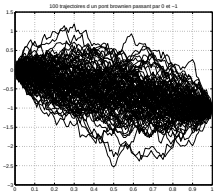
Randal DOUC

MONTE CARLO AND ADVANCED NUMERICAL SIMULATION METHODS


Telecom Sudparis, Institut Polytechnique de Paris.

Contents

Plan of action

1. Day 1: 4H30
    a. Lecture 1H30: Recap on Measures, Integration, Random Variables, independence.
    b. Tutorials 1H30: Exercises.
    c. Computer Session 1H30: Histograms and sampling of a random variable.
2. Day 2: 4H30
    a. Lecture 1H30: LLN, Central Limit Theorem, Confidence Intervals.
    b. Tutorials 1H30: Exact sampling, the quantile function. Rejection sampling: exercise.
    c. Computer Session 1H30: LLN, CLT and confidence intervals. Exact sampling.
3. Day 3: 3H00
    a. Lecture 1H30: Approximate Sampling, Importance Sampling. Monte Carlo by Markov chains.
    b. Tutorials 1H30: Approximate Sampling, Markov chains, exercise.
4. Day 4: 4H30
    a. Lecture 1H30: Variance reduction (I): antithetic, control variates.
    b. Tutorials 1H30: Markov chains and antithetic variables.
    c. Computer Session 1H30: Importance sampling and MCMC.
5. Day 5: 4H30
    a. Lecture 1H30: Variance reduction (II): conditioning, stratified sampling
    b. Tutorials 1H30: Exercises on variance reduction.
    c. Computer Session 1H30: Variance reduction (II)

Chapter 1

Preliminaries

These lecture notes are built from various lecture notes, thoses of Bernard Lapeyre and Denis Talay, Eric Moulines and Gersende Fort, Emmanuel Temam, Jérôme Lelong, Bruno Bouchard, etc., all mixed up with various personal touches. It has been actively proofread by Cyrille Dubarry.

Do not hesitate to point out the errors or typos that still remain and to propose any improvements on the content of this course.

Chapter 2

Refresh on Mesure theory and Lebesgue integration

Contents

Keywords: *Sigma-fields, measures, probability measures, measurable functions, random variables, expectations. Weak convergence, almost sure convergence, Law of Large numbers, central limit theorems, confidence intervals.*

In this chapter, we provide a short recap on measure and integration theory with a special focus on their application in probability theory.

2.1 Sigma-fields, Measures and Probability

Let us start with the definition of sigma-fields...

Definition 2.1. Let Ω be a given set. We say that a family of sets $\mathscr{F} \subset \mathscr{P}(\Omega)$ is a sigma-field on Ω if and only if the three following properties are satisfied

(i) $\Omega \in \mathscr{F}$
(ii) if $A \in \mathscr{F}$ then $\bar{A} = \Omega \setminus A \in \mathscr{F}$.
(iii) if for all $i \in \mathbb{N}$, $A_i \in \mathscr{F}$ then $\cap_{i\in\mathbb{N}} A_i \in \mathscr{F}$

We then say that $(\Omega, \mathscr{F})$ is a measurable space.

A sigma-field is stable by complementary sets, countable intersection, countable union and also by taking the "set difference" $\setminus$ in the sense that if $A, B \in \mathscr{F}$, then $A \setminus B \in \mathscr{F}$ (indeed, just write $A \setminus B = A \cap B^c$.)

► **Q 1**: Do those properties have special names?
The second property is often called *stability by complementary sets* and the last one *stability by countable intersection*. You may also find in the literature some other *equivalent* definitions:

(i) $\emptyset \in \mathscr{F}$
(ii) if $A \in \mathscr{F}$ then $\bar{A} = \Omega \setminus A \in \mathscr{F}$.
(iii) if for all $i \in \mathbb{N}$, $A_i \in \mathscr{F}$ then $\cup_{i\in\mathbb{N}} A_i \in \mathscr{F}$

But I prefer the way it is expressed in Theorem 2.1.

► **Q 2**: Why do you need these properties?

In the theory of probability, a set A will correspond typically correspond to an event that may occur, it can be expressed as a constraint with respect to all the possibilities. The fact that we ask the stability by complementary sets or by countable intersections corresponds to considering either the complementary event or the fact that all the events A_i are satisfied.

► **Q 3**: Do you have any examples?

The smallest sigma-field is $\mathscr{F} = \{\Omega, \emptyset\}$ and the largest one is $\mathscr{P}(\Omega)$. Often, sigma-fields we are interested in are generated by some family of sets...

► **Q 4**: What do you mean ?

Say that you are interested in a family of sets $\mathscr{C} \subset \mathscr{P}(\Omega)$ but unfortunately, some of the 3 properties that define sigma-fields are not satisfied for $\mathscr{C}$. In that case, we can still include $\mathscr{C}$ into a larger family so that it is a sigma-field. We can even consider the "smallest" one (in a sense to be defined), which contains all the sets in $\mathscr{C}$.

Definition 2.2. Let $\mathscr{C} \subset \mathscr{P}(\Omega)$. There exists a sigma-field, named $\sigma(\mathscr{C})$ which contains $\mathscr{C}$ and which is minimal for the inclusion, that is, any other sigma-field that contains $\mathscr{C}$ also contains $\sigma(\Omega)$. We then say that $\sigma(\mathscr{C})$ is the sigma-field generated by $\mathscr{C}$.

A good exercise is to prove the property that appears in the above definition. This can be done by defining $\sigma(\mathscr{C})$ as the collection of all the sets which are common wrt to all the sigma fields that contain $\mathscr{C}$. In other words, define

$$\mathscr{A} = \cap\{\mathscr{T} : \mathscr{C} \subset \mathscr{T} \text{ and } \mathscr{T} \text{ is a sigma-field}\}$$

Even though $\mathscr{A}$ is defined with an uncountable intersection, you may check that $\mathscr{A}$ is sigma-field, that this sigma-field contains all the sets in $\mathscr{C}$ and that any other sigma-field that contains all the sets in $\mathscr{C}$ necessarily contains all the sets in $\mathscr{A}$. All in all, $\mathscr{A}$ is the smallest sigma-field satisfying this property and we can call it $\sigma(\mathscr{C})$.

► **Q 5**: What is your favorite example?

An important example is the case of open sets. If Ω is $\mathbb{R}^k$ and $\mathscr{C}$ is the family of open sets on Ω, then the sigma-field generated by open sets is called the Borel sigma-field and is noted $\mathscr{B}(\Omega)$.

► **Q 6**: For different family of sets, if you consider the sigma-fields generated by each of them, do you systematically find different sigma-fields?

Of course not... In practice, if we have two family of sets $\mathscr{C} \subset \mathscr{P}(\Omega)$ and $\mathscr{D} \subset \mathscr{P}(\Omega)$ and if we want to check if $\sigma(\mathscr{C}) = \sigma(\mathscr{D})$ then a necessary and sufficient condition for getting that is to check successively that $\mathscr{D} \subset \sigma(\mathscr{C})$ and $\mathscr{C} \subset \sigma(\mathscr{D})$. You can use this property for checking that the sigma-field generated by open sets (i.e. the Borel sigma-field) is also the sigma-field generated by closed sets.

► **Q 7**: That's nice. You can prove it easily?

Yes, please do so. It's a good way to check that everything is understandable.

Definition 2.3. Let $(\Omega, \mathscr{F})$ be a measurable space. We say that a function $\mu : \mathscr{F} \to \bar{\mathbb{R}}^+ := \mathbb{R}^+ \cup \{\infty\}$ is a measure if it satisfies the *sigma-additivity property*, that is for any family of sets (A_i) such that $A_i \in \mathscr{F}$ for any $i \in \mathbb{N}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$\mu\left(\cup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i) \tag{2.1}$$

We then say that $(\Omega, \mathscr{F}, \mu)$ is a measured space. Moreover, if $\mu(\Omega) = 1$ then we say that μ is a probability measure.

Note that in the right hand side of (2.1), we sum quantities in $\bar{\mathbb{R}}^+$ that is, we use the convention that if $a \in \mathbb{R}^+$, $a + \infty = \infty$ and $\infty + \infty = \infty$.

► **Q 8**: The measure μ evaluated on a set A can be infinite?

Yes of course, $\mu(A)$ takes its values between 0 and $\mu(\Omega)$ actually... But if you consider a measure of probability (which is nothing but a particular measure), then the values of $\mu(A)$ are between 0 and 1.

► **Q 9**: Some useful properties?

(i) If $A \subset B$, then $\mu(A) \leq \mu(B)$.
(ii) $\mu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.
(iii) $\mu(\cup_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mu(\cup_{i=1}^{n} A_i)$
(iv) if $\mu(A_1) < \infty$, then, $\mu(\cap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mu(\cap_{i=1}^{n} A_i)$

► **Q 10**: What are the typical measures I will deal with?
There are at least two fundamental examples of measures:

(i) The Dirac measure on a which is defined by $\delta_a(A) = 0$ if $a \notin A$ and 1 otherwise.
(ii) The Lebesgue measure λ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ which is defined with the following characterizing property: it is the only measure such that for any segment $A = [a,b]$, we have $\lambda(A) = b - a$. Then, you can show quite easily that the Lebesgue measure of any interval (the interval may be closed, open or none of them) is the length of the interval (which thus may be infinite, take $A = [1, \infty[$ for example).

From these measures, you can construct other measures, for example by multiplying them by some non negative measurable functions.

2.2 Integrals, random variables, expectation

► **Q 11**: Measurable functions?
Yes. The definition is below.

Definition 2.4. If $(A, \mathscr{A})$ and $(B, \mathscr{B})$ are measurable sets. We say that $h : A \to B$ is a $\mathscr{A}/\mathscr{B}$ measurable function if and only if for any $B \in \mathscr{B}$, $f^{-1}(B) \in \mathscr{A}$.

Of course, if $\mathscr{B} = \sigma(\mathscr{C})$, then instead of checking for any $B \in \mathscr{B}$, $f^{-1}(B) \in \mathscr{A}$, we may only check for any $C \in \mathscr{C}$, $f^{-1}(C) \in \mathscr{A}$. For a given measurable function f, we may define $\sigma\left\{f^{-1}(B) : B \in \mathscr{B}\right\}$, it turns out that it is a sigma field, called $\sigma(f)$. Measurable functions are linked with random variables...
► **Q 12**: Can you be more precise?
Here is the definition of a random variable.

Definition 2.5. Let $(\Omega, \mathscr{F})$ be a measurable space and consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. A random variable $X : \Omega \to \mathbb{R}$ is, by definition, a $\mathscr{F}/\mathscr{B}(\mathbb{R})$ measurable function.

► **Q 13**: You mean that a random variable is nothing more than a measurable function?
Yes, and it is always real valued. If X takes its values in $\mathbb{R}^k$, we don't speak about random variables but about random vectors and/or (in some books) about random elements. Now that we have defined general measures with the important particular case of the probability measures, we can define the *integral* associated to a measure μ and if μ is probability measure, we will defined the *expectation* of this random variable.

Recall that a measure μ associates to a set $A \in \mathscr{F}$, a real number $\mu(A)$. Now, we want μ to associate to any real-valued $\mathscr{F}/\mathscr{B}(\mathbb{R})$-measurable function f a real number called $\mu(f)$ or $\int_\Omega f(w)\mu(\mathrm{d}w)$. There is here an abuse of notation... Stricto sensu: $\mu(A)$ is well defined but $\mu(f)$ is an abuse of notation because between brackets, we write a function f and not a set A. To avoid confusions, most of the time we write $\int f(w)\mu(\mathrm{d}w)$ or $\int f \mathrm{d}\mu$ (to avoid w) instead of $\mu(f)$ but all these notation of course refer to the same object.

Actually, it will be not be possible to define $\mu(f)$ for any measurable function... Let us be more precise. The construction of the integral wrt μ will be done progressively. We start with the $\mu(\mathbb{1}_A)$. By definition, we set:

$$\mu(\mathbb{1}_A) = \mu(A)$$

Then, we will define

$$\mu(\sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}) = \sum_{i=1}^{n} \alpha_i \mu(A_i)$$

Then, we define for any measurable *non-negative* function f,

$$\mu(f) = \sup \left\{ \mu(\sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}) : \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i} \leq f \right\}$$

Then, for any measurable function such that $\mu(|f|) < \infty$, we set

$$\mu(f) = \mu(f^+) - \mu(f^-)$$

► **Q 14**: You always integrate on the whole space?
Yes, but if you integrate a function f on a subset $\Omega_0 \in \mathscr{F}$ where $\Omega_0 \subset \Omega$, then by definition, it just means $\int f(w)\mathbb{1}_{\Omega_0}(w)\mu(\mathrm{d}w)$. That is you integrate on the whole space but thanks to the indicator function $\mathbb{1}_{\Omega_0}$, only the values of f on Ω_0 are meaningful.

► **Q 15**: You told me that two examples of measures were important.
So what?
► **Q 16**: What are the integrals associated to those measures?
Our two important examples of integrals constructed from measures μ are

(i) Integrals associated to Dirac measures... We can show that for any $a \in \Omega$ and any measurable function f, $\int f(w)\delta_a(\mathrm{d}w) = f(a)$.
(ii) Integrals associated to the Lebesgue measure... This is a common case, and instead of writing $\int f(w)\lambda(\mathrm{d}w)$, we usually write $\int f(w)\mathrm{d}w$.

► **Q 17**: OK for the construction of the integral but what are the essential properties?
I guess the very essential ones allow to interchange limit and integral. Two of them are essential:

(i) The monotone convergence theorem: if $\{f_n,\ n \in \mathbb{N}\} n \in \mathbb{N}$ is a family of measurable *non-negative* functions and $f_n \leq f_{n+1}$ for all large enough n, then $\int \lim_{n\to\infty} f_n(w)\mu(\mathrm{d}w) = \lim_{n\to\infty} \int f_n(w)\mu(\mathrm{d}w)$
(ii) The Lebesgue dominated convergence theorem: if $\{f_n : n \in \mathbb{N}\}$ is a family of measurable functions such that $\lim_{n\to\infty} f_n(w)$ exists for μ-almost all $w \in \Omega$ and if $|f_n| \leq h$ where $\int h d\mu < \infty$, then $\int \lim_{n\to\infty} f_n(w)\mu(\mathrm{d}w) = \lim_{n\to\infty} \int f_n(w)\mu(\mathrm{d}w)$

But there are also essential properties: the linearity of the integral, or if $f \leq g$ then, $\mu(f) \leq \mu(g)$, or $|\mu(f)| \leq \mu(|f|)$. Or if $\mu(|f|) < \infty$ then, $|f(w)| < \infty$ for μ-almost all $w \in \Omega$.

► **Q 18**: You said that when you multiply a non-negative measurable function and a measure, it is a measure... What do you mean exactly?
If f is a non-negative measurable function, then, the measure $f\mathrm{d}\mu$ is defined by: $A \mapsto \int \mathbb{1}_A(w)f(w)\mu(\mathrm{d}w)$. Therefore, with the two typical measures (δ_a and λ), you can define so many different measures $f\mathrm{d}\delta_0$ or $f\mathrm{d}\lambda$ by using a non-negative measurable function f.

► **Q 19**: You told me there is some link between the expectation operator associated to a probability measure and the integral associated to a measure.
The expectation is defined in the same way as integrals associated to some measure: it is constructed from a measured space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable X (that is a $\mathscr{F}/\mathscr{B}(\mathbb{R})$-measurable function). Then, by definition, $\mathbb{E}[X]$ is just the integral associated to the measure $\mathbb{P}$ taken at the random variable X, that is, $\mathbb{E}[X] = \int_\Omega X(w)\mathbb{P}(\mathrm{d}w)$.

Definition 2.6. The law of a random variable X on $(\Omega, \mathscr{F}, \mathbb{P})$ is the measure μ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ defined by: $\mu : A \mapsto \mathbb{P}(X \in A)$.

In this definition, we use $\mathbb{P}(X \in A)$ which means $\mathbb{P}(\Omega_0)$ where $\Omega_0 = \{w \in \Omega : X(w) \in A\} := \{X \in A\}$. We also call this measure, the push-forward measure of $\mathbb{P}$ through the function X. It defines a measure on the arrival sigma field.

► **Q 20**: How do you check that two random variables are independent?
Actually, two random vectors X, Y defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ but taking values in $\mathbb{R}^p$ and $\mathbb{R}^d$ are independent if and only one the equivalent properties are satisfied:

(i) For all $(A,B) \in \mathscr{B}(\mathbb{R}^p) \times \mathscr{B}(\mathbb{R}^d)$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

(ii) For all bounded or non-negative $\mathscr{B}(\mathbb{R}^p)/\mathscr{B}(\mathbb{R})$-measurable functions $f : \mathbb{R}^p \to \mathbb{R}$ and for all $\mathscr{B}(\mathbb{R}^d)/\mathscr{B}(\mathbb{R}))$-measurable bounded or non-negative functions $g : \mathbb{R}^d \to \mathbb{R}$, we have

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

(iii) for all $(u,v) \in \mathbb{R}^p \times \mathbb{R}^d$,

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{i}u^T X + \mathrm{i}v^T Y}\right] = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}u^T X}\right]\mathbb{E}\left[\mathrm{e}^{\mathrm{i}v^T Y}\right]$$

(iv) for all $(u,v) \in \mathbb{R}^p \times \mathbb{R}^d$ and all $(x,y) \in \mathbb{R}^2$,

$$\mathbb{P}(u^T X \le x, v^T Y \le y) = \mathbb{P}(u^T X \le x)\mathbb{P}(v^T Y \le y)$$

We have $X \overset{\mathscr{L}}{=} Y$ (where X and Y are two random variables) if and only if any of the following conditions holds true:

(i) $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for any $A \in \mathscr{B}(\mathbb{R})$
(ii) $\mathbb{P}(X \le t) = \mathbb{P}(Y \le t)$ for any $t \in \mathbb{R}$. Note that $t \mapsto \mathbb{P}(X \le t)$ is the cumulative distribution function for the random variable X.
(iii) $\mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right] = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uY}\right]$ for any $u \in \mathbb{R}$. Note that $u \mapsto \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right]$ is the characteristic function for the random variable X.

