

# MACS 201 : Hilbert spaces and probability

## 1 Hilbert spaces

**Def.** Let  $\mathcal{H}$  be a complex linear space. An **inner-product** on  $\mathcal{H}$  is a function  $\langle \cdot | \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbf{C}$  which satisfies the following properties :

- (i)  $\forall (x, y) \in \mathcal{H} \times \mathcal{H}, \langle x | y \rangle = \overline{\langle y | x \rangle},$
- (ii)  $\forall x, y, z \in \mathcal{H}, \forall (\alpha, \beta) \in \mathbf{C} \times \mathbf{C}, \langle \alpha x + \beta y | z \rangle = \alpha \langle x | z \rangle + \beta \langle y | z \rangle,$
- (iii)  $\forall x \in \mathcal{H}, (\langle x | x \rangle = 0) \iff (x = 0)$

Then  $\|\cdot\| : x \mapsto \sqrt{\langle x | x \rangle} \geq 0$  defines a norm on  $\mathcal{H}$ . Both are continuous.

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

- a) *Cauchy-Schwarz inequality* :  $|\langle x | y \rangle| \leq \|x\| \cdot \|y\|,$
- b) *triangular inequality* :  $|||x| - |y||| \leq \|x - y\| \leq \|x\| + \|y\|,$
- c) *Parallelogram inequality* :  $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$

**Def.** An inner-product space  $\mathcal{H}$  is called an **Hilbert space** if it is complete.

**Prop.** For all measured space  $(\Omega, \mathcal{F}, \mu)$ , the space  $L^2(\Omega, \mathcal{F}, \mu)$  endowed with  $\langle f | g \rangle = \int f \bar{g} d\mu$  is a Hilbert space.

**Def.** 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  $\forall s \in \mathcal{S}, x \perp s$ . Also we write  $\mathcal{S} \perp \mathcal{T}$  if all vectors in  $\mathcal{S}$  are orthogonal to  $\mathcal{T}$ .

**Not.** If  $\mathcal{H} = \mathcal{A} + \mathcal{B}$  and  $\mathcal{A} \perp \mathcal{B}$  we will denote  $\mathcal{H} = \mathcal{A} \oplus \mathcal{B}$ .

**Def.** Let  $\mathcal{E}$  be a subset of an Hilbert space  $\mathcal{H}$ . The orthogonal set of  $\mathcal{E}$  is  $\mathcal{E}^\perp = \{x \in \mathcal{H} \mid \forall y \in \mathcal{E}, \langle x | y \rangle = 0\}$ .

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

### Orthogonal and orthonormal bases

**Def.** Let  $E$  be a subset of  $\mathcal{H}$ . It is an orthogonal set if for all  $(x, y) \in E \times E, x \neq y, x \perp y$ . If moreover  $\forall x \in E, \|x\| = 1$ , we say that  $E$  is orthonormal.

**Th.** Let  $(e_i)_{i \geq 1}$  be an orthonormal sequence of an Hilbert space  $\mathcal{H}$  and let  $(\alpha_i)_{i \geq 1} \in \mathbf{C}^{\mathbf{N}}$ . The series  $\sum_{i=1}^{\infty} \alpha_i e_i$  converges in  $\mathcal{H}$  if and only if  $\sum_i |\alpha_i|^2 < \infty$ , in which case  $\|\sum_{i=1}^{\infty} \alpha_i e_i\|^2 = \sum_{i=1}^{\infty} |\alpha_i|^2$ .

**Prop.** Let  $x \in \mathcal{H}$  (Hilbert space) and  $E = \{e_1, \dots, e_n\}$  a finite orthonormal set of vectors. Then  $\|x - \sum_{k=1}^n \langle x | e_k \rangle e_k\|^2 = \|x\|^2 - \sum_{k=1}^n |\langle x | e_k \rangle|^2 = \inf\{\|x - y\|^2, y \in \text{Span}(e_1, \dots, e_n)\}.$

**Cor** (Bessel inequality). Let  $(e_i)_{i \geq 1}$  be an orthonormal sequence of a Hilbert space  $\mathcal{H}$ . Then  $\forall x \in \mathcal{H}, \sum_{i=1}^{\infty} |\langle x | e_i \rangle|^2 \leq \|x\|^2$ .

**Def.** A subset  $E$  of a Hilbert space  $\mathcal{H}$  is said **dense** if  $\overline{\text{Span}(E)} = \mathcal{H}$ . An orthonormal dense sequence is called a Hilbert basis.

**Prop.** Consider the measured space  $(\Omega, \mathcal{F}, \mu)$  and the Hilbert space  $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mu)$ ,  $\overline{\text{Span}(\mathbf{1}_A, A \in \mathcal{F})} = \mathcal{H}$ .

**Th.** Let  $(e_i)_{i \geq 1}$  be a Hilbert basis of the Hilbert space  $\mathcal{H}$ . Then  $\forall x \in \mathcal{H}, x = \sum_{i=1}^{\infty} \langle x | e_i \rangle e_i$ .

**Th.** Let  $(e_i)_{i \geq 1}$  be an orthonormal sequence of the Hilbert space  $\mathcal{H}$ . The following assertions are equivalent :

- (i)  $(e_i)_{i \geq 1}$  is a Hilbert basis,
- (ii) if some  $x \in \mathcal{H}$  satisfies  $\forall i \geq 1, \langle x | e_i \rangle = 0$  then  $x = 0$ ,
- (iii)  $\forall x \in \mathcal{H}, \|x\|^2 = \sum_{i=1}^{\infty} |\langle x | e_i \rangle|^2$ .

**Th.** A Hilbert space  $\mathcal{H}$  is separable (i.e. contains a countable dense subset) if and only if it admits a Hilbert basis.

### Fourier series

Let  $\psi_n : x \mapsto \frac{1}{\sqrt{2\pi}} e^{inx}, n \in \mathbf{Z}$ . Let  $L^1(\mathbf{T})$  denote the set of  $2\pi$ -periodic locally integrable functions. For  $f \in L^1(\mathbf{T})$ , set  $\forall n \in \mathbf{N}, f_n = \sum_{k=-n}^n (\int_{\mathbf{T}} f \bar{\psi}_k) \psi_k$ .

**Th.** Let  $f$  be a continuous  $2\pi$ -periodic function. Then the Cesaro sequence  $\frac{1}{n} \sum_{k=0}^{n-1} f_k$  converges uniformly to  $f$ .

**Cor.** Let  $\mu$  be a finite measure on the Borel sets of  $\mathbf{T} = \mathbf{R}/(2\pi\mathbf{Z})$ . The sequence  $(\phi_n)_{n \in \mathbf{Z}}$  is dense in the Hilbert space  $L^2(\mathbf{T}, \mathcal{B}(\mathbf{T}), \mu)$ .

**Cor.** The sequence  $(\phi_n)_{n \in \mathbf{Z}}$  is a Hilbert basis in  $L^2(\mathbf{T})$ . In particular,  $\forall f \in L^2(\mathbf{T}), f = \sum_{k=-\infty}^{\infty} \alpha_k \phi_k$  with  $\alpha_k = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{T}} f(x) e^{-ikx} dx$  when the infinite sum converges in  $L^2(\mathbf{T})$ . The Parseval identity then reads  $\int_{\mathbf{T}} |f(x)|^2 dx = \sum_{k=-\infty}^{\infty} |\alpha_k|^2$ .

## Projection and orthogonality principle

**Th** (Projection theorem). Let  $\mathcal{E}$  be a closed convex subset of a Hilbert space  $\mathcal{H}$  and  $x \in \mathcal{H}$ . Then the following holds :

- (i) There exists a unique vector  $\text{proj}(x | \mathcal{E}) \in \mathcal{E}$  such that  $\|x - \text{proj}(x | \mathcal{E})\| = \inf_{w \in \mathcal{E}} \|x - w\|$ .
- (ii) If moreover  $\mathcal{E}$  is a linear subspace,  $\text{proj}(x | \mathcal{E})$  is the unique  $\hat{x} \in \mathcal{E}$  such that  $x - \hat{x} \in \mathcal{E}^\perp$ . It is called the orthogonal projection of  $x$  onto  $\mathcal{E}$ .

**Prop.** Let  $\mathcal{H}$  be a Hilbert space and  $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2$  closed subspaces of  $\mathcal{H}$ . Then the following assertions hold.

- (i) Suppose that  $\mathcal{E} = \overline{\text{Span}((e_k)_{k \in \mathbb{N}})}$  with  $(e_k)$  being an orthonormal sequence. Then  $\text{proj}(h | \mathcal{E}) = \sum_{k=0}^{\infty} \langle h | e_k \rangle e_k$ .
- (ii) The function  $\text{proj}(\cdot | \mathcal{H}) : x \mapsto \text{proj}(x | \mathcal{E})$  is linear and continuous on  $\mathcal{H}$ .
- (iii)  $\|x\|^2 = \|\text{proj}(x | \mathcal{E})\|^2 + \|x - \text{proj}(x | \mathcal{E})\|^2$
- (iv)  $(x \in \mathcal{E} \iff \text{proj}(x | \mathcal{E}) = x)$  and  $(x \in \mathcal{E}^\perp \iff \text{proj}(x | \mathcal{E}) = 0)$
- (v) If  $\mathcal{E}_1 \subset \mathcal{E}_2$  then  $\forall x \in \mathcal{H}, \text{proj}(\text{proj}(x | \mathcal{E}_2) | \mathcal{E}_1) = \text{proj}(x | \mathcal{E}_1)$
- (vi) If  $\mathcal{E}_1 \perp \mathcal{E}_2$  then  $\forall x \in \mathcal{H}, \text{proj}\left(x | \mathcal{E}_1 \oplus \mathcal{E}_2\right) = \text{proj}(x | \mathcal{E}_1) + \text{proj}(x | \mathcal{E}_2)$

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

1. Denote  $M_{-\infty} = \bigcap_n M_n$ . Then  $\forall h \in \mathcal{H}, \text{proj}(h | M_{-\infty}) = \lim_{n \rightarrow -\infty} \text{proj}(h | M_n)$ .
2. Denote  $M_\infty = \overline{\bigcup_n M_n}$ . Then  $\forall h \in \mathcal{H}, \text{proj}(h | M_\infty) = \lim_{n \rightarrow \infty} \text{proj}(h | M_n)$ .

