the *canonical process*.

Definition 4.2.6 (Canonical process)**.** *Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ the measurable space of corresponding paths. The canonical functions defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ is the collection of measurable functions $(\xi_t)_{t\in T}$ defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ and valued in $(\mathsf{X}, \mathcal{X})$ as $\xi_t(\omega) = \omega_t$ for all $\omega = (\omega_t)_{t\in t} \in \mathsf{X}^T$.*

When $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ is endowed with a probability measure $\mathbb{P}$, then the canonical process *$(\xi_t)_{t\in T}$ defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T}, \mathbb{P})$ has distribution $\mathbb{P}$.*

So far we have considered that the random process $X = (X_t)_{t\in T}$ is given as defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The only case where we have precised how all X_t are defined is in Definition 4.2.6, where we only need to have a probability $\mathbb{P}$ on the space $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to construct the whole process with the desired distribution. And we have seen that having this distribution is equivalent to have all the fidi distributions.

In practice, fidi distributions are much easier to deal with. Thus the following question arises:

Given a collection of distributions $(\nu_I)_{I\in\mathcal{I}}$ such that for all $I \in \mathcal{I}$, ν_I is a probability measure on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$, can we define a probability $\mathbb{P}$ on $(\mathsf{X}^T, \mathcal{X}^{\otimes T}$ such that $(\nu_I)_{I\in\mathcal{I}}$ are the fidi distributions associated to $\mathbb{P}$, $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$ for all $I \in \mathcal{I}$?

To answer this question. Suppose first that $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$ for all $I \in \mathcal{I}$. Let $J \subset I$ two finite subsets. Let us denote by $\Pi_{I,J}$ the canonical projection of X^I onto X^J defined by

$$\Pi_{I,J}[x] = (x_t)_{t\in J} \quad \text{for all} \quad x = (x_t)_{t\in I} \in \mathsf{X}^I \ . \tag{4.4}$$

The canonical projection is only preserving the vector entries that correspond to the indices of the subset J. Then, since $\Pi_J = \Pi_{I,J} \circ \Pi_I$, we get that $\mathbb{P} \circ \Pi_J^{-1} = \mathbb{P} \circ \Pi_I^{-1} \circ \Pi_{I,J}^{-1}$ and thus

$$\nu_J = \nu_I \circ \Pi_{I,J}^{-1} \qquad \text{for all } J \subset I \ . \tag{4.5}$$

This relationship is the formal translation of the fact that the fidi dimensional distribution of $J \subset I$ is obtained from that of I by integrating with respect to the variables X_t, where t belongs to the complementary set of J in I. This property induce a particular structure in the collection of fidi distributions. In particular, they must at least satisfy the compatibility condition (4.5). We shall soon see that this condition is in fact sufficient.

The following theorem shows how from the collection of all fidi distributions, one can get back to a unique probability measure on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, provided that Condition (4.5) holds and that $(\mathsf{X}, \mathcal{X})$ satisfies some topological conditions, such as $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ (for more general topological spaces, in particular infinite dimensional ones, see [4, Theorem 12.1.2])). Note that only the existence part of this result is new as the uniqueness part is already a consequence of Theorem 4.2.1.

Theorem 4.2.2 (Kolmogorov)**.** *Set $\mathsf{X} = \mathbb{R}^p$ and $\mathcal{X} = \mathcal{B}(\mathbb{R}^p)$ for some $p \geq 1$. Let $\mathcal{I}$ be the set of all finite subsets of T. Let $(\nu_I)_{I\in\mathcal{I}}$ be such that, for all $I \in \mathcal{I}$, ν_I is a probability measure on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ and satisfies Condition* (4.5)*. Then there exists a unique probability measure $\mathbb{P}$ on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ such that, for all $I \in \mathcal{I}$, $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$.*

The following example is a simple application of this theorem.

Example 4.2.1 (Independent random variables)**.** *Let $(\nu_t)_{t\in T}$ be a collection of probability measures on $(\mathsf{X}, \mathcal{X})$. For all $I \in \mathcal{I}$, set*

$$\nu_I = \bigotimes_{t\in I} \nu_t \ , \tag{4.6}$$

where $\otimes$ denotes the tensor product of measures, that is, ν_I is the distribution of a vector with independent entries and marginal distributions given by ν_t, $t \in I$. It is easy to see that one defines a compatible collection of probability measures $(\nu_I)_{I\in\mathcal{I}}$ in the sense that Condition (4.5) holds. Hence, setting $\Omega = \mathsf{X}^T$, $X_t(\omega) = \omega_t$ and $\mathcal{F} = \sigma(X_t, t \in T)$, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $(X_t)_{t\in T}$ is a collection of independent random variables and $X_t \sim \nu_t$ for all $t \in T$.

4.2.3 Gaussian processes

We now introduce an important class of random processes that can be seen as an extension of Gaussian vectors to the infinite-dimensional case. Let us recall first the definition of Gaussian random variables, univariate and then multivariate. More details can be found in [5, Chapter 16].

Definition 4.2.7 (Gaussian variable)**.** *The real valued random variable X is Gaussian if its characteristic function satisfies :*

$$\phi_X(u) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right] = \exp(\mathrm{i}\mu u - \sigma^2 u^2/2)$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.

One can show that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. If $\sigma \neq 0$, then X admits a probability density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{4.7}$$

If $\sigma = 0$, then $X = \mu$ a.s. This definition can be extended to random vectors as follows.

Definition 4.2.8 (Gaussian vector)**.** *A random vector $[X_1, \ldots, X_n]^T$ valued in $\mathbb{R}^n$ is a Gaussian vector if any linear combination of $X_1, \ldots, X_n$ is a Gaussian variable.*

Let μ denote the mean vector of $[X_1, \ldots, X_n]^T$ and Γ its covariance matrix. Then, for all $u \in \mathbb{R}^n$, the random variable $Y = \sum_{k=1}^n u_k X_k = u^T X$ is Gaussian. It follows that its distribution is determined by its mean and variance which can be expressed as

$$\mathbb{E}[Y] = \sum_{k=1}^n u_k \mathbb{E}[X_k] = u^T\mu \quad \text{and} \quad \mathrm{Var}(Y) = \sum_{j,k=1}^n u_j u_k \,\mathrm{Cov}(X_j, X_k) = u^T\Gamma u$$

Thus, the characteristic function of $[X_1, \ldots, X_n]^T$ can be written using μ and Γ as

$$\phi_X(u) = \mathbb{E}\left[\exp(\mathrm{i}u^T X)\right] = \mathbb{E}\left[\exp(\mathrm{i}Y)\right] = \exp\left(\mathrm{i}u^T\mu - \frac{1}{2}u^T\Gamma u\right) \tag{4.8}$$

Conversely, if a n-dimensional random vector X has a characteristic function of this form, we immediately obtain that X is a Gaussian vector from the characteristic function of its scalar products. This property yields the following proposition.

Proposition 4.2.3. *The probability distribution of an n-dimensional Gaussian vector X is determined by its mean vector and covariance matrix Γ. We will denote*

$$X \sim \mathcal{N}(\mu, \Gamma) .$$

Conversely, for all vector $\mu \in \mathbb{R}^n$ and all non-negative symmetric matrix Γ, the distribution $X \sim \mathcal{N}(\mu, \Gamma)$ is well defined.

Proof. The first part of the result follows directly from (4.8). It also yields the following lemma.

Lemma 4.2.4. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ being a $n \times n$ non-negative symmetric matrix. Then for all $p \times n$ matrix A and $\mu' \in \mathbb{R}^n$, we have $\mu' + AX \sim \mathcal{N}(\mu' + A\mu, A\Gamma A^T)$.*

Let us now show the second (converse) part. First it holds for $n = 1$ as we showed previously. The case where Γ is diagonal follows easily. Indeed, let σ_i^2, $i = 1, \ldots, n$ denote the diagonal entries of Γ and set $\mu = [\mu_1, \ldots, \mu_n]^T$. Then take X_1, ...,X_n independent such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. We then get $X \sim \mathcal{N}(\mu, \Gamma)$ by writing its characteristic function. To conclude the proof of Proposition 4.2.3, just observe that all non-negative symmetric matrix Γ can be written as $\Gamma = U\Sigma U^T$ with Σ diagonal with non-negative entries and U orthogonal. Thus taking $Y \sim \mathcal{N}(0, \Sigma)$ and setting $X = \mu + UY$, the above lemma implies that $X \sim \mathcal{N}(\mu, \Gamma)$, which concludes the proof. □

The following proposition is easy to get (see [5, Corollaire 16.1]).

Proposition 4.2.5. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ a $n \times n$ non-negative symmetric matrix. Then X has independent components if and only if Γ is diagonal.*

Using the same path as in the proof of Proposition 4.2.3, i.e. by considering the cases where Γ is diagonal and using the diagionalization in an orthogonal basis to get the general case, one gets the following result (see [5, Corollaire 16.2]).

Proposition 4.2.6. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ a $n \times n$ non-negative symmetric matrix. If Γ is full rank, the probability distribution of X admits a density defined in $\mathbb{R}^n$ by*

$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Gamma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^n .$$

If Γ's rank $r < n$, that is, Γ has an $n-r$-dimensional null space, X belongs, with probability 1, to an r-dimensional affine subspace of $\mathbb{R}^n$. Indeed, there are $r - n$ linearly independent vectors a_i such that $\mathrm{Cov}(a_i^T X) = 0$ and thus $a_i^T X = a_i^T \mu$ a.s. Obviously X does not admit a density function in this case.

Having recalled the classical results on Gaussian vectors, we now introduce the definition of *Gaussian processes.*

Definition 4.2.9 (Gaussian processes)**.** *A real-valued random process $X = (X_t)_{t\in T}$ is called a Gaussian process if, for all finite set of indices $I = \{t_1, t_2, \cdots, t_n\}$, $[X_{t_1}, X_{t_2}, \cdots, X_{t_n}]^T$ is a Gaussian vector.*

Thus a Gaussian vector $[X_1, \ldots, X_n]^T$ may itself be seen as a Gaussian process $(X_t)_{t\in\{1,\ldots,n\}}$. This definition therefore has an interest in the case where T has an infinite cardinality. According to (4.8), the collection of fidi distributions is characterized by the mean function $\mu : t \in T \mapsto \mu(t) \in \mathbb{R}$ and the covariance function $\gamma : (t, s) \in (T \times T) \mapsto \gamma(t, s) \in \mathbb{R}$. Moreover, for all finite set of indices $I = \{t_1, t_2, \ldots, t_n\}$, the matrix Γ_I with entries $\Gamma_I(m, k) = \gamma(t_m, t_k)$, with $1 \leq m, k \leq n$, is a covariance matrix of a random vector of dimension n. It is therefore nonnegative symmetric. Conversely, given a function $\mu : t \in T \mapsto m(t) \in \mathbb{R}$ and a function $\gamma : (t, s) \in (T \times T) \mapsto \gamma(t, s) \in \mathbb{R}$ such that, for all finite set of indices I, the matrix Γ_I is nonnegative symmetric, we can define, for all finite set of indices $I = \{t_1, t_2, \cdots, t_n\}$, a Gaussian probability ν_I on $\mathbb{R}^n$ by

$$\nu_I \stackrel{\text{def}}{=} \mathcal{N}(\mu_I, \Gamma_I) \tag{4.9}$$

where $\mu_I = [\mu(t_1), \ldots, \mu(t_n)]^T$. The so defined collection $(\nu_I)_{I\in\mathcal{I}}$, satisfies the compatibility conditions and this implies, by Theorem 4.2.2, the following result.

Theorem 4.2.7. *Let T be any set of indices, μ a real valued function defined on T and γ a real valued function defined on $T \times T$ such that all restrictions Γ_I to the set $I \times I$ with $I \subseteq T$ finite are nonnegative symmetric matrices. Then one can define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a Gaussian process $(X_t)_{t\in T}$ defined on this space with* mean μ *and* covariance function *γ, that is such that, for all $s, t \in T$,*

$$\mu(t) = \mathbb{E}[X_t] \quad \text{and} \quad \gamma(s, t) = \mathbb{E}[(X_s - \mu(s))(X_t - \mu(t))] .$$

As a consequence we can extend the usual notation $\mathcal{N}(\mu, \gamma)$ as follows.

Definition 4.2.10 (Gaussian process fidi distributions)**.** *Let T be any index set. Let μ be any real valued function on T and γ any real valued function defined on $T \times T$ satisfying the condition of Theorem 4.2.7. We denote by $\mathcal{N}(\mu, \gamma)$ the law of the Gaussian process with mean μ and covariance γ in the sense of fidi distributions.*

4.3 Strict stationarity of a random process in discrete time

4.3.1 Definition

Stationarity plays a central role in stochastic modelling. We will distinguish two versions of this property, *strict stationarity* which says that the distribution of the random process is invariant by shifting the time origin and a *weak stationarity*, which imposes that only the first and second moments are invariant, with the additional assumption that these moments exist.

Definition 4.3.1 (Shift and backshift operators)**.** *Suppose that $T = \mathbb{Z}$ or $T = \mathbb{N}$. We denote by* S *and call the* shift operator *the mapping $\mathsf{X}^T \to \mathsf{X}^T$ defined by*

$$\mathrm{S}(x) = (x_{t+1})_{t\in T} \quad \textit{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T \ .$$

For all $\tau \in T$, we define S^τ by

$$\mathrm{S}^\tau(x) = (x_{t+\tau})_{t\in T} \quad \textit{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T \ .$$

The operator $\mathrm{B} = \mathrm{S}^{-1}$ is called the backshift operator*.*

Definition 4.3.2 (Strict stationarity)**.** *Set $T = \mathbb{Z}$ or $T = \mathbb{N}$. A random process $(X_t)_{t\in T}$ is strictly stationary if X and $\mathrm{S}\circ X$ have the same law, i.e. $\mathbb{P}^{\mathrm{S}\circ X} = \mathbb{P}^X$.*

Since the law is characterized by fidi distributions, one has $\mathbb{P}^{\mathrm{S}\circ X} = \mathbb{P}^X$ if and only if

$$\mathbb{P}^{\mathrm{S}\circ X} \circ \Pi_I^{-1} = \mathbb{P}^X \circ \Pi_I^{-1}$$

for all finite subset $I \in \mathcal{I}$. Now $\mathbb{P}^{\mathrm{S}\circ X} \circ \Pi_I^{-1} = \mathbb{P}^X \circ (\Pi_I \circ \mathrm{S})^{-1}$ and $\Pi_I \circ \mathrm{S} = \Pi_{I+1}$, where $I + 1 = \{t + 1, t \in I\}$. We conclude that $\{X_t, t \in T\}$ is *strictly stationary* if and only if, for all finite set $I \in \mathcal{I}$,

$$\mathbb{P}^{X_I} = \mathbb{P}_{I+1} \ .$$

Also observe that the strict stationarity implies that X and $\mathrm{S}^\tau \circ X$ has the same law for all $\tau \in T$ and thus $\mathbb{P}^{X_I} = \mathbb{P}_{I+\tau}$, where $I + \tau = \{t + \tau, t \in I\}$.

Example 4.3.1 (I.i.d process)**.** *Let $(Z_t)_{t\in T}$ be a sequence of* independent and identically distributed *(i.i.d) with values in $\mathbb{R}^d$. Then $(Z_t)_{t\in T}$ is a strictly stationary process , since, for all finite set $I = \{t_1, < t_2 < \cdots < t_n\}$ and all Borel set $A_1, \ldots, A_n$ of $\mathbb{R}^d$, we have*

$$\mathbb{P}(Z_{t_1} \in A_1, \cdots, Z_{t_n} \in A_n) = \prod_{j=1}^{n} \mathbb{P}(Z_0 \in A_j) \ ,$$

which does not depend on $t_1, \ldots, t_n$. Observe that, from Example 4.2.1, for all probability ν on $\mathbb{R}^d$, we can define a random process $(Z_t)_{t\in T}$ which is i.i.d. with marginal distribution *ν, that is, such that $Z_t \sim \nu$ for all $t \in T$.*

4.3.2 Stationarity preserving transformations

In this section, we set $T = \mathbb{Z}$, $\mathsf{X} = \mathbb{C}^d$ et $\mathcal{X} = \mathcal{B}(\mathbb{C}^d)$ for some integer $d \geq 1$. Let us start with an illustrating example.

Example 4.3.2 (Moving transformation of an i.i.d. process)**.** *Let Z be an i.i.d. process (see Example 4.3.1). Let k be an integer and g a measurable function from $\mathbb{R}^k$ to $\mathbb{R}$. One can check that the process $(X_t)_{t\in\mathbb{Z}}$ defined by*

$$X_t = g(Z_t, Z_{t-1}, \cdots, Z_{t-k+1})$$

also is a stationary random process in the strict sense. On the other hand, the obtained process is not i.i.d. in general since for $k \geq 1$, $X_t, X_{t+1}, \ldots, X_{t+k-1}$ are identically distributed but are in general dependent variables as they all depend on the same random variables Z_t. Nevertheless such a process is said to be k-dependent *because $(X_s)_{s\leq t}$ and $(X_s)_{s>t+k}$ are independent for all t. The m-dependent processes can be used to approximate a large class of dependent processes to study the asymptotic behavior of statistics such as usual the sample mean.*

Observe that in this example, to derive the stationarity of X, it is not necessary to use that Z is i.i.d., only that it is stationary. In fact, to check stationarity, it is often convenient to reason directly on the laws of the trajectories using the notion of filtering.

Definition 4.3.3. *Let ϕ be a measurable function from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{Y}^T, \mathcal{Y}^{\otimes T})$ and $X = (X_t)_{t\in T}$ be a process with values in $(\mathsf{X}, \mathcal{X})$. A ϕ- filtering with input X and output Y means that the random process $Y = (Y_t)_{t\in T}$ is defined as $Y = \phi \circ X$, or, equivalently, $Y_t = \Pi_t(\phi(X))$ for all $t \in T$, where Π_t is a shorthand notation for $\Pi_{\{t\}}$ defined in (4.3). Thus Y takes its values in $(\mathsf{Y}, \mathcal{Y})$. If ϕ is linear, we will say that Y is obtained by* linear filtering *of X.*

In Example 4.3.2, X is obtained by ϕ-filtering Z (non-linearly, unless g is a linear form) with $\phi : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ defined by

$$\phi\big((x_t)_{t\in\mathbb{Z}}\big) = \big(g(x_t, x_{t-1}, \ldots, x_{t-k+1})\big)_{t\in\mathbb{Z}} .$$

Example 4.3.3 (Shift)**.** *A very basic linear filtering is obtained with $\phi = \mathrm{S}$ where S is the shift operator of Definition 4.3.1. In this case $Y_t = X_{t+1}$ for all $t \in \mathbb{Z}$.*

Example 4.3.4 (Finite impulse response filter (FIR))**.** *Let $n \geq 1$ and $t_1 < \cdots < t_n$ in $\mathbb{Z}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$. Then $\phi = \sum_i \alpha_i \mathrm{S}^{-t_i}$ defines a linear filtering and for any input $X = (X_t)_{t\in\mathbb{Z}}$, the output is given by*

$$Y_t = \sum_{i=1}^{n} \alpha_i X_{t-t_i}, \quad t \in \mathbb{Z} .$$

Example 4.3.5 (Differencing operator)**.** *A particular case is the* differencing operator $\mathbf{1} - \mathrm{S}^{-1}$ *where $\mathbf{1}$ denotes the identity on X^T. The output then reads as*

$$Y_t = X_t - X_{t-1}, \quad t \in \mathbb{Z} .$$

One can iterate this operator so that $Y = (\mathbf{1} - \mathrm{S}^{-1})^k X$ is given by

$$Y_t = \sum_{j=0}^{k} \binom{k}{j} (-1)^j X_{t-j}, \quad t \in \mathbb{Z} .$$

Example 4.3.6 (Time reversion)**.** *Let $X = \{X_t, t \in \mathbb{Z}\}$ be a random process. Time reversion then set the output as*

$$Y_t = X_{-t}, \quad t \in \mathbb{Z}\,.$$

Note that in all previous examples the operators introduced preserve the strict stationarity, that is to say, if the input X is strictly stationary then so is the output Y. It is easy to construct a linear filtering which does not preserve the strict stationarity, for example, $y = \phi(x)$ with $y_t = x_t$ for t even and $Y_t = x_t + 1$ for t odd. A property stronger than the conservation of stationarity and very easy to verify is given by the following definition.

Definition 4.3.4. *A ϕ-filter is* shift invariant *if ϕ commutes with* S, $\phi \circ \mathrm{S} = \mathrm{S} \circ \phi$.[1]

It is easy to show that a shift-invariant filter preserves the strict stationarity. However it is a stronger property. The time reversion is an example of a filter that is not shift-invariant, although it does preserve the strict stationarity. Indeed, in this case, we have $\phi \circ \mathrm{S} = \mathrm{S}^{-1} \circ \phi$. All the other examples above are shift-invariant.

Remark 4.3.1. *A* shift invariant *ϕ-filter is entirely determined by its composition with the canonical projection Π_0 defined as in (4.3) but with $I = \{0\}$. Indeed, let $\phi_0 = \Pi_0 \circ \phi$. Then for all $s \in \mathbb{Z}$, $\Pi_s \circ \phi = \Pi_0 \circ \mathrm{S}^s \circ \phi = \Pi_0 \circ \phi \circ \mathrm{S}^s$. Since for all $x \in \mathsf{X}^T$, $\phi(x)$ is the sequence $(\Pi_s \circ \phi)_{s \in T}$, we get the result.*

4.4 Stopping Times

In this section, we consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$ and an adapted process $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$. The σ-field $\mathcal{F}_\infty$ is defined as the smallest one containing all $\mathcal{F}_k$, $k \in \mathbb{N}$, that is

$$\mathcal{F}_\infty = \bigvee_{k\in\mathbb{N}} \mathcal{F}_k\,.$$

In many examples, $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is the natural filtration of some given process $(Y_m)_{m\in\mathbb{N}}$.

The term *stopping time* is an expression from gambling. A game of chance which evolves in time (for example an infinite sequence of coin tosses) can be adequately represented by a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$, the sub-$\sigma$-fields $\mathcal{F}_n$ giving the information on the results of the game available to the player at time n. A stopping rule for the player thus consists of giving a rule for leaving the game at time n, based at each time on the information at his disposal at that time. The time τ of stopping the game by such a rule is a stopping time. Note that stopping times may take the value $+\infty$, corresponding to the case where the game does not stop.

Definition 4.4.1 (Stopping times)**.** *A random variable τ from Ω to $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ is called a stopping time on the filtered measurable space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}})$ if, for all $k \in \mathbb{N}$, $\{\tau \le k\} \in \mathcal{F}_k$. The family $\mathcal{F}_\tau$ of events $A \in \mathcal{F}_\infty$ such that, for every $k \in \mathbb{N}$, $A \cap \{\tau \le k\} \in \mathcal{F}_k$, is called the σ-field of events prior to time τ.*

It can be easily checked that $\mathcal{F}_\tau$ is indeed a σ-field (Exercise 4.5). Since $\{\tau = n\} = \{\tau \le n\} \setminus \{\tau \le n-1\}$, one can replace $\{\tau \le n\}$ by $\{\tau = n\}$ in the definition of the stopping time τ

[1] There is a slight hidden discrepancy in this definition: if ϕ is defined from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{Y}^T, \mathcal{Y}^{\otimes T})$ with $\mathsf{X} \neq \mathsf{Y}$ then the notation S refers to two different shifts: one on X^T and the other one on Y^T.

and in the definition of the σ-field $\mathcal{F}_\tau$. It may sometimes be useful to note that the constant random variables are also stopping times. In such a case, there exists some $n \in \mathbb{N}$ such that $\tau(\omega) = n$ for every $\omega \in \Omega$, and $\mathcal{F}_\tau = \mathcal{F}_n$.

For any stopping time τ, the event $\{\tau = \infty\}$ belongs to $\mathcal{F}_\infty$, for it is the complement of the union of the event $\{\tau = n\}$, $n \in \mathbb{N}$, which all belong to $\mathcal{F}_\infty$. It follows that $B \cap \{\tau = \infty\} \in \mathcal{F}_\infty$ for all $B \in \mathcal{F}_\tau$, showing that $\tau : \Omega \to \bar{\mathbb{N}}$ is $\mathcal{F}_\infty$ measurable.

Definition 4.4.2 (Hitting times)**.** *For $A \in \mathcal{X}$, the first hitting time τ_A and σ_A of the set A are the $\bar{\mathbb{N}}$-valued random-variables respectively defined on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ by*

$$\tau_A : (x_n)_{n\in\mathbb{N}} \mapsto \inf\{n \geq 0 \,:\, x_n \in A\} \ ,$$
$$\sigma_A : (x_n)_{n\in\mathbb{N}} \mapsto \inf\{n \geq 1 \,:\, x_n \in A\} \ ,$$

where, by convention, $\inf \emptyset = +\infty$. The successive positive hitting times $\sigma_A^{(n)}$, $n \geq 0$, are defined inductively by $\sigma_A^{(0)} = 0$ and for all $k \geq 0$,

$$\sigma_A^{(k+1)} : x = (x_n)_{n\in\mathbb{N}} \mapsto \inf\left\{n > \sigma_A^{(k)}(x) \,:\, x_n \in A\right\} \ . \tag{4.10}$$

Hitting times are stopping times with respect to the canonical filtration, since, for all $n \in \mathbb{N}$,

$$\{\tau_A \leq n\} = \bigcup_{k=0}^{n} \{\xi_k \in A\} \in \mathcal{F}_n \ ,$$

and

$$\{\sigma_A \leq 0\} = \emptyset \quad \text{and if } n \geq 1 \,,\ \{\sigma_A \leq n\} = \bigcup_{k=1}^{n} \{\xi_k \in A\} \in \mathcal{F}_n^\xi \ .$$

The stopping times property is preserved through many standard operations, as shown in the following useful result.

Proposition 4.4.1. *Let $(\Omega,\ \mathcal{F},\ (\mathcal{F}_k)_{k\in\mathbb{N}},\ \mathbb{P})$ be a filtered probability space and τ and σ be two stopping times for the filtration $(\mathcal{F}_n)_{n\geq 0}$. Denote by $\mathcal{F}_\tau$ and $\mathcal{F}_\sigma$ the σ-fields of the events prior to τ and σ, respectively. Then,*

(i) $\tau \wedge \sigma$, $\tau \vee \sigma$ and $\tau + \sigma$ are stopping times,

(ii) if $\tau \leq \sigma$, then $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$,

(iii) $\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$,

(iv) $\{\tau < \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$, $\{\tau = \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.

Proof.

(i) Let $n \in \mathbb{N}$. We show that the events $\{\tau \wedge \sigma \leq n\}$, $\{\tau \vee \sigma \leq n\}$ and $\{\tau + \sigma \leq n\}$ belong to $\mathcal{F}_n$. Since

$$\{\tau \wedge \sigma \leq n\} = \{\tau \leq n\} \cup \{\sigma \leq n\}$$

and τ and σ are stopping times, $\{\tau \leq n\}$ and $\{\sigma \leq n\}$ belong to $\mathcal{F}_n$; therefore $\{\tau \wedge \sigma \leq n\} \in \mathcal{F}_n$. Similarly, $\{\tau \vee \sigma \leq n\} = \{\tau \leq n\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$. Finally,

$$\{\tau + \sigma \leq n\} = \bigcup_{k=0}^{n} \{\tau \leq k\} \cap \{\sigma \leq n - k\} \ .$$

Now, for $0 \leq k \leq n$, $\{\tau \leq k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\sigma \leq n-k\} \in \mathcal{F}_{n-k} \subset \mathcal{F}_n$; hence $\{\tau+\sigma \leq n\} \in \mathcal{F}_n$.

(ii) Let $A \in \mathcal{F}_\tau$ and $n \in \mathbb{N}$. As $\{\sigma \leq n\} \subset \{\tau \leq n\}$, $A \cap \{\sigma \leq n\} = A \cap \{\tau \leq n\} \cap \{\sigma \leq n\}$. Now $A \cap \{\tau \leq n\} \in \mathcal{F}_n$ and $\{\sigma \leq n\} \in \mathcal{F}_n$ (σ is a stopping time); therefore $A \cap \{\tau \leq n\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$ and $A \cap \{\sigma \leq n\} \in \mathcal{F}_n$. Thus $A \in \mathcal{F}_\sigma$.

(iii) It follows from (ii) that $\mathcal{F}_{\tau\wedge\sigma} \subset \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. Conversely, let $A \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. Obviously $A \subset \mathcal{F}_\infty$. To prove that $A \in \mathcal{F}_{\tau\wedge\sigma}$, one must show that, for every $k \geq 0$, $A \cap \{\tau \wedge \sigma \leq k\} \in \mathcal{F}_k$. We have $A \cap \{\tau \leq k\} \in \mathcal{F}_k$ and $A \cap \{\sigma \leq k\} \in \mathcal{F}_k$. Hence, since $\{\tau \wedge \sigma \leq k\} = \{\tau \leq k\} \cup \{\sigma \leq k\}$, we get

$$A \cap \{\tau \wedge \sigma \leq k\} = A \cap (\{\tau \leq k\} \cup \{\sigma \leq k\}) = (A \cap \{\tau \leq k\}) \cup (A \cap \{\sigma \leq k\}) \in \mathcal{F}_k \ .$$

(iv) Let $n \in \mathbb{N}$. We write

$$\{\tau < \sigma\} \cap \{\tau \leq n\} = \bigcup_{k=0}^{n} \{\tau = k\} \cap \{\sigma > k\} \ .$$

Now, for $0 \leq k \leq n$, $\{\tau = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\sigma > k\} = \{\sigma \leq k\}^c \in \mathcal{F}_k \subset \mathcal{F}_n$. Therefore, $\{\tau < \sigma\} \cap \{\tau \leq n\} \in \mathcal{F}_n$, showing that $\{\tau < \sigma\} \in \mathcal{F}_\tau$. Similarly,

$$\{\tau < \sigma\} \cap \{\sigma \leq n\} = \bigcup_{k=0}^{n} \{\sigma = k\} \cap \{\tau < k\}$$

and since, for $0 \leq k \leq n$, $\{\sigma = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\tau < k\} = \{\tau \leq k-1\} \in \mathcal{F}_{k-1} \subset \mathcal{F}_n$, it also holds $\{\tau < \sigma\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$ so that $\{\tau < \sigma\} \in \mathcal{F}_\sigma$. Finally, $\{\tau < \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. The last statement of the proposition follows from

$$\{\tau = \sigma\} = \{\tau < \sigma\}^c \cap \{\sigma < \tau\}^c \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma \ .$$

□

4.5 Exercises

Exercise 4.1. Let X be a Gaussian vector, A_1 and A_2 two linear applications. Let us set $X_1 = A_1X$ and $X_2 = A_2X$. Give the distribution of (X_1, X_2) and a necessary and sufficient condition for X_1 and X_2 to be independent.

Exercise 4.2. Let X be a Gaussian random variable, with zero mean and unit variance, $X \sim \mathcal{N}(0,1)$. Let $Y = X\mathbf{1}_{\{U=1\}} - X\mathbf{1}_{\{U=0\}}$ where U is a Bernoulli random variable with parameter $1/2$ independent of X. Show that $Y \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}\,(X,Y) = 0$ but also that X and Y are not independent.

Exercise 4.3. Let $n \geq 1$ and Γ be a $n \times n$ nonnegative definite hermitian matrix.

1. Find a Gaussian vector X valued in $\mathbb{R}^n$ and a unitary matrix U such that UX has covariance matrix Γ. [Hint : take a look at the proof of Proposition 4.2.3].
2. Show that
$$\Sigma := \frac{1}{2}\begin{bmatrix} \mathrm{Re}(\Gamma) & -\mathrm{Im}(\Gamma) \\ \mathrm{Im}(\Gamma) & \mathrm{Re}(\Gamma) \end{bmatrix}$$
is a real valued $(2n) \times (2n)$ nonnegative definite symmetric matrix.

Let X and Y be two n-dimensional Gaussian vectors such that

$$\begin{bmatrix} X & Y \end{bmatrix}^T \sim \mathcal{N}\,(0, \Sigma)\ .$$

3. What is the covariance matrix of $Z = X + \mathrm{i}Y$?
4. Compute $\mathbb{E}\left[ZZ^T\right]$.

The random variable Z is called a centered circularly-symmetric normal vector.

Let now T be an arbitrary index set, $\mu : I \to \mathbb{C}$ and $\gamma : T^2 \to \mathbb{C}$ such that for all finite subset $I \subset T$, the matrix $\Gamma_I = [\gamma(s,t)]_{s,t\in I}$ is a nonnegative definite hermitian matrix.

5. Use the previous questions to show that there exists a random process $(X_t)_{t\in T}$ valued in $\mathbb{C}$ such that, for all $s, t \in T$,
$$\mathbb{E}\,[X_t] = \mu(t) \quad \text{and} \quad \mathrm{Cov}\,(X_s, X_t) = \gamma(s,t)\ .$$

Exercise 4.4. Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. real valued random variables. Determine in each of the following cases, if the defined process is strongly stationary.

1. $Y_t = a + b\varepsilon_t + c\varepsilon_{t-1}$ (a, b, c real numbers).
2. $Y_t = a + b\varepsilon_t + c\varepsilon_{t+1}$.
3. $Y_t = \sum_{j=0}^{+\infty} \rho^j \varepsilon_{t-j}$ for $|\rho| < 1$, assuming that $\mathbb{E}\,[|\epsilon_0|] < \infty$.
4. $Y_t = \varepsilon_t\varepsilon_{t-1}$.
5. $Y_t = (-1)^t\,\varepsilon_t$, $Z_t = \varepsilon_t + Y_t$.

Exercise 4.5. Let τ be a stopping time on the filtered measurable space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}})$. Show that $\mathcal{F}_\tau$ in Definition 4.4.1 is a sub-σ-field of $\mathcal{F}$.

Exercise 4.6. Let τ be a $(\mathcal{F}_n)$-stopping time on $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Let Y be a real random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Show that Y is $\mathcal{F}_\tau$-measurable if and only if $Y\mathbb{1}_{\{\tau \leq n\}}$ is $\mathcal{F}_n$-measurable for all $n \in \mathbb{N}$.

2. Show that $A \in \mathcal{F}$ is $\mathcal{F}_\tau$-measurable if and only if $A \cap \{\tau = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}$.

3. Let $n \in \mathbb{N}$ and Y be a real random variable $\mathcal{F}_n$-measurable. Prove that $Y\mathbb{1}_{\{\tau=n\}}$ is $\mathcal{F}_\tau$-measurable.

4. Let X be an integrable real random variable. Show that for all $n \in \overline{\mathbb{N}}$,

$$\mathbb{E}\left[X\mathbb{1}_{\{\tau=n\}} \middle| \mathcal{F}_\tau\right] = \mathbb{1}_{\{\tau=n\}}\mathbb{E}\left[X \middle| \mathcal{F}_n\right] .$$

Chapter 5

Weakly stationary processes

In this chapter, we focus on second order properties of univariate time series, that is, on their means and covariance functions. It turns out that the stationarity induces a particular structure of the covariances of a time series that can be exploited to provide a spectral representation of the time series. Finally we will conclude the chapter with the Wold decomposition, which basically shows that any weakly stationary processes, up to an additive deterministic-like component, can be expressed by linearly filtering a white noise (the innovation process).

5.1 L^2 processes

The definition of the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$, or simply L^2, is recalled in Example 1.1.4.

Definition 5.1.1 (L^2 univariate time series)**.** *The process $X = (X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{C}$ is an L^2 process if $X_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ for all $t \in T$.*

The *mean function* defined on T by $\mu(t) = \mathbb{E}[X_t]$ takes its values in $\mathbb{C}$ and the *covariance function* is defined on $T \times T$ by

$$\gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathbb{E}\left[(X_s - \mu(s))\overline{(X_t - \mu(t))}\right] ,$$

which takes its values in $\mathbb{C}$. We will sometimes use the notation μ_X and γ_X, the subscript X indicating the process used in these definitions. For all $s \in T$, $\gamma(s,s)$ is a variance and is thus nonnegative. More generally, the following properties hold.

Proposition 5.1.1. *Let Γ be the covariance function of a L^2 process $X = (X_t)_{t\in T}$ with values in $\mathbb{C}^d$. The following properties hold.*

(i) Hermitian symmetry: for all $s, t \in T$,

$$\gamma(s,t) = \overline{\gamma(t,s)} \tag{5.1}$$

(ii) Nonnegativity: for all $n \geq 1$, $t_1, \ldots, t_n \in T$ and $a_1, \cdots, a_n \in \mathbb{C}$,

$$\sum_{1\leq k,m\leq n} \overline{a_k}\gamma(t_k, t_m)a_m \geq 0 \tag{5.2}$$

Conversely, if γ satisfy these two properties, there exists an L^2 process $X = (X_t)_{t\in T}$ with values in $\mathbb{C}$ with covariance function γ.

Proof. Relation (5.1) is immediate. To show (5.2), define the linear combination $Y = \sum_{k=1}^{n} \overline{a_k} X_{t_k}$. Y is a complex valued random variable. Using that the Cov operator is hermitian, we get

$$\mathrm{Var}\,(Y) = \sum_{1\leq k,m\leq n} \overline{a_k}\gamma(t_k, t_m)a_m$$

which implies (5.2).

The converse assertion follows from Exercise 4.3. □

5.2 Univariate weakly stationary time series

From now on, in this chapter, we take $T = \mathbb{Z}$. If an L^2 process is strictly stationary, then its first and second order properties must satisfy certain properties. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a strictly stationary L^2 process with values in $\mathbb{C}$. Then its mean function is constant, since its marginal distribution is invariant. Moreover its covariance function Γ satisfies $\gamma(s,t) = \gamma(s-t, 0)$ for all $s, t \in \mathbb{Z}$ since the bi-dimensional marginals are also invariant by a translation of time. A weakly stationary process inherits these properties but is not necessary strictly stationary, as in the following definition.

Definition 5.2.1 ((Univariate) weakly stationary time series)**.** *Let $\mu \in \mathbb{C}$ and $\gamma : \mathbb{Z} \to \mathbb{C}$. A process $(X_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}$ is said* weakly stationary *with mean μ and autocovariance function γ if all the following assertions hold:*

(i) X is an L^2 process, i.e. $\mathbb{E}\left[|X_t|^2\right] < +\infty$,

(ii) for all $t \in \mathbb{Z}$, $\mathbb{E}\left[X_t\right] = \mu$,

(iii) for all $(s,t) \in \mathbb{Z}\times\mathbb{Z}$, $\mathrm{Cov}\,(X_s, X_t) = \gamma(s-t)$.

By definition the autocovariance function of a weakly stationary process is defined on T instead of T^2 for the covariance function in the general case.

As already mentioned a strictly stationary L^2 process is weakly stationary. The converse implication is of course not true in general. It is true however for Gaussian processes defined in Section 4.2.3, see Proposition 4.2.3.

Here we considered time series valued in $\mathbb{C}$, hence called univariate. We will also consider weakly stationary time series valued in a general Hilbert space. In the case where this space is $\mathbb{C}^d$, the obtained time series is called *multivariate*.

5.2.1 Properties of the autocovariance function

The properties of Proposition 5.1.1 imply the following ones in the case of a weakly stationary process.

Proposition 5.2.1. *The autocovariance function $\gamma : \mathbb{Z} \to \mathbb{C}$ of a complex valued weakly stationary process satisfies the following properties.*

(i) Hermitian symmetry : for all $s \in \mathbb{Z}$,

$$\gamma(-s) = \overline{\gamma(s)}$$

(ii) Nonnegative definiteness : for all integer $n \geq 1$ and $a_1, \cdots, a_n \in \mathbb{C}$,

$$\sum_{s=1}^{n}\sum_{t=1}^{n} \overline{a_s}\gamma(s-t)a_t \geq 0$$

The autocovariance matrix Γ_n of n consecutive samples $X_1, \ldots, X_n$ of the time series has a particular structure, namely it is constant on its diagonals, $(\Gamma_n)_{ij} = \gamma(i-j)$,

$$\begin{aligned}\Gamma_n &= \mathrm{Cov}([X_1 \ \ldots \ X_n]^T) \\ &= \begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(1-n) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(2-n) \\ \vdots & & & \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix}\end{aligned} \tag{5.3}$$

On says that Γ_n is a *Toeplitz* matrix . Since $\gamma(0)$ is generally non-zero (note that otherwise X_t is zero a.s. for all t), it can be convenient to normalize the autocovariance function in the following way.

Definition 5.2.2 (Autocorrelation function)**.** *Let X be a weakly stationary process with autocovariance function γ such that $\gamma(0) \neq 0$. The* autocorrelation function *of X is defined as*

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \quad \tau \in \mathbb{Z}\ .$$

It is normalized in the sense that $\rho(0) = 1$ and $|\rho(s)| \leq 1$ for all $s \in \mathbb{Z}$.

The last assertion follows from the Cauchy-Schwarz inequality (see Theorem 1.1.1),

$$|\gamma(s)| = |\mathrm{Cov}\,(X_s, X_0)| \leq \sqrt{\mathrm{Var}\,(X_s)\,\mathrm{Var}\,(X_0)} = \gamma(0)\ ,$$

the last equality following from the weakly stationary assumption.

Let us give some simple examples of weakly stationary processes. We first examine a very particular case.

Definition 5.2.3 (White noise)**.** *A weak white noise is a centered weakly stationary process whose autocovariance function satisfies $\gamma(0) = \sigma^2 > 0$ and $\gamma(s) = 0$ for all $s \neq 0$. We will denote $(X_t) \sim \mathrm{WN}(0, \sigma^2)$. When a weak white noise is an i.i.d. process, it is called a* strong white noise. *We will denote $(X_t) \sim \mathrm{IID}(0, \sigma^2)$.*

Of course a strong white noise is a weak white noise. However the converse is in general not true. The two definitions only coincide for Gaussian processes because in this case the independence is equivalent to being uncorrelated.

Example 5.2.1 (MA(1) process)**.** *Define, for all $t \in \mathbb{Z}$,*

$$X_t = Z_t + \theta Z_{t-1}\ , \tag{5.4}$$

where $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ and $\theta \in \mathbb{R}$. Then $\mathbb{E}\,[X_t] = 0$ and the autocovariance function reads

$$\gamma(s) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } s = 0, \\ \sigma^2\theta & \text{if } s = \pm 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

Such a weakly stationary process is called a Moving Average of order 1 MA(1).

Example 5.2.2 (Harmonic process). *Let $(A_k)_{1\leq k\leq N}$ be N real valued L^2 random variables. Denote $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$. Let $(\Phi_k)_{1\leq k\leq N}$ be N i.i.d. random variables with a uniform distribution on $[-\pi,\pi]$, and independent of $(A_k)_{1\leq k\leq N}$. Define*

$$X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k) \;, \tag{5.6}$$

where $(\lambda_k)_{1\leq k\leq N} \in [-\pi,\pi]$ are N frequencies. The process (X_t) is called an harmonic process. *It satisfies $\mathbb{E}[X_t] = 0$ and, for all $s,t\in\mathbb{Z}$,*

$$\mathbb{E}[X_s X_t] = \frac{1}{2}\sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k(s-t)) \;.$$

It is thus a weakly stationary process.