2.3 Convergence

We start with some notation. In what follows,

- i.i.d means independent and identically distributed.
- r.v. means random variables.
- for $r,s \in \mathbb{N}$ such that $r \le s$, we write $[r:s] = \{r, r+1, \ldots, s\}$,
- $X \perp\!\!\!\perp Y$ means X and Y are independent random variables,
- $X \overset{\mathscr{L}}{=} Y$ means X and Y have the same law.
- $\liminf_n a_n = \lim_{n\to\infty}(\inf_{k\ge n} a_k)$ and similarly, $\limsup_n a_n = \lim_{n\to\infty}\left(\sup_{k\ge n} a_k\right)$. Moreover, $\lim_n a_n$ exists if and only if $\liminf_n a_n = \limsup_n a_n$.
- for any $a \in \mathbb{R}$, $a^+ = \max(a,0)$ and $a^- = \max(-a,0) = -\min(a,0)$ and we have $|a| = a^+ + a^-$ and $a = a^+ - a^-$.

Moreover, the following notions of convergence for random variables is used throughout these lecture notes.

► $\boxed{X_n \overset{w}{\Rightarrow} X}$ means *convergence in distribution* (or "convergence en loi" in French). It is equivalent to any of the following statements.

(a) for all bounded continuous functions h, we have $\lim_n \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$.
(b) for all $A \in \mathscr{B}(\mathbb{R})$ such that $\mathbb{P}(X \in \partial A) = 0$, we have $\lim_n \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$.

(c) for all $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$, we have $\lim_n \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x)$.
(d) for all $u \in \mathbb{R}$, we have $\lim_n \mathbb{E}\left[e^{iuX_n}\right] = \mathbb{E}\left[e^{iuX}\right]$

By abuse of terminology, we may also say that $\boxed{X_n \textbf{ weakly converges to } X}$ instead of saying the distribution of X_n converges weakly to the distribution of X.

- $\boxed{X_n \overset{\mathbb{P}\text{-prob}}{\longrightarrow} X}$ means *convergence in probability*: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

- $\boxed{X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X}$ means *almost sure convergence*:

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1\ .$$

$\boxed{\text{Almost sure convergence implies convergence in probability}}$. We also recall the following properties.

(i) If $X_n \overset{w}{\Rightarrow} X$ then for all continuous functions f, $f(X_n) \overset{w}{\Rightarrow} f(X)$. Note that this property holds, when f is continuous (and not necessarily bounded), for example $f(u) = u^2$ so that $X_n^2 \overset{w}{\Rightarrow} X^2$.
(ii) **The Slutsky Lemma** If $X_n \overset{\mathbb{P}\text{-prob}}{\longrightarrow} c$ where c is a constant and if $Y_n \overset{w}{\Rightarrow} Y$, then $(X_n, Y_n) \overset{w}{\Rightarrow} (c, Y)$ that is for all continuous functions f, $f(X_n, Y_n) \overset{w}{\Rightarrow} f(c, Y)$.
(iii) $X \sim \mathrm{N}(0,1)$ iif for all $u \ge 0$, $\mathbb{E}\left[e^{iuX}\right] = e^{-u^2/2}$. Moreover, $X \sim \mathrm{N}(\mu, \sigma^2)$ iif for all $u \ge 0$, $\mathbb{E}\left[e^{iuX}\right] = e^{-u^2\mathbb{V}\mathrm{ar}(X)/2 + iu\mathbb{E}(X)}$ and in that case, $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$ and $\mu = \mathbb{E}(X)$.

2.4 Some usual distributions

Name	Acronym	Parameter	density function: $f_X(x)$	cdf: $F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u$	Other properties
Gaussian	$\mathrm{N}(\mu, \sigma^2)$	(μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$	No explicit expression	$\mathbb{E}[X] = \mu$, $\mathbb{V}\mathrm{ar}X = \sigma^2$
Exponential	$\exp(\lambda)$	$\lambda > 0$	$\lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$	$\left(1 - e^{-\lambda x}\right) \mathbb{1}_{\mathbb{R}^+}(x)$	$\mathbb{E}[X] = 1/\lambda$, $\mathbb{V}\mathrm{ar}X = 1/\lambda^2$
Gamma	$\Gamma(k, \theta)$	$(k, \theta) \in \left(\mathbb{R}_+^*\right)^2$	$\frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}$	$\frac{\Gamma_{x/\theta}(k)}{\Gamma(k)}$	

In the above description,

(i) if $X_i \sim \Gamma(k_i, \theta)$ and (X_i) are independent, then $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n k_i, \theta)$.
(ii)

$$\Gamma(k) = \begin{cases} \int_0^\infty t^{k-1} e^{-t} \mathrm{d}t & \text{if } k \in \mathbb{R}_+^* \\ k! & \text{if } k \in \mathbb{N}. \end{cases} \quad (\blacktriangleright \textsc{Gamma function})$$

$$\Gamma_x(k) = \int_0^x t^{k-1} e^{-t} \mathrm{d}t \ (\blacktriangleright \textsc{incomplete Gamma function})$$

Chapter 3

Introduction to the Monte Carlo methods

Contents

Keywords: *LLN,CLT, Slutski's lemma, unbiased estimator, Delta-method, confidence intervals.*

Many probabilistic issues that arise in statistical applications boil down to the calculation of expectations. For example, in Bayesian inference for hidden models, we might be interested in the expectation of some function of the parameter wrt to a posteriori distribution. In that context, it's quite natural to focus on numerical methods, which allow to calculate these expectations, considering that in most cases closed-form formulae are not available.

3.1 Principle of the method

Monte Carlo methods are used to numerically calculate expectations. These methods are based on the celebrated Strong Law of the Large Numbers.

Theorem 3.1. (►Strong Law of Large Numbers (LLN)). *Let $(X_n)_n$ be i.i.d random variables with the same law as X. If $\mathbb{E}[|X|] < \infty$, then*

$$\bar{X}_N = N^{-1} \sum_{i=1}^{N} X_i \xrightarrow{\mathbb{P}\text{-}a.s.} \mathbb{E}[X]$$

The convergence also holds in L^1: $\mathbb{E}\left[|\bar{X}_N - \mathbb{E}(X)|\right] \to 0$.

Remark 3.2 *The integrability assumption is mandatory, indeed we can show that if you consider an i.i.d. sequence of r.v. according to a Cauchy distribution, then the empirical mean does not converge. Please write an R program for illustrating this property.*

Proof. We will prove only the a.s. convergence. We start with an elementary result:

Lemma 3.3 A preliminary result: *Let (Y_i) be iid random variables such that $\mathbb{E}[|Y_1|] < \infty$ and $\mathbb{E}[Y_1] > 0$, then a.s.,*

$$\liminf_n S_n/n \geq 0$$

where $S_n = \sum_{i=1}^n Y_i$.

► *(Proof of the lemma)* Set $L_n = \inf(S_k, k \in [1:n])$, $L_\infty = \inf(S_k, k \in \mathbb{N}^*)$, $A = \{L_\infty = -\infty\}$. Let $\theta(y_1, y_2, \dots,) = (y_2, y_3, \dots)$ be the shift operator. Then, a.s.,

$$\begin{aligned} L_n &= S_1 + \inf(0, S_2 - S_1, \dots, S_n - S_1) = Y_1 + \inf(0, L_{n-1} \circ \theta) \\ &\geq Y_1 + \inf(0, L_n \circ \theta) = Y_1 - L_n^- \circ \theta. \end{aligned}$$

where the inequality follows from the fact that $n \mapsto L_n$ is nonincreasing. This implies a.s. (since $L_n^- \circ \theta$ is a.s. finite)

$$1_A Y_1 \leq 1_A L_n + 1_A L_n^- \circ \theta$$

Taking the expectation on both sides and then, using $\mathbb{P}(1_A = 1_A \circ \theta) = 1$, and the strong stationarity of the sequence:

$$\mathbb{E}[1_A Y_1] \leq \mathbb{E}[1_A L_n] + \mathbb{E}\left[1_A \circ \theta \, L_n^- \circ \theta\right] = \mathbb{E}[1_A L_n] + \mathbb{E}\left[1_A L_n^-\right] = \mathbb{E}\left[1_A L_n^+\right] \to 0$$

where the right-hand side tends to 0 by the dominated convergence theorem since a.s. $\lim_n 1_A L_n^+ = 1_A L_\infty^+ = 0$ and $0 \leq 1_A L_n^+ \leq Y_1^+$. Finally $\mathbb{E}[1_A Y_1] \leq 0$. Therefore, noting that $1_A \circ \theta$ is independent from Y_1,

$$0 \geq \mathbb{E}[1_A Y_1] = \mathbb{E}[1_A \circ \theta \; Y_1] = \mathbb{E}[1_A \circ \theta]\,\mathbb{E}[Y_1] = \underbrace{\mathbb{E}[1_A]}_{\geq 0}\underbrace{\mathbb{E}[Y_1]}_{>0}$$

This implies $\mathbb{P}(A) = 0$ and the lemma is proved.◄

(Proof of the Theorem.) We now turn to the proof of the LLN. Without loss of generality, we assume that $\mathbb{E}[X_1] = 0$. Applying the lemma with $Y_i = X_i + \varepsilon$ (where $\varepsilon > 0$), we get $\liminf n^{-1} \sum_{i=1}^n X_i \geq -\varepsilon \quad a.s.$ And applying again the lemma with $Y_i = -X_i + \varepsilon$, we get a.s., $\limsup n^{-1} \sum_{i=1}^n X_i \leq \varepsilon$ which finishes the proof since ε is arbitrary. ■

Since an estimator of $\mathbb{E}[X]$ is a random variable, it is important to provide confidence intervals. Indeed, two particular samplings may lead to very different estimates. The following theorem gives a result on the convergence speed of the estimator and then allows to provide confidence intervals. In practice, we should always balance the obtained estimator with the width of the confidence interval in order to justify the relevance of the announced result.

Theorem 3.4. (►CENTRAL LIMIT THEOREM (CLT)) *Let $(X_n)_n$ be a sequence of i.i.d random variables with the same law as X. If $\mathbb{E}[X^2] < \infty$, then*

$$\frac{\bar{X}_N - \mathbb{E}[X]}{\sqrt{\frac{\sigma^2}{N}}} \overset{w}{\Rightarrow} \mathcal{N}(0,1)$$

where $\sigma^2 = \mathbb{V}ar(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X - \mathbb{E}[X]]^2$.

PROOF. Replacing if necessary X_i by $(X_i - \mathbb{E}[X])/\sigma$, we can assume that $\mathbb{E}[X] = 0$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] = 1$. In such a case, we only need to prove that

$$Z_N = \sum_{i=1}^N \frac{X_i}{\sqrt{N}} \overset{w}{\Rightarrow} \mathcal{N}(0,1)$$

To this aim, (as for all weak convergence results), it is sufficient to show that the characteristic function of Z_N, $u \mapsto \mathbb{E}\left(\mathrm{e}^{\mathrm{i}uZ_N}\right)$ tends to the one of $\mathcal{N}(0,1)$ that is $u \mapsto \mathrm{e}^{-u^2/2}$. Write

$$\mathbb{E}\left(\mathrm{e}^{\mathrm{i}uZ_N}\right) = \mathbb{E}\left(\mathrm{e}^{\mathrm{i}u\sum_{i=1}^N X_i/\sqrt{N}}\right) = \left[\mathbb{E}\left(\mathrm{e}^{\mathrm{i}uX/\sqrt{N}}\right)\right]^N$$

$$\overset{\text{Taylor Exp.}}{\approx} \left[\mathbb{E}\left(1 + \frac{\mathrm{i}u}{\sqrt{N}}X - \frac{u^2}{2N}X^2 + \text{remainder terms}\right)\right]^N \approx \left[1 + \frac{\mathrm{i}u}{\sqrt{N}}\underbrace{\mathbb{E}[X]}_{0} - \frac{u^2}{2N}\underbrace{\mathbb{E}[X^2]}_{1}\right]^N = \left(1 - \frac{u^2}{2N}\right)^N \to \mathrm{e}^{-u^2/2}$$

This completes the proof. Of course, the difficult part (the one we forgot to mention) consists in showing that the remainder terms do not come into play when N tends toward infinity (in the notation $\approx$). But the take-home message from this proof is that the CLT is obtained from the convergence of the characteristic function with a Taylor expansion of order 2. ■

Nevertheless, in many practical situations $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$ is an unknown quantity. We are then tempted to replace σ^2 by an estimator. A natural question is: does the previous theorem remain valid? The answer is yes, as shown by the following proposition.

Proposition 3.5 *Let $(X_n)_n$ be a series of r.v. i.i.d of the same law as X, such that $\mathbb{E}[X^2] < \infty$ and $\sigma^2 > 0$. Noting*

$$\tilde{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^{N}\left[X_i - \mathbb{E}[X]\right]^2, \qquad \hat{\sigma}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[X_i - \bar{X}_N\right]^2$$

then

$$\frac{\bar{X}_N - \mathbb{E}[X]}{\sqrt{\frac{\tilde{\sigma}_N^2}{N}}} \overset{w}{\Rightarrow} \mathcal{N}(0.1), \qquad \frac{\bar{X}_N - \mathbb{E}[X]}{\sqrt{\frac{\hat{\sigma}_N^2}{N}}} \overset{w}{\Rightarrow} \mathcal{N}(0.1) \tag{3.1}$$

Lemma 3.6 *If $\mathbb{E}[X^2]$, the estimators $\tilde{\sigma}_N^2$ and $\hat{\sigma}_N^2$ are* **unbiased** *and* **strongly convergent** *estimators of σ^2, i.e.*

$$\mathbb{E}[\tilde{\sigma}_N^2] = \mathbb{E}[\hat{\sigma}_N^2] = \sigma^2,$$
$$\tilde{\sigma}_N^2 \overset{\mathbb{P}\text{-}a.s.}{\longrightarrow} \sigma^2, \quad \hat{\sigma}_N^2 \overset{\mathbb{P}\text{-}a.s.}{\longrightarrow} \sigma^2$$

PROOF. Clearly, $\mathbb{E}[\tilde{\sigma}_N^2] = \mathbb{E}[X_1 - \mathbb{E}[X_1]]^2 = \sigma^2$. By the LGN, $\tilde{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^{N}\{X_i - \mathbb{E}[X]\}^2 \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \mathbb{E}[X - \mathbb{E}[X]]^2 = \sigma^2$. Moreover, introducing $\mathbb{E}[X]$ in the square term $[X_i - \bar{X}_N]^2$, we get after straightforward algebra:

$$\hat{\sigma}_N^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N}(X_i - \mathbb{E}[X])^2 - 2\underbrace{\sum_{i=1}^{N}(X_i - \mathbb{E}[X])}_{N(\bar{X}_N - \mathbb{E}[X])}(\bar{X}_N - \mathbb{E}[X]) + N(\bar{X}_N - \mathbb{E}[X])^2\right) = \frac{\sum_{i=1}^{N}(X_i - \mathbb{E}[X])^2}{N-1} - \frac{N}{N-1}(\bar{X}_N - \mathbb{E}[X])^2 \tag{3.2}$$

The LGN shows that $\bar{X}_N \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \mathbb{E}[X]$ and we deduce by applying again the LGN on the expression of $\hat{\sigma}_N^2$ obtained in (3.2), that $\hat{\sigma}_N^2 \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \sigma^2$. Finally, by (3.2),

$$\mathbb{E}(\hat{\sigma}_N^2) = \frac{1}{N-1}\left(N\sigma^2 - N\mathbb{V}\mathrm{ar}\bar{X}_N\right) = \frac{1}{N-1}\left(N\sigma^2 - N\frac{\sigma^2}{N}\right) = \sigma^2$$

■

Remark 3.7 *If $\mathbb{E}[X]$ is known, it is better to use $\tilde{\sigma}_N^2$. Otherwise, we use $\hat{\sigma}_N^2$.*

The proof of the proposition 3.5 is a simple consequence of Slutski's Lemma (please, check it!), which we will now state without proof. Slutski's Lemma in many practical situations allows to obtain weak convergences of quantities of interest from other weak convergences provided that the quantities of interest only differ by variables that converges in probability to constants.

Theorem 3.8. (►SLUTSKI'S LEMMA) *If*

i) $X_n \overset{w}{\Rightarrow} X$,
ii) $Y_n \overset{\mathbb{P}\text{-}prob}{\longrightarrow} a$ where a is a constant

then for any continuous (and not necessarily bounded) function f

$$f(X_n, Y_n) \overset{w}{\Rightarrow} f(X, a)$$

Another result is extremely useful for obtaining other weak convergence results associated with $g(X_n)$ from a weak convergence result associated to X_n.

