Lecture 6 : Basic Probability Tools

1 Probability space

1.1 Introduction with examples

A probability space is defined as a triplet $(\Omega, \mathcal{F}, P)$, where:

- Ω is the sample space.
- $\mathcal{F}$ is the event space.
- P is the probability measure such that $P(\Omega) = 1$.

and a random variable is a mapping from Ω to some other space (in our application it will be $\mathbb{R}^p$ for a certain $p \in \mathbb{N}$). We will define those object rigorously in the sequel, let us first present some examples to understand relevant notions and difficulties.

Example 6.1. *Let $\mathbb{P}$ be a probability uniformly distributed on $\Omega \equiv [0, 2\pi)$ it is in bijection with the sphere:*

$$\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2, x^2 + y^2 = 1\}.$$

through the correspondence $\Phi : \Omega \to \mathbb{S}^1$ satisfying $\Phi(\omega) = (\cos\omega, \sin(\omega))$. For any random variable $X : \Omega \to \mathbb{R}^p$, one can define the expectation:

$$\mathbb{E}[X] \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega),$$

(this is the Lebesgue integral along the measure $\mathbb{P}$). In this example we set that $\mathbb{P}$ was the uniform measure on $[0, 2\pi)$, therefore:

$$\mathbb{E}[X] = \frac{1}{2\pi} \int_0^{2\pi} X(\omega) d\omega.$$

In particular, let us introduce the random variables $X, Y : \Omega \to \mathbb{R}$ defined for any $\omega \in \Omega$ as:

$$X(\omega) = \cos(\omega) \qquad \text{and} \qquad Y(\omega) = \sin(\omega).$$

then one has:

$$\mathbb{E}[XY] \equiv \int_\Omega \cos(\omega)\sin(\omega) d\mathbb{P}(\omega) = \frac{1}{2} \int_0^{2\pi} \sin(2\omega) d\omega = 0 = \mathbb{E}[X]\mathbb{E}[Y].$$

One could then be tempted to think that X and Y are independent (for a definition that will be given later). But that is not the case. It can be intuitively anticipated from the relation

$$\forall \omega \in \Omega : \qquad X(\omega)^2 + Y(\omega)^2 = cos(\omega)^2 + \sin(\omega)^2 = 1.$$

But one can more simply remark that:

$$\mathbb{P}\left(X \geq \frac{1}{2}\right) = \mathbb{P}\left(X^{-1}\left(\left[\frac{1}{2}, +\infty\right[\right)\right) = \mathbb{P}\left(\left[0, \frac{\pi}{3}\right] \cup \left[\frac{5\pi}{3}, 2\pi\right[\right) = \frac{2\pi}{3 \cdot 2\pi} = \frac{1}{3},$$

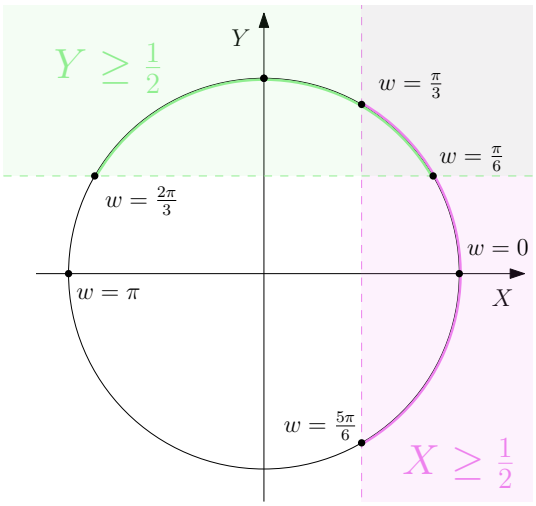

Figure 6.1: Representation of $\mathbb{P}(X \geq \frac{1}{2}, Y \geq \frac{1}{2})$, $\mathbb{P}(X \geq \frac{1}{2})$ and $\mathbb{P}(Y \geq \frac{1}{2})$.

the same way, $\mathbb{P}(Y \geq \frac{1}{2}) = \frac{1}{3}$*, but:*

$$\begin{aligned}\mathbb{P}(X \geq \frac{1}{2}, Y \geq \frac{1}{2}) &= \mathbb{P}(X^{-1}\left(\left[\frac{1}{2}, +\infty\right[\right) \cap Y^{-1}\left(\left[\frac{1}{2}, +\infty\right[\right)) \\ &= \mathbb{P}\left(\left[\frac{\pi}{6}, \frac{\pi}{3}\right]\right) = \frac{\pi}{6 \cdot 2\pi} = \frac{1}{12} \\ &\neq \frac{1}{9} = \mathbb{P}(X \geq \frac{1}{2})\mathbb{P}(Y \geq \frac{1}{2})\end{aligned}$$

This implies that X *and* Y *are two dependent variables, as expected.*

Example 6.2. *In statistical learning, the common model is:*

$$y = f(x) + \varepsilon,$$

where $y \in \mathbb{R}$*,* $x \in \mathbb{R}^p$ *and* $\varepsilon \in \mathbb{R}$*. In fact,* x, y *and* ε *are all random variables. A first modelling could be to consider the sample set* $\Omega = \mathbb{R}^p$ *with a certain probability law* $\mathbb{P}$ *defined on* Ω *and characteristic of the distribution of* x *(for example, the canonical Gaussian distribution). But then all random variables would write* $g(x)$ *and thus depend on* x*. This does not allow us to correctly model the noise random variable* ε*, which is traditionally assumed to be independent with* x*. Therefore, to correctly model this kind of settings, one has to consider a conceptual sample set* Ω *(also called "universe"), which is endowed with a certain probability law* $\mathbb{P}$ *and which allows to define all the random variables appearing in the problem. In the model studied,* $Y : \Omega \to \mathbb{R}$*,* $X : \Omega \to \mathbb{R}^p$ *and* $\varepsilon : \Omega \to \mathbb{R}$*. This conceptual model allows correctly formulate independence between* X *and* ε*. The exact nature of* Ω *is then almost never investigated, all calculations and inferences are done only thanks to the properties satisfied by* Ω *and* $\mathbb{P}$*, which we will define below.*

1.2 Definition of a Probability space

The sample space Ω is a general set and does not have to satisfy particular properties. We provide below the properties of the event space $\mathcal{F}$ and of the probability measure $\mathbb{P}$ that are crucial to be able to model random behaviors. The set of all subsets of Ω is traditionnaly denoted 2^Ω because there is a correspondance between the subsets of Ω and the mappings from Ω to $\{0, 1\}$. Given such a mapping $f : \omega \to \{0, 1\}$, one can indeed introduce the subset $A_f = \{x \in \Omega, f(x) = 1\}$ and given $A \subset \Omega$, one can define uniquely the mapping $f : \omega \to \{0, 1\}$ satisfying $f(x) = 1$ if $x \in \Omega$ and $f(x) = 0$ otherwise.

Definition 1 (Event space)**.** *Let* Ω *be a sample space. A set of subsets of* Ω*,* $\mathcal{F} \subset 2^\Omega$ *is alled an event space iif it satisfies the following properties:*

1. Ω ***belongs to*** $\mathcal{F}$*: The sample space itself is an element of* $\mathcal{F}$*, i.e.,* $\Omega \in \mathcal{F}$*.*
2. ***Closed under complementation****: If* $A \in \mathcal{F}$*, then the complement of* A *with respect to* Ω*, denoted* $A^c = \Omega \setminus A$*, is also in* $\mathcal{F}$*, i.e.,* $A^c \in \mathcal{F}$*.*
3. ***Closed under countable unions****: If* $A_1, A_2, A_3, \dots$ *are elements of* $\mathcal{F}$*, then their union is also in* $\mathcal{F}$*, i.e.,*

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Remark 6.3. *Considering the case* $\Omega = \mathbb{R}$*, note that the set* $I = \{]a, b[, a, b \in \mathbb{R} \cup \{-\infty, +\infty\}\}$ *is not a valid set of event. One has* $\Omega \in I$ *but the closeness under complementation is not satisfied since* $]0, \infty[\in I$ *but* $]-\infty, 0] = \mathbb{R} \setminus]0, \infty[\notin I$*. Performing iteratively the complementation and countable unions of elements of* I *can though produce an event space called the Borel algebra. The Borel Algebra is the smallest event space that contains all the intervals of* $\mathbb{R}$*. One can define similarly the Borel algebra in any* $\mathbb{R}^p$ *with* $p \in \mathbb{N}$ *as being the smallest events space containing all the open sets (and thus the closed sets) of* $\mathbb{R}^p$*. It is denoted* $\mathcal{B}(\mathbb{R}^p)$*. Without further specification, that will be the event space that we will consider in our examples.*

Definition 2 (Probability law)**.** *Let us consider* Ω *be a sample space, and a set of event* $\mathcal{F} \subset 2^{\Omega}$*. A probability law is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that satisfies the following properties:*

1. ***Non-negativity****:* $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{F}$*,*
2. ***Normalization****:* $\mathbb{P}(\Omega) = 1$*,*
3. ***Countable additivity****: If* $A_1, A_2, A_3, \dots \in \mathcal{F}$ *are mutually disjoint, then:*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Lemma 6.4. *Considering a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and given an event* $A \subset \mathcal{F}$*:*

$$\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$$

Proof. Since $A \subset \mathcal{F}$, $A \cap (\Omega \setminus A) = \emptyset$ one can express:

$$\mathbb{P}(\Omega \setminus A) + \mathbb{P}(A) = \mathbb{P}(A \cup (\omega \setminus A)) = \mathbb{P}(\Omega) = 1,$$

□

Lemma 6.5. *Considering a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and two event* $A \subset \mathcal{F}$*:*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Proof. Since $A \cup B = (A \setminus (A \cap B)) \cup (A \cap B) \cup (B \setminus (A \cap B)$ and all the left-hand sets are disjoint, one has the identity:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \setminus (A \cap B)) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus (A \cap B)).$$

One can then conclude thanks to the identities $\mathbb{P}(A \setminus (A \cap B)) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$ and $\mathbb{P}(B \setminus (A \cap B)) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. □

Definition 3. *Given two probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a dimension* $p \in \mathbb{N}$*, a random variable is a mapping* $X : \Omega \to \mathbb{R}^p$ *such that for all open set* $B \subset \mathbb{R}^p$*,* $X^{-1}(B)$ *is an event of* $\mathcal{F}$ *(*$X^{-1}(B) \equiv \{\omega \in \Omega, X(\omega) \in B\}$*).*

In measure theory, one would call random variables measurable mappings because the reciprocal image of Borel sets (that form the events space of $\mathbb{R}^p$) are in $\mathcal{F}$.