Example 5.2.3 (Random walk). *Let $(S_t)_{t\in\mathbb{N}}$ be a random process defined by $S_0 = 0$ and, for all $t\in\mathbb{N}^*$, by $S_t = X_1 + \cdots + X_t$, where (X_t) is a strong white noise with variance σ^2. Such a process is called a* random walk. *We have $\mathbb{E}[S_t] = 0$, $\mathbb{E}\left[S_t^2\right] = t\sigma^2$ and for all $s\leq t\in\mathbb{N}$,*

$$\mathbb{E}[S_s S_t] = \mathbb{E}[(S_s + X_{s+1} + \cdots + X_t)S_s] = s\,\sigma^2$$

The process $(S_t)_{t\in\mathbb{N}}$ is not weakly stationary.

Example 5.2.4 (Continued from Example 5.2.1). *Consider the function χ defined on $\mathbb{Z}$ by*

$$\chi(s) = \begin{cases} 1 & \text{if } s = 0, \\ \rho & \text{if } s = \pm 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.7}$$

where $\rho\in\mathbb{R}$. It is the autocovariance function of a real valued process if and only if $\rho\in[-1/2,1/2]$. We know from Example 5.2.1 that χ is the autocovariance function of a real valued MA(1) process if and only if $\sigma^2(1+\theta^2) = 1$ and $\sigma^2\theta = \rho$ for some $\theta\in\mathbb{R}$. If $|\rho|\leq 1/2$, the solutions to this equation are

$$\theta = (2\rho)^{-1}(1 \pm \sqrt{1-4\rho^2}) \quad \text{and} \quad \sigma^2 = (1+\theta^2)^{-1} \;.$$

If $|\rho| > 1/2$, there are no real solutions. In fact, in this case, it can even be shown that there is no real valued weakly stationary process whose autocovariance is χ, see Exercise 5.3.

Some simple transformations of processes preserve the weak stationarity. Linearity is crucial in this case since otherwise the second order properties of the output cannot solely depend on the second order properties of the input.

Example 5.2.5 (Invariance of the autocovariance function under time reversion (continued from Example 4.3.6). *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a weakly stationary process with mean μ_X and autocovariance function γ_X. Denote, for all $t\in\mathbb{Z}$, $Y_t = X_{-t}$ as in Example 4.3.6. Then (Y_t) is weakly stationary with same mean as X and autocovariance function $\gamma_Y = \overline{\gamma_X}$.*

$$\begin{aligned} &\mathbb{E}[Y_t] = \mathbb{E}[X_{-t}] = \mu_X \;, \\ &\mathrm{Cov}(Y_{t+h}, Y_t) = \mathrm{Cov}(X_{-t-h}, X_{-t}) = \gamma_X(-h) = \overline{\gamma_X(h)} \;. \end{aligned}$$

5.2.2 Empirical mean and autocovariance function

Suppose that we observe n consecutive samples of a real valued weakly stationary time series $X = (X_t)$. Can we have a rough idea of the second order parameters of X μ and γ ? This is an estimation problem. The first step for answering this question is to provide estimators of μ and γ. Since these quantities are defined using an expectation $\mathbb{E}$, a quite natural approach is to replace this expectation by an empirical sum over the observed data. This yields the *empirical mean*

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \ , \tag{5.8}$$

and the *empirical autocovariance* and *autocorrelation* functions

$$\hat{\gamma}_n(h) = \frac{1}{n}\sum_{k=1}^{n-|h|}(X_k - \hat{\mu}_n)(X_{k+|h|} - \hat{\mu}_n) \quad \text{and} \quad \hat{\rho}_n(h) = \hat{\gamma}_n(h)/\hat{\gamma}_n(0) \ . \tag{5.9}$$

Let us examine how such estimators look like on some examples.

Example 5.2.6 (Heartbeats (Continued from Example 4.1.1))**.** *Take the data displayed in Figure 4.1, which roughly looks stationary. Its empirical autocorrelation is displayed in Figure 5.1. We observe a positive correlation in the sense that the obtained values are significantly above the x-axis, at least if one compares with the empirical correlation obtained from a sample of a Gaussian white noise with the same length.*

A positive autocorrelation $\rho(h)$ has a simple interpretation: it means that X_t and X_{t+h} have a tendency of being on the same side of their means with a higher probability. A more precise interpretation is to observe that, in the sense of the Hilbert space L^2 (see Chapter 1),

$$\mathrm{proj}\,(X_{t+h} - \mu | \,\mathrm{Span}\,(X_t - \mu)) = \rho(h)(X_t - \mu) \ ,$$

and the error has variance $\gamma(0)(1 - |\rho(h)|^2)$ (see Exercise 5.5). In practice, we do not have access to the exact computation of these quantities from a single sample $X_1, \ldots, X_n$. We can however let t varies at fixed h, hoping that the evolution in t more or less mimic the variation in ω. In Figure 5.2, we plot X_t VS X_{t+1} and indeed see this phenomenon: $\hat{\rho}(1) = 0.966$ indicate that X_{t+1} is very well approximated by a linear function of X_t, as can be observed in this figure.

5.3 Spectral measure

Recall that $\mathbb{T}$ denotes any interval congruent to $[0, 2\pi)$. We denote by $\mathcal{B}(\mathbb{T})$ the associated Borel σ-field. The Herglotz theorem shows that the autocovariance function of a weakly stationary process X is entirely determined by a finite nonnegative measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. This measure is called the *spectral measure* of X.

Theorem 5.3.1 (Herglotz)**.** *A sequence $(\gamma(h))_{h\in\mathbb{Z}}$ is a nonnegative definite hermitian sequence in the sense of Proposition 5.2.1 if and only if there exists a finite nonnegative measure ν on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ such that :*

$$\gamma(h) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda}\nu(\mathrm{d}\lambda), \ \forall h \in \mathbb{Z} \ . \tag{5.10}$$

Moreover this relation defines ν uniquely.

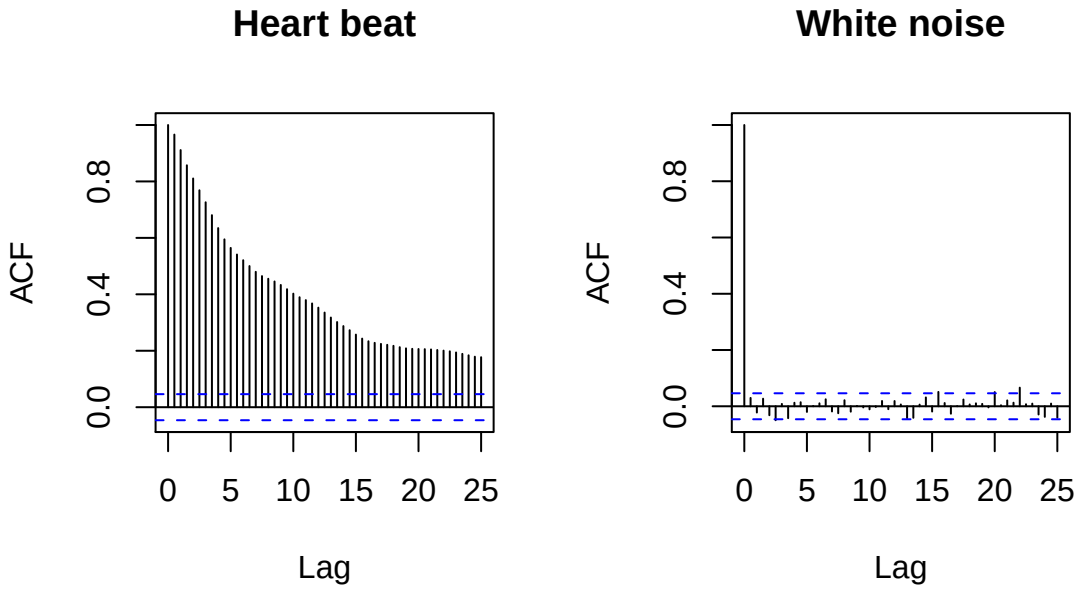

Figure 5.1: *Left : empirical autocorrelation* $\hat{\rho}_n(h)$ *of heartbeat data for* $h = 0, \ldots, 100$. *Right : the same from a simulated white noise sample with same length.*

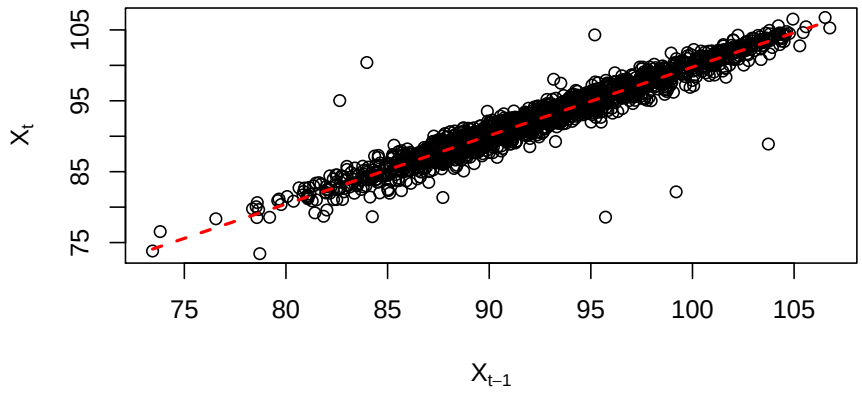

Figure 5.2: *Each point is a couple* (X_{t-1}, X_t), *where* $X_1, \ldots, X_n$ *is the heartbeat data sample. The dashed line is the best approximation of* X_t *as a linear function of* X_{t-1}.

Remark 5.3.1. *By Proposition 5.2.1, Theorem 5.3.1 applies to all γ which is an autocovariance function of a weakly stationary process X. In this case ν (also denoted ν_X) is called the* spectral measure *of X. If ν admits a density f, it is called the* spectral density function.

Proof. Suppose first that $\gamma(n)$ satisfies (5.10) with ν as in the theorem. Then γ is an hermitian function. Let us show it is a nonnegative definite hermitian function. Fix a positive integer n. For all $a_k \in \mathbb{C}$, $1 \le k \le n$, we have

$$\sum_{k,m} a_k \overline{a_m} \gamma(k-m) = \int_{\mathbb{T}} \sum_{k,m} a_k \overline{a_m} \mathrm{e}^{\mathrm{i}k\lambda} \mathrm{e}^{-im\lambda} \nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} \left| \sum_k a_k \mathrm{e}^{\mathrm{i}k\lambda} \right|^2 \nu(\mathrm{d}\lambda) \ge 0 \, .$$

Hence γ is nonnegative definite.

Conversely, suppose that γ is a nonnegative definite hermitian sequence. For all $n \ge 1$, define the function

$$\begin{aligned} f_n(\lambda) &= \frac{1}{2\pi n} \sum_{k=1}^{n} \sum_{m=1}^{n} \gamma(k-m) \mathrm{e}^{-\mathrm{i}k\lambda} \mathrm{e}^{\mathrm{i}m\lambda} \\ &= \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right) \gamma(k) \mathrm{e}^{-\mathrm{i}k\lambda} \, . \end{aligned}$$

Since γ is nonnegative definite, we get from the first equality that $f_n(\lambda) \ge 0$, for all $\lambda \in \mathbb{T}$. Define ν_n as the nonnegative measure with density f_n on $\mathbb{T}$. We get that

$$\begin{aligned} \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu_n(\mathrm{d}\lambda) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} f_n(\lambda) \mathrm{d}\lambda &= \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right) \gamma(k) \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}(h-k)\lambda} \mathrm{d}\lambda \\ &= \begin{cases} \left(1 - \frac{|h|}{n}\right) \gamma(h), & \text{if } |h| < n \, , \\ 0, & \text{otherwise} \, . \end{cases} \end{aligned} \tag{5.11}$$

We can multiply the sequence (ν_n) by a constant to obtain a sequence of probability measures. Thus Theorem A.2.3 implies that there exists a nonnegative measure ν and a subsequence (ν_{n_k}) of (ν_n) such that

$$\lim_{k \to \infty} \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu_{n_k}(\mathrm{d}\lambda) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu(\mathrm{d}\lambda), \, .$$

Using (5.11) and taking the limit of the subsequence, we get that

$$\gamma(h) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu(\mathrm{d}\lambda), \ \forall h \in \mathbb{Z} \, .$$

Let us conclude with the uniqueness of ν. Suppose that another nonnegative measure ξ satisfies for all $h \in \mathbb{Z}$: $\int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \mu(\mathrm{d}\lambda)$. Then by Theorem 1.3.1, we obtain that $\int_{\mathbb{T}} g(\lambda) \nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} g(\lambda) \mu(\mathrm{d}\lambda)$ for all continuous (2π)-periodic function g. This implies $\nu = \mu$. □

Corollary 5.3.2 (The ℓ^2 case)**.** *Let $(\gamma(h))_{h \in \mathbb{Z}} \in \ell^2(\mathbb{Z})$. Then it is a nonnegative definite hermitian sequence in the sense of Proposition 5.2.1 if and only if*

$$f(\lambda) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \gamma(h) \mathrm{e}^{-\mathrm{i}h\lambda} \, ,$$

where the convergence holds in $L^2(\mathbb{T})$, is nonnegative for almost every λ.

Proof. It suffices to apply Theorem 5.3.1 and Corollary 1.3.3. □

The proof also shows that f is the spectral density function associated to γ.

Example 5.3.1 (MA(1), Continued from Example 5.2.4)**.** *Consider Example 5.2.4. Then* $(\chi(h))$ *is in* $\ell^1(\mathbb{Z})$ *and*

$$f(\lambda) = \frac{1}{2\pi} \sum_h \chi(h) \mathrm{e}^{-\mathrm{i}h\lambda} = \frac{1}{2\pi}(1 + 2\rho\cos(\lambda)) \ .$$

Thus we obtain that χ *is nonnegative definite if and only if* $|\rho| \leq 1/2$*. An exemple of such a spectral density function is displayed in Figure 5.3.*

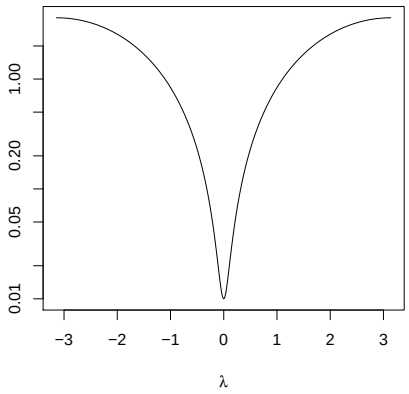

Figure 5.3: *Spectral density function (in logarithmic scale) of an MA(1) process, as given by (5.4) with* $\sigma = 1$ *and* $\theta = -0.9$*.*

Example 5.3.2 (Spectral density function of a white noise)**.** *Recall the definition of a white noise, Definition 5.2.3. We easily get that the white noise* $\mathrm{IID}(0, \sigma^2)$ *admits a spectral density function given by*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \ ,$$

that is, a constant spectral density function. Hence the name "white noise", referring to white color that corresponds to a constant frequency spectrum.

Example 5.3.3 (Spectral measure of an harmonic process, continued from Example 5.2.2)**.** *The autocovariance function of* X *is given by (see Example 5.2.2)*

$$\gamma(h) = \frac{1}{2} \sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k h) \ , \tag{5.12}$$

where $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$*. Observing that*

$$\cos(\lambda_k h) = \frac{1}{2} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}h\lambda} (\delta_{\lambda_k}(\mathrm{d}\lambda) + \delta_{-\lambda_k}(\mathrm{d}\lambda))$$

where $\delta_{x_0}(\mathrm{d}\lambda)$ denote the Dirac mass at point x_0, the spectral measure of X reads

$$\nu(\mathrm{d}\lambda) = \frac{1}{4}\sum_{k=1}^{N} \sigma_k^2 \delta_{\lambda_k}(\mathrm{d}\lambda) + \frac{1}{4}\sum_{k=1}^{N} \sigma_k^2 \delta_{-\lambda_k}(\mathrm{d}\lambda) \; .$$

We get a sum of Dirac masses with weights σ_k^2 and located at the frequencies of the harmonic functions.

Harmonic processes have singular properties. The autocovariance function in (5.12) implies that covariance matrices Γ_n are expressed as a sum of $2N$ matrices with rank 1. Thus Γ_n is not invertible as soon as $n > 2N$ and thus harmonic process fall in the following class of process.

Definition 5.3.1 (Linearly predictable processes)**.** *A weakly stationary process X is called* linearly predictable *if there exists $n \geq 1$ such that for all $t \geq n$, $X_t \in \mathrm{Span}\,(X_1, \ldots, X_n)$ (in the L^2 sense).*

One can wonder whether the other given examples are linearly predictable. The answer is given by the following result, whose proof is left to the reader (see Exercise 5.8).

Proposition 5.3.3. *Let γ be the autocovariance function of a weakly stationary process X. If $\gamma(0) \neq 0$ and $\gamma(t) \to 0$ as $t \to \infty$ then X is not linearly predictable.*

5.4 Spectral representation of weakly stationary processes

5.4.1 Random fields with orthogonal increments

In this section, we let $(\mathbb{X}, \mathcal{X})$ denote any measurable space.

Definition 5.4.1 (Random fields with orthogonal increments)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A* random field with orthogonal increments *W on $(\mathbb{X}, \mathcal{X})$ is a L^2 random process indexed on $\mathcal{X}$, say $W = (W(A))_{A\in\mathcal{X}}$ such that*

(i) For all $A \in \mathcal{X}$, $\mathbb{E}\,[W(A)] = 0$.

(ii) For all $A, B \in \mathcal{X}$ such that $A \cap B = \emptyset$, $W(A)$ and $W(B)$ are uncorrelated and $W(A \cup B) = W(A) + W(B)$;

(iii) For all nonincreasing sequence $(A_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$, we have $\mathrm{Var}\,(W(A_n)) \to 0$.

Lemma 5.4.1. *Let W a random field with orthogonal increments on $(\mathbb{X}, \mathcal{X})$. Let $A \in \mathcal{X}$, and set $\eta(A) = \mathrm{Var}\,(W(A))$. Then η is a finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$. Moreover, for all $A, B \in \mathcal{X}$, $\mathrm{Cov}\,(W(A), W(B)) = \eta(A \cap B)$.*

Proof. To show that η is a measure, it is sufficient to show that η is additive and continuous, that is, for all nonincreasing sequence $(A_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$, we have $\eta(A_n) = 0$. These two properties follow from (*ii*) and (*iii*) of Definition 5.4.1.

Observe that $A = (A \setminus B) \cup (A \cap B)$ and $B = (B \setminus A) \cup (A \cap B)$ and that $A \setminus B$, $B \setminus A$ and $A \cap B$ are disjoint sets. By (*ii*) in Definition 5.4.1, we have $W(A) = W(A \setminus B) + W(A \cap B)$

and $W(B) = W(B \setminus A) + W(A \cap B)$ and, moreover, $W(A \setminus B)$, $W(B \setminus A)$ are $W(A \cap B)$ are uncorrelated. Consequently,

$$\mathrm{Cov}\left(W(A), W(B)\right) = \mathrm{Var}\left(W(A \cap B)\right) = \eta(A \cap B) \ ,$$

which concludes the proof. □

The measure η is called the *intensity measure* of W. The previous lemma comes with the converse following result.

Lemma 5.4.2. *Let W be a L^2 random process indexed by $\mathcal{X}$ such that, for all $A \in \mathcal{X}$, $\mathbb{E}\left[W(A)\right] = 0$. Suppose that there exists a measure η on $(\mathbb{X}, \mathcal{X})$ such that, for all $A, B \in \mathcal{X}$, $\mathrm{Cov}\left(W(A), W(B)\right) = \eta(A \cap B)$. Then W is a random field with orthogonal increments on $(\mathbb{X}, \mathcal{X})$ with intensity measure η.*

Proof. Let $A, B \in \mathcal{X}$ such that $A \cap B = \emptyset$. We have that

$$\begin{aligned}\mathrm{Var}\left(W(A \cup B) - W(A) - W(B)\right) \\ = \eta(A \cup B) + \eta(A) + \eta(B) - 2\eta(A) - 2\eta(B) + 2\eta(A \cap B) = 0 \ ,\end{aligned}$$

where we used $\eta(A \cup B) = \eta(A) + \eta(B)$ and $\eta(A \cap B) = 0$. Thus $W(A \cup B) = W(A) + W(B)$ and the additivity property (ii) is satisfied.

Consider a nonincreasing sequence $(A_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$. Since η is a measure, we have $\mathrm{Var}\left(W(A_n)\right) = \eta\left(A_n\right) \to 0$ which gives (iii). Hence the result. □

Example 5.4.1 (Randomly weighted sum of Dirac masses)**.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a centered complex valued L^2 random process defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Denote for all $n \geq 0$, $\sigma_n^2 = \mathrm{Var}\left(Y_n\right)$ and assume that $(\sigma_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ and $\mathrm{Cov}\left(Y_n, Y_k\right) = 0$ for $n \neq k$. Let $(\lambda_n)_{n\in\mathbb{N}} \subset \mathbb{X}$. Define the process W indexed by $\mathcal{X}$ as*

$$W = \sum_{n=0}^{\infty} Y_n \delta_{\lambda_n} \ ,$$

where δ_x is the Dirac mass at x. Then, for all $A, B \in \mathcal{X}$,

$$\mathrm{Cov}\left(W(A), W(B)\right) = \sum_{n=0}^{\infty} \sigma_n^2 \mathbb{1}_A(\lambda_n)\mathbb{1}_B(\lambda_n) = \eta(A \cap B) \ ,$$

where

$$\eta(A) = \sum_{n=0}^{\infty} \sigma_n^2 \delta_{\lambda_n}(A) \ .$$

By Lemma 5.4.2, W is a random field with orthogonal increments on $(\mathbb{X}, \mathcal{X})$ with intensity measure η.

5.4.2 Stochastic integral

Theorem 5.4.3. *Let W be a random field with orthogonal increments on $(\mathbb{X}, \mathcal{X})$ with intensity measure η. Then there exists a unique isometric operator w from $L^2(\mathbb{X}, \mathcal{X}, \eta)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathrm{w}(\mathbb{1}_A) = W(A)$ for all $A \in \mathcal{X}$.*

For all $f \in L^2(\mathbb{X}, \mathcal{X}, \eta)$, we further have $\mathbb{E}\left[\mathrm{w}(f)\right] = 0$ and we have

$$\mathrm{w}(L^2(\mathbb{X}, \mathcal{X}, \eta)) = \overline{\mathrm{Span}}\left(W(A), A \in \mathcal{X}\right) \ ,$$

where the closure is understood in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Proof. Set $\mathcal{H} = L^2(\mathbb{X}, \mathcal{X}, \eta)$ and $\mathcal{I} = L^2(\Omega, \mathcal{F}, \mathbb{P})$. For $A, B \in \mathcal{X}$, we have

$$\langle \mathbb{1}_A, \mathbb{1}_B \rangle_{\mathcal{H}} = \int \mathbb{1}_A \mathbb{1}_B \mathrm{d}\eta = \langle W(A), W(B) \rangle_{\mathcal{I}} \ .$$

Since by Proposition 1.2.4, we have

$$\overline{\mathrm{Span}}\,(\mathbb{1}_A, A \in \mathcal{X}) = L^2(\mathbb{X}, \mathcal{X}, \eta) \ ,$$

the result follows from the extension theorem for isometric operators (see Theorem 1.6.2). □

It can be convenient to use the same notation for W and the isometric operator w. Since $f \mapsto \mathrm{w}(f)$ is a linear operator on a set of unctions, it is also common to use the integral notation although this integral relies on a particular L^2 construction (and thus do not satisfy the usual nice properties of the classical integral),

$$\int f \, \mathrm{d}W = \int f(\lambda) \, \mathrm{d}W(\lambda) = W(f) = \mathrm{w}(f) \ , \tag{5.13}$$

which satisfies, for all $(f, g) \in L^2(\mathbb{X}, \mathcal{X}, \eta)$ and $(u, v) \in \mathbb{C} \times \mathbb{C}$,

$$\int (uf + vg) \, \mathrm{d}W = u \int f \, \mathrm{d}W + v \int g \, \mathrm{d}W \ .$$

and

$$\mathbb{E}\left[\left(\int f \, \mathrm{d}W \right) \overline{\left(\int g \, \mathrm{d}W \right)} \right] = \int f \bar{g} \, \mathrm{d}\eta \ .$$

Moreover; since $\mathbb{E}\left[W(A)\right] = 0$ for all $A \in \mathcal{X}$ and $\mathbb{E}$ is continuous on $L^2(\mathbb{X}, \mathcal{X}, \eta)$, we have

$$\mathbb{E}\left[\int f \, \mathrm{d}W \right] = 0 \ .$$

We will call $\int f \, \mathrm{d}W$ the *stochastic integral* of f with respect to W.

Interestingly, for any finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$, all isometric operators from $L^2(\mathbb{X}, \mathcal{X}, \nu)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ can be interpreted as a stochastic integral with intensity measure ν.

Theorem 5.4.4. *Let ν be a finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$ and* J *an isometric operator from $L^2(\mathbb{X}, \mathcal{X}, \nu)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that for all $f \in L^2(\mathbb{X}, \mathcal{X}, \nu)$, $\mathbb{E}\left[\mathrm{J}(f)\right] = 0$. Then there exists a random field W with orthogonal increments on $\mathbb{X}$ with intensity measure ν such that, for all $f \in L^2(\mathbb{X}, \mathcal{X}, \nu)$, $\mathrm{J}(f) = \int_{\mathbb{X}} f \, \mathrm{d}W$.*

Proof. To obtain W, we must set, for all $A \in \mathcal{X}$, $W(A) = \mathrm{J}(\mathbb{1}_A)$. Since J is an isometric operator, for all $A, B \in \mathcal{X}$,

$$\mathrm{Cov}\left(W(A), W(B)\right) = \langle \mathrm{J}(\mathbb{1}_A), \mathrm{J}(\mathbb{1}_B) \rangle_{L^2(\Omega, \mathcal{F}, \mathbb{P})} = \langle \mathbb{1}_A, \mathbb{1}_B \rangle_{L^2(\mathbb{X}, \mathcal{X}, \nu)} = \nu(A \cap B) \ .$$

and by Lemma 5.4.2, W is a random field with orthogonal increments on $\mathbb{X}$ with intensity measure ν and it is the unique one whose integral coincides with J on $\{\mathbb{1}_A, \, A \in \mathcal{X}\}$. Since both are isometric operators and

$$\overline{\mathrm{Span}}\,(\mathbb{1}_A, \, A \in \mathcal{X}) = L^2(\mathbb{X}, \mathcal{X}, \nu) \ ,$$

they coincide on the whole space $L^2(\mathbb{X}, \mathcal{X}, \nu)$, which achieves the proof. □

5.4.3 Spectral representation based on the spectral field

We now introduce the spectral field associated to a weakly stationary process. It is a random field with orthogonal increments defined on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. We first start from such a random field.

Proposition 5.4.5. *Let W be a random field with orthogonal increments defined on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ with intensity measure η. Then, the sequence $(X_t)_{t\in\mathbb{Z}}$ defined by*

$$X_t = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}t\lambda} \, \mathrm{d}W(\lambda) \, ,$$

is a centered weakly stationary process with spectral measure η.

Proof. Define $f_t(\lambda) = \mathrm{e}^{\mathrm{i}t\lambda}$ for all $t \in \mathbb{Z}$, so that $f_t \in L^2(\mathbb{T}, \eta)$. Since the stochastic integral is an isometric operator, we have, for all $(s,t) \in \mathbb{Z}^2$,

$$\mathrm{Cov}\,(X_s, X_t) = \mathbb{E}\left[X_s \bar{X}_t\right] = \langle W(f_s), W(f_t)\rangle_{L^2(\Omega,\mathcal{F},\mathbb{P})} = \langle f_s, f_t\rangle_{L^2(\mathbb{T},\eta)} = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}(s-t)\lambda}\eta(\mathrm{d}\lambda) \, .$$

Hence the result. □

Conversely, let us show that any centered weakly stationary process can be expressed as in Proposition 5.4.5.

Definition 5.4.2 (Linear closure of a random process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a L^2 process. Its* linear closure*, denoted by $\mathcal{H}_\infty^X$ is defined as*

$$\mathcal{H}_\infty^X = \overline{\mathrm{Span}}\,(X_t,\, t \in \mathbb{Z}) \, ,$$

where the closure is understood in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

In other words, $\mathcal{H}_\infty^X$ is the space of all L^2 random variables that can be obtained as an L^2 limit of a sequence of finite linear combinations of $(X_t)_{t\in\mathbb{Z}}$.

Theorem 5.4.6 (Spectral representtion)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with spectral measure ν. Then there exists a random field $\hat{X}$ with orthogonal increments defined on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ with intensity measure ν, which we call the* spectral field*, such that, for all $t \in \mathbb{Z}$,*

$$X_t = \int \mathrm{e}^{\mathrm{i}t\lambda} \, \mathrm{d}\hat{X}(\lambda) \, .$$

Moreover, the mapping $f \mapsto \int f \, \mathrm{d}\hat{X}$ defines the unique unitary operator from $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ to $\mathcal{H}_\infty^X$ that maps each function $\lambda \mapsto \mathrm{e}^{\mathrm{i}t\lambda}$ to X_t.

Proof. Set $\mathcal{H} = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$, $\mathcal{I} = L^2(\Omega, \mathcal{F}, \mathbb{P})$ and, for all $t \in \mathbb{Z}$, $f_t(\lambda) = \mathrm{e}^{\mathrm{i}t\lambda}$. We shall consider the sequences $(f_t)_{t\in\mathbb{Z}}$ and $(X_t)_{t\in\mathbb{Z}}$ in $\mathcal{H}$ and $\mathcal{I}$, respectively. By Corollary 1.3.2, we have $\overline{\mathrm{Span}}\,(f_t,\ t \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ and by Theorem 5.3.1, for all $(s,t) \in \mathbb{Z}^2$,

$$\langle f_s, f_t\rangle_{L^2(\mathbb{T},\mathcal{B}(\mathbb{T}),\nu)} = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}s\lambda}\mathrm{e}^{-\mathrm{i}t\lambda}\nu(\mathrm{d}\lambda) = \mathrm{Cov}\,(X_s, X_t) = \langle X_s, X_t\rangle_{L^2(\Omega,\mathcal{F},\mathbb{P})} \, .$$

By Theorem 1.6.2, there exists a unique isometric operator S_X from $\overline{\mathrm{Span}}\,(f_t,\ t \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu) = \mathcal{H}$ to $\mathcal{I}$ such that, for all $t \in \mathbb{Z}$, $\mathrm{S}_X(f_t) = X_t$. As an operator from $\mathcal{H}$ to $\mathrm{S}_X(\mathcal{H}) = \overline{\mathrm{Span}}\,(X_t,\, t \in \mathbb{Z}) = \mathcal{H}_\infty^X$, it is unitary (since then it is isometric and surjective).

Applying Theorem 5.4.4, we obtain the random field $\hat{X}$, which satisfies all the claimed properties. This concludes the proof. □

The reader may legitimately wonder where all this abstract framework take us to. In fact it will be very useful to express standard linear filters on weakly stationary processes. The reason is the following. Using the spectral field $\hat{X}$, we obtain a *spectral representation* of any random variable $Y \in \mathcal{H}_\infty^X$ as the stochastic integral of a function $f \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$,

$$Y = \int f \, \mathrm{d}\hat{X} \ .$$

Moreover this defines f uniquely since we have a bijection. It follows that a linear operator on $\mathcal{H}_\infty^X$ can equivalently be seen as a linear operator on $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$.

Let $H_\infty^X = \mathrm{Span}\,(X_t,\, t \in \mathbb{Z})$, so that its closure in L^2 is $\mathcal{H}_\infty^X$. Suppose that X is not linearly predictable, so that any element $Y \in H_\infty^X$ has a unique representation

$$Y = \sum_{t \in \mathbb{Z}} \lambda_t X_t \ ,$$

where $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ with finite support.

Take now a shift-invariant linear filter F on $\mathbb{C}^{\mathbb{Z}}$. A linear operator $\tilde{\mathrm{F}}$ is obtained on H_∞^X by setting, for any $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ with finite support,

$$\tilde{\mathrm{F}}\left(\sum_{t\in\mathbb{Z}} \lambda_t X_t\right) = \Pi_0 \circ \mathrm{F}\left(\sum_{t\in\mathbb{Z}} \lambda_t S^t(X)\right) = \sum_{t\in\mathbb{Z}} \lambda_t \Pi_t \circ \mathrm{F}(X_t) \ ,$$

since $\Pi_0 \circ \mathrm{F} \circ S^t = \Pi_t \circ \mathrm{F}$. In other words, $\tilde{\mathrm{F}}(X_t) = \Pi_t \circ \mathrm{F}(X)$ and it is extended linearly to H_∞^X. If $\tilde{\mathrm{F}}$ is continuous on H_∞^X as a subset of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, then it admits a unique continuous extension to $\mathcal{H}_\infty^X$. Using the spectral representation, the operator $\tilde{\mathrm{F}}$ can as well be studied on $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$, in particular its continuity properties.

Unfortunately the confusion between the original operator F and its induction $\tilde{\mathrm{F}}$ on $\mathcal{H}_\infty^X$ is widespread in the literature on time series.

Example 5.4.2 (Backshift operator)**.** *Consider the backshift operator* $\mathrm{B} = \mathrm{S}^{-1}$*, see Definition 4.3.1. In this case, the induced operator on* H_∞^X *is defined by*

$$\tilde{\mathrm{B}}(X_t) = X_{t-1} \ ,$$

At first sight, it seems hard to express this operator for any $Y \in H_\infty^X$*. However using the spectral representation,*

$$Y = \int f \, \mathrm{d}\hat{X} \ ,$$

we immediately get

$$\tilde{\mathrm{B}}(Y) = \int f(\lambda) \mathrm{e}^{-\mathrm{i}\lambda} \, \hat{X}(\mathrm{d}\lambda) \ ,$$

first for any $f : \lambda \mapsto \mathrm{e}^{\mathrm{i}t\lambda}$ *with* $t \in \mathbb{Z}$*, then for* f *being a trigonometric polynomial* $\lambda \mapsto \sum_{t\in\mathbb{Z}} \lambda_t \mathrm{e}^{\mathrm{i}t\lambda}$*, where* $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ *with finite support, and finally for any* $f \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ *by continuity extension.*

5.5 Innovation process

In this section, we let $X = (X_t)_{t\in\mathbb{Z}}$ denote a centered weakly stationary processes. We shall define the Wold decomposition of X. This decomposition mainly relies on the concept of innovations. Let

$$\mathcal{H}_t^X = \overline{\mathrm{Span}}\,(X_s, s \le t)$$

denote the *linear past* of a given random process $X = (X_t)_{t\in\mathbb{Z}}$ up to time t. It is related to the already mentioned space $\mathcal{H}_\infty^X$ as follows

$$\mathcal{H}_\infty^X = \overline{\bigcup_{t\in\mathbb{Z}} \mathcal{H}_t^X}\ .$$

Let us introduce the *innovations* of a weakly stationary process.

Definition 5.5.1 (Innovation process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. We call* innovation process *the process $\epsilon = (\epsilon_t)_{t\in\mathbb{Z}}$ defined by*

$$\epsilon_t = X_t - \mathrm{proj}\left(X_t | \mathcal{H}_{t-1}^X\right)\ . \tag{5.14}$$

By the orthogonal principle (see Theorem 1.4.1), each ϵ_t is characterized by the fact that $X_t - \epsilon_t \in \mathcal{H}_{t-1}^X$ (which implies $\epsilon_t \in \mathcal{H}_t^X$) and $\epsilon_t \perp \mathcal{H}_{t-1}^X$. As a consequence $(\epsilon_t)_{t\in\mathbb{Z}}$ is a centered orthogonal sequence. We shall see below that it is in fact a white noise, that is, the variance of the innovation

$$\sigma^2 = \|\epsilon_t\|^2 = \mathbb{E}\left[|\epsilon_t|^2\right] \tag{5.15}$$

does not depend on t.

Example 5.5.1 (Innovation process of a white noise)**.** *The innovation process of a white noise $X \sim \mathrm{WN}(0, \sigma^2)$ is $\epsilon = X$.*

Example 5.5.2 (Innovation process of a MA(1), continued from Example 5.2.1)**.** *Consider the process X defined in Example 5.2.1. Observe that $Z_t \perp \mathcal{H}_{t-1}^X$. Thus, if $\theta Z_{t-1} \in \mathcal{H}_{t-1}^X$, we immediately get that $\epsilon_t = Z_t$. The questions are thus: is Z_{t-1} in $\mathcal{H}_{t-1}^X$? and, if not, what can be done to compute ϵ_t ?*

Because the projection in (5.14) is done on an infinite dimension space, it is interesting to compute it as a limit of finite dimensional projections. To this end, define, for $p \ge 0$, the finite dimensional space

$$\mathcal{H}_{t,p}^X = \mathrm{Span}\,(X_s, t-p < s \le t)\ ,$$

and observe that $(\mathcal{H}_{t,p}^X)_p$ is an increasing sequence of linear space whose union has closure $\mathcal{H}_t^X$. Hence by Property (ii) in Theorem 1.4.3, we have, for any L^2 variable Y,

$$\lim_{p\to\infty} \mathrm{proj}\left(Y | \mathcal{H}_{t,p}^X\right) = \mathrm{proj}\left(Y | \mathcal{H}_t^X\right)\ , \tag{5.16}$$

where the limit holds in the L^2 sense.

Definition 5.5.2 (Prediction coefficients)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. We call the* predictor of order p *the random variable* $\mathrm{proj}\left(X_t | \mathcal{H}_{t-1,p}^X\right)$ *and the* partial innovation process of order p *the process $\epsilon_p^+ = \left(\epsilon_{t,p}^+\right)_{t\in\mathbb{Z}}$ defined by*

$$\epsilon_{t,p}^+ = X_t - \mathrm{proj}\left(X_t | \mathcal{H}_{t-1,p}^X\right)\ .$$

The prediction coefficients *are any coefficients* $\phi_p^+ = \left(\phi_{k,p}^+\right)_{k=1,\dots,p}$ *which satisfy, for all* $t \in \mathbb{Z}$,

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}^X\right) = \sum_{k=1}^{p} \phi_{k,p}^+ X_{t-k} \ . \tag{5.17}$$

Observe that, by the orthogonality principle, (5.17) is equivalent to

$$\Gamma_p^+ \phi_p^+ = \gamma_p^+ \ , \tag{5.18}$$

where $\gamma_p^+ = [\gamma(1), \gamma(2), \cdots, \gamma(p)]^T$ and

$$\begin{aligned}\Gamma_p^+ &= \operatorname{Cov}\left([X_{t-1} \ \dots \ X_{t-p}]^T\right)^T \\ &= \begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & & \gamma(-p+1) \\ \gamma(1) & \gamma(0) & \gamma(-1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(-1) \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix} ,\end{aligned}$$

Observing that Equation (5.18) does not depend on t and that the orthogonal projection is always well defined, such coefficients $\left(\phi_{k,p}^+\right)_{k=1,\dots,p}$ always exist. However they are uniquely defined if and only if Γ_p^+ is invertible.

Let us now compute the variance of the order-p prediction error $\epsilon_{t,p}^+$, denoted as

$$\sigma_p^2 = \| X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \|^2 = \mathbb{E}\left[\left| X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \right|^2 \right] \ . \tag{5.19}$$

By (5.17) and Proposition 1.4.2, we have

$$\begin{aligned}\sigma_p^2 &= \langle X_t, X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \rangle \\ &= \gamma(0) - \sum_{k=1}^{p} \overline{\phi_{k,p}^+} \gamma(k) \\ &= \gamma(0) - (\phi_p^+)^H \gamma_p^+ \ . \end{aligned} \tag{5.20}$$

Equations (5.18) and (5.20) are called *Yule-Walker equations*. An important consequence of these equations is that σ_p^2 does not depend on t, and since (5.16) implies

$$\sigma^2 = \lim_{p \to \infty} \sigma_p^2 \ ,$$

we obtain that, as claimed above, the variance of the innovation defined in (5.15) is also independent of t. So we can state the following result.

Corollary 5.5.1. *The innovation process of a centered weakly stationary process* X *is a (centered) weak white noise. Its variance is called the* innovation variance of the process X.

The innovation variance is not necessarily positive, that is, the innovation process can be zero a.s., as shown by the following example.

Example 5.5.3 (Innovations of the harmonic process (continued from Example 5.2.2))**.** *Consider the harmonic process $X_t = A\cos(\lambda_0 t + \Phi)$ where A is a centered random variable with finite variance σ_A^2 and Φ is a random variable, independent of A, with uniform distribution on $(0, 2\pi)$. Then X is a centered weakly stationary process with autocovariance function $\gamma(\tau) = (\sigma_A^2/2)\cos(\lambda_0\tau)$. The prediction coefficients or order 2 are given by*

$$\begin{bmatrix}\phi_{1,2}^+\\ \phi_{2,2}^+\end{bmatrix} = \begin{bmatrix}1 & \cos(\lambda_0)\\ \cos(\lambda_0) & 1\end{bmatrix}^{-1}\begin{bmatrix}\cos(\lambda_0)\\ \cos(2\lambda_0)\end{bmatrix} = \begin{bmatrix}2\cos(\lambda_0)\\ -1\end{bmatrix}$$

We then obtain that $\sigma_2^2 = \|X_t - \mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,2}^X\right)\|^2 = 0$ and thus

$$X_t = \mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,2}^X\right) = 2\cos(\lambda_0)X_{t-1} - X_{t-2} \in \mathcal{H}_{t-1}^X$$

Hence in this case the innovation process is zero: one can exactly predict the value of X_t from its past.

The latter example indicates that the harmonic process is *deterministic*, according to the following definition.

Definition 5.5.3 (Regular/deterministic process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. If the variance of its innovation process is zero, we say that X is* deterministic*. Otherwise, we say that X is* regular*.*

Let us define the intersection of the whole past of the process X as

$$\mathcal{H}_{-\infty}^X = \bigcap_{t\in\mathbb{Z}} \mathcal{H}_t^X \ .$$

Note that this (closed) linear space may not be null. Take a deterministic process X such as the harmonic process above. Then $X_t \in \mathcal{H}_{t-1}^X$, which implies that $\mathcal{H}_t^X = \mathcal{H}_{t-1}^X$. Thus, for a deterministic process, we have, for all t, $\mathcal{H}_{-\infty}^X = \mathcal{H}_t^X$, and thus also, $\mathcal{H}_{-\infty}^X = \mathcal{H}_\infty^X$, which is of course never null unless $X = 0$ a.s.