Theorem 3.9. ($\blacktriangleright\Delta$-METHOD) *If*

i) $\sqrt{n}(X_n - a) \overset{w}{\Rightarrow} Z$,
ii) $x \mapsto g(x)$ *is differentiable at the point* $x = a$

then

$$\sqrt{n}\,(g(X_n) - g(a)) \overset{w}{\Rightarrow} g'(a)Z$$

PROOF. First we notice that i) implies that $X_n \overset{\mathbb{P}\text{-prob}}{\longrightarrow} a$. This yields $\frac{g(X_n)-g(a)}{X_n-a} \overset{\mathbb{P}\text{-prob}}{\longrightarrow} g'(a)$. We then apply Slutski's lemma, noting that

$$\sqrt{n}\,(g(X_n) - g(a)) = \underbrace{\sqrt{n}(X_n - a)}_{\overset{w}{\Rightarrow} Z} \underbrace{\left(\frac{g(X_n) - g(a)}{X_n - a}\right)}_{\overset{\mathbb{P}\text{-prob}}{\longrightarrow} g'(a)}$$

■

3.2 On the use of the CLT for Monte Carlo methods

Let X be a random variable and f be a measurable function such that $\mathbb{E}(f^2(X)) < \infty$. We wish to approximate $\mathbb{E}(f(X))$ by a Monte Carlo method. We can draw a N-sample $(X_1, ..., X_N)$ according to the law of X, and then set

$$S_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \quad V_N = \frac{1}{N-1}\sum_{i=1}^{N} (f(X_i) - S_N)^2$$

Due to the LLN, S_n converges a.s. to $\mathbb{E}(f(X))$ and by Proposition 3.5,

$$\sqrt{N}\frac{S_N - \mathbb{E}(f(X))}{\sqrt{V_N}} \overset{w}{\Rightarrow} \mathcal{N}(0,1)$$

This last result then gives a confidence interval on the estimator S_n. Indeed, weak convergence implies that

$$\mathbb{P}\left(\sqrt{N}\frac{S_N - \mathbb{E}(f(X))}{\sqrt{V_N}} \in [-a, a]\right) \to \mathbb{P}(|G| \le a)$$

where $G \sim \mathcal{N}(0,1)$. So, $\left[S_N - a\sqrt{\frac{V_N}{N}}; S_N + a\sqrt{\frac{V_N}{N}}\right]$ is an confidence interval for $\mathbb{E}(f(X))$ of asymptotic level α if a is the quantile of order $1 - \alpha/2$ of the law $\mathcal{N}(0,1)$, i.e. $\mathbb{P}(|G| \le a) = 1 - \alpha/2$ with $G \sim \mathcal{N}(0,1)$. In many examples, we take $\alpha = 0.05$ and in this case $a \approx 1.96$.

3.3 Take-home message

a) The assumptions of the LLN and CLT should be perfectly known.
b) The student should know exactly which estimators are biased or unbiased. The proof should be known.
c) Use of the Slutski lemma and of the delta method to obtain other weak convergence from the CLT.
d) Diverse methods to obtain a confidence interval.

3.4 Highlights

Monte Carlo methods. Source: Wikipedia

The term "Monte Carlo method" was coined in the 1940s by physicists working on nuclear weapon projects in the Los Alamos National Laboratory.

Enrico Fermi in the 1930s and Stanislaw Ulam in 1946 first had the idea. Ulam later contacted John Von Neumann to work on it.

Physicists at Los Alamos Scientific Laboratory were investigating radiation shielding and the distance that neutrons would likely travel through various materials. Despite having most of the necessary data, such as the average distance a neutron would travel in a substance before it collided with an atomic nucleus or how much energy the neutron was likely to give off following a collision, the problem could not be solved with analytical calculations. John von Neumann and Stanislaw Ulam suggested that the problem be solved by modeling the experiment on a computer using chance. Being secret, their work required a code name. Von Neumann chose the name "Monte Carlo". The name is a reference to the Monte Carlo Casino in Monaco where Ulam's uncle would borrow money to gamble.

Random methods of computation and experimentation (generally considered forms of stochastic simulation) can be arguably traced back to the earliest pioneers of probability theory (see, e.g., Buffon's needle, and the work on small samples by William Sealy Gosset), but are more specifically traced to the pre-electronic computing era. The general difference usually described about a Monte Carlo form of simulation is that it systematically "inverts" the typical mode of simulation, treating deterministic problems by first finding a probabilistic analog (see Simulated annealing). Previous methods of simulation and statistical sampling generally did the opposite: using simulation to test a previously understood deterministic problem. Though examples of an "inverted" approach do exist historically, they were not considered a general method until the popularity of the Monte Carlo method spread.

Monte Carlo methods were central to the simulations required for the Manhattan Project, though were severely limited by the computational tools at the time. Therefore, it was only after electronic computers were first built (from 1945 on) that Monte Carlo methods began to be studied in depth. In the 1950s they were used at Los Alamos for early work relating to the development of the hydrogen bomb, and became popularized in the fields of physics, physical chemistry, and operations research. The Rand Corporation and the U.S. Air Force were two of the major organizations responsible for funding and disseminating information on Monte Carlo methods during this time, and they began to find a wide application in many different fields.

Uses of Monte Carlo methods require large amounts of random numbers, and it was their use that spurred the development of pseudorandom number generators, which were far quicker to use than the tables of random numbers which had been previously used for statistical sampling.

Chapter 4

Exact or approximate sampling

Contents

Keywords: *Cumulative distribution function, generalized inverse, the rejection sampling, sampling by mapping, sampling from a conditional law, importance sampling.*

4.1 Exact Sampling

4.1.1 The inverse cumulative distribution function

Let Y be a real random variable with cumulative distribution function $F_Y : t \mapsto F_Y(t) = \mathbb{P}(Y \leq t)$. Immediate properties of the function F_Y are as follows: F_Y is non-decreasing, right-continuous, has a left-limit, tends toward 0 in $-\infty$ and toward 1 in ∞. If F_Y is strictly increasing, it is easy to define the inverse function of F_Y. Unfortunately, it may be constant on some intervals (for example when Y is a discrete r.v.). As a result, we need to define a generalized inverse, which is also valid for non-decreasing functions (which can be constant over certain intervals):

Definition 4.1. (►GENERALIZED INVERSE) We define *the generalized inverse* of F_Y by: for any $x \in [0,1]$,

$$F_Y^{-1}(x) = \inf\{y \in \mathbb{R} : F_Y(y) \geq x\}$$

Of course, if F_Y is invertible then the generalized inverse is the inverse in the classical sense. The generalized inverse F_Y^{-1} of the cumulative fucntion F_Y is also called the **quantile function**.

Proposition 4.2 *If $U \sim \mathscr{U}[0.1]$, then $F_Y^{-1}(U) \overset{\mathscr{L}}{\equiv} Y$.*

PROOF. Using that the cumulative distribution functions are right-continuous, the following (intuitive) equivalence can be proved:

$$\forall u, v \in \mathbb{R}, \quad \{F_Y^{-1}(u) \leq v\} \Longleftrightarrow \{u \leq F_Y(v)\}.$$

So, if U is a r.v. with uniform distribution on $[0,1]$, we have the equality

$$P(F_Y^{-1}(U) \leq y) = P(U \leq F_Y(y)) = F_Y(y),$$

which completes the proof. ■

As a byproduct of this Proposition, we can propose a sampling method by using the inverse cumulative distribution function (in the case where this function is explicitly available). We formalize it in the following Corollary.

Corollary 4.3 *Let f be a probability density. Let $F(t) = \int_{-\infty}^{t} f(u)\mathrm{d}u$ with generalized inverse F^{-1}. Suppose F^{-1} is explicitly available. Then we can draw a r.v. Y with density f by the following Algorithm 1:*

Algorithm 1 Sampling by the inverse cumulative distribution function

1: Draw $U \sim \mathscr{U}[0.1]$
2: Set $Y = F^{-1}(U)$

Example 4.4 (►SIMULATION OF A DISCRETE LAW) *Let Y be a random variable with support on $(y_k)_{k\in\mathbb{N}}$, such that $P(Y = y_k) = p_k$. If $U \sim \mathscr{U}[0.1]$, then*

$$X = y_0 1_{U \leq p_0} + \sum_{k \geq 1} y_k 1_{\sum_{i=0}^{k-1} p_i < U \leq \sum_{i=0}^{k} p_i} \overset{\mathscr{L}}{\equiv} Y$$

Indeed, we can easily check that for $k > 0$, we have

$$\mathbb{P}(X = y_k) = \mathbb{P}\left(\sum_{i=0}^{k-1} p_i < U \leq \sum_{i=0}^{k} p_i\right) = p_k$$

This intuitive sampling method is actually a sampling by inversion of the cumulative distribution function (please, check it).

A typical example is to draw Y according to the Bernoulli law with parameter p by drawing $U \sim \mathscr{U}[0.1]$ and setting $Y = 1_{U \leq p}$. Another important example is to sample a binomial law $\mathscr{B}(n,p)$ by setting $Y = \sum_{i=1}^{n} 1_{U_i \leq p}$ where U_i is $\mathscr{U}[0.1]$. Apart from the examples of discrete law, the generalized inverse is reduced in most other situations to the classical inverse. And if available in a closed-form, it possible to use this sampling method. Let us look at an example of a "continuous" law.

Example 4.5 *The exponential distribution of parameter $\lambda > 0$ has a density of $f(x) = \lambda \exp(-\lambda x) 1_{\mathbb{R}^+}(x)$. The associated cumulative distribution function is $F(t) = -\exp(-\lambda t) + 1$ for $t \geq 0$ with inverse $F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$. So, if $U \sim \mathscr{U}]0.1[$, $-\frac{1}{\lambda}\ln(1-U) \sim \exp(\lambda)$. Now $1-U$ also follows $\mathscr{U}]0.1[$. Since it is better to use as few calculations as possible, we set $F^{-1}(1-U) = -\frac{1}{\lambda}\ln(U) \sim \exp(\lambda)$.*

4.1.2 The rejection sampling

The inverse cumulative distribution function does not always admit an explicit form. We now present a method widely used in practice which consists in (1) proposing candidates according to another law and then (2) in defining a stopping rule, namely, a criterion which if met, allows us to select a variable which ("surprisingly") will be distributed according to the target distribution.

Proposition 4.6 *Let f and g be two probability densities satisfying for any $x \in \mathbb{R}$, $\boxed{f(x) \leq Mg(x)}$. We consider (U_i) and (X_i) two sequences of random variables such that*

i) $U_i \sim \mathscr{U}[0.1]$ and $X_i \sim g$
ii) the sequence of random variables $(U_s)_{s\geq 1}$ and $(X_t)_{t\geq 1}$ are independent.

Let $\nu = \inf\left\{i \in \mathbb{N},\ U_i \leq \frac{f(X_i)}{Mg(X_i)}\right\}$ and $Y = X_\nu$. Then,

a) $\nu \sim \mathbb{G}eom(M^{-1})$ and $Y \sim f$.
b) ν and Y are independent variables.

We encourage the reader to read the following proof, rich in technical lessons.

PROOF. Note that the random vector $\begin{pmatrix} U \\ X \end{pmatrix}$ has the density $(u,x) \mapsto 1_{[0,1]}(u)g(x)$. This allows to write, knowing that the vectors $\begin{pmatrix} U_k \\ X_k \end{pmatrix}$ are i.i.d.,

$$\begin{aligned}
\mathbb{P}(Y \in A,\ \nu = k) &= \mathbb{P}(X_k \in A,\ \nu = k) \\
&= \mathbb{P}\left(U_1 > \frac{f(X_1)}{Mg(X_1)}, \ldots, U_{k-1} > \frac{f(X_{k-1})}{Mg(X_{k-1})},\ U_k \leq \frac{f(X_k)}{Mg(X_k)},\ X_k \in A\right) \\
&= \mathbb{P}\left(U_1 > \frac{f(X_1)}{Mg(X_1)}\right)^{k-1} \mathbb{P}\left(U_k \leq \frac{f(X_k)}{Mg(X_k)},\ X_k \in A\right) \\
&= \left(\iint_{\{u > \frac{f(x)}{Mg(x)}\}} 1_{[0,1]}(u)g(x)\ \mathrm{d}u\mathrm{d}x\right)^{k-1} \iint_{\{u \leq \frac{f(x)}{Mg(x)}\}\cap\{x\in A\}} 1_{[0,1]}(u)g(x)\ \mathrm{d}u\mathrm{d}x \\
&= \left(\int \left(1 - \frac{f(x)}{Mg(x)}\right) g(x)\ \mathrm{d}x\right)^{k-1} \int_{\{x\in A\}} \frac{f(x)}{Mg(x)} g(x)\ \mathrm{d}x \\
&= \left(1 - \frac{1}{M}\right)^{k-1} \frac{1}{M} \int_{\{x\in A\}} f(x)\mathrm{d}x
\end{aligned} \tag{4.1}$$

where the penultimate line comes from integration over u and the last line comes from the fact that f and g are probability densities. We then have by marginalization

$$\mathbb{P}(\nu = k) = \mathbb{P}(Y \in \mathbb{R},\ \nu = k) = \left(1 - \frac{1}{M}\right)^{k-1} \frac{1}{M}$$

so $\nu \sim \mathbb{G}\mathrm{eom}(M^{-1})$, and therefore $\mathbb{P}(\nu < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(\nu = k) = 1$. This implies:

$$\mathbb{P}(Y \in A) = \mathbb{P}(Y \in A,\ \nu < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(Y \in A,\ \nu = k) = \sum_{k=1}^{\infty} \left(1 - \frac{1}{M}\right)^{k-1} \frac{1}{M} \int_{\{x\in A\}} f(x)\mathrm{d}x = \int_{\{x\in A\}} f(x)\mathrm{d}x$$

Therefore $Y \sim f$. Finally, (4.1) shows that ν and Y are independent random variables. ■

The immediate corollary of this proposition will be the rejection sampling method:

Corollary 4.7 (►REJECTION ALGORITHM) *Let f and g be two densities such that*

i) we can sample according to the density g,
ii) there is M such that for every $x \in \mathbb{R}$, $f(x) \leq Mg(x)$
iii) for any x such that $g(x) > 0$, the quantity $\frac{f(x)}{Mg(x)}$ is explicitly available

Then, we can sample a random variable Y according to the density f by the algorithm 2.

Algorithm 2 The rejection sampling

1: draw $X \sim g$, $U \sim \mathscr{U}[0.1]$ independently
2: **while** $U \geq \frac{f(X)}{Mg(X)}$ **do** draw $X \sim g$, $U \sim \mathscr{U}[0.1]$ independently
3: **end while**
4: Set $Y = X$

A fundamental difference with sampling by the inverse cumulative distribution function is the number of samplings required before producing a single r.v. distributed according to the target distribution: this waiting time before acceptance is here a random variable ν of geometric distribution $\mathbb{G}\text{eom}(M^{-1})$. The expectation of a geometric distribution is the inverse of its parameter, which gives $\mathbb{E}(\nu) = M$. The function f being given, we are therefore interested in a density g such that $f \leq Mg$ with M as small as possible (to minimize the waiting time before acceptance). Of course, we always have $M \geq 1$ (to see this, just integrate $f(x) \leq Mg(x)$ and note that f and g are probability densities) but we can't reach the limit $M = 1$. Otherwise, we would have $f(x) \leq g(x)$ and by integration $\int \underbrace{g(x) - f(x)}_{\geq 0} \mathrm{d}x = 1 - 1 = 0$, hence $f = g$ and if we can draw according to g, that means we can draw according to f!

4.1.3 Sampling from a conditional distribution

4.1.3.1 Link with the rejection sampling

Let X be a r.v. with density g, which can be sampled. If you want to sample a r.v. according to the distribution of X conditionally on the event $\{X \in [a,b]\}$, an intuitive idea would be to sample following g and to keep only the candidates that fall into $[a,b]$. We will show that this method actually corresponds to a particular case of the rejection sampling. The distribution of X conditionally on the event $\{X \in [a,b]\}$ has the density

$$f(x) = 1_{[a,b]}(x) \frac{g(x)}{\int_a^b g(u)\mathrm{d}u}.$$

Indeed, write

$$\mathbb{P}(X \in A | X \in [a,b]) = \frac{\mathbb{P}(X \in A \cap [a,b])}{\mathbb{P}(X \in [a,b])} = \frac{\int_A 1_{[a,b]}(x) g(x) \mathrm{d}x}{\int_a^b g(u)\mathrm{d}u} = \int_A f(x)\mathrm{d}x.$$

Now, note that

$$f(x) \leq Mg(x) \quad \text{avec} \quad M = \frac{1}{\int_a^b g(u)\mathrm{d}u}$$

Therefore, applying the rejection sampling, draw $U \sim \mathscr{U}[0,1]$ and $X \sim g$ independently and set $Y = X$ only if

$$U \leq \frac{f(X)}{Mg(X)} = 1_{[a,b]}(X) = \begin{cases} 1 & \text{if } X \in [a,b] \\ 0 & \text{otherwise} \end{cases}.$$

Finally if $X \in [a,b]$, the candidate is accepted with probability 1, and if $X \notin [a,b]$, it is refused with probability 1. This is exactly the intuitive idea of the algorithm. In this example, $M = 1/\int_a^b g(u)\mathrm{d}u$ is not available in closed-form in general; still, as seen before, the ratio $f(X)/(Mg(X))$ can be calculated and this allows to apply the rejection sampling. Indeed if $\mathbb{P}(X \in [a,b])$ is small then, we will wait long before accepting a candidate. To apply the rejection sampling with a target density with support inside $[a,b]$, it is ofter better to propose according to a density with support inside $[a,b]$.

4.1.3.2 Link with sampling by inversion of the cumulative distribution function

If $\boxed{F_X \text{ et } F_X^{-1} \text{ are available in closed-form}}$, it is more efficient to sample a random variable according to the conditional law of X with respect to the event $\{X \in [a,b]\}$ in the following way

i) draw $U \sim \mathscr{U}[0,1]$
ii) set $Y = F_X^{-1}\left(F_X(a) + U[F_X(b) - F_X(a)]\right)$.

Algorithm 3 Sample from a conditional distribution

1: draw $U \sim \mathscr{U}[0,1]$
2: Set $Y = F_X^{-1}\left(F_X(a) + U[F_X(b) - F_X(a)]\right)$.

Indeed, write for all $t \in [a,b]$,

$$\begin{aligned}\mathbb{P}(Y \leq t) &= \mathbb{P}(F_X^{-1}(F_X(a) + U[F_X(b) - F_X(a)]) \leq t) = \mathbb{P}\left(F_X(a) + U[F_X(b) - F_X(a)] \leq F_X(t)\right) \\ &= \mathbb{P}\left(U \leq \frac{F_X(t) - F_X(a)}{F_X(b) - F_X(a)}\right) = \frac{F_X(t) - F_X(a)}{F_X(b) - F_X(a)} = \frac{\int_a^t g(x)\mathrm{d}x}{\int_a^b g(u)\mathrm{d}u} = \int_{]-\infty,t[} f(x)\mathrm{d}x\end{aligned}$$

An alternative way to see this is the following: the target density is

$$g(x) = \frac{f(x)\mathbb{1}_{x\in[a,b]}}{\int_a^b f(t)dt} = \frac{f(x)\mathbb{1}_{x\in[a,b]}}{F_X(b) - F_X(a)}$$

This target density has the cumulative distribution function G defined by for $x \in [a,b]$,

$$G(x) = \frac{\int_a^x f(t)dt}{F_X(b) - F_X(a)} = \frac{F(x) - F(a)}{F_X(b) - F_X(a)}$$

Since $\frac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)} = u$ is equivalent to $F_X(x) = F_X(a) + u\left[F_X(b) - F_X(a)\right]$, we deduce that G has a generalized inverse which is:

$$G^{-1}(u) = F_X^{-1}\left(F_X(a) + u\left[F_X(b) - F_X(a)\right]\right).$$

This completes the proof.