2 Example of Probability spaces and typical laws

We consider below a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given a random variable $X : \Omega \to \mathbb{R}$, the "law" of X, $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ is defined $\forall B \in \mathcal{B}(\mathbb{R})$ as:

$$\mathbb{P}_X(B) \equiv \mathbb{P}\left(X^{-1}(B)\right).$$

Then $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ form a Probability space.

Note that two random variable $X, Y : \Omega \to \mathbb{R}$ can have the same distribution (i.e. for all Borel set $B \subset \mathbb{R}$, $\mathbb{P}_X(B) = \mathbb{P}_Y(B)$) but still be different. Indeed $\mathbb{P}(X^{-1}(B)) = \mathbb{P}(Y^{-1}(B))$ does not imply $X^{-1}(B) = Y^{-1}(B)$. For instance in Example 6.2, X and Y have the same distribution but are not equal.

2.1 Discrete distributions

For discrete subset $\mathcal{X} \subset \mathbb{R}$, the events space (the intersection of $\mathcal{B}(\mathbb{R})$ and $2^{\mathcal{X}}$) is exactly the set of all the subsets of $\mathcal{X}$: $2^{\mathcal{X}}$, it is then sufficient to define the values of $\mathbb{P}_X(x) \equiv \mathbb{P}_X(\{x\}) = \mathbb{P}(X^{-1}(\{x\})$ for all $x \in \mathcal{X}$.

- A **Bernoulli** random variable $X : \Omega \to \{0,1\}$ satisfies:

  $$\mathbb{P}(X = 1) = p \qquad \text{and} \qquad \mathbb{P}(X = 0) = 1 - p$$

  for a certain $p \in [0,1]$. We denote $X \sim \text{Ber}(p)$. When $p = \frac{1}{2}$, it can represent the result after throwing a coin (head or tail). The expectation computes:

  $$\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot 0 = p.$$

- A Rademacher random variable $X : \Omega \to \{-1, 1\}$ takes with equal probability the values 1 and -1:

  $$\mathbb{P}(X = 1) = \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}(X = -1) = \frac{1}{2}.$$

  It is very similar to Bernoulli random variables ($\frac{X+1}{2} \sim \text{Ber}(\frac{1}{2})$) and used to compute the famous "Rademacher complexity" that will be presented later in the course.

- Given $n \in \mathbb{N}$, a **Binomial** random variable $X : \Omega \to [n]$ satisfies:

  $$\forall k \in [n] : \mathbb{P}(X = k) = \binom{k}{n} p^k (1-p)^k.$$

  We denote $X \sim \text{Bin}(p, n)$. Given n independent random variables $X_1, \dots X_n \sim \text{Ber}(p)$, one has $X_1 + \cdots + X_n \sim \text{Bin}(p, n)$:

  $$\begin{aligned}\mathbb{P}(X_1 + \cdots + X_n = k) &= \sum_{I \subset [n], |I| = k} \mathbb{P}(\forall i \in I : \ X_i = 1, \forall j \in [n] \setminus I : X_j = 0) \\ &= \sum_{I \subset [n], |I| = k} \prod_{i \in I} \mathbb{P}(X_i = 1) \prod_{j \in [n] \setminus I} \mathbb{P}(X_j = 0) = \binom{k}{n} p^k (1-p)^{n-k}.\end{aligned}$$

  The expectation computes:

  $$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = np.$$

- A random variable $X : \omega \to \mathbb{N}$ follows a **Poisson** distribution of mean $\lambda \in \mathbb{R}$, and we denote $X \sim \text{Poi}(\lambda)$ iif:

  $$\forall k \in \mathbb{N} : \mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k}.$$

  It usually represents the number of random events that happen in a given interval of time. The expectation computes:

  $$\mathbb{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

2.2 Continuous random variables

To present laws of continuous random variables, it is convenient to introduce first the notion of probability density. It rely on an important result of measure theory: Radon-Nikodym Theorem.

Definition 4. *Given two measures $\mu, \nu : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$, one says that ν is absolutely continuous under μ and we denote $\nu \ll \mu$ iif:*

$$\forall B \in \mathcal{B}(\mathbb{R}) : \qquad \mu(B) = 0 \quad \Longrightarrow \quad \nu(B) = 0$$

Theorem 6.6 (Radon-Nikodym)**.** *Given two measures $\mu, \nu : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$, if $\nu \ll \mu$ then there exists a measurable mapping $f : \mathbb{R} \to \mathbb{R}$ such that:*

$$\forall B \in \mathcal{B}(\mathbb{R}) : \qquad \nu(B) = \int_B f(x) d\mu(x).$$

Continuous random variables are taking value in $\mathbb{R}$ that are continuous under the Lebesgue measure λ (the uniform measure on $\mathbb{R}$). That means (see Definition 4) that for any Borel set $B \subset \mathcal{B}(\mathbb{R})$:

$$\lambda(B) = \int_B dt = 0 \qquad \Longrightarrow \qquad \mathbb{P}_X(B) = 0.$$

Applying Theorem 6.6, one can then deduce the existence of a measurable mapping $p_X : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{P}_X(A) = \int_A p_X(x) dx,$$

p_X is called the density of X, and one some times denotes $d\mathbb{P}_X(x) = p_X(x)dx(x)$. Below we will present classical examples of laws of random variables $X : \Omega \to \mathbb{R}$ by expressing their density function p_X.

- A random variable $X : \omega \to \mathbb{R}$ is **Uniformly** distributed on an interval $[a, b]$, and we denote $X \sim \text{Unif}([a, b])$ iif:

  $$\forall x \in \mathbb{R} : p_X(x) = \frac{\mathbb{1}_{[a,b]}}{b-a},$$

  where, given $A \subset \mathbb{R}$, $\mathbb{1}_{x \in A} = 1$ if $x \in A$ and $\mathbb{1}_{x \in A} = 0$ otherwise.

  The expectation computes:

  $$\mathbb{E}[X] = \int_{\mathbb{R}} x d\mathbb{P}_X(x) = \frac{1}{b-a} \int_a^b x dx = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{b+a}{2}.$$

- A random variable $X : \omega \to \mathbb{R}$ follows a **Beta distribution** of parameters $a, b > 0$, and we denote $X \sim \text{Beta}(a, b)$ iif

  $$\forall x \in \mathbb{R} : p_X(x) = \frac{\mathbb{1}_{[0,1]}}{B(a,b)} x^{a-1}(1-x)^{b-1},$$

  where, $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}$. It is widely used in Bayesian statistics, because, as we will see, it is stable through posterior inferences.

  To express $B(a, b)$ in the case $a, b \in \mathbb{N}^*$ we will use the following recurrence relation obtained from an integration by parts when $a \geq 2$:

  $$\begin{aligned} B(a,b) &= \frac{a-1}{b} \int_0^1 x^{a-2}(1-x)^b dx - \frac{1}{b} \left[x^{a-1}(1-x)^b \right]_0^1 = \frac{a-1}{b} B(a-1, b+1) - 0 \\ &= \cdots = \frac{(a-1)(a-2)\cdots 1}{b(b+1)\cdots(b+a-2)} B(1, b+a-1) = \frac{(a-1)!(b-1)!}{(b+a-2)!} \int_0^1 (1-x)^{b+a-2} dx = \frac{(a-1)!(b-1)!}{(b+a-1)!}. \end{aligned}$$

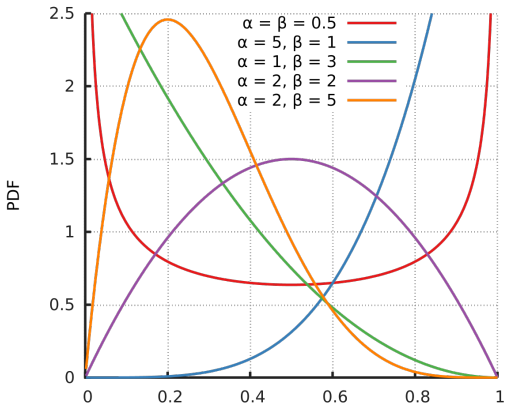

Figure 6.2: Depending on the values of $a, b > 0$, the expectation $\frac{a}{a+b}$ is either the point of highest density either the point of lowest density, in that last case the distribution has two modes in 0 and 1.

This identity can be extended to real positive values of $a, b > 0$ thanks to the Gamma function defined as $\forall x > 0$: $\Gamma(x) = \int_0^\infty t^{z-1} e^{-t} dt$, indeed, for any $n \in \mathbb{N}^*$, $\Gamma(n) = n!$. One then can show that:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The expectation computes:

$$\mathbb{E}[X] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} B(a+1,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{b+a}.$$

- The density of a Gaussian (or Normal) random variable $X : \omega \to \mathbb{R}$ with mean μ and variance σ^2 expresses:

$$\forall x \in \mathbb{R} : p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

then we denote $X \sim \mathrm{N}(\mu, \sigma^2)$. It is straightforward to check that $\mu = \mathbb{E}[X]$ and $\sigma = \mathbb{E}[(x-\mu)(X-\mu)]$ once one knows the identity $\int e^{t^2/2} = \sqrt{2\pi}$. Its cumulative distribution function is classically denoted "erf":

$$\mathrm{erf} : t \mapsto \frac{2}{\sqrt{\pi}} \int_{-\infty}^{t} e^{-u^2} du,$$

(if $X \sim \mathcal{N}(0,1)$, $\mathbb{P}(X \leq t) = \mathrm{erf}(t)$). The normal distribution is of particular interest because of the central limit Theorem (see Theorem 6.14 given later in the lecture).