Example 5.5.4 (Constant process)**.** *A very simple example of deterministic process is obtained by taking $\lambda_0 = 0$ in Example 5.5.3. In other words, $X_t = X_0$ for all $t \in \mathbb{Z}$.*

For a regular process, things are a little bit more involved. For the white noise, it is clear that $\mathcal{H}_{-\infty}^X = \{0\}$. In this case, we say that X is *purely non-deterministic*. However not every regular process is purely nondeterministic. Observe indeed that for two uncorrelated centered and weakly stationary process X and Y, setting $Z = X + Y$, which is also centered and weakly stationary, we have, for all $t \in \mathbb{Z}$

$$\mathcal{H}_t^Z \subseteq \mathcal{H}_t^X \overset{\perp}{\oplus} \mathcal{H}_t^Y \ .$$

This implies that

$$\mathcal{H}_{-\infty}^Z \subseteq \mathcal{H}_{-\infty}^X \overset{\perp}{\oplus} \mathcal{H}_{-\infty}^Y \ . \tag{5.21}$$

Also, by Proposition 1.4.2, the innovation variance of Z is larger than the sum of the innovations variances of X and Y. From these facts, we have that the sum of two uncorrelated processes is regular if at least one of them is regular and it is purely non-deterministic if both are purely non-deterministic. A regular process which is not purely nondeterministic can easily be obtained as follows.

Example 5.5.5 (Uncorrelated sum of a white noise with a constant process)**.** *Define $Z = X + Y$ with $X \sim \mathrm{WN}(0, \sigma^2)$ and $Y_t = Y_0$ for all t, where Y_0 is centered with positive variance and uncorrelated with $(X_t)_{t\in\mathbb{Z}}$. Then by (5.21), $\mathcal{H}^Z_{-\infty} \subseteq \mathrm{Span}\,(Y_0)$. Moreover, it can be shown (see Exercise 5.9) that $Y_0 \in \mathcal{H}^Z_{-\infty}$ and thus $X_t = Z_t - Y_0 \in \mathcal{H}^Z_t$. Hence we obtain $\mathcal{H}^Z_{-\infty} = \mathrm{Span}\,(Y_0)$, so that Z is not purely non-deterministic and Z has innovation X, so that Z is regular.*

In fact, the Wold decomposition indicates that the configuration of Example 5.5.5 is the only one: every regular process is the sum of two uncorrelated processes: one which is deterministic, the other which is purely nondeterministic. Before stating this result we introduce the following coefficients, defined for any regular process X,

$$\psi_s = \frac{\langle X_t, \epsilon_{t-s}\rangle}{\sigma^2}\ , \tag{5.22}$$

where ϵ is the innovation process and σ^2 its variance. By weak stationarity of X, this coefficient do no depend on t but only on k, since

$$\begin{aligned}\langle X_t, \epsilon_{t-k}\rangle &= \gamma(k) - \mathrm{Cov}\left(X_t, \mathrm{proj}\left(X_{t-k}|\,\mathcal{H}^X_{t-k-1}\right)\right)\\ &= \gamma(k) - \lim_{p\to\infty} \mathrm{Cov}\left(X_t, \mathrm{proj}\left(X_{t-k}|\,\mathcal{H}^X_{t-k-1,p}\right)\right)\\ &= \gamma(k) - \lim_{p\to\infty} \sum_{j=1}^{p} \phi_{j,p}\gamma(k+j)\ .\end{aligned}$$

It is easy to show that $\psi_0 = 1$. Moreover, since ϵ is a white noise, we have, for all $t \in \mathbb{Z}$,

$$\mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t) = \sum_{k\geq 0} \psi_k \epsilon_{t-k}\ .$$

We can now state the Wold decomposition.

Theorem 5.5.2 (Wold decomposition)**.** *Let X be a regular process and let ϵ be its innovation process and σ^2 its innovation variance, so that $\epsilon \sim \mathrm{WN}(0, \sigma^2)$. Define the L^2 centered process U as*

$$U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}\ ,$$

where ψ_k is defined by (5.22). Defined the L^2 centered process V by the following equation:

$$X_t = U_t + V_t, \quad \textit{for all } t \in \mathbb{Z}. \tag{5.23}$$

Then the following assertions hold.

(i) We have $U_t = \mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t)$ and $V_t = \mathrm{proj}\left(X_t|\,\mathcal{H}^X_{-\infty}\right)$.

(ii) ϵ and V are uncorrelated: for all (t, s), $\langle V_t, \epsilon_s\rangle = 0$.

(iii) U is a purely non-deterministic process and has same innovation as X. Moreover, $\mathcal{H}^\epsilon_t = \mathcal{H}^U_t$ for all $t \in \mathbb{Z}$.

(iv) V is a deterministic process and $\mathcal{H}^V_{-\infty} = \mathcal{H}^X_{-\infty}$.

Proof. The proof mainly relies on the facts established above and on Theorem 1.4.3. Notice that, for all $s \leq t$ we have

$$\mathcal{H}_s^X \overset{\perp}{\oplus} \mathrm{Span}\,(\epsilon_k,\, s < k \leq t) = \mathcal{H}_t^X \ .$$

Then, by Theorem 1.4.3, we get that

$$\mathcal{H}_{-\infty}^X \overset{\perp}{\oplus} \mathcal{H}_t^\epsilon = \mathcal{H}_t^X \ .$$

The facts then easily follow. The details are left to the reader (see Exercise 5.10). □

5.6 Exercises

Exercise 5.1. Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ be two second order stationary processes that are uncorrelated in the sense that X_t and Y_s are uncorrelated for all t, s. Show that $Z_t = X_t + Y_t$ is a second order stationary process. Compute its autocovariance function, given the autocovariance functions of X and Y. Do the same for the spectral measures.

Exercise 5.2. Consider the processes of Exercise 4.4, with the additional assumption that $(\varepsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$. Determine in each case, if the defined process is weakly stationary. In the case of Question 4, consider also $Z_t = Y_t^2$ under the assumption $\mathbb{E}[\varepsilon_0^4] < \infty$.

Exercise 5.3. Define χ as in (5.7).

1. For which values of ρ is χ an autocovariance function ? [Hint : use the Hertglotz theorem].
2. Exhibit a Gaussian process with autocovariance function χ.

Exercise 5.4. For $t \geq 2$, define

$$\Sigma_2 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \ldots, \Sigma_t = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$$

1. For which values of ρ, is Σ_t guaranteed to be a covariance matrix for all values of t [Hint: write Σ_t as $\alpha I + A$ where A has a simple eigenvalue decomposition]?
2. Define a stationary process whose finite-dimensional covariance matrices coincide with Σ_t (for all $t \geq 1$).

Exercise 5.5. Let X and Y two L^2 centered random variables. Define

$$\rho = \frac{\mathrm{Cov}\,(X, Y)}{\mathrm{Var}\,(Y)}\ ,$$

with the convention $0/0 = 0$. Show that

$$\mathrm{proj}\,(X|\,\mathrm{Span}\,(Y)) = \rho\, Y \quad \text{and} \quad \mathbb{E}\left[(X - \mathrm{proj}\,(X|\,\mathrm{Span}\,(Y)))^2\right] = \mathrm{Var}\,(X) - |\rho|^2\,\mathrm{Var}\,(Y)\ .$$

Exercise 5.6. Let (Y_t) be a weakly stationary process with spectral density f such that $0 \leq m \leq f(\lambda) \leq M < \infty$ for all $\lambda \in \mathbb{R}$. For $n \geq 1$, denote by Γ_n the covariance matrix of $[Y_1, \ldots, Y_n]^T$. Show that the eigenvalues of Γ_n belong to the interval $[2\pi m, 2\pi M]$.

Exercise 5.7. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with spectral density f and denote by $\hat{X}$ its spectral representation field, so that, for all $t \in \mathbb{Z}$,

$$X_t = \int \mathrm{e}^{\mathrm{i}t\lambda}\, \mathrm{d}\hat{X}(\lambda)\ .$$

Assume that f is two times continuously differentiable and that $f(0) = 0$. Define, for all $t \geq 0$,

$$Y_t = X_{-t} + X_{-t+1} + \cdots + X_0\ .$$

1. Build an example of such a process X of the form $X_t = \epsilon_t + a\epsilon_{t-1}$ with $\epsilon \sim \mathrm{WN}(0,1)$ and $a \in \mathbb{R}$.
2. Determine g_t such that $Y_t = \int g_t \, \mathrm{d}\hat{X}$.
3. Compute
$$\lim_{n\to\infty} \int_{\mathbb{T}} \left| \frac{1}{n} \sum_{k=1}^{n} \mathrm{e}^{-\mathrm{i}k\lambda} \right|^2 \mathrm{d}\lambda$$
4. Show that
$$Z = \int (1 - \mathrm{e}^{-\mathrm{i}\lambda})^{-1} \, \mathrm{d}\hat{X}(\lambda) \, .$$
is well defined in $\mathcal{H}_\infty^X$.
5. Deduce from the previous questions that
$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y_t = Z \quad \text{in } L^2.$$
6. Show this result directly in the particular case exhibited in Question 1.

Exercise 5.8. Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with covariance function γ. Denote
$$\Gamma_t = \mathrm{Cov}\left(\left[X_1, \ldots, X_t\right]^T, = \right) \left[\gamma(i-j)\right]_{1\le i,j\le t} \qquad \text{for all } t \ge 1.$$
We temporarily assume that there exists $k \ge 1$ such that Γ_k is invertible but Γ_{k+1} is not.

1. Show that we can write X_n as $\sum_{t=1}^{k} \alpha_t^{(n)} X_t$, where $\alpha^{(n)} \in \mathbb{R}^k$, for all $n \ge k+1$.
2. Show that the vectors $\alpha^{(n)}$ are bounded independently of n.

Suppose now that $\gamma(0) > 0$ and $\gamma(t) \to 0$ as $t \to \infty$.

3. Show that, for all $t \ge 1$, Γ_t is invertible.
4. Deduce that Proposition 5.3.3 holds.

Exercise 5.9. Define $Z = X + Y$ with $X \sim \mathrm{WN}(0, \sigma^2)$ and $Y_t = Y_0$ for all t, where Y_0 is centered with positive variance and uncorrelated with $(X_t)_{t\in\mathbb{Z}}$.

1. Show that $\mathcal{H}_{-\infty}^Z \subseteq \mathrm{Span}\,(Y_0)$. [Hint : see Example 5.5.5]

Define, for all $t \in \mathbb{Z}$ and $n \ge 1$,
$$T_{t,n} = \frac{1}{n} \sum_{k=1}^{n} Z_{t-k}$$

2. What is the L^2 limit of $T_{t,n}$ as $n \to \infty$?
3. Deduce that $\mathcal{H}_{-\infty}^Z = \mathrm{Span}\,(Y_0)$.

Exercise 5.10. Define $(X_t)_{t\in\mathbb{Z}}$, $(U_t)_{t\in\mathbb{Z}}$ and $(V_t)_{t\in\mathbb{Z}}$ as in Theorem 5.5.2.

1. Show that
$$\mathcal{H}^X_{-\infty} \overset{\perp}{\oplus} \mathcal{H}^\epsilon_t = \mathcal{H}^X_t \ .$$
2. Deduce that $U_t = \mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t)$, $V_t = \mathrm{proj}\,\big(X_t|\,\mathcal{H}^X_{-\infty}\big)$ and that U and V are uncorrelated.
3. Show that $\mathcal{H}^X_{-\infty} = \mathcal{H}^V_t$ and $\mathcal{H}^\epsilon_t = \mathcal{H}^U_t$ for all $t \in \mathbb{Z}$. [Hint : observe that $\mathcal{H}^X_t \subset \mathcal{H}^U_t \oplus \mathcal{H}^V_t$ and use the previous questions]
4. Conclude the proof of Theorem 5.5.2.

Chapter 6

Martingales in discrete time

6.1 Definitions and Elementary properties

Definition 6.1.1 (Martingale)**.** *Let* $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ *be a filtered probability space and* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *a real integrable adapted process. We say that* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is*

(i) *a* martingale *if for all* $n \in \mathbb{N}$, $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$, $\mathbb{P}$-a.s.

(ii) *a* submartingale *if for all* $n \in \mathbb{N}$, $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n$, $\mathbb{P}$-a.s.

(iii) *a* supermartingale *if for all* $n \in \mathbb{N}$, $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$, $\mathbb{P}$-a.s.

We will also say that $(X_n)_{n\in\mathbb{N}}$ is a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-martingale. When the filtration is note mentioned, the usual convention is to take the natural filtration $\left(\mathcal{F}_n^X\right)_{n\in\mathbb{N}}$ of $(X_n)_{n\in\mathbb{N}}$.

We have the following immediate properties (showing them is a simple but probably useful exercise).

Proposition 6.1.1. *The following properties hold.*

(i) $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale if and only if it is a submartingale and a supermartingale.*

(ii) *If* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale (resp. submartingales, supermartingales) and* Y_0 *is* $\mathcal{F}_0$*-measurable, then* $((Y_0 + X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale (resp. submartingales, supermartingales).*

(iii) *The sum of two martingales (resp. submartingales, supermartingales) both defined with the same filtration is a martingale (resp. submartingales, supermartingales) with the same filtration.*

(iv) *If* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a submartingale if and only if* $((-X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a supermartingale.*

(v) *If* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale (resp. submartingale, supermartingale) then for all* $0 \leq m \leq n$,

$$\mathbb{E}[X_n|\mathcal{F}_m] = \quad \text{(resp. } \geq, \leq) \quad X_m \quad \mathbb{P}\text{-a.s.}$$

(vi) *If* $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale (resp. submartingale, supermartingale) then for all* $n \in \mathbb{N}$,

$$\mathbb{E}[X_n] = \quad \text{(resp. } \geq, \leq) \quad \mathbb{E}[X_0] \ .$$

Let us give some classical examples of martingales.

Example 6.1.1 (Partial sums of an independent process)**.** *Let $(\epsilon_n)_{n\in\mathbb{N}}$ be a zero mean adapted independent process and define, for all $n \in \mathbb{N}$, the partial sum S_n by*

$$S_n = \sum_{k=0}^{n} \epsilon_k \ .$$

*Then $((S_n, \mathcal{F}_n^\epsilon))_{n\in\mathbb{N}}$ is a martingale. If $(\epsilon_n)_{n\in\mathbb{N}}$ is i.i.d., then $(S_n)_{n\in\mathbb{N}}$ is often called a*random walk *(although this term is sometimes referring only to the integer valued case where ϵ is only valued in $\{-1, 1\}$).*

Example 6.1.2 (Closed martingale)**.** *Let Y be an L^1 random variable on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$. Define, for all $n \in \mathbb{N}$,*

$$Y_n = \mathbb{E}\left[Y \middle| \mathcal{F}_n\right] \ .$$

Then $((Y_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale.

We will see later that not all martingales can be defined as a closed martingale and those that can be defined as such are easily characterized.

We denote the differencing operator on $\mathbb{R}^{\mathbb{N}}$ by Δ, for all $x = (x_n)_{n\in\mathbb{N}}$, Δx is defined in $\mathbb{R}^{\mathbb{N}^*}$ by

$$\Delta x_n = x_n - x_{n-1} \quad \text{for all} \quad n \in \mathbb{N}^* \ .$$

Definition 6.1.2 (Martingale difference)**.** *A process $((Z_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ is called a* martingale difference *if it is adapted, integrable and $\mathbb{E}\left[Z_n \middle| \mathcal{F}_{n-1}\right] = 0$ for all $n \in \mathbb{N}^*$, $\mathbb{P}$-a.s.*

Obviously if $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale, then ΔX is a martingale difference. Also, if $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is an adapted integrable process and $((\Delta X_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ is a martingale difference, then $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale.

In general a deterministic transformation of a martingale is not a martingale, except if the transformation is linear. If the transformation is convex (or concave), one can still derive something.

Proposition 6.1.2. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be a martingale (resp. a submartingale) valued in an open interval $I \subset \mathbb{R}$ and f be a convex (resp. a convex increasing) $I \to \mathbb{R}$ function such that $(f(X_n))_{n\in\mathbb{N}}$ is L^1. Then $((f(X_n), \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale.*

Proof. See Exercise 6.5. □

Two important examples of application of this result are given in the following.

Corollary 6.1.3. *The following facts hold.*

(i) If, for some $p \geq 1$, $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is an L^p martingale, then $((|X_n|^p, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale.

(ii) If $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale, then for any $a \in \mathbb{R}$, $(((X_n - a)_+, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale.

6.2 Martingale transform and optional stopping theorem

Along with martingales, predictable processes are used to decompose many processes of interest. In particular they are used to define the *martingale transform*, which is the discrete time analog of the stochastic integral.

Definition 6.2.1 (Predictable process and martingale transform)**.** *The process $(A_n)_{n\in\mathbb{N}^*}$ is said to be* predictable *with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if for all $n \geq 1$, A_n est $\mathcal{F}_{n-1}$ -measurable. Given an adapted process $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$, the A-*martingale transform *of X, denoted by $A\cdot X = ((A\cdot X)_n)_{n\in\mathbb{N}}$ is the $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted process recursively defined by setting $(A\cdot X)_0 = 0$ and, for all $n \in \mathbb{N}^*$,*

$$(A\cdot X)_n = (A\cdot X)_{n-1} + A_n\,\Delta X_n\ . \tag{6.1}$$

The following result can be readily checked.

Proposition 6.2.1. *Let $(A_n)_{n\in\mathbb{N}^*}$ be an L^∞ predictable process with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and let $X = ((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an adapted process. Then the following assertions hold.*

(i) If X is a martingale , then $(((A\cdot X)_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale.

(ii) If X is a submartingale (resp. supermartingale) and A is non-negative then $(((A\cdot X)_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale (resp. supermartingale).

The following definition is extensively used for martingales.

Definition 6.2.2 (Stopped process)**.** *If $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is an adapted process and ν is a stopping time for the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, the* stopped process *$X^\nu$ is defined by $X^\nu_n = X_{n\wedge\nu}$.*

Many derivation on martingales are based on intermediary results first established on stopped martingales, thanks to the following result.

Proposition 6.2.2. *If $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale (resp. submartingale, supermartingale) and ν is a stopping time for the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, then the stopped process X^ν is a martingale (resp. submartingale, supermartingale) with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

Proof. Set $A_n = \mathbb{1}_{\{\tau\geq n\}}$. Since $\{\tau \geq n\} = \{\tau \leq n-1\}^c \in \mathcal{F}_{n-1}$, $(A_n)_{n\in\mathbb{N}^*}$ is predictable and the proof follows by noting that $X^\nu_n = X_0 + (A\cdot X)_n$ and by applying Proposition 6.2.1. □

From this result, we deduce that, for any $0 \leq m < n$, we have for a martingale (resp. submartingale, supermartingale) $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ and a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-stopping time ν,

$$\mathbb{E}\left[X_{n\wedge\nu}\middle|\,\mathcal{F}_m\right] = \quad (\text{resp. } \geq, \leq) \quad X_{m\wedge\nu} \quad \mathbb{P}\text{-a.s.}$$

Since $X_{m\wedge\nu}$ is $\mathcal{F}_{m\wedge\nu}$-measurable and $\mathcal{F}_{m\wedge\nu} \subseteq \mathcal{F}_m$, we thus get that

$$\mathbb{E}\left[X_{n\wedge\nu}\middle|\,\mathcal{F}_{m\wedge\nu}\right] = \quad (\text{resp. } \geq, \leq) \quad X_{m\wedge\nu} \quad \mathbb{P}\text{-a.s.}$$

Observe that, in this formula, $n\wedge\nu$ and $m\wedge\nu$ are two stopping times satisfying

$$0 \leq m\wedge\nu \leq n\wedge\nu \leq n\ .$$

It turns out that the formula continues to hold for any two stopping times satisfying $\underline{\nu} \leq \overline{\nu} \leq n$ for some $n \in \mathbb{N}$.

Theorem 6.2.3 (Optional stopping theorem)**.** *Suppose that $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale (resp. submartingale, supermartingale). Let $\underline{\nu}$ and $\overline{\nu}$ be two L^∞ $(\mathcal{F}_n)_{n\in\mathbb{N}}$-stopping times such that*

$$\underline{\nu} \leq \overline{\nu} \qquad \mathbb{P}\text{-a.s.}$$

Then we have

$$\mathbb{E}\left[X_{\overline{\nu}} | \mathcal{F}_{\underline{\nu}}\right] = \quad (resp.\ \geq, \leq) \quad X_{\underline{\nu}} \quad \mathbb{P}\text{-a.s.} \tag{6.2}$$

Proof. First observe that if there exists $n \in \mathbb{N}$ such that $\overline{\nu} \leq n$ $\mathbb{P}$-a.s., then $X_n^{\overline{\nu}} = X^{\overline{\nu}}$ so that in Formula (6.2), we can replace X by $X^{\overline{\nu}}$ and $\overline{\nu}$ by n. By Proposition 6.2.2, $X^{\overline{\nu}}$ and $\overline{\nu}$ has the same martingale (resp. submartingale, supermartingale) property as X.

Hence it is sufficient to prove the result in the special case where there exists $n \in \mathbb{N}$ such that $\overline{\nu} = n$ for some $n \in \mathbb{N}$. Let $A \in \mathcal{F}_{\underline{\nu}}$. Since $\underline{\nu} \leq n$ $\mathbb{P}$-a.s., and $A \cap \{\underline{\nu} = k\} \in \mathcal{F}_k$ for all $k \geq 0$, we have

$$\begin{aligned}
\mathbb{E}\left[X_n \mathbb{1}_A\right] &= \sum_{k=0}^{n} \mathbb{E}\left[X_n \mathbb{1}_{A\cap\{\underline{\nu}=k\}}\right] \\
&= \sum_{k=0}^{n} \mathbb{E}\left[\mathbb{E}\left[X_n | \mathcal{F}_k\right] \mathbb{1}_{A\cap\{\underline{\nu}=k\}}\right] \\
&= \quad (\text{resp. } \geq, \leq) \quad \sum_{k=0}^{n} \mathbb{E}\left[X_k \mathbb{1}_{A\cap\{\underline{\nu}=k\}}\right] \\
&= \quad (\text{resp. } \geq, \leq) \quad \mathbb{E}\left[X_{\underline{\nu}} \mathbb{1}_A\right] .
\end{aligned}$$

This shows (6.2) in the case $\overline{\nu} = n$ and the proof is concluded. □

Interestingly the stopping theorem allows us to extend Property (vi) of Proposition 6.1.1 to L^∞ stopping times and then we obtain a characterization of martingales.

Proposition 6.2.4. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an L^1 adapted process. It is a martingale if and only if, for all L^∞ stopping times ν, we have*

$$\mathbb{E}\left[X_\nu\right] = \mathbb{E}\left[X_0\right] \tag{6.3}$$

Proof. We apply the optional stopping theorem with $\underline{\nu} = 0$ and $\overline{\nu} = \nu$ and obtain the proof of the "only if" part.

To prove the "if" part, suppose that (6.3) holds for all L^∞ stopping times ν. Let $n \in \mathbb{N}$ and $A \in \mathcal{F}_n$. Then we can define

$$\nu = n\mathbb{1}_A + (n+1)\mathbb{1}_{A^c}$$

If $k < n$ or $k > n+1$, we have $\{\nu = k\} = \emptyset \in \mathcal{F}_k$ and, if $k = n, n+1$ $\{\nu = k\} \in \mathcal{F}_n \subseteq \mathcal{F}_k$. Hence we have

$$\mathbb{E}\left[X_\nu\right] = \mathbb{E}\left[X_n \mathbb{1}_A\right] + \mathbb{E}\left[X_{n+1}\mathbb{1}_{A^c}\right] = \mathbb{E}\left[(X_n - X_{n+1})\mathbb{1}_A\right] + \mathbb{E}\left[X_{n+1}\right] .$$

But since ν is bounded by $n+1$, (6.3) implies that $\mathbb{E}\left[X_\nu\right] = \mathbb{E}\left[X_0\right] = \mathbb{E}\left[X_{n+1}\right]$ and thus

$$\mathbb{E}\left[(X_n - X_{n+1})\mathbb{1}_A\right] = 0 .$$

This shows that $\mathbb{E}\left[X_{n+1} | \mathcal{F}_n\right] = X_n$ and thus $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a martingale. □

Well known applications of the optional stopping theorem are the Wald identities.

Proposition 6.2.5 (Wald identities)**.** *Let $(X_n)_{n\in\mathbb{N}^*}$ be an i.i.d. sequence in L^1 and σ an L^1 stopping time with respect to $\mathcal{F}^X$. Then we have*

$$\mathbb{E}\left[\sum_{k=1}^{\sigma} X_k\right] = \mathbb{E}\left[\sigma\right]\mathbb{E}\left[X_1\right] \ .$$

Suppose moreover that $\mathbb{E}\left[X_1\right] = 0$. Then

$$\mathbb{E}\left[\left(\sum_{k=1}^{\sigma} X_k\right)^2\right] = \mathbb{E}\left[\sigma\right]\mathbb{E}\left[X_1^2\right] \ .$$

Proof. See Exercises 6.7 and 6.10. □

6.3 Doob decomposition

The following proposition is based on an elementary decomposition, which is mostly used for submartingale.

Proposition 6.3.1. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an L^1 (integrable) adapted process. Then there exists a unique martingale $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ and a unique $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$–predictable process $(A_n)_{n\in\mathbb{N}^*}$ such that, setting $A_0 = 0$, we have $X_n = M_n + A_n$ for all $n \in \mathbb{N}$. Moreover the process $(A_n)_{n\in\mathbb{N}}$ is recursively defined by setting $A_0 = 0$ and, for all $n \in \mathbb{N}^*$,*

$$A_n = A_{n-1} + \mathbb{E}\left[\Delta X_n \middle| \mathcal{F}_{n-1}\right] \ .$$

The decomposition $X = A + M$ is called the Doob decomposition *of X.*

Proof. Suppose that $X = M + A$ with $A_0 = 0$ and $((A_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ a predictable process. Then $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is an adapted integrable process. Thus M is a martingale if and only of ΔM is a martingale difference. Now, we have, for all $n \in \mathbb{N}^*$,

$$\Delta A_n = \mathbb{E}\left[\Delta A_n \middle| \mathcal{F}_{n-1}\right] = \mathbb{E}\left[\Delta X_n \middle| \mathcal{F}_{n-1}\right] - \mathbb{E}\left[\Delta M_n \middle| \mathcal{F}_{n-1}\right] \ .$$

And we conclude that ΔM is a martingale difference if and only if, for all $n \in \mathbb{N}^*$,

$$\Delta A_n = \mathbb{E}\left[\Delta X_n \middle| \mathcal{F}_{n-1}\right] \ .$$

This proves the existence and uniqueness of the decomposition $X = M + A$ with $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ martingale, $A_0 = 0$ and $((A_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ predictable. □

The main interest of this decomposition is to provide a characterization of martingales, supermartingales and submartingales through the monotonousness of its predictable part.

Proposition 6.3.2. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an L^1 (integrable) adapted process and $X = M + A$ its Doob decomposition. Then*

(i) $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a supermartingale if and only if $(A_n)_{n\in\mathbb{N}}$ is non-increasing $\mathbb{P}$-a.s.*;*

(ii) $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a submartingale if and only if* $(A_n)_{n\in\mathbb{N}}$ *is non-decreasing* $\mathbb{P}$-a.s.*;*

(iii) $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *is a martingale if and only if* $(A_n)_{n\in\mathbb{N}}$ *is the zero process* $\mathbb{P}$-a.s..

Proof. Observe that, for all $n \in \mathbb{N}^*$,

$$\Delta A_n = \mathbb{E}\left[\Delta A_n \middle| \mathcal{F}_{n-1}\right] = \mathbb{E}\left[\Delta X_n \middle| \mathcal{F}_{n-1}\right] \qquad \mathbb{P}\text{-a.s.}$$

The result immediately follows. □

The Doob decomposition is extensively used for studying L^2 martingales. For such a martingale, according to Corollary 6.1.3 its square is a submartingale and the the following definition applies.

Definition 6.3.1 (Predictable Quadratic variation)**.** *Let* $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *be an* L^2 *martingale. The predictable part of the Doob decomposition of* $\left((M_n^2, \mathcal{F}_n)\right)_{n\in\mathbb{N}}$ *is called the* predictable quadratic variation *of* M *and denoted by* $\langle M\rangle = (\langle M\rangle_n)_{n\in\mathbb{N}}$*. Since* $\left((M_n^2, \mathcal{F}_n)\right)_{n\in\mathbb{N}}$ *is a submartingale,* $\langle M\rangle$ *is a non-decreasing on* $\mathbb{N}$ *(and non-negative since it starts at zero).*

By the recursive formula of Proposition 6.3.1, we have that, for all $n \in \mathbb{N}^*$,

$$\langle M\rangle_n = \sum_{j=1}^{n} \mathbb{E}\left[(\Delta M_j)^2 \middle| \mathcal{F}_{j-1}\right] , \tag{6.4}$$

and for all $0 \le \ell \le k$,

$$\mathbb{E}\left[(M_k - M_\ell)^2 \middle| \mathcal{F}_\ell\right] = \mathbb{E}\left[M_k^2 - M_\ell^2 \middle| \mathcal{F}_\ell\right] = \mathbb{E}\left[\langle M\rangle_k - \langle M\rangle_\ell \middle| \mathcal{F}_\ell\right] .$$

If $M_0 = 0$ $\mathbb{P}$-a.s. , then, for all $n \in \mathbb{N}^*$, $\mathbb{E}\left[M_n^2\right] = \mathbb{E}\left[\langle M\rangle_n\right]$, that is, the L^2-norm of M_n coincides with the mean of $\langle M\rangle_n$.

Note however that the *predictable* quadratic variation process $\langle M\rangle$ of a martingale M should not be confused with the *quadratic variation* or *square bracket* process $[M]$ defined by $[M]_0 = 0$ and, for all $n \in \mathbb{N}^*$,

$$\Delta[M]_n = (M_n - M_{n-1})^2 .$$

Although we obviously have, for an L^2 martingale M,

$$\mathbb{E}\left[[M]_n\right] = \mathbb{E}\left[\langle M\rangle_n\right] = \mathbb{E}\left[M_n^2\right] - \mathbb{E}\left[M_0^2\right] ,$$

these two processes do not coincide in general since $[M]$ may not be predictable.

6.4 Maximal inequalities

Proposition 6.4.1. *Let* $X = ((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ *be an* L^1 *adapted process. Denote, for all* $n \in \mathbb{N}$*,*

$$X_n^* = \max\left(X_0, \ldots, X_n\right) .$$

The following assertions hold.

(i) If X is a submartingale, then, for all $a \geq 0$ and all $n \in \mathbb{N}$, we have

$$a\mathbb{P}(X_n^* \geq a) \leq \mathbb{E}\left[X_n \mathbb{1}_{\{X_n^* \geq a\}}\right] .$$

(ii) If X is a non-negative submartingale, then, for all $a \geq 0$ and all $n \in \mathbb{N}$,

$$a\mathbb{P}\left(X_n^* \geq a\right) \leq \mathbb{E}\left[X_n\right] .$$

(iii) If X is a non-negative supermartingale, then, for all $a \geq 0$,

$$a\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n \geq a\right) \leq \mathbb{E}\left[X_0\right] .$$

(iv) (Maximal Doob inequality) If $|X| = ((|X_n|, \mathcal{F}_n)_n)_{n\in\mathbb{N}}$ is a submartingale, then, for all $n \in \mathbb{N}$ and $p > 1$,

$$\|\max(|X_0|, \ldots, |X_n|)\|_p \leq \frac{p}{p-1} \|X_n\|_p$$

Proof. Let $a \geq 0$. We introduce the stopping time

$$\tau = \inf\left\{k \in \mathbb{N} \,:\, X_k \geq a\right\} .$$

Hence, for all $n \in \mathbb{N}$, we have

$$\{X_n^* \geq a\} = \{\tau \leq n\} ,$$

and, on the event $\{\tau < \infty\}$, $a \leq X_\tau$. Hence, we get, for all $n \in \mathbb{N}$,

$$a\mathbb{1}_{\{X_n^* \geq a\}} = a\mathbb{1}_{\{\tau \leq n\}} \leq X_\tau \mathbb{1}_{\{\tau \leq n\}} = X_{\tau\wedge n}\mathbb{1}_{\{\tau \leq n\}} . \tag{6.5}$$

Suppose now that X is a submartingale. Then, for all $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[X_{\tau\wedge n}\mathbb{1}_{\{\tau\leq n\}}\right] &\leq \mathbb{E}\left[\mathbb{E}\left[X_n\middle|\,\mathcal{F}_{\tau\wedge n}\right]\mathbb{1}_{\{\tau\leq n\}}\right] \quad \text{(using the optional stoping theorem)} \\
&= \mathbb{E}\left[\mathbb{E}\left[X_n\mathbb{1}_{\{\tau\leq n\}}\middle|\,\mathcal{F}_{\tau\wedge n}\right]\right] \\
&= \mathbb{E}\left[X_n\mathbb{1}_{\{\tau\leq n\}}\right] = \mathbb{E}\left[X_n\mathbb{1}_{\{X_n^*\geq a\}}\right] .
\end{aligned}$$

This, with (6.5), proves (i). If X is moreover non-negative then $\mathbb{E}\left[X_n\mathbb{1}_{\{X_n^*\geq a\}}\right] \leq \mathbb{E}\left[X_n\right]$ and (ii) follows.

Let us now assume that X is a non-negative supermatingale. In this case, we get directly from (6.5) that, for all $n \in \mathbb{N}$,

$$a\mathbb{P}(X_n^* \geq a) \leq \mathbb{E}\left[X_{\tau\wedge n}\right] \leq \mathbb{E}\left[X_{\tau\wedge 0}\right] = \mathbb{E}\left[X_0\right] .$$

To get (iii), we observe that for all $a > 0$ (the case $a = 0$ is trivial),

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n \geq a\right) = \lim_{b\uparrow a} \mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n > b\right) ,$$

and, for all $b > 0$,

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n > b\right) = \lim_{n\to\infty} \mathbb{P}\left(X_n^* > b\right) \leq \frac{1}{b}\mathbb{E}\left[X_0\right] .$$

(using the next-to-last displayed equation). This concludes the proof of (iii).

We conclude with the proof of (iv). We denote $|X_n|^* = \max(|X_0|, \dots, |X_n|)$. Let $n \in \mathbb{N}$ and $p > 1$. Then we have

$$\||X_n|^*\|_p = p \int_0^\infty a^{p-1} \mathbb{P}\left(|X_n^*| > a\right) \, \mathrm{d}a$$

Since $|X|$ is a submartingale, we can apply (i) and get that

$$\begin{aligned} \||X_n|^*\|_p &\leq p \int_0^\infty a^{p-2} \mathbb{E}\left[|X_n| \mathbb{1}_{\{|X_n|^* \geq a\}}\right] \, \mathrm{d}a \\ &= p\mathbb{E}\left[|X_n| \int_0^\infty a^{p-2} \mathbb{1}_{\{|X_n|^* \geq a\}} \, \mathrm{d}a\right] \\ &= \frac{p}{p-1} \mathbb{E}\left[|X_n| |X_n|^{*p-1}\right] \, . \end{aligned}$$

Note that $1/p + (p-1)/p = 1$, hence the Hölder inequality gives that

$$\mathbb{E}\left[|X_n| |X_n|^{*p-1}\right] \leq \||X_n|\|_p \left\||X_n|^{*p-1}\right\|_{p/p-1} = \|X_n\|_p \, \||X_n|^*\|_p^{p-1} \, .$$

Plugging this inequality in the previous display, we get that

$$\||X_n|^*\|_p \leq \frac{p}{p-1} \|X_n\|_p \, ,$$

which achieves the proof. □

In particular, if $X = ((X_n, \mathcal{F}_n))_{n \in \mathbb{N}}$ is a martingale then by 6.1.3, $|X|$ is submartingale and the Doob maximal inequality (iv) of Proposition 6.4.1 applies. For $p = 2$ it gives that, for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\max_{0 \leq k \leq n} X_k^2\right] \leq 4\mathbb{E}\left[X_n^2\right] \, . \tag{6.6}$$

6.5 Asymptotic behavior

6.5.1 Martingales bounded in L^2

We say that, for a given $p \geq 1$, a process $(X_n)_{n \in \mathbb{N}}$ is bounded in L^p if

$$\sup_{n \in \mathbb{N}} \|X_n\|_p < \infty \, .$$

We have the following result.

Theorem 6.5.1. *Let $M = ((M_n, \mathcal{F}_n))_{n \in \mathbb{N}}$ be a martingale bounded in L^2. Then there exists $M_\infty \in L^2$ such that, as $n \to \infty$, M_n converges in L^2 and $\mathbb{P}$-a.s. to M_∞.*

We first state a simple result, whose proof is left as an exercise.

Lemma 6.5.2. *Let $M = ((M_n, \mathcal{F}_n))_{n \in \mathbb{N}}$ be an L^2 martingale. Then ΔM is an orthogonal sequence in L^2.*

Proof of Theorem 6.5.1. By Lemma 6.5.2, we have for all $0 \leq n \leq p$,

$$\|M_p - M_n\|_2^2 = \left\| \sum_{k=n+1}^{p} \Delta M_k \right\|_2^2 = \sum_{k=n+1}^{p} \|\Delta M_k\|_2^2 \ .$$

Moreover, we have

$$\sum_{k \in \mathbb{N}^*} \|\Delta M_k\|_2^2 = \lim_{p \to \infty} \|M_p - M_0\|_2^2 \leq 2 \sup_{p \in \mathbb{N}} \|M_p\|_2^2 < \infty \ .$$

The last two displayed equations imply that M is a Cauchy sequence in L^2. Hence there exists $M_\infty \in L^2$ such that, as $n \to \infty$, M_n converges in L^2 to M_∞.

To conclude the proof it only remains to show that M is also a Cauchy sequence $\mathbb{P}$-a.s.This is true if, for all $\epsilon > 0$ we have

$$\mathbb{P}\left(\bigcap_{n \in \mathbb{N}} \left\{ \sup_{q \geq p \geq n} |M_q - M_p| > \epsilon \right\} \right) = \lim_{n \to \infty} \mathbb{P}\left(\sup_{q \geq p \geq n} |M_q - M_p| > \epsilon \right) = 0 \ .$$

Observe that, for all $n \in \mathbb{N}$,

$$\sup_{q \geq p \geq n} |M_q - M_p| \leq 2 \sup_{p \geq n} |M_p - M_n| \ .$$

Hence we finally have to show that, for all $\epsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{p \geq n} |M_p - M_n| > \epsilon \right) = 0 \ .$$

Observe that, for any $n \in \mathbb{N}$, $((|M_{n+k} - M_n|, \mathcal{F}_{n+k}))_{k \in \mathbb{N}}$ is a non-negative submartingale. By Proposition 6.4.1 (ii), we have that, for any $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{P}\left(\sup_{p \geq n} |M_p - M_n| > \epsilon \right) &= \lim_{k \to \infty} \mathbb{P}\left(\sup_{n \leq p \leq n+k} |M_p - M_n| > \epsilon \right) \\ &\leq \frac{1}{\epsilon} \limsup_{k \to \infty} \mathbb{E}\left[|M_{n+k} - M_n| \right] \\ &\leq \frac{1}{\epsilon} \|M_\infty - M_n\|_2 \ , \end{aligned}$$

which converges to 0 as $n \to \infty$ and thus the proof is concluded. □

6.5.2 Submartingales bounded in L^1

Given a process $X = (X_n)_{n \in \mathbb{N}}$, for all $n \in \mathbb{N}$ and $a < b \in \mathbb{R}$, we denote by $\mathrm{U}_n(a, b)$ the *number of upcrossings* from a to b up to time n, that is,

$$\mathrm{U}_n^X(a, b) = \#\left\{ 0 < k \leq n \ : \ \exists i \in \{0, \dots, k-1\} \text{ s.t. } X_i \leq a \text{ and } \tau_{b,i} = k \right\} \ ,$$

where

$$\tau_{b,i} = \inf\left\{ j \geq i \ : \ X_j \geq b \right\}$$

is the first time after i when the process gets over b.

It turns out that the mean number of upcrossings can be adequately bounded for submartingales.

Proposition 6.5.3. *Suppose that $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale. Then, for all $n \in \mathbb{N}$ and $a < b \in \mathbb{R}$, we have*

$$(b-a)\mathbb{E}\left[\mathrm{U}_n^X(a,b)\right] \leq \mathbb{E}\left[(X_n - a)_+\right] - \mathbb{E}\left[(X_0 - a)_+\right] \ . \tag{6.7}$$

Proof. We introduce some definitions for given $a < b \in \mathbb{R}$. First we introduce recursive stopping times of successive crossings of levels a (downward) and b (upward). Set $\sigma_0^+ = -1$. For $k \in \mathbb{N}^*$, define

$$\sigma_k^- = \inf\left\{m > \sigma_{k-1}^+ \ : \ X_m \leq a\right\} \quad \text{and} \quad \sigma_k^+ = \inf\left\{m > \sigma_k^- \ : \ X_m \geq b\right\} \ ,$$

which all are $(\mathcal{F}_n)_{n\in\mathbb{N}}$-stopping times. Following this definition, we have, for all $n \in \mathbb{N}$,

$$\mathrm{U}_n^X(a,b) = \#\left\{k \in \mathbb{N}^* \ : \ \sigma_k^+ \leq n\right\} \ . \tag{6.8}$$

Moreover, since, for all $k \in \mathbb{N}^*$ and $m \in \mathbb{N}$,

$$\{\sigma_k^- < m \leq \sigma_k^+\} = \{\sigma_k^- \leq m-1\} \cap \{\sigma_k^+ \leq m-1\}^c \in \mathcal{F}_{m-1} \ ,$$

the process $H = (H_n)_{n\in\mathbb{N}^*}$ defined by

$$H = \sum_{k=1}^{\infty} \mathbb{1}_{\{\sigma_k^- + 1, \ldots, \sigma_k^+\}}$$

is a $(\mathcal{F}_n)_{n\in\mathbb{N}}$-predictable process (valued in $\{0,1\}$).

Define, for all $n \in \mathbb{N}$,

$$Y_n = a \vee X_n = a + (X_n - a)_+ \ .$$

Proposition 6.1.2 implies that $((Y_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale. Observe that, for all $n \in \mathbb{N}$,

$$(H \cdot Y)_n = \sum_{k=1}^{\infty} \left(\mathbb{1}_{\{\sigma_k^- + 1, \ldots, \sigma_k^+\}} \cdot Y\right)_n = \sum_{k=1}^{\infty} \mathbb{1}_{\{\sigma_k^- < n\}} \left(Y_{\sigma_k^+ \wedge n} - Y_{\sigma_k^-}\right) \ .$$

Note that for all $n \in \mathbb{N}$ and $k \in \mathbb{N}^*$ such that $\sigma_k^- < n$,

$$Y_{\sigma_k^+ \wedge n} - Y_{\sigma_k^-} = Y_{\sigma_k^+ \wedge n} - a \geq 0 \ ,$$

and, if moreover $\sigma_k^+ \leq n$, then

$$Y_{\sigma_k^+ \wedge n} = Y_{\sigma_k^+} \geq b \ .$$

Hence we get that, for all $n \in \mathbb{N}$,

$$(H \cdot Y)_n \geq \sum_{k=1}^{\infty} \mathbb{1}_{\{\sigma_k^+ \leq n\}} (b-a) = (b-a)\mathrm{U}_n^X(a,b) \ , \tag{6.9}$$

where we have used (6.8).