4.1.4 Other sampling methods

4.1.4.1 Sampling by mapping

We also can try to show that the target distribution is the one of random variables, which are obtained by a mapping (typically C^1-mappings but not always) of r.v. which can be sampled easily (typically r.v. with uniform distribution).

Lemma 4.8 CHANGE OF VARIABLES- *Let ϕ be C^1-mapping from an open set O of $\mathbb{R}^d$ to an open set O' of $\mathbb{R}^d$. Assume that $V \sim g$ where g has a support in O, then $U = \phi(V)$ has the density $u \mapsto f(u) = g \circ \phi^{-1}(u) |\det(\nabla \phi^{-1}(u))|$*

It is useless to learn this lemma by heart since we can recover it quite intuitively by the change of variable formula for multiple integrals:

$$\mathbb{E}(h(U)) = \mathbb{E}(h \circ \phi(V)) = \int h \circ \underbrace{\phi(v)}_{u} g(v) \mathrm{d}v = \int h(u) \times g(\phi^{-1}(u)) \underbrace{\left|\frac{\partial v}{\partial u}\right|}_{|\det \nabla(\phi^{-1}(u))|} \mathrm{d}u = \int h(u) f(u) \mathrm{d}u$$

Thanks to this lemma, we can state the following proposition which proposes a sampling method for obtaining a couple of independent gaussian variables.

Proposition 4.9 BOX MULLER- *Let U and V be two i.i.d.r.v. with distribution $\mathscr{U}[0,1]$. Set*

$$X = \sqrt{-2\ln U}\cos(2\pi V), \quad Y = \sqrt{-2\ln U}\sin(2\pi V)$$

Then X and Y are i.i.d. with distribution $\mathscr{N}(0,1)$.

PROOF. We can easily show by the change of variable formula that if X and Y are i.i.d. of distribution $\mathscr{N}(0,1)$ then, denoting by (R,θ) the polar coordonates of (X,Y), the couple of r.v. (R,θ) is independent, $R^2 \sim \exp(1/2)$ and $\theta \sim \mathscr{U}[0,2\pi]$. ■

To draw gaussian r.v., we can also use the inversion of the cumulative distribution function of $\mathscr{N}(0,1)$. Set $\mathscr{N}(t) = \int_{-\infty}^{t} (\mathrm{e}^{-u^2/2}/\sqrt{2\pi})\mathrm{d}u$ which is not explicit (nor its inverse) but it may happen that the numerical values of $\mathscr{N}^{-1}(t)$ can be approximated at any order in some libraries and in that case, we can draw a gaussian distribution by using the inverse cumulative distribution function.

4.1.4.2 Sampling a gaussian vector

The Box and Muller algorithm allows to sample easily a vector of distribution $\boxed{\mathscr{N}(0,I)}$ where I is the identity matrix. If we then intend to sample a general gaussian vector of distribution $\mathscr{N}(\mu,\Sigma)$ then note that Σ being symmetric, and nonnegative, there exists a real valued triangular matrix A such that $AA^T = \Sigma$ (see the **Cholesky** decomposition), then the random vector $\boxed{\mu + AG}$ with $G \sim \mathscr{N}(0,I)$ is distributed according to $\boxed{\mathscr{N}(\mu, \underbrace{\Sigma}_{AA^T})}$.

4.1.4.3 Sampling by marginalisation

Other methods for sampling from a given distributions exist: for example, it happens that the target density f can be written as $f(u) = \int g(u,v)\mathrm{d}v$ where g is itself a density. In that case, if we can sample according to

g then we sample $(U,V) \sim g$ and U will have the density f. Often, $g(u,v) = h(v)p(u|v)$ and h and p can be sampled. We draw $V \sim h$. Then, setting $V = v$, we draw $U \sim p(\cdot|v)$. When V is a discrete-valued random variable , we then say that f is a *mixing distribution*.

4.2 Approximate sampling

4.2.1 Importance Sampling

In most situations, we do not really intend to sample according to a distribution but we want an appproximation of the expectation of a functional of a r.v. with a given law, i.e. $\mathbb{E}_f(h(Y)) = \int h(y)f(y)\mathrm{d}y$ (for the notation $\mathbb{E}_f$, see the footnote [1]). In that case, Importance Sampling (IS) boils down to sample $(X_1,\ldots,X_N)$ according to a proposal density $g > 0$, $X_i \sim g$, and then, to use the approximation

$$\mathbb{E}_f(h(Y)) = \int h(y)f(y)\mathrm{d}y \approx \bar{S}_N = \frac{\sum_{i=1}^N \frac{f(X_i)}{g(X_i)}h(X_i)}{N}$$

Conditions for applyiing importance sample are then

i) we have $\int_{g(x)=0} f(x)\mathrm{d}x = 0$.
ii) we can draw from g.
iii) for all x, $f(x)/g(x)$ has a closed-form expression.

The first assumption implies

$$\mathbb{E}_g(f(X)h(X)/g(X)) = \int \frac{f(x)}{g(x)}h(x)g(x)\mathrm{d}x = \int_{g(x)\neq 0} h(x)f(x)\mathrm{d}x = \mathbb{E}_f(h(Y)).$$

The importance sampling estimator $\bar{S}_N$ is clearly unbiased $\mathbb{E}(\bar{S}_N) = \mathbb{E}_f(h(Y))$, strongly convergent under the assumption $\mathbb{E}_g\left[\frac{f(X)}{g(X)}|h(X)|\right] = \mathbb{E}_f[|h(X)|] < \infty$. The LLN indeed show that

$$\frac{\sum_{i=1}^N \frac{f(X_i)}{g(X_i)}h(X_i)}{N} \xrightarrow{\mathbb{P}\text{-a.s.}} \int \frac{f(x)}{g(x)}h(x)g(x)\mathrm{d}x = \int h(y)f(y)\mathrm{d}y$$

We can see Importance Sampling as a method where we draw X_i according to a proposal distribution g and the error is then corrected, (because we have not drawn according to f) by associating to each X_i a weight $f(X_i)/g(X_i)$.

We can have the impression that IS can be applied more often than the rejection sampling since in IS, we do not need to assume that $\sup_x \frac{f(x)}{g(x)} < \infty$ (this condition can be a bit restrictive sometimes)... It is true but the advantage of the rejection sampling is that it produces an exact sample according to f!

4.2.1.1 Optimisation of the proposal density

Under the assumption that $\mathbb{E}_g\left[\frac{f^2(X)}{g^2(X)}h^2(X)\right] = \mathbb{E}_f\left[\frac{f(X)}{g(X)}h^2(X)\right] < \infty$, the CLT gives the quality of the approximation as follows

$$\sqrt{N}\left(\frac{\sum_{i=1}^N \frac{f(X_i)}{g(X_i)}h(X_i)}{N} - \int h(y)f(y)\mathrm{d}y\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}_g\left(\frac{f(x)}{g(x)}h(x)\right)\right)$$

[1] When we put an f under the expectation, this means that we take the expectation with respect to a random variable of density f

To target a density g, that gives rise to the most precise approximation, we should then minimize the quantity $\mathbb{V}\mathrm{ar}_g\left(\frac{f(x)}{g(x)}h(x)\right)$.

Proposition 4.10

$$\inf\left\{\mathbb{V}ar_g\left(\frac{f(x)}{g(x)}h(x)\right); g \text{ density}\right\} = \left(\int f(x)|h|(x)\mathrm{d}x\right)^2 - \left(\int f(x)h(x)\mathrm{d}x\right)^2$$

The infimum is attained with a density $g^\star$ defined by $g^\star(x) = f(x)|h(x)|/\int f(u)|h(u)|\mathrm{d}u$.

PROOF. Write

$$\mathbb{V}\mathrm{ar}_g\left(\frac{f(x)}{g(x)}h(x)\right) = \int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x - \left(\int h(y)f(y)\mathrm{d}y\right)^2$$

The second term does not depend on g, we can thus minimize the first term of the right-hand side. By the Cauchy-Schwarz inequality;

$$\left(\int f(x)|h(x)|\mathrm{d}x\right)^2 = \left(\int \frac{f(x)}{\sqrt{g(x)}}|h(x)|\sqrt{g(x)}\mathrm{d}x\right)^2 \leq \left(\int \frac{f^2(x)}{g(x)}|h(x)|^2\,\mathrm{d}x\right)\left(\underbrace{\int (\sqrt{g(x)})^2\mathrm{d}x}_{1}\right) = \int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x \tag{4.2}$$

Moreover, the equality in the Cauchy Schwarz inequality holds for $\sqrt{g^\star(x)} \propto \frac{f(x)}{\sqrt{g^\star(x)}}|h(x)|$, which can also be written as, knowing that $g^\star$ is a density: $g^\star(x) = f(x)|h(x)|/\int f(u)|h(u)|\mathrm{d}u$.

An alternative (and more direct) proof is as follows: Note that

$$\mathbb{V}\mathrm{ar}_g\left(\frac{f(x)}{g(x)}|h(x)|\right) = \int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x - \left(\int |h(y)|f(y)\mathrm{d}y\right)^2$$

And since the variance is non negative, we get

$$\int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x \geq \left(\int |h(y)|f(y)\mathrm{d}y\right)^2$$

This shows (4.2). This inequality becomes an equality for $g^\star$ such that $\mathbb{V}\mathrm{ar}_{g^\star}\left(\frac{f(x)}{g^\star(x)}|h(x)|\right) = 0$, that is, if $x \mapsto \frac{f(x)}{g^\star(x)}|h(x)|$ is g-a.s. a constant, that is if $g^\star \propto |h|f$ which corresponds to $g^\star(x) = f(x)|h(x)|/\int f(u)|h(u)|\mathrm{d}u$. Finally, with this definition of $g^\star$, we have

$$\int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x \geq \int \frac{f(x)}{g^\star(x)}h^2(x)f(x)\mathrm{d}x$$

■

Unfortunately, this lemma has no immediate practical consequences since it is not obvious to know how to sample according to the density distribution $g^\star(x) = f(x)|h(x)|/\int f(u)|h(u)|\mathrm{d}u$ and even if it were the case, to be able to use sampling importance, it would have been necessary to know how to calculate explicitly

$$\frac{f(x)}{g^\star(x)} = \frac{\int f(u)|h(u)||\mathrm{d}u}{|h(x)|}$$

which we don't usually know how to do, since we're trying to give a numerical value to $\int f(u)h(u)\mathrm{d}u$. However, the message of this lemma is that, when considering the optimal approximation of $\mathbb{E}_f(h(Y))$, the density that corresponds to the minimal variance is not necessarily f.

4.2.2 Other methods for approximate sampling

There are many other approximate sampling algorithms, in particular the MCMC algorithms (Monte Carlo by Markov Chains) whose general principle consists in building a Markov chain whose stationary distribution is of density f. The approximation of $\mathbb{E}_f(h(X))$ by $N^{-1}\sum_{i=1}^{N} h(X_i)$ then comes from a law of large numbers for a Markov chain (we can no longer use the usual LLN because the r.v. are not i.i.d.).

Other hybrid methods judiciously combine Importance Sampling and MCMC algorithms. That being said, the question remains open for choosing the most efficient estimation method that approximates $\mathbb{E}_f(h(X))$; it is still the subject of intensive research, particularly when the variable X evolves in a high dimmensional space.

4.3 Take-home message

a) Sampling by the inverse cumulative distribution function: the student should know how to show it when the cumulative distribution function is invertible. He should know how to distinguish discrete and continuous variables.
b) Know how to re-prove the proposition that justifies the rejection sampling. The rejection algorithm requires a number of random samplings before producing a sample from the target distribution.
c) Know how to make change of variables to do the mapping sampling.
d) On the use of importance sampling: the student should know the justification for the convergence and the CLT. Optimization of asymptotic variance.

Chapter 5

Metropolis-Hastings algorithms

Keywords: *Markov chains, Markov property, Metropolis Hastings, canonical space.*

This chapter introduces very basic tools in the Markov Chain Monte Carlo (MCMC) theory. We only focus on the most fundamental properties that will be useful for a first understanding of Metropolis-Hastings (MH) algorithms. Let us start with a gentle and smooth introduction to Markov chains.

5.1 Main notation

Let $(\mathsf{X}, \mathscr{X})$ be a measurable space, i.e. $\mathscr{X}$ is a σ-algebra on X, and consider the following notations.

- $\mathsf{M}_+(\mathsf{X})$ is the set of (non-negative) measures on $(\mathsf{X}, \mathscr{X})$.
- $\mathsf{M}_1(\mathsf{X})$ is the set of probability measures on $(\mathsf{X}, \mathscr{X})$.
- $\mathsf{F}(\mathsf{X})$ is the set of real-valued measurable functions f on X and $\mathsf{F}_+(\mathsf{X})$ the set of non-negative measurable functions on X.
- If $k \leq \ell$, $u_{k:\ell}$ means $(u_k, \ldots, u_\ell)$ and $u_{k:\infty}$ means $(u_{k+\ell})_{\ell \in \mathbb{N}}$.

Other notation will be introduced progressively.

5.2 Definitions

We first describe a Markov kernel, which will then be fundamental for the definition of a Markov chain.

Definition 5.1. We say that $P : \mathsf{X} \times \mathscr{X} \to \mathbb{R}^+$ is a Markov kernel, if for all $(x, A) \in \mathsf{X} \times \mathscr{X}$,

- $\mathsf{X} \ni y \mapsto P(y, A)$ is $\mathscr{X}/\mathscr{B}(\mathbb{R}^+)$ measurable,
- $\mathscr{X} \ni B \mapsto P(x, B)$ is a probability measure on $(\mathsf{X}, \mathscr{X})$.

In words, for all $(x, A) \in \mathsf{X} \times \mathscr{X}$, as a function of the first component only, $P(\cdot, A)$ is measurable and as a function of the second component only, $P(x, \cdot)$ is a probability measure. In particular, $P(x, \mathsf{X}) = 1$ for all $x \in \mathsf{X}$. Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, \mathrm{d}y)$. For example,

$$P(x, A) = \int_{\mathsf{X}} \mathbb{1}_A(y) P(x, \mathrm{d}y) = \int_A P(x, \mathrm{d}y) \, .$$

In almost all the course, a Markov kernel P allows to move a point x from a measurable space $(\mathsf{X},\mathscr{X})$ to another point on the same measurable space, that is, P is defined on $\mathsf{X}\times\mathscr{X}$ but we can more generally define a Markov kernel from a measurable space $(\mathsf{X},\mathscr{X})$ to another measurable space $(\mathsf{Y},\mathscr{Y})$. In such case, P will be a Markov kernel on $\mathsf{X}\times\mathscr{Y}$. We can now move on to the definition of a Markov chain.

Definition 5.2. Let $\{X_k : k\in\mathbb{N}\}$ be a sequence of random variables on the same probability space $(\Omega,\mathscr{G},\mathbb{P})$ and taking values on X, we say that $\{X_k : k\in\mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution $\nu\in\mathsf{M}_1(\mathsf{X})$ if and only if

(i) for all $(k,A)\in\mathbb{N}\times\mathscr{X}$, $\mathbb{P}(X_{k+1}\in A|X_{0:k})=P(X_k,A)$, $\mathbb{P}$-a.s.
(ii) $\mathbb{P}(X_0\in A)=\nu(A)$.

Note that in the definition we consider $\mathbb{P}(X_{k+1}\in A|X_{0:k})$, that is, the conditional probability is with respect to the sigma-field $\sigma(X_{0:k})$. We can actually replace $\sigma(X_{0:k})$ by $\mathscr{F}_k$ as soon as we know that $(X_k)_{k\geq 0}$ is $(\mathscr{F}_k)_{k\geq 0}$-adapted.

What does it mean exactly? Well, recall that if $\{\mathscr{F}_k : k\in\mathbb{N}\}$ is a sequence of embedded sigma-fields on X (that is, $\mathscr{F}_k\subset\mathscr{F}_{k+1}$ for all k), then $\{\mathscr{F}_k : k\in\mathbb{N}\}$ is called a **filtration** on X and we say that $(X_k)_{k\geq 0}$ is $(\mathscr{F}_k)_{k\geq 0}$-adapted if X_k is $\mathscr{G}/\mathscr{F}_k$-measurable for all $k\in\mathbb{N}$. Of course, the most natural filtration for $\{X_k : k\in\mathbb{N}\}$ is indeed $\mathscr{F}_k=\sigma(X_{0:k})$ and unsurprisingly, we call it the **natural filtration**. But other possibilities exist, where $\mathscr{F}_k$ is enlarged to include some other variables alongside with $X_{0:k}$. For example, let $\{Y_k : k\in\mathbb{N}\}$ be any other sequence of random variables on $(\Omega,\mathscr{G},\mathbb{P})$ and taking values in X (we do not assume anything on the relation between $\{X_k : k\in\mathbb{N}\}$ and $\{Y_k : k\in\mathbb{N}\}$). A typical example corresponds to $X_{k+1}=g(X_{0:k},Y_{0:k})$, but we do not even need to assume that for the moment. Set $\mathscr{F}_k=\sigma(X_{0:k},Y_{0:k})$ and assume that

$$\mathbb{P}(X_{k+1}\in A|\mathscr{F}_k)=P(X_k,A),\quad \mathbb{P}-\text{a.s.} \tag{5.1}$$

Since X_ℓ is $\mathscr{F}_\ell$-measurable and $\mathscr{F}_\ell\subset\mathscr{F}_k$ for $\ell\leq k$, we deduce that $\sigma(X_{0:k})\subset\mathscr{F}_k$. This allows to apply the tower property, which yields

$$\begin{aligned}\mathbb{P}(X_{k+1}\in A|X_{0:k}) &= \mathbb{E}\left[\mathbb{P}(X_{k+1}\in A|\mathscr{F}_k)|X_{0:k}\right]\\ &= \mathbb{E}\left[P(X_k,A)|X_{0:k}\right]=P(X_k,A),\quad \mathbb{P}-\text{a.s.}\end{aligned}$$

and therefore if we assume (5.1), then as soon as $(X_k)_{k\geq 0}$ is $(\mathscr{F}_k)_{k\geq 0}$-adapted, we can conclude that $\{X_k : k\in\mathbb{N}\}$ is a Markov chain with Markov kernel P. Why is it useful? Well, sometimes you define iteratively X_{k+1} using other variables rather than $X_{0:k}$ only and therefore, considering $\mathbb{P}(X_{k+1}\in A|\mathscr{F}_k)$ is easier to deal with. Let us see it in action through a very simple example.