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

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

**Th** (Riesz representation theorem). Let  $\mathcal{H}$  be a Hilbert space. Then  $F : \mathcal{H} \rightarrow \mathbb{C}$  is a non-zero continuous linear form if and only if  $\exists x \in \mathcal{H} \setminus \{0\}, \forall y \in \mathcal{H}, F(y) = \langle y | x \rangle$ .

## Unitary Operator

**Def.** Let  $\mathcal{H}$  and  $\mathcal{I}$  be two Hilbert spaces. An **isometric** operator  $S : \mathcal{H} \rightarrow \mathcal{I}$  is a linear application such that  $\forall (v, w) \in \mathcal{H}^2, \langle Sv | Sw \rangle_{\mathcal{I}} = \langle v | w \rangle_{\mathcal{H}}$ . If it is moreover bijective, it is a **unitary** operator. In this case we also says that  $\mathcal{H}$  and  $\mathcal{I}$  are isomorphic.

**Th.** 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  $\mathbb{C}^n$ .

**Th.** Let  $\mathcal{H}$  and  $\mathcal{I}$  be two Hilbert spaces and  $\mathcal{G}$  a subspace of  $\mathcal{H}$ .

- (i) Let  $S : \mathcal{G} \rightarrow \mathcal{I}$  be isometric on  $\mathcal{G}$ . Then  $S$  admits a unique isometric extension  $\bar{S} : \bar{\mathcal{G}} \rightarrow \mathcal{I}$  and  $\bar{S}(\bar{\mathcal{G}})$  is the closure of  $S(\mathcal{G})$  in  $\mathcal{I}$ .
- (ii) Let  $(v_t)_{t \in T}$  and  $(w_t)_{t \in T}$  be two set of vectors in  $\mathcal{H}$  and  $\mathcal{I}$  indexed by an arbitrary index set  $T$ . Suppose  $\forall (s, t) \in T^2, \langle v_t | v_s \rangle_{\mathcal{H}} = \langle w_t | w_s \rangle_{\mathcal{I}}$ . Then, there exists a unique isometric operator  $S : \overline{\text{Span}((v_t)_{t \in T})} \rightarrow \overline{\text{Span}((w_t)_{t \in T})}$  such that  $\forall t \in T, Sv_t = w_t$ . Moreover,  $S(\overline{\text{Span}((v_t)_{t \in T})}) = \overline{\text{Span}((w_t)_{t \in T})}$ .

## 2 Probability

Let  $(\Omega, \mathcal{F}, \mathbf{P})$  be a probability space.

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

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

**Not.** For  $p > 0$ , we denote by  $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbf{P})$  the space of random variables  $X$  such that  $\mathbf{E}(|X|^p) < \infty$  and by  $L^p(\Omega, \mathcal{F}, \mathbf{P})$  the one identifying random variables that are equal  $\mathbf{P}$ -a.s.

## Conditional calculus

**Lem.** Let  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbf{P})$  and  $\mathcal{G}$  a sub- $\sigma$ -field of  $\mathcal{F}$ . Then there exists  $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbf{P})$  such that

$$\forall A \in \mathcal{G}, \mathbf{E}(X \mathbf{1}_A) = \mathbf{E}(Y \mathbf{1}_A) \quad (1)$$

Moreover the following assertions hold.

- (i) If  $Y' \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbf{P})$  also satisfies (1) then  $Y' = Y$   $\mathbf{P}$ -a.s.
- (ii) If  $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$ , then  $Y = \text{proj}(X | L^2(\Omega, \mathcal{G}, \mathbf{P}))$ .
- (iii) (1) continues to hold extended as  $\mathbf{E}(XZ) = \mathbf{E}(YZ)$  for all  $\mathcal{G}$ -measurable r.v.  $Z$  such that  $\mathbf{E}(|XZ|) < \infty$ .

**Def.** Let  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbf{P})$  and  $\mathcal{G}$  a sub- $\sigma$ -field of  $\mathcal{F}$ . The unique  $Y \in L^1(\Omega, \mathcal{G}, \mathbf{P})$  defined by (1) is called the **conditional expectation** of  $X$  given  $\mathcal{G}$ , and denoted by  $Y = \mathbf{E}(x | \mathcal{G})$ .

**Prop.** Suppose that  $X, Y, Z, (X_n)_{n \geq 1} \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ . The following hold  $\mathbf{P}$ -a.s.

- (i) (linearity)  $\forall a, b \in \mathbf{R}, \mathbf{E}(aX + bY \mid \mathcal{G}) = a\mathbf{E}(X \mid \mathcal{G}) + b\mathbf{E}(Y \mid \mathcal{G})$
- (ii) If  $X$  is  $\mathcal{G}$ -measurable,  $\mathbf{E}(X \mid \mathcal{G}) = X$
- (iii) If  $\mathcal{G} = \{\emptyset, \Omega\}$  is the trivial  $\sigma$ -field, then  $\mathbf{E}(X \mid \mathcal{G}) = \mathbf{E}(X)$
- (iv) If  $X$  is independent of  $\mathcal{G}$  then  $\mathbf{E}(X \mid \mathcal{G}) = \mathbf{E}(X)$
- (v) (positivity) If  $X \leq Y$  then  $\mathbf{E}(X \mid \mathcal{G}) \leq \mathbf{E}(Y \mid \mathcal{G})$
- (vi)  $\mathbf{E}(X \mid \mathcal{G}) \vee \mathbf{E}(Y \mid \mathcal{G}) \leq \mathbf{E}(X \vee Y \mid \mathcal{G})$ ,  $\mathbf{E}(X \mid \mathcal{G})_+ \leq \mathbf{E}(X_+ \mid \mathcal{G})$  and  $|\mathbf{E}(X \mid \mathcal{G})| \leq \mathbf{E}(|X| \mid \mathcal{G})$
- (vii) (tower property) If  $\mathcal{H}$  is a sub- $\sigma$ -field of  $\mathcal{F}$  such that  $\mathcal{G} \subset \mathcal{H}$  then  $\mathbf{E}(\mathbf{E}(X \mid \mathcal{H}) \mid \mathcal{G}) = \mathbf{E}(X \mid \mathcal{G})$
- (viii) The expectation is not modified by conditional expectation :  $\mathbf{E}(\mathbf{E}(X \mid \mathcal{G})) = \mathbf{E}(X)$
- (ix) If  $X$  is  $\mathcal{G}$ -measurable and  $XY \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ , then  $\mathbf{E}(XY \mid \mathcal{G}) = X \cdot \mathbf{E}(Y \mid \mathcal{G})$

**Def.** Let  $Y$  be a r.v. and  $\sigma(X)$  the sub- $\sigma$ -field generated by a r.v.  $X$ . If  $\mathbf{E}(Y \mid \sigma(X))$  is well-defined, it is written as  $\mathbf{E}(Y \mid X)$  and is called the **conditional expectation** of  $Y$  given  $X$ .

**Def.** Let  $\mathcal{G}$  be a sub- $\sigma$ -field of  $\mathcal{F}$ . For any event  $A \in \mathcal{F}$ , we denote  $\mathbf{P}(A \mid \mathcal{G}) = \mathbf{E}(1_A \mid \mathcal{G})$ . The mapping  $A \mapsto \mathbf{P}(A \mid \mathcal{G})$  is called a **version of the conditional probability** of  $A$  given  $\mathcal{G}$ .

**Def.** Let  $\mathcal{G}$  be a sub- $\sigma$ -field of  $\mathcal{F}$ . A **regular version** of the conditional probability of  $\mathbf{P}$  given  $\mathcal{G}$  is a function  $\mathbf{P}^{\mathcal{G}}: \Omega \times \mathcal{F} \rightarrow [0; 1]$  such that

- (i) For all  $A \in \mathcal{F}$ ,  $\mathbf{P}^{\mathcal{G}}(A): \omega \mapsto \mathbf{P}^{\mathcal{G}}(\omega, A)$  is  $\mathcal{G}$ -measurable and is a version of the conditional probability of  $A$  given  $\mathcal{G}$ ,  $\mathbf{P}^{\mathcal{G}}(A) = \mathbf{P}(A \mid \mathcal{G})$ .
- (ii) For all  $\omega \in \Omega$ , the mapping  $A \mapsto \mathbf{P}^{\mathcal{G}}(\omega, A)$  is a probability on  $\mathcal{F}$ .

**Lemma.** Let  $\mathbf{P}^{\mathcal{G}}$  be a regular version of the conditional probability of  $\mathbf{P}$  given  $\mathcal{G}$  and let  $Y \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ . Then  $\mathbf{E}(Y \mid \mathcal{G}) = \mathbf{E}^{\mathcal{G}}(Y)$   $\mathbf{P}$ -a.s., with  $\mathbf{E}^{\mathcal{G}}(Y): \omega \mapsto \int Y(\omega') \mathbf{P}^{\mathcal{G}}(\omega, d\omega')$ .