Now observe that H and $1 - H$ are two predictable processes valued in $\{0,1\}$ and that, for all $n \in \mathbb{N}$,

$$Y_n - Y_0 = (H \cdot Y)_n + ((1-H) \cdot Y)_n \ .$$

By Proposition 6.2.1, this two terms define submartingales starting at zero, hence have non-negative expectations. In particular, we get that, for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[(H \cdot Y)_n\right] \leq \mathbb{E}\left[Y_n - Y_0\right] = \mathbb{E}\left[(X_n - a)_+\right] - \mathbb{E}\left[(X_0 - a)_+\right] \ .$$

With (6.9), we conclude that (6.7) holds. □

Proposition 6.5.3 is the essential step for proving the following fundamental result.

Theorem 6.5.4 (Martingale convergence theorem)**.** *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be a submartingale. Then it is bounded in L^1 if and only if it satisfies*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[(X_n)_+\right] < \infty \,. \tag{6.10}$$

Moreover, in this case, there exists an L^1 random variable X_∞ such that $X_n \xrightarrow{\mathbb{P}a.s.} X_\infty$.

Proof. We first observe that if $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1 then it satisfies (6.10).

Next, we note that, for all $n \in \mathbb{N}$, $\mathbb{E}\left[X_n\right] \geq \mathbb{E}\left[X_0\right]$ and thus

$$\mathbb{E}\left[|X_n|\right] = 2\mathbb{E}\left[(X_n)+\right] - \mathbb{E}\left[X_n\right] \leq 2\mathbb{E}\left[(X_n)+\right] - \mathbb{E}\left[X_0\right] \,.$$

Hence (6.10) implies that $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1.

We now prove the $\mathbb{P}$-a.s.convergence to an L^1 r.v. Observe that for any sequence $x = (x_n)_{n\in\mathbb{N}}$,

$$\liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n \Rightarrow \exists a < b \in \mathbb{Q} \,,\ \mathrm{U}^x_\infty(a,b) = \infty \,. \tag{6.11}$$

Applying Proposition 6.5.3, we get that, for all $n \in \mathbb{N}$ and all $a < b$, we have

$$(b-a)\mathbb{E}\left[\mathrm{U}^X_n(a,b)\right] \leq \mathbb{E}\left[(X_n - a)_+\right] \leq \sup_{n\in\mathbb{N}} \mathbb{E}\left[(X_n)_+\right] + |a| \,.$$

Hence for all $a < b \in \mathbb{R}$, $\mathbb{E}\left[\mathrm{U}^X_\infty(a,b)\right] < \infty$ which implies that $\mathrm{U}^X_\infty(a,b) < \infty$ $\mathbb{P}$-a.s. and, by (6.11), we get that

$$\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n \qquad \mathbb{P}\text{-a.s.} \,,$$

that is, X_n converges $\mathbb{P}$-a.s., but possibly to $-\infty$ or ∞. Let us denote X_∞ the limit (valued in $\bar{\mathbb{R}}$). Now, Fatou's lemma implies that

$$\mathbb{E}\left[|X_\infty|\right] = \mathbb{E}\left[\liminf_{n\to\infty} |X_n|\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[|X_n|\right] \,,$$

which is finite since $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1. Hence X_∞ is L^1, which concludes the proof. □

Corollary 6.5.5. *If $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a submartingale, a supermartinagle, or a martingale bounded in L^1 then there exists an L^1 random variable X_∞ such that $X_n \xrightarrow{\mathbb{P}a.s.} X_\infty$.*

Proof. The assumption of this corollary implies that X or $-X$ satisfies the assumption of Theorem 6.5.4. □

6.5.3 Closed martingales

Recall the closed martingales introduced in Example 6.1.2. The goal of this section is to characterize the martingales that can be defined as closed martingales.

To this end we will need the following definition.

Definition 6.5.1 (Uniform integrability)**.** *A process $(X_t)_{t\in T}$ is said to be* uniformly integrable *(U.I.) if*

$$\lim_{m\to\infty} \sup_{t\in T} \mathbb{E}\left[|X_t| \mathbb{1}_{\{|X_t|\geq m\}}\right] = 0 \,.$$

It is standard to give this definition for a process, although this definition, as the tightness, can be formulated relying only on the distributions of the X_t's (i.e.one does not even need the r.v.'s X_t to be defined on the same probability space), namely,

$$\lim_{m\to\infty} \sup_{t\in T} \int |x| \mathbb{1}_{\{|x|\geq m\}} \, \mathbb{P}^{X_t}(\mathrm{d}x) = 0 \, .$$

Remark 6.5.1. *Observe that for all $m > 0$ and r.v. X_t, we can write*

$$\mathbb{E}\left[|X_t|\right] \leq m + \mathbb{E}\left[|X_t| \mathbb{1}_{\{|X_t|\geq m\}}\right] \, .$$

So if $(X_t)_{t\in T}$ is U.I., it is bounded in L^1. The converse implication is of course not true. Take, as a counterexample, for all $n \in \mathbb{N}$, $X_n = 2^n U_n$ with $U_n \sim \mathbf{Ber}(2^{-n})$.

Before stating the characterization of closed martingales using uniform integrability, we give some general results related to U.I. processes.

Proposition 6.5.6. *Suppose that $(X_n)_{n\in\mathbb{N}}$ is U.I. Then, we have*

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[X_n\right] \leq \limsup_{n\to\infty} \mathbb{E}\left[X_n\right] \leq \mathbb{E}\left[\limsup_{n\to\infty} X_n\right] \, .$$

Proof. It is sufficient to prove the first inequality (the second is obvious and the third is obtained by replacing X_n by $-X_n$).

For all $m > 0$, we have

$$\liminf_{n\to\infty} X_n \leq \liminf_{n\to\infty} (X_n \vee (-m)) = -m + \liminf_{n\to\infty} \left[m + (X_n \vee (-m))\right] \, .$$

By Fatou's lemma, we have, for all $m > 0$,

$$\mathbb{E}\left[\liminf_{n\to\infty} \left[m + (X_n \vee (-m))\right]\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[\left[m + (X_n \vee (-m))\right]\right] \, .$$

The last two displays thus yield, for all $m > 0$,

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[(X_n \vee (-m))\right] \, . \tag{6.12}$$

Now, since $(X_n)_{n\in\mathbb{N}}$ is U.I. and

$$|X_n - (X_n \vee (-m))| = |X_n + m| \, \mathbb{1}_{\{X_n < -m\}} \leq |X_n| \mathbb{1}_{\{|X_n|>m\}} \, ,$$

we have

$$\lim_{m\to\infty} \sup_{n\in\mathbb{N}} |\mathbb{E}\left[X_n\right] - \mathbb{E}\left[(X_n \vee (-m))\right]| = 0 \, .$$

It follows that $\liminf_{n\to\infty} \mathbb{E}\left[(X_n \vee (-m))\right]$ can be made arbitrarily close to $\liminf_{n\to\infty} \mathbb{E}\left[X_n\right]$ by taking m large enough. With (6.12), we conclude that

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[X_n\right] \, .$$

The proof is concluded. □

Theorem 6.5.7. *Suppose that $(X_n)_{n\in\mathbb{N}}$ is U.I. and $X_n \xrightarrow{a.s.} X_\infty$, then X_∞ is L^1 and X_n converges to X_∞ in L^1.*

Proof. The sequence of the positive parts of the X_n's is U.I. and converges to the positive part of X_∞ $\mathbb{P}$-a.s.. The same can be said of the negative parts. Hence we can assume in the following that $(X_n)_{n\in\mathbb{N}}$ is non-negative without loss of generality.

Now, by Proposition 6.5.6, we get

$$\mathbb{E}[X_\infty] = \mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \le \liminf_{n\to\infty} \mathbb{E}[X_n] \ ,$$

which is finite, since $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1. Hence X_∞ is L^1. It follows (see Exercise 6.14) that $(|X_n - X_\infty|)_{n\in\mathbb{N}}$ is U.I. and converges to 0 a.s. Hence using Proposition 6.5.6 again,

$$\limsup_{n\to\infty} \mathbb{E}[|X_n - X_\infty|] \le \mathbb{E}\left[\limsup_{n\to\infty} |X_n - X_\infty|\right] = 0 \ ,$$

that is, X_n converges to X_∞ in L^1. □

Lemma 6.5.8. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then the process $(\mathbb{E}[X|\mathcal{G}])_{\mathcal{G}\subset\mathcal{F}}$ (indexed by all sub-σ-field of $\mathcal{F}$) is U.I.*

The proof of this lemma requires the following result.

Lemma 6.5.9. *Let $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\lim_{\delta\to 0} \sup\{\mathbb{E}[|Y|\mathbb{1}_A] \ : \ A \in \mathcal{F}, \ \mathbb{P}(A) \le \delta\} = 0 \ .$$

Proof. Since $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, the mapping $\nu : A \mapsto \mathbb{E}[|Y|\mathbb{1}_A]$ defines a finite measure on $(\Omega, \mathcal{F})$. We prove this lemma by contradiction. Suppose that there exists $\epsilon > 0$ and a sequence $(A_n)_{n\in\mathbb{N}} \in \mathcal{F}^{\mathbb{N}}$ such that, for all $n \in \mathbb{N}$, $\mathbb{P}(A_n) \le 2^{-n}$ and $\mathbb{E}[|Y|\mathbb{1}_{A_n}] \ge \epsilon$. Define, for all $n \in \mathbb{N}$,

$$B_n = \bigcup_{k\ge n} A_k \ .$$

Then, for all $n \in \mathbb{N}$, $\mathbb{P}(B_n) \le 2^{1-n}$ and $\mathbb{E}[|Y|\mathbb{1}_{B_n}] \ge \epsilon$. However, by dominated convergence,

$$\lim_{n\to\infty} \mathbb{E}[|Y|\mathbb{1}_{B_n}] = \mathbb{E}[|Y|\mathbb{1}_{\cap_n B_n}] = 0 \ ,$$

since $\mathbb{P}(\cap_n B_n) = \lim_{n\to\infty} \mathbb{P}(B_n) = 0$. We obtain a contradiction and the proof is concluded. □

We can now prove Lemma 6.5.8.

Proof of Lemma 6.5.8. Let $m > 0$ and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. We have that

$$\begin{aligned}\mathbb{E}\left[|\mathbb{E}[X|\mathcal{G}]|\mathbb{1}_{\{|\mathbb{E}[X|\mathcal{G}]|\ge m\}}\right] &\le \mathbb{E}\left[\mathbb{E}[|X||\mathcal{G}]\,\mathbb{1}_{\{|\mathbb{E}[X|\mathcal{G}]|\ge m\}}\right] \\ &= \mathbb{E}\left[|X|\mathbb{1}_{\{|\mathbb{E}[X|\mathcal{G}]|\ge m\}}\right] \ .\end{aligned}$$

Now observe that, by the Markov inequality

$$\mathbb{P}(|\mathbb{E}[X|\mathcal{G}]| \ge m) \le m^{-1}\mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|] \le m^{-1}\mathbb{E}[|X|] \ .$$

We get that, for all $m > 0$,

$$\sup\left\{\mathbb{E}\left[|\mathbb{E}[X|\mathcal{G}]|\mathbb{1}_{\{|\mathbb{E}[X|\mathcal{G}]|\geq m\}}\right] : \mathcal{G} \subset \mathcal{F}\right\} \leq \sup\left\{\mathbb{E}[|X|\mathbb{1}_A] : A \in \mathcal{F}, \mathbb{P}(A) \leq m^{-1}\|X\|_1\right\} .$$

Applying Lemma 6.5.9, we obtain that this goes to 0 as $m \to \infty$, which concludes the proof. □

We can now state the result characterizing closed martingales.

Theorem 6.5.10. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be a martingale. The three following properties are equivalent*

(i) The sequence $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.

(ii) The sequence $(X_n)_{n\in\mathbb{N}}$ converges in L^1.

(iii) The sequence $(X_n)_{n\in\mathbb{N}}$ is a closed martingale.

If these equivalent assumptions hold, we say that $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is a regular martingale. *Moreover, the following assertions follows.*

(a) The L^1 limit X_∞ in (ii) is also a $\mathbb{P}$-a.s. limit and we have, for all $n \in \mathbb{N}$,

$$X_n = \mathbb{E}[X_\infty|\mathcal{F}_n] \qquad \mathbb{P}\text{-a.s.}$$

(b) An L^1 random variable X satisfies $X_n = \mathbb{E}[X|\mathcal{F}_n]$ for all $n \in \mathbb{N}$ if and only if

$$X_\infty = \mathbb{E}\left[X|\bigvee_{n\in\mathbb{N}}\mathcal{F}_n\right] \qquad \mathbb{P}\text{-a.s.} \tag{6.13}$$

Proof. First Lemma 6.5.8 directly shows that (iii)⇒(i).

If (i) holds then $(X_n)_{n\in\mathbb{N}}$ is bounded in L^1 and Corollary 6.5.5 shows that it converges $\mathbb{P}$-a.s. By Theorem 6.5.7, we then obtain (ii).

Suppose now that (ii) holds and let X_∞ denote the L_1 limit. Now since for all $k \in \mathbb{N}$ and all $n \geq k$,

$$\mathbb{E}[X_n|\mathcal{F}_k] = X_k$$

and X_n converges to X_∞ in L^1 we get that, for all $k \in \mathbb{N}$,

$$\mathbb{E}[X_\infty|\mathcal{F}_k] = X_k .$$

Hence (iii) holds.

We have proved the equivalence of (i), (ii) and (iii), and also that Assertion (a) holds as a byproduct.

It only remains to prove (b). If (6.13) holds then, for all $n \in \mathbb{N}$, we have

$$\mathbb{E}[X|\mathcal{F}_n] = \mathbb{E}\left[\mathbb{E}\left[X|\bigvee_{k\in\mathbb{N}}\mathcal{F}_k\right]\middle|\mathcal{F}_n\right] = \mathbb{E}[X_\infty|\mathcal{F}_n] = X_n ,$$

where we used (a). Suppose now that $\mathbb{E}[X|\mathcal{F}_n] = X_n$ for all $n \in \mathbb{N}$ and let us prove that (6.13) holds. Let $n \in \mathbb{N}$ and $A \in \mathcal{F}_n$. We have, for, all $p \geq n$,

$$\mathbb{E}[1_A X] = \mathbb{E}[\mathbb{E}[\mathbb{1}_A X|\mathcal{F}_p]] = \mathbb{E}[\mathbb{1}_A\mathbb{E}[X|\mathcal{F}_p]] = \mathbb{E}[1_A X_p] .$$

Hence, letting $p \to \infty$, $\mathbb{E}[1_A X] = \mathbb{E}[1_A X_\infty]$. The π-λ theorem (Theorem 2.0.1) then implies that $\mathbb{E}[\mathbb{1}_A X] = \mathbb{E}[\mathbb{1}_A X_\infty]$ for all $A \in \mathcal{F}_\infty := \bigvee_{n\in\mathbb{N}} \mathcal{F}_n$, showing that $X_\infty = \mathbb{E}[X|\,\mathcal{F}_\infty]$, $\mathbb{P}$-a.s. □

A nice application of Theorem 6.5.10 is the 0-1 law, see Exercise 6.22.

Inspecting the proof of (i)⇒(ii)⇒(iii), we see that it is still working out if $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ is only a submartingale, resulting in the following statement.

Theorem 6.5.11. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be a U.I. submartingale. Then $(X_n)_{n\in\mathbb{N}}$ converges in L^1 and* $\mathbb{P}$-a.s.*to an L^1 random variables X_∞. In addition, for all $n \in \mathbb{N}$, $X_n \le \mathbb{E}[X_\infty|\,\mathcal{F}_n]$, $\mathbb{P}$-a.s.*

6.6 Application to convergence theorems

Convergence theorems for martingales allows us to derive a reasonably simple proof of the law of large numbers.

Theorem 6.6.1 (Law of large numbers)**.** *Let $(X_n)_{n\in n\in\mathbb{N}^*}$ be a sequence of i.i.d. L^1 random variables. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mathbb{E}[X_1] \qquad \mathbb{P}\text{-a.s.}$$

Before proving this result, we give a useful lemma for deterministic sequences.

Lemma 6.6.2 (Kronecker's lemma)**.** *Let $(b_n)_{n\in\mathbb{N}}$ be a non-decreasing diverging sequence such that $b_0 = 0$. Define $a = \Delta b$, the (non-negative) sequence of its increments. Let $(x_n)_{n\in\mathbb{N}^*}$ be a sequence valued on $\mathbb{R}$ and set, for all $n \ge 1$,*

$$s_n = \sum_{k=1}^{n} x_k \ .$$

(i) If $(x_n)_{n\in\mathbb{N}^*}$ converges to $x_\infty \in \mathbb{R}$, then

$$\lim_{n\to\infty} b_n^{-1} \sum_{k=1}^{n} a_k x_k = x_\infty \ .$$

(ii) If $(s_n)_{n\in\mathbb{N}^*}$ converges in $\mathbb{R}$, then

$$\lim_{n\to\infty} b_n^{-1} \sum_{k=1}^{n} b_k x_k = 0 \ .$$

Proof. The proof of (i) is similar to that of the Cesaro's lemma. Since, for all $n \in \mathbb{N}^*$, provided that $b_n > 0$,

$$b_n^{-1} \sum_{k=1}^{n} a_k = 1 \ ,$$

we have

$$\left| b_n^{-1} \sum_{k=1}^{n} a_k x_k - x_\infty \right| = \left| b_n^{-1} \sum_{k=1}^{n} a_k (x_k - x_\infty) \right|$$

Take $1 < m < n$ with $b_n > 0$. We get that

$$\left| b_n^{-1} \sum_{k=1}^{n} a_k x_k - x_\infty \right| \leq b_n^{-1} \left| \sum_{k=1}^{m} a_k (x_k - x_\infty) \right| + b_n^{-1}(b_n - b_m) \sup_{m<k\leq n} |x_k - x_\infty| \, .$$

Hence, for any $m > 1$, we have

$$\limsup_{n\to\infty} \left| b_n^{-1} \sum_{k=1}^{n} a_k x_k - x_\infty \right| \leq \sup_{k>m} |x_k - x_\infty| \, .$$

Assertion (i) follows.

For all $n \in \mathbb{N}^*$ with $b_n > 0$, we can write, setting $s_0 = 0$,

$$\begin{aligned} b_n^{-1} \sum_{k=1}^{n} b_k x_k &= b_n^{-1} \sum_{k=1}^{n} b_k (s_k - s_{k-1}) \\ &= s_n - b_n^{-1} \sum_{k=0}^{n-1} a_{k+1} s_k \, . \end{aligned}$$

If $(s_n)_{n\in\mathbb{N}^*}$ converges to $s_\infty \in \mathbb{R}$, we get from (i) that, for all $m \geq 1$,

$$\lim_{n\to\infty} b_n^{-1} \sum_{k=0}^{n-1} a_{k+1} s_k = s_\infty \, .$$

Hence, with the previous display, we obtain (ii). □

Proof of Theorem 6.6.1. We first observe that, by replacing X_k by $X_k - \mathbb{E}[X_1]$, we can assume in the following that $\mathbb{E}[X_1] = 0$ without loss of generality.

Now the proof has several successive steps.

Step 1 Let us show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k \mathbb{1}_{\{|X_k|>k\}} = 0 \qquad \mathbb{P}\text{-a.s.} \tag{6.14}$$

Clearly, this is implied by

$$\#\{k \in \mathbb{N}^* \, : \, |X_k| > k\} < \infty \qquad \mathbb{P}\text{-a.s.} \, ,$$

which in turn is equivalent to

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}^*} \bigcup_{k\geq n} \{|X_k| > k\} \right) = 0 \, .$$

Now this probability is zero since, for all $n \in \mathbb{N}^*$,

$$\mathbb{P}\left(\bigcup_{k\geq n} \{|X_k| > k\} \right) \leq \sum_{k\geq n} \mathbb{P}(|X_k| > k) = \mathbb{E}\left[\sum_{k\geq n} \mathbb{1}_{\{|X_1|>k\}} \right] \leq \mathbb{E}\left[(|X_1| - n)_+ \right] \, ,$$

which, by dominated convergence, has limit 0 as $n \to \infty$. Hence we have shown (6.14).

Step 2 Let us show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[X_k \mathbb{1}_{\{|X_k|\le k\}}\right] = 0 \qquad \mathbb{P}\text{-a.s.} \tag{6.15}$$

For all $n \in \mathbb{N}^*$, we have

$$\begin{aligned}\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[X_k \mathbb{1}_{\{|X_k|\le k\}}\right] &= \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[X_1 \mathbb{1}_{\{|X_1|\le k\}}\right] \\ &= \mathbb{E}\left[X_1 \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{|X_1|\le k\}}\right] ,\end{aligned}$$

which tends to $\mathbb{E}[X_1] = 0$ as $n \to \infty$ by dominated convergence. Hence (6.15) holds.

Step 3 From **Step 1** and **Step 2**, it only remains to show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left(X_k \mathbb{1}_{\{|X_k|\le k\}} - \mathbb{E}\left[X_k \mathbb{1}_{\{|X_k|\le k\}}\right]\right) = 0 \qquad \mathbb{P}\text{-a.s.} \tag{6.16}$$

Define $M_0 = 0$ and, for all $n \in \mathbb{N}^*$,

$$M_n = \sum_{k=1}^{n} k^{-1} \left(X_k \mathbb{1}_{\{|X_k|\le k\}} - \mathbb{E}\left[X_k \mathbb{1}_{\{|X_k|\le k\}}\right]\right) .$$

We observe that M_n is an L^2 martingale and, for all $n \in \mathbb{N}^*$,

$$\begin{aligned}\mathbb{E}\left[M_n^2\right] &= \sum_{k=1}^{n} k^{-2} \operatorname{Var}\left(X_k \mathbb{1}_{\{|X_k|\le k\}}\right) \\ &\le \sum_{k=1}^{n} k^{-2} \mathbb{E}\left[X_1^2 \mathbb{1}_{\{|X_1|\le k\}}\right] \\ &= \mathbb{E}\left[X_1^2 \sum_{k=1}^{n} k^{-2} \mathbb{1}_{\{|X_1|\le k\}}\right] \\ &\le \mathbb{E}\left[X_1^2 \left(1 + \sum_{k=2}^{n} (k(k-1))^{-1} \mathbb{1}_{\{|X_1|\le k\}}\right)\right] \\ &\le \mathbb{E}\left[X_1^2 \left(1 + \{(|X_1| \vee 2) - 1\}^{-1}\right)\right]\end{aligned}$$

Since X_1 is L^1 this upper bound is finite, which shows that M is bounded in L^2. Hence, by Theorem 6.5.1, it converge $\mathbb{P}$-a.s. By Lemma 6.6.2 (with $b_k = k$ and $x = \Delta M$), (6.16) follows and the proof is concluded.

□

We now derive an a.s. convergence theorem for L^2 martingales.

Theorem 6.6.3. *Let $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an L^2 martingale with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $M_0 = 0$, $\mathbb{P}$-a.s. Then $(M_n)_{n\in\mathbb{N}}$ converges in $\mathbb{R}$, $\mathbb{P}$-a.s. on the event*

$$A = \left\{ \sum_{n=0}^{\infty} \mathbb{E}\left[(\Delta M_{n+1})^2 \middle| \mathcal{F}_n \right] < \infty \right\} .$$

It is easy to see that if M is bounded in L^2, then A has probability 1; hence we recover the fact that it converges $\mathbb{P}$-a.s. in this case, as already established in Theorem 6.5.1. In other words the condition that A has probability 1 is a weaker assumption than that used in Theorem 6.5.1 but we note that it does not provide the L^2 convergence.

Proof of Theorem 6.6.3. Let $K > 0$. Define the stopping time

$$\tau_K = \inf \left\{ k \geq 1 , \sum_{n=0}^{k} \mathbb{E}\left[(\Delta M_{n+1})^2 \middle| \mathcal{F}_n \right] > K \right\} . \tag{6.17}$$

Proposition 6.2.2 shows that M^{τ_K} is a martingale, and $M^{\tau_K} = H \cdot M$ with $((H_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ the predictable process defined by $H_n = \mathbb{1}_{\{n\leq\tau_K\}}$ for all $n \in \mathbb{N}^*$. Then for all $n \in \mathbb{N}$,

$$\|M_n^{\tau_K}\|_2^2 = \mathbb{E}\left[\langle M^{\tau_K}\rangle_n\right] ,$$

and

$$\langle M^{\tau_K}\rangle_n = \sum_{k=1}^{n} H_k \mathbb{E}\left[\Delta M_k \middle| \mathcal{F}_{k-1}\right] = \sum_{k=1}^{n\wedge\tau_K} \mathbb{E}\left[\Delta M_k \middle| \mathcal{F}_{k-1}\right] \leq K .$$

We conclude that M^{τ_K} is bounded in L^2 and, by Theorem 6.5.1, $(M_n^{\tau_K})_{n\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. and in L^2. Therefore, $(M_n)_{n\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. on the event $\{\tau_K = \infty\}$, and thus also on the event

$$\bigcup_{K\in\mathbb{N}^*} \{\tau_K = \infty\}$$

which coincides with A. □

We then get the following result, for any sequence bounded in L^2.

Corollary 6.6.4. *Let $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ be an adapted process bounded in L^2. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}\right]) = 0 \quad \mathbb{P}\text{-a.s.}$$

Proof. Define the martingale $((M_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ by $M_0 = 0$ and, for all $n \in \mathbb{N}*$

$$M_n = \sum_{k=1}^{n} k^{-1} (X_k - \mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}\right])$$

By Lemma 6.6.2 (with $b_k = k$ and $s = M$), it is sufficient to show that $(M_n)_{n\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. By Theorem 6.6.3, it is thus sufficient to show that the set A defined in this theorem has probability 1. We have

$$\sum_{n=0}^{\infty} \mathbb{E}\left[(\Delta M_{n+1})^2 \middle| \mathcal{F}_n \right] = \sum_{k=1}^{\infty} k^{-2} \mathbb{E}\left[(X_k - \mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}\right])^2 \middle| \mathcal{F}_{k-1} \right] \leq \sum_{k=1}^{\infty} k^{-2} \mathbb{E}\left[X_k^2 \middle| \mathcal{F}_{k-1} \right] .$$

But since $(M_n)_{n\in\mathbb{N}}$ is bounded in L^2, this upper bound has finite expectation, hence is finite $\mathbb{P}$-a.s., which conclude the proof. □

6.7 Exercises

Exercise 6.1. Let $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ be a supermartingale such that $\mathbb{E}[X_n]$ is constant. Show that $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ is a martingale.

Exercise 6.2. Let $(\epsilon_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with mean zero and variances given by $\mathrm{Var}[\epsilon_n] = \sigma_n^2$. Let $S_0 = T_0 = 0$ and for all $n \geq 1$,

$$S_n = \sum_{i=1}^{n} \epsilon_i \quad \text{et} \quad T_n = \sum_{i=1}^{n} \sigma_i^2 \, .$$

Show that $S_n^2 - T_n$ is a martingale.

Exercise 6.3. Soient $\{X_n\}_{n\geq 0}$ et $\{Y_n\}_{n\geq 0}$ deux martingales de carré intégrable (définies sur un même espace filtré).

1. Montrer que, pour $m \leq n$, on a $\mathbb{E}[X_m Y_n | \mathcal{F}_m] = X_m Y_m$ p.s.
2. Montrer que $\mathbb{E}[X_n Y_n] - \mathbb{E}[X_0 Y_0] = \sum_{k=1}^{n} \mathbb{E}[(X_k - X_{k-1})(Y_k - Y_{k-1})]$.

Exercise 6.4 (Inégalité d'Azuma). Soit $((X_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ une martingale. On suppose qu'il existe c_1, c_2, ... des réels positifs tels que

$$|X_k - X_{k-1}| \leq c_k.$$

On pose $Y_k = X_k - X_{k-1}$.

1. Montrer que, pour $t > 0$ et $-c \leq x \leq c$,
$$\mathrm{e}^{tx} \leq \frac{\mathrm{e}^{ct} + \mathrm{e}^{-ct}}{2} + \frac{\mathrm{e}^{ct} - \mathrm{e}^{-ct}}{2c}\, x.$$
2. On note $f_c(x)$ le terme de droite de l'inégalité précédente. Soit Z une v.a. telle que $|Z| \leq c$. Montrer que pour toute tribu $\mathcal{F}$,
$$\mathbb{E}\left[\mathrm{e}^{tZ} \middle| \mathcal{F}\right] \leq f_c(\mathbb{E}[Z|\mathcal{F}]).$$
3. Soit $t > 0$, montrer que pour tout a,
$$\mathbb{P}(Y_1 + \ldots + Y_m \geq a) \leq \mathrm{e}^{-ta}\mathbb{E}\left[\mathrm{e}^{tY_1 + \ldots + tY_m}\right].$$
4. Montrer que
$$\mathbb{E}\left[\mathrm{e}^{tY_n} \middle| \mathcal{F}_{n-1}\right] \leq f_{c_n}(0) \leq \mathrm{e}^{(c_n t)^2/2}.$$
5. Montrer que
$$\mathbb{P}(X_m - X_0 \geq a) \leq \exp\left(-at + \frac{t^2}{2}\sum_{k=1}^{m} c_k^2\right). \tag{6.18}$$
6. Trouver la valeur de t qui minimise le terme de droite de (6.18). En déduire que
$$\mathbb{P}(X_m - X_0 \geq a) \leq \exp\left(-\frac{a^2}{2\sum_{k=1}^{m} c_k^2}\right).$$

Exercise 6.5 (Jensen's inequality). Let X be a L^1 random variable valued in $]a,\ b[$ with $-\infty \leq a < b \leq +\infty$ and let $\Phi : (a,\ b) \to \mathbb{R}$ be a convex function. Assume that $\Phi(X)$ is L^1.

1. Show that
$$\Phi(\mathbb{E}[X]) \leq \mathbb{E}[\Phi(X)].$$
Hint: Use that Φ admits a left (and a right) derivative $\Phi'_<$ at any point $x \in (a,b)$, and, for all $y \in (a,b)$, $\Phi(y) \geq \Phi'_<(x)\,(y-x) + \Phi(x)$.

2. Let $\mathcal{G}$ be a sub σ-field. Show that
$$\Phi(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[\Phi(X) \mid \mathcal{G}] \quad \mathbb{P}\text{-a.s.}$$
Hint: Show first that the inequality holds on $\{|\Phi(\mathbb{E}[X \mid \mathcal{G}])| \vee |\Phi'(\mathbb{E}[X \mid \mathcal{G}])| \leq n\}$ for any given $n \geq 1$.

3. Prove Proposition 6.1.2.

Exercise 6.6. Soit $(Y_n)_{n\in\mathbb{N}}$ une suite de v.a. indépendantes et de même loi telles que $\mathbb{P}\{Y_k = 1\} = \mathbb{P}\{Y_k = -1\} = 1/2$. On pose $\mathcal{F}_0 = \{\emptyset,\ \Omega\}$, $\mathcal{F}_n = \sigma(Y_1,\ \ldots,\ Y_n)$, $S_0 = 0$ et $S_n = Y_1 + \ldots + Y_n$ pour $n \geq 1$. On note $\mathrm{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$ et on considère le processus défini par $M_0 = 0$ et, pour $n \geq 1$,
$$M_n = \sum_{k=1}^{n} \mathrm{sgn}(S_{k-1}) Y_k.$$

1. Quel est le compensateur de la sous-martingale (S_n^2)?

2. Montrer que (M_n) est une martingale et calculer le compensateur de (M_n^2).

3. Quelle est la décomposition de Doob de $(|S_n|)$? Montrer que M_n est mesurable par rapport à la tribu $\sigma(|S_1|,\ \ldots,\ |S_n|)$.

Exercise 6.7 (Première identité de Wald). Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. de moyenne $\mu \in \mathbb{R}$. Soit τ un temps d'arrêt intégrable. Montrer que
$$\mathbb{E}\left[\sum_{k=1}^{\tau} X_k\right] = \mathbb{E}\left[\tau\right]\mu\ .$$
On pourra commencé par se placer dans le cas où $X_k \geq 0$.

Exercise 6.8. On considère une surmartingale $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, (X_n)_{n\geq 0}, \mathbb{P})$. On suppose qu'il existe une constante M telle que, pour tout $n \geq 1$,
$$\mathbb{E}\left[|X_n - X_{n-1}||\,\mathcal{F}_{n-1}\right] \leq M \text{ p.s.}$$

1. Montrer que si $(V_n)_{n\geq 1}$ est un processus positif tel que pour tout $n \geq 0$, V_n soit $\mathcal{F}_{n-1}$ mesurable, on a :
$$\mathbb{E}\left(\sum_{n=1}^{\infty} V_n |X_n - X_{n-1}|\right) \leq M\mathbb{E}\left(\sum_{n=1}^{\infty} V_n\right)$$

2. Soit ν un temps d'arrêt intégrable. On utilise le fait que :

$$\mathbb{E}(\nu) = \sum_{n\geq 1} \mathbb{P}(\nu \geq n)$$

   (a) Déduire de la question 1 que $\mathbb{E}\left[\sum_{n\geq 1} \mathbb{1}_{\{\nu\geq n\}} |X_n - X_{n-1}|\right] < +\infty$.

   (b) Que vaut $\sum_{n\geq 1} \mathbb{1}_{\{\nu\geq n\}}(X_n - X_{n-1})$? En déduire que X_ν est intégrable.

3. Montrer que $(X_{\nu\wedge p})_{p\geq 0}$ tend vers X_ν dans L^1 lorsque p tend vers l'infini.

4. En déduire que si $\nu_1 \leq \nu_2$ sont deux temps d'arrêt avec ν_2 intégrable, on a :

$$\mathbb{E}\left(X_{\nu_2} | \mathcal{F}_{\nu_1}\right) \leq X_{\nu_1}$$

   On peut se servir, après l'avoir prouvé, du fait que si $A \in \mathcal{F}_{\nu_1}$, alors $A \cap \{\nu_1 \leq k\} \in \mathcal{F}_{\nu_1\wedge k}$.

Exercise 6.9. Sur un espace de probabilité $(\Omega,\ \mathcal{F},\ \mathcal{P})$, on considère des v.a. de Bernoulli de moyenne 1/2 indépendantes, X_n^i, avec $n \geq 1$ et $i \in \{1,\ 2\}$. On pose $S_n^i = \sum_{k=1}^n X_k^i$ et $\nu_i = \inf\{n \geq 1 :\ S_n^i = a\}$ où $a \geq 1$ est un entier. On pose $\nu = \nu_1 \wedge \nu_2$.

1. Montrer que $\mathbb{P}(\nu_i < +\infty) = 1$ pour $i = 1,\ 2$.

2. Pour tout $(i, j) \in \{1,\ 2\}^2$ et tout $n \in \mathbb{N}$, on pose:

$$\begin{aligned} M_n^i &= 2S_n^i - n, \\ M_n^{i,j} &= (2S_n^i - n)(2S_n^j - n) - n\delta_{i,j}, \end{aligned}$$

   où $\delta_{i,j} = \mathbb{1}_{\{i=j\}}$. Montrer que (M_n^i) et $(M_n^{i,j})$ sont des $\mathcal{F}_n$-martingales, avec

$$\mathcal{F}_n = \sigma(X_k^i :\ i \in \{1, 2\},\ k \leq n).$$

3. Montrer que $\mathbb{E}[\nu] \leq 2a$.

4. Montrer que $\mathbb{E}[M_\nu^{i,j}] = 0$.

5. Montrer que $\mathbb{E}[|S_\nu^1 - S_\nu^2|] \leq \sqrt{a}$, en considérant la martingale $M_n^{1,1} - 2M_n^{1,2} + M_n^{2,2}$.

Exercise 6.10 (Seconde identité de Wald). Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. centrées de variance finie σ^2. Soit τ un temps d'arrêt intégrable. Montrer que

$$\mathrm{Var} \sum_{k=1}^{\tau} X_k = \mathbb{E}\left[\tau\right] \sigma^2 \ .$$

Exercise 6.11. Soit (S_n) une marche aléatoire simple sur $\mathbb{Z}$ telle que $S_0 = 0$ et $S_n = U_1 + \cdots + U_n$ pour $n \geq 1$, où les U_i sont indépendantes et de même loi, telles que $0 < \mathbb{P}\{U_i = 1\} = p = 1 - \mathbb{P}\{U_i = -1\} = 1 - q < 1$.

1. Soit $Z_n = (q/p)^{S_n}$. Montrer que $\{Z_n\}_{n\geq 0}$ est une martingale positive.

2. Déduire d'une inégalité maximale appliquée à la martingale $(Z_n)_{n\geq 0}$ que, pour tout $k \in \mathbb{N}$,
$$\mathbb{P}\left\{\sup_{n\geq 0} S_n \geq k\right\} \leq \left(\frac{p}{q}\right)^k,$$
et que lorsque $q > p$
$$\mathbb{E}\left[\sup_{n\geq 0} S_n\right] \leq \frac{p}{q-p}.$$

Exercise 6.12. On considère une suite $(X_n)_{n\geq 1}$ de v.a. réelles indépendantes de même loi normale $\mathcal{N}(m,\ \sigma^2)$ avec $m < 0$. On pose $S_0 = 0$, $S_n = \sum_{i=1}^n X_i$, $\mathcal{F}_n = \sigma(S_0,\ \ldots,\ S_n)$ et
$$W = \sup_{n\geq 0} S_n.$$
Le but de cet exercise est d'établir certaines propriétés de la variable aléatoire W.

1. Montrer que $\mathbb{P}\{W < +\infty\} = +1$.
2. On rappelle que, pour $X_1 \sim \mathcal{N}(m, \sigma^2)$, $\mathbb{E}[\mathrm{e}^{\lambda X_1}] = \mathrm{e}^{\lambda^2\sigma^2/2}\mathrm{e}^{\lambda m}$. Que vaut $\mathbb{E}[\mathrm{e}^{\lambda S_{n+1}} \mid \mathcal{F}_n]$?
3. Montrer qu'il existe un unique $\lambda_0 > 0$ tel que $(\exp(\lambda_0 S_n))_{n\in\mathbb{N}}$ soit une martingale.
4. Montrer que, pour tout $a > 1$, on a:
$$\mathbb{P}\{\mathrm{e}^{\lambda_0 W} > a\} \leq \frac{1}{a}$$
et que pour tout $t > 0$, $\mathbb{P}\{W > t\} \leq \mathrm{e}^{-\lambda_0 t}$.
5. Montrer que:
$$\mathbb{E}[\mathrm{e}^{\lambda W}] = 1 + \lambda \int_0^{+\infty} \mathrm{e}^{\lambda t}\mathbb{P}\{W > t\}dt.$$
En déduire que pour tout $\lambda < \lambda_0$, $\mathbb{E}[\mathrm{e}^{\lambda W}] < +\infty$. En particulier, la v.a. W a des moments à tous les ordres.

Exercise 6.13. On considère une martingale de carré intégrable $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, (M_n)_{n\geq 0}, \mathbb{P})$ telle que $M_0 = 0$ p.s. On note $A_n = \langle M\rangle_n$ le processus croissant associé (*i.e.* le compensateur de la suite $\{M_n^2\}_{n\geq 0}$). Soit $a \in \mathbb{R}$, on pose $\tau_a = \inf\{n \in \mathbb{N};\ A_{n+1} > a^2\}$.

1. Montrer que τ_a est un temps d'arrêt.
2. Montrer que:
$$\mathbb{P}\left(\sup_{n\geq 0} |M_{n\wedge\tau_a}| > a\right) \leq a^{-2}\mathbb{E}[A_\infty \wedge a^2].$$
3. Montrer que:
$$\mathbb{P}\left\{\sup_{n\geq 0} |M_n| > a\right\} \leq \mathbb{P}\left\{A_\infty > a^2\right\} + \mathbb{P}\left\{\sup_{n\in\mathbb{N}} |M_{n\wedge\tau_a}| > a\right\}. \tag{6.19}$$

4. Soit X une v.a. positive. Montrer en appliquant le théorème de Fubini, que pour tout λ

$$\begin{aligned}\int_0^{\lambda} \mathbb{P}\{X > t\}dt &= \mathbb{E}[X \wedge \lambda],\\ \int_0^{+\infty} \frac{\mathbb{E}[X \wedge a^2]}{a^2} da &= 2\mathbb{E}[\sqrt{X}].\end{aligned}$$

5. Montrer que $\mathbb{E}[\sup_{n\geq 0} |M_n|] \leq 3\mathbb{E}[\sqrt{A_\infty}]$. On pourra intégrer (6.19) par rapport à a entre 0 et l'infini.

6. Soit $(Y_n)_{n\geq 1}$ une suite de v.a. centrées indépendantes de même loi et de carré intégrable. On pose $S_0 = 0$ et pour $n \geq 1$, $S_n = Y_1 + \ldots + Y_n$. On définit la filtration $\mathcal{F}_0 = \{\emptyset,\ \Omega\}$ et $\mathcal{F}_n = \sigma(Y_1,\ \ldots,\ Y_n)$ pour $n \geq 1$. Montrer que, si τ est un temps d'arrêt et que $\mathbb{E}[\sqrt{\tau}] < +\infty$ alors $\mathbb{E}[S_\tau] = 0$.

Exercise 6.14. Let $X = (X_t)_{t\in T}$ and $Y = (Y_t)_{t\in T}$ be two real valued processes indexed by the same (arbitrary) set T. Define $U = X + Y$. Suppose that X and Y are U.I. We show in this exercise that U is then also U.I.

1. Let $p, q > 0$. Show that, for all $t \in T$,

$$|X_t|\mathbb{1}_{\{|U_t|>p+q\}} \leq |X_t|\mathbb{1}_{\{|X_t|>p\}} + p\mathbb{1}_{\{|Y_t|>q\}}\ .$$

2. Show that

$$\limsup_{q\to\infty} \sup_{t\in T} \mathbb{P}\left(|Y_t| > q\right) = 0\ .$$

3. For any $p > 0$, Deduce a bound of

$$\limsup_{q\to\infty} \sup_{t\in T} \mathbb{E}\left[|X_t|\mathbb{1}_{\{|U_t|>p+q\}}\right]\ .$$

4. Conclude that U is U.I.

Dans toute la suite, si (M_n) désigne une martingale de carré intégrable, le compensateur de la sous-martingale (M_n^2) est noté $\langle M\rangle_n$.