Example 5.3 *Let $\{\varepsilon_k : k\geq 1\}$ be i.i.d. random variables on $\mathbb{R}^p$ with density f with respect to the Lebesgue measure on $\mathbb{R}^p$, and let $X_0\sim\mu$. We assume that X_0 is independent of $\{\varepsilon_k : k\in\mathbb{N}\}$. Define*

$$X_{k+1}=aX_k+b\varepsilon_{k+1},\quad k\in\mathbb{N}.$$

Set $\mathscr{F}_0=\sigma(X_0)$ and for $k\geq 1$, $\mathscr{F}_k=\sigma(X_0,\varepsilon_{1:k})$. Since X_ℓ is a deterministic function of X_0 and $\varepsilon_{1:\ell}$, we deduce that $(X_k)_{k\geq 0}$ is $(\mathscr{F}_k)_{k\geq 0}$-adapted. Therefore, we only need to check (5.1). Now, for any non-negative or bounded measurable function h on X,

$$\mathbb{E}\left[h(X_{k+1})|\mathscr{F}_k\right]=\int_{-\infty}^{\infty}h(\underbrace{aX_k+b\varepsilon}_{y})f(\varepsilon)\mathrm{d}\varepsilon=\int_{-\infty}^{\infty}h(y)\underbrace{f\left(\frac{y-aX_k}{b}\right)\frac{1}{b^p}\mathrm{d}y}_{P(X_k,\mathrm{d}y)},$$

where the last equality follows from an adequate change of variable. Therefore, $\{X_k : k\in\mathbb{N}\}$ is a Markov chain with Markov kernel

$$(x,A)\mapsto P(x,A)=\int_A f\left(\frac{y-ax}{b}\right)\frac{1}{b^p}\mathrm{d}y.$$

In this example, we can check that $\mathscr{F}_k$ is actually the natural filtration of $\{X_k : k \in \mathbb{N}\}$ but we even do not need to check this property for getting that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain.

What have we learned so far? At this stage, we are able to solve typical exercises where some random variables are given and the question is to determine whether these random variables form a Markov chain or not and if yes, what is the expression of the associated Markov kernel.

5.2.1 Additional notation

We are now ready (and eager) to absorb frantically other notation... For all $\mu \in \mathsf{M}_+(\mathsf{X})$, all Markov kernels P, Q on $\mathsf{X} \times \mathscr{X}$, and all measurable non-negative or bounded functions on h on X, we use the following convention and notation.

- μP is the (positive) measure: $\mathscr{X} \ni A \mapsto \mu P(A) = \int \mu(\mathrm{d}x) P(x,A)$,
- PQ is the Markov kernel: $(x,A) \mapsto \int_{\mathsf{X}} P(x,\mathrm{d}y) Q(y,A)$,
- Ph is the measurable function $x \mapsto \int_{\mathsf{X}} P(x,\mathrm{d}y) h(y)$.

It is easy to check that if μ is a probability measure, then μP is also a probability measure (since $\mu P(\mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) P(x,\mathsf{X}) = \int_{\mathsf{X}} \mu(\mathrm{d}x) = 1$). With this notation, using Fubini's theorem,

$$\begin{aligned}\mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h \\ &= \mu((PQ)h) = \int \cdots \int_{\mathsf{X}^3} \mu(\mathrm{d}x) P(x,\mathrm{d}y) Q(y,\mathrm{d}z) h(z).\end{aligned}$$

Therefore, all these parenthesis can be discarded and we can write μPQh without any ambiguity. That is excellent news because it is simpler to deal with expressions without all these parenthesis. To sum up, measures act on the left side of a Markov kernel whereas functions acts on the right side. To make sure you have mastered all the notation, check your understanding with the following equalities $\delta_x P(A) = P(x,A) = P\mathbb{1}_A(x)$.

To finish up with notation, we now define the iterates of a Markov kernel P, which will come in very handy thereafter: for a given Markov kernel P on $\mathsf{X} \times \mathscr{X}$, define $P^0 = I$ where I is the identity kernel: $(x,A) \mapsto \mathbb{1}_A(x)$, and set for $k \geq 0$, $P^{k+1} = P^k P$.

Lemma 5.4 *Let $\{X_k : k \in \mathbb{N}\}$ be a Markov chain on the same probability space $(\Omega, \mathscr{G}, \mathbb{P})$ and taking values on X, with Markov kernel P and with initial distribution $\nu \in \mathsf{M}_1(\mathsf{X})$. Then, for any $n \in \mathbb{N}$, the law of $X_{0:n}$ is $\nu(\mathrm{d}x_0) \prod_{i=0}^{n-1} P(x_i, \mathrm{d}x_{i+1})$ (with the convention that $\prod_{i=0}^{-1} = 1$).*

PROOF. Recall that for all $(n,A) \in \mathbb{N} \times \mathscr{X}$, $\mathbb{P}(X_{n+1} \in A | X_{0:n}) = P(X_n, A)$, $\mathbb{P}$-a.s. or equivalently for all non-negative measurable functions h_{k+1} on X, $\mathbb{E}[h_{n+1}(X_{n+1}) | X_{0:n}] = Ph_{n+1}(X_n)$ $\mathbb{P}$-a.s. We now show by induction that for all $n \in \mathbb{N}$,

$$(H_n) \text{ the law of } X_{0:n} \text{ is } \nu(\mathrm{d}x_0) \prod_{i=0}^{n-1} P(x_i, \mathrm{d}x_{i+1}).$$

We first note that (H_0) is true since by assumption, $X_0 \sim \nu$. Assume now that (H_n) holds for some $n \in \mathbb{N}$. Then, for all non-negative measurable functions $h_0, \ldots, h_{n+1}$ on X, the tower property yields

$$\mathbb{E}\left[\prod_{i=0}^{n+1} h_i(X_i)\right] = \mathbb{E}\left[\left(\prod_{i=0}^{n} h_i(X_i)\right) \mathbb{E}[h_{n+1}(X_{n+1}) | X_{0:n}]\right] = \mathbb{E}\left[\left(\prod_{i=0}^{n} h_i(X_i)\right) Ph_{n+1}(X_n)\right],$$

and since the inner term in the rhs only depends on $X_{0:n}$, we can apply (H_n) and thus,

$$\begin{aligned}\mathbb{E}\left[\prod_{i=0}^{n+1} h_i(X_i)\right] &= \int \cdots \int_{\mathsf{X}^{n+1}} \left[\nu(\mathrm{d}x_0) \prod_{i=0}^{n-1} P(x_i, \mathrm{d}x_{i+1})\right] \left(\prod_{i=0}^{n} h_i(x_i)\right) Ph_{n+1}(x_n) \\ &= \int \cdots \int_{\mathsf{X}^{n+1}} \left[\nu(\mathrm{d}x_0) \prod_{i=0}^{n} P(x_i, \mathrm{d}x_{i+1})\right] \left(\prod_{i=0}^{n+1} h_i(x_i)\right),\end{aligned}$$

showing that the law of $X_{0:n+1}$ is $\nu(\mathrm{d}x_0)\prod_{i=0}^{n} P(x_i,\mathrm{d}x_{i+1})$ and (H_{n+1}) is thus proved. ■

As a consequence, the marginal law of X_n is given by integrating $\nu(\mathrm{d}x_0)\prod_{i=0}^{n-1} P(x_i,\mathrm{d}x_{i+1})$ over $x_{0:n-1}$ and thus, for all $A \in \mathscr{X}$,

$$\mathbb{P}(X_n \in A) = \int \cdots \int_{\mathsf{X}^{n+1}} \mathbb{1}_A(X_n)\nu(\mathrm{d}x_0)\prod_{i=0}^{n-1} P(x_i,\mathrm{d}x_{i+1}) = \nu P^n(A),$$

that is, νP^n is the distribution of X_n or in a compact notation, $\boxed{X_n \sim \nu P^n}$.

5.3 Canonical space

So far, the situation is the following: the $\{X_k : k \in \mathbb{N}\}$ are already given and we check the two items in the definition of a Markov chain (Theorem 5.2) with Markov kernel P and initial distribution ν. We now turn to the reverse situation where a couple (ν, P) of initial distribution and Markov kernel are given beforehand and we intend to construct the random variables $\{X_k : k \in \mathbb{N}\}$ on some convenient (common) probability space $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\{X_k : k \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν.

5.3.1 A simpler problem

We start with an easier problem where we only want to construct $\{X_k : k \in [0 : n-1]\}$ where n is some given positive integer. That is, we only consider **a finite range of integers** such that the first item in Theorem 5.2 is satisfied. For a given $\nu \in \mathsf{M}_1(\mathsf{X})$, define the triplet $(\Omega_n, \mathscr{G}_n, \mathbb{P}_{\nu,n})$ as follows:

- $\Omega_n = \mathsf{X}^{n+1}$, $\mathscr{G}_n = \mathscr{X}^{\otimes(n+1)}$ and $\mathbb{P}_{\nu,n}$ is the probability measure defined on $(\Omega_n, \mathscr{G}_n)$ by

$$\mathscr{G}_n \ni A \mapsto \mathbb{P}_{\nu,n}(A) = \int \cdots \int_{\mathsf{X}^{n+1}} \mathbb{1}_A(\omega_{0:n})\nu(\mathrm{d}\omega_0)\prod_{i=1}^{n} P(\omega_{i-1},\mathrm{d}\omega_i),$$

and for $\omega \in \Omega_n$, set $X_k(\omega) = \omega_k$, (that $X_k(\omega)$ is the projection of the k-th component of ω).

We aim to show that for all $A \in \mathscr{X}$ and all $k \in [0 : n-1]$, we have $\mathbb{P}_{\nu,n}(X_{k+1} \in A | X_{0:k-1}) = P(X_k, A)$, $\mathbb{P}_{\nu,n} - a.s.$ To do so, write for any $k \in [0 : n-1]$, any non-negative measurable function h on X^{k+1} and any $A \in \mathscr{X}$

$$\begin{aligned}
\mathbb{E}_{\nu,n}\left[h(X_{0:k})\mathbb{1}_A(X_{k+1})\right] &= \int \cdots \int_{\mathsf{X}^{k+2}} h(\omega_{0:k})\mathbb{1}_A(\omega_{k+1})\nu(\mathrm{d}\omega_0)\prod_{i=1}^{k+1} P(\omega_{i-1},\mathrm{d}\omega_i) \\
&= \int \cdots \int_{\mathsf{X}^{k+2}} h(\omega_{0:k})P(\omega_k, A)\nu(\mathrm{d}\omega_0)\prod_{i=1}^{k} P(\omega_{i-1},\mathrm{d}\omega_i) \\
&= \mathbb{E}_{\nu,n}\left[h(X_{0:k})P(X_k, A)\right].
\end{aligned}$$

Since h is arbitrary, this is equivalent to saying that $\mathbb{P}_{\nu,n}(X_{k+1} \in A | X_{0:k}) = P(X_k, A)\,, \quad \mathbb{P} - a.s.$

5.3.2 The general case

We now consider the general case where $k \in \mathbb{N}$ instead of $k \in [0 : n-1]$. Define the coordinate process (X_n) by $X_n(\omega) = \omega_n$ for all $\omega \in \mathsf{X}^{\mathbb{N}}$. We will sometimes use $X_{k:\ell} : \omega \mapsto (\omega_k, \ldots, \omega_\ell)$ for $k \le \ell$ and by extension $X_{k:\infty} : \omega \mapsto (\omega_k, \ldots, \omega_\ell, \omega_{\ell+1}, \ldots)$. In particular, $X_{0:\infty}(\omega) = I(\omega) = \omega$ where I is the identity function.

Theorem 5.5. ***(The canonical space)*** *Let $(\mathsf{X},\mathscr{X})$ be a measurable space and let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$. For every probability measure $\nu\in\mathsf{M}_1(\mathsf{X})$, there exists a unique probability measure $\mathbb{P}_\nu$ on the canonical space $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}})$ such that, under $\mathbb{P}_\nu$, the coordinate process $\{X_n\,:\,n\in\mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν.*

This result is often referred to as the *Ionescu-Tulcea theorem*. Its proof goes far beyond the scope of this course and we will admit it here. Some other, much simpler proofs exist and are based on the Kolmogorov extension theorem, but they hold at the price of additive assumptions on the space (in contrast, we only assume here that $(\mathsf{X},\mathscr{X})$ is a measurable space, which is quite minimal). In the canonical representation, we therefore set $\Omega=\mathsf{X}^{\mathbb{N}}$, $\mathscr{G}=\mathscr{X}^{\otimes\mathbb{N}}$ and $\mathbb{P}=\mathbb{P}_\nu$. In the particular case where the initial distribution ν is a Dirac mass, we use the compact notation $\boxed{\mathbb{P}_x=\mathbb{P}_{\delta_x}}$. Thus, the theorem allows to define not only one probability measure but a family of probability measures $(\mathbb{P}_\nu)_{\nu\in\mathsf{M}_1(\mathsf{X})}$ on the space of trajectories.

What are the relations between the probability measures $(\mathbb{P}_\nu)_{\nu\in\mathsf{M}_1(\mathsf{X})}$? A consequence of this theorem is that for all $A\in\mathscr{X}^{\otimes(n+1)}$, $\mathbb{P}_\nu(X_{0:n}\in A)=\int_A\nu(\mathrm{d}\omega_0)\prod_{i=1}^n P(\omega_{i-1},\mathrm{d}\omega_i)$. Replacing ν by δ_{x_0} and comparing the two obtained expressions, we get

$$\mathbb{P}_\nu(X_{0:n}\in A)=\int_{\mathsf{X}}\nu(\mathrm{d}x_0)\mathbb{P}_{x_0}(X_{0:n}\in A).$$

We can actually extend this result to any $A\in\mathscr{X}^{\otimes\mathbb{N}}$ by replacing the $n+1$ -tuple $X_{0:n}$ by the (infinite) trajectory $X_{0:\infty}$. First note the following equalities: $A=\{\omega\in A\}=\{\omega\in\Omega\,:\,X_{0:\infty}(\omega)\in A\}=\{X_{0:\infty}\in A\}$.

We then obtain the following identity: for all $A\in\mathscr{X}^{\otimes\mathbb{N}}$,

$$\mathbb{P}_\nu(A)=\mathbb{P}_\nu(X_{0:\infty}\in A)=\int_{\mathsf{X}}\nu(\mathrm{d}x_0)\mathbb{P}_{x_0}(X_{0:\infty}\in A)=\int_{\mathsf{X}}\nu(\mathrm{d}x_0)\mathbb{P}_{x_0}(A)\tag{5.2}$$

5.3.3 The Markov property.

Define the shift operator S by $S:\mathsf{X}^{\mathbb{N}}\ni\omega\mapsto\omega'\in\mathsf{X}^{\mathbb{N}}$ where $\omega=(\omega_i)_{i\in\mathbb{N}}$ and $\omega'=(\omega_{i+1})_{i\in\mathbb{N}}$

Theorem 5.6. ***(The Markov property)*** *For any $\nu\in\mathsf{M}_1(\mathsf{X})$, any non-negative or bounded function h on $\mathsf{X}^{\mathbb{N}}$ and any $n\in\mathbb{N}$,*

$$\mathbb{E}_\nu\left[h\circ S^k|\mathscr{F}_k\right]=\mathbb{E}_{X_k}[h]\,,\quad\mathbb{P}_\nu-a.s.\tag{5.3}$$

where $\mathscr{F}_k=\sigma(X_{0:k})$.

The expression of the Markov property, as it stands, may seem a bit cryptic. Recalling the definition of $X_{k:\infty}$, we have $\mathbb{P}_\nu-a.s.$,

$$\begin{aligned}\mathbb{E}_\nu\left[h(X_{k:\infty})|\mathscr{F}_k\right]&=\mathbb{E}_\nu\left[h\circ S^k\circ X_{0:\infty}|\mathscr{F}_k\right]=\mathbb{E}_\nu\left[h\circ S^k\circ I|\mathscr{F}_k\right]\\&=\mathbb{E}_\nu\left[h\circ S^k|\mathscr{F}_k\right]=\mathbb{E}_{X_k}[h]=\mathbb{E}_{X_k}[h\circ I]\\&=\mathbb{E}_{X_k}[h(X_{0:\infty})].\end{aligned}$$

Therefore, for any $\nu \in \mathsf{M}_1(\mathsf{X})$,

$$\boxed{\mathbb{E}_\nu\left[h(X_{k:\infty})|\mathscr{F}_k\right] = \mathbb{E}_{X_k}\left[h(X_{0:\infty})\right], \quad \mathbb{P}_\nu - a.s.}$$

is another **equivalent expression of the Markov property.** This expression may seem easier to deal with but the reader has to be at ease with both formulations. A stronger version of the Markov property exists and is called (as expected) the strong Markov property: its statement is (5.3) with the exception that k is replaced by a stopping time τ and that the identity only holds on the event $\{\tau < \infty\}$. The strong Markov property is extremely important in Markov Chain theory but, quite surprisingly, it is not needed in this basic course.