**Def.** Let  $\mathcal{G}$  be a sub- $\sigma$ -field of  $\mathcal{F}$ . Let  $(Y, \mathcal{Y})$  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  $\mathbf{P}^{Y|\mathcal{G}}: \Omega \times \mathcal{Y} \rightarrow [0; 1]$  such that

- (i) For all  $A \in \mathcal{Y}$ ,  $\omega \mapsto \mathbf{P}^{Y|\mathcal{G}}(\omega, A)$  is  $\mathcal{G}$  measurable and is a version of conditional distribution of  $Y$  given  $\mathcal{G}$ ,  $\mathbf{P}^{Y|\mathcal{G}}(\cdot, A) = \mathbf{P}(Y \in A \mid \mathcal{G})$   $\mathbf{P}$ -a.s.
- (ii) For every  $\omega$ ,  $A \mapsto \mathbf{P}^{Y|\mathcal{G}}(\omega, A)$  is a probability on  $\mathcal{Y}$ .

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

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

$Q$  is said to be finite if  $\forall x \in X, Q(x, Y) < \infty$ . It is called a **probability kernel** if  $\forall x \in X, Q(x, Y) = 1$ . It is called a **Markov kernel** if it is a probability kernel on  $X \times \mathcal{X}$ .

**Def.** Let  $X$  and  $Y$  be random variables with values in the measure spaces  $(X, \mathcal{X})$  and  $(Y, \mathcal{Y})$  respectively. A **regular version of the conditional distribution of  $Y$  given  $X$**  is a probability kernel  $\mathbf{P}^{X|Y}: X \times \mathcal{Y} \rightarrow [0; 1]$  such that  $\forall A \in \mathcal{Y}, \mathbf{P}^{X|Y}(X, A) = \mathbf{P}(Y \in A \mid X)$   $\mathbf{P}$ -a.s.

**Th.** Let  $\mathcal{G}$  be sub- $\sigma$ -field of  $\mathcal{F}$ . Let  $d \geq 1$  and  $Y$  be an  $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ -valued random variable. Then, there exists a regular version of the conditional distribution of  $Y$  given  $\mathcal{G}$ ,  $\mathbf{P}^{Y|\mathcal{G}}$ , and this version is unique in the sense that for any other regular version  $\bar{\mathbf{P}}^{Y|\mathcal{G}}$  of this distribution, for  $\mathbf{P}$ -almost every  $\omega$  it holds that  $\forall F \in \mathcal{F}, \mathbf{P}^{Y|\mathcal{G}}(\omega, F) = \bar{\mathbf{P}}^{Y|\mathcal{G}}(\omega, F)$ . Moreover, if  $\mathcal{G} = \sigma(X)$  for some r.v.  $X$  with values in a measurable space  $(X, \mathcal{X})$ , there also exists a unique regular version (hence a probability kernel)  $\mathbf{P}^{Y|X}$ .

**Lemma.** Let  $\mathbf{P}^{Y|X}$  be a regular version of the conditional expectation of  $Y$  given  $X$ . Then, for any real-valued measurable function  $g$  on  $Y$  such that  $\mathbf{E}(|g(Y)|) < \infty$ , we have  $\mathbf{E}(g(Y) \mid X) = \int g(Y) \mathbf{P}^{Y|X}(X, dy)$ ,  $\mathbf{P}$ -a.s.

**Prop.** Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two jointly Gaussian vectors, respectively valued in  $\mathbf{R}^p$  and  $\mathbf{R}^q$ . Then the following holds.

- (i) If  $\text{Cov}(\mathbf{Y})$  is invertible, then  $\hat{\mathbf{X}} := \text{proj}(\mathbf{X} \mid \text{Span}(1, \mathbf{Y}))$  is given by  $\hat{\mathbf{X}} = \mathbf{E}(\mathbf{X}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) \text{Cov}(\mathbf{Y})^{-1}(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))$ , and  $\text{Cov}(\mathbf{X} - \hat{\mathbf{X}}) = \text{Cov}(\mathbf{X}) - \text{Cov}(\mathbf{X}, \mathbf{Y}) \text{Cov}(\mathbf{Y})^{-1} \text{Cov}(\mathbf{Y}, \mathbf{X})$ .
- (ii) We have  $\mathbf{E}(\mathbf{X} \mid \mathbf{Y}) = \text{proj}(\mathbf{X} \mid \text{Span}(1, \mathbf{Y}))$ .
- (iii) Let  $\hat{\mathbf{X}} = \mathbf{E}(\mathbf{X} \mid \mathbf{Y})$ . Then  $\text{Cov}(\mathbf{X} - \hat{\mathbf{X}}) = \mathbf{E}(\mathbf{X}(\mathbf{X} - \hat{\mathbf{X}})^T) = \mathbf{E}((\mathbf{X} - \hat{\mathbf{X}})\mathbf{X}^T)$  and  $\mathbf{P}^{Y|X}(\mathbf{X}, \cdot) = \mathcal{N}(\hat{\mathbf{X}}, \text{Cov}(\mathbf{X} - \hat{\mathbf{X}}))$ .



### Radon-Nikodym derivative

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

**Def.** Let  $\lambda$  be a measure on  $(\Omega, \mathcal{F})$ . We say that a  $\sigma$ -finite measure  $\mu$  is **absolutely continuous** with respect to  $\lambda$  or that  $\lambda$  dominates  $\mu$  and we write  $\mu \ll \lambda$  if  $\forall A \in \mathcal{F}, (\lambda(A) = 0) \implies (\mu(A) = 0)$ .

**Th (Radon-Nikodym theorem).** Let  $\lambda, \mu \in \mathbf{M}_+(\Omega, \mathcal{F})$  be  $\sigma$ -finite measures such that  $\mu \ll \lambda$ . Then, there exists a non-negative Borel function  $\phi$  such that  $\forall A \in \mathcal{F}, \mu(A) = \int_A \phi d\lambda$ .

**Def.** Let  $(X, Y)$  be two random elements admitting a density  $f$  with respect to measure  $\xi \otimes \xi'$  on  $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ . Then the function  $(x, y) \mapsto f(y | x) = \frac{f(x, y)}{\int f(x, y') d\xi'(y')}$  is called the **conditional density** of  $Y$  given  $X$ .

**Th.** Let  $(X, Y)$  be two random elements admitting a density  $f: X \times Y \rightarrow \mathbf{R}_+$  with respect to  $\xi \otimes \xi'$  on  $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ . Then,  $\forall x \in X, \forall A \in \mathcal{Y}, \mathbf{P}^{Y|X}(x, A) = \int_A f(y | x) \xi'(dy)$ .

**Lem.** Let  $P$  and  $Q$  be two probabilities on the measurable space  $(\Omega, \mathcal{F})$  and let  $\nu \in \mathbf{M}_+(\Omega, \mathcal{F})$  dominate both  $P$  and  $Q$  (e.g.  $\nu = P + Q$ ). Let  $f_P$  and  $f_Q$  denote the densities of  $P$  and  $Q$  with respect to  $\nu$ . Then,  $\text{KL}(P||Q) = \int \ln \left( \frac{f_P}{f_Q} \right) dP$  is always well defined and takes values in  $[0; \infty]$ . Moreover we have :

(i) If  $Q$  does not dominate  $P$  then  $\text{KL}(P||Q) = \infty$ .

(ii) If  $P \ll Q$  then  $\text{KL}(P||Q) = \int \ln \left( \frac{dP}{dQ} \right) dP$  (may be finite or infinite).

(iii) We have  $\text{KL}(P||Q) = 0 \iff P = Q$ .

**Def.** The quantity  $\text{KL}(P||Q)$  is called the **Kullback-Leibler divergence** between  $P$  and  $Q$ .

**Th.** Let  $P$  and  $Q$  be two probabilities on the measurable space  $(\Omega, \mathcal{F})$  and  $X$  a measurable mapping from  $(\Omega, \mathcal{F})$  to  $(X, \mathcal{X})$ . Then we have  $\text{KL}(P^X||Q^X) \leq \text{KL}(P||Q)$ .

**Rem.** Recall that  $\forall A \in \mathcal{X}, P^X(A) = \int_{X^{-1}(A)} dP$  while  $\forall F \in \mathcal{F}, P(F) = \int_F dP$ .

## 3 Mathematical statistics

### Statistical modeling

**Def.** 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  $(X, \mathcal{X})$ . We say that  $\mathcal{P}$  is a **statistical model** for the observation variable  $X$  and denote  $\mathcal{P}^X = (P^X)_{P \in \mathcal{P}}$  the corresponding collection of probability distributions.

It is usual in statistics to consider  $\Omega = X, \mathcal{F} = \mathcal{X}$  and  $X(\omega) = \omega$ , in which case  $\forall P \in \mathcal{P}, P = P^X$ .

**Def.** Let  $\nu \in \mathbf{M}_+(X, \mathcal{X})$  and  $\mathcal{P}$  be a statistical model for  $X$ . We say that  $\mathcal{P}$  is a  $\nu$ -dominated model for  $X$ , or that  $\mathcal{P}^X$  is  $\nu$ -dominated, if  $\forall P \in \mathcal{P}, P^X \ll \nu$ .