Exercise 6.15 (La "Martingale classique"). Un joueur mise sur les résultats des jets indépendants d'une pièce **équilibrée** (*i.e.* la probabilité d'obtenir Pile est égale à 1/2). A chaque tour, il mise une somme $S > 0$. Si la pièce tombe sur Pile, son capital augmente de S, si elle tombe sur Face, le joueur perd sa mise et donc son capital diminue de S. Une stratégie populaire en France au XVIIIe siècle est appelée "La Martingale classique". Elle est définie comme suit:

- le joueur s'arrête de jouer la première fois qu'il gagne, *i.e.* dès le premier Pile, ses mises suivantes sont nulles et **son capital n'évolue plus**;
- il double sa mise a chaque tour, c'est-à-dire qu'il mise la somme $S_n = 2^n$ au n-ième tour, tant qu'il n'a pas gagné.

Soit Y_n le capital du joueur au temps n (*i.e.* après n jets de la pièce). On admettra que le capital initial est nul ($Y_0 = 0$, au premier tour il mise donc $S_1 = 2$ euros qu'il doit emprunter), et que le joueur a le droit de s'endetter d'une somme illimitée, c'est-à-dire que Y_n peut devenir négatif, arbitrairement grand en valeur absolue. On désigne par U_n le résultat du n-ième lancer, $U_n = 1$ si le résultat est Pile et $U_n = 0$ sinon, et par $\tau = \inf\{n \geq 1 : U_n = 1\}$ l'instant (aléatoire) du premier "Pile" obtenu. On considère la filtration $\mathcal{F} = \{\mathcal{F}_n\}$ où $\mathcal{F}_n$ désigne la tribu engendrée par les résultats (aléatoires) des n premiers lancers, $\mathcal{F}_n = \sigma(U_k : k \leq n)$.

1. Quelle est la loi de la durée τ du jeu? Montrer qu'il s'agit d'un $\mathcal{F}$-temps d'arrêt. Justifier de façon heuristique la stratégie de la "Martingale classique"?

2. Montrer que la stratégie (S_n) de "La Martingale classique" est $\mathcal{F}$-prévisible.

3. Montrer que $(Y_n)_{n\in\mathbb{N}}$ est une martingale et plus précisément la suite arrêtée au temps d'arrêt τ d'une $\mathcal{F}$-martingale que l'on précisera.

4. Déterminer le processus croissant $\langle Y\rangle_n$. Calculer $\mathbb{E}[\langle Y\rangle_n]$ et discuter la convergence de Y_n dans L^2.

5. Exprimer Y_n en fonction de τ et n, préciser sa loi, et discuter la convergence presque-sûre de Y_n. Quelle est sa limite presque-sûre?

6. La suite Y_n converge-t-elle dans L^1?

7. On suppose maintenant que la banque n'admet pas que le joueur s'endette de plus qu'une valeur limite L (on pourra supposer que $L = 2^k$ pour un $k \geq 1$). Par conséquent, le joueur est obligé de s'arrêter dès que son capital au temps n est strictement inférieur a $-L + 2^{n+1}$. Notons Z_n le capital du joueur à l'instant n.

   (a) Soit N la durée du jeu (*i.e.* le nombre de fois que le joueur mise une somme non nulle). Montrer que N est un temps d'arrêt et préciser sa loi.

   (b) Le processus Z_n est-il une martingale?

   (c) Discuter la convergence presque-sûre et dans L^1 de Z_n et commenter les résultats.

Exercise 6.16. Soit $X_0 = 1$. On définit une suite $(X_n)_{n\in\mathbb{N}}$ récursivement en supposant que pour tout $n \geq 1$, X_n suit la loi uniforme sur $[0,\ 2X_{n-1}]$ conditionnellement à $\sigma(X_k : k \leq n-1)$, c'est-à-dire que l'on peut définir la suite (X_n) de façon récursive en posant: $\forall n \geq 1$, $X_n = 2U_nX_{n-1}$ où les U_n sont i.i.d. de loi uniforme sur $[0,\ 1]$.

1. Montrer que (X_n) est une martingale de carré intégrable par rapport à la filtration $\mathcal{F}_n = \sigma(U_k : k \leq n))$.

2. Calculer le processus croissant $\langle X\rangle_n$ et $\mathbb{E}[\langle X\rangle_n]$. Discuter la convergence de X_n dans L^2. (Indication : on exprimera X_n en fonction de $U_1, \ldots, U_n$ pour tout $n \geq 1$)

3. Discuter la convergence presque-sûre de X_n.

4. Déterminer la limite presque-sûre de X_n. (Indication : Considérer $Y_n = \log(X_n)$ et songer à appliquer la loi forte des grands nombres)

5. Discuter la convergence de X_n dans L^1.

Exercise 6.17. Un joueur dispose initialement de la somme $X_0 = 1$. Il joue à un jeu de hasard, dans lequel il mise à chaque tour une proportion λ de son capital, avec $0 < \lambda \leq 1$. Il a une chance sur deux de doubler sa mise, sinon il perd sa mise. Précisément, l'évolution du capital X_n en fonction du temps n est décrite par: $\forall n \in \mathbb{N}$,

$$X_{n+1} = (1 - \lambda)X_n + \lambda X_n \xi_{n+1},$$

où les ξ_n sont i.i.d., avec $\mathbb{P}\{\xi_n = 2\} = \mathbb{P}\{\xi_n = 0\} = 1/2$.

1. Montrer que (X_n) est une $\mathcal{F}$-martingale de carré intégrable, avec $\mathcal{F}_n = \sigma(\xi_k,\ k \leq n)$.
2. Calculer $\mathbb{E}[X_n]$.
3. Discuter la convergence presque-sûre de X_n lorsque $n \to +\infty$.
4. Calculer $\mathbb{E}[X_n^2]$ par récurrence sur n.
5. Que peut-on en déduire sur la convergence dans L^2 de (X_n)?
6. Déterminer le processus croissant $\langle X \rangle_n$.
7. On suppose désormais que le joueur mise a chaque tour la totalité de son capital, c'est-à-dire $\lambda = 1$.
   (a) Calculer explicitement la loi de X_n.
   (b) Déterminer la limite presque-sûre de (X_n).
   (c) Discuter la convergence de X_n dans L^1. Les X_n sont-ils uniformement intégrables?

Exercise 6.18. Soit $(F_n, \mathcal{F}_n)_{n\geq 0}$ une martingale bornée dans L^1, $F_n^+ = \max(F_n, 0)$ et $F_n^- = -\min(F_n, 0)$.

1. Montrer que $(F_n^+)_{n\geq 0}$ et $(F_n^-)_{n\geq 0}$ sont des sous-martingales.
2. On note $F_n^+ = M_n + A_n$ et $F_n^- = N_n + B_n$ la décomposition de Doob-Meyer de ces deux sous-martingales (A_n et B_n sont des processus prévisibles, $A_0 = B_0 = 0$ et M_n et N_n sont des martingales). Montrer que $A_n = B_n$ $\mathbb{P}$-p.s. [on montrera que $A_n - B_n$ est une martingale]
3. Montrer que $\lim_{n\to\infty} A_n$ existe. On note A_∞ cette limite. Montrer que $\mathbb{E}[A_\infty] < \infty$.
4. On pose $F_n^\oplus = M_n + \mathbb{E}[A_\infty \mid \mathcal{F}_n]$ et $F_n^\ominus = N_n + \mathbb{E}[A_\infty \mid \mathcal{F}_n]$. Montrer que $(F_n^\oplus)_{n\geq 0}$ et $(F_n^\ominus)_{n\geq 0}$ sont des martingales positives.
5. Montrer que $F_n = F_n^\oplus - F_n^\ominus$.
6. Montrer que $F_n^\oplus$ est bornée dans L^1.

On a donc démontré un résultat dû à Krickeberg (1956). Toute martingale bornée dans L^1 est la différence de deux martingales positives bornées dans L^1.

Exercise 6.19. Soit $(F_n)_{n\geq 0}$ un processus adapté et $(V_n)_{n\geq 1}$ un processus prévisible, tous deux par rapport à une même filtration $(\mathcal{F}_n)_{n\geq 0}$. On pose

$$(V\cdot F)_n = \sum_{k=1}^{n} V_k(F_k - F_{k-1}) \ ,$$

la transformée martingale de $(F_k)_{k\geq 0}$ (un équivalent à temps discret de l'intégrale stochastique). On pose pour $n \geq 1$, $D_n = F_n - F_{n-1}$. On suppose dans la suite que $F_0 = 0$ et $\sup_k |V_k| \leq 1$, $\mathbb{P}$-p.s.

1. On suppose tout d'abord que $(F_n)_{n\geq 0}$ est une martingale bornée dans L^2. Montrer que $\mathbb{E}\left[(V\cdot F)_n^2\right] \leq \mathbb{E}\left[F_n^2\right]$. En déduire que $(V\cdot F)_n$ converge $\mathbb{P}$-p.s. et dans L^2.

2. On suppose que $(F_n)_{n\geq 0}$ est une sous-martingale positive bornée, *i.e.* il existe $M > 0$ tel que $\sup_n |F_n| \leq M$, $\mathbb{P}$-p.s.

   (a) Montrer que $\mathbb{E}\left[F_n^2\right] \geq \mathbb{E}\left[F_{n-1}^2\right] + \mathbb{E}\left[D_n^2\right]$.

   (b) En déduire que $\sum_{k=1}^{n} \mathbb{E}\left[D_k^2\right] \leq \mathbb{E}\left[F_n^2\right]$.

   (c) On pose $\hat{F}_n = \sum_{k=1}^{n} \hat{D}_k$ où $\hat{D}_1 = D_1$ et pour $k \geq 2$, $\hat{D}_k = D_k - \mathbb{E}\left[D_k|\,\mathcal{F}_{k-1}\right]$. Montrer que $(\hat{F}_n)_{n\geq 0}$ et $((V\cdot\hat{F})_n)_{n\geq 0}$ sont des martingales.

   (d) Montrer que $\mathbb{E}\left[\hat{F}_n^2\right] \leq \mathbb{E}\left[F_n^2\right]$ et $\mathbb{E}\left[((V\cdot\hat{F})_n)^2\right] \leq \mathbb{E}\left[F_n^2\right]$. En déduire que $(\hat{F}_n)_{n\geq 0}$ et $((V\cdot\hat{F})_n)_{n\geq 0}$ convergent p.s. et dans L^2. En déduire que $\sum_{k=1}^{n} \mathbb{E}\left[D_k|\,\mathcal{F}_{k-1}\right]$ converge p.s. vers une variable aléatoire finie p.s. puis que $\sum_{k=1}^{n} V_k\mathbb{E}\left[D_k|\,\mathcal{F}_{k-1}\right]$ converge p.s. vers une v.a. finie p.s.

   (e) En déduire que $((V\cdot F)_n)_{n\geq 0}$ converge p.s.

3. Soit $(F_n)_{n\geq 0}$ une sous-martingale positive bornée dans L^1 et à accroissements bornés: *i.e.* il existe $M > 0$ tel que $\sup_n |D_n| \leq M$, $\mathbb{P}$-p.s. Pour $c \geq 0$, on note

   $$\tau_c = \inf\{n \geq 0, F_n \geq c\} \ .$$

   On note $F_n^{\tau_c} = F_{n\wedge\tau_c}$.

   (a) Montrer que $((V\cdot F^{\tau_c})_n)_{n\geq 0}$ converge p.s. En déduire que $((V\cdot F)_n)_{n\geq 0}$ converge p.s. sur l'événement $\{\tau_c = \infty\}$.

   (b) Montrer que pour toute suite $(c_n)_{n\geq 0}$ de nombres positifs tels que $\lim_{n\to\infty} c_n = \infty$,

   $$\mathbb{P}\bigcup_{n=1}^{\infty}\{\tau_{c_n} = \infty\} = 1 \ .$$

   (c) En déduire que $((V\cdot F)_n)_{n\geq 0}$ converge p.s.

4. Montrer en utilisant les résultats de l'exercise 6.18 que $((V\cdot F)_n)_{n\geq 0}$ converge p.s. si $(F_n)_{n\geq 0}$ est une martingale bornée dans L^1.

Exercise 6.20. Soit $(S_n)_{n\geq 0}$ une marche aléatoire symétrique, $S_0 = 0$ et $S_n = S_{n-1} + \epsilon_n$, où $(\epsilon_n)_{n\geq 0}$ est une suite de variables aléatoires i.i.d. de loi donnée par $\mathbb{P}(\epsilon_1 = 1) = 1/2$, $\mathbb{P}(\epsilon_1 = -1) = 1/2$. Soit d'autre part $a \in \mathbb{N}^*$.

1. Montrer que $S_n^2 - n$ est une martingale.
2. On pose $T = \inf\{n \geq 0, S_n \notin]-a, a[\}$. Montrer que $\mathbb{E}[T \wedge n] \leq a^2$ et en déduire que $\mathbb{E}\left[S_T^2\right] = \mathbb{E}[T] = a^2$.
3. Calculer les constantes b et c telles que
$$Y_n = S_n^4 - 6nS_n^2 + bn^2 + cn$$
est une martingale.
4. En déduire une expression de $\mathbb{E}\left[T^2\right]$.

Exercise 6.21. Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité muni d'une filtration $\{\mathcal{F}_n\}_{n\geq 0}$ et $X_1,\ X_2,\ \ldots$ une suite de v.a. de carré intégrable sur Ω telles que: $\forall n \in \mathbb{N}$, $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0$. On pose $V_0 = 0$ et, pour $n \geq 1$, $V_n = \mathbb{E}[X_n^2 \mid \mathcal{F}_{n-1}]$. Soient $(Z_n,\ n \geq 1)$ des v.a. telles que Z_n soit $\mathcal{F}_{n-1}$ mesurable. On pose: $M_0 = A_0 = 0$ et, pour $n \geq 1$,
$$M_n = \sum_{k=1}^{n} Z_k X_k, \quad A_n = \sum_{k=1}^{n} Z_k^2 V_k, \quad A_\infty = \lim \uparrow A_n.$$
On suppose que, pour tout n, $\mathbb{E}[A_n] < +\infty$.

1. Montrer que M_n est une martingale de carré intégrable.
2. Montrer que $Y_n = M_n^2 - A_n$ est une martingale.
3. Montrer que si ν est un temps d'arrêt borné, $\mathbb{E}[M_\nu] = 0$ et $\mathbb{E}[M_\nu^2] = \mathbb{E}[A_\nu]$.
4. Soit ν un temps d'arrêt tel que $\mathbb{E}[A_\nu] < +\infty$.
   (a) Montrer que $M_{\nu\wedge n}$ converge p.s. et dans L^2. En déduire que, sur $\{\nu = +\infty\}$, $M_\infty = \lim M_n$ existe p.s. .
   (b) Montrer que $\mathbb{E}[M_\nu] = 0$ et que $\mathbb{E}[M_\nu^2] = \mathbb{E}[A_\nu]$.
   (c) Montrer que $\mathbb{P}\{\sup_{n\leq\nu} |M_n| \geq \rho\} \leq \frac{1}{\rho^2}\mathbb{E}[A_\nu]$.

Exercise 6.22 (0-1 law). Let $(X_n)_{n\in\mathbb{N}}$ be a process of independent random variables. Define its natural tail σ-field by
$$\mathcal{T} = \bigcap_{n\in\mathbb{N}} \sigma(X_p, p \geq n)\ .$$

1. Compute $\mathbb{E}\left[\mathbb{1}_A \middle| \mathcal{F}_n^X\right]$ for all $n \in \mathbb{N}$ and $A \in \mathcal{T}$.
2. Deduce that any $A \in \mathcal{T}$ has probability 0 or 1.

Part III

Introduction to Markov chains

Chapter 7

Markov Chains: basic definitions

7.1 Definition using conditioning

Definition 7.1.1 (Markov Chain)**.** *Let* $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$ *be a filtered probability space and* $(\mathsf{X}, \mathcal{X})$ *be a measurable space. An adapted stochastic process* $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ *on* X *is a* Markov chain *if, for all* $k \in \mathbb{N}$ *and* $A \in \mathcal{X}$*,*

$$\mathbb{P}[X_{k+1} \in A \,|\, \mathcal{F}_k] = \mathbb{P}[X_{k+1} \in A \,|\, X_k] \quad \mathbb{P}\text{-a.s.} \tag{7.1}$$

Let $(\mathcal{G}_k)_{k\in\mathbb{N}}$ be another filtration such that for all $k \in \mathbb{N}$, $\mathcal{G}_k \subset \mathcal{F}_k$. If $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ is a Markov chain and $(X_k)_{k\in\mathbb{N}}$ is adapted to the filtration $(\mathcal{G}_k)_{k\in\mathbb{N}}$, then $((X_k, \mathcal{G}_k))_{k\in\mathbb{N}}$ is also a Markov chain. In particular a Markov chain $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ is always a Markov chain with respect to the natural filtration $\left(\mathcal{F}_k^X\right)_{k\in\mathbb{N}}$.

Clearly the condition given in Definition 7.1.1 is equivalent to having that, for all $Y \in L^1(\Omega, \sigma(X_{k+1}), \mathbb{P})$,

$$\mathbb{E}[Y \,|\, \mathcal{F}_k] = \mathbb{E}[Y \,|\, X_k] \quad \mathbb{P}\text{-a.s.} \tag{7.2}$$

In fact one can even extend Y to the whole future as shown in the following result.

Proposition 7.1.1. *Let* $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ *be a an adapted stochastic process. The following properties are equivalent:*

(i) $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ *is a Markov chain,*

(ii) for all $k \in \mathbb{N}$ *and* $Y \in L^1(\Omega, \sigma(X_l,\, l \geq k), \mathbb{P})$*,*

$$\mathbb{E}[Y \,|\, \mathcal{F}_k] = \mathbb{E}[Y \,|\, X_k] \quad \mathbb{P}\text{-a.s.} \tag{7.3}$$

(iii) for all $k \in \mathbb{N}$*,* $Y \in L^1(\Omega, \sigma(X_l,\, l \geq k), \mathbb{P})$ *and* $Z \in L^\infty(\Omega, \mathcal{F}_k, \mathbb{P})$*,*

$$\mathbb{E}[YZ \,|\, X_k] = \mathbb{E}[Y \,|\, X_k]\, \mathbb{E}[Z \,|\, X_k] \quad \mathbb{P}\text{-a.s.} \tag{7.4}$$

Proof.

(i) $\Rightarrow$ (ii) Consider the property

($\mathcal{P}_n$): (7.3) holds for all $Y = \prod_{j=0}^n Y_j$, where $Y_j \in L^\infty(\Omega, \sigma(X_{k+j}), \mathbb{P})$ for all $j \geq 0$.

We first prove by induction that $\mathcal{P}_n$ holds for all $n \geq 0$. $(\mathcal{P}_0)$ is trivially true. Assume now that $(\mathcal{P}_n)$ holds. The Markov property (7.2) yields

$$\begin{aligned}\mathbb{E}\left[Y_0 \dots Y_n Y_{n+1} \middle| \mathcal{F}_k\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_0 \dots Y_n Y_{n+1} \middle| \mathcal{F}_{k+n}\right] \middle| \mathcal{F}_k\right] \\ &= \mathbb{E}\left[Y_0 \dots Y_n \mathbb{E}\left[Y_{n+1} \middle| \mathcal{F}_{k+n}\right] \middle| \mathcal{F}_k\right] \\ &= \mathbb{E}\left[Y_0 \dots Y_n \mathbb{E}\left[Y_{n+1} \middle| X_{k+n}\right] \middle| \mathcal{F}_k\right] \ .\end{aligned}$$

Since $Y_n \mathbb{E}\left[Y_{n+1} \middle| X_{k+n}\right] \in L^\infty(\Omega, \sigma(X_{k+n}), \mathbb{P})$, we can apply the induction assumption $\mathcal{P}_n$ and get that

$$\begin{aligned}\mathbb{E}\left[Y_0 \dots Y_n\, Y_{n+1} \middle| \mathcal{F}_k\right] &= \mathbb{E}\left[Y_0 \dots Y_n \mathbb{E}\left[Y_{n+1} \middle| X_{k+n}\right] \middle| X_k\right] \\ &= \mathbb{E}\left[Y_0 \dots Y_n \mathbb{E}\left[Y_{n+1} \middle| \mathcal{F}_{k+n}\right] \middle| X_k\right] \\ &= \mathbb{E}\left[Y_0 \dots Y_n\, Y_{n+1} \middle| X_k\right] \ ,\end{aligned}$$

showing that $(\mathcal{P}_{n+1})$ holds. Therefore, $(\mathcal{P}_n)$ is true for all $n \in \mathbb{N}$. Consider the set

$$\mathcal{H} = \left\{ Y \in L^1(\Omega, \sigma(X_j, j \geq k), \mathbb{P}) \,:\, \mathbb{E}\left[Y \middle| \mathcal{F}_k\right] = \mathbb{E}\left[Y \middle| X_k\right] \quad \mathbb{P}\text{-a.s.} \right\} \ .$$

It is easily seen that $\mathcal{H}$ is a vector space. In addition, if Y_n converges to Y in the L^1-sense with $Y_n \in \mathcal{H}$ for all n, then $Y \in \mathcal{H}$. Hence it suffices to show that $\mathbb{1}_A \in \mathcal{H}$ for all $A \in \sigma(X_j, j \geq k)$ to get that $\mathcal{H} = L^1(\Omega, \sigma(X_j, j \geq k), \mathbb{P})$. By the $\pi - \lambda$-theorem it suffices to prove that $\mathbb{1}_A \in \mathcal{H}$ for all A in a π-system that generates $\sigma(X_j, j \geq k)$. This is a consequence of having $\mathcal{P}_n$ for all $n \geq 0$, which concludes the proof of (ii).

(ii) $\Rightarrow$ (iii) If $Y \in L^1(\Omega, \sigma(X_l,\, l \geq k), \mathbb{P})$ and $Z \in L^\infty(\Omega, \mathcal{F}_k, \mathbb{P})$, then (ii) yields

$$\mathbb{E}\left[YZ \middle| \mathcal{F}_k\right] = Z\mathbb{E}\left[Y \middle| \mathcal{F}_k\right] = Z\mathbb{E}\left[Y \middle| X_k\right] \quad \mathbb{P}\text{-a.s.}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[YZ \middle| X_k\right] &= \mathbb{E}\left[\mathbb{E}\left[YZ \middle| \mathcal{F}_k\right] \middle| X_k\right] = \mathbb{E}\left[Z\mathbb{E}\left[Y \middle| X_k\right] \middle| X_k\right] \\ &= \mathbb{E}\left[Z \middle| X_k\right] \mathbb{E}\left[Y \middle| X_k\right] \quad \mathbb{P}\text{-a.s.}\end{aligned}$$

(iii) $\Rightarrow$ (i) If $A \in \mathsf{X}$ and $B \in \mathcal{F}_k$, we obtain from (iii) that

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A(X_{k+1})\mathbb{1}_B\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_A(X_{k+1})\mathbb{1}_B \middle| X_k\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_A(X_{k+1}) \middle| X_k\right] \mathbb{E}\left[\mathbb{1}_B \middle| X_k\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_A(X_{k+1}) \middle| X_k\right] \mathbb{1}_B \middle| X_k\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_A(X_{k+1}) \middle| X_k\right] \mathbb{1}_B\right] \ ,\end{aligned}$$

showing (i).

□

Condition (7.4) means that the future of a Markov chain is conditionnally independent of its past, given its present state.

An important caveat must be made; the Markov property is not hereditary. If $(X_k)_{k\in\mathbb{N}}$ is a Markov chain on X and f is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$, then, unless f is one-to-one, $\{f(X_k), k \in \mathbb{N}\}$ is not necessarily a Markov chain. In particular, if X is a product space, $\mathsf{X} = \mathsf{X}_1 \times \mathsf{X}_2$ and $X_k = (X_{1,k}, X_{2,k})$, $k \geq 0$, then the sequence $(X_{1,k})_{k\in\mathbb{N}}$ may not be a Markov chain, see Exercise 7.2.

7.2 How to use kernels

The most convenient way to define a Markov chain is to rely on Markov kernels (recall Definition 2.1.6). Indeed, if for instance $\mathsf{X} = \mathbb{R}^d$, Eq. (7.1) says that the regular version of the conditional distribution of X_{k+1} given $\mathcal{F}_k$ is given by a Markov kernel. If this kernel does not depend on k, we will say that the Markov chain $(X_k)_{k\in\mathbb{N}}$ is *homogeneous*. As we will see later, in this case, the finite distributions of $X = (X_k)_{k\in\mathbb{N}}$ are entirely characterized by this kernel and the distribution of X_0, called the *initial distribution*.

Before that, we need some basic facts explaining how to handle kernels (notation conventions, composition, tensor product and so on).

7.2.1 Kernel seen as a functional operator

We have seen in Section 2.1.3 that a kernel N on $\mathsf{X}\times\mathcal{Y}$ apply to a measure $\mu \in \mathbb{M}_+(\mathsf{X},\mathcal{X})$, giving a new measure, denoted by μN, belonging to $\mathbb{M}_+(\mathsf{Y},\mathcal{Y})$. Similarly N can also operates on functions. More precisely, given a function $f : \mathsf{Y} \to \bar{\mathbb{R}}$, measurable from $(\mathsf{Y},\mathcal{Y})$ to $(\bar{\mathbb{R}},\mathcal{B}(\bar{\mathbb{R}}))$, we denote by Nf the function defined on X by

$$Nf : x \mapsto \int_{\mathsf{Y}} N(x,\mathrm{d}y) f(y) \;, \qquad x \in \mathsf{X} \;,$$

whenever this integral is well defined (for instance if f is non-negative). Some properties of f are preserved when applied to N, as shown in the following result.

Proposition 7.2.1. *Let N be a kernel on $\mathsf{X}\times\mathcal{Y}$. Then $Nf \in \mathrm{F}_+(\mathsf{X},\mathcal{X})$ for all $f \in \mathrm{F}_+(\mathsf{Y},\mathcal{Y})$.*

Proof. Let $f \in \mathrm{F}_+(\mathsf{Y},\mathcal{Y})$. Of course, by definition, $x \mapsto Nf(x)$ takes non-negative values. We now show that it is measurable. Assume first that f is a simple nonnegative function, i.e. $f = \sum_{i\in I} \beta_i \mathbb{1}_{B_i}$ for a finite collection of nonnegative numbers β_i and sets $B_i \in \mathcal{Y}$. Then, for $x \in \mathsf{X}$, $Nf(x) = \sum_{i\in I} \beta_i N(x,B_i)$, and by the property (i) of Definition 2.1.6, the function Nf is measurable. Let now $f \in \mathrm{F}_+(\mathsf{X},\mathcal{X})$ and let $(f_n)_{n\in\mathbb{N}}$ be a increasing sequence of measurable nonnegative simple functions such that $\lim_{n\to\infty} f_n(x) = f(x)$. Then, by the monotone convergence theorem, for all $x \in \mathsf{X}$, it holds that

$$Nf(x) = \lim_{n\to\infty} Nf_n(x) \;.$$

Therefore, Nf is the pointwise limit of a sequence of nonnegative measurable functions, hence is measurable. □

Note that we use the same symbol N for the kernel defined on $\mathsf{X}\times\mathcal{Y}$ and the resulting $\mathrm{F}_+(\mathsf{Y},\mathcal{Y}) \to \mathrm{F}_+(\mathsf{X},\mathcal{X})$ operator defined by $f \mapsto Nf$. This is justified in the sense that this operator can be seen as an extension from indicator functions $\mathbb{1}_A$ to any non-negative functions. We will also use the notation $N(x,f)$ for $Nf(x)$ and $N(x,\mathbb{1}_A)$ or $N\mathbb{1}_A(x)$ for $N(x,A)$.

Interestingly if one is given an $\mathrm{F}_+(\mathsf{Y},\mathcal{Y}) \to \mathrm{F}_+(\mathsf{X},\mathcal{X})$ operator, it is easy to check if it can be associated to a kernel on $\mathsf{X}\times\mathcal{Y}$. We say that an operator $\mathrm{F}_+(\mathsf{Y},\mathcal{Y}) \to \mathrm{F}_+(\mathsf{X},\mathcal{X})$ is additive and positively homogeneous, i.e. for all $f,g \in \mathrm{F}_+(\mathsf{Y},\mathcal{Y})$ and $\alpha \in \mathbb{R}_+$, it holds that $N(f+g) = Nf + Ng$ and $N(\alpha f) = \alpha Nf$. In addition, if $(f_n)_{n\in\mathbb{N}} \subset \mathrm{F}_+(\mathsf{Y},\mathcal{Y})$ is a non-decreasing sequence of functions, $\lim_{n\to\infty} Nf_n = N(\lim_{n\to\infty} f_n)$ by the monotone convergence theorem. The following lemma shows that every additive and positive homogeneous function satisfying this property may be associated to a kernel.

Lemma 7.2.2. *Let $M : \mathrm{F}_+(\mathsf{Y}, \mathcal{Y}) \to \mathrm{F}_+(\mathsf{X}, \mathcal{X})$ be an additive and positively homogeneous function such that $\lim_{n\to\infty} M(f_n) = M(\lim_{n\to\infty} f_n)$ for every non-decreasing sequence $(f_n)_{n\in\mathbb{N}}$ of functions in $\mathrm{F}_+(\mathsf{Y}, \mathcal{Y})$. Then the function N defined on $\mathsf{X} \times \mathcal{Y}$ by $N(x, A) = M(\mathbb{1}_A)(x)$ is a kernel and $M(f)(x) = \int_{\mathsf{Y}} N(x, \mathrm{d}y) f(y)$ for all for $f \in \mathrm{F}_+(\mathsf{Y}, \mathcal{Y})$.*

Proof. Since M is additive, for each $x \in \mathsf{X}$, the function $A \mapsto N(x, A)$ is additive. Indeed, for $n \in \mathbb{N}^*$, and pairwise disjoints $A_1, \ldots, A_n \in \mathcal{X}$, we have

$$N\left(x, \bigcup_{i=1}^{n} A_i\right) = M\left(\sum_{i=1}^{n} \mathbb{1}_{A_i}\right)(x) = \sum_{i=1}^{n} M(\mathbb{1}_{A_i})(x) = \sum_{i=1}^{n} N(x, A_i) \; .$$

Let $(A_i)_{i\in\mathbb{N}} \subset \mathcal{X}$ be a sequence of pairwise disjoints sets. Then, by additivity and the monotone convergence property of M, we have, for all $x \in \mathsf{X}$,

$$N\left(x, \bigcup_{i=1}^{\infty} A_i\right) = M\left(\sum_{i=1}^{\infty} \mathbb{1}_{A_i}\right)(x) = \sum_{i=1}^{\infty} M(\mathbb{1}_{A_i})(x) = \sum_{i=1}^{\infty} N(x, A_i) \; .$$

This proves that $A \mapsto N(x, A)$ is a measure on $(\mathsf{X}, \mathcal{X})$. Thus N is a kernel on $\mathsf{X} \times \mathcal{Y}$. The last statement follows since any function $f \in \mathrm{F}_+(\mathsf{X}, \mathcal{X})$ is a increasing limit of simple functions. □

7.2.2 Composition of kernels

We just explained how kernels act on a function to obtain a function. It is a simple extension to see that if one replace the measure by an other kernel, we then obtain a new measure.

Proposition 7.2.3 (Composition of kernels)**.** *Let M and N be two kernels on $\mathsf{X} \times \mathcal{Y}$ and $\mathsf{Y} \times \mathcal{Z}$. There exists a kernel MN on $\mathsf{X} \times \mathcal{Z}$, called the composition or the product of M and N such that for all $x \in \mathsf{X}$ and $A \in \mathcal{Z}$ by*

$$MN(x, A) = \int_{\mathsf{Y}} M(x, \mathrm{d}y) N(y, A) \; . \tag{7.5}$$

Proof. For any $A \in \mathcal{Z}$, $y \mapsto N(y, A)$ is a measurable function, and by Proposition 7.2.1, $x \mapsto \int_{\mathsf{Y}} M(x, \mathrm{d}y) N(y, A)$ is a measurable function. For any $x \in \mathsf{X}$, $M(x, \cdot)$ is a measure on $(\mathsf{Y}, \mathcal{Y})$ and by Proposition 2.1.9, $A \mapsto \int M(x, \mathrm{d}y) N(y, A)$ is a measure on $(\mathsf{Z}, \mathcal{Z})$. Hence, the function $(x, A) \mapsto \int_{\mathsf{X}} M(x, \mathrm{d}y) N(y, A)$ is a kernel on $\mathsf{X} \times \mathcal{Z}$. □

Since MN is a kernel on $\mathsf{X} \times \mathcal{Z}$, for any $f \in \mathrm{F}_+(\mathsf{Z}, \mathcal{Z})$, we can define the function $MNf : x \mapsto MNf(x)$, which by Proposition 7.2.1 belongs to $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$. On the other hand, $Nf : y \mapsto Nf(y)$ is a function belonging to $\mathrm{F}_+(\mathsf{Y}, \mathcal{Y})$, and since M is a kernel on $\mathsf{X} \times \mathcal{Y}$, we may consider the function $x \mapsto M[Nf](x)$. A natural question to ask is whether these two quantities are equal.

Proposition 7.2.4. *Let M be a kernel on $\mathsf{X} \times \mathcal{Y}$ and N be a kernel on $\mathsf{Y} \times \mathcal{Z}$. Then, for each $x \in \mathsf{X}$, and $f \in \mathrm{F}_+(\mathsf{Z}, \mathcal{Z})$,*

$$MNf(x) = M[Nf](x) \; . \tag{7.6}$$

Proof. Let $f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$, $A_i \in \mathcal{Z}$, be a simple function. Then,

$$MNf(x) = \sum_{i=1}^{n} \alpha_i MN(x, A_i) = \int_{\mathsf{Y}} M(x, \mathrm{d}y) \sum_{i=1}^{n} \alpha_i N(y, A_i) = M[Nf](x) \; . \tag{7.7}$$

Let $f \in \mathrm{F}_+(\mathsf{Z}, \mathcal{Z})$. The function f is the pointwise limit of a increasing sequence of nonnegative simple functions $(f_n)_{n\in\mathbb{N}}$. Note that the sequence of functions $(Nf_n)_{n\in\mathbb{N}}$ is also increasing , and by the monotone convergence theorem, $\lim_{n\to\infty} Nf_n(x) = Nf(x)$. Therefore, applying again the monotone convergence theorem and (7.7), we have

$$MNf(x) = \lim_{n\to\infty} MNf_n(x) = \lim_{n\to\infty} M[Nf_n](x) = M[Nf](x) \; .$$

□

Given a Markov kernel N on $\mathsf{X}\times\mathcal{X}$, we may define the n-th power of this kernel iteratively. For $x \in \mathsf{X}$ and $A \in \mathcal{X}$, we set $N^0(x,\ A) = \delta_x(A)$ and for $n \geq 1$, we define inductively N^n by

$$N^n(x, A) = \int_{\mathsf{X}} N(x, \mathrm{d}y)\, N^{n-1}(y, A) \; . \tag{7.8}$$

For integers $k, n \geq 0$ this yields the Chapman-Kolmogorov equation.

$$N^{n+k} = N^n\, N^k \; . \tag{7.9}$$

In the case of a discrete state space X, a kernel N can be seen as a matrix with non negative entries indexed by X. Then the k-th power of the kernel N^k defined in (7.8) is simply the k-th power of the matrix N.

7.2.3 Tensor products of kernels

Composition of kernels is the successive application of two kernels, that is we move from

$$\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$$

to

$$\mathsf{X} \to \mathsf{Z} \; .$$

Tensor product also starts from the scheme

$$\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$$

but it yields to

$$\mathsf{X} \to (\mathsf{Y} \times \mathsf{Z}) \; .$$

Proposition 7.2.5. *Let M be a probability kernel on $\mathsf{X}\times\mathcal{Y}$ and N be a probability kernel on $\mathsf{Y}\times\mathcal{Z}$. Then, there exists a kernel $M\otimes N$ on $\mathsf{X}\times(\mathcal{Y}\otimes\mathcal{Z})$, called the tensor product of M and N, such that, for all $f \in \mathbb{F}_+(\mathsf{Y}\times\mathsf{Z}, \mathcal{Y}\otimes\mathcal{Z})$,*

$$M\otimes Nf(x) = \int_{\mathsf{Y}} \left(\int_{\mathsf{Z}} f(y,z) N(y, \mathrm{d}z) \right) M(x, \mathrm{d}y) \; . \tag{7.10}$$

In other words we can write

$$\int_{\mathsf{Y}} \left(\int_{\mathsf{Z}} f(y,z) N(y, \mathrm{d}z) \right) M(x, \mathrm{d}y) = \int_{\mathsf{Y}\times\mathsf{Z}} f(v)\ M \otimes N(x, \mathrm{d}v)\ ,$$

which is often directly written as

$$\int_{\mathsf{Y}\times\mathsf{Z}} f(y,z) M(x, \mathrm{d}y) N(y, \mathrm{d}z) \tag{7.11}$$

Proof. It is the goal of this proof to show that the integral in (7.11) over $\mathsf{Y} \times \mathsf{Z}$ makes sense. At the moment, since we only starts with two kernels M and N, such an integral can only be understood by successively integrating y and z or z and y. Integrating first with respect to y does not make an obvious sense because of the presence of $N(y, \mathrm{d}z)$, since we only know how to integrate *functions* not kernels. That is why, the only definition of $M \otimes N$ we can start with is (7.10), providing that the integrated functions are indeed measurable. When applying the usual construction of product measures and prove the Tonelli theorem (see [7, Section 8, P. 160]), one shows that $z \mapsto f(y,z)$ is measurable for all y and that $y \mapsto \int f(y,z)\mathrm{d}\mu(z)$ is also measurable for any given σ-finite measure μ. Adapting this, one can prove similarly that $y \mapsto \int_{\mathsf{Z}} f(y,z) N(y, \mathrm{d}z)$ is measurable. Hence in (7.10) one can integrate first with respect to z and then with respect to y and obtain a measurable function on $x \in \mathsf{X}$.

Now consider the so obtained mapping $I : \mathrm{F}_+(\mathsf{Y} \times \mathsf{Z}, \mathcal{Y} \otimes \mathcal{Z}) \to \mathrm{F}_+(\mathsf{X}, \mathcal{X})$ defined by

$$If(x) = \int_{\mathsf{Y}} \left(\int_{\mathsf{Z}} f(y,z) N(y, \mathrm{d}z) \right) M(x, \mathrm{d}y)\ .$$

The mapping I is an additive positively homogeneous application. In addition, for any non-decreasing sequence $(f_n)_{n\in\mathbb{N}}$, we have $I[\lim_{n\to\infty} f_n] = \lim_{n\to\infty} I(f_n)$ by applying the monotone convergence Theorem, twice successively. Therefore, Lemma 7.2.2 shows that (7.10) defines a kernel on $\mathsf{X} \times (\mathcal{Y} \otimes \mathcal{Z})$. □

Observe that in the special case where $f = g \otimes h$, we can write

$$M \otimes N(g \otimes h) = M(g \times Nh)\ .$$

By the usual characterization theorem on π-systems (Theorem 2.0.2), if $M \otimes N$ is a probability measure, it is in fact sufficient to characterize it by computing $M \otimes N$ on $\mathbb{1}_A \otimes \mathbb{1}_B$, thus by compute=ing $M(\mathbb{1}_A \times N\mathbb{1}_B)$, for all $A \in \mathcal{Y}$ and $B \in \mathcal{Z}$. This useful in the following lemma.

Lemma 7.2.6. *Let $(\mathsf{X}_k, \mathcal{X}_k)$, $k = 0,1,2,3$ be measurable spaces. Let M_k be probability kernels on $\mathsf{X}_k \times \mathcal{X}_{k+1}$ for $k = 0,1,2,3$.*

(i) The kernel $M_0 \otimes M_1$ is a probability kernel on $\mathsf{X}_0 \times (\mathcal{X}_1 \otimes \mathcal{X}_2)$.

(ii) We have

$$[M_0 \otimes (M_1 \otimes M_2)]\,(\mathbb{1}_{\mathsf{X}_1} \otimes \cdot) = (M_0 M_1) \otimes M_2\ .$$

(iii) We have

$$M_0 \otimes (M_1 \otimes M_2) = (M_0 \otimes M_1) \otimes \bar{M}_2\ ,$$

where $\bar{M}_2$ is the kernel defined on $(\mathsf{X}_1 \times \mathsf{X}_2) \times \mathcal{X}_3$ defined by $\bar{M}_2 f = \mathbb{1}_{\mathsf{X}_1} \otimes (M_2 f)$.

The proof is left as an exercise (Exercise 7.1). Note that in (ii), parentheses are used to precise that in the succession of tensor product $M \otimes (N \otimes P)$, one needs to go from right to left (first $N \otimes P$, then $M\otimes$ the result of $N \otimes P$). In fact in a succession of tensor products $M \otimes N \otimes P$, only this order makes sense, hence the parentheses will be dropped. In other words when writing $M_0 \otimes M_1 \otimes \cdots \otimes M_n$ for kernels M_k on $\mathsf{X}_k \times \mathcal{X}_{k+1}$ for $k = 0, \ldots, n$ and measurable spaces $(\mathsf{X}_k, \mathcal{X}_k)$, $k = 0, \ldots, n+1$, one always means this product from right to left :

$$M_0 \otimes \cdots \otimes M_n = M_0 \left(\otimes M_1 \left(\otimes \ldots (M_{n-1} \otimes M_n) \ldots\right)\right) .$$

Consequently, for all $f = f_1 \otimes \cdots \otimes f_{n+1}$ with $f_k \in \mathrm{F}_+(\mathsf{X}_k, \mathcal{X}_k)$ for $k = 1, \ldots, n+1$, we have

$$M_0 \otimes \cdots \otimes M_n(f) = M_0 \left(f_1 \times M_1 \left(f_2 \times \ldots (f_{n-1} \times M_{n-1} \left(f_n \times M_n f_{n+1}\right)) \ldots\right)\right) . \tag{7.12}$$

And in the case of probability kernel, this completely characterizes the kernel, even restricting oneself to $f_k = \mathbb{1}_{A_k}$ with $A_k \in \mathcal{X}_k$ for all k. Note that we obtain as in Lemma 7.2.6 (ii) that, for probability kernels, marginalizing on the first $1 \leq m < n$ components gives, for all $f \in \mathrm{F}_+(\prod_{i=m+1}^n \mathsf{X}_i, \bigotimes_{i=m+1}^n \mathcal{X}_i)$

$$M_0 \otimes \cdots \otimes M_n(\mathbb{1}_{\mathsf{X}_1 \times \cdots \times \mathsf{X}_m} \otimes f) = (M_0 \ldots M_m) \otimes M_{m+1} \cdots \otimes M_n(f) . \tag{7.13}$$

In particular, the n-th tensorial power $P^{\otimes n}$ of a Markov kernel P on $X \times \mathcal{X}$ is the kernel on $(\mathsf{X}, \mathcal{X}^{\otimes n})$ defined by

$$P^{\otimes n} f(x) = \int_{\mathsf{X}^n} f(x_1, \ldots, x_n) P(x, \mathrm{d}x_1) P(x_1, \mathrm{d}x_2) \cdots P(x_{n-1}, \mathrm{d}x_n) , \tag{7.14}$$

where again this integral is defined by integrating first in x_n, then in x_{n-1} and so on down to x_1. The following application of (7.13) will be useful. For any probabilimeasure on $(\mathsf{X}, \mathcal{X})$ and any Markov kernel P on $X \times \mathcal{X}$, we have, for all $1 \leq m < n$,

$$\nu \otimes P^{\otimes n}(\mathsf{X}^m \times A) = (\nu P^m) \otimes P^{\otimes(n-m)}(A) \qquad \text{for all } A \in \mathcal{X}^{\otimes(n-m)} . \tag{7.15}$$

7.3 Homogeneous Markov chains

The finite dimensional (fidi) distributions of a Markov chain can entirely be described by a succession of kernels (provided that conditional distributions all admit regular versions). We here provide some details of how this description works in the special case of Homogeneous Markov chains.