5.4 At this point...

a) I can write perfectly expression of Markov kernels if it is asked.
b) I understand the different notation μPQf...
c) I perfectly understand the canonical space.
d) I can understand (5.2).
e) I perfectly understand the Markov property (5.3).

5.5 Metropolis-Hastings algorithms

5.5.1 Invariant probability measures: existence

Definition 5.7. We say that $\pi \in \mathsf{M}_1(\mathsf{X})$ is an invariant probability measure for the Markov kernel P on $\mathsf{X} \times \mathscr{X}$ if $\pi P = \pi$.

In words, if (X_k) is a Markov chain with Markov kernel P and assuming that $X_0 \sim \pi$, then for all $k \geq 1$, we have $X_k \sim \pi$. (This is due to the fact that applying P^k on both sides of $\pi P = \pi$ shows that $\pi P^{k+1} = \pi P^k$and therefore, for all $k \in \mathbb{N}$, $\pi P^k = \pi$). This result on the (marginal) distribution of X_k may be extended to n-tuples.

More precisely, it can be readily checked that if π is an *invariant probability measure* for P, then the sequence of random variables $\{X_k : k \in \mathbb{N}\}$ is a *strongly stationary sequence* under $\mathbb{P}_\pi$ (in the sense that for all $n, p \in \mathbb{N}^*$, and all n-tuple $k_{1:n}$, the random vector $(X_{k_1}, \ldots, X_{k_n})$ follows the same distribution as $(X_{k_1+p}, \ldots, X_{k_n+p})$).

We now introduce the notion of reversibility for a Markov kernel. This will be of crucial importance for designing Markov kernels with a given invariant probability measure.

Definition 5.8. Let $\pi \in \mathsf{M}_1(\mathsf{X})$ and P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$. We say that P is π-reversible if and only if (with infinitesimal notation)

$$\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \pi(\mathrm{d}y)P(y,\mathrm{d}x), \tag{5.4}$$

that is, for all measurable bounded or non-negative functions h on $\left(\mathsf{X}^2, \mathscr{X}^{\otimes 2}\right)$,

$$\iint_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \iint_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}y)P(y,\mathrm{d}x). \tag{5.5}$$

In words, a Markov kernel P is π-reversible if and only if the probability measure $\pi(\mathrm{d}x)P(x,\mathrm{d}y)$ is symmetric with respect to (x,y).

Proposition 5.9 *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$. Let $\pi\in\mathsf{M}_1(\mathsf{X})$ such that P is π-reversible, then the Markov kernel P is π-invariant.*

PROOF. For any $A\in\mathscr{X}$, we have by the reversibility relation

$$\pi P(A) = \iint_{\mathsf{X}^2} \mathbb{1}_A(y)\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \iint_{\mathsf{X}^2} \mathbb{1}_A(y)\pi(\mathrm{d}y)P(y,\mathrm{d}x) = \int_A \pi(\mathrm{d}y)\underbrace{P(y,\mathsf{X})}_{1} = \pi(A),$$

which finishes the proof. ■

What are the consequences? If you want to check easily that a kernel P is π-invariant, it is sufficient to check that it is π-reversible.

5.5.2 Metropolis-Hastings (MH) algorithms

In this section, we are given a probability measure $\pi\in\mathsf{M}_1(\mathsf{X})$ and the idea now is to construct a Markov chain $\{X_k : k\in\mathbb{N}\}$ admitting π as invariant probability measure, in which case we say that π is a target distribution. In other words, we try to find a Markov kernel P on $\mathsf{X}\times\mathscr{X}$ such that P is π-invariant. The reason for that is that an invariant probability measure will be a good candidate for the "limiting" distribution of $\{X_k : k\in\mathbb{N}\}$ (in some sense to be defined) and this in turn, will allow us to provide an approximation $\pi(h)$:

$$\pi(h) = \int_{\mathsf{X}} h(x)\pi(\mathrm{d}x) \approx n^{-1}\sum_{k=0}^{n-1} h(X_i).$$

5.5.2.1 Construction of the kernel

For simplicity we now assume that π has a density with respect to some dominating σ-finite measure λ and by abuse of notation, we also denote by π this density, that is we write $\pi(\mathrm{d}x)=\pi(x)\lambda(\mathrm{d}x)$ and we assume that this density π is **positive**.

Moreover, let Q be Markov kernel on $\mathsf{X}\times\mathscr{X}$ such that $Q(x,\mathrm{d}y)=q(x,y)\lambda(\mathrm{d}y)$, that is, for any $x\in\mathsf{X}$, $Q(x,\cdot)$ is also dominated by λ and denoting by $q(x,\cdot)$ this density, we assume for simplicity that $q(x,y)$ is **positive** for all $x,y\in\mathsf{X}$. At this stage, there is almost no link between Q and the target distribution π.

For a given function $\alpha:\mathsf{X}^2\to[0,1]$, consider the following Algorithm 4.

In words, Q allows to propose a candidate for the next value of the Markov chain (X_k) and this candidate will be accepted or refused according to a probability that depends on the function α.

We will now choose conveniently α in such a way that (X_k) is a Markov chain with invariant probability measure π. The concept of π-reversibility will help us. To do so, let us assume that for all $x,y\in\mathsf{X}$,

$$\pi(x)\alpha(x,y)q(x,y) = \pi(y)\alpha(y,x)q(y,x), \tag{5.6}$$

and let us show that it implies that the Markov kernel P associated to (X_k) is π-reversible.

Algorithm 4 The Metropolis-Hastings algorithm

Input: n
Output: $X_0, \dots, X_n$
At $t = 0$, draw X_0 according to some arbitrary distribution
for $t \leftarrow 0, \dots, n-1$ **do**
    Draw independently $Y_{t+1} \sim Q(X_t, \cdot)$ and $U_{t+1} \sim \mathrm{Unif}(0,1)$
    Set $X_{t+1} = \begin{cases} Y_{t+1} & \text{if } U_{t+1} \leq \alpha(X_t, Y_{t+1}) \\ X_t & \text{otherwise} \end{cases}$
end for

First, we write down the Markov kernel associated to (X_k): in passing, this is an excellent opportunity to check if we are able to express explicitly a Markov kernel by analyzing conveniently the update transition. Denote $\mathscr{F}_t = \sigma(X_0, U_{1:t}, Y_{1:t})$ and note that (X_t) is adapted to the filtration $(\mathscr{F}_t)$ (which is equivalent to $\sigma(X_{0,t}) \subset \mathscr{F}_t$). Then, setting $\bar{\alpha}(x) = 1 - \int_{\mathsf{X}} Q(x, \mathrm{d}y)\alpha(x,y)$, we have for any bounded or non-negative measurable function h on X and any $t \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[h(X_{t+1})|\mathscr{F}_t\right] &= \mathbb{E}\left[\mathbb{1}_{\{U_{t+1} < \alpha(X_t, Y_{t+1})\}} h(Y_{t+1})|\mathscr{F}_t\right] + \mathbb{E}\left[\mathbb{1}_{\{U_{t+1} \geq \alpha(X_t, Y_{t+1})\}}|\mathscr{F}_t\right] h(X_t) \\
&= \int_{\mathsf{X}} Q(X_t, \mathrm{d}y)\alpha(X_t, y)h(y) + \bar{\alpha}(X_t)h(X_t) \\
&= \int_{\mathsf{X}} \underbrace{\left[Q(X_t, \mathrm{d}y)\alpha(X_t, y) + \bar{\alpha}(X_t)\delta_{X_t}(\mathrm{d}y)\right]}_{P^{MH}_{\langle \pi, Q\rangle}(X_t, \mathrm{d}y)} h(y) = P^{MH}_{\langle \pi, Q\rangle} h(X_t).
\end{aligned}$$

Therefore, $\{X_t : t \in \mathbb{N}\}$ is a Markov chain with Markov kernel

$$P^{MH}_{\langle \pi, Q\rangle}(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x,y) + \bar{\alpha}(x)\delta_x(\mathrm{d}y) \tag{5.7}$$

Lemma 5.10 *The Markov kernel $P^{MH}_{\langle \pi, Q\rangle}$ is π-reversible if and only if*

$$\pi(\mathrm{d}x)Q(x, \mathrm{d}y)\alpha(x,y) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)\alpha(y,x). \tag{5.8}$$

In this literature, (5.8) is often called the **detailed balance condition.** PROOF. First, note that

$$\pi(\mathrm{d}x)\bar{\alpha}(x)\delta_x(\mathrm{d}y) = \pi(\mathrm{d}y)\bar{\alpha}(y)\delta_y(\mathrm{d}x) \tag{5.9}$$

Indeed, for any measurable function h on X^2, we have

$$\begin{aligned}
\iint_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}x)\bar{\alpha}(x)\delta_x(\mathrm{d}y) &= \int_{\mathsf{X}} h(x,x)\pi(\mathrm{d}x)\bar{\alpha}(x) \\
&= \int_{\mathsf{X}} h(y,y)\pi(\mathrm{d}y)\bar{\alpha}(y) = \iint_{\mathsf{X}^2} h(x,y)\pi(\mathrm{d}y)\bar{\alpha}(y)\delta_y(\mathrm{d}x).
\end{aligned}$$

Combining (5.7) with (5.9), we obtain that $P^{MH}_{\langle \pi, Q\rangle}$ is π-reversible if and only if the detailed balance condition (5.8) is satisfied. This completes the proof. ■

5.5.2.2 Acceptance probability

We now make use of Lemma 5.10 in order to find an explicit expression of the acceptance probability α. We have the following lemma.

Lemma 5.11 *Denote* $\alpha^{MH}(x,y) = \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right)$ *and* $\alpha^{b}(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)+\pi(y)q(y,x)}$. *Then, any* $\alpha \in \left\{\alpha^{MH}, \alpha^{b}\right\}$ *satisfies the detailed balance condition* (5.8). *Moreover, any other* $\alpha \in [0,1]$ *that satisfies the detailed balance condition is dominated by* α^{MH} *in the sense that: for* $\lambda^{\otimes 2}$*-almost all* $x, y \in \mathsf{X}$,

$$\alpha(x,y) \leq \alpha^{MH}(x,y) \tag{5.10}$$

PROOF. The fact that any $\alpha \in \left\{\alpha^{MH}, \alpha^{b}\right\}$ satisfies $\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x)$ for $\lambda^{\otimes 2}$-almost all $x, y \in \mathsf{X}$, is immediate, by replacing α by its expression. It remains to check (5.10) . Assume now that for $\lambda^{\otimes 2}$-almost all $x, y \in \mathsf{X}$,

$$\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x).$$

then, using that $\alpha(y,x) \leq 1$ shows that $\alpha(x,y) \leq \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$. Moreover, $\alpha(x,y) \leq 1$ and this finally implies

$$\alpha(x,y) \leq \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) = \alpha^{MH}(x,y),$$

which completes the proof. ■

According to (5.10), α^{MH} is actually the highest acceptance probability among the acceptance probabilities such that $P^{MH}_{\langle \pi, Q\rangle}$ is π-reversible and therefore, this acceptance is widely used in practice (in the sense that we expect that a Markov kernel that accepts often, explores the space more rapidly and therefore is preferable to another one with less acceptance probability). In what follows, unless otherwise stated, ***we implicitly assume that the Markov kernel*** $P^{MH}_{\langle \pi, Q\rangle}$ ***is associated to the acceptance probability*** α^{MH}.

Example 5.12 ***(The independence sampler)*** *If the proposition kernel is* $Q(x, \mathrm{d}y) = q(y)\lambda(\mathrm{d}y)$ *where* q *is a density wrt* λ *on* X, *then at each time step, the proposed candidate is drawn irrespective of the current value of the Markov chain (this is because,* $Q(x, \mathrm{d}y)$ *does not depend on* x*), that is, in the step 2(a) of Algorithm 4, we draw* $Y_{t+1} \sim q(\cdot)$. *In such case, the acceptance probability is* $\alpha(x,y) = \min\left(\frac{\pi(y)q(x)}{\pi(x)q(y)}, 1\right)$ *and the Metropolis-Hastings algorithm is called the* Independence Sampler.

Another important example is the following.

Example 5.13 ***(The random walk MH sampler)*** *If* $\mathsf{X} = \mathbb{R}^p$ *and if the proposition kernel is* $Q(x, \mathrm{d}y) = q(y-x)\lambda(\mathrm{d}y)$ *where* q *is a symmetric density wrt* λ *on* X, *(by symmetric, we mean that* $q(u) = q(-u)$ *for all* $u \in \mathsf{X}$*) then at each time step in Algorithm 4, we draw a candidate* $Y_{t+1} \sim q(y - X_k)\,\lambda(\mathrm{d}y)$. *In such case, the acceptance probability is* $\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$ *and the associated algorithm is called the* (symmetric) Random Walk Metropolis-Hasting. *Another way of writing the proposition update is* $Y_{t+1} = X_k + \eta_k$ *where* $\eta_k \sim q(\cdot)$.

5.6 Invariant probability measure: uniqueness

We start with a very simple lemma that will be useful for finding sufficient conditions for uniqueness.

Lemma 5.14 *If* P *admits two distinct invariant probability measures, it also admits distinct invariant probability measures* π_0 *and* π_1 *that are mutually singular, i.e., such that there exists* $A \in \mathscr{X}$ *such that* $\pi_0(A) = \pi_1(A^c) = 0$.

PROOF. Let ζ_0, ζ_1 be two distinct invariant probability measures for P. Both have densities with respect to some common dominating measure (for example, taking $\zeta = \zeta_1 + \zeta_2$, we have that ζ dominates both ζ_0 and ζ_1, which can be seen from

the implication $\zeta(A)=0 \Rightarrow (\zeta_1(A)=0 \text{ and } \zeta_2(A)=0)$ for any $A\in\mathscr{X}$ and according to the Radon Nikodym theorem, if a measure dominates another one, the latter has a density with respect to the former). Write then $\zeta_0(\mathrm{d}x)=f_0(x)\zeta(\mathrm{d}x)$ and $\zeta_1(\mathrm{d}x)=f_1(x)\zeta(\mathrm{d}x)$ where f_0,f_1 are non-negative measurable functions on X. Define the positive part $(\zeta_1-\zeta_0)^+$and the negative part $(\zeta_1-\zeta_0)^-$ of the signed measure $\zeta_1-\zeta_0$ by $(\zeta_1-\zeta_0)^+(\mathrm{d}x)=[f_1(x)-f_0(x)]^+\zeta(\mathrm{d}x)$ and $(\zeta_1-\zeta_0)^-(\mathrm{d}x)=[f_1(x)-f_0(x)]^-\zeta(\mathrm{d}x)$. Then,

$$\begin{aligned}(\zeta_1-\zeta_0)^+P\mathbb{1}_A &= \int_{\mathsf{X}}\zeta(\mathrm{d}x)[f_1(x)-f_0(x)]^+P(x,A)\\ &\geq \int_{\mathsf{X}}\zeta(\mathrm{d}x)[f_1(x)-f_0(x)]P(x,A)\\ &\geq \zeta_1P(A)-\zeta_0P(A)=\zeta_1(A)-\zeta_0(A).\end{aligned}$$

Therefore, $(\zeta_1-\zeta_0)^+P$ is a (non-negative) measure that is greater than the signed measure $\zeta_1-\zeta_0$. Since the positive part $(\zeta_1-\zeta_0)^+$ is also the smallest (non-negative) measure that is greater than $\zeta_1-\zeta_0$, we conclude that $(\zeta_1-\zeta_0)^+\leq(\zeta_1-\zeta_0)^+P$. The measure $(\zeta_1-\zeta_0)^+P-(\zeta_1-\zeta_0)^+$ is therefore non-negative and we have

$$[(\zeta_1-\zeta_0)^+P-(\zeta_1-\zeta_0)^+](\mathsf{X})=\int_{\mathsf{X}}(\zeta_1-\zeta_0)^+(\mathrm{d}x)\underbrace{P(x,\mathsf{X})}_{1}-(\zeta_1-\zeta_0)^+(\mathsf{X})=0.$$

Finally, $(\zeta_1-\zeta_0)^+=(\zeta_1-\zeta_0)^+P$. The probability measure $\pi_0=\frac{(\zeta_1-\zeta_0)^+}{(\zeta_1-\zeta_0)^+(\mathsf{X})}$ is thus an invariant probability measure for P. Replacing $(\zeta_1-\zeta_0)^+$ by $(\zeta_1-\zeta_0)^-$, we obtain in the same way that $\pi_1=\frac{(\zeta_1-\zeta_0)^-}{(\zeta_1-\zeta_0)^-(\mathsf{X})}$ is an invariant probability measure. We can easily check that taking $A=\{f_0\geq f_1\}$, we have $\pi_0(A)=\pi_1(A^c)=0$, showing that these probability measures are mutually singular. ■

To be rigorous, in the course of the proof, we actually need some results on the positive and negative part of a signed-measure. The interested reader may work on the following exercise to fully understand the previous proof:

Exercise 5.15. Define $\mathsf{M}_s(\mathsf{X})$ the set of signed-measures. Let $\mu\in\mathsf{M}_s(\mathsf{X})$ and assume that $\mu\preceq\zeta$ where $\zeta\in\mathsf{M}_+(\mathsf{X})$ (in the sense that we have the implication: if for some $A\in\mathscr{X}$, $\zeta(A)=0$, then $\mu(A)=0$). According to the Radon-Nikodym theorem, there exists a measurable function h such that $\mu(\mathrm{d}x)=h(x)\zeta(\mathrm{d}x)$. Define $\mu^+(\mathrm{d}x)=|h(x)|\zeta(\mathrm{d}x)$.