**Lem (Halmos and Savage).** Let  $\nu \in \mathbf{M}_+(X, \mathcal{X})$ . Consider a  $\nu$ -dominated model  $\mathcal{P}$  for the variable  $X$ . Then there exists a countable collection  $(P_n)_{n \geq 1}$  in  $\mathcal{P}$  such that  $\mathcal{P}^X$  is also dominated by  $\mu = \sum_{n \geq 1} 2^{-n} P_n^X$ .

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

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

**Def.** Let  $\mathcal{P}$  be a statistical model for  $X$ . A **statistic** in this context is any random variable  $T$  valued in  $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$  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  $t(P) \in \mathbf{R}^d$ , it is called an **estimator** of  $t(P)$ . In this case, the **bias** of  $T$  for estimating  $t(P)$  is defined as  $\text{Bias}(T, P) = \int T dP - t(P)$  whenever  $\int |T| dP < \infty$ . We say that  $T$  is an **unbiased estimator** of  $t(P)$  if  $\forall P \in \mathcal{P}, \int T dP = t(P)$ . The **quadratic risk** or **mean squared error** (in the case  $d = 1$ ) is defined by  $\text{MSE}(T, P) = \int (T - t(P))^2 dP = \text{Var}(T) + \text{Bias}(T, P)^2$ .

**Def.** Let  $T$  be a statistic valued in  $(\mathbf{R}^d, \mathbf{R}^{\mathcal{D}})$  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  $Q \subset \mathbf{R}^d \times \mathcal{X}$  such that, for all  $P \in \mathcal{P}, Q$  is a regular version of  $P^{X|T}$ .

**Lem.** Let  $S$  be a sufficient statistic associated to the Markov kernel  $Q$  and let  $T = g(X)$  be an unbiased estimator of the parameter  $t(P)$  (both real valued). Define  $T^R = \int g(x) Q(S, dx)$ . 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,  $\forall P \in \mathcal{P}, \text{MSE}(T^R, P) \leq \text{MSE}(T, P)$ .

**Th (Fisher Factorization theorem).** Let  $\nu \in \mathbf{M}_+(X, \mathcal{X})$ . Consider a  $\nu$ -dominated model  $\mathcal{P}$  for  $X$  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  $X$  such that  $\forall P \in \mathcal{P}$ , there exists a Borel function  $f_P: \mathbf{R}^d \rightarrow \mathbf{R}_+$  such that  $\frac{dP^X}{d\nu} = h \cdot f_P \circ g$ .

**Def.** Consider a  $\nu$ -dominated model  $\mathcal{P}$  for  $X$ . For all  $P \in \mathcal{P}$ , let us denote by  $f_P$  the density of  $P^X$  with respect to  $\nu$ . The **likelihood function** is defined as  $P \mapsto f_P \circ X$  on  $P \in \mathcal{P}$ .

Then,  $f_{P_1}(X) \geq f_{P_2}(X)$  is an indication that  $\text{KL}(P_*^X \| P_1^X) \leq \text{KL}(P_*^X \| P_2^X)$  where  $P_*$  is the true distribution of  $X$ .

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

With a parametric model we define the likelihood function directly on  $\Theta$ ,  $\theta \mapsto f_\theta \circ X$  where  $f_\theta$  denotes the density of  $P_\theta$  with respect to  $\nu$ .

**Def.** A statistic  $\hat{\theta}_n$  valued in  $\Theta$  such that  $f_{\hat{\theta}_n} \circ X = \max_{\theta \in \Theta} f_\theta \circ X$  is called a **maximum likelihood estimator (MLE)**.

### Statistical testing

We define two hypothesis, respectively called the *null hypothesis* and the *alternative hypothesis*.

- $(H_0)$  the observation variable  $X$  has distribution  $P^X$  with  $P \in \mathcal{P}_0$ ,
- $(H_1)$   $X$  has distribution  $P^X$  with  $P \in \mathcal{P}_1$ ,

with  $\{\mathcal{P}_0, \mathcal{P}_1\}$  a partition of a statistical model  $\mathcal{P}$ .  $(H_i)$  is simple if  $\mathcal{P}_i$  reduces to one point.

**Def.** A **statistical test** is a statistic  $\delta$  with values in  $\{0, 1\}$ . If  $\delta = 0$  we say that we accept  $(H_0)$ . Otherwise we reject it.

To evaluate the performance of a test  $\delta$ , two type of risks are considered :

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

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

**Def.** Let  $\alpha \in [0; 1]$ . We say that a test  $\delta$  is of level  $\alpha$  if  $\sup_{P \in \mathcal{P}_0} P(\delta = 1) \leq \alpha$ . We say that  $\delta$  is uniformly more powerful than  $\delta'$  for level  $\alpha$  if both are of level  $\alpha$  and  $\forall P \in \mathcal{P}_1, P(\delta = 1) \geq P(\delta' = 1)$ .

### Simple hypotheses

We consider  $\mathcal{P}_0 = \{P_0\}$  and  $\mathcal{P}_1 = \{P_1\}$ , with  $f_0$  and  $f_1$  the densities of  $P_0^X$  and  $P_1^X$  with respect to a common dominating measure.

**Def.** 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  $t$ .

**Th.** Denote by  $T$  the likelihood ratio corresponding to  $\mathcal{P}_0$  and  $\mathcal{P}_1$ . 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  $\delta'$  for the level  $\alpha_t$ . Moreover, if  $\delta'$  is of level  $\alpha_t$  and as powerful as  $\delta$ , then they coincide on the set  $\{T \neq t\}$   $P_i$ -a.s. for  $i \in \{0, 1\}$ .

### Fisher information matrix

We consider a parametric  $\nu$ -dominated model  $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$  for the observation variable  $X$  valued in  $(X, \mathcal{X})$ , and denote by  $f_\theta$  the density of  $P_\theta$  with respect to  $\nu$ . We assume that  $\Theta$  is an open subset of  $\mathbf{R}^n$  and denote by  $\|f\| := \left( \int_X |f(x)|^2 \nu(dx) \right)^{\frac{1}{2}}$  the norm of the Hilbert space  $L^2(X, \mathcal{X}, \nu)$ . Observe that  $\forall \theta \in \Theta, \xi_\theta = \sqrt{f_\theta} \in L^2(X, \mathcal{X}, \nu)$ .

**Def.** We say that  $\mathcal{P}$  is **Hellinger differentiable** at  $\theta$  if  $\theta' \mapsto \xi_{\theta'}$  defined from  $\Theta \rightarrow L^2(X, \mathcal{X}, \nu)$  admits a derivative at  $\theta$  :  $\exists \dot{\xi}_\theta \in (L^2(X, \mathcal{X}, \nu))^d, \lim_{\theta' \rightarrow \theta} \frac{1}{|\theta' - \theta|} \|\xi_{\theta'} - \xi_\theta - \dot{\xi}_\theta^T(\theta' - \theta)\| = 0$ .

**Lem.** Let  $\theta \in \Theta$  and  $V \subset \Theta$  be a neighborhood of  $\theta$ . Suppose that for  $\nu$ -a.e.  $x$  and all  $\theta' \in V$ , we can write  $\xi_{\theta'}(x) = \xi_\theta(x) + \int_{t=0}^1 g_{t\theta' + (1-t)\theta}^T(x)(\theta' - \theta) dt$ , where, for all  $x \in X$ ,  $g$  satisfies one of the following assertions,

- (i) we have  $\lim_{\epsilon \downarrow 0} \left\| \sup_{|\theta' - \theta| \leq \epsilon} |g_{\theta'} - g_\theta| \right\| = 0$ ,
- (ii) for  $\nu$ -a.e.  $x$ ,  $\theta' \mapsto g_{\theta'}(x)$  is continuous and  $\exists \epsilon > 0, \left\| \sup_{|\theta' - \theta| \leq \epsilon} |g_{\theta'}| \right\| < \infty$ .

Then  $\mathcal{P}$  is Hellinger differentiable at  $\theta$  with derivative  $g_\theta$ .

The derivative of  $\theta \mapsto \ln f_\theta(X)$  is called the score function.

**Lem.** Suppose that  $A := \{f_\theta > 0\}$  does not depend on  $\theta$  and  $\forall x \in A, \theta \mapsto \ln f_\theta(x)$  is continuously differentiable on  $\Theta$  with derivative  $\theta \mapsto \dot{l}_\theta(x)$ . Suppose moreover that  $\forall \theta \in \Theta$  there exists a neighborhood  $V$  of  $\theta$  such that  $\int \sup_{\theta' \in V} \left( |\dot{l}_\theta(x)|^2 f_\theta(x) \right) \nu(dx) < \infty$ . Then  $\mathcal{P}$  is Hellinger differentiable with Hellinger derivative given by  $\dot{\xi}_\theta(x) = \frac{1}{2} \dot{l}_\theta(x) \xi_\theta(x) \mathbf{1}_A(x)$ .

**Def.** Let  $\mathcal{P}$  be Hellinger differentiable with Hellinger derivative  $\dot{\xi}_\theta$ . The **Fisher information matrix** is defined as  $\mathcal{I}(\theta) := 4 \int_X \dot{\xi}_\theta(x) \dot{\xi}_\theta(x)^T \nu(dx)$ .

With the conditions of the previous lemma we have  $\mathcal{I}(\theta) = \mathbf{E}_\theta \left[ (\dot{l}_\theta(X))^2 \right]$ .