Definition 7.3.1 (Homogeneous Markov Chain)**.** *Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. Let ν be a probability measure on $(\mathsf{X}, \mathcal{X})$ and let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. Let $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space. An adapted stochastic process $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ is called a homogeneous Markov chain with Markov kernel P and initial distribution ν if, for all $k \geq 0$ and $A \in \mathcal{X}$,*

(i) $\mathbb{P}(X_0 \in A) = \nu(A)$,

(ii) $\mathbb{P}\left[X_{k+1} \in A \,|\, \mathcal{F}_k\right] = P(X_k, A)$ $\mathbb{P}$-a.s.

Remark 7.3.1. *Condition (ii) is equivalent to saying that the conditional distribution of X_{k+1} given $\mathcal{F}_k$ admits a regular version given by $\mathbb{P}^{X_{k+1}|\mathcal{F}_k} = P(X_k, \cdot)$.*

Remark 7.3.2. *Assume that $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ is a homogeneous Markov chain. Then, $\left((X_k, \mathcal{F}_k^X)\right)_{k\in\mathbb{N}}$ is also a homogeneous Markov chain. Unless specified otherwise, we will always consider the natural filtration and we will simply write that $(X_k)_{k\in\mathbb{N}}$ is a homogeneous Markov chain.*

Proposition 7.3.1. *Let P be a Markov kernel on $(\mathsf{X}, \mathcal{X})$ and ν be a probability measure on $\mathsf{X} \times \mathcal{X}$. An X-valued random process $(X_k)_{k\in\mathbb{N}}$ is a homogeneous Markov chain with kernel P and initial distribution ν if and only if, for all $k \in \mathbb{N}$, the distribution $(X_0, \dots, X_k)$ is $\nu \otimes P^{\otimes k}$. Moreover it then follows that X_k has distribution νP^k.*

Proof of Proposition 7.3.1. Note that, applying Lemma 7.2.6(iii), recursively we get that, for all $k \geq 0$,

$$\nu \otimes P^{\otimes(k+1)} = (\nu \otimes P^{\otimes k}) \otimes \bar{P}^{(k)} \ , \tag{7.16}$$

where $\bar{P}^{(k)}$ is the kernel on $\mathsf{X}^{k+1} \times \mathcal{X}$ by $\bar{P}^{(k)} f = \mathbb{1}_{\mathsf{X}^k} \otimes (Pf)$ for all $f \in \mathrm{F}_+(\mathsf{X}, \mathcal{X})$.

Next by Theorem 2.1.10 and (7.16), we get that, for any $k \geq 0$, if $\mathbb{P}^{(X_0,\dots,X_k)} = \nu \otimes P^{\otimes k}$, then we have $\mathbb{P}^{(X_0,\dots,X_k,X_{k+1})} = \nu \otimes P^{\otimes(k+1)}$ if and only if $\mathbb{P}^{(X_0,\dots,X_k)} = \nu \otimes P^{\otimes k}$ and $\mathbb{P}^{X_{k+1}|(X_0,\dots,X_k)} = \bar{P}^{(k)}$ and the latter exactly means that, for all $A \in \mathcal{X}$,

$$\mathbb{P}\left[X_{k+1} \in A \,|\, X_0, \dots, X_k\right] = \bar{P}^{(k)}((X_0, \dots, X_k), A) = P(X_k, A) \ .$$

From this we easily get the claimed equivalence by easy induction on k.

It remains to prove the last assertion. Using that $\mathbb{P}^{X_k} = \mathbb{P}^{(X_0,\dots,X_k)}(\mathsf{X}^k \times \cdot)$ and applying (7.12) with $M_0 = \nu$, $M_1 = \cdots = M_k = P$, $f_1 = \cdots = f_k = \mathbb{1}_{\mathsf{X}}$ and $f_{k+1} = \mathbb{1}_A$ with $A \in \mathcal{X}$, we get, whenever $\mathbb{P}^{(X_0,\dots,X_k)} = \nu \otimes P^{\otimes k}$, $\mathbb{P}^{X_k}(A) = \nu P^k \mathbb{1}_A$ and the proof is concluded. □

Under appropriate conditions on the structure of the state space X, every homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$ with value in X may be represented as a functional autoregressive process, i.e., $X_{k+1} = f(X_k, Z_{k+1})$ where $(Z_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of random variables with values in a measurable space $(\mathsf{Z}, \mathcal{Z})$, X_0 is independent of $(Z_k)_{k\in\mathbb{N}}$ and f is a measurable function from $(\mathsf{X} \times \mathsf{Z}, \mathcal{X} \otimes \mathcal{Z})$ into $(\mathsf{X}, \mathcal{X})$.

This can be easily proved for a real valued Markov chain $(X_k)_{k\in\mathbb{N}}$ with initial distribution ν and Markov kernel P. Let X be a real-valued random variable and let $F(x) = \mathbb{P}(X \leq x)$ be the cumulative distribution function of X. Let F^{-1} be the quantile function, defined as the generalized inverse of F by

$$F^{-1}(u) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq u\} \ . \tag{7.17}$$

The right continuity of F implies that $u \leq F(x) \Leftrightarrow F^{-1}(u) \leq x$. Therefore, if Z is uniformly distributed on $[0, 1]$, $F^{-1}(Z)$ has the same distribution as X, since $\mathbb{P}(F^{-1}(Z) \leq t) = \mathbb{P}(Z \leq F(t)) = F(t) = \mathbb{P}(X \leq t)$.

Define $F_0(t) = \nu((-\infty, t])$ and $g = F_0^{-1}$. Consider the function F from $\mathbb{R} \times \mathbb{R}$ to $[0, 1]$ defined by $F(x, x') = P(x, (-\infty, x'])$. Then, for each $x \in \mathbb{R}$, $F(x, \cdot)$ is a cumulative distribution function. Let the associated quantile function $f(x, \cdot)$ be defined by

$$f(x, u) = \inf\left\{x' \in \mathbb{R} \,:\, F(x, x') \geq u\right\} \ . \tag{7.18}$$

The function $(x, u) \mapsto f(x, u)$ is Borel measurable since $(x, x') \mapsto F(x, x')$ is itself a Borel measurable function (see Exercise 7.3). If Z is uniformly distributed on $[0, 1]$, then, for all $x \in \mathbb{R}$ and $A \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}\left(f(x, Z) \in A\right) = P(x, A) \ .$$

Let $(Z_k)_{k\in\mathbb{N}}$ be a sequence of i.i.d. random variables, uniformly distributed on $[0,1]$. Define a sequence of random variables $(X_k)_{k\in\mathbb{N}}$ by $X_0 = g(Z_0)$ and for $k \geq 0$,

$$X_{k+1} = f(X_k, Z_{k+1}) \, .$$

Then, $(X_k)_{k\in\mathbb{N}}$ is a Markov chain with Markov kernel P and initial distribution ν.

We state without proof a general result for reference only; it will not be needed in the sequel.

Theorem 7.3.2. *Let $\mathsf{X} = \mathbb{R}^d$ with $d \geq 1$ endowed with $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$. Let P be a Markov kernel on $\mathsf{X}\times\mathcal{X}$ and ν be a probability on $(\mathsf{X},\mathcal{X})$. Let $(Z_k)_{k\in\mathbb{N}}$ be a sequence of i.i.d. random variables uniformly distributed on $[0,1]$. There exists a measurable application g from $([0,1], \mathcal{B}([0,1]))$ to $(\mathsf{X},\mathcal{X})$ and a measurable application f from $(\mathsf{X}\times[0,1], \mathcal{X}\otimes\mathcal{B}([0,1]))$ to $(\mathsf{X},\mathcal{X})$ such that the sequence $(X_k)_{k\in\mathbb{N}}$ defined by $X_0 = g(Z_0)$ and $X_{k+1} = f(X_k, Z_{k+1})$ for $k \geq 0$, is a homogeneous Markov chain with initial distribution ν and Markov kernel P.*

From now on, we will almost uniquely deal with homogeneous Markov chain and we will, for simplicity, omit to mention *homogeneous* in the statements.

7.4 The canonical Chain

Consider a homogeneous Markov chain $X = (X_k)_{k\in\mathbb{N}}$ on the state space $(\mathsf{X},\mathcal{X})$ and defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$. Then, by Proposition 7.3.1, the initial distribution ν and the transition kernel P of the Markov chain entirely determines the fidi distributions of the process X, which in turns determines the distribution $\mathbb{P}^X$, see Theorem 4.2.1.

The question arises now whether for any Markov kernel P on $\mathsf{X}\times\mathcal{X}$ and $\nu \in \mathbb{M}_1(\mathsf{X},\mathcal{X})$, on can indeed define a homogeneous Markov chain $X = (X_k)_{k\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$, with initial distribution ν and transition kernel P, or, equivalently, so that $\mathbb{P}^X$ indeed corresponds to the fidi distributions given in Proposition 7.3.1.

The answer is given by using the canonical process $\xi = (\xi_k)_{k\in\mathbb{N}}$ of Definition 4.2.6 (with $T = \mathbb{N}$), and applying the Kolmogorov theorem (Theorem 4.2.2).

Theorem 7.4.1. *Let $\mathsf{X} = \mathbb{R}^p$ and $\mathcal{X} = \mathcal{B}(\mathsf{X}^p)$. Let P be a Markov kernel P on $\mathsf{X}\times\mathcal{X}$ and $\nu \in \mathbb{M}_1(\mathsf{X},\mathcal{X})$. Then, there exists a unique probability $\mathbb{P}_{P,\nu}$ on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ such that the canonical process $\xi = (\xi_k)_{k\in\mathbb{N}}$ on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ is a Markov chain with initial distribution ν and kernel P.*

Proof. Set, for $k \in \mathbb{N}$ and $f \in \mathrm{F}_+(\mathsf{X}^{k+1}, \mathcal{X}^{\otimes(k+1)})$, $\mu_k = \nu \otimes P^{\otimes k}$. From this one deduce μ_I for all finite set $I \subset \mathbb{N}$ and Condition (4.5) holds. By Kolmogorov's theorem (Theorem 4.2.2), we get in particular the existence and uniqueness of $\mathbb{P}_{P,\nu}$ on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ such that for all $n \geq 0$ and all $A_0, \ldots, A_n \in \mathcal{X}$,

$$\mathbb{P}_{P,\nu}(A_0 \times \cdots \times A_n \times \mathsf{X}^{\{n+1,n+2,\ldots\}}) = \nu \otimes P^{\otimes n}(A_0 \times \cdots \times A_n) \, .$$

Hence, by Proposition 7.3.1, the canonical process is a homogeneous Markov chain with kernel P and initial distribution ν. □

Remark 7.4.1. *The same result of course holds for a more general topological spaces* X*, provided that the Kolmogorov theorem continues to hold. From now on, we will repeatedly assume that the Kolmogorov theorem applies without mentioning any assumption on* X*, in order to be able to apply Theorem 7.4.1. In other words the fact that Theorem 7.4.1 will be implicit whenever we consider a homogeneous Markov chain.*

This construction gives rise to so called the *canonical chain*.

Definition 7.4.1 (Canonical Markov Chain)**.** *The canonical Markov chain with kernel* P *on* $\mathsf{X}\times\mathcal{X}$ *is the canonical process* $\xi = (\xi_k)_{k\in\mathbb{N}}$ *on the canonical filtered space* $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}_k^{\xi}\right)_{k\in\mathbb{N}}\right)$ *endowed with the collection of probability measures* $\{\mathbb{P}_{P,\nu}, \nu \in \mathbb{M}_1(\mathsf{X},\mathcal{X})\}$ *given by Theorem 7.4.1.*

The notation $\mathbb{P}_{P,\nu}$ is useful when one considers a homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$ with a fixed kernel P but with an initial distribution ν that can change from one line to another. Similarly, the expectation associated to $\mathbb{P}_{P,\nu}$ is denoted by $\mathbb{E}_{P,\nu}$ and for $x \in \mathsf{X}$, $\mathbb{P}_{P,x}$ and $\mathbb{E}_{P,x}$ are usual shorthand symbols for $\mathbb{P}_{P,\delta_x}$ and $\mathbb{E}_{P,\delta_x}$. Finally we will omit P from the notation when there is no ambiguity about the Markov kernel.

Proposition 7.4.2. *Consider the canonical filtered space* $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}_k^{\xi}\right)_{k\in\mathbb{N}}\right)$ *endowed with the collection of probability measures* $\{\mathbb{P}_{P,\nu}, \nu \in \mathbb{M}_1(\mathsf{X},\mathcal{X})\}$ *defined as above from the Markov kernel* P *on* $(\mathsf{X},\mathcal{X})$*. Then for all* $A \in \mathcal{X}^{\otimes\mathbb{N}}$*, the function* $x \mapsto \mathbb{P}_{P,x}(A)$ *is* $\mathcal{X}$*-measurable. Moreover, for all* $\nu \in \mathbb{M}_1(\mathsf{X},\mathcal{X})$ *and* $A \in \mathcal{X}^{\otimes\mathbb{N}}$*, we have*

$$\mathbb{P}_{P,\nu}(A) = \int_{\mathsf{X}} \mathbb{P}_{P,x}(A)\, \nu(\mathrm{d}x)\ . \tag{7.19}$$

Proof. Let $\mathcal{M}$ be the class of sets $A \in \mathcal{X}^{\otimes\mathbb{N}}$ satisfying that $x \mapsto \mathbb{P}_{P,x}(A)$ is $\mathcal{X}$-measurable and for which (7.19) holds. The set $\mathcal{M}$ is a λ-system and contains all the sets of the form $\prod_{i=0}^{n} A_i \times \mathsf{X}^{\{n+1,n+2,\dots\}}$, $A_i \in \mathcal{X}$, $n \in \mathbb{N}$, by Proposition 7.3.1. Hence Theorem 2.0.1 gives that $\mathcal{M} = \mathcal{X}^{\otimes\mathbb{N}}$. □

From this result we can introduce, we see that the mapping $(x, A) \mapsto \mathbb{P}_{P,x}(A)$ defines a probability kernel, hence we introduce the following definition.

Definition 7.4.2 (Canonical kernel)**.** *Consider the canonical filtered space* $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}_k^{\xi}\right)_{k\in\mathbb{N}}\right)$ *and let* $\mathbb{P}_{P,x}$ *be defined as above from the Markov kernel* P *on* $(\mathsf{X},\mathcal{X})$ *for all* $x \in \mathsf{X}$*. Then the probability kernel* $(x, A) \mapsto \mathbb{P}_{P,x}(A)$ *on* $\mathsf{X} \times \mathcal{X}^{\otimes\mathbb{N}}$ *is called the* canonical kernel *and is denoted by* $\mathbb{K}_P$ *hereafter.*

Using this definition, we can interpret the right-hand side of (7.19) as a composition of ν with the canonical kernel, that is by writing

$$\mathbb{P}_{P,\nu} = \nu\mathbb{K}_P\ , \tag{7.20}$$

from which it also follows that, for all $Y \in \mathrm{F}_+(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$,

$$\mathbb{E}_{P,\nu}[Y] = \nu\mathbb{K}_P(Y) = \int_{\mathsf{X}} \mathbb{E}_{P,x}[Y]\, \nu(\mathrm{d}x)\ . \tag{7.21}$$

And from this, it also follows that a real-valued Borel function Y on $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}\right)$ is in $L^1(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_{P,\nu})$ if and only if $\int_{\mathsf{X}} \mathbb{E}_{P,x}[|Y|]\, \nu(\mathrm{d}x) < \infty$, in which case (7.21) continues to hold.

The canonical kernel is also convenient for describing conditional distributions as indicated by the following results, which will be of high interest in the next chapter.

Theorem 7.4.3. *Consider the canonical filtered space* $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}_k^{\xi}\right)_{k\in\mathbb{N}}\right)$ *endowed with the collection of probability measures* $(\mathbb{P}_\nu)_{\nu\in\mathbb{M}_1(\mathsf{X},\mathcal{X})}$ *defined as above from the Markov kernel* P *on* $(\mathsf{X}, \mathcal{X})$*. Then, for all* $\nu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$*, under* $\mathbb{P}_{P,\nu}$*, the canonical kernel* $\mathbb{K}_P$ *is a regular version of the conditional distribution of the identity random variable* $\xi = (\xi_k)_{k\geq 0} : \omega \mapsto \omega$ *given* ξ_0*,* $\mathbb{P}_{P,\nu}^{\xi|\xi_0} = \mathbb{K}_P$*.*

Proof. For all $B \in \mathcal{X}^{\otimes\mathbb{N}}$ and $A \in \mathcal{X}$, we have

$$\mathbb{E}_{P,\nu}\left[\mathbb{1}_A(\xi_0)\mathbb{1}_B\right] = \int_{\mathsf{X}} \mathbb{E}_{P,x}\left[\mathbb{1}_A(\xi_0)\mathbb{1}_B\right] \nu(\mathrm{d}x) = \int_{\mathsf{X}} \mathbb{1}_A(x)\mathbb{E}_{P,x}\left[\mathbb{1}_B\right] \nu(\mathrm{d}x) = \mathbb{E}_{P,\nu}\left[\mathbb{1}_A(\xi_0)\mathbb{E}_{P,\xi_0}\left[\mathbb{1}_B\right]\right] ,$$

since $\xi_0 = x$ $\mathbb{P}_{P,x}$-a.s. and $\xi_0 \sim \nu$ under $\mathbb{P}_{P,\nu}$. Now the latter equation can be interpreted as

$$\mathbb{E}_{P,\nu}\left[\mathbb{1}_C\mathbb{1}_B\right] = \mathbb{E}_{P,\nu}\left[\mathbb{1}_C\mathbb{K}_P(\xi_0, B)\right] \qquad \text{for all } C \in \sigma(\xi_0) ,$$

which achieves the proof. □

We thus get , for all $Y \in L^1(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_{P,\nu})$,

$$\mathbb{E}_{P,\nu}[Y|\xi_0] = \mathbb{K}_P Y(\xi_0) = \mathbb{E}_{P,\xi_0}[Y] \quad \mathbb{P}_{P,\nu}\text{-a.s. .}$$

Note also that using 2.1.10, since under $\mathbb{P}_{P,\nu}$, ξ_0 has distribution ν, the assertion of Theorem 7.4.3 can be summarized by the following formula

$$\mathbb{P}_{P,\nu}^{(\xi_0,\xi)} = \nu \otimes \mathbb{K}_P . \tag{7.22}$$

It is interesting to compare this formula with (7.20), having in mind that $\mathbb{K}_P$ has the specific property to be of the form $\mathbb{K}_P(x, A \times B) = \delta_x(A)\mathbb{K}_P(x, \mathsf{X} \times B)$ for all $A \in \mathcal{X}$ and $B \in \mathcal{X}^{\otimes\mathbb{N}^*}$, which is more or less what is exploited in the proof of Theorem 7.4.3.

An event might be almost surely true with respect to one such probability measure $\mathbb{P}_{P,\mu}$ and almost surely wrong with respect to another one $\mathbb{P}_{P,\nu}$. This is obvious for an event of the form $\{\xi_0 \in A\}$. However for certain kernels and events (typically involving the limit behavior of ξ_n as $n \to \infty$), the following definition can be useful.

Definition 7.4.3. *An event is* $\mathbb{P}_{P,*}$-a.s. *if it is* $\mathbb{P}_{P,\nu}$-a.s. *for all initial distribution* ν*.*

The following Lemma is an immediate consequence of (7.19) in Proposition 7.4.2.

Lemma 7.4.4. *Let* $A \in \mathcal{X}^{\otimes\mathbb{N}}$*. We have* $X \in A$ $\mathbb{P}_{P,*}$-a.s. *if and only if it is true* $\mathbb{P}_{P,x}$-a.s. *for all* $x \in \mathsf{X}$*.*

7.5 Complements

7.5.1 Sampled kernel and resolvent kernel

Definition 7.5.1 (Sampled kernel, m-skeleton, resolvent kernel)**.** *Let a be a probability measure on $\mathbb{N}$, that is a sequence $\{a(n),\ n \in \mathbb{N}\}$ such that $a(n) \geq 0$ for all n and $\sum_{k=0}^{\infty} a(n) = 1$. Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. The* sampled kernel *$K_a$ is defined by*

$$K_a(x, A) = \sum_{n=0}^{\infty} a(n) P^n(x, A) \ , \quad x \in \mathsf{X},\ A \in \mathcal{X} \ . \tag{7.23}$$

(i) If $a = \delta_m$ for an integer $m \in \mathbb{N}$, then $K_{\delta_m} = P^m$ is the kernel of the so-called m-skeleton.

(ii) If a_ϵ is the geometric distribution with mean $1/\epsilon$, $\epsilon > 0$, i.e.

$$a_\epsilon(n) = (1 - \epsilon)\, \epsilon^n \ , \quad n \in \mathbb{N} \ , \tag{7.24}$$

then K_{a_ϵ} is referred to as the resolvent *kernel.*

If a and b are two sequences of real numbers, then $a * b$ is the convolution of a and b defined, for $k \in \mathbb{N}$ by

$$a * b(n) = \sum_{k=0}^{n} a(k) b(n - k) \ .$$

Lemma 7.5.1. *If a and b are probability measures on $\mathbb{N}$, then the sampled kernels K_a and K_b satisfy the generalized Chapman-Kolmogorov equations*

$$K_{a*b}(x, A) = \int K_a(x, \mathrm{d}y) K_b(y, A) \tag{7.25}$$

Proof. Applying the definition of the sampled kernel and the Chapman-Kolmogorov equation (7.9) yields (note that all the terms in the sum below are nonnegative)

$$\begin{aligned} K_{a*b}(x, A) &= \sum_{n=0}^{\infty} P^n(x, A) a * b(n) = \sum_{n=0}^{\infty} P^n(x, A) \sum_{m=0}^{n} a(m) b(n - m) \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{n} \int P^m(x, \mathrm{d}y) P^{n-m}(y, A) a(m) b(n - m) \\ &= \int \sum_{m=0}^{\infty} P^m(x, \mathrm{d}y) a(m) \sum_{n=m}^{\infty} P^{n-m}(y, A) b(n - m) = \int K_a(x, \mathrm{d}y) K_b(y, A) \ . \end{aligned}$$

□

7.5.2 Markov chains of order p

Definition 7.5.2 (Markov Chain of order p)**.** *Let $p \geq 1$ be an integer. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. Let $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space. An adapted stochastic process $((X_k, \mathcal{F}_k))_{k\in\mathbb{N}}$ is called a Markov chain of order p if the process $\{(X_k, \ldots, X_{k+p-1}), k \in \mathbb{N}\}$ is a Markov chain with values in X^p.*

Let $(X_k)_{k\in\mathbb{N}}$ be a Markov chain of order $p \geq 2$ and let K_p be the kernel of the chain $(\mathbf{X}_k)_{k\in\mathbb{N}}$ with $\mathbf{X}_k = X_{k:k+p-1}$, that is

$$\mathbb{P}\left[\mathbf{X}_1 \in A_1 \times \cdots \times A_p \,|\, \mathbf{X}_0\right] = K_p(\mathbf{X}_0, A_1 \times \cdots \times A_p) \quad \mathbb{P}\text{-a.s.}$$

Since $\mathbf{X}_0$ and $\mathbf{X}_1$ have $p-1$ common components, the kernel K_p has a particular form. More precisely, defining the kernel K on $\mathsf{X}^p \times \mathcal{X}$ by

$$K(x_{0:p-1}, A) = K_p(x_{0:p-1}, \mathsf{X}^{p-1} \times A)$$

we obtain that

$$K_p(x_{0:p-1}, A_1 \times \cdots \times A_p) = \delta_{x_1}(A_1) \cdots \delta_{x_{p-1}}(A_{p-1}) K(x_{0:p-1}, A_p) \; .$$

We thus see that an equivalent definition of a homogeneous Markov chain of order p is the existence of a kernel K on $\mathsf{X}^p \times \mathcal{X}$ such that for all $n \geq 0$,

$$\mathbb{E}\left[X_{n+p} \in A | \, \mathcal{F}_{n+p-1}^X\right] = K(x_{0:p-1}, A) \quad \mathbb{P}\text{-a.s.}$$

Similarly to Theorem 7.3.2, if $\mathsf{X} = \mathbb{R}^d$ with $d \geq 1$ endowed with $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$, a Markov chain of order p can be expressed as a functional autoregressive process of order p, i.e. there exists an i.i.d. sequence $(Z_k)_{k\in\mathbb{N}}$ of random variables uniformly distributed on $[0,1]$ and a measurable function F defined on $\mathsf{X}^p \times [0,1]$ such that

$$X_{k+p+1} = F(X_{k+1:k+p}, Z_{k+p+1}) \; .$$

7.6 Exercises

Exercise 7.1. Prove Lemma 7.2.6.

Exercise 7.2. Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. Let P and Q be two Markov and probability kernels on $\mathsf{X} \times \mathcal{X}$ and $\mathsf{X} \times \mathcal{Y}$, respectively. Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be a filtration, $(X_k, \mathcal{F}_k)_{k\in\mathbb{N}}$ be a homogeneous Markov chain with kernel P and suppose that $(Y_k)_{k\in\mathbb{N}}$ is $(\mathcal{F}_k)_{k\in\mathbb{N}}$-adapted and such that, for all $k \geq 0$, the conditional distribution of Y_k given $\mathcal{F}_{k-1} \vee \sigma(X_k)$ is $Q(X_k, \cdot)$. Set $Z_k = (X_k, Y_k)$ for all $k \geq 0$.

1. In which case $(X_k)_{k\in\mathbb{N}}$ is i.i.d., depending on the initial distribution μ taken for X_0 ?
2. Show that $(Z_k)_{k\in\mathbb{N}}$ is a homogeneous Markov with kernel R defined by
$$R((x, y), A) = P \otimes Q(x, A) \quad x \in \mathsf{X},\ y \in \mathsf{Y},\ A \in \mathcal{X} \otimes \mathcal{Y}\ .$$
3. Determine all possible kernels P and Q when $\mathsf{X} = \mathsf{Y} = \{0, 1\}$ using $(a, b), (c, d) \in [0, 1]^2$.

We let p and q denote the corresponding Markov matrices on $\{0, 1\}^2$.

4. Show that for all $k \geq 1$, $\mathbb{E}\left[Y_k \middle| X_{k-1}, Y_{k-1}\right] = pq(X_{k-1}, 1)$.
5. Is $(Y_k, \mathcal{F}_k)_{k\in\mathbb{N}}$ a Markov chain ?

Exercise 7.3. Let $(X_k)_{k\in\mathbb{N}}$ be a homogeneous Markov chain valued in $\mathbb{R}$ with Markov kernel P. Define F on $\mathbb{R}^2$ by
$$F(x, x') = P(x, (-\infty, x'])\ .$$

1. Write $F(x, x')$ using $\mathbb{P}_{P,x}$ and X_1.
2. Show that, for all $(x, x') \in \mathbb{R}^2$, $F(x, x') = \inf_{q\in\mathbb{Q}} \left(F(x, q)\mathbb{1}_{\{x'\leq q\}} + \mathbb{1}_{\{x'>q\}}\right)$.
3. Deduce that F is a Borel function.

Exercise 7.4. Soit $(Z_n)_{n\geq 0}$ une suite i.i.d. de variables aléatoires à valeurs dans $\mathbb{N}$, de loi μ. On considère une suite de variables aléatoires $(X_n)_{n\geq 0}$ à valeurs dans $\mathbb{N}$ definie pour $n \geq 1$ par :
$$X_{n+1} = \begin{cases} X_n - 1 & \text{si } X_n \geq 1 \\ Z_n + 1 & \text{si } X_n = 0\ , \end{cases}$$
avec X_0 indépendant de $(Z_n)_{n\geq 0}$. Montrer que $(X_n)_{n\geq 0}$ est une chaîne de Markov homogène sur $\mathbb{N}$ de matrice de transition à déterminer.

Exercise 7.5. Let f and g be two measurable functions defied on $\mathbb{R}$ and respectively valued in $\mathbb{R}$ and $\mathbb{R}_+^*$. Let $(Z_n)_{n\geq 0}$ be an i.i.d. sequence of real valued random variables with probability distribution μ. Let us define for all $n \geq 1$,
$$X_n = f(X_{n-1}) + g(X_{n-1})Z_n\ ,$$
given some real valued random variable X_0 independent of $(Z_n)_{n\geq 0}$.

1. Show that $(X_n)_{n\geq 0}$ is a Markov chain on $\mathbb{R}$, and determine its transition kernel Q.
2. Show that if μ admits a density with respect to the Lebesgue measure, then Q admits a transition density function and determine this density function.

Chapter 8

Shifting Markov chains

8.1 Invariant Measures and Stationarity

Definition 8.1.1 (Invariant measure)**.** *Let P be a Markov kernel on $(\mathsf{X}, \mathcal{X})$. A non zero σ-finite positive measure $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ is said to be invariant with respect to P (or P-invariant) if $\mu = \mu P$.*

In general, there may exist more than one invariant measure, or none if X is not finite. As a trivial example, consider $\mathsf{X} = \mathbb{N}$ and $P(x, \{x+1\}) = 1$.

If an invariant measure is finite, it may be normalized to an invariant probability measure. The fundamental role of an invariant probability measure is illustrated by the following result.

Theorem 8.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let P be a Markov kernel on a measurable space $(\mathsf{X}, \mathcal{X})$. A Markov chain $(X_k)_{k\in\mathbb{N}}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with kernel P is a stationary process if and only if its initial distribution is invariant with respect to P.*

Proof. If the chain $(X_k)_{k\in\mathbb{Z}}$ is stationary, then the marginal distribution is constant. In particular, the distribution of X_1 is equal to the distribution of X_0, which precisely means that $\pi P = \pi$. Thus π is invariant. Conversely, if $\pi P = \pi$, then $\pi P^h = \pi$ for all $h \geq 1$. Then, for all integers h and n, by Proposition 7.3.1, the distribution of $(X_h, \ldots, X_{h+n})$ is $\pi P^h \otimes P^{\otimes n}$. Since $\pi P^h = \pi$, it does not depend on h. □

Definition 8.1.2 (Absorbing set)**.** *A non-empty set $B \in \mathcal{X}$ is called absorbing if $P(x, B) = 1$ for all $x \in B$.*

Hence once a Markov chain falls in the absorbing set B, it remains in this set. The following result sheds some light on the set of invariant measures.

Theorem 8.1.2. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. Then,*

(i) The set of invariant probability measures for P is a convex subset of the convex cone $\mathbb{M}_+(\mathsf{X}, \mathcal{X})$.

(ii) Let π be an invariant probability and $\mathsf{X}_1 \subset \mathsf{X}$ with $\pi(\mathsf{X}_1) = 1$. There exists $B \subset \mathsf{X}_1$ such that $\pi(B) = 1$ and $P(x, B) = 1$ for all $x \in B$ (i.e. B is absorbing for P).

Proof. (i) P is a linear operator on $\mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Therefore, if π, π' are two invariant probability measures for P, then for every scalar $a \in [0, 1]$, using first the linearity and then the invariance,

$$(a\pi + (1-a)\pi')P = a\pi P + (1-a)\pi' P = a\pi + (1-a)\pi' .$$

(ii) The invariance of π implies that

$$\pi(\mathsf{X}_1) = 1 = \int_{\mathsf{X}_1} P(x,\ \mathsf{X}_1)\pi(\mathrm{d}x) .$$

Therefore, there exists a set $\mathsf{X}_2 \in \mathcal{X}$ such that,

$$\mathsf{X}_2 \subset \mathsf{X}_1 ,\ \pi(\mathsf{X}_2) = 1 \quad \text{and} \quad P(x,\ \mathsf{X}_1) = 1 ,\ \text{for all } x \in \mathsf{X}_2.$$

Repeating the above argument, we obtain a non-increasing sequence $\{\mathsf{X}_i, i \geq 1\}$ of sets $\mathsf{X}_i \in \mathcal{X}$ such that $\pi(\mathsf{X}_i) = 1$ for all $i = 1, 2, \ldots$, and $P(x,\ \mathsf{X}_i) = 1$, for all $x \in \mathsf{X}_{i+1}$. Define $B \stackrel{\text{def}}{=} \bigcap_{i=1}^{\infty} \mathsf{X}_i \in \mathcal{X}$. The set B is nonempty because

$$\pi(B) = \pi\left(\bigcap_{i=1}^{\infty} \mathsf{X}_i\right) = \lim_{i\to\infty} \pi(\mathsf{X}_i) = 1 .$$

The set B is absorbing for P because for any $x \in B$,

$$P(x,\ B) = P\left(x,\ \bigcap_{i=1}^{\infty} \mathsf{X}_i\right) = \lim_{i\to\infty} P(x,\ \mathsf{X}_i) = 1.$$

□

Example 8.1.1. *Take* $\mathsf{X} = \mathbb{R}$ *and* $P(x, \cdot) = \delta_{x+1}$ *if* $x \leq 0$ *and* $P(x, \cdot)$ *is a unit intensity exponential distribution otherwise. Then* $\mathbb{R}_+$ *is an absorbing set.*

8.2 The Shift operator and the Markov property

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $(\xi_k)_{k\in\mathbb{N}}$ be the canonical process on $\mathsf{X}^{\mathbb{N}}$, as defined in Definition 4.2.6 with $T = \mathbb{N}$,

$$\xi_k : w = (w_0, w_1, w_2, \ldots) \mapsto \xi_k(w) = w_k .$$

Recall that in Definition 4.3.1, in the case where the index set T is $\mathbb{N}$, we introduced the shift operator $\mathrm{S} : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$, defined by

$$\mathrm{S} : w = (w_0, w_1, w_2, \ldots) \mapsto \mathrm{S}(w) = (w_1, w_2, \ldots) .$$

Also recall the definition of the product σ-field $\mathcal{X}^{\otimes\mathbb{N}}$ in (4.2) as the smallest σ-field containing cylinders. It can also be viewed as the smallest σ-field generated by the canonical process, $\mathcal{X}^{\otimes\mathbb{N}} = \sigma(\xi_k, k \in \mathbb{N})$. Since the reciprocal image of a cylinder by S is a cylinder, we get that S is measurable from $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ to itself.

As usual S^0 denotes the identity operator and for $k \geq 1$, S^k the k-th composition iterate of S. Then, for $(j,k) \in \mathbb{N}^2$, it holds that

$$\xi_k \circ \mathrm{S}^j = \xi_{j+k} \ .$$

We will consider moreover a Markov kernel P on $\mathsf{X} \times \mathcal{X}$ and use the notation $\mathbb{P}_{P,\nu}$ for the corresponding probability measure on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ with initial distribution $\nu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ and transition kernel P (see Theorem 7.4.1 and Definition 7.4.1). We will also use the notation $\mathbb{K}_P$ introduced in Definition 7.4.2.

We have the following lemma which says that S^k can be seen as a random variable defined on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_{P,\nu})$ and valued in $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ and as such, it defines a Markov chain with kernel P and initial distribution νP^k.

Lemma 8.2.1. *For all $k \geq 0$, S^k is measurable as a function from $(\mathsf{X}^{\mathbb{N}}, \sigma(\xi_j, j \geq k))$ to $(\mathsf{X}^{\mathbb{N}}, \mathcal{F}_\infty)$. Moreover, for all initial distribution $\nu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ and Markov kernel P on $\mathsf{X} \times \mathcal{X}$, we have $\mathbb{P}^{\mathrm{S}^k}_{P,\nu} = \mathbb{P}_{P,\nu P^k}$, or equivalently, $(\nu \mathbb{K}_P)^{\mathrm{S}^k} = \nu P^k \mathbb{K}_P$.*

Proof. Let $k \geq 0$. Consider a cylinder $B = A_0 \times \cdots \times A_n \times \mathsf{X}^{\{n+1, n+2, \dots\}}$ with $n \geq 1$ and $A_i \in \mathcal{X}$ for $i = 0, \dots, n$. Then we have

$$\begin{aligned}
(\mathrm{S}^k)^{-1}(B) &= \{\omega = (\omega_t)_{t\in\mathbb{N}} \ : \ \omega_{k+i} \in A_i,\, i = 0, \dots, n\} \\
&= \mathsf{X}^{\{0,\dots,k-1\}} \times \left(\prod_{i=k}^{k+n} A_i \right) \times \mathsf{X}^{\{n+1,n+2,\dots\}} \\
&= \xi_k^{-1}(A_0) \cap \cdots \cap \xi_{k+n}^{-1}(A_n) \\
&\in \sigma(\xi_j, j \geq k) \ .
\end{aligned}$$

Hence the first assertion. Now, we moreover have that

$$\begin{aligned}
\nu\mathbb{K}_P \circ (\mathrm{S}^k)^{-1}(B) = \nu \otimes P^{\otimes(k+n)} \left(\mathsf{X}^{\{0,\dots,k-1\}} \times \left(\prod_{i=k}^{k+n} A_i \right) \right) \\
= (\nu P^k) \otimes P^{\otimes n} \left(\prod_{i=0}^{n} A_i \right) = (\nu P^k)\mathbb{K}_P(B) \ ,
\end{aligned}$$

which proves the second assertion. □

We can use this to rewrite the Markov property in terms of the conditional distribution of the shifted trajectories.

Proposition 8.2.2 (Homogeneous Markov property)**.** *Consider the canonical filtered space $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}^\xi_k\right)_{k\in\mathbb{N}}\right)$. For all $\nu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ and transition kernel P on $\mathsf{X} \times \mathcal{X}$, we have*

$$\mathbb{P}^{(\xi_k, \mathrm{S}^k)}_{P,\nu} = (\nu P^k) \otimes \mathbb{K}_P \ . \tag{8.1}$$

Moreover, the conditional distribution of S^k given $\mathcal{F}^\xi_k$ under $\mathbb{P}_{P,\nu}$ admits $(\omega, A) \mapsto \mathbb{P}_{P,\xi_k(\omega)}(A) = \mathbb{K}_P(\xi_k(\omega), A)$ as a regular version. That is, we have $\mathbb{P}_{P,\nu}$-a.s., for all $A \in \mathcal{X}^{\otimes\mathbb{N}}$, $\mathbb{P}^{\mathrm{S}^k \,|\mathcal{F}^\xi_k}_{P,\nu}(A) = \mathbb{P}_{P,\xi_k}(A) = \mathbb{K}_P \mathbb{1}_A(\xi_k)$.

Proof. Note that $\xi_k = \xi_0 \circ \mathrm{S}^k$. Hence the distribution of (ξ_k, S^k) under $\mathbb{P}_{P,\nu}$ is the same as that of (ξ_0, ξ) under $\mathbb{P}^{\mathrm{S}^k}_{P,\nu}$, which is equal to $\mathbb{P}_{P,\nu P^k}$ by Lemma 8.2.1. In other words, we have

$$\mathbb{P}^{(\xi_k,\mathrm{S}^k)}_{P,\nu} = \mathbb{P}^{(\xi_0,\xi)}_{P,\nu P^k} \ .$$

Hence (8.1) follow from (7.22).

Let now $A \in \mathcal{X}^{\otimes\mathbb{N}}$. Under $\mathbb{P}_{P,\nu}$, ξ is homogeneous Markov chain, hence by Proposition 7.1.1(ii) and using that S^k is $\sigma(\xi_l,\, l \geq k)$-measurable (see Lemma 8.2.1), we have

$$\mathbb{E}_{P,\nu}\left[\mathbb{1}_A(\mathrm{S}^k)\middle|\, \mathcal{F}_k\right] = \mathbb{E}_{P,\nu}\left[\mathbb{1}_A(\mathrm{S}^k)\middle|\, \xi_k\right] = \mathbb{K}_P \mathbb{1}_A(\xi_k) \ .$$

where the second inequality follows from (8.1) and Theorem 2.1.10. The proof is concluded. □

An immediate consequence of Proposition 8.2.2 is that, for all measurable real-valued mapping Y on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ we have $Y \circ \mathrm{S}^k \in L^1(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_{P,\nu})$ if and only if $\mathbb{K}_P Y \in L^1(\mathsf{X}, \mathcal{X}, \nu P^k)$ and if this assertion holds, we have

$$\mathbb{E}_{P,\nu}\left[Y \circ \mathrm{S}^k\middle|\, \mathcal{F}_k\right] = \mathbb{E}_{P,\xi_k}[Y] = \mathbb{K}_P Y(\xi_k) \qquad \mathbb{P}_{P,\nu}\text{-a.s.} \tag{8.2}$$

We illustrate the homogeneous Markov property through a simple example.

Example 8.2.1. *Assume that there exists a set $C \in \mathcal{X}$ such that, for all $x \notin C$, $\mathbb{P}_{P,x}(\sigma_C < \infty) = 1$, where σ_C denotes the positive hitting time of the canonical chain to C. Then, for all $x \in C$, we have*

$$\begin{aligned}
\mathbb{P}_{P,x}(\sigma_C < \infty) &= \mathbb{P}_{P,x}(\xi_1 \in C) + \mathbb{P}_{P,x}(\xi_1 \notin C,\, \sigma_C \circ \mathrm{S}^1 < \infty) \\
&= \mathbb{P}_{P,x}(\xi_1 \in C) + \mathbb{E}_{P,x}[\mathbb{1}_{\{\xi_1 \notin C\}} \mathbb{P}_{P,\xi_1}(\sigma_C < \infty)] \\
&= \mathbb{P}_{P,x}(\xi_1 \in C) + \mathbb{P}_{P,x}(\xi_1 \notin C) = 1 \ .
\end{aligned}$$

Therefore, for all $x \in \mathsf{X}$, $\mathbb{P}_{P,x}(\sigma_C < \infty) = 1$.

Let τ be a stopping time defined on the canonical filtered space $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}^{\xi}_k\right)_{k\in\mathbb{N}}\right)$. Define S^τ and ξ_τ for all $\omega = (\omega)_{n\in\mathbb{N}} \in \{\tau < \infty\}$ by

$$\mathrm{S}^\tau(w) = \mathrm{S}^{\tau(w)}(w) = \left(\omega_{\tau(\omega)+n}\right)_{n\in\mathbb{N}} \quad \text{and} \quad \xi_\tau(\omega) = \xi_{\tau(\omega)}(\omega) = \omega_{\tau(\omega)} \ . \tag{8.3}$$

With this definition, we have $\xi_\tau = \xi_k$ on $\{\tau = k\}$ and $\xi_k \circ \mathrm{S}^\tau = \xi_{\tau+k}$ on $\{\tau < \infty\}$.