- A random variable $X : \omega \to \mathbb{R}$ follows a **Student's t distribution** (or simply a t-distribution) of degree of freedom ν iif:

$$\forall x \in \mathbb{R} : p_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

and we denote $X \sim t(\nu)$. Note that $t(1)$ is the Cauchy distribution and $t(\nu)$ tends to $\mathcal{N}(0,1)$ when ν tends to ∞. Student random variable of degree of freedom $\nu = n-1$ with $n \in \mathbb{N}$ can be constructed

followingly. Consider n independent random variables $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ for given $\mu, \sigma \in \mathbb{R}$ and denote the sample mean and the unbiased estimate of the variance respectively:

$$\hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \hat{\sigma} \equiv \sqrt{\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}.$$

Then it can be shown that:

$$\sqrt{n-1}\cdot\frac{\hat{\mu}-\mu}{\hat{\sigma}} \sim t(n-1),$$

this remark is at the core of the Student's t test. The Student distribution is commonly used as an example of heavy tailed distribution, since one has:

$$X \sim t(\nu) \quad \Longrightarrow \quad \mathbb{E}[|X|^\nu] = \infty, \quad \forall r < \nu : \mathbb{E}[|X|^r] \leq \infty.$$

2.3 Multivariate continuous distributions

Continuous distribution in $\mathbb{R}^p$ are distributions absolutely continuous under the Lebesgue measure λ_p of $\mathbb{R}^p$ ($\forall B \in \mathcal{B}(\mathbb{R}^p)$, $\lambda_p(B) = \int_B dx_1 \cdots dx_p$). As for distribution in $\mathbb{R}$, Radon-Nikodym Theorem ensures they have a density. Given a random vector $X : \Omega \to \mathbb{R}^p$, we define its marginals as being the random variables $X_i : \Omega \to \mathbb{R}$, $i \in [p]$ satisfying:

$$\forall \omega \in \Omega : \quad X(\omega) = (X_1(\omega), \dots, X_p(\omega)) \in \mathbb{R}^p$$

. then, the density of X is some times denoted $p_{X_1,\dots,X_p} : \mathbb{R}^p \to \mathbb{R}_+$. It satisfies for all Borel set $B \in \mathcal{B}(\mathbb{R}^p)$:

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \int_B p_X(x) d\lambda_p(x) = \int_B p_{X_1,\dots,X_p}(x_1, \dots, x_p) dx_1 \cdots dx_p. \tag{6.1}$$

To give an example, consider the Gaussian random vector $X : \Omega \to \mathbb{R}^p$ of mean $\mu \in \mathbb{R}^p$ and covariance $\Sigma \in \mathcal{M}_p$ has a density:

$$\forall x \in \mathbb{R}^p : \qquad p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\Sigma)} \exp\left(-(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

then we denote $X \sim \mathcal{N}(\mu, \Sigma)$. Given a random vector $X = (X_1, \dots, X_p) : \Omega \to \mathbb{R}^p$, the random variables $X_i : \Omega \to \mathbb{R}$ are called the **marginals** of X they are usually not independent.

Lemma 6.7. *Given two continuous (possibly dependent) random vectors $X : \Omega \to \mathbb{R}^p$ and $y : \Omega \to \mathbb{R}^q$, with respective density p_X and p_Y, and for any $y \in \mathbb{R}^q$:*

$$\int_{\mathbb{R}^p} p_{X,Y}(x, y) dx = p_Y(y).$$

Proof. We know that for any $B \in \mathcal{B}(\mathbb{R}^q)$:

$$\int_{y\in B} p_Y(y) dy = \mathbb{P}(X \in \mathbb{R}^p, Y \in B) = \int_{y\in B}\left(\int_{x\in\mathbb{R}^p} p_{X,Y}(x, y) dx\right) dy,$$

thus, Radon-Nikodym Theorem allows us to conclude that:

$$\int_{x\in\mathbb{R}^p} p_{X,Y}(x, y) dx = p_Y(y)$$

for almost[1] all $y \in \mathbb{R}^q$. □

[1] That means that if we denote $A \equiv \{y \in \mathbb{R}^q : int_{x\in\mathbb{R}^p} p_{X,Y}(x, y) dx \neq p_Y(y)\}$, then $\lambda_q(A) = 0$, where λ_q is the Lebesgue measure in $\mathbb{R}^q$

In high dimension, there are some non intuitive phenomena that happens. It is important to have them in mind when doing machine learning where the most difficult problems are high dimensional. Sometimes called the curse of dimension, it can become a blessing when it is well understood (notably thanks to random matrix theory tools). In high dimension happens the so-called concentration of measure phenomenon that we illustrate below with the example of the norm of Gaussian vectors (but the same phenomenon happens for other functionals and other random vectors). One of the main contributors to the Theory of Concentration of the Measure, Michel Talagrand, explained it followingly: "A random variable that depends (in a "smooth" way) on the influence of many independent variable (but not too much on any of them) is essentially constant".

Let us consider here the case of a Gaussian vector $X = (X_1, \dots, X_n) \sim \mathcal{N}(0, I_n)$ and study the behavior of the random variable:

$$D = \|X\|^2 = \sum_{i=1}^{n} X_i^2.$$

We will show that the standard deviation of D is small as compared to its expectation. The variance of a random variable $Y : \Omega \to \mathbb{R}$ is commonly denoted:

$$\mathbb{V}[Y] \equiv \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

The expectation of D computes:

$$\mathbb{E}[D] = \sum_{i=1}^{n} \mathbb{E}[X_i^2] = 1,$$

and to compute the standard deviation, one can use the following lemma.

Lemma 6.8. *Given two independent variables $X, Y : \Omega \to \mathbb{R}$:*

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

Proof. Let us simply compute:

$$\begin{aligned}\mathbb{V}[X+Y] &= \mathbb{E}[(X + Y - \mathbb{E}[X] - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] - 2\underbrace{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}_{=0} + \mathbb{E}[(YY - \mathbb{E}[YY])^2] \ = \mathbb{V}[X] + \mathbb{V}[Y].\end{aligned}$$

□

Returning to our problem, one can show:

$$\mathbb{E}[D] = \sum_{i=1}^{n} \mathbb{V}[X_i^2] = \sum_{i=1}^{n} \mathbb{E}[(X_i^2 - \mathbb{E}[X_i^2])^2] = \sum_{i=1}^{n} \mathbb{E}[X_i^4] - \mathbb{E}[X_i^2]^2 = 2n,$$

since the fourth moment of $Y \sim \mathcal{N}(0, 1)$ can be computed and equals $\mathbb{E}[Y^4] = 3\mathbb{E}[Y^2] = 3$. Therefore, one has:

$$\frac{\sqrt{\mathbb{V}[D]}}{\mathbb{E}[D]} = \frac{\sqrt{2n}}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

One says that the squared norm of $X \sim \mathcal{N}(0, I_n)$ concentrates around n. The same holds for the non squared norm. And actually drawings of X will be more likely to lie close to the sphere $\sqrt{n}\mathbb{S}^{n-1} \equiv \{x \in \mathbb{R}^n, \|X\| = \sqrt{n}\}$ (see Figure 6.3, **(Right)** below). Note that the law of D is absolutely continuous under the Lebesgue

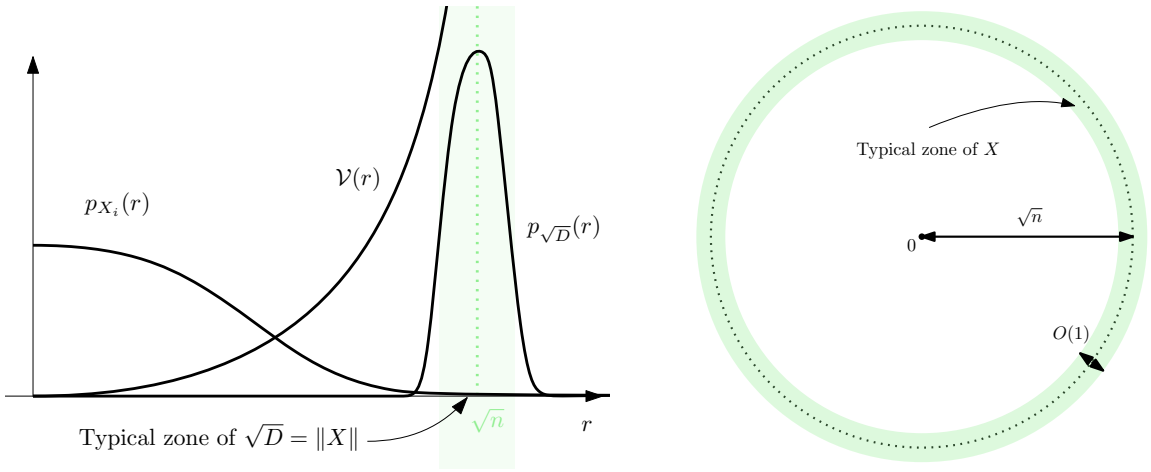

Figure 6.3: **(Left)** Schematic view of the density of one marginal X_i, of the norm $\sqrt{D}$ and of the volume $V(r)$ depending on the distance to 0: r. **(Right)** Schematic view of the distribution of $X \sim \mathcal{N}(0, I_n)$, it concentrates around the sphere $\sqrt{n}\mathbb{S}^{n-1}$.