**Th.** Let  $\mathcal{P}$  be Hellinger differentiable with Hellinger derivative  $\dot{\xi}_\theta$ . Let  $T = g(X)$  be a scalar statistic such that, for some  $\epsilon > 0$ ,  $\sup_{|\theta' - \theta| \leq \epsilon} \mathbf{E}_\theta(T^2) < \infty$ . Define  $\psi: \theta \rightarrow \mathbf{E}_\theta(T)$ . Then  $\psi$  is differentiable at  $\theta$  and, if  $\mathcal{I}(\theta)$  is positive definite, we have  $\text{Var}_\theta(T) \geq \dot{\psi}(\theta)^\top \mathcal{I}(\theta)^{-1} \dot{\psi}(\theta)$ .

**Def.** Let  $T$  be as in the previous theorem. If  $\forall \theta \in \Theta$ ,  $\text{Var}_\theta(T) = \dot{\psi}(\theta)^\top \mathcal{I}(\theta)^{-1} \dot{\psi}(\theta)$ , we say that  $T$  is an efficient estimator of  $\psi(\theta)$ .

## 4 Random processes

### Random processes

We consider a probability space  $(\Omega, \mathcal{F}, \mathbf{P})$ , an index  $T$  and a measurable space  $(X, \mathcal{X})$  called the observation space.

**Def.** A **random process** defined on  $(\Omega, \mathcal{F}, \mathbf{P})$ , indexed on  $T$  and valued in  $(X, \mathcal{X})$  is a collection  $(X_t)_{t \in T}$  of r.v. defined on  $(\Omega, \mathcal{F}, \mathbf{P})$  and taking values in  $(X, \mathcal{X})$ .

**Def.** For each  $\omega \in \Omega$ , the application  $t \mapsto X_t(\omega)$  is called the **path** associated to the experiment  $\omega$ .

**Def.** A **filtration** of a measurable space  $(\Omega, \mathcal{F})$  is an increasing sequence  $(\mathcal{F}_t)_{t \in T}$  of sub- $\sigma$ -fields of  $\mathcal{F}$ . A **filtered probability space**  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in T}, \mathbf{P})$  is a probability space endowed with a filtration. A random process  $(X_t)_{t \in T}$  defined on  $(\Omega, \mathcal{F}, \mathbf{P})$  is said to be **adapted** to the filtration if for each  $t \in T$ ,  $X_t$  is  $\mathcal{F}_t$ -measurable. Then we note  $((X_t, \mathcal{F}_t))_{t \in T}$ .

**Def.** The **natural filtration** of a process  $(X_t)_{t \in T}$  is the smallest filtration with respect to which  $(X_t)_{t \in T}$  is adapted, i.e.  $\forall t \in T$ ,  $\mathcal{F}_t^X = \sigma(X_s, s \leq t)$ .

**Def.** We call **finite dimensional distributions**, or **fidi distributions**, of the process  $X$  the collection of probability measures  $(\mathbf{P}_I)_{I \in \mathcal{I}}$  where  $\mathbf{P}_I$  denotes the probability distribution of the random vector  $\{X_t, t \in I\}$ .

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  $\forall x = (x_t)_{t \in I} \in X^I$ ,  $\Pi_{I,J}(x) = (x_t)_{t \in J}$ . Then  $\mathbf{P}_I \circ \Pi_{I,J}^{-1} = \mathbf{P}_J$  (compatibility condition). We denote  $\Pi_I = \Pi_{T,I}$  and  $\Pi_s = \Pi_{\{s\}}$  where  $s \in T$ .

**Th** (Kolmogorov). Let  $\mathcal{I}$  be the set of all finite subsets of  $T$ . Suppose that, for all  $I \in \mathcal{I}$ ,  $\nu_I$  is a probability measure on  $(X^I, \mathcal{X}^{\otimes I})$  and that the collection  $\{\nu_I, I \in \mathcal{I}\}$  satisfies  $\forall I, J \in \mathcal{I}, I \subset J, \nu_I \circ \Pi_{I,J}^{-1} = \nu_J$ . Then there exists a unique probability measure  $\mathbf{P}$  on  $(X^T, \mathcal{X}^{\otimes T})$  such that,  $\forall I \in \mathcal{I}, \nu_I = \mathbf{P} \circ \Pi_I^{-1}$ .

**Def.** Let  $X = (X_t)_{t \in T}$  be a random process defined on  $(\Omega, \mathcal{F}, \mathbf{P})$ . The **law in the sense of fidi distribution** is the image measure  $\mathbf{P}^X$ , that is, the unique probability measure defined on  $(X^T, \mathcal{X}^{\otimes T})$  that satisfies  $\forall I \in \mathcal{I}, \mathbf{P}^X \circ \Pi_I^{-1} = \mathbf{P}_I$ , i.e.  $\forall (A_t)_{t \in I} \in \mathcal{X}^I$ ,  $\mathbf{P}^X(\prod_{t \in I} A_t \times X^{T \setminus I}) = \mathbf{P}(X_t \in A_t, t \in I)$ .

**Def.** The canonical functions defined on  $(X^T, \mathcal{X}^{\otimes T})$  is the collection of measurable functions  $(\xi_t)_{t \in T}$  valued in  $(X, \mathcal{X})$  as  $\forall \omega = (\omega_t)_{t \in T} \in X^T$ ,  $\xi_t(\omega) = \omega_t$ . When  $(X^T, \mathcal{X}^{\otimes T})$  is endowed with the image measure  $\mathbf{P}^X$  then the **canonical process**  $(\xi_t)_{t \in T}$  defined on  $(X^T, \mathcal{X}^{\otimes T}, \mathbf{P}^X)$  has the same fidi distribution as  $X$ .

### Gaussian processes

**Def.** The real valued r.v.  $X$  is Gaussian if its characteristic function satisfies  $\phi_X(u) = \mathbf{E}(e^{iuX}) = \exp(i\mu u - \sigma^2 u^2/2)$  where  $\mu \in \mathbf{R}$  and  $\sigma \in \mathbf{R}_+$ .

If  $\sigma \neq 0$  then  $X$  admits a probability density function  $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ .

**Def.** A random vector  $[X_1, \dots, X_n]^\top$  valued in  $\mathbf{R}^n$  is a Gaussian vector if any linear combination of  $X_1, \dots, X_n$  is a Gaussian variable.

Let  $\mu$  denote the mean vector of  $[X_1, \dots, X_n]^\top$  and  $\Gamma$  its covariance matrix. Then  $\forall u \in \mathbf{R}^n$ ,  $Y = u^\top X$  is Gaussian,  $\mathbf{E}(Y) = u^\top \mu$  and  $\text{Var}(Y) = u^\top \Gamma u$ . Thus  $\phi_X(u) = \mathbf{E}[\exp(iu^\top X)] = \exp(iu^\top \mu - \frac{1}{2}u^\top \Gamma u)$ .

**Prop.** The probability distribution of an  $n$ -dimensional Gaussian vector  $X$  is determined by its mean vector and covariance matrix  $\Gamma$ . We denote  $X \sim \mathcal{N}(\mu, \Gamma)$ . Conversely, for all vector  $\mu \in \mathbf{R}^n$  and all non-negative symmetric matrix  $\Gamma$ , the distribution  $\mathcal{N}(\mu, \Gamma)$  is well defined.

**Lem.** Let  $X \sim \mathcal{N}(\mu, \Gamma)$  with  $\mu \in \mathbf{R}^n$  and  $\Gamma$  a  $n \times n$  non-negative symmetric matrix. Then  $X$  has independent components if and only if  $\Gamma$  is diagonal.

**Prop.** Let  $X \sim \mathcal{N}(\mu, \Gamma)$  with  $\mu \in \mathbf{R}^n$  and  $\Gamma$  a  $n \times n$  non-negative symmetric matrix. If  $\Gamma$  is full rank, the probability distribution of  $X$  admits a density defined in  $\mathbf{R}^n$  by  $p(x) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(\Gamma)}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Gamma^{-1}(x - \mu)\right)$ .

**Def.** A real-valued random process  $X = (X_t)_{t \in T}$  is called a **Gaussian process** if, for all finite set of indices  $I = \{t_1, \dots, t_n\}$ ,  $[X_{t_1}, \dots, X_{t_n}]^\top$  is a Gaussian vector.

**Th.** Let  $T$  be any set of indices,  $\mu: T \rightarrow \mathbf{R}$  and  $\gamma: T \times T \rightarrow \mathbf{R}$  such that all restrictions  $\Gamma_I$  to the set  $I \times I$  with  $I \subset T$  finite are nonnegative symmetric matrices. Then one can define a probability space  $(\Omega, \mathcal{F}, \mathbf{P})$  and a gaussian process  $(X_t)_{t \in T}$  defined on this space with mean  $\mu$  and covariance function  $\gamma$ , that is such that,  $\forall s, t \in T$ ,  $\mu(t) = \mathbf{E}[X_t]$  and  $\gamma(s, t) = \mathbf{E}[(X_s - \mu(s))(X_t - \mu(t))]$ .



**Def.** Let  $T$ ,  $\mu$  and  $\gamma$  be as above. We denote by  $\mathcal{N}(\mu, \gamma)$  the law of the Gaussian process with mean  $\mu$  and covariance  $\gamma$  in the sense of fidi distribution.

### Strict stationarity of a random process in discrete time

Suppose that  $T = \mathbf{Z}$  or  $T = \mathbf{N}$ .