These variables are only defined on $\{\tau < \infty\}$ and we will consider the conditional expectations involving such variable. To this end we first explain why this actually makes sense. Consider a L^1 r.v. X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-field $\mathcal{G} \subset \mathcal{F}$. Let moreover $A \in \mathcal{G}$. We say that

$$\mathbb{E}[X|\,\mathcal{G}] = Y \quad \text{on } A \quad \mathbb{P}\text{-a.s.}$$

if $\mathbb{E}[X|\,\mathcal{G}](\omega) = Y(\omega)$ for $\mathbb{P}$-a.e. $\omega \in A$, that is, if $\mathbb{P}(\{\mathbb{E}[X|\,\mathcal{G}] \neq Y\} \cap A) = 0$. It is easy to show that this is true if and only if $Y\mathbb{1}_A$ is $\mathcal{G}$-measurable and, for all $B \in \mathcal{G}$,

$$\mathbb{E}[X\mathbb{1}_{B\cap A}] = \mathbb{E}[Y\mathbb{1}_{B\cap A}] \ .$$

We thus see that to prove this kind of result, we only need to define X and Y on A.

Proposition 8.2.3 (Strong Markov property)**.** *Consider a stopping time τ defined on the canonical filtered space $\left(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \left(\mathcal{F}_k^{\xi}\right)_{k\in\mathbb{N}}\right)$. Define S^{τ} and ξ_{τ} on $\{\tau<\infty\}$ as above. For all $\nu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ and transition kernel P on $\mathsf{X}\times\mathcal{X}$, the conditional distribution of S^{τ} given $\mathcal{F}_{\tau}^{\xi}$ under $\mathbb{P}_{P,\nu}$ admits $(\omega, A) \mapsto \mathbb{P}_{P,\xi_{\tau}(\omega)}(A) = \mathbb{K}_P(\xi_{\tau}(\omega), A)$ as a regular version on $\omega \in \{\tau<\infty\}$.*

Proof. Note that on $\{\tau<\infty\}$, we have

$$\xi_{\tau} = \sum_{k\geq 0} \xi_k \mathbb{1}_{\{\tau=k\}} ,$$

which is $\mathcal{F}_{\tau}^{\xi}$-measurable. We thus only need to show that for all $B \in \mathcal{F}_{\tau}^{\xi}$ and $A \in \mathcal{X}^{\otimes\mathbb{N}}$, we have

$$\mathbb{E}_{P,\nu}[\mathbb{1}_A(\mathrm{S}^{\tau})\mathbb{1}_{B\cap\{\tau<\infty\}}] = \mathbb{E}_{P,\nu}[\mathbb{K}_P(\xi_{\tau}, A)\mathbb{1}_{B\cap\{\tau<\infty\}}] . \tag{8.4}$$

Now, let us observe that

$$\mathbb{E}_{P,\nu}[\mathbb{1}_A(\mathrm{S}^{\tau})\mathbb{1}_{B\cap\{\tau<\infty\}}] = \sum_{k=0}^{\infty} \mathbb{E}_{P,\nu}[\mathbb{1}_A(\mathrm{S}^{k})\mathbb{1}_{B\cap\{\tau=k\}}] .$$

By Proposition 8.2.2, since $B\cap\{\tau=k\} \in \mathcal{F}_k$, we have for all $k\geq 0$,

$$\mathbb{E}_{P,\nu}[\mathbb{1}_A(\mathrm{S}^{k})\mathbb{1}_{B\cap\{\tau=k\}}] = \mathbb{E}_{P,\nu}\left[\mathbb{E}_{P,\nu}\left[\mathbb{1}_A(\mathrm{S}^{k})\middle|\mathcal{F}_k\right]\mathbb{1}_{B\cap\{\tau=k\}}\right] = \mathbb{E}_{P,\nu}\left[\mathbb{K}_P(\xi_k, A)\mathbb{1}_{B\cap\{\tau=k\}}\right] .$$

Inserting this in the previous sum we obtain (8.4). □

The following result is useful for defining successive positive hitting times.

Proposition 8.2.4. *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be the natural filtration of the canonical process $(\xi_n)_{n\in\mathbb{N}}$ on $\mathsf{X}^{\mathbb{N}}$. Let τ and σ be two stopping times with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

(i) For all integers $n, m \in \mathbb{N}^2$, $(\mathrm{S}^n)^{-1}(\mathcal{F}_m) = \sigma(\xi_n, \ \ldots, \ \xi_{n+m})$.

(ii) For every positive integer k, $k+\tau\circ\mathrm{S}^k$ is a stopping time.

(iii) $\rho = \sigma + \tau\circ\mathrm{S}^{\sigma}$ is a stopping time. On $\{\sigma<\infty, \tau<\infty\}$, we have $\xi_{\tau}\circ\mathrm{S}^{\sigma} = \xi_{\rho}$.

Proof.

(i) For all $A\in\mathcal{X}$, and all integers $k, n \in \mathbb{N}^2$,

$$(\mathrm{S}^n)^{-1}\{\xi_k \in A\} = \{\xi_k\circ\mathrm{S}^n \in A\} = \{\xi_{k+n} \in A\} .$$

Since the σ-field $\mathcal{F}_m$ is generated by the events of the form

$$B = \{\xi_0 \in A_0\} \cap \{\xi_1 \in A_1\} \cap \cdots \cap \{\xi_m \in A_m\} ,$$

the σ-field $(\mathrm{S}^n)^{-1}(\mathcal{F}_m)$ is generated by the events

$$\begin{aligned}(\mathrm{S}^n)^{-1}(B) &= (\mathrm{S}^n)^{-1}\{\xi_0 \in A_0\} \cap (\mathrm{S}^n)^{-1}\{\xi_1 \in A_1\} \cap \cdots \cap (\mathrm{S}^n)^{-1}\{\xi_m \in A_m\} \\ &= \{\xi_n \in A_0\} \cap \{\xi_{n+1} \in A_1\} \cap \cdots \cap \{\xi_{n+m} \in A_m\} .\end{aligned}$$

These events generate the σ-field $\sigma(\xi_n, \ \ldots, \ \xi_{n+m}) \subset \mathcal{F}_{n+m}$.

(ii) Since τ is a stopping time, $\{\tau = m-k\} \in \mathcal{F}_{m-k}$ and by (i), it also holds that $(\mathrm{S}^k)^{-1}\{\tau = m-k\} \in \mathcal{F}_m$. Thus,

$$\{k + \tau \circ \mathrm{S}^k = m\} = \{\tau \circ \mathrm{S}^k = m-k\} = (\mathrm{S}^k)^{-1}\{\tau = m-k\} \in \mathcal{F}_m \ .$$

This implies that $k + \tau \circ \mathrm{S}^k$ is a stopping time.

(iii) From the definition of ρ, we obtain

$$\begin{aligned}\{\rho = m\} = \{\sigma + \tau \circ \mathrm{S}^\sigma = m\} &= \bigcup_{k=0}^{m} \{k + \tau \circ \mathrm{S}^k = m,\ \sigma = k\} \\ &= \bigcup_{k=0}^{m} \{k + \tau \circ \mathrm{S}^k = m\} \cap \{\sigma = k\}.\end{aligned}$$

Since σ is a stopping time and since $k + \tau \circ \mathrm{S}^k$ is a stopping time for each k (see (ii)), we obtain that $\{\rho = m\} \in \mathcal{F}_m$. Thus ρ is a stopping time. By construction, if $\tau(\omega)$ and $\sigma(\omega)$ are finite, we have

$$\xi_\tau \circ \mathrm{S}^\sigma(\omega) = \xi_{\tau \circ \mathrm{S}^{\sigma(\omega)}}(\mathrm{S}^\sigma(\omega)) = \xi_{\sigma + \tau \circ \mathrm{S}^\sigma}(\omega) \ .$$

□

Lemma 8.2.5. *The successive hitting and positive hitting times to a measurable set A are stopping times with respect to the canonical filtration. In addition, $\sigma_A = 1 + \tau_A \circ \mathrm{S}^1$ and for $n \geq 0$, $\sigma_A^{(n+1)} = \sigma_A^{(n)} + \sigma_A \circ \mathrm{S}^{\sigma_A^{(n)}}$ on $\{\sigma_A^{(n)} < \infty\}$.*

Proof. The proof is a straightforward application of Proposition 8.2.4 (iii). □

8.3 Construction of invariant measure using recurrent states

In this section we explain how to construct an invariant measure ν for a Markov kernel P, based on the existence of a *recurrence state*, which we now introduce.

Definition 8.3.1 ((Positive) recurrent state)**.** *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. We say that $x \in \mathsf{X}$ is a recurrent state of P if the positive hitting time to $\{x\}$, denoted by σ_x, satisfies $\mathbb{P}_{P,x}(\sigma_x < \infty) = 1$. A recurrent state is called* positive *recurrent if moreover $\mathbb{E}_{P,x}[\sigma_x] < \infty$.*

Recurrent state are useful in the following construction.

Definition 8.3.2 (Occupation measure)**.** *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $x \in \mathsf{X}$, we define the* occupation measure $\nu_{P,x}$ *by*

$$\nu_{P,x}(A) = \mathbb{E}_{P,x}\left[\sum_{k=1}^{\sigma_x} \mathbb{1}_A(\xi_k)\right] \ , \qquad A \in \mathcal{X} \ ,$$

where σ_x denotes the positive hitting time to $\{x\}$.

It is easy to check that ν is indeed σ-additive on $\mathcal{X}$, hence is a measure, and we have for all $g \in \mathrm{F}_+(\mathsf{X}, \mathcal{X})$,

$$\int g \, \nu_{P,x}(\mathrm{d}x) = \mathbb{E}_{P,x}\left[\sum_{k=1}^{\sigma_x} g(\xi_k)\right] . \tag{8.5}$$

Moreover we immediately have

$$\nu_{P,x}(\mathsf{X}) = \mathbb{E}_{P,x}\left[\sigma_x\right] , \tag{8.6}$$

hence $\nu_{P,x}$ is not a probability measure and it is finite if and only if x is a positive recurrent state.

We have the following result.

Theorem 8.3.1. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $x \in \mathsf{X}$ If x is a recurrent state for P, then the occupation measure $\nu_{P,x}$ is invariant for P, $\nu_{P,x}P = \nu_{P,x}$.*

Proof. Let $A \in \mathcal{X}$. We have, since σ_x is a stopping time

$$\begin{aligned}
\nu_{P,x}(A) &= \mathbb{E}_{P,x}\left[\sum_{k=1}^{\infty} \mathbb{1}_A(\xi_k)\mathbb{1}_{\{k\le\sigma_x\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}_{P,x}\left[\mathbb{1}_A(\xi_k)\mathbb{1}_{\{\sigma_x\le k-1\}^c}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}_{P,x}\left[\mathbb{E}_{P,x}\left[\mathbb{1}_A(\xi_k)|\,\mathcal{F}_{k-1}^{\xi}\right]\mathbb{1}_{\{k\le\sigma_x\}}\right] .
\end{aligned}$$

By Proposition 8.2.2, since $\xi_k = \xi_1 \circ \mathrm{S}^{k-1}$, we have

$$\mathbb{E}_x\left[\mathbb{1}_A(\xi_k)|\,\mathcal{F}_{k-1}\right] = \mathbb{K}_P(\mathbb{1}_A \circ \xi_1)(\xi_{k-1}) \quad \mathbb{P}_{P,x}\text{-a.s.}$$

And since $\mathbb{K}_P(\mathbb{1}_A \circ \xi_1) = P\mathbb{1}_A$, we obtain that

$$\nu_{P,x}(A) = \sum_{k=1}^{\infty} \mathbb{E}_{P,x}\left[P(\xi_{k-1}, A)\mathbb{1}_{\{k\le\sigma_x\}}\right] = \mathbb{E}_{P,x}\left[\sum_{k=1}^{\sigma_x} P(\xi_{k-1}, A)\right] = \mathbb{E}_{P,x}\left[\sum_{j=0}^{\sigma_x-1} P(\xi_j, A)\right] .$$

Now observe that on $\{\xi_0 = x,\, \sigma_x < \infty\}$, for any nonnegative function g defined on X, we have

$$\sum_{j=0}^{\sigma_x-1} g(\xi_j) = g(x) + \sum_{j=1}^{\sigma_x-1} g(\xi_j) = \sum_{j=1}^{\sigma_x} g(\xi_j) .$$

Hence, we finally get, since $\mathbb{P}_{P,x}(\xi_0 = x,\, \sigma_x < \infty) = 1$,

$$\nu_{P,x}(A) = \mathbb{E}_{P,x}\left[\sum_{j=1}^{\sigma_x} P(\xi_j, A)\right] = \nu_{P,x}P(A) ,$$

where we applied (8.5). Since this is true for all $A \in \mathcal{X}$, it concludes the proof. □

We immediately have the following consequence from Theorem 8.3.1 and (8.6).

Corollary 8.3.2. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $x \in \mathsf{X}$ If x is a positive recurrent state for P, then the normalized occupation measure $\left(\mathbb{E}_{P,x}\left[\sigma_x\right]\right)^{-1} \nu_{P,x} \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ is an invariant initial distribution for P.*

Unfortunately assuming the existence of a (positive) recurrent state excludes many interesting cases. Indeed observe that if $P(x, \cdot)$ is diffuse then $\sigma_x = \infty$ $[\mathbb{P}_{P,y}]\mathbb{P}$-a.s. for all y! So, for instance, any Markov kernel for which $P(x, \cdot)$ admits a density with respect to the Lebesgue measure is excluded from such an assumption. Different constructions are of course available in this interesting case but they are not investigated here.

8.4 The finite state-space case

In this section, we always assume that the state-space X is a finite or countable set and $\mathcal{X}$ is the class of all subsets of X. By Corollary 8.3.2, the occupation measure of Definition 8.3.2 then provides an invariant distribution if we can find a positive recurrent state. It is thus of interest to classify states in recurrent ones and non-recurrent ones (they are then called *transiant* states). The first step is to decompose the space in subsets where all states "communicate".

Definition 8.4.1. *Let $x, y \in \mathsf{X}$. We say that x communicates with y and write $x \leftrightarrow y$ ifthere exist n and n' in $\mathbb{N}$ such that*

$$P^n(x, \{y\}) > 0 \quad \textit{and} \quad P^{n'}(y, \{x\}) > 0 \, .$$

It is straightforward to check that $\leftrightarrow$ defines an equivalence relationship on X.

Here, we will limit ourselves in the simplest possible case.

Definition 8.4.2. *We say that P is irreducible if $\leftrightarrow$ has a unique equivalent class.*

In this case, we have a direct way to construct an invariant probability measure for the kernel P.

Proposition 8.4.1. *If P is irreducible and X is finite, then for all $x, y \in \mathsf{X}$, we have $\mathbb{E}_{P,x}[\sigma_y] < \infty$.*

Proof. Because P is irreducible, we have $\left\{k \geq 1 \, : \, P^k(x, \{y\}) > 0\right\} \neq \emptyset$ for all $x, y \in \mathsf{X}$, and we set

$$n(x, y) = \min \left\{k \geq 1 \, : \, P^k(x, \{y\}) > 0\right\} \, .$$

Since X is finite,

$$n = \max_{x,y \in \mathsf{X}} n(x, y)$$

is well defined and finite and also

$$\alpha = \inf_{x,y \in \mathsf{X}} P^{n(x,y)}(x, \{y\}) > 0 \, .$$

We write, for all $x, y \in \mathsf{X}$ and $p \geq 1$,

$$\begin{aligned}
\mathbb{P}_{P,x}\left(\sigma_y > n\, p\right) &= \mathbb{P}_{P,x}\left(\sigma_y > n\,(p-1),\ \xi_k \neq y \text{ for } n\,(p-1) < k \leq n\, p\right) \\
&= \mathbb{E}_{P,x}\left[\mathbb{1}_{\{\sigma_y > n\,(p-1)\}} \, \mathbb{P}_{P,x}\left(\xi_k \neq y \text{ for } n\,(p-1) < k \leq n\, p \,|\, \mathcal{F}_{n\,(p-1)}\right)\right] \\
&= \mathbb{E}_{P,x}\left[\mathbb{1}_{\{\sigma_y > n\,(p-1)\}} \, \mathbb{P}_{P,\xi_{n\,(p-1)}}\left(\xi_j \neq y \text{ for } 0 < j \leq n\right)\right] \, .
\end{aligned}$$

Now observe that, by definition of n and α, for all $x, y \in \mathsf{X}$,

$$\mathbb{P}_{P,x}\left(\xi_j \neq y \text{ for } 0 < j \leq n\right) \leq \mathbb{P}_{P,x}\left(\xi_{n(x,y)} \neq y\right) \leq 1 - \alpha \, .$$

Hence we obtain that, for all $x, y \in \mathsf{X}$ and $p \geq 1$,

$$\mathbb{P}_{P,x}\left(\sigma_y > n\, p\right) \leq (1-\alpha)\mathbb{P}_{P,x}\left(\sigma_y > n\,(p-1)\right) \leq (1-\alpha)^p \, .$$

From which we conclude that $\mathbb{P}_{P,x}\left(\sigma_y < \infty\right) = 1$. Similarly, we obtain

$$\mathbb{E}_{P,x}[\sigma_y] \leq \mathbb{E}_{P,x}\left[n\left(1 + \sum_{p=1}^{\infty} \mathbb{1}_{\{\sigma_y > np\}}\right)\right] \leq n\left(1 + \sum_{p=1}^{\infty}(1-\alpha)^p\right) < \infty \, .$$

□

We conclude from Corollary 8.3.2 that an irreducible Markov chain on a finite state space always admits an invariant probability measure. It can be shown moreover that there is only one invariant probability measure in this case.

8.5 Exercises

Exercise 8.1. Let P be a Markov kernel on a general state space $(E, \mathcal{F})$. Assume that there exist a probability measure μ on $(E, \mathcal{F})$ and a mapping $\lambda : \mathcal{F} \to [0, 1]$ such that,

$$\lim_{n\to\infty} \sup_{A\in\mathcal{F}} |\mu P^n(A) - \lambda(A)| = 0 \; . \tag{8.7}$$

1. Show that λ is a probability measure on $(E, \mathcal{F})$.

We can show that (8.7) holds if and only if

$$\lim_{n\to\infty} \sup_{|g|\le 1} \left| \int_E g(x)\mu P^n(\mathrm{d}x) - \int_E g(y)\lambda(\mathrm{d}y) \right| = 0 \; . \tag{8.8}$$

2. Show that, if (8.8) holds, then λ is invariant for P.

3. Assume now that all probability measure μ on $(E, \mathcal{F})$ satisfies (8.7), with the same λ. Show that P admits a unique invariant probability measure.

Exercise 8.2. Let $d \in \mathbb{N}^*$. Let P be a Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Assume that there exists a probability measure μ on $\mathbb{R}^d$, such that the sequence of probability measures $(\mu P^n)_{n\ge 0}$ converges weakly to some probability measure λ_μ. Assume moreover that for all f continuous and bounded, then Pf is also continuous and bounded.

1. Show that λ_μ is invariant for P.

2. What can be said if P admits a unique invariant probability measure.

Assume now that for all probability measure μ on $\mathbb{R}^d$, $(\mu P^n)_{n\ge 0}$ weakly converges to the same probability measure λ.

3. Show that P admits a unique invariant probability measure.

Exercise 8.3. Let $(\xi_n)_{n\in\mathbb{N}}$ denote the canonical process valued in a countable state space E. Consider the sequence of successive return time to $y \in E$ defined by $\sigma_y^{(0)} = \inf\{n > 0 \; : \; \xi_n = y\}$ and for $p \ge 1$,

$$\sigma_y^{(p)} = \begin{cases} \sigma_y^{(p-1)} + \sigma_y^{(0)} \circ \mathrm{S}^{\sigma_y^{(p-1)}} & \text{if } \sigma_y^{(p-1)} < \infty \; , \\ \infty & \text{otherwise} \; , \end{cases}$$

where S is the shift operator. Also, denote by $N_y = \sum_{n\ge 0} \mathbb{1}_y(\xi_n)$.

1. Show that we can write, for all $p \ge 1$,

$$\sigma_y^{(p)} = \inf\{n > \sigma_y^{(p-1)} \; : \; \xi_n = y\} \; .$$

2. Show that for all $y \in E$, $\left(\mathcal{F}_{\sigma_y^{(p)}}\right)_{p\ge 0}$ is a filtration.

Let P be a Markov kernel/transition matrix on E.

3. Show that, for all y and all probability ν on E, under $\mathbb{P}_{P,\nu}$, the sequence of random variables $(\sigma_y^{(p)})_{p\in\mathbb{N}}$ is a homogeneous Markov chain with respect to $\left(\mathcal{F}_{\sigma_y^{(p)}}\right)_{p\ge 0}$, valued in $\overline{\mathbb{N}^*} = \mathbb{N}^* \cup \{\infty\}$.

4. Determine the corresponding transition matrix using the distribution μ_y of $\sigma_y^{(0)}$.

We now assume that $y \in E$ with $\mathbb{P}_{P,y}(\sigma_y^{(0)} < \infty) = 1$.

5. Show that for all $p \geq 1$, $\mathbb{P}_{P,y}(\sigma_y^{(p)} < \infty) = 1$. Deduce from this result that $\mathbb{P}_{P,y}(N_y = \infty) = 1$.

6. Show that under $\mathbb{P}_{P,y}$, the sequence of random variables $(\sigma_y^{(p+1)} - \sigma_y^{(p)})_{p\geq 0}$ is i.i.d. , with common distribution μ_y.

Exercise 8.4. We use the same definition and assumptions as in Exercise 8.3.

1. Show for all $x, y \in E$, $x \neq y$,

$$\mathbb{P}_{P,x}\left(N_y = m\right) = \begin{cases} \mathbb{P}_{P,x}\left(\sigma_y^{(0)} < \infty\right) \mathbb{P}_{P,y}\left(\sigma_y^{(0)} = \infty\right) \left(\mathbb{P}_{P,y}\left(\sigma_y^{(0)} < \infty\right)\right)^{m-1} & \text{if } m \geq 1 \\ \mathbb{P}_{P,x}\left(\sigma_y^{(0)} = \infty\right) & \text{otherwise} \end{cases} .$$

2. Show for all $x \in E$,

$$\mathbb{P}_{P,x}\left(N_x = m\right) = \begin{cases} \mathbb{P}_{P,x}\left(\sigma_x^{(0)} = \infty\right) \left(\mathbb{P}_{P,x}\left(\sigma_x^{(0)} < \infty\right)\right)^{m-1} & \text{if } m \geq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Exercise 8.5. Let $\xi = (\xi_n)_{n\in\mathbb{N}}$ denote the canonical process valued in $E = \{0, 1\}$ and the transition matrix on E, K given by

$$K = \begin{pmatrix} 1/4 & 3/4 \\ 1/2 & 1/2 \end{pmatrix}$$

1. Show that K admits a unique invariant probability measure μ.

Let F defined on $E^{\mathbb{N}}$ by

$$F(x) = x_0 + x_1 .$$

2. Compute for all $i \in E$, $\mathbb{E}_{K,i}[F(\xi)]$.

3. Deduce from this result that for all $n \geq 1$ and $i \in E$, $\mathbb{E}_{K,i}[F(\mathrm{S}^n) \mid \mathcal{F}_n^{\xi}]$.

Let $\sigma = \inf\{n \geq 1 \mid \xi_n = 1\}$.

4. Compute $\mathbb{P}_{K,i}(\sigma = n)$ pour $i = 0, 1$ and all $n \geq 1$.

5. Justify that for all $i \in E$, $\mathbb{P}_{K,i}(\sigma < \infty) = 1$.

6. Compute $\mathbb{E}_{K,i}[F(\mathrm{S}^{\sigma}) \mid \mathcal{F}_{\sigma}^{\xi}]$.

Exercise 8.6. Let $\xi = (\xi)_{n\in\mathbb{N}}$ denote the canonical process valued in $E = \{-1, 0, 1\}$ and let the transition matrix K be given by

$$K = \begin{pmatrix} 1/4 & 1/4 & 1/2 \\ 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

Let F defined on $E^{\mathbb{N}}$ by

$$F(x) = x_1 - x_2 .$$

1. Compute for all $i \in E$, $\mathbb{E}_{K,i}[F(\xi)]$
2. Deduce from this result that for all $n \geq 1$ and $i \in E$, $\mathbb{E}_{K,i}[F(\xi \circ \mathrm{S}^n) \mid \mathcal{F}_n^\xi]$.
3. Let $\sigma = \inf\{n \geq 1 \mid \xi_n = 1\}$. Compute for all $i \in E$, $\mathbb{E}_{K,i}[\mathbb{1}_{\{\sigma<\infty\}} F(\xi \circ \mathrm{S}^\sigma) \mid \mathcal{F}_\sigma^\xi]$.

Exercise 8.7. On considère ici l'espace d'état fini $E = \{1, 2, 3\}$, et la matrice de transition

$$K = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

Soit F la fonctionnelle définie sur $E^{\mathbb{N}}$ par

$$F(x) = \mathbb{1}_{\{\sum_{j=1}^{3} \mathbb{1}_{\{1\}}(x_j)=1\}} \cdot$$

1. Que représente F ?
2. Calculer pour tout $i \in E$, $\mathbb{E}_{K,i}[F(\xi)]$.
3. En déduire pour tout $n \geq 1$ et $i \in E$, $\mathbb{E}_{K,i}[F(\xi \circ \mathrm{S}^n) \mid \mathcal{F}_n^\xi]$.
4. Soit $\sigma = \inf\{n \geq 1 \mid \xi_n = 1\}$. Justifier que pour tout $i \in E$, $\mathbb{P}_{K,i}(\sigma < \infty) = 1$, et calculer $\mathbb{E}_{K,i}[F(\xi \circ \mathrm{S}^\sigma) \mid \mathcal{F}_\sigma^\xi]$.

Chapter 9

Complements: classical examples

9.1 Reversible Markov chains

Definition 9.1.1. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. A σ-finite measure ξ on $\mathcal{X}$ is said to be reversible with respect to P if for all $(A, B) \in \mathcal{X} \times \mathcal{X}$*

$$\xi \otimes P(A \times B) = \xi \otimes P(B \times A) \; . \tag{9.1}$$

Equivalently, revesibility means that for all bounded measurable functions f defined on $(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$,

$$\iint_{\mathsf{X}\times\mathsf{X}} \xi(\mathrm{d}x) P(x, \mathrm{d}x') f(x, x') = \iint_{\mathsf{X}\times\mathsf{X}} \xi(\mathrm{d}x) P(x, \mathrm{d}x') f(x', x) \; . \tag{9.2}$$

If X is a countyable state-space, a (finite or σ-finite) measure ξ is reversible with respect to to P if and only if, for all $(x, x') \in \mathsf{X} \times \mathsf{X}$,

$$\xi(\{x\}) P(x, \{x'\}) = \xi(\{x'\}) P(x', \{x\}) \; . \tag{9.3}$$

This condition is referred to as the *detailed balance condition*, often written as

$$\xi(x) P(x, x') = \xi(x') P(x', x) \; .$$

(in the countable case, measures on singletons are identified to their density functions).

If $(X_k)_{k\in\mathbb{N}}$ is a Markov chain with kernel P and initial distribution ξ, the reversibility condition (9.1) precisely means that (X_0, X_1) and (X_1, X_0) have the same distribution. This implies in particular that the distribution of X_1 is the same as that of X_0, and this means that ξ is P-invariant. Thus we see that reversibility implies invariance. This property can be extended to all finite dimensional distributions.

Proposition 9.1.1. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $\xi \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$. If ξ is reversible with respect to P, then*

(i) ξ is P-invariant

(ii) the process $(X_k)_{k\in\mathbb{N}}$ is reversible, i.e. for any $p \in \mathbb{N}$, $(X_0, \ldots, X_p)$ and $(X_p, \ldots, X_0)$ have the same distribution.

Proof. (i) Using (9.1) with $A = \mathsf{X}$ and $B \in \mathcal{X}$, we get

$$\xi P(B) = \xi \otimes P(\mathsf{X} \times B) = \xi \otimes P(B \times \mathsf{X}) = \int \xi(\mathrm{d}x) \mathbb{1}_B(x) P(x, \mathsf{X}) = \xi(B) \; .$$

(ii) We show by induction on k that for $(X_0, \ldots, X_k)$ and $(X_k, \ldots, X_0)$ have the same distribution under $\mathbb{P}_\xi$, that is for all $A_0, \ldots, A_k \in \mathcal{X}$,

$$\xi \otimes P^{\otimes k}(A_0 \times \cdots \times A_k) = \xi \otimes P^{\otimes k}(A_k \times \cdots \times A_k) \; .$$

The reversibility gives this property for $k = 1$. Let now $k \geq 2$. We can write, using manipulations ion tensor products and the induction hypothesis,

$$\begin{aligned}
\xi \otimes P^{\otimes k}(A_0 \times \cdots \times A_k) &= \xi \otimes P^{\otimes (k-1)}(\mathbb{1}_{A_0} \otimes \cdots \otimes \mathbb{1}_{A_{k-2}} \otimes (\mathbb{1}_{A_{k-1}} \times P\mathbb{1}_{A_k})) \\
&= \xi \otimes P^{\otimes (k-1)}((\mathbb{1}_{A_{k-1}} \times P\mathbb{1}_{A_k}) \otimes \mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \\
&= \xi \left(\mathbb{1}_{A_{k-1}} \times P\mathbb{1}_{A_k} \times P^{\otimes (k-1)}(\mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \\
&= \xi \left(\left(\mathbb{1}_{A_{k-1}} \times P^{\otimes (k-1)}(\mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \times P\mathbb{1}_{A_k} \right) \\
&= \xi \otimes P \left(\left(\mathbb{1}_{A_{k-1}} \times P^{\otimes (k-1)}(\mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \otimes \mathbb{1}_{A_k} \right) \\
&= \xi \otimes P \left(\mathbb{1}_{A_k} \otimes \left(\mathbb{1}_{A_{k-1}} \times P^{\otimes (k-1)}(\mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \right) \; ,
\end{aligned}$$

where in the last line, we use the induction hypothesis already proven for $k = 1$. Continuing from there, we get

$$\begin{aligned}
\xi \otimes P^{\otimes k}(A_0 \times \cdots \times A_k) &= \xi \left(\mathbb{1}_{A_k} \times P \left(\mathbb{1}_{A_{k-1}} \times P^{\otimes (k-1)}(\mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \right) \\
&= \xi \left(\mathbb{1}_{A_k} \times P^{\otimes k} \left(\mathbb{1}_{A_{k-1}} \times \mathbb{1}_{A_{k-2}} \otimes \cdots \otimes \mathbb{1}_{A_0}) \right) \right) \\
&= \xi \otimes P^{\otimes k}(A_k \times \cdots \times A_0) \; ,
\end{aligned}$$

which concludes the proof.

□

9.2 Discrete time renewal process

Basic definitions

Definition 9.2.1 (Renewal process)**.** *Let $(Y_n)_{n \in \mathbb{N}^*}$ be a sequence of i.i.d. positive integer-valued random variables with distribution $b = \{b(n),\ n \in \mathbb{N}^*\}$. Let Y_0 be a nonnegative integer-valued random variable, independent of the sequence $(Y_n)_{n \in \mathbb{N}^*}$.*

- *The process $(\Sigma_n)_{n \in \mathbb{N}}$ defined by*

$$S_n = \sum_{i=0}^{n} Y_i \tag{9.4}$$

is called a renewal process. The S_n's are called the renewals or the epochs of the renewal process. The common distribution of $(Y_n)_{n \in \mathbb{N}^*}$, $\mathbb{P}(Y_1 = k) = b(k)$ for $k \in \mathbb{N}^*$ is called the waiting time distribution.

- *The first renewal occurs at Y_0, the delay of the process. The renewal process is called pure or zero-delayed if $Y_0 = 0$; it is called delayed otherwise. The distribution of Y_0 is the delay distribution.*

The random variable Y_n is the duration between two successive events. It is easily seen that $(S_n)_{n\in\mathbb{N}}$ is a Markov chain on $\mathbb{N}$ with initial distribution a (the *delay distribution*) and transition kernel P given

$$P(i,j) = \begin{cases} b(j-i) & \text{if } j > i\ , \\ 0 & \text{otherwise.} \end{cases} \tag{9.5}$$

The dependence of P on the waiting time distribution b is implicit. The notation $\mathbb{P}_a^P$ stands for the distribution induced on $(\mathbb{N}^{\mathbb{N}}, \mathcal{P}(\mathbb{N})^{\otimes\mathbb{N}})$ by the Markov chain $(S_n)_{n\in\mathbb{N}}$ with Markov kernel P and initial distribution a; $\mathbb{P}_i^P$ is a short hand notation for $\mathbb{P}_{\delta_i}^P$ and $\mathbb{E}_a^P$, $\mathbb{E}_i^P$ are used for the expectation with respect to $\mathbb{P}_a^P$ and $\mathbb{P}_i^P$, respectively. There are several Markov kernels associated to a given renewal process and this is why it is important to clearly indicate in the notation which Markov chain (or kernel) we are considering.

It is often convenient to indicate the epochs by a random sequence $(V_n)_{n\in\mathbb{N}}$, where $V_n = 1$ if n is a renewal time (i.e. $\{V_n = 1\} = \bigcup_{m=0}^{\infty}\{S_m = n\}$) and $V_n = 0$ otherwise. The renewal sequence $v^{(a)} = \left(v_k^{(a)}\right)_{k\in\mathbb{N}}$ associated to the renewal times is defined through

$$v_k^{(a)} = \mathbb{P}_a^P(V_k = 1)\ , \quad k \geq 0\ . \tag{9.6}$$

For $i \in \mathbb{N}$, we write $v^{(i)}$ for $v^{(\delta_i)}$. The renewal sequence associated to the zero-delayed renewal is denoted $u = (u_k)_{k\in\mathbb{N}}$:

$$u_k = \mathbb{P}_0^P(V_k = 1)\ , \quad k \geq 0\ . \tag{9.7}$$

Note that $u = v^{(0)}$,In this case $\mathbb{P}_0^P(S_0 = 0) = 1$, we have $V_0 = 1$ $\mathbb{P}_0^P$-a.s. Since $Y_1, Y_2, \ldots, Y_m$ are i.i.d. positive integer-valued random variables with common distribution b, the distribution of $Y_1 + \cdots + Y_m$ is b^{*m}, the m-fold convolution of b, defined recursively by

$$b^{*0} = \delta\ ,\ b^{*1} = b^*\ ,\ b^{*m} = b^{*(m-1)} * b\ , \quad m \geq 1\ , \tag{9.8}$$

where $\{\delta(n),\ n \in \mathbb{N}\}$ is the sequence defined $\delta(0) = 1$ and $\delta(n) = 0$ for $n \geq 1$. For the zero-delayed renewal, $b^{*m}(k)$ is the probability that the $(m+1)$-th epoch occurs at time k. The zero-delayed renewal sequence may thus be expressed as a function of the convolution powers of the waiting time distribution b,

$$u_n = \mathbb{P}_0^P(V_n = 1) = \sum_{m=0}^{n} \mathbb{P}_0^P(S_m = n) = \sum_{m=0}^{n} b^{*m}(n)\ . \tag{9.9}$$

Note that the number of terms in the sum is less than n, because $\mathbb{P}_0^P(S_m = n) = 0$ for all $m > n$. By splitting the event on the second epoch, the renewal sequence u_n may be expressed

as

$$u_n = \mathbb{P}_0^P(V_n = 1) = \mathbb{P}(Y_1 = n) + \sum_{k=1}^{n-1} \sum_{m=2}^{n} \mathbb{P}(Y_1 = k)\mathbb{P}(Y_2 + \cdots + Y_m = n - k)$$
$$= \mathbb{P}(Y_1 = n) + \sum_{k=1}^{n-1} \mathbb{P}(Y_1 = k)\mathbb{P}_0^P(V_{n-k} = 1)$$
$$= b(n) + \sum_{k=1}^{n-1} b(k)u_{n-k} \; .$$

Since $u_0 = 1$, the latter relation may be rewritten as $u_n = \sum_{k=1}^{n} b(k)u_{n-k} = (b * u)_n$ for $n \in \mathbb{N}^*$ or more concisely,

$$u = \delta + b * u \; . \tag{9.10}$$

For $s \in \mathbb{C}$, denote by $U(s) = \sum_{n=0}^{\infty} u_n s^n$ and $B(s) = \sum_{n=1}^{\infty} b(n)s^n$ the generating functions of the zero-delayed renewal sequence $(u_n)_{n\in\mathbb{N}}$ and of the waiting time distribution $\{b(n),\ n \in \mathbb{N}^*\}$. These series are absolutely convergent for $|s| < 1$. These generating function satisfy

$$U(s) = 1 + B(s)U(s) \quad \Rightarrow \quad U(s) = \frac{1}{1 - B(s)} \; . \tag{9.11}$$

The delayed renewal sequence $\left(v_n^{(a)}\right)_{n\in\mathbb{N}}$ can easily be related to the zero-delayed renewal sequence $(u_n)_{n\in\mathbb{N}}$ by

$$v_n^{(a)} = \mathbb{P}_a(V_n = 1) = \mathbb{P}_a(Y_0 = n) + \sum_{k=1}^{n-1} \mathbb{P}_a(Y_0 = k) \sum_{m=1}^{n-1} \mathbb{P}(Y_1 + \cdots + Y_m = n - k)$$
$$= a(n) + \sum_{k=1}^{n-1} a(k)u_{n-k} = \sum_{k=0}^{n} a(k)u_{n-k} \; .$$

Again delayed renewal sequence may be more concisely rewritten as

$$v^{(a)} = a * u \; . \tag{9.12}$$

Plugging $u = \delta + b * u$ into the previous relation leads to the *renewal equation.*

$$v^{(a)} = a + b * v^{(a)} \; . \tag{9.13}$$

Denoting by $V^{(a)}(s) = \sum_{k=0}^{\infty} v_k^{(a)} s^k$ and $A(s) = \sum_{k=0}^{\infty} a(k)s^k$, (9.13) yields

$$V^{(a)}(s) = A(s) + B(s)V^{(a)}(s) \quad \Rightarrow \quad V^{(a)}(s) = \frac{A(s)}{1 - B(s)} \; . \tag{9.14}$$

If the *mean waiting time* (or *mean recurrence time* is finite $\sum_{j=1}^{\infty} jb(j) < \infty$, the delay distribution $\{a(k),\ k \in \mathbb{N}\}$ may be chosen in such a way that the delayed renewal sequence $(v_k)_{k\in\mathbb{N}}$ is constant. Suppose that $v_k = c$ for all $k \in \mathbb{N}$ where c is some positive constant that will be chosen later. Then, $V(s) = c(1 - s)^{-1}$ and $A(s) = c(1 - B(s))(1 - s)^{-1}$. Inverting the generating function yields, for any $k \geq 0$,

$$a(k) = c\left(1 - \sum_{j=1}^{k} b(j)\right) = c \sum_{j=k+1}^{\infty} b(j) = c\mathbb{P}(Y_1 \geq k + 1) \; .$$

Since $1 = \sum_{k=0}^{\infty} a(k)$, the constant should be chosen so that $1 = c\sum_{k=1}^{\infty} \mathbb{P}(Y_1 \geq k) = c\mathbb{E}[Y_1]$, and thus, the constant c is the renewal intensity defined as the inverse the mean waiting (or recurrence) time $c = 1/\mathbb{E}[Y_1]$.

Proposition 9.2.1. *Assume that the mean waiting time is finite, i.e.* $\sum_{j=1}^{\infty} jb(j) < \infty$. *The unique delay distribution yielding a constant delayed renewal sequence is given by*

$$\bar{a}(k) = \frac{\sum_{j=k+1}^{\infty} b(j)}{\sum_{j=1}^{\infty} jb(j)} , \quad k \geq 0 . \tag{9.15}$$

In that case, the constant value taken by the renewal sequence $\left(v_n^{(\bar{a})}\right)_{n\in\mathbb{N}}$ *is equal to the renewal intensity* $\lambda = (\sum_{j=1}^{\infty} jb(j))^{-1}$.

Backward and forward recurrence times

Renewal processes and sequences have close connections to Markov chains as discussed below. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain on a discrete state space X with transition kernel K Choose a particular state $\alpha \in \mathsf{X}$. Denote by $Y_k = \sigma_\alpha^{(k+1)} - \sigma_\alpha^{(k)}$, $k \geq 1$, the waiting time between the $(k+1)$-th and the k-th return times to the state α. Then $(Y_k)_{k\in\mathbb{N}^*}$ is i.i.d. under $\mathbb{P}_\alpha$ according to the strong Markov property, the common distribution b is given by $b(k) = \mathbb{P}_\alpha(\sigma_\alpha = k)$, $k > 1$, and the zero-delayed renewal sequence is $u_k = P^k(\alpha, \alpha)$.

