

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 2π -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 2π -periodic function. Then the Cesaro sequence $\frac{1}{n} \sum_{k=0}^{n-1} f_k$ converges uniformly to f .

Cor. Let μ 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 (π - λ theorem). If $\mathcal{A} \subset \mathcal{C}$ with \mathcal{A} a π -system and \mathcal{C} a λ -system, then $\sigma(\mathcal{A}) = \mathcal{C}$.

Th (Characterization of probability measures). Let \mathcal{C} be a π -system on Ω and $\mathcal{F} = \sigma(\mathcal{C})$ the smallest σ -field containing \mathcal{C} . Then a probability measure μ on (Ω, \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- σ -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- σ -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 σ -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- σ -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- σ -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- σ -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- σ -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} .

Lem. 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- σ -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 ω , $A \mapsto \mathbf{P}^{Y|\mathcal{G}}(\omega, A)$ 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 $\mathbf{P}^{Y|X} : X \times \mathcal{Y} \rightarrow [0; 1]$ such that $\forall A \in \mathcal{Y}, \mathbf{P}^{Y|X}(X, A) = \mathbf{P}(Y \in A \mid X)$ \mathbf{P} -a.s.

Th. Let \mathcal{G} be sub- σ -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 ω 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) $\mathbf{P}^{Y|X}$.

Lem. 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 $\text{Cov}(\mathbf{X} - \hat{\mathbf{X}}) = \mathbf{E} \left((\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^{\top} \right) = \mathbf{E} \left((\mathbf{X} - \hat{\mathbf{X}}) \mathbf{X}^{\top} \right)$ and $\mathbf{P}^{Y|X}(\mathbf{X}, \cdot) = \mathcal{N} \left(\hat{\mathbf{X}}, \text{Cov}(\mathbf{X} - \hat{\mathbf{X}}) \right)$.

Radon-Nikodym derivative

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

Def. Let λ be a measure on (Ω, \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 $\forall A \in \mathcal{F}, (\lambda(A) = 0) \implies (\mu(A) = 0)$.

Th (Radon-Nikodym theorem). Let $\lambda, \mu \in \mathbf{M}_+(\Omega, \mathcal{F})$ be σ -finite measures such that $\mu \ll \lambda$. Then, there exists a non-negative Borel function ϕ 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 (Ω, \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 ν . 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 (Ω, \mathcal{F}) and X a measurable mapping from (Ω, \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 (Ω, \mathcal{F}) be a measurable space and \mathcal{P} a collection of probabilities on this space. Let X be a measurable function from (Ω, \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 ν -dominated model for X , or that \mathcal{P}^X is ν -dominated, if $\forall P \in \mathcal{P}, P^X \ll \nu$.

Lem. Let $\nu \in \mathbf{M}_+(X, \mathcal{X})$. Consider a ν -dominated model \mathcal{P} for the variable X . Then there exists a countable collection $(P_n)_{n \geq 1}$ 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 Θ 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 $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ 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 $Q \subset \mathbf{R}^d \times \mathcal{X}$ 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 ν -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 ν -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 ν . 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 Θ , $\theta \mapsto f_\theta \circ X$ where f_θ denotes the density of P_θ with respect to ν .

Def. 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)**.