**Def.** We denote by  $S: \mathcal{X}^T \rightarrow \mathcal{X}^T$  and call the **shift operator** the mapping defined by  $\forall x = (x_t)_{t \in T}, S(x) = (x_{t+1})_{t \in T}$ . For all  $\tau \in T$  we define  $S^\tau$  by  $S^\tau(x) = (x_{t+\tau})_{t \in T}$ . The operator  $B = S^{-1}$  is called the **backshift operator**.

**Def.** A random process  $(X_t)_{t \in T}$  is **strictly stationary** if  $X$  and  $S \circ X$  have the same law, i.e.  $\mathbf{P}^{S \circ X} = \mathbf{P}^X$ .

### Stationarity preserving transformations

In this section, we set  $T = \mathbf{Z}$ ,  $\mathcal{X} = \mathbf{C}^d$  and  $\mathcal{X} = \mathcal{B}(\mathbf{C}^d)$  for some integer  $d \geq 1$ .

**Def.** Let  $\phi$  be a measurable function from  $(\mathcal{X}^T, \mathcal{X}^{\otimes T})$  to  $(\mathcal{Y}^T, \mathcal{Y}^{\otimes T})$  and  $X = (X_t)_{t \in T}$  be a process with values in  $(\mathcal{X}, \mathcal{X})$ . A  $\phi$ -filtering process  $Y = (Y_t)_{t \in T}$  is defined as  $\forall t, Y_t = \phi \circ X$  or, equivalently,  $Y_t = \Pi_t(\phi(X))$ . Thus  $Y$  makes its values in  $(\mathcal{Y}, \mathcal{Y})$ . If  $\phi$  is linear, we will say that  $Y$  is obtained by linear filtering of  $X$ .

**Def.** A  $\phi$ -filter is **shift invariant** if  $\phi$  commutes with  $S$ .

**Rem.** A shift invariant  $\phi$ -filter preserves the strict stationarity and is entirely determined by its composition with the canonical projection  $\Pi_0$

## 5 Weakly stationary processes

### $L^2$ processes

**Def.** The process  $X = (X_t)_{t \in T}$  defined on  $(\Omega, \mathcal{F}, \mathbf{P})$  with values in  $\mathbf{C}^d$  is an  $L^2$  process if  $\forall t \in T, X_t = L^2(\Omega, \mathcal{F}, \mathbf{P})$ . Its **mean function** is defined on  $T$  by  $\mu(t) = \mathbf{E}(X_t)$  and the **covariance function** is defined on  $T \times T$  by  $\Gamma(s, t) = \text{Cov}(X_s, X_t) = \mathbf{E}((X_s - \mu(s))(X_t - \mu(t))^H)$ .

**Prop.** Let  $\Gamma$  be the covariance function of a  $L^2$  process  $X = (X_t)_{t \in T}$  with values in  $\mathbf{C}^d$ . The following properties hold.

- (i) *Hermitian symmetry* :  $\forall s, t \in T, \Gamma(s, t) = \Gamma(t, s)^H$ .
- (ii) *Nonnegativity* :  $\forall n \in \mathbf{N}^*, t_1, \dots, t_n \in T, a_1, \dots, a_n \in \mathbf{C}^d, \sum_{1 \leq k, m \leq n} a_k^H \Gamma(t_k, t_m) a_m \geq 0$ .

Conversely, if  $\Gamma$  satisfies these two properties, there exists an  $L^2$  process with values in  $\mathbf{C}^d$  with covariance function  $\Gamma$ .

In the scalar case ( $d = 1$ ), we also use the notation  $\gamma(s, t)$ .

### Weakly stationary processes

**Def.** Let  $\mu \in \mathbf{C}^d$  and  $\Gamma: \mathbf{Z} \rightarrow \mathbf{C}^{d \times d}$ . A process  $(X_t)_{t \in \mathbf{Z}}$  with values in  $\mathbf{C}^d$  is said **weakly stationary** with mean  $\mu$  and autocovariance function  $\Gamma$  if all the following assertions hold :

- (i)  $X$  is an  $L^2$  process, i.e.  $\mathbf{E}(|X_t|^2) < +\infty$ ,
- (ii)  $\forall t \in \mathbf{Z}, \mathbf{E}(X_t) = \mu$ ,
- (iii)  $\forall (s, t) \in \mathbf{Z} \times \mathbf{Z}, \text{Cov}(X_s, X_t) = \Gamma(s - t)$ .

A strictly stationary  $L^2$  process is weakly stationary.

**Prop.** The autocovariance function  $\gamma: \mathbf{Z} \rightarrow \mathbf{C}$  of a complex valued weakly stationary process satisfies the following :

- (i) *Hermitian symmetry* :  $\forall s \in \mathbf{Z}, \gamma(-s) = \overline{\gamma(s)}$ .
- (ii) *Nonnegative definiteness* :  $\forall i \in \mathbf{N}^*, a_1, \dots, a_n \in \mathbf{C}, \sum_{s=1}^n \sum_{t=1}^n \overline{a_s} \gamma(s - t) a_t \geq 0$ .

**Def.** Let  $X$  be a weakly stationary process with autocovariance function  $\gamma$  such that  $\gamma(0) \neq 0$ . The **autocorrelation function** of  $X$  is defined as  $\forall \tau \in \mathbf{Z}, \rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$ . It is normalized in the sense that  $\rho(0) = 1$  and  $\forall s \in \mathbf{Z}, |\rho(s)| \leq 1$ .

**Def.** A **weak white noise** is a centered weakly stationary process whose autocovariance function satisfies  $\gamma(0) = \sigma^2 > 0$  and  $\forall s \neq 0, \gamma(s) = 0$ . We will denote  $(X_t) \sim \text{WN}(0, \sigma^2)$ . When a white noise is an i.i.d. process it is called a **strong white noise**. We will denote  $(X_t) \sim \text{IID}(0, \sigma^2)$ .

### Spectral measure

**Th.** (Herglotz) A sequence  $(\gamma(h))_{h \in \mathbf{Z}}$  is a nonnegative definite hermitian sequence iff there exists a finite nonnegative measure  $\nu$  on  $(\mathbf{T}, \mathbf{T})$  such that

$$\forall h \in \mathbf{Z}, \gamma(h) = \int_{\mathbf{T}} e^{ih\lambda} \nu(d\lambda).$$

Moreover this relation defines  $\nu$  uniquely.

**Rem.** This applies to all autocovariance function of a weakly stationary process  $X$ . In this case  $\nu$  is called the **spectral measure** of  $X$ . If  $\nu$  admits a density  $f$ , it is called the **spectral density function**.

**Cor.** Let  $(\gamma(h))_{h \in \mathbf{Z}} \in l^2(\mathbf{Z})$ . Then it is a nonnegative definite hermitian sequence iff for almost every  $\lambda$ ,  $f(\lambda) = \frac{1}{2\pi} \sum_{h \in \mathbf{Z}} \gamma(h) e^{-ih\lambda}$  is nonnegative, where the convergence holds in  $L^2(\mathbf{T})$ .

**Def.** A weakly stationary process  $X$  is called **linearly predictable** if  $\exists n \geq 1, \forall t \geq n, X_t \in \text{Span}(X_1, \dots, X_n)$  (in the  $L^2$  sense).

**Prop.** Let  $\gamma$  be the autocovariance function of a weakly stationary process  $X$ . If  $\gamma(0) \neq 0$  and  $\gamma(t) \xrightarrow{t \rightarrow \infty} 0$  then  $X$  is not linearly predictable.

### Spectral representation of a weakly stationary process

**Def.** A random fields with orthogonal increments  $W$  on  $(X, \mathcal{X})$  is a  $L^2$  random process indexed on  $\mathcal{X}$ , say  $W = (W(A))_{A \in \mathcal{X}}$  such that

- (i)  $\forall A \in \mathcal{X}, \mathbf{E}(W(A)) = 0$ ,
- (ii)  $\forall A, B \in \mathcal{X}$  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  $(A_n)_{n \in \mathbf{N}} \subset \mathcal{X}, \bigcap_{n=0}^{\infty} A_n = \emptyset$ , we have  $\text{Var}(W(A_n)) \rightarrow 0$ .

**Lem.** Let  $W$  a random field with orthogonal increments on  $(X, \mathcal{X})$ . Let  $A \in \mathcal{X}$  and set  $\nu(A) = \text{Var}(W(A))$ . Then  $\nu$  is a finite nonnegative measure on  $(X, \mathcal{X})$ . Moreover  $\forall A, B \in \mathcal{X}, \text{Cov}(W(A), W(B)) = \nu(A \cap B)$ .

The measure  $\nu$  is called the **intensity measure** of  $W$ .

**Lem.** Let  $W$  be a  $L^2$  random process indexed by  $\mathcal{X}$  such that  $\forall A \in \mathcal{X}, \mathbf{E}(W(A)) = 0$ . Suppose  $\exists \nu, \forall A, B \in \mathcal{X}, \text{Cov}(W(A), W(B)) = \nu(A \cap B)$ . Then  $W$  is a random field with orthogonal increments on with intensity measure  $\nu$ .

**Th.** Let  $W$  be a random field with orthogonal increments with intensity measure  $\nu$ . Then there exists a unique isometric operator  $w$  from  $L^2(X, \mathcal{X}, \nu)$  to  $L^2(\Omega, \mathcal{F}, \mathbf{P})$  such that  $\forall A \in \mathcal{X}, w(\mathbf{1}_A) = W(A)$ . For all  $f \in L^2(X, \mathcal{X}, \nu)$  we further have  $\mathbf{E}(w(f)) = 0$  and  $w(L^2(X, \mathcal{X}, \nu)) = \overline{\text{Span}(W(A), A \in \mathcal{X})}$ .