measure on $\mathbb{R}$, one can therefore, thanks to Radon-Nikodym Theorem, introduce its density that satisfies for any interval $[a, b]$ of $\mathbb{R}$ with $0 \leq a \leq b$:

$$\begin{aligned}\int_a^b p_D(r)dr &= \mathbb{P}_D([a,b]) = \mathbb{P}(D \in [a,b]) \\ &= \int_{a \leq \|x\|^2 \leq b} p_X(x) d\lambda_n(x) \\ &= \frac{1}{(2\pi)^{n/2}} \int_{a \leq x_1^2 + \cdots + x_n^2 \leq b} e^{-\frac{1}{2}(x_1^2 + \cdots + x_n^2)} dx_1 \cdots dx_n.\end{aligned}$$

With the use of n-dimensional spherical coordinates, some integral calculus (which is out of the scope of this course) can then allow us to show that one can choose:

$$p_D(r) = \frac{r^{n-1}}{(2\pi)^{n/2}} e^{-\frac{r^2}{2}}.$$

This density is the product of a volume term $\mathcal{V}(r) = \frac{r^{n-1}}{(2\pi)^{(n-1)/2}}$ and the marginal density $p_{X_i}(r) = \frac{1}{(2\pi)^{1/2}} e^{-\frac{r^2}{2}}$. Those two influences implies that the maximum is reached around n (see Figure 6.3, **(Left)** that provides an intuition). Indeed one can compute the mode (i.e. $\text{Argmax}_{r>0} p_D(r)$) followingly:

$$p_D'(r) = 0 \quad \iff \quad (n-1) r^{n-2} r^{n-1} e^{-\frac{r^2}{2}} = r^n e^{-\frac{r^2}{2}} \quad \iff \quad r = (n-1).$$

Note that the mode $n-1$ is not equal to the mean n but still it is very close. More details of this phenomenon will be given in the lecture about concentration of the measure.

3 From sampling to estimates

3.1 Sampling with inverse distribution function

Definition 5. *Given a random variable $X : \Omega \to \mathbb{R}$, the cumulative distribution function – "the **cdf**" – of X, $F_X : \mathbb{R} \to [0,1]$ is the mapping defined for all $x \in \mathbb{R}$ as:*

$$F_X(x) = \mathbb{P}(X \leq t)$$

Lemma 6.9. *For continuous random variables $X : \Omega \to \mathbb{R}$, there is a one-to-one correspondence between the density p_X (also called the probability density function – "the **pdf**"), and the cdf F_X:*

$$\forall x \in \mathbb{R} : \qquad F_X(x) = \int_{-\infty}^{x} p_X(t)dt \quad \text{and} \quad p_X(x) = F'_X(x).$$

Lemma 6.10. *Given a random variable $U \sim Unif([0,1])$ and a strictly increasing and on-to[2] (thus invertible) mapping $F : \mathbb{R} \to [0,1]$ the cumulative distribution function of $F^{-1}(U)$ is exactly F:*

$$F_{F^{-1}(U)} = F$$

Proof. Given $u \in \mathbb{R}$, let us simply compute:

$$\mathbb{P}(F^{-1}(U) \geq u) = \mathbb{P}(U \geq F(u)) = \int_0^{F(u)} dt = F(u).$$

□

This Lemma is of high importance because it allows to simulate all the classical random variables thanks to the "pseudo random number generator" which is a method present in all computers in order to sample independent random variables uniformly in $[0,1]$. One then just has to apply F_X^{-1} to such samples to obtain independent samples of a given random variable $X : \Omega \to \mathbb{R}$.

In practice, more efficient methods like the "Box-Muller method" are generally employed to sample multidimensional random vectors. But this method is arguably too elaborate for a presentation in this course.

3.2 Rejection sampling

Let us consider a simple but representative example.

Example 6.11. *One wants to sample a variable $Y : \Omega \to \mathbb{R}$ having a Gaussian truncated density:*

$$f : y \mapsto \frac{e^{-y^2/2}\mathbb{1}_{[a,+\infty)}}{\int_a^{+\infty} e^{-t^2/2}dt},$$

for a given $a \in \mathbb{R}$. In the case of the Gaussian distribution, one can use precise numerical estimate of the erf function to compute $F_Y : t \mapsto \int_{-\infty}^{t} f(y)dy$. But let's say that we do not have those numerical tools (or that the distribution that is truncated is more complex than the Gaussian) there still exists a naive way to sample Y which is the following:

1. *Sample $X \sim \mathcal{N}(0,1)$ (it has the density $t \mapsto \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$),*
2. *if $X \geq a$ accept it, otherwise reject it.*

It can be shown that the set of the accepted samples following this procedure will indeed have as density f. However, if a is too big (let's say $a = 4.5$) then most of the sample of X will satisfy $X \leq 4.5$, there is indeed approximately a chance out of 10000 that $X \geq 4.5$. Therefore this procedure looks highly inefficient, and we propose below a more relevant sampling method.

In the previous subsection, given a random variable $Y : \omega \mapsto \mathbb{R}$, one has to know F_X in order to sample X. As pictured in Example 6.11, it often happens in statistical problems that one just has partial information on the density of X and, more importantly, that there is no close-form for the integration of this density: F_X. Typically, one knows that the density f is proportional to a certain function $\tilde{f} : \mathbb{R} \to \mathbb{R}_+$ ($\tilde{f} : y \mapsto e^{-y^2/2}\mathbb{1}_{[a,+\infty)}$ in Example 6.11). This proportional relation is denoted $f \propto \tilde{f}$, and that means that there exists a constant $C > 0$ such that $\forall x \in \mathbb{R}$: $f(x) = Cf(x)$. Of course, in that case, since $\int p_X = 1$, one has $C = \frac{1}{\int \tilde{f}}$.

[2]Here it means that $F(\mathbb{R}) = [0,1]$.

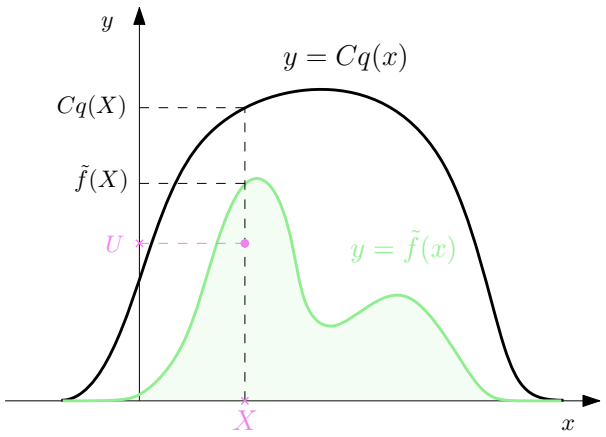

Figure 6.4: Two possible samples of U and X. $U \leq \tilde{f}(X)$ so we accept X.

Given a density $f \propto \tilde{f}$ (we have knowledge on $\tilde{f}$ but not on f) rejection sampling relies on the existence of an other density $q : \mathbb{R} \to \mathbb{R}$ such that there exists a constant $C > 0$ satisfying:

$$\forall x \in \mathbb{R} : \qquad f(x) \leq Cq(x).$$

Sample data set of the rejection sampling method is obtained iteratively following on each iteration those three steps:

1. sample x from the distribution X with density q,
2. sample u from $U \sim \text{Unif}([0, Cq(x)])$,
3. if $u \leq \tilde{f}(x)$ keep the value x in the final sample data set ("Accept") and if $u > \tilde{f}(x)$ reject the value x (see Figure 3.2).

Doing this iteration 5 times would be formalized with 5 independent random variables X_1, X_2, X_3, X_4, X_5 all having density q, 5 independent variables U_1, U_2, U_3, U_4, U_5 respectively dependent on X_1, X_2, X_3, X_4, X_5 since $\forall i \in [5]$, $U_i \sim \text{Unif}([0, Cq(X_i)])$. Then if, for a particular sample $w_1, w_2, w_3, w_4, w_5 \in \Omega$, one has:

$$U(w_1) \leq \tilde{f}(X(w_1)), \quad U(w_2) > \tilde{f}(X(w_2)), \quad U(w_3) \leq \tilde{f}(X(w_3)), \quad U(w_4) \leq \tilde{f}(X(w_4)), \quad U(w_5) > \tilde{f}(X(w_5)),$$

the kept values would be $\{X(w_1), X(w_3), X(w_4)\}$. This selection is done under the condition $U \leq \tilde{f}(X)$ (i.e. under the event $\{w \in \Omega, U(\omega) \leq \tilde{f}(X(\omega))\}$), therefore, the density of the associated random variable is the conditional density:

$$x \mapsto p_X(x \mid U \leq \tilde{f}(X)).$$

Proposition 6.12. *The random variable output from rejection sampling has density f.*

The following proof relies on the notation for conditional densities. Since the computation rules are quite intuitive, we allow ourselves to present them here although the rigorous definitions of such objects will be provided in Section 4.