Given a waiting time distribution b, we may then construct a Markov chain $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{N}$ such that the delay distribution b is equal to the distribution of the return time of the Markov chain to some state. In this section, we consider a zero-delayed renewal process ($Y_0 = 0$) and we denote by A_k (for $k \in \mathbb{N}$) the time before the next renewal, also called the *residual lifetime*, or the *forward recurrence time chain*:

$$A_k = \inf_{\{n\,:\,S_n > k\}} S_n - k . \tag{9.16}$$

Observe that if $A_k = j$ for some $j > 1$, then $A_{k+1} = j - 1$ and if $A_k = 1$, then a renewal

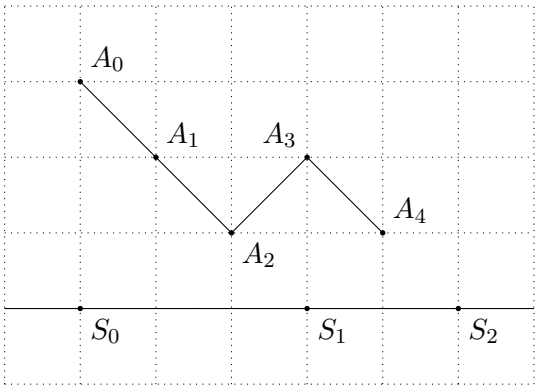

Figure 9.1: An example of residual lifetime process.

occurs at time $k+1$. Set $C = \{A_0 = i_0, \dots, A_{k-1} = i_{k-1}\}$ where $i_\ell \in \mathbb{N}^*$. Then, we have

$$\begin{aligned}\mathbb{P}(C, A_k = 1, A_{k+1} = j) &= \sum_{m=1}^{k+1} \mathbb{P}(C, A_k = 1, S_m = k+1, S_{m+1} - S_m = j) \\ &= \sum_{m=1}^{k+1} \mathbb{P}(C, A_k = 1, S_m = k+1)\mathbb{P}(S_{m+1} - S_m = j) \\ &= \mathbb{P}(C, A_k = 1)b(j) \ .\end{aligned}$$

Finally, $(A_k)_{k\in\mathbb{N}}$ is a Markov chain on $\mathbb{N}^*$ with transition matrix Q defined by

$$Q(1, j) = b(j) \qquad j \geq 1 \ , \tag{9.17a}$$

$$Q(j, j-1) = 1 \qquad j \geq 2 \ . \tag{9.17b}$$

If the mean waiting time is finite, i.e. $\sum_{j=1}^{\infty} jb(j) < \infty$ then the probability measure $\tilde{\pi}$ on $\mathbb{N}^*$ defined by

$$\tilde{\pi}(j) = \frac{\sum_{\ell=j}^{\infty} b(\ell)}{\sum_{q=1}^{\infty} qb(q)} \ , \ j \geq 1 \ , \tag{9.18}$$

is invariant for Q. Note indeed that, for $j \geq 1$,

$$\begin{aligned}\tilde{\pi}Q(j) &= \sum_{i=1}^{\infty} \tilde{\pi}(i)Q(i, j) = \tilde{\pi}(1)b(j) + \tilde{\pi}(j+1) \\ &= \left(\mathbb{E}\left[Y_1\right]\right)^{-1} \left(b(j) + \sum_{\ell=j+1}^{\infty} b(\ell) \right) = \tilde{\pi}(j) \ .\end{aligned}$$

For $k \in \mathbb{N}$, let B_k be the time spent since the last renewal, also called the *age process*, or the *backward recurrence time chain*, defined by

$$B_k = k - \sup_{\{n \,:\, S_n \leq k\}} S_n \ . \tag{9.19}$$

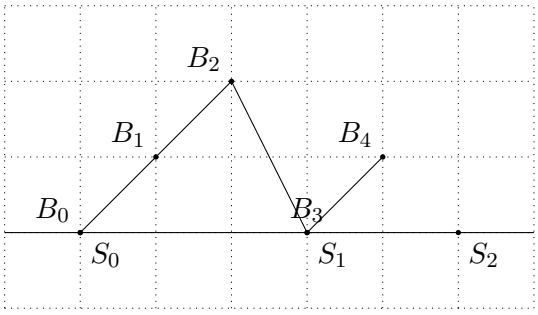

Figure 9.2: An example of age process.

Note that $B_k = 0$ if k is a renewal time and $B_k = k - S_n$ if $S_n < k < S_{n+1}$. Thus $A_k + B_k = S_{n+1} - S_n$ if $S_n \leq k \leq S_{n+1}$, that is, $A_k + B_k$ is the duration of the current renewal period.

First note that, if $B_k = i$ for some $i \in \mathbb{N}$, then $B_{k+1} = i+1$ or 0 and $B_{k-1} = i-1, \dots, B_{k-i} = 0$ and $k-i$ is a renewal time. Setting $D = \{B_0 = i_0, \dots, B_{k-i-1} = i_{k-i-1}\}$. Then

$$\begin{aligned}\mathbb{P}(B_0 = i_0, \dots, B_{k-1} = i_{k-1}, B_k = i) &= \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i, S_{m+1} - S_m > i) \\ &= \mathbb{P}(Y_1 > i) \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i) \; .\end{aligned}$$

Moreover,

$$\begin{aligned}&\mathbb{P}(B_0 = i_0, \dots, B_{k-1} = i_{k-1}, B_k = i, B_{k+1} = i+1) \\ &\quad = \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i, S_{m+1} - S_m > i+1) \\ &\quad = \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i) \mathbb{P}(S_{m+1} - S_m > i+1) \\ &\quad = \mathbb{P}(Y_1 > i+1) \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i) \; .\end{aligned}$$

Along the same lines, we obtain

$$\mathbb{P}(B_0 = i_0, \dots, B_{k-1} = i_{k-1}, B_k = i, B_{k+1} = 0) = \mathbb{P}(Y_1 = i+1) \sum_{m=0}^{k} \mathbb{P}(D, S_m = k-i) \; .$$

Altogether, this implies that $(B_k)_{k\in\mathbb{N}}$ is a nonnegative integer-valued Markov chain with transition matrix R defined for $n \in \mathbb{N}$ by

$$R(n, n+1) = \mathbb{P}\left[Y_1 > n+1 \,|\, Y_1 > n\right] \; , \tag{9.20a}$$

$$R(n, 0) = \mathbb{P}\left[Y_1 = n+1 \,|\, Y_1 > n\right] \; . \tag{9.20b}$$

If the mean waiting time is finite, then $\bar{\pi}(j) = (\mathbb{E}[Y_1])^{-1} \sum_{\ell=j+1}^{\infty} b(\ell)$, $j \geq 0$, is R-invariant. Indeed, for $j \geq 1$, we get

$$\begin{aligned}\bar{\pi}R(j) = \bar{\pi}(j-1)R(j-1, j) &= \frac{\sum_{\ell=j}^{\infty} b(\ell)}{\mathbb{E}[Y_1]} \frac{\sum_{\ell=j+1}^{\infty} b(\ell)}{\sum_{\ell=j}^{\infty} b(\ell)} \\ &= \frac{\sum_{\ell=j+1}^{\infty} b(\ell)}{\mathbb{E}[Y_1]} = \bar{\pi}(j) \; ,\end{aligned}$$

and, for $j = 0$,

$$\bar{\pi}R(0) = \sum_{i=0}^{\infty} \frac{\sum_{\ell=i+1}^{\infty} b(\ell)}{\mathbb{E}[Y_1]} \frac{b(i+1)}{\sum_{\ell=i+1}^{\infty} b(\ell)} = \frac{1}{\mathbb{E}[Y_1]} = \bar{\pi}(0) \; .$$

9.3 Time Series Examples

Many time series models are Markov chains or can be embedded into Markov chains or use Markov chains as a building block. We now describe some classical time series models in the framework of Markov chains.

9.3.1 Autoregressive processes

Example 9.3.1 (AR(1) process)**.** *The AR(1) process $(X_k)_{k\in\mathbb{N}}$ is defined recursively as follows: the value of the process X_k at time k, is an affine combination of the previous value X_{k-1} and an innovation (noise), i.e.*

$$X_k = \mu + \Phi X_{k-1} + Z_k \,, \tag{9.21}$$

where $(Z_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional random vectors, independent of X_0. The AR(1) model can be seen as an extension of the random walk model

$$X_k = X_{k-1} + Z_k \,. \tag{9.22}$$

We assume that $\mathbb{E}\left[|Z_1|\right] < \infty$ and $\mathbb{E}\left[Z_1\right] = 0$. Then $(X_k)_{k\in\mathbb{N}}$ is a Markov chain with kernel

$$P(x, A) = \mathbb{P}(Z_1 + \Phi x + \mu \in A)\,, \quad A \in \mathcal{B}(\mathbb{R}^d) \,. \tag{9.23}$$

Equivalently, for all $h \in \mathrm{F}_+(\mathsf{X}, \mathcal{X})$ and $x \in \mathsf{X}$,

$$Ph(x) = \mathbb{E}\left[h(\mu + \Phi x + Z_1)\right] \,.$$

Example 9.3.2 (AR(p) process)**.** *The AR(1) process can be generalized by assuming that the current value is obtained as an affine combination of the p preceding values of the process and a random disturbance. Let $(Z_k)_{k\in\mathbb{N}}$ be a sequence of i.i.d. real valued random variables. Let $\phi_1, \ldots, \phi_p$ be real numbers and let $X_0, X_{-1}, \ldots, X_{-p+1}$ be random variables, independent of the sequence $(Z_k)_{k\in\mathbb{N}}$. The AR(p) process $(X_k)_{k\in\mathbb{N}}$ is defined by the recursion*

$$X_k = \mu + \phi_1 X_{k-1} + \phi_2 X_{k-2} + \cdots + \phi_p X_{k-p} + Z_k \,, \quad k \geq p \,. \tag{9.24}$$

The sequence $(X_k)_{k\in\mathbb{N}}$ is not a Markov chain, but can easily be embedded in a Markov chain. The vector process $\mathbf{X}_k = (X_k, X_{k-1}, \ldots, X_{k-p+1})$ is a vector autoregressive process of order 1, defined by the recursion:

$$\mathbf{X}_k = B\mu + \Phi \mathbf{X}_{k-1} + BZ_k \tag{9.25}$$

with

$$\Phi = \begin{pmatrix} \phi_1 & \cdots & \cdots & \phi_p \\ 1 & 0 & & 0 \\ \vdots & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix} \,, \quad B = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \,.$$

Thus $(\mathbf{X}_k)_{k\in\mathbb{N}}$ is an $\mathbb{R}^p$ valued Markov chain with kernel

$$P(\mathbf{x}, A) = \mathbb{P}(B\mu + BZ_0 + \Phi \mathbf{x} \in A) \tag{9.26}$$

for $\mathbf{x} \in \mathbb{R}^p$ and $A \in \mathcal{B}(\mathbb{R}^p)$.

9.3.2 Simple extensions of autoregressive processes

In the AR(1) model, the conditional expectation of the value of the process at time k is an affine function of the previous value: $\mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}^X\right] = \mu + \phi X_{k-1}$. In addition, provided that $\mathbb{E}\left[Z_1^2\right] < \infty$ in (9.21), the conditional variance is almost-surely constant since $\mathbb{E}\left[(X_k - \mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}^X\right])^2 \middle| \mathcal{F}_{k-1}^X\right] = \mathbb{E}\left[Z_1^2\right]$ $\mathbb{P}$-a.s.. We say that the model is a *linear Markov conditionally homoscedastic* model. Of course, these assumptions can be relaxed in several directions. We might first consider models which are still conditionally homoscedastic, but for which the conditional expectation of X_k given the past is a non-linear function of the past observation X_{k-1}, leading to the following model.

Example 9.3.3 (FAR(1) process)**.** *The functional autoregressive is given by*

$$X_k = f(X_{k-1}) + Z_k \,, \tag{9.27}$$

where $(Z_k)_{k\in\mathbb{N}^*}$ be a sequence of integrable zero-mean i.i.d. real-valued random variable independent of X_0 and $f : \mathbb{R} \to \mathbb{R}$ is a measurable function. With this definition, the conditional expectation of X_k given $\mathcal{F}_{k-1}^X$ is given by $f(X_{k-1}) = \mathbb{E}\left[X_k \middle| \mathcal{F}_{k-1}^X\right]$. The kernel of this chain is given by

$$P(x, A) = \mathbb{P}(Z_1 - f(x) \in A) \,, \quad A \in \mathcal{B}(\mathbb{R}) \,.$$

For any $h \in \mathrm{F}_+(\mathsf{X}, \mathcal{X})$, we get

$$Ph(x) = \mathbb{E}\left[h(f(x) + Z_1)\right] \,.$$

A completely different direction for extending AR processes is to introduce *conditional heteroscedasticity.*

Example 9.3.4 (ARCH(p))**.** *It has been generally acknowledged in the econometrics and applied financial literature that many financial time series such as log-returns of share prices, stock indices, and exchange rates, exhibit stochastic volatility and heavy-tailedness. These features cannot be adequately modelled via a linear time series model. Nonlinear models, such as the ARCH model and the bilinear models (see Example 9.3.5) have been proposed to capture these and other characteristics. In order for a linear time series model to possess heavy-tailed marginal distributions, it is necessary for the input noise sequence to be heavy-tailed. For non-linear models, heavy-tailed marginals can be obtained when the system is injected with light-tailed marginals such as with normal noise. An Autoregressive Conditional Heteroscedastic model of order p, ARCH(p), is defined as a solution of the recursion*

$$X_k = \sigma_k \, Z_k \,, \tag{9.28a}$$

$$\sigma_k^2 = \alpha_0 + \alpha_1 X_{k-1}^2 + \cdots + \alpha_p X_{k-p}^2 \,, \tag{9.28b}$$

where the coefficients $\alpha_j \geq 0$, $j \in \{0, \ldots, p\}$ are non-negative and $(Z_k)_{k\in\mathbb{Z}}$ is a sequence of i.i.d. random variable with zero mean (often assumed to be standard Gaussian). The ARCH(p) process is a Markov chain of order p. Assume that Z_1 has a bounded continuous density g with respect to Lebesgue's measure on $\mathbb{R}$. Then, for $h \in \mathbb{F}_+(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ we get

$$\begin{aligned} &Ph(x_1, \ldots, x_p) \\ &\quad = \mathbb{E}\left[h(\sqrt{\alpha_0 + \alpha_1 x_1^2 + \cdots + \alpha_p x_p^2} Z_1)\right] \\ &\quad = \int h(y) \frac{1}{\sqrt{\alpha_0 + \alpha_1 x_1^2 + \cdots + \alpha_p x_p^2}} g\left(\frac{y}{\sqrt{\alpha_0 + \alpha_1 x_1^2 + \cdots + \alpha_p x_p^2}}\right) \mathrm{d}y \,. \end{aligned}$$

The kernel therefore has a density with respect to Lebesgue's measure given by

$$p(x_1,\dots,x_p;y)=\frac{1}{\sqrt{\alpha_0+\alpha_1x_1^2+\cdots+\alpha_px_p^2}}g\left(\frac{y}{\sqrt{\alpha_0+\alpha_1x_1^2+\cdots+\alpha_px_p^2}}\right)\ .$$

We will latter see that it is relatively easy to discuss the properties of this model, which is used widely in financial econometric.

Example 9.3.5 (Simple Markov bilinear model)**.** *The simple Markov bilinear process is defined by the recursion*

$$X_k=aX_{k-1}+(1+bX_{k-1})Z_k\ , \tag{9.29}$$

where a and b are scalar and $(Z_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of random variables which are independent of X_0. Assuming that $\mathbb{E}\left[Z_1^2\right]=1$ and $\mathbb{E}\left[Z_1\right]=0$, the conditional expectation of X_k is linear $\mathbb{E}\left[X_k|\,\mathcal{F}_{k-1}^X\right]=\theta X_{k-1}$ like an AR(1) model but the model is conditionally heteroscedastic with conditional variance given by $\mathrm{Var}\left(X_k\,|\,\mathcal{F}_{k-1}^X\right)=(1+bX_{k-1})^2$. It might be seen as a AR(1) process with ARCH(1) error. This model is useful for modeling financial time series in which the current volatility depends on the past value, including on its sign. This asymmetry has been pointed out to be a characteristic feature of financial time series.

Example 9.3.6 (Self-Exciting threshold AR model)**.** *Self-exciting threshold AR (SETAR) models have been widely employed as a model for nonlinear time series. Threshold models are piecewise linear AR models for which the linear relationship varies according to delayed values of the process (hence the term* self-exciting*). In this class of models, it is hypothesized that different autoregressive processes may operate and that the change between the various AR is governed by threshold values and a time lag. A ℓ-regimes TAR model has the form*

$$X_k=\begin{cases}\phi_0^{(1)}+\sum_{i=1}^{p_1}\phi_i^{(1)}X_{k-i}+\sigma^{(1)}Z_k^{(1)} & \text{if } X_{k-d}\le r_1\ ,\\ \phi_0^{(2)}+\sum_{i=1}^{p_2}\phi_i^{(2)}X_{k-i}+\sigma^{(2)}Z_k^{(2)} & \text{if } r_1<X_{k-d}\le r_2\ ,\\ \quad\vdots & \quad\vdots\\ \phi_0^{(\ell)}+\sum_{i=1}^{p_\ell}\phi_i^{(\ell)}X_{k-i}+\sigma^{(\ell)}Z_k^{(\ell)} & \text{if } r_{\ell-1}<X_{k-d}\ ,\end{cases} \tag{9.30}$$

where $(Z_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of real-valued random variables, the positive integer d is a specified delay, and $-\infty<r_1<\cdots<r_{k-1}<\infty$ is a partition of $\mathsf{X}=\mathbb{R}$. These models allow for changes in the AR coefficients over time, and those changes are determined by comparing previous values (back-shifted by a time lag equal to d) to fixed threshold values. Each different AR model is referred to as a regime*. In the definition above, the values p_j of the order of AR models can differ in each regime, although in many applications, they are equal.*

The model can be generalized to include the possibility that the regimes depend on a collection of the past values of the process, or that the regimes depend on an exogenous variable (in which case the model is not self-exciting).

The popularity of TAR models is due to their being relatively simple to specify, estimate, and interpret as compared to many other nonlinear time series models. In addition, despite its apparent simplicity, the class of TAR models can reproduce many nonlinear phenomena such as stable and unstable limit cycles, jump resonance, harmonic distortion, modulation effects, chaos and so on.

Example 9.3.7 (Smooth Transition Autoregressive model (STAR)). *Another possible extension of the autoregressive models is to replace the* hard *thresholds by* smooth *functions, therefore avoiding discontinuity in the autoregressive coefficients. As an example, consider the following smooth autoregressive model (which might be seen as an extension of a 2-regimes STAR model)*

$$X_k = \sum_{i=1}^{p} \{\phi_i + \pi_i u(X_{k-d})\} X_{k-i} + Z_k \ , \tag{9.31}$$

where u is either given by

$$u(x) = (1 + \exp\{-\gamma(x-c)\})^{-1} \ .$$

leading to the logistic smooth transition autoregressive model or

$$u(x) = 1 - \exp\{-\gamma(x-c)^2\} \tag{9.32}$$

leading to the exponential smooth transition autoregressive model. In both cases, γ is a positive constant and $c \in \mathbb{R}$ In this case, the autoregressive part retains an additive form, but the coefficients entering the regression vary smoothly from one regime to the other with X_{k-d}.

9.4 Discrete State Space Examples

9.4.1 Random walks

Let $(Z_n)_{n\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables with values in $\mathbb{Z}^d$ and distribution ν. Let X_0 be a random variable in $\mathbb{Z}^d$ independent of $(Z_n)_{n\in\mathbb{N}^*}$. A random walk with jump distribution ν is a process $(X_k)_{k\in\mathbb{N}}$ defined by X_0 and the recursion

$$X_{n+1} = X_n + Z_{n+1} \ .$$

This is a Markov chain with kernel P defined on $\mathbb{Z}^d \times \mathbb{Z}^d$ by

$$P(x,y) = \nu(y-x) \, , \quad x, y \in \mathbb{Z}^d \ .$$

The random walk is *spatially homogeneous* in the sense that $P(x,y) = P(0, y-x) = \nu(y-x)$, showing that the kernel is determined by the jump distribution ν. The random walk is said to be *symmetric* if $\nu(x) = \nu(-x)$ for all $x \in \mathsf{X}$.

Example 9.4.1 (Simple random walk). *A* simple random walks *on $\mathbb{Z}^d$ is a specific example of spatially homogeneous and symmetric random walk in which the increment distribution gives equal weight $1/(2d)$ to the points $x \in \mathbb{Z}^d$ satisfying $|x| = 1$. The transition kernel of the* d-dimensional simple random walk *is given by $P(x,y) = 1/(2d)$ if $|y-x| = 1$ and $P(x,y) = 0$ otherwise.*

Example 9.4.2 (Bernoulli random walk). *The Bernoulli random walk on $\mathbb{N}$ is defined by*

$$P(x, x+1) = p \, , \quad P(x, x-1) = q \, , \quad p \geq 0 \, , \quad q \geq 0, p+q = 1 \ .$$

For $n \in \mathbb{N}$, $P^n(0,x)$ is the probability of an n-step transition from 0 to x (the probability that a "particle", starting at zero, finds itself after n iterations at x). Suppose that n and x are

both even or odd and that $|x| \leq n$ (otherwise $P^n(0,x) = 0$). Then $P^n(0,x)$ is the probability of $1/2(x+n)$ successes in n independent Bernoulli trials, where the probability of success is p. Therefore

$$P^n(0,x) = p^{(n+x)/2} q^{(n-x)/2} \binom{n}{(n+x)/2},$$

where the sum $n+x$ is even and $|x| \leq n$ and $P^n(0,x) = 0$ otherwise.

Example 9.4.3 (Reflected and absorbed random walk on $\mathbb{N}$)**.** *If the simple random walk is restricted to the nonnegative integers, then the zero state is called a barrier. If $P(0,1) = 1$, then it is called a reflecting barrier. If $P(0,0) = 1$ then 0 is called an absorbing barrier. Once the particle reaches zero it remains there forever. If $P(0,1) > 0$ and $P(0,0) > 0$, then 0 is called a partially reflecting barrier.*

If the simple random walk is restricted to a finite number of states, say $0, 1, 2, \ldots, a$, then both the states 0 and a may be reflecting, absorbing, or partially reflecting barriers.

A simple random walk on $\mathbb{Z}$ can be used to describe the fortune of a gambler engaged in a series of games whose outcome is winning or losing one unit. In that case $P(x,x) = 0$. If the player cannot be indebted or borrow money, then the state 0 is an absorbing barrier and this event is called the gambler's ruin. A reflecting barrier at $a > 0$ means that the gambler will cash in all their gain above a and an absorbing barrier at a means that the gambler will stop playing as soon as their fortune reaches the level a.

9.4.2 Population models

Example 9.4.4 (Level-dependent quasi Birth-and-death process)**.** *A level-dependent quasi birth-and-death process is a Markov chain on the finite or infinite subset $\mathsf{X} = \{a,\ a+1,\ \ldots,\ b\}$ ($-\infty \leq a < b \leq \infty$) of $\mathbb{Z}$ with kernel P defined by*

$$P(x, x+1) = p_x\,, \quad P(x, x-1) = q_x\,, \quad P(x,x) = r_x\,,$$

with $p_x + q_x + r_x = 1$ for all $x \in \mathsf{X}$.

This process can be used to describe the position of a particle moving on a grid, which at each step may only remain at the same state or move to an adjacent state with a probability possibly depending on the state.

If $P(0,0) = 1$ and $p_x + q_x = 1$ for $x > 0$, this process may be considered as a model for the size of a population, recorded each time it changes, p_x being the probability that a birth occurs before a death when the size of the population is x. Birth-and-death have many applications in demography, queueing theory, performance engineering or biology. They may be used to study the size of a population, the number of diseases within a population or the number of customers waiting in a queue for a service.

Example 9.4.5 (Ehrenfest's Urn)**.** *This model, also called the dog-flea model, is a Markov chain on a finite state space $\{0, \ldots, N\}$ where $N > 1$ is a fixed integer. Balls (or particles) numbered 1 to N are divided among two urns A and B. At each step, an integer i is drawn at random and the ball numbered i is moved to the other urn. The number X_n of balls at time n is a Markov chain on $\{0, \ldots, N\}$ with transition matrix P defined by*

$$P(i, i+1) = \frac{N-i}{N}\,, \quad i = 0, \ldots, N-1\,,$$
$$P(i, i-1) = \frac{i}{N}\,, \quad i = 1, \ldots, N\,.$$

The states 0 *and* N *are reflecting barriers. The Binomial distribution* $B(N, 1/2)$ *is reversible with respect to the kernel* P*. Indeed, for all* $i = 0, \ldots, N-1$*,*

$$\binom{N}{i}\frac{N-i}{N} = \frac{N!(N-i)}{i!(N-i)!N} = \binom{N}{i+1}\frac{i+1}{N} .$$

This is the detailed balance condition of Definition 9.1.1. Thus the binomial distribution $B(N, 1/2)$ *is invariant.*

For $n \geq 1$*,*

$$\mathbb{E}\left[X_n \middle| X_{n-1}\right] = (X_{n-1}+1)\frac{N-X_n}{N} + (X_{n-1}-1)\frac{X_n}{N} = X_{n-1}(1-2/N)+1 .$$

Set $m_n(x) = \mathbb{E}_x[X_n]$ *for* $x \in \{0, \ldots, N\}$ *and* $a = 1-2/N$*, this yields*

$$m_n(x) = am_{n-1}(x) + 1 .$$

The solution of this recurrence equation is

$$m_n(x) = xa^n + \frac{1-a^n}{1-a} ,$$

and since $0 < a < 1$*, this yields that* $\lim_{n\to\infty} \mathbb{E}_x[X_n] = 1/(1-a) = N/2$*, which is the expectation of the stationary distribution.*

Example 9.4.6 (Wright-Fisher model)**.** *The Wright-Fisher model is an idealized genetics model used to investigate the fluctuation of gene frequency in a population of constant size under the influence of mutation and selection. The model describes a simple haploid random reproduction disregarding selective forces and mutation pressure. The size of the population is set to* N *individuals of two types* 1 *and* 2*. Let* X_n *be the number of individuals of type* 1 *at time* n*. Then* $(X_n)_{n\in\mathbb{N}}$ *is a Markov chain with state-space* $\mathsf{X} = \{0, 1, \ldots, N\}$ *and transition matrix*

$$P(j,k) = \binom{N}{k}\left(\frac{j}{N}\right)^k\left(1-\frac{j}{N}\right)^{N-k} ,$$

with the usual convention $0^0 = 1$*. In words, given that the number of type* 1 *individuals at the current generation is* j*, the number of type* 1 *individuals at the next generation follows a binomial distribution with success probability* j/N*. Looking backwards, this can be interpreted as having each of the individual in the next generation 'pick their parents at random' from the current population. The states* 0 *and* N *are absorbing.*

Example 9.4.7 (Galton-Watson process)**.** *The Galton-Watson process is a branching process arising from Francis Galton's investigation of the extinction of family names. The process models family names as patrilineal (passed from father to son); offsprings are either male or female, and names become extinct holders of the family name die without male descendants. This model has been applied in many different applications including the survival probabilities for a new mutant gene, the initiation of a nuclear chain reaction, the dynamics of disease outbreaks in their first generations of spread, or the chances of extinction of small population of organisms.*

A Galton-Watson process is a stochastic process $(X_n)_{n\in\mathbb{N}}$ *which evolves according to the recursion* $X_0 = 1$ *and*

$$X_{n+1} = \sum_{j=1}^{X_n} \xi_j^{(n+1)} \ , \tag{9.33}$$

where $\{\xi_j^{(n+1)} : n, j \in \mathbb{N}\}$ *is a set of i.i.d. nonegative integer-valued random variables with distribution* ν*. The random variable* X_n *can be thought of as the number of descendants (along the male line) in the n-th generation, and* $\{\xi_j^{(n+1)}, j = 1, \dots, X_n\}$ *represents the number of (male) children of the j-th descendant of the n-th generation.*

The conditional distribution of X_{n+1} *given the past depends only on the current size of the population* X_n *and the the number of offsprings of each individual* $\{\xi_j^{(n+1)}\}_{j=1}^{X_n}$ *which are conditionally independent given the past. The process* $(X_n)_{n\in\mathbb{N}}$ *is therefore an homogeneous Markov chain whose transition matrix is given by* $P(0,0) = 1$ *and for* $j \in \mathbb{N}^*$ *and* $k \in \mathbb{N}$*,*

$$P(j,k) = \sum_{(k_1,\dots,k_j)\in\mathbb{N}^j, k_1+\cdots+k_j=k} \nu(k_1)\nu(k_2)\cdots\nu(k_j) = \nu^{*j}(k) \ .$$

Example 9.4.8 (INAR process)**.** *An INAR (INteger AutoRegressive) process is a Galton Walton process with immigration, defined by the recursion* $X_0 = 1$ *and*

$$X_{n+1} = \sum_{j=1}^{X_n} \xi_j^{(n+1)} + Y_{n+1} \ , \tag{9.34}$$

where $\{\xi_j^{(n)},\ j, n \in \mathbb{N}^*\}$ *are i.i.d. integer-valued random variables and* $(Y_n)_{n\in\mathbb{N}^*}$ *is a sequence of i.i.d. integer-valued random variables, independent of* $\{\xi_j^{(n)}\}$*. The random variable* Y_{n+1} *represents the "immigrants", that is the part of the* $(n+1)$*-th generation which does not descend from the n-th generation. Contrary to the previous model, the state 0 is not absorbing.*

Let ν *be the distribution of* ξ_1^1 *and* μ *be the distribution of* Y_1*. Then the transition matrix of the INAR process is given, for* $j \in \mathbb{N}$ *and* $k \in \mathbb{N}$*, by*

$$P(j,k) = \mu * \nu^{*j}(k) \ .$$

9.4.3 Queueing and storage models

Example 9.4.9 (Discrete Time queueing system)**.** *Clients arrive for service and enter in a queue. During each time interval a single customer is served, provided that at least one customer is present in the queue. We assume that the number of arrivals during the n-th service period is a sequence of i.i.d. integer-valued random variable* $(Z_n)_{n\in\mathbb{N}}$*, independent of the initial state* X_0 *and whose distribution is given by*

$$\mathbb{P}(Z_n = k) = a_k \geq 0 \ , \quad k \in \mathbb{N} \ , \quad \sum_{k=0}^{\infty} a_k = 1 \ .$$

The state of the queue at the start of each period is defined to be the number of clients waiting for service, which is given by

$$X_{n+1} = (X_n - 1)_+ + Z_{n+1} \ .$$

The Markov kernel of the chain is given by $P(0, y) = a_y$ *for* $y \in \mathbb{N}$ *and for* $x \geq 1$,

$$P(x, y) = \begin{cases} a_{y-x+1} & \text{if } y \geq x - 1 \text{ ,} \\ 0 & \text{otherwise.} \end{cases}$$

Appendix A

Convergence of random elements in a metric space

In this appendix we provide the main definitions and results concerning the convergence of a sequence of random elements valued in a metric space. The strong convergence and the convergence in probability are not more difficult in this setting than in the case of vector valued random variables. The weak convergence is more delicate as some topology properties of the metric space have to be considered. A classical reference for the weak convergence in metric spaces is [1]. Here we provide a brief account of the essential classical definitions and results. The detailed proofs can be found in [1].

From now on, we let (X, d) be a metric space. We note $\mathrm{C}_b(\mathsf{X})$ (resp. $\mathrm{Lip}_b(\mathsf{X})$) the space of real-valued bounded continuous functions (resp. bounded and Lipschitz) on (X, d). We denote by $\mathcal{B}(\mathsf{X})$ the Borel σ-fields on X and by $\mathbb{M}_1(\mathsf{X})$ the set of probability measures on $\mathcal{B}(\mathsf{X})$.

A.1 Definitions and characterizations

As mentioned above, the weak convergence is in general more delicate to handle than other convergences. An additional difficulty is that it is often presented as a "convergence" of a sequence of random variables, but the word "convergence" is not rigorous in such a presentation. In fact the weak convergence defines a convergence for the sequence of the marginal distributions, thus, for a sequence valued in $\mathbb{M}_1(\mathsf{X})$, the set of probability measures on X.

The term weak convergence is opposed to strong convergence which, in contrast, makes sense only for a sequence of random variables.

Definition A.1.1 (Strong convergence)**.** *Let X, X_n, $n \geq 1$ be random variables valued in $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will say that X_n strongly converges to X and denote $X_n \xrightarrow{\text{a.s.}} X$ in (X, d) (or simply $X_n \xrightarrow{\text{a.s.}} X$ if no ambiguity occurs) if $d(X_n, X) \to 0$ almost surely.*

A basic criterion for proving strong convergence is based on the Borel Cantelli lemma.

Lemma A.1.1 (Borel Cantelli's Lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of measurable sets. Then,*

$$\sum_{k\in\mathbb{N}} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}(\limsup A_n) = 0 \,.$$

In particular, if X, X_n, $n \geq 1$ *are random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that, for any* $\epsilon > 0$,

$$\sum_{k \in \mathbb{N}} \mathbb{P}(d(X_n, X) > \epsilon) < \infty \ ,$$

then $X_n \xrightarrow{\text{a.s.}} X$.

The convergence in probability also applies to a sequence of random variables. It is weaker than the strong convergence.

Definition A.1.2 (Convergence in probability)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *We will say that* X_n *converges in probability to* X *and denote* $X_n \xrightarrow{P} X$ *in* (X, d) *(or simply* $X_n \xrightarrow{P} X$ *if no ambiguity occurs) if* $\mathbb{P}(d(X_n, X) > \epsilon) \to 0$ *for any* $\epsilon > 0$.

It is straightforward to verify that the convergence in probability can be characterized with the strong convergence as follows.

Theorem A.1.2. *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Then we have* $X_n \xrightarrow{P} X$ *if and only if for all subsequence* (X_{α_n}), *there is a subsequence* $(X_{\alpha_{\beta_n}})$ *such that* $X_{\alpha_{\beta_n}} \xrightarrow{\text{a.s.}} X$.

The following result shows that any probability measure μ defined on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is *regular*, in the sense that it can be defined for all $A \in \mathcal{B}(\mathsf{X})$ by

$$\mu(A) = \inf \{\mu(U) \ : \ U \text{ open set } \supset A\} = \sup \{\mu(F) \ : \ F \text{ closed set} \subset A\} \ . \tag{A.1}$$

Proposition A.1.3. *Let* $\mu \in \mathbb{M}_1(\mathsf{X})$. *Then (A.1) holds for all* $A \in \mathcal{B}(\mathsf{X})$.

Definition A.1.3 (Weak convergence of probability measures)**.** *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$. *We say that* μ_n *converges weakly to* μ *if, for all* $f \in \mathrm{C}_b(\mathsf{X})$, $\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$.

Weak convergence is also often used for a sequence of random variables.

Definition A.1.4 (Weak convergence of random variables)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. *We will say that* X_n *converges weakly to* X *and denote* $X_n \Longrightarrow X$ *in* (X, d) *(or simply* $X_n \Longrightarrow X$ *if no ambiguity occurs) if* μ_n *converges weakly to* μ, *where* μ_n *is the probability distribution of* X_n *and* μ *is the probability distribution of* X.

The following theorem provides various characterizations of the weak convergence. It is often referred to as the *Portmanteau theorem*.

Theorem A.1.4. *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$. *The following properties are equivalent:*

(i) μ_n *converges weakly to* μ,

(ii) for all $f \in \mathrm{Lip}_b(\mathsf{X})$, $\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$,

(iii) for all closed set F, $\limsup_n \mu_n(F) \leq \mu(F)$,

(iv) for all open set U, $\liminf_n \mu_n(U) \geq \mu(U)$,

(v) for all $B \in \mathcal{B}(\mathsf{X})$ *such that* $\mu(\partial B) = 0$, $\lim_n \mu_n(B) = \mu(B)$.

Let (Y, δ) be a metric space. For all measurable $h : \mathsf{X} \to \mathsf{Y}$, we denote

$$D_h \stackrel{\text{def}}{=} \{x \in \mathsf{X} : \ h \text{ is discontinuous at } x\} \ . \tag{A.2}$$

The following theorem is often referred to as the *continuous mapping theorem.*

Theorem A.1.5. *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$ *and* $h : \mathsf{X} \to \mathsf{Y}$ *be measurable. We assume that* μ_n *converges weakly to* μ *and that* $\mu(D_h) = 0$. *Then* $\mu_n \circ h^{-1}$ *converges weakly to* $\mu \circ h^{-1}$.

The weak convergence is equivalent to the convergence of integrals of bounded continuous functions. The case of unbounded continuous functions is treated in the following result.

Proposition A.1.6. *Assume that* μ_n *converges weakly to* μ. *Let* f *be a continuous function such that* $\lim_{a\to\infty} \sup_n \int_{|f|>a} |f| \, \mathrm{d}\mu_n = 0$. *Then* f *is* μ*-integrable and* $\int f d\mu_n \to \int f d\mu$.

We now provide a statement expressed with random variables for convenience and add the equivalent statement for the strong convergence and the convergence in probability. It is a direct application of Theorem A.1.5 and Theorem A.1.2.

Theorem A.1.7 (Continuous mapping theorem for the three convergences)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $h : \mathsf{X} \to \mathsf{Y}$ *measurable and define* D_h *as in (A.2). Assume that* $\mathbb{P}(X \in D_h) = 0$. *Then the following assertions hold.*

*(*i*) If* $X_n \xrightarrow{\text{a.s.}} X$, *then* $h(X_n) \xrightarrow{\text{a.s.}} h(X)$.

*(*ii*) If* $X_n \xrightarrow{P} X$, *then* $h(X_n) \xrightarrow{P} h(X)$.

*(*iii*) If* $X_n \Longrightarrow X$, *then* $h(X_n) \Longrightarrow h(X)$.

Let us recall briefly some standard results on the weak convergence, strong convergence and convergence in probability.

Theorem A.1.8. *Let* (X, d) *and* (Y, δ) *be two metric space. We equip* $\mathsf{X} \times \mathsf{Y}$ *with the metric* $d + \delta$. *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* Y_n, $n \geq 1$ *be random variables valued in* $(\mathsf{Y}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *The following assertions hold.*

*(*i*) If* $X_n \xrightarrow{\text{a.s.}} X$, *then* $X_n \xrightarrow{P} X$.

*(*ii*) If* $X_n \xrightarrow{P} X$, *then* $X_n \Longrightarrow X$.

*(*iii*) For all* $c \in \mathsf{X}$, $X_n \xrightarrow{P} c$ *if and only if* $X_n \Longrightarrow c$,

*(*iv*) Suppose that the spaces* (X, d) *and* (Y, δ) *coincide. If* $X_n \Longrightarrow X$ *and* $d(X_n, Y_n) \xrightarrow{P} 0$, *then* $Y_n \Longrightarrow X$.

*(*v*) For all* $c \in \mathsf{X}$, *if* $X_n \Longrightarrow X$ *and* $Y_n \xrightarrow{P} c$, *then* $(X_n, Y_n) \Longrightarrow (X, c)$.

*(*vi*) If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then* $(X_n, Y_n) \xrightarrow{P} (X, Y)$.

The following classical lemma can be useful.

Lemma A.1.9. *Let $(Z_{n,m})_{n,m\geq 1}$ be an array of random variables in X. Suppose that for all $m \geq 1$, $Z_{n,m}$ converges weakly to Z_m as $n \to \infty$ and that Z_m converges weakly to Z as $m \to \infty$. Let now $(X_n)_{n\geq 1}$ be random variables in X such that, for all $\epsilon > 0$,*

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}(d(X_n, Z_{m,n}) > \epsilon) = 0 \, .$$

Then X_n converges weakly to Z as $n \to \infty$.

Proof. Let $f \in \mathrm{Lip}_b(\mathsf{X})$ so that $|f(x) - f(y)| \leq K\, d(x, y)$ and $|f(x)| \leq C$ for all $x, y \in \mathsf{X}$. Then we write

$$\begin{aligned}\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right] = \mathbb{E}\left[f(X_n) - f(Z_{m,n})\right] \\ + \left[\mathbb{E}\left[f(Z_{m,n})\right] - \mathbb{E}\left[f(Z_m)\right]\right] + \left[\mathbb{E}\left[f(Z_m)\right] - \mathbb{E}\left[f(Z)\right]\right] . \quad \text{(A.3)}\end{aligned}$$

Then, for all $\epsilon > 0$, since $|f(X_n) - f(Z_{m,n})| \leq K\,\epsilon$ if $d(X_n, Z_{m,n}) \leq \epsilon$ and $|f(X_n) - f(Z_{m,n})| \leq C$ otherwise, we have,

$$\mathbb{E}\left[|f(X_n) - f(Z_{m,n})|\right] \leq K\,\epsilon + C\,\mathbb{P}(d(X_n, Z_{m,n}) > \epsilon)$$

By Theorem A.1.4 and using the assumptions of the lemma, we get that, for some large enough m,

$$\limsup_{n\to\infty} \left|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right]\right| \leq (K\,\epsilon + C)\epsilon + 0 + \epsilon \, .$$

Hence, since $\epsilon > 0$ can be taken arbitrarily small, $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(Z)\right]$ and we conclude with Theorem A.1.4. □

A.2 Some topology results

An important fact about the weak convergence on $\mathbb{M}_1(\mathsf{X})$ is that it is metrizable, provided that X is separable. This is shown in the two following results.

Let us denote by $\mathcal{S}$ the class of closed sets of X and, for $A \subset \mathsf{X}$ and $\alpha > 0$, $A^\alpha = \{x \in \mathsf{X},\, d(x, A) < \alpha\}$. A^α is an open set and $A^\alpha \downarrow \overline{A}$ if $\alpha \downarrow 0$. We set, for $\lambda, \mu \in \mathbb{M}_1(\mathsf{X})$,

$$\boldsymbol{\rho}(\lambda, \mu) = \inf\left\{\alpha > 0 \,:\, \lambda(F) \leq \mu(F^\alpha) + \alpha \text{ for all } F \in \mathcal{S}\right\} . \quad \text{(A.4)}$$

The following result shows that $\boldsymbol{\rho}$ is indeed a metric, which is not completely obvious from (A.4).

Lemma A.2.1. *$\boldsymbol{\rho}$ defined in (A.4) is a metric on $\mathbb{M}_1(\mathsf{X})$.*

The following result indicates that the metric $\boldsymbol{\rho}$ defines the topology of the weak convergence whenever (X, d) is separable.

Proposition A.2.2. *Assume that (X, d) is separable. Let $(\mu_n)_{n\in\mathbb{N}} \subset \mathbb{M}_1(\mathsf{X})$ and $\mu \in \mathbb{M}_1(\mathsf{X})$. Then $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ iff $\boldsymbol{\rho}(\mu_n, \mu) \to 0$. Moreover $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is separable.*

In the following, we assume that (X, d) is separable, so that, by Proposition A.2.2, $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is a separable metric space associated to the weak convergence. As a consequence, a subset $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is compact if it is sequentially compact.

The relative compactness of a subset of $\mathbb{M}_1(\mathsf{X})$ can be related to its *tightness*, that is, coarsely speaking, the property of all the measures of this subset to be almost supported on the same compact subset of X.

Definition A.2.1. *Let Γ be a subset of $\mathbb{M}_1(\mathsf{X})$.*

(i) We say that Γ is tight if for all $\epsilon > 0$, there exists a compact set $K \subset \mathsf{X}$ such that $\mu(K) \geq 1 - \epsilon$ for all $\mu \in \Gamma$.

(ii) We say that Γ is relatively compact if every sequence of elements in Γ contains a weakly convergent subsequence, or, equivalently if $\overline{\Gamma}$ is compact.

The following result is often referred to as the *Prokhorov theorem*.

Theorem A.2.3. *Let (X, d) be separable. Then if $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is tight, it is relatively compact.*

This theorem has the following converse result in the case where (X, d) is complete.

Theorem A.2.4. *Let (X, d) be separable and complete. If $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is relatively compact, then it is tight.*

Since singletons are compact, a direct but important consequence of this theorem is that any $\{\mu\} \subset \mathbb{M}_1(\mathsf{X})$ is tight.

Let us conclude this section with a last topological result.

Theorem A.2.5. *Let (X, d) be separable and complete. Then $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is separable and complete.*

Index

Bibliography

[1] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

[3] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, 2018.

[4] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[5] J. Jacod and P. Protter. *Probability essentials*. Universitext. Springer-Verlag, Berlin, second edition, 2003.

[6] H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[7] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[8] N. Young. *An introduction to Hilbert space*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1988.