**Th.** Let  $\nu$  be a finite nonnegative measure on  $(X, \mathcal{X})$  and  $J: L^2(X, \mathcal{X}, \nu) \rightarrow L^2(\Omega, \mathcal{F}, \mathbf{P})$  an isometric operator such that  $\forall f, \mathbf{E}(J(f)) = 0$ . Then there exists a random field  $W$  with orthogonal increments on  $X$  with intensity measure  $\nu$  such that  $\forall f, J(f) = \int_X f dW$ .

**Prop.** Let  $W$  be a r.f.o.i. on  $(\mathbf{T}, \mathcal{B}(\mathbf{T}))$  with intensity measure  $\nu$ . Then the sequence  $(X_t)_{t \in \mathbf{Z}}$  defined by  $X_t = \int_{\mathbf{T}} e^{it\lambda} dW(\lambda)$  is a centered weakly stationary process with spectral measure  $\nu$ .

**Def.** Let  $X$  be a  $L^2$  process. Its **linear closure** is defined as  $\mathcal{H}_{\infty}^X = \overline{\text{Span}(X_t, t \in \mathbf{Z})}$  (closure in  $L^2(\Omega, \mathcal{F}, \mathbf{P})$ ).

**Th.** Let  $X$  be a centered weakly stationary process with spectral measure  $\nu$ . Then there exists a r.f.o.i.  $\hat{X}$  on  $(\mathbf{T}, \mathcal{B}(\mathbf{T}))$  with intensity measure  $\nu$ , called the **spectral field**, such that  $\forall t \in \mathbf{Z}, X_t = \int e^{it\lambda} d\hat{X}(\lambda)$ . Moreover, the mapping  $f \mapsto \int f d\hat{X}$  defines the unique operator from  $L^2(\mathbf{T}, \mathcal{B}(\mathbf{T}), \nu)$  to  $\mathcal{H}_{\infty}^X$  that maps each function  $\lambda \mapsto e^{it\lambda}$  to  $X_t$ .

### Innovation process

Let  $X = (X_t)_{t \in \mathbf{Z}}$  denote a centered weakly stationary process. Let  $\mathcal{H}_t^X = \overline{\text{Span}(X_s, s \leq t)}$  denote the **linear past** of  $X$  up to time  $t$ .

**Def.** We call **innovation process** the process  $\epsilon = (\epsilon_t)_{t \in \mathbf{Z}}$  defined by  $\epsilon_t = X_t - \text{proj}(X_t | \mathcal{H}_{t-1}^X)$ .

**Def.** We call **predictor of order  $p$**  the random variable  $\text{proj}(X_t | \mathcal{H}_{t-1,p}^X)$  and the **partial innovation process** of order  $p$  the process  $\epsilon_p^+ = (\epsilon_{t,p}^+)_{t \in \mathbf{Z}}$  defined by  $\epsilon_{t,p}^+ = X_t - \text{proj}(X_t | \mathcal{H}_{t-1,p}^X)$ . The **prediction coefficients** are any coefficients  $\phi_p^+ = (\phi_{k,p}^+)_{k \in \llbracket 1;p \rrbracket}$  which satisfy  $\forall t \in \mathbf{Z}, \text{proj}(X_t | \mathcal{H}_{t-1,p}^X) = \sum_{k=1}^p \phi_{k,p}^+ X_{t-k}$ .

**Cor.** The innovation process of a centered weakly stationary process is a (centered) weak white noise. Its variance is called the **innovation variance** of the process.

**Def.** If the variance of its innovation process is zero, we say that  $X$  is **deterministic**. Otherwise we say that  $X$  is **regular**.

We define the intersection of the whole past of  $X$  as  $\mathcal{H}_{-\infty}^X = \bigcap_{t \in \mathbf{Z}} \mathcal{H}_t^X$ . We denote  $\psi_s = \frac{\langle X_t | \epsilon_{t-s} \rangle}{\sigma^2}$ .

**Th (Wold decomposition).** Let  $X$  be a regular process,  $\epsilon$  its innovation process and  $\sigma^2$  its innovation variance, so that  $\epsilon \sim \text{WN}(0, \sigma^2)$ . Define the  $L^2$  centered process  $U$  as  $U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}$ . Define the  $L^2$  centered process  $V$  by  $\forall t \in \mathbf{Z}, X_t = U_t + V_t$ . Then the following assertions hold,

- (i)  $U_t = \text{proj}(X_t | \mathcal{H}_t^{\epsilon})$  and  $V_t = \text{proj}(X_t | \mathcal{H}_{-\infty}^X)$ ,
- (ii)  $\epsilon$  and  $V$  are uncorrelated :  $\forall (t, s), \langle V_t | \epsilon_s \rangle = 0$ ,
- (iii)  $U$  is a purely non-deterministic process and has same innovation as  $X$ . Moreover  $\forall t \in \mathbf{Z}, \mathcal{H}_t^{\epsilon} = \mathcal{H}_t^U$ .
- (iv)  $V$  is a deterministic process and  $\mathcal{H}_{-\infty}^V = \mathcal{H}_{-\infty}^X$ .

## 6 Markov chains : basic definitions

### Definition using conditioning

**Def.** Let  $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbf{N}}, \mathbf{P})$  be a filtered probability space and  $(X, \mathcal{X})$  be a measurable space. An adapted stochastic process  $((X_k, \mathcal{F}_k))_{k \in \mathbf{N}}$  on  $X$  is a **Markov chain** if  $\forall k \in \mathbf{N}, \forall A \in \mathcal{X}, \mathbf{P}(X_{k+1} \in A | \mathcal{F}_k) = \mathbf{P}(X_{k+1} \in A | X_k)$ .



**Prop.** Let  $((X_k, \mathcal{F}_k))_{k \in \mathbb{N}}$  be an adapted stochastic process. The following properties are equivalent :

- (i)  $((X_k, \mathcal{F}_k))_{k \in \mathbb{N}}$  is a Markov chain,
- (ii)  $\forall k \in \mathbb{N}, \forall Y \in L^1(\Omega, \sigma(X_l, l \geq k), \mathbf{P}), \mathbf{E}(Y | \mathcal{F}_k) = \mathbf{E}(Y | X_k)$   $\mathbf{P}$ -a.s.
- (iii)  $\forall k \in \mathbb{N}, \forall Y \in L^1(\Omega, \sigma(X_l, l \geq k), \mathbf{P}), \forall Z \in L^\infty(\Omega, \mathcal{F}_k, \mathbf{P}), \mathbf{E}(YZ | X_k) = \mathbf{E}(Y | X_k)\mathbf{E}(Z | X_k)$   $\mathbf{P}$ -a.s.

#### How to use kernels

Let  $N$  be a kernel and  $f$  be a measurable function defined on  $\mathcal{Y}$ . We denote by  $Nf$  the function defined on  $\mathcal{X}$  by  $Nf: x \mapsto \int_{\mathcal{Y}} N(x, dy)f(y)$  whenever this integral is well-defined (for instance if  $f$  is non-negative).

**Prop.** Let  $N$  be a kernel on  $\mathcal{X} \times \mathcal{Y}$ . Then  $\forall f \in \mathbf{F}_+(\mathcal{Y}, \mathcal{Y}), Nf \in \mathbf{F}_+(\mathcal{X}, \mathcal{X})$ . Moreover, if  $N$  is a probability kernel, then  $\forall f \in \mathbf{F}_b(\mathcal{X}, \mathcal{X}), Nf \in \mathbf{F}_b(\mathcal{Y}, \mathcal{Y})$ .

**Lem.** Let  $M: \mathbf{F}_+(\mathcal{Y}, \mathcal{Y}) \rightarrow \mathbf{F}_+(\mathcal{X}, \mathcal{X})$  be an additive and positively homogeneous function such that  $\lim_{n \rightarrow \infty} M(f_n) = M(\lim_{n \rightarrow \infty} f_n)$  for every non-decreasing sequence  $(f_n)_{n \in \mathbb{N}}$ . Then the function  $N$  defined on  $\mathcal{X} \times \mathcal{Y}$  by  $N(x, A) = M(\mathbf{1}_A)(x)$  is a kernel and  $\forall f \in \mathbf{F}_+(\mathcal{Y}, \mathcal{Y}), M(f)(x) = \int_{\mathcal{Y}} N(x, dy)f(y)$ .

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

**Prop.** Let  $M$  and  $N$  be two kernels on  $\mathcal{X} \times \mathcal{Y}$  and  $\mathcal{Y} \times \mathcal{Z}$ . There exists a kernel  $MN$  on  $\mathcal{X} \times \mathcal{Z}$ , called the **composition** or the **product** of  $M$  and  $N$  such that  $\forall x \in \mathcal{X}, \forall A \in \mathcal{Z}, MN(x, A) = \int_{\mathcal{Y}} M(x, dy)N(y, A)$ .

**Prop.** Let  $M$  and  $N$  be kernels on  $\mathcal{X} \times \mathcal{Y}$  and  $\mathcal{Y} \times \mathcal{Z}$ . Then  $\forall x \in \mathcal{X}, \forall f \in \mathbf{F}_+(\mathcal{Z}, \mathcal{Z}), MNf(x) = M[Nf](x)$ .