Proof. One wants to compute:

$$p(Y = y | U \leq \tilde{f}(X)) = \frac{p(X = y, U \leq \tilde{f}(X))}{\mathbb{P}(U \leq \tilde{f}(X))} = \frac{p(U \leq \tilde{f}(y) | X = y) p_X(y)}{\mathbb{P}(U \leq \tilde{f}(X))}$$

By definition, $p_X(y) = q(y)$,

$$p(U \leq \tilde{f}(x_0) | X = y) = \frac{1}{Cq(y)} \int_0^{\tilde{f}(y)} dt = \frac{\tilde{f}(y)}{Cq(y)},$$

and:

$$\mathbb{P}(U \leq \tilde{f}(X)) = \int_{\mathbb{R}} p(U \leq \tilde{f}(x_0)|X = y)p_X(x)dx = \frac{\int \tilde{f}}{C}.$$

Finally one gets as expected:

$$p(Y = y \mid U \leq \tilde{f}(X)) = \frac{\tilde{f}(y)}{\int \tilde{f}} = f(y).$$

□

3.3 Monte Carlo estimation - problem setting

A common problem in statistical learning is to estimate expectations $\mu \equiv \mathbb{E}[h(X)]$ for a random variable $X : \Omega \to \mathbb{R}$ having a density f and a certain measurable mapping $h : \Omega \to \mathbb{R}$. The law of large numbers gives us:

Theorem 6.13 (Law of large numbers)**.** *Given any sequence of i.i.d. random variables $X_1, X_2, \ldots : \Omega \to \mathbb{R}$, for all $\varepsilon > 0$:*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right) \underset{n\to\infty}{0}.$$

Therefore a first naive idea is to use as estimator for $\mu = \mathbb{E}[X]$ the random variable $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If there were no computation limitation, Theorem 6.13 would ensure that for n big enough, $\hat{\mu}$ can be as close to μ as needed. However the big limitation of Theorem 6.13 is that it does not provides the speed of convergence, and we will see in Example 6.15 that the variance (and thus the speed) varies a lot depending on the estimator. Actually, X_1 is already an estimator for μ since $\mathbb{E}[X_1] = \mu$, however the big improvement with $\hat{\mu}$ is that (when the variance is bounded):

$$\mathbb{V}[\hat{\mu}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[X_1].$$

Therefore when n tends to ∞, the random variable $\hat{\mu}$ concentrates around μ with a standard deviation of order $1/\sqrt{n}$. This speed is arguably too slow for a lot of applications, that is why we will present in this subsection other methods to devise efficient estimators.

When n is big enough, the central limit Theorem provides a limiting confidence interval.

Theorem 6.14 (Central limit)**.** *Given a sequence of independent identically distributed random variables $(X_i)_{i\in\mathbb{N}} : \Omega \to \mathbb{R}^{\mathbb{N}}$ such that $(\forall i \in \mathbb{N})$ $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[(X_i - \mu)^2] = \sigma^2 \leq \infty$ then for all $t \in \mathbb{R}$:*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \leq t\right) = erf(t)\,.$$

Denoting the sample (here biased) standard deviation:

$$\hat{\sigma} \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \hat{\mu})^2},$$

one has, for n big enough, the pseudo confidence interval:

$$\mathbb{P}\left(\hat{\mu} - 1.96\frac{\hat{\sigma}}{n} \leq \mu \leq \hat{\mu} + 1.96\frac{\hat{\sigma}}{n}\right) \geq 0.95.$$

Once again, the question is to know what does it means exactly for n to be "big enough".

Let us give some examples to understand the issues it better.

Example 6.15. *Let us consider a Cauchy random vector $X \sim \frac{1}{\pi}\frac{1}{1+x^2}$ and we want to estimate the probability:*

$$\theta \equiv \mathbb{P}(X \geq 2)$$

we will provide below several estimators and compute for all of them the variance to be able to compare them. They all depend on a sampling of n i.i.d. random variables $X_1, \dots, X_n \sim \frac{1}{\pi}\frac{1}{1+x^2}$.

- $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \geq 2}$. *Of course $\mathbb{E}[\hat{\theta}_1] = \mathbb{P}(X \geq 2)$ and the variance computes:*

$$\mathbb{V}[\hat{\theta}_1] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}[\mathbb{1}_{X_i \geq 2}] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\mathbb{1}^2_{X_i \geq 2}] - \mathbb{E}[\mathbb{1}_{X_i \geq 2}]^2 = \frac{1}{n}(\theta - \theta^2),$$

  since $\mathbb{1}^2_{X_i \geq 2} = \mathbb{1}_{X_i \geq 2}$. Numerically (integrating the Cauchy distribution), that gives us $\mathbb{V}[\hat{\theta}_1] \approx \frac{0.120}{n}$.

- *Noting that $\theta = \frac{1}{\pi}\int_2^{+\infty}\frac{1}{1+x^2} = \frac{1}{2\pi}\left(\int_{-\infty}^{-2}\frac{1}{1+x^2} + \int_2^{+\infty}\frac{1}{1+x^2}\right)$, one may rather choose $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{|X_i| \geq 2}$. Once again, $\mathbb{E}[\hat{\theta}_2] = \mathbb{P}(X \geq 2) = \theta$ and the variance here computes:*

$$\mathbb{V}[\hat{\theta}_2] = \frac{1}{4n}\mathbb{E}[\mathbb{1}_{|X_i| \geq 2}] - \mathbb{E}[\mathbb{1}_{|X_i| \geq 2}]^2 = \frac{1}{2n}(\theta - 4\theta^2),$$

  That gives us $\mathbb{V}[\hat{\theta}_2] \approx \frac{0.58}{n} \leq \mathbb{V}[\theta_1]$.

- *Going further, we note that $\theta = \frac{1}{2}\left(1 - \frac{1}{\pi}\int_{-2}^2 \frac{1}{1+x^2}\right)$ and then rather choose $\hat{\theta}_3 \equiv \frac{1}{2}(1 - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{|X_i| \leq 2})$. With this choice one can reduce a bit more the variance and obtain $\mathbb{V}[\hat{\theta}_3] \approx \frac{0.002}{n}$.*

- *Finally with the change of variable $x \to \frac{1}{x}$, one has the new identity $\theta = \frac{1}{\pi}\int_0^{\frac{1}{2}}\frac{1/x^2}{1+1/x^2}dx = \frac{1}{\pi}\int_0^{\frac{1}{2}}\frac{1}{1+x^2}dx$ and the associated estimator would be $\hat{\theta}_4 \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{0 \leq X_i \leq \frac{1}{2}}$. Here an estimation of the variance gives us $\mathbb{V}[\hat{\theta}_4] \approx \frac{10^{-4}}{n}$.*

With those example we saw that on practical examples, it is generally possible to find some small tricks to find estimator with comparatively small standard deviation (from $\hat{\theta}_1$ to $\hat{\theta}_4$ the standard deviation reduces by a factor $\sqrt{10^3} \approx 32$ – that means that one needs 32 times less samples to estimate θ with the same precision error). We will see in next subsection some more systematic techniques to optimize estimators

3.4 Importance sampling

Considering a measurable mapping $h : \mathbb{R} \to \mathbb{R}$ and a density $f : \mathbb{R} \to [0, 1]$, we try to estimate $\int_{\mathbb{R}} h(x)f(x)dx$. Importance sampling relies on the following identity, true for any $g : \mathbb{R} \to \mathbb{R}$, measurable and such that $g(x) = 0 \implies f(x) = 0$ (the measure induced by f is absolutely continuous under the measure induced by g):

$$\int_{\mathbb{R}} h(x)f(x)dx = \int_{\mathbb{R}} \frac{h(x)f(x)}{g(x)}g(x)dx.$$

The idea is then to play on g in order to find the best estimator possible. The easiest way to compare estimators is found through a comparison of the variance, and the optimal estimator to that respect is given by next theorem.

Theorem 6.16 (Rubenstein)**.** *Given a measurable mapping $h : \mathbb{R} \to \mathbb{R}$ and a density $f : \mathbb{R} \to \mathbb{R}_+$, the density $g : \mathbb{R} \to \mathbb{R}_+$ that minimizes the variance of $\frac{h(X)f(X)}{g(X)}$ for $X \sim g$ is:*

$$g^* : x \mapsto \frac{|h(x)|f(x)}{\int_{\mathbb{R}} |h(t)|f(t)dt}.$$

This theorem is proven thanks to Jensen inequality.

Lemma 6.17 (Jensen inequality)**.** *Given a convex mapping $\phi : \mathbb{R} \to \mathbb{R}$ and a random variable $X : \Omega \to \mathbb{R}$ one has the inequality:*

$$\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\phi(X)].$$

For concave mappings $\phi : \mathbb{R} \to \mathbb{R}$, one has the converse identity.

Proof of Theorem 6.16. Given a random variable $X \sim g$, one can first note that:

$$\mathbb{V}\left[\frac{h(X)f(X)}{g(X)}\right] = \mathbb{E}\left[\left(\frac{h(X)f(X)}{g(X)}\right)^2\right] - \mathbb{E}\left[\frac{h(X)f(X)}{g(X)}\right]^2.$$

The term $\mathbb{E}\left[\frac{h(X)f(X)}{g(X)}\right] = \int_{\mathbb{R}} h(x)f(x)dx$ is independent of g, one therefore just tries to minimize the first term that satisfies thanks to Jensen inequality:

$$\mathbb{E}\left[\left(\frac{h(X)f(X)}{g(X)}\right)^2\right] \geq \mathbb{E}\left[\frac{|h(X)|f(X)}{g(X)}\right]^2.$$

To conclude, first remark that $\mathbb{E}\left[\frac{|h(X)|f(X)}{g(X)}\right]^2$ is independent of g and that it is equal to $\mathbb{E}\left[\left(\frac{h(X)f(X)}{g(X)}\right)^2\right]$ when $g = g^*$. That exactly implies that g^* is the density that minimize the variance. Be careful that $x \mapsto \frac{h(x)f(x)}{\int_{\mathbb{R}} h(t)f(t)dt}$ is not a density, we need to take the absolute value of $h(x)$ to ensure that $g^*(x)$ is positive. □

4 Conditioning

4.1 Independence

We are still placed in a probability space $(\omega, \mathcal{F}, \mathbb{P})$.

Definition 6. *Two events $A, B \in \mathcal{F}$ are said to be independent iif:*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Remark 6.18.