1. Show that the measure $\mu^+$is well-defined (in the sense that the measure $|h(x)|\zeta(\mathrm{d}x)$ does not depend on the measure ζ, provided that the ζ dominates μ.)
2. Show that for ζ-almost all $x\in\mathsf{X}$, $|h(x)|\leq 1$.
3. Assume that there exists $\nu\in\mathsf{M}_+(\mathsf{X})$ such that for all $A\in\mathscr{X}$, we have $\mu(A)\leq\nu(A)$. Show that $\mu^+(A)\leq\nu(A)$ for all $A\in\mathscr{X}$.

We now make use of Lemma 5.14 in order to give a sufficient condition for uniqueness.

Proposition 5.16 *Assume that there exists a non-null measure $\mu\in\mathsf{M}_+(\mathsf{X})$ satisfying the following property:*

- *For all $A\in\mathscr{X}$ such that $\mu(A)>0$ and for all $x\in\mathsf{X}$, there exists $n\in\mathbb{N}$ such that $P^n(x,A)>0$.*

Then, P admits at most one invariant probability measure.

If the assumption of Proposition 5.16 holds, we say that P is **μ-irreducible** and in such case, μ is called an **irreducibility measure** for P. PROOF. The proof is by contradiction. Assume that there exists two distinct invariant probability measures. According to Lemma 5.14, we can consider two invariant probability measures π_1 and π_2 that are mutually singular. Under the assumptions of the Proposition, let $A\in\mathscr{X}$ such that $\mu(A)>0$. Then, for any $i\in\{1,2\}$,we have

$$0<\int_{\mathsf{X}}\pi_i(\mathrm{d}x)\underbrace{\sum_{n=0}^{\infty}P^n(x,A)}_{>0}=\sum_{n=0}^{\infty}\pi_iP^n(A)=\sum_{n=0}^{\infty}\pi_i(A),$$

which in turn implies that $\pi_i(A) > 0$. The contraposed implication gives that if for some $i \in \{1,2\}$, $\pi_i(A) = 0$, then $\mu(A) = 0$. Now, since $\{\pi_i : i \in \{1,2\}\}$ are mutually singular, there exists $A \in \mathscr{X}$ such that $\pi_1(A) = \pi_2(A^c) = 0$ and this shows that $\mu(A) = \mu(A^c) = 0$ which is impossible. ■

5.6.1 Application to Metropolis-Hastings algorithms.

We have already seen that $P^{MH}_{\langle\pi,Q\rangle}$ is π-invariant and we have assumed that $Q(x,\mathrm{d}y) = q(x,y)\lambda(\mathrm{d}y)$ and $\pi(\mathrm{d}y) = \pi(y)\lambda(\mathrm{d}y)$ and for simplicity, we said that for all $x,y \in \mathsf{X}$, $q(x,y) > 0$ and $\pi(y) > 0$. This in turn implies that $\alpha(x,y) = \alpha^{MH}(x,y) > 0$ and therefore if $\lambda(A) > 0$, then for all $x \in \mathsf{X}$,

$$P(x,A) \geq \int_A \underbrace{q(x,y)\alpha(x,y)}_{>0}\lambda(\mathrm{d}y) > 0.$$

This shows that P is λ-irreducible and therefore π **is the unique invariant probability measure for** P**.**

5.7 After studying this chapter...

a) I can understand the link between reversibility and invariant measure.
b) If asked, I can check that MH chains are reversible.
c) I understand the form of the acceptance probability.
d) Independence sampling and random walk MH have no secrets for me and I am able to implement them.
e) If asked, I can show that an invariant measure is unique by using Proposition 5.16
f) I find that Metropolis-Hasting algorithms are magical.

Chapter 6

Variance reduction techniques

Contents

Keywords: *Importance sampling, antithetic variates, stratification, control variates, conditioning, weak discrepancy, Van der Corput sequences, Halton sequences.*

We have seen in the previous chapters some tools for sampling exactly or approximately according to a target distribution defined by a density f or by the solution of a SDE (in particular, we have seen discretization schemes for some SDE), the final goal remaining to give the most precise numerical value of a quantity of the type $\mathbb{E}[h(Y)] = \int h(y)f(y)\mathrm{d}y$ where Y has the density f.

This chapter is devoted to various methods that allow to produce estimators $\bar{S}_n$ of $\mathbb{E}[h(Y)]$ such that the quadratic deviation for a fixed n is the smallest possible. All the estimators we will see here will be unbiased and strongly convergent: $\mathbb{E}\left[\bar{S}_n\right] = \mathbb{E}[h(Y)]$ et $\bar{S}_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \mathbb{E}[h(Y)]$, so that the standard deviation writes

$$\mathbb{E}\left[\bar{S}_n - \underbrace{\mathbb{E}[h(Y)]}_{\mathbb{E}[\bar{S}_n]}\right]^2$$

which is also the variance $\mathbb{V}\mathrm{ar}(\bar{S}_n)$. Let us now look at some sampling methods for reducing the variance compared to a standard Monte Carlo estimator.

6.1 Importance Sampling

We refer the reader to Section (4.2.1) where the importance sampling is introduced and where it is shown that the asymptotic variance of the importance sampling estimator did not necessarily reach its optimal variance when the instrumental distribution is the target distribution. Let us recall the method. Consider a r.v. $Y \sim f$ where f is a density. Suppose we want to approximate $\mathbb{E}[h(Y)]$. An (instrumental) density g is then selected. And we draw an n-sample $(X_1, \ldots, X_n)$ according to the distribution of density g and we set

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} h(X_i)$$

On the contrary to the chapter 4, we will focus on the non-asymptotic variance (instead of the asymptotic variance) of $\bar{S}_n$.

$$\mathbb{V}\mathrm{ar}(\bar{S}_n) = \frac{\mathbb{V}\mathrm{ar}\left(\frac{f(X)}{g(X)}h(X)\right)}{n} = \frac{\int f(x)\frac{f(x)}{g(x)}h^2(x)\mathrm{d}x - (\int h(x)f(x)\mathrm{d}x)^2}{n}$$

We thus obtain a reduction of the variance in comparison with a standard Monte Carlo estimator iff

$$\int f(x)\frac{f(x)}{g(x)}h^2(x)\mathrm{d}x \leq \int f(x)h^2(x)$$

6.2 Antithetic variates

Suppose we wish to evaluate $\mathbb{E}(h(U))$ where $U \sim \mathscr{U}[0,1]$ by a Monte Carlo method. If we have an n-sample $(U_1,\ldots,U_n)$ of i.i.d. r.v. according to the distribution of U, a classical method would be to consider

$$S_n = \frac{1}{n}\sum_{i=1}^{n} h(U_i).$$

Nevertheless, one can note that if $U \sim \mathscr{U}[0,1]$ then $1-U$ also follows the distribution $\mathscr{U}[0,1]$. It is then tempting to consider

$$S'_n = \frac{1}{n}\sum_{i=1}^{n} h(1-U_i).$$

or

$$\bar{S}_{2n} = \frac{1}{2n}\sum_{i=1}^{n} (h(U_i)+h(1-U_i)) \tag{6.1}$$

Clearly, $\bar{S}_{2n}$ is an *unbiased* and *strongly convergent* estimator of $\mathbb{E}(h(U))$. Let us calculate now the variance of $\bar{S}_{2n}$:

$$\mathbb{V}\mathrm{ar}(\bar{S}_{2n}) = \frac{1}{4n}\left(2\mathbb{V}\mathrm{ar}(h(U)) + 2\mathbb{C}\mathrm{ov}(h(U),h(1-U))\right) = \mathbb{V}\mathrm{ar}(S_{2n}) + \frac{1}{2n}\mathbb{C}\mathrm{ov}(h(U),h(1-U))$$

If $\mathbb{C}\mathrm{ov}(h(U),h(1-U))$ is negative, then the speed of convergence of $\bar{S}_{2n}$ is better than the one of S_{2n}. The following lemma gives a sufficient condition that ensures the negativity of the covariance of two functions of the same random variable.

Lemma 6.1 CORRELATION INEQUALITY- *Let X be a r.v and h, g be two monotone functions such that one is decreasing and the other one increasing, then $\mathbb{C}ov(h(X),g(X)) \leq 0$.*

PROOF. Assume first that $\mathbb{E}(g(X)) = 0$ and $\mathbb{E}(h(X)) = 0$. Let x and y be two real numbers. We have

$$(h(x)-h(y))(g(x)-g(y)) \leq 0.$$

If X and Y are independent, with the same distribution, then

$$\mathbb{E}\left[(h(X)-h(Y))(g(X)-g(Y))\right] \leq 0.$$

Expanding this quantity, we obtain

$$\mathbb{E}[h(X)g(X)] + \mathbb{E}[h(Y)g(Y)] - \mathbb{E}[h(X)]\mathbb{E}[g(Y)] - \mathbb{E}[h(Y)]\mathbb{E}[g(X)] \leq 0$$

which is the desired result by noting that $X \overset{\mathscr{L}}{\equiv} Y$. The proof is completed for $\mathbb{E}(g(X)) = \mathbb{E}(h(X)) = 0$. For the general case, we set $\bar{g}(x) = g(x) - \mathbb{E}(g(X))$ and $\bar{h}(x) = h(x) - \mathbb{E}(h(X))$ and we apply the previous result. ■

Corollary 6.2 *If $U \sim \mathscr{U}[0,1]$ and if h is nondecreasing then, $\mathbb{C}ov(h(U), h(1-U)) \leq 0$.*

PROOF. Take $g(x) = h(1-x)$ in the previous lemma. ■

Example 6.3 *Let $\psi(x) = (\lambda e^{\sigma x} - K)^+$. We want to approximate $\mathbb{E}[\psi(G)]$ or $G \sim \mathscr{N}(0.1)$ by a Monte Carlo method (of course, we know that Black and Scholes provides an exact closed-form formula, without any approximation). In this case, noting that G and $-G$ have the same law, one can choose an estimator of the form*

$$\bar{S}_n = \frac{1}{2n} \sum_{i=1}^{n} (\psi(G_i) + \psi(-G_i))$$

where $G_i \overset{i.i.d.}{\sim} \mathscr{N}(0,1)$. Clearly the proof of $\mathbb{C}ov(\psi(G), \psi(-G)) \leq 0$ is a direct consequence of of the Lemma 6.1 with $f(x) = \psi(x)$ and $g(x) = \psi(-x)$.

More generally, we can use a Monte Carlo method with antithetic variate to approximate $\mathbb{E}[h(Y)]$ if $h : \mathbb{R} \to \mathbb{R}$ is monotone and if there is a decreasing function φ such as $h(Y) \overset{\mathscr{L}}{\equiv} h \circ \varphi(Y)$.

Exercise 6.4. GENERALISATION- Let Y be a r.v. with values in $\mathbb{R}^d$. Let $h : \mathbb{R}^d \to \mathbb{R}$ and $\tilde{h} : \mathbb{R}^d \to \mathbb{R}$ such that h is increasing wrt to any of its coordinates and $\tilde{h}$ is decreasing wrt to any of its coordinates. Assume furthermore that

$$h(Y) \overset{\mathscr{L}}{\equiv} \tilde{h}(Y).$$

Show that under some integrability assumptions (that should be explicit), the quantity $\frac{1}{2n} \sum_{i=1}^{n} [h(Y_i) + \tilde{h}(Y_i)]$ is an unbiased and strongly convergent estimator, with a lower variance than the one of the standard Monte Carlo estimator $\frac{1}{n} \sum_{i=1}^{n} h(Y_i)$.

6.3 Control Variates

In order to reduce the variance of the Monte Carlo method for the calculation of $\mathbb{E}(Y)$, another approach consists in finding a r.v X, such that its expectation **is available in closed-form** and such that $\boxed{\mathbb{V}\text{ar}(Y - \alpha X) \leq \mathbb{V}\text{ar}(Y)}$ for some $\alpha \in \mathbb{R}$.

We then consider $(Y_1, ..., Y_n)$ an n-sample according to the distribution of Y and $(X_1, ..., X_n)$ an n-sample according to the distribution of X. We then set

$$\bar{S}_n = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \alpha(X_i - \mathbb{E}[X])] = \frac{1}{n} \sum_{i=1}^{n} Y_i - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] \right)$$

Of course, $\bar{S}_n$ is an *unbiased* and *strongly convergent* estimator of $\mathbb{E}(Y)$. Indeed by the LLN, we have $\bar{S}_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} \mathbb{E}(Y)$. Now, setting $S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$

$$\mathbb{V}\text{ar}(\bar{S}_n) = \frac{\mathbb{V}\text{ar}[Y - \alpha(X - \mathbb{E}[X])]}{n} = \frac{\mathbb{V}\text{ar}[Y - \alpha X]}{n} \leq \frac{\mathbb{V}\text{ar}[Y]}{n} = \mathbb{V}\text{ar}(S_n)$$

which shows that the variance is smaller than the one obtained by a standard Monte Carlo method. We then say that X is a *control variate* for Y.

Unfortunately, there is no general method for creating a control variate X starting from Y. This is on a case-by-case basis. However, to obtain a "good" control variate X, we must keep in mind that $Y-\alpha X$ should "vary" less than Y alone and so, X must "follow" more or less the evolution of Y - replacing if necessary X by $-X$, (if Y is large, also should be X and if Y small, also should be X) while X still has the advantage over Y to have an expectation which can be explicitely calculated.

6.3.0.1 Optimisation of α

Once the control variate X chosen, the problem is then to select the most efficient coefficient α. We have the following lemma which turns out to be an orthogonal projection result:

Lemma 6.5 *Let Y and X be two square integrable r.v integrable, then $\mathbb{V}ar(Y-\alpha^\star X)=\inf_{\alpha\in\mathbb{R}}\mathbb{V}ar(Y-\alpha X)$ where we have set*

$$\alpha^\star=\frac{\mathbb{C}ov(Y,X)}{\mathbb{V}ar(X)}$$

PROOF. Note that, using the notation $\|\cdot\|=\sqrt{<\cdot,\cdot>}$ where $<U,V>=\mathbb{E}(UV)$ is the usual scalar product on L^2 random variables,

$$\mathbb{V}\mathrm{ar}(Y-\alpha X)=\mathbb{V}\mathrm{ar}[Y-\mathbb{E}(Y)-\alpha(X-\mathbb{E}[X])]=\|Y-\mathbb{E}(Y)-\alpha(X-\mathbb{E}[X])\|^2$$

Therefore, the optimum coefficient $\alpha=\alpha^\star$ in obtained by searching in the space $\mathscr{H}=\{\alpha[X-\mathbb{E}[X]];\alpha\in\mathbb{R}\}$ the closest element from $Y-\mathbb{E}(Y)$ for the norm associated to a scalar product: this is exactly the orthogonal projection of the vector $Y-\mathbb{E}(Y)$ on the subspace $\mathscr{H}$. We then obtain (by the classical formaula for the projection on an orthogonal basis, knowing that here, the basis consists in only one vector!)

$$\mathscr{P}_{\perp}^{Y-\mathbb{E}(Y)}[\mathscr{H}]=<Y-\mathbb{E}(Y),\frac{X-\mathbb{E}[X]}{\|X-\mathbb{E}[X]\|}>\frac{X-\mathbb{E}[X]}{\|X-\mathbb{E}[X]\|}=\frac{<Y-\mathbb{E}(Y),X-\mathbb{E}[X]>}{<X-\mathbb{E}[X],X-\mathbb{E}[X]>}(X-\mathbb{E}[X])=\frac{\mathbb{C}\mathrm{ov}(Y,X)}{\mathbb{V}\mathrm{ar}(X)}(X-\mathbb{E}[X])$$

hence the value of $\alpha^\star$. ■

From this lemma, the estimator issued from the control variate technique writes

$$\bar{S}_n=\frac{1}{n}\sum_{i=1}^n Y_i+\frac{\mathbb{C}\mathrm{ov}(Y,X)}{\mathbb{V}\mathrm{ar}(X)}\left(\frac{1}{n}\sum_{i=1}^n X_i-\mathbb{E}[X]\right)$$

For this estimator to be really used in practise, it is therefore necessary to know explicitly not only $\mathbb{E}[X]$ but also $\mathbb{V}\mathrm{ar}(X)$ and $\mathbb{C}\mathrm{ov}(Y,X)$; and the latter is not known in general since $\mathbb{E}(Y)$ is not known itself (this is precisely what we have been looking for since the beginning of this course). We are therefore forced to estimate this covariance, taking care that this additional estimation error does not harm the decrease in variance.

Remark 6.6 *We can see the Monte Carlo methods with antithetic variates as some particular cases of control variates technique. To see this, let us take again the example we used for antithetic variates (6.1):*

$$\frac{1}{n}\sum_{i=1}^n\frac{h(U_i)+h(1-U_i)}{2}=\frac{1}{n}\sum_{i=1}^n h(U_i)-\underbrace{\frac{h(U_i)-h(1-U_i)}{2}}_{X_i}$$

where $\mathbb{E}(X_i)=0$, which is thus known exactly and this situation finally falls down into the control variates technique.

Exercise 6.7. THE CALL-PUT PARITY- In a simplified manner, the payoff of a selling option (a Call) can be written as $(\lambda e^{\sigma G}-K)_+$ where $G\sim\mathscr{N}(0,1)$. Noting that $(\lambda e^{\sigma G}-K)_+-(K-\lambda e^{\sigma G})_+=\lambda e^{\sigma G}-K$, propose a control variate technique for the calculation of $\mathbb{E}(\lambda e^{\sigma G}-K)_+$.

Exercise 6.8. KEMNA AND VORST TECHNIQUE FOR ASIAN OPTIONS.- We want to approximate the price of an asian put in a Black and Scholes model where the price of the risky asset is given by $S_t = x\mathrm{e}^{(r-\frac{\sigma^2}{2})t+\sigma W_t}$. The payoff of the asian put under the risk neutral probability is equal to

$$P = \mathbb{E}\left[\mathrm{e}^{-rT}\left(K - \frac{1}{T}\int_0^T S_u \mathrm{d}u\right)_+\right]$$

Estimate this expectation using a Monte Carlo method with control variates: $X = \mathrm{e}^{-rT}(K - \exp(Z))_+$ where $Z = \frac{1}{T}\int_0^T \ln S_u \mathrm{d}u$. Explain the method and calculate $\mathbb{E}[X]$.