**Prop.** Let  $M$  be a kernel on  $\mathcal{X} \times \mathcal{Y}$  and  $N$  be a kernel on  $\mathcal{Y} \times \mathcal{Z}$ . Then, there exists a kernel  $M \otimes N$  on  $\mathcal{X} \times (\mathcal{Y} \otimes \mathcal{Z})$ , called the **tensor product** of  $M$  and  $N$ , such that

$$\forall f \in \mathbf{F}_+(\mathcal{Y} \times \mathcal{Z}, \mathcal{Y} \otimes \mathcal{Z}), M \otimes Nf(x) = \int_{\mathcal{Y} \times \mathcal{Z}} f(v)M \otimes N(x, dv) = \int_{\mathcal{Y}} \left( \int_{\mathcal{Z}} f(y, z)N(y, dz) \right) M(x, dy).$$

**Lem.** Let  $M$  be a kernel on  $\mathcal{X} \times \mathcal{Y}$  and  $N$  be a kernel on  $\mathcal{Y} \times \mathcal{Z}$ .

- (i) If  $M$  and  $N$  are both finite kernels, then  $M \otimes N$  is a finite kernel.
- (ii) If  $M$  and  $N$  are both probability kernels, then  $M \otimes N$  is a probability kernel.
- (iii) Then tensor product of kernels is associative :  $(M \otimes N) \otimes P = M \otimes (N \otimes P)$ .

#### Homogeneous Markov chains

**Def.** Let  $(\mathcal{X}, \mathcal{X})$  be a measurable space,  $\nu$  a probability measure on  $(\mathcal{X}, \mathcal{X})$ ,  $P$  a Markov kernel on  $\mathcal{X} \times \mathcal{X}$  and  $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbb{N}}, \mathbf{P})$  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  $\nu$  is

$$\forall k \in \mathbb{N}, \forall A \in \mathcal{X}, \quad \mathbf{P}(X_0 \in A) = \nu(A) \quad \text{and} \quad \mathbf{P}(X_{k+1} \in A | \mathcal{F}_k) = P(X_k, A) \quad \mathbf{P}\text{-a.s.}$$

**Not.** Unless specify, we consider the natural filtration and write that  $(X_k)_{k \in \mathbb{N}}$  is a homogeneous Markov chain.

**Prop.** An  $\mathcal{X}$ -valued random process  $(X_k)_{k \in \mathbb{N}}$  is a homogeneous Markov chain with kernel  $P$  and initial distribution  $\nu$  if and only if, for all  $k \in \mathbb{N}$ , the distribution of  $(X_1, \dots, X_n)$  is  $\nu \otimes P^{\otimes k}$ .

We also get that the distribution of  $X_k$  is  $\nu P^k$ .

**Th.** Let  $\mathcal{X} = \mathbf{R}^d$  with  $d \geq 1$  endowed with  $\mathcal{X} = \mathcal{B}(\mathbf{R}^d)$ ,  $P$  a Markov kernel on  $(\mathcal{X}, \mathcal{X})$  and  $\nu$  a probability measure on  $(\mathcal{X}, \mathcal{X})$ . 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  $(\mathcal{X}, \mathcal{X})$  and a measurable application  $f$  from  $(\mathcal{X} \times [0; 1], \mathcal{X} \otimes \mathcal{B}([0; 1]))$  to  $(\mathcal{X}, \mathcal{X})$  such that the sequence  $(X_k)_{k \in \mathbb{N}}$  defined by  $X_0 = g(Z_0)$  and  $\forall k \geq 0, X_{k+1} = f(X_k, Z_{k+1})$  is a homogeneous Markov chain with initial distribution  $\nu$  and Markov kernel  $P$ .

#### The canonical Chain, notation $\mathbf{P}_\xi$ and $\mathbf{E}_\xi$

**Th.** Let  $P$  be a Markov kernel on  $\mathcal{X} \times \mathcal{X}$  and  $\nu \in \mathbf{M}_1(\mathcal{X}, \mathcal{X})$ . Then, there exists a unique probability  $\mathbf{P}_\nu$  on  $(\mathcal{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$  is a Markov chain with initial distribution  $\nu$  and kernel  $P$ .

**Def.** The **canonical Markov chain** with kernel  $P$  on  $\mathcal{X} \times \mathcal{X}$  is the canonical process  $(X_n)_{n \in \mathbb{N}}$  on the canonical filtered space  $(\mathcal{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, (\mathcal{F}_k)_{k \in \mathbb{N}})$  endowed with the collection of probability measures  $\{\mathbf{P}_\nu, \nu \in \mathbf{M}_1(\mathcal{X}, \mathcal{X})\}$  given by the previous theorem.

The expectation associated to  $\mathbf{P}_\nu$  is denoted by  $\mathbf{E}_\nu$  and for  $x \in \mathcal{X}$ ,  $\mathbf{P}_x$  and  $\mathbf{E}_x$  are shorthand for  $\mathbf{P}_{\delta_x}$  and  $\mathbf{E}_{\delta_x}$ .

**Prop.** (i)  $\forall A \in \mathcal{X}^{\otimes}$ , the function  $x \mapsto \mathbf{P}_x(A)$  is  $\mathcal{X}$ -measurable.

(ii)  $\forall \xi \in \mathbf{M}_1(\mathcal{X}, \mathcal{X}), \forall A \in \mathcal{X}^{\otimes}, \mathbf{P}_\xi(A) = \int_{\mathcal{X}} \mathbf{P}_x(A)\xi(dx)$

(iii)  $\forall \xi \in \mathbf{M}_1(\mathcal{X}, \forall Y \in \mathbf{F}_+(\mathcal{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}), \mathbf{E}_\xi(Y) = \int_{\mathcal{X}} \mathbf{E}_x(Y)\xi(dx)$

(iv)  $\forall \xi \in \mathbf{M}_1(\mathcal{X}, \forall Y \in L^1(\mathcal{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbf{P}_\xi), \mathbf{E}(|Y|) < \infty$  for  $\xi$ -a.e.  $x$  and  $\mathbf{E}_\xi(Y) = \int_{\mathcal{X}} \mathbf{E}_x(Y)\xi(dx)$

(v)  $\forall \xi \in \mathbf{M}_1(\mathcal{X}, \forall Y \in L^1(\mathcal{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, \mathbf{P}_\xi), \mathbf{E}_\xi(Y | X_0) = \mathbf{E}_{X_0}(Y)$   $\mathbf{P}_\xi$ -a.s.

**Def.** An event is  $\mathbf{P}_*$ -a.s. if it is  $\mathbf{P}_\nu$ -a.s. for all initial distribution  $\nu$ .

**Lem.** Let  $A \in \mathcal{X}^{\otimes \mathbb{N}}$ . We have  $X \in A$   $\mathbf{P}$ -a.s. if and only if it is true  $\mathbf{P}_x$ -a.s. for all  $x \in \mathcal{X}$ .

## 7 Stationary Markov chains

### Invariant measures and stationarity

Let  $P$  be a Markov kernel on  $(X, \mathcal{X})$ .

**Def.** A non zero  $\sigma$ -finite positive measure  $\mu \in \mathbf{M}_+(X, \mathcal{X})$  is said to be **invariant** with respect to  $P$  (or  $P$ -invariant) if  $\mu = \mu P$ .

**Th.** A Markov chain  $(X_k)_{k \in \mathbf{N}}$  defined on  $(\Omega, \mathcal{F}, \mathbf{P})$  with kernel  $P$  is a stationary process iff its initial distribution is invariant with respect to  $P$ .

**Def.** A non-empty set  $B \in \mathcal{X}$  is called **absorbing** if  $\forall x \in B, P(x, B) = 1$ .

**Th.** (i) The set of invariant probability measures for  $P$  is a convex subset of the convex cone  $\mathbf{M}_+(X, \mathcal{X})$ .

(ii) Let  $\pi$  be an invariant probability and  $X_1 \subset X$  with  $\pi(X_1) = 1$ . There exists  $B \subset X_1$  such that  $\pi(B) = 1$  and  $B$  is absorbing for  $P$ .

### Stopping times

**Def.** A random variable  $\tau: \Omega \rightarrow \bar{\mathbf{N}} = \mathbf{N} \cup \{\infty\}$  is called a **stopping time** if  $\forall k \in \mathbf{N}, \{\tau \leq k\} \in \mathcal{F}_k$ . The family  $\mathcal{F}_\tau$  of events  $A$  such that  $\forall k \in \mathbf{N}, A \cap \{\tau \leq k\} \in \mathcal{F}_k$  is called the  $\sigma$ -field of events prior to times  $\tau$ .

**Def.** For  $A \in \mathcal{X}$ , the **first hitting time** of the set  $A$  by the process  $(X_n)_{n \in \mathbf{N}}$  is  $\tau_A = \inf\{n \geq 0 \mid X_n \in A\}$  and **return time** is  $\sigma_A = \inf\{n \geq 1 \mid X_n \in A\}$ , where, by convention,  $\inf \emptyset = +\infty$ . The successive return times  $\sigma_A^{(n)}$  are defined successively by  $\sigma_A^{(0)} = 0$  and  $\forall k \geq 0, \sigma_A^{(k+1)} = \inf\{n > \sigma_A^{(k)} \mid X_n \in A\}$ .

**Prop.** Let  $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbf{N}}, \mathbf{P})$  be a filtered space and  $\tau$  and  $\sigma$  be two stopping times for  $(\mathcal{F}_k)_{k \in \mathbf{N}}$ . Denote by  $\mathcal{F}_\tau$  and  $\mathcal{F}_\sigma$  the  $\sigma$ -fields of the events prior to  $\tau$  and  $\sigma$  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)  $\{\tau < \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$  and  $\{\tau = \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$ .