- *Events $A \in \mathcal{F}$ of null probability are independent with any other events $B \in \mathcal{F}$ since $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) = 0$*
- *Two distinct events of non null probability are not independent since $\mathbb{P}(A)\mathbb{P}(B) \neq 0 = \mathbb{P}(A \cap B)$*
- *(See Figure 6.5). Let us consider the case $\Omega = [0,1] \times [0,1]$ and $\mathbb{P}$ is uniformly distributed on Ω. Given $A \in \mathcal{F}$, $\mathbb{P}(A)$ is exactly the area of A. Given any intervals $[x, x+a], [y, y+b] \subset [0,1]$, the events $A = [x, x+a] \times [0,1]$ and $B = [0,1] \times [y, y+b]$ are independent since one has:*

$$\mathbb{P}(A \cap B) = a \cdot b = (a \cdot 1)(1 \cdot b) = \mathbb{P}(A)\mathbb{P}(B).$$

The independence between random variables relies on the notion of independence of events.

Definition 7. *Two random variables $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$ are said to be independent iif for all Borel sets $A \in \mathcal{B}(\mathbb{R}^p)$ and $B \in \mathcal{B}(\mathbb{R}^q)$, $X^{-1}(A)$ and $Y^{-1}(B)$ are independent.*

Given an event $A \in \mathcal{F}$, we say that X is independent with A if A is independent with $X^{-1}(E)$ for all $E \in \mathcal{B}(\mathbb{R})$.

Example 6.19.

- *We already presented in Example 6.2 two non independent random variables when $\Omega = [0, 2\pi)$, $X : \omega \mapsto \cos(\omega)$ and $Y : \omega \mapsto \sin(\omega)$.*

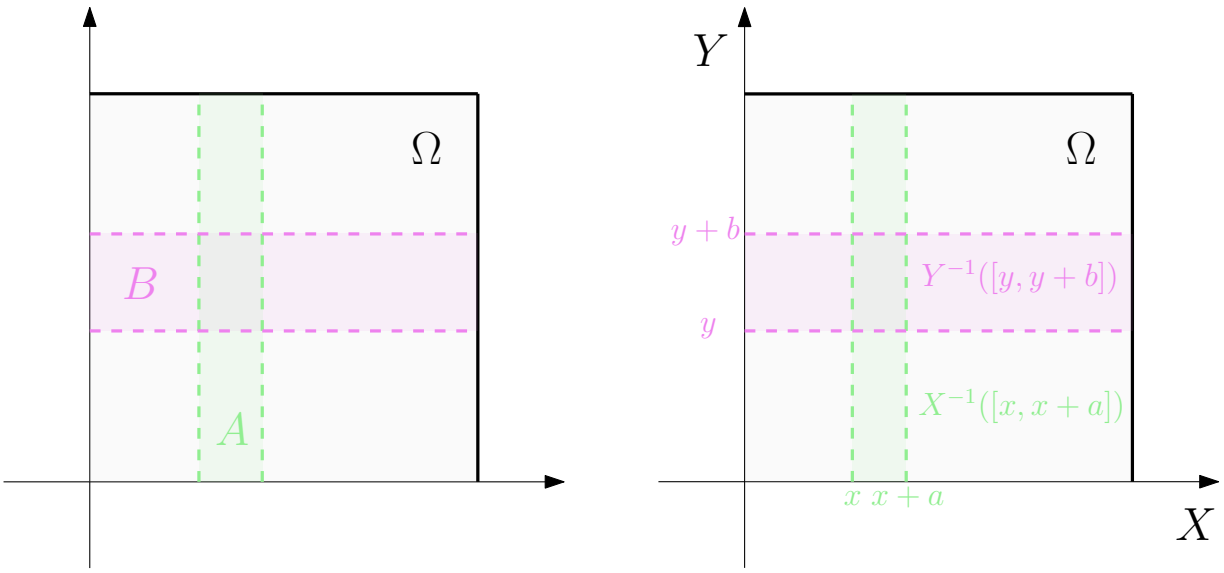

Figure 6.5: **(left)** Independence of A and B towards the uniform distribution on $\Omega = [0,1]^2$. **(right)** Independence of $X : (\omega_1, \omega_2) \mapsto \omega_1$ and $Y : (\omega_1, \omega_2) \mapsto \omega_2$.

- *(See Figure 6.5). To present a simple example of two independent random variables, one can consider again the uniform probability on $\omega = [0,1] \times [0,1]$ setting $X : (\omega_1, \omega_2) \mapsto \omega_1$ and $Y : (\omega_1, \omega_2) \mapsto \omega_2$. Then for any intervals $[x, x+a], [y, y+b] \subset [0,1]$, one knows that $X^{-1}([x, x+a]) = [x, x+a] \times [0,1]$ and $Y^{-1}([y, y+b]) = [0,1] \times [y, y+b]$ are independent. Since by definition of the Borel space, all events are union, intersections or complementary of intervals, one can deduce that the independence between $X^{-1}(A)$ and $Y^{-1}(B)$ generalizes to any Borel set $A, B \in \mathcal{B}([0,1])$.*

Proposition 6.20. *Two random variables $X : \omega \to \mathbb{R}^p$ and $X : \omega \to \mathbb{R}^q$ are independent iif for all measurable bounded mappings $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^q \to \mathbb{R}$:*

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

Proof. Let us do the proof in the real case $p = q = 1$.

The sufficient condition is easily satisfied since one can consider for any Borel sets $A, B \in \mathcal{B}(\mathbb{R})$ the mappings $f : x \mapsto \mathbb{1}_A$ and $g : y \mapsto \mathbb{1}_B$, then one has:

$$\mathbb{P}(X^{-1}(A) \cap Y^{-1}(B)) = \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)] = \mathbb{P}(X^{-1}(A))\mathbb{P}(Y^{-1}(B)). \tag{6.2}$$

For the necessary condition, one can first show it for mappings $f : x \mapsto \mathbb{1}_A$ and $g : y \mapsto \mathbb{1}_B$ for certain events $A, B \in \mathcal{B}(\mathbb{R})$ exactly with the same identity as (6.2). Then one can approximate any general bounded measurable mapping $f : \mathbb{R} \to \mathbb{R}$ with a series $\sum_{i=1}^n a_i \mathbb{1}_{A_i}$ for a certain sequence of parameters $(a_i)_{i \in \mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ and a certain sequence of events $(A_i)_{i \in \mathbb{N}} \in (\mathbb{R})^{\mathbb{N}}$. The same way we can approximate g with $\sum_{i=1}^n b_i \mathbb{1}_{B_i}$. One can then compute thanks to the linearity of the expectation:

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \lim_{n \to \infty} \mathbb{E}\left[\sum_{i=1}^n a_i \mathbb{1}_{A_i} \sum_{j=1}^n b_j \mathbb{1}_{B_j}\right] \\
&= \lim_{n \to \infty} \sum_{i=1}^n \sum_{j=1}^n a_i b_j \mathbb{E}[\mathbb{1}_{A_i} \mathbb{1}_{B_j}] \\
&= \lim_{n \to \infty} \sum_{i=1}^n \sum_{j=1}^n a_i b_j \mathbb{E}[\mathbb{1}_{A_i}]\mathbb{E}[\mathbb{1}_{B_j}] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].
\end{aligned}$$

□

This theoretical definition of independence aims at formalizing the drawing of independent events like two tossing of a coin. The result of the first try can be represented by a first random variable $X_1 : \Omega \to \{0, 1\}$ and the second try by an other *independent* random variable $X_2 : \Omega \to \{0, 1\}$. This formalism is only allowed because it is always possible to consider an independent copy of any independent vector.

Generally two random variables are not independent, a tool is though available to understand their dependence in that case: conditioning.

4.2 Conditioning

One can find a lot of different notations of conditional probabilities like:

$\mathbb{P}(A\|B)$	$\mathbb{P}(A \mid Y \in F)$	$p(A \mid Y = y)$
$p(X = x\|B)$ or $p_X(x\|B)$	$p(X = x \mid Y \in F)$ or $p_X(x\|Y \in B)$	$p(X = x \mid Y = y)$ or $p_{X,Y}(x\|y)$,

for $A, B \in \mathcal{F}$ two events, $X, Y \to \mathbb{R}$, two continuous random variables, $y, x \in \mathbb{R}$ two scalars and a Borel set $F \in \mathcal{B}(\mathbb{R})$. We will define all of these notation below. The two first one are defined easily and intuitively:

- $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ if $\mathbb{P}(B) \neq 0$, $\mathbb{P}(A|B) = 0$ otherwise.
- Naturally, $\mathbb{P}(A \mid Y \in F) = \mathbb{P}(A \mid Y^{-1}(F))$.

To define the other notations that are actually conditional *densities*, one has to define first:

$$p(X = x, A) \text{ also denoted } p_X(x, A) \qquad \text{and} \qquad p(X = x, Y = y) \text{ also denoted } p_{X,Y}(x, y).$$

The latter notation was already introduced in (6.1), in Subsection 2.3, it just represents the joint density of X and Y. The first notation requires more work to be rigorously defined. In the case where $X^{-1}(X(A)) = A$, it is simply:

$$p_X(x,\ A) = p_X(x)\mathbb{1}_A(x) \tag{6.3}$$

The following remark gives the definition in the general case.

Remark 6.21 (Elaborate definition of $p_X(x,\ A)$)**.** *Let us introduce the measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, 1]$ satisfying for all $E \in \mathcal{B}(\mathbb{R})$:*

$$\mu(E) = \mathbb{P}(X \in E,\ A) = \mathbb{P}(X^{-1}(E) \cap A).$$

We know that this measure is absolutely continuous under $\mathbb{P}_X$. Indeed: $\mathbb{P}_X(E) = 0 \implies \mathbb{P}(X^{-1}(E) \cap A) \leq \mathbb{P}_X(E) = 0$. Therefore, Radon-Nikodym Theorem provides the existence of a mapping $f : \mathbb{R} \to \mathbb{R}_+$ such that:

$$\forall F \in \mathcal{B}(\mathbb{R}) : \quad \mathbb{P}(X \in E,\ A) = \mu(E) = \int_{\mathbb{R}} f(x) d\mathbb{P}_X(x) = \int_{\mathbb{R}} f(x) p_X(x) dx.$$

The mapping $x \mapsto f(x)p_X(x)$ is then denoted $p(X = x, A)$.

Note that if $\mathbb{P}(A) < 1$, it is not a density since:

$$\int_{\mathbb{R}} p(X = x, A) dx = \mathbb{P}(X \in \mathbb{R},\ A) = \mathbb{P}(A) \tag{6.4}$$

We can know define:

- $p(A \mid Y = y) = \frac{p_Y(y,\ A)}{p_Y(y)}$ when $p_Y(y) \neq 0$ and $p(A \mid Y = y) = 0$ otherwise.
- $p(X = x|B) = \frac{p_X(x,B)}{\mathbb{P}(B)}$ if $\mathbb{P}(B) \neq 0$ and $p(X = x|B) = 0$ otherwise
- $p(X = x \mid Y \in F) = p(X = x \mid Y^{-1}(F))$

- $p(X = x \mid Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$ if $p_Y(y) \neq 0$, and $p(X = x \mid Y = y) = 0$ otherwise..

First let us show that those conditioned densities are probability densities in the following sense.

Lemma 6.22. *Given two continuous (possibly dependent) random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$, with respective density p_X and p_Y, and an event $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$, the conditional densities $x \mapsto p(X = x|A)$ and $x \mapsto p(X = x|Y = y)$ are both density functions.*

Proof. One just needs to check that $\int_{\mathbb{R}^p} p(X = x|A)dx = \int_{\mathbb{R}^p} p(X = x|Y = y)dx = 1$. Let us first compute:

$$\int_{\mathbb{R}^p} p(X = x|A)dx = \frac{1}{p(A)} \int_{\mathbb{R}^p} p_X(x, A)dx = 1$$

thanks to (6.4). Second:

$$\int_{\mathbb{R}^p} p(X = x|Y = y)dx = \frac{1}{p(Y = y)} \int_{\mathbb{R}^p} p(X = x, Y = y)dx = 1,$$

thanks to Lemma 6.7. □

One often uses the notation $\mathbb{P}(X = x|Y)$ (or $p_X(x|Y)$ for continuous variables X) which are actually random variables that could be seen as measurable mappings of Y:

$$\mathbb{P}(X = x|Y) : \omega \mapsto \mathbb{P}(X = x|Y = Y(\omega)) \qquad (\text{or} \quad p_X(x|Y) : \omega \mapsto p_X(x|Y = Y(\omega)) \text{ if } X \text{ is continuous}$$

. In a sens those conditional probabilities provide exactly the distribution of X if Y was fixed. As any other random variables, one can integrate them to obtain the following lemma.

Lemma 6.23. *Given three random variables $X, Z, Y : \Omega \to \mathbb{R}$ with X discrete and Z continuous:*

$$\mathbb{E}[\mathbb{P}(X = x|Y)] = \mathbb{P}(X = x) \qquad \text{and} \qquad \mathbb{E}[p_Z(z|Y)] = p_Z(z).$$

Proof. Let us simply do the proof when Y is continuous:

$$\mathbb{E}[p_Z(z|Y)] = \int_{y\in\mathbb{R}} p_Z(z|Y = y)p_Y(y)dy = \int_{y\in\mathbb{R}} p_Z(z, Y = y)dy = p_Z(z),$$

by definition of conditioned probabilities and joint probabilities. □

The following lemma justifies the fact that conditioning is only relevant for non independent random variables and events

Lemma 6.24. *Given two independents events $A, B \in \mathcal{F}$ such that $\mathbb{P}(B) \neq 0$:*

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

given two independent continuous random variables $X, Y : \Omega \to \mathbb{R}$ and two scalars $x, y \in \mathbb{R}$ such that $p_Y(y) \neq 0$:

$$p_{X,Y}(x|y) = p_X(x).$$

Besides, if X is independent with B:

$$p_X(x|B) = p_X(x).$$

To study conditioning with Gaussian random vectors, one needs the following important linear algebra result.

Lemma 6.25 (Schur)**.** *The inverse of a block matrix $M = \left(\begin{smallmatrix} E & F \\ G & H \end{smallmatrix}\right)$ expresses:*

$$M^{-1} = \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix}$$

where $M/H \equiv E - FH^{-1}G$.

Corollary 6.26. *the inverse of a block matrix* $M = \left(\begin{smallmatrix} E & F \\ G & H \end{smallmatrix}\right)$*, decomposes:*

$$M^{-1} = \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} \cdot \begin{pmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \cdot \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix}$$

and therefore $\det(\Sigma) = \det(H)\det(M/H)$*.*

Proof. Let us simply compute:

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} \cdot \begin{pmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \cdot \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} &= \begin{pmatrix} (M/H)^{-1} & 0 \\ -H^{-1}G(M/H)^{-1} & H^{-1} \end{pmatrix} \cdot \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix}. \end{aligned}$$

One can then conclude with Lemma 6.25. □

Example 6.27. *Given a Gaussian random vector* $X \sim \mathcal{N}(\mu, \Sigma)$ *where* $\mu \in \mathbb{R}^{p+q}$ *and* $\Sigma \in \mathbb{R}^{(p+q)\times(p+q)}$*, one can be interested in decomposing the vector* X *in two parts* $X_1 : \Omega \to \mathbb{R}^p$ *and* $X_2 : \Omega \to \mathbb{R}^q$ *to try and express the density of* X_1 *conditioned on* X_2*. For that we will need the block decomposition:*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad\qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^p$*,* $\mu_2 \in \mathbb{R}^q$*,* $\Sigma_{11} \in \mathbb{R}^{p\times p}$*,* $\Sigma_{12} \in \mathbb{R}^{p\times q}$*,* $\Sigma_{21} \in \mathbb{R}^{q\times p}$ *and* $\Sigma_{22} \in \mathbb{R}^{q\times q}$*. Above we just defined conditional densities for distributions in* $\mathbb{R}$ *but the definitions extend naturally to continuous distributions in* $\mathbb{R}^p$*, for* $p \geq 2$*.*

Let us express:

$$p_{X_1,X_2}(x_1, x_2) = \frac{1}{(2\pi)^{\frac{p+q}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \cdot \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \cdot \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right).$$

Let us first look at the terms depending on x_1, x_2 *and simply express thanks to Corollary 6.26:*

$$\begin{aligned} p_{X_1,X_2}(x_1, x_2) &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \cdot \begin{pmatrix} I_p & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_q \end{pmatrix} \cdot \begin{pmatrix} (\Sigma/\Sigma_{22}) & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \cdot \begin{pmatrix} I_p & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_q \end{pmatrix} \cdot \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(x_1 - \mu_{1|2}\right)^T \Sigma_{1|2}^{-1}\left(x_1 - \mu_{1|2}\right)\right) \cdot \exp\left(\frac{1}{2}(x_2 - \mu_2)^T\Sigma_{22}^{-1}(x_2 - \mu_2)\right). \end{aligned}$$

with:

- $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$
- $\Sigma_{1|2} = \Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma 21$*.*

We know also from Corollary 6.26 that:

$$\frac{1}{(2\pi)^{\frac{p+q}{2}}\sqrt{|\Sigma|}} = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma/\Sigma_{22}|}} \frac{1}{(2\pi)^{\frac{q}{2}}\sqrt{|\Sigma_{22}|}}.$$

Finally one can deduce from the identity $p_{X_1,X_2}(x_1, x_2) = p_{X_1,X_2}(x_1|x_2)p_{X_2}(x_2)$ *that:*

$$p_{X_1,X_2}(x_1|x_2) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma_{1|2}|}} \exp\left(-\frac{1}{2}\left(x_1 - \mu_{1|2}\right)^T \Sigma_{1|2}^{-1}\left(x_1 - \mu_{1|2}\right)\right),$$

Fixing $x_2 \in \mathbb{R}$*, we recognize here a Gaussian distribution* $\mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$*.*

4.3 Conditional Expectation and conditional Independence

Given two, say, continuous, random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$, and an event $A \in \mathcal{F}$, the conditional expectation of X knowing A and the conditional independence of X and Y knowing A are defined the same way as without the condition, the densities $p_X(x|A)$, $p_Y(y|A)$, and $p_{X,Y}(x,y|A)$ given $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ are merely replacing the densities $p_X(x)$, $p_Y(y)$, and $p_{X,Y}(x,y)$ in these definition. For simplicity, the definitions below are only provided for continuous random vectors, they can be naturally extended to more general distribution.

Definition 8 (Conditional expectation)**.** *Given an event $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$ and a random vector $X : \Omega \to \mathbb{R}^p$ the conditional expectation of X knowing A is defined as:*

$$\mathbb{E}[X|A] = \frac{1}{\mathbb{P}(A)} \int_A X(\omega) d\mathbb{P}(x), = \frac{1}{\mathbb{P}(A)} \int_{\mathbb{R}^p} x p_X(x, A) dx,$$

Given a supplementary continuous random variable $Y : \Omega \to \mathbb{R}^q$ and $y \in \mathbb{R}^q$:

$$\mathbb{E}[Y|X = x] = \int_{\mathbb{R}^q} y p_{Y,X}(y|x) dy.$$

Following these definitions, a frequently used object is the random vector:

$$\begin{aligned} \mathbb{E}[Y|X] : \Omega &\longrightarrow \mathbb{R}^q \\ \omega &\longmapsto \mathbb{E}[Y|X = X(\omega)]. \end{aligned}$$

It satisfies the following property:

Lemma 6.28. *Given two random vector $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$:*

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[X]$$

and for all $f : \mathbb{R}^p \to \mathbb{R}$, measurable:

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

Proof. With the definitions, one can express:

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}[Y|X]\right] &= \int_{\mathbb{R}^p} \left(\int_{\mathbb{R}^q} y p_{Y,X}(y|x) dx \right) p_X(x) dx \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} y p_{Y,X}(y,x) dy dx \quad = \mathbb{E}[Y]. \end{aligned}$$

Besides, given $\omega \in \Omega$:

$$\begin{aligned} \mathbb{E}[Y f(X)|X](\omega) &= \int_{\mathbb{R}^q} y f(x) p_{X,Y}((x,y)|X = X(\omega)) dx dy = \int_{\mathbb{R}^q} y f(x) \frac{p_{X,Y}((x,y), X = X(\omega))}{p_X(X(\omega))} dx dy \\ &= \frac{f(X(\omega))}{p_X(X(\omega))} \int_{\mathbb{R}^q} y p_Y(y, X = X(\omega)) dy = f(X(\omega)) \int_{\mathbb{R}^q} y p_Y(y|X = X(\omega)) dy \\ &= f(X(\omega))\mathbb{E}[Y|X = X(\omega)], \end{aligned}$$

thanks to Lemma 6.7. □

Lemma 6.29. *Given two independent random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$:*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y],$$

Proof. We know from Lemma 6.24 that for all $\omega \in \Omega$, $p_Y(y|X = X(\omega)) = p_{Y,X}(y|X(\omega)) = p_Y(y)$, therefore:

$$\mathbb{E}[Y|X = x] = \int_{\mathbb{R}^q} y p_{Y,X}(y|x) dy = \int_{\mathbb{R}^q} y p_Y(y) dy = \mathbb{E}[Y].$$

□

Example 6.30. *Given four parameters $\mu, \nu, \sigma, \theta \in \mathbb{R}$, $\sigma, \theta > 0$ and considering $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z \sim (\nu, \theta^2)$, the random variable $Y = X + Z$ satisfies:*

$$\mathbb{E}[Y] = \mu + \nu \qquad \mathbb{E}[Y \mid X] = X + \nu$$
$$\mathbb{V}[Y] = \sigma^2 + \theta^2 \qquad \mathbb{E}[(Y - X - \nu)^2 \mid X] = \theta^2,$$

one then denotes $Y \sim \mathcal{N}(X + \nu, \theta^2 \mid X)$.

In the previous example it is interesting to see that the variance of Y, conditionally on X, $\mathbb{V}[Y|X] \equiv \mathbb{E}[(Y - \mathbb{E}[Y|X])^2 \mid X]$, is independent of X. It is a simple consequence of the fact that $Y = X + Z$ and that X is independent with Z. We formalize below the notion of conditional independence.

Definition 9 (Conditional independence)**.** *Given three events $A, B, C \in \mathcal{F}$, one says that A is independent of B conditionally on C iif:*

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C).$$

Given three independent random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$ and $Z : \Omega \to \mathbb{R}^r$, one says that Y is independent of Y conditionally on X iif for any measurable mappings $f : \mathbb{R}^r \to \mathbb{R}$ and $g : \mathbb{R}^q \to \mathbb{R}$:

$$\mathbb{E}\left[f(Z)g(Y) \mid X\right] = \mathbb{E}\left[f(Z) \mid X\right]\mathbb{E}\left[g(Y) \mid X\right].$$

Example 6.31. *Let us give an example of two independent random vectors that are not independent conditionally on a well chosen event. The example is pictured on Figure 6.6. For that let us consider the sample set $\Omega = [0,1] \times [0,1]$ and the random variables $X, Y : \Omega \to [0,1]$ satisfying for all $(\omega_1, \omega_2) \in \Omega$, $X((\omega_1, \omega_2)) = \omega_1$ and $Y((\omega_1, \omega_2)) = \omega_2$. Considering $A = \{(\omega_1, \omega_2) \in \Omega, \omega_2 \leq \omega_2\}$ and following the introduction of $p_{X,Y}((x,y), A)$ in[3] (6.3), let us express:*

$$p_{X,Y}((x,y), A) = \mathbb{1}_A((x,y))p_{X,Y}(x,y) = \mathbb{1}_A((x,y)).$$

Therefore one can compute:

$$p_X(x, A) = \int_0^1 p_{X,Y}((x,y), A)dy = \int_0^1 \mathbb{1}_A((x,y))dy = \int_0^x dy = x$$

(be careful, here we do not have $p_X(x, A) = \mathbb{1}_{X^{-1}(A)}(x)p_X(x)$ because $X^{-1}(X(A)) \neq A$). The same way, one has $p_Y(y, A) = 1 - y$ and knowing that $\mathbb{P}(A) = \frac{1}{2}$, one can deduce:

$$p_X(x, A) = 2x \qquad \text{and} \qquad p_Y(y, A) = 2(1 - y).$$

One can then integrate to get the conditional expectations:

$$\mathbb{E}[X|A] = \int_0^1 x p_X(x, A)dx = 2\int_0^1 x^2 dx = \frac{2}{3} \qquad \text{and} \qquad \mathbb{E}[Y|A] = \frac{1}{3}.$$

However:

$$\begin{aligned}\mathbb{E}[XY|A] &= \int_0^1 \int_0^1 xy p_{X,Y}((x,y)|A)dxdy \\ &= \frac{1}{\mathbb{P}(A)}\int_0^1 x\left(\int_0^x ydy\right)dx = \frac{1}{\mathbb{P}(A)}\int_0^1 \frac{x^3}{2}dx = \frac{1}{4} \neq \frac{2}{9} = \mathbb{E}[X|A]\ \mathbb{E}[Y|A].\end{aligned}$$

We see that X, Y are not independent conditionally on A.

Lemma 6.32. *Given two random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$, for any measurable mapping $f : \mathbb{R}^p \to \mathbb{R}$, $f(X)$ is independent of Y, conditionally on X.*

[3] Here $A = (X,Y)^{-1}((X,Y)(A))$, where (X,Y) is a random vector satisfying for all $\omega \in \Omega$, $(X,Y)(\omega) = (X(\omega), Y(\omega))$.

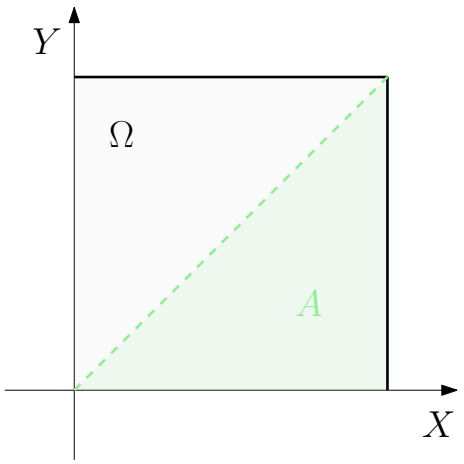

Figure 6.6: For the uniform measure on $\Omega = [0,1]^2$, the random variables X and Y are independent but they are not independent conditionally on $A = \{\omega \in \Omega, Y(\omega) \leq X(\omega)\}$.

Proof. It is a simple consequence of Lemma 6.28, given two measurable mappings $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^q \to \mathbb{R}$:

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

□

Lemma 6.33. *Given three independent random vectors $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$ and $Z : \Omega \to \mathbb{R}^r$, Y and Z are also conditionally independent with X.*

Remark 6.34. *In the setting of Lemma 6.33, one can infer for instance that for any measurable mappings $\phi : \mathbb{R}^p \to \mathbb{R}^q$ and $\psi : \mathbb{R}^p \to \mathbb{R}^r$, $Y + \phi(X)$ and $Z + \psi(X)$ are independent conditionally on X.*

Proof of Lemma 6.33. It is a simple consequence of Lemma 6.29, for any mappings $f : \mathbb{R}^r \to \mathbb{R}$ and $g : \mathbb{R}^q \to \mathbb{R}$, $f(Z)$ and $g(Y)$ are both independents of X and therefore:

$$\mathbb{E}\left[f(Z)g(Y)|X\right] = \mathbb{E}\left[f(Z)g(Y)\right] = \mathbb{E}\left[f(Z)\right]\mathbb{E}\left[g(Y)\right] = \mathbb{E}\left[f(Z)|X\right]\ \mathbb{E}\left[g(Y)|X\right].$$

□

With the spirit of Remark 6.34, we are then able to construct a lot of conditionally independent random vectors, then their behavior is very similar to classically independent random vectors, one can for instance provide the proposition on the summation of the variance of conditionally independent random vectors very similar to Lemma 6.8.

Lemma 6.35. *Given two random variables $Z, Y : \Omega \to \mathbb{R}$ independent conditionally on a third random vector $X : \Omega \to \mathbb{R}$:*

$$\mathbb{V}[Z + Y|X] = \mathbb{V}[Z|X] + \mathbb{V}[Y|X]$$

Proof. Let us simply compute:

$$\begin{aligned}
\mathbb{V}[Z+Y|X] &= \mathbb{E}[(Z+Y-\mathbb{E}[Z|X]-\mathbb{E}[Y|X])^2|X] \\
&= \mathbb{E}[(Z-\mathbb{E}[Z|X])^2|X] - 2\mathbb{E}[(Z-\mathbb{E}[Z|X])(Y-\mathbb{E}[Y|X])|X] + \mathbb{E}[(Y-\mathbb{E}[Y|X])^2|X] \\
&= \mathbb{E}[(Z-\mathbb{E}[Z|X])^2|X] - 2\underbrace{\mathbb{E}[Z-\mathbb{E}[Z|X]|X]}_{=0}\ \underbrace{\mathbb{E}[Y-\mathbb{E}[Y|X]|X]}_{=0} + \mathbb{E}[(Y-\mathbb{E}[Y|X])^2|X] \\
&= \mathbb{V}[Z|X] + \mathbb{V}[Y|X].
\end{aligned}$$

□