Exercise 6.9. BASKET OPTIONS- Consider d assets such that the price at time T writes in the following way (for $i = 1, \ldots, d$),

$$S_T^i = x_i \exp\left[\left(\underbrace{r - \frac{1}{2}\sum_{j=1}^p \sigma_{ij}^2}_{\beta_i}\right)T + \sum_{j=1}^p \sigma_{ij} W_T^j\right]$$

We want approximate the price of a basket put: $\mathbb{E}(K-Y)_+$ where $Y = \sum_{i=1}^d a_i S_T^i$ (with $a_i \geq 0$ and $\sum_{i=1}^d a_i = 1$). For this, we set $m = \sum_{i=1}^d a_i x_i$ and we approximate Y/m with $X = \exp\left[\sum_{i=1}^d \frac{a_i x_i}{m}\left(\beta_i T + \sum_{j=1}^p \sigma_{ij} W_T^j\right)\right]$. Explain the method, give the control variate associated to X and calculate the expectation.

6.4 Conditioning

We wish to approximate $\mathbb{E}(g(X,Y))$ where the random vector (X,Y) has the density $(x,y) \mapsto f(x,y)$. Suppose that the function h defined by $\boxed{h(X) = \mathbb{E}[g(X,Y)|X]}$ is **available in a closed-form** and $\mathbb{E}|g(X,Y)| < \infty$; assume in addition that the distribution of X can be *sampled*. Then, the quantity $\mathbb{E}(g(X,Y))$ can be approximated by the estimator

$$\bar{S}_n = \frac{1}{N}\sum_{i=1}^N h(X_i)$$

where (X_i) are i.i.d. with density $x \mapsto \int f(x,y)\mathrm{d}y$ (obtained by marginalization). $\bar{S}_n$ is a *strongly convergent* and *unbiased* estimator of $\mathbb{E}(g(X,Y))$ with a lower variance than

$$S_n = \frac{1}{N}\sum_{i=1}^N g(X_i, Y_i)$$

where the (X_i, Y_i) are i.i.d. with density f. Indeed, the consistency of the estimator is given by the LLN combined with $\mathbb{E}(h(X)) = \mathbb{E}[\mathbb{E}(g(X,Y)|X)] = \mathbb{E}(g(X,Y))$. Moreover, by the tower property and the formula for conditional variances:

$$\mathbb{V}\mathrm{ar}(g(X,Y)) = \mathbb{V}\mathrm{ar}[\underbrace{\mathbb{E}(g(X,Y)|X)}_{h(X)}] + \underbrace{\mathbb{E}[\mathbb{V}\mathrm{ar}(g(X,Y)|X)]}_{\geq 0} \geq \mathbb{V}\mathrm{ar}(h(X))$$

Remark 6.10 *According to this last inequality, the variance of conditional expectation is always less than the variance (without conditioning). This property is also used in statistics under the name of Rao-Blackwell's method or "Rao-Blackwellisation", it is not the subject of this course but we advise the reader to make the link between these methods and the Rao-Blackwell estimator whose definition and properties can be found in the literature...*

Exercise 6.11. A STOCHASTIC VOLATILITY MODEL- CONDITIONING- We consider a financial asset such that the price satisfies the SDE:

$$dS_t = S_t(rdt + \sigma_t dW_t)$$

where the process $(\sigma_t)_{t\geq 0}$ is assumed continuous, random and independent from $(W_t)_{t\geq 0}$. Then, S_t writes

$$S_t = x\exp\left(rt - \int_0^t \frac{\sigma_s^2}{2}\,ds + \int_0^t \sigma_s dW_s\right)$$

We want to approximate $\mathbb{E}(\mathrm{e}^{-rT}(S_T-K)_+)$, the price of an europeran call with strike K.

1. Show that conditionally on all the trajectory (σ_t), $\int_0^t \sigma_s dW_s$ follows a gaussian distribution with a variance to be defined.
2. Deduce that $\mathbb{E}\left[\mathrm{e}^{-rT}(S_T-K)_+ \middle| (\sigma_t)_{0\leq t\leq T}\right]$ can be explicitely written as a Black and Scholes formula.
3. Explain the estimation of a call by a Monte Carlo method with conditioning.

6.5 Stratified sampling

Let $Y \sim f$ be a r.v. with valueds in $\mathbb{R}^d$. We wish to approximate $\mathbb{E}[h(Y)]$. We note that

$$\mathbb{E}[h(Y)] = \int h(x)f(x)\mathrm{d}x = \sum_{i=1}^{p} \underbrace{\left(\int_{S_i} f(u)\mathrm{d}u\right)}_{\alpha_i} \int h(x) \underbrace{\left[\frac{f(x)1_{S_i}(x)}{\int_{S_i} f(u)\mathrm{d}u}\right]}_{\text{densité de } Y|_{Y\in S_i}} \mathrm{d}x$$

where $(S_i)_{1\leq i\leq p}$ are "regions" or "strata" of $\mathbb{R}^d$; in mathematical terms, the $(S_i)_{1\leq i\leq p}$ form a partition of $\mathbb{R}^d$. If the values of $\boxed{(\alpha_i)_{i=1,\dots,d} \text{ are known}}$ and if we can draw from the distribution of Y conditionally on $\{Y\in S_i\}$ then, we can set:

$$\bar{S}_n = \sum_{i=1}^{p} \alpha_i \left[\frac{\sum_{j=1}^{n_i} h(Y_{i,j})}{n_i}\right],$$

with the conditions

i) $\sum_{i=1}^{p} n_i = n$

ii) $Y_{i,j} \overset{\mathscr{L}}{=} Y|_{Y\in S_i}$ and $(Y_{i,j})$ are independent

Moreover, set $S_n = \frac{\sum_{i=1}^{n} h(X_i)}{n}$ with $X_i \overset{i.i.d.}{\sim} f$. Using the notation

$$\sigma_i^2 = \mathbb{V}\mathrm{ar}(h(Y)|Y\in S_i), \quad \mu_i = \mathbb{E}[h(Y)|Y\in S_i]$$

we get

$$\mathbb{V}\mathrm{ar}(\bar{S}_n) = \sum_{i=1}^{p} \alpha_i^2 \frac{\sigma_i^2}{n_i}, \quad \mathbb{V}\mathrm{ar}(S_n) = \frac{1}{n}\left[\underbrace{\mathbb{E}[h^2(Y)]}_{\sum_{i=1}^{p}\alpha_i(\sigma_i^2+\mu_i^2)} - \left(\underbrace{\mathbb{E}[h(Y)]}_{\sum_{i=1}^{p}\alpha_i\mu_i}\right)^2\right] \tag{6.2}$$

We now consider two subproblems:

i) For a given number of simulations n, what value should we take for n_i, i.e. how many samples n_i should we use in each region S_i?

ii) For given choice of the allocation numbers (n_i), is the resulting variance really lower than a usual Monte Carlo method?

6.5.0.1 Proportional allocation

Chooose $\boxed{\frac{n_i}{n} = \alpha_i}$. Then, (6.2) writes

$$\mathbb{V}\mathrm{ar}(\bar{S}_n) = \frac{1}{n}\sum_{i=1}^{p} \alpha_i \sigma_i^2$$

$$\mathbb{V}\mathrm{ar}(S_n) = \frac{1}{n}\sum_{i=1}^{p} \alpha_i(\sigma_i^2+\mu_i^2) - \left(\sum_{i=1}^{p}\alpha_i\mu_i\right)^2 = \mathbb{V}\mathrm{ar}(\bar{S}_n) + \frac{1}{n}\left(\sum_{i=1}^{p}\alpha_i\mu_i^2 - \left(\sum_{i=1}^{p}\alpha_i\mu_i\right)^2\right) \geq \mathbb{V}\mathrm{ar}(\bar{S}_n)$$

where the last equality follows from $\sum_{i=1}^{p}\alpha_i\mu_i^2 - \left(\sum_{i=1}^{p}\alpha_i\mu_i\right)^2 \geq 0$ since (α_i) defines a probability distribution on $\{1,\dots,p\}$. The stratified sampling thus induces a lower variance in the case of proportional allocation.

6.5.0.2 Optimal allocation

We now aim at finding the optimal $(n_i^\star)$ such that (see (6.2)),

$$\sum_{i=1}^{p}\alpha_i^2\frac{\sigma_i^2}{n_i^\star} = \inf\left\{\sum_{i=1}^{p}\alpha_i^2\frac{\sigma_i^2}{n_i}; \sum_{i=1}^{p} n_i = n\right\}$$

Set $\boxed{\frac{n_i^\star}{n} = \frac{\alpha_i\sigma_i}{\sum_{j=1}^{p}\alpha_j\sigma_j}}$ and let us check that this choice is optimal. Indeed by Cauchy-Schwarz's inequality,

$$\sum_{i=1}^{p}\alpha_i^2\frac{\sigma_i^2}{n_i^\star} = \frac{1}{n}\left(\sum_{i=1}^{p}\alpha_i\sigma_i\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{p}\sqrt{n_i}\frac{\alpha_i\sigma_i}{\sqrt{n_i}}\right)^2 \leq \frac{1}{n}\left(\sum_{i=1}^{p}n_i\right)\left(\sum_{i=1}^{p}\left(\frac{\alpha_i\sigma_i}{\sqrt{n_i}}\right)^2\right) = \sum_{i=1}^{p}\alpha_i^2\frac{\sigma_i^2}{n_i}$$

which indeed shows the optimality of the $n_i^\star$. That being said, the proportional allocation has the advantage (when compared to the optimal allocation) to be independent of the function h whereas the $n_i^\star$ are defined in terms of σ_i, which in turn, depend on h.

Remark 6.12 *The meticulous reader can note that the proof of the optimal allocation shares some strong similarities with the one of the optimal variance in the Importance Sampling techniques (see Proposition 4.10). Let us gather the two optimisation problem in the following table (by setting $\beta_i = n_i/n$):*

ALLOCATION OPTIMALE-	$\inf\left\{\sum_{i=1}^{p}\frac{\alpha_i^2\sigma_i^2}{\beta_i}\ ;\quad \beta_i\geq 0,\ \sum_{i=1}^{p}\beta_i = 1\right\}$
VARIANCE OPTIMALE POUR L'IS-	$\inf\left\{\int\frac{f^2(x)h^2(x)}{g(x)}\mathrm{d}x\ ;\quad g(x)\geq 0,\ \int g(x)\mathrm{d}x = 1\right\}$

And we can see that the two optimisation problems are the same, one being the continuous version of the other one.

6.6 Quasi Monte Carlo methods

6.6.1 Weak discrepancy sequences

We are looking for (x_i) with values in $[0,1]$ such that

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \to_{n\to\infty} \mathbb{E}(h(U))$$

where $U \sim \mathscr{U}[0,1]$ and such that the convergence is quicker than the one given by the CLT. The sequence (x_i) being deterministic, we say that the approximation by $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ is of quasi-Monte Carlo type.

Definition 6.13. "SUITES EQUIRÉPARTIES"- A sequence $(x_i)_i$ with values in $[0,1]^d$ is **"equirépartie"** on $[0,1]^d$ if it satisfies one of the three equivalent properties :

i) For all bounded continuous functions h on $[0,1]^d$, $n^{-1}\sum_{i=1}^{n} h(x_i) \to \int_{[0,1]^d} h(u)\mathrm{d}u$

ii) $\forall y = (y^1,\dots,y^d) \in [0,1]^d$, $n^{-1}\sum_{i=1}^{n} 1_{[0,y^1]\times\ldots\times[0,y^d]}(x_i) \to \text{Volume}([0,y^1]\times\ldots\times[0,y^d]) = \prod_{j=1}^{d} y^j$.

iii) Defining the *discrepancy at the origin*, $\mathscr{D}_n^*$, of the sequence (x_i) by

$$\mathscr{D}_n^* = \sup_{y=(y^1,\dots,y^d)\in[0,1]^d} \left| n^{-1}\sum_{i=1}^{n} 1_{[0,y^1]\times\ldots\times[0,y^d]}(x_i) - \text{Volume}([0,y^1]\times\ldots\times[0,y^d]) \right|.$$

we have $\mathscr{D}_n^* \to 0$.

Remark 6.14 *The sequences are obtained as the output of r.v. with uniform distribution on $[0,1]^d$ are "equirépartie". They indeed satisfy the first condition due to the LLN.*

There exist also other "equirépartie" sequences (x_n) such that for a certain class of functions h, we have the following result

$$\left| \frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbb{E}(h(U)) \right| \leq c(h)\frac{(\ln n)^d}{n} \tag{6.3}$$

Such sequences are called *sequences with weak discrepancy*. The speed of convergence is therefore of order $(\ln(n))^d/n$ which is much better than the ones obtained by standard Monte Carlo methods (where the speed, given by the CLT, is of order $1/\sqrt{n}$.

6.6.1.1 The Van der Corput sequence

We first give an example in dimension 1. Let p be a prime number. Each number n can be written in basis p as a sequence $(a_{i,n})_{i\geq 0}$ of integers in $\{0,\dots,p-1\}$, eventually equal to 0 such that

$$n = \sum_{i=0}^{\infty} a_{i,n} p^i \quad \text{où} \quad a_{i,n} \in \{0,\dots,p-1\}$$

Set $\boxed{x_n^p = \sum_{i=0}^{\infty} \frac{a_{i,n}}{p^{i+1}}}$. This sequence $(x_n^p)_{n\geq 0}$ is known as a Van Der Corput sequence, it is "equirépartie" on $[0,1]$ with weak discrepancy.
The support of the Van Der Corput sequence associated to the number p is the set of the extreme values of intervals obtained by splitting the interval $[0,1]$ into p intervals of the same length and by repeating the procedure on each subinterval.

6.6.1.2 Halton Sequence

It is a generalization of the Van der Corput sequence in dimension d. We consider $\mathbb{R}^d$ and let $(p_1,...,p_d)$ be the d first prime numbers, then the sequence of vectors $\boxed{(x_n^{p_1},\dots,x_n^{p_d})}$ (where x_n^p is the n-th term in the Van der Corput sequence associated to the prime number p) is an "equirépartie" sequence on $[0,1]^d$ with weak discrepancy.

There also exists other weak discrepancy sequences as the Faure sequence or the Sobol sequences, whose description and properties can be easily found in the literature.

Algorithm 5 Calcul de x_n^p

Input: $n,\ p$.
$power = p$.
$vdc = (n \bmod p)/power$.
$n = \text{floor}\ (n/p)$.
while $n > 0$ **do** $power = power * p$
    $vdc = vdc + (n \bmod p)/power$
    $n = \text{floor}\ (n/p)$.
end while
Output: vdc.

Remark 6.15 *We stress that these sequences are not random, it is therefore not possible to consider confidence intervals using the CLT, which does not mean anything for a deterministic sequence. Nevertheless, the inequality* (6.3) *gives an upper-bound for the error on the condition that we know exactly* $c(h)$*, which is not the case in general.*

6.7 Take-home message

a) Be able to distinguish all the means for reducing the variance: importance sampling, antithetic variates, control variates, conditioning, stratified sampling.
b) For each method, be able to show the convergence properties and the asymptotic normality.
c) Optimality for the importance sampling, for the control variates and the stratified sampling.
d) Terminology of the quasi Monte Carlo sequences. Definition of the Van der Corput sequences and of the Halton sequences.

6.8 Highlights

6.8.0.1 Van der Corput. source: Wikipedia

Johannes Gualtherus van der Corput (Rotterdam, September 4, 1890 - Amsterdam, September 16, 1975) was a Dutch mathematician, working in the field of analytic number theory.

He was appointed professor at the University of Groningen in 1923, and at the University of Amsterdam in 1946. He was one of the founders of the Mathematisch Centrum in Amsterdam, of which he also was the first director. From 1953 on he worked in the United States at the University of California, Berkeley and the University of Wisconsin-Madison. Among his students were J. F. Koksma and J. Popken.

Chapter 7

Exercises

Exercise 7.1. Let $p \in \mathbb{R}_*^+$.

1. Let $\Omega = [0,1]$, $\mathscr{F} = \mathscr{B}(\Omega)$ and $\mathbb{P} = \lambda|_\Omega$ the Lebesgue measure restricted to $[0,1]$. Define $X : \omega \mapsto X(\omega) = \mathbb{1}_{[0,p]}(\omega)$. What is the law of X?
2. Let $\Omega = \{0,1\}$, $\mathscr{F} = \mathscr{P}(\Omega)$ and $\mathbb{P}$ such that $\mathbb{P}(\{1\}) = 1 - \mathbb{P}(\{0\}) = p$. Let $Y : \omega \mapsto Y(\omega) = \omega$. Show that the law of Y is the same as the one of X.

Exercise 7.2. Show that the Borel sigma field $\mathscr{B}(\mathbb{R})$ is generated by

(i) the closed sets
(ii) $\{]a,b] : a,b \in \mathbb{R}\}$
(iii) $\{]-\infty,b] : a,b \in \mathbb{R}\}$

Exercise 7.3. Show that any continuous function $f : \mathbb{R} \to \mathbb{R}$ is $\mathscr{B}(\mathbb{R})/\mathscr{B}(\mathbb{R})$-measurable.

Exercise 7.4. Let $\tau = \{A \in \mathscr{B}(\mathbb{R}) : A = -A\}$ where $-A = \{-x : x \in A\}$.

1. Show that $A \in \tau$ if and only if we have $(x \in A) \Rightarrow (-x \in A)$.
2. Show that τ is a sigma field on $\mathbb{R}$.
3. Consider the mappings $f : x \mapsto \mathrm{e}^x$, $g : x \mapsto x^3$ and $h : x \mapsto \cos(x)$.
   (a) Are they $\mathscr{B}(\mathbb{R})/\tau$-measurable?
   (b) Are they $\tau/\mathscr{B}(\mathbb{R})$-measurable?
   (c) Are they τ/τ-measurables?
4. Find all the functions that are $\tau/\mathscr{B}(\mathbb{R})$-measurable.
5. Find all the functions that are τ/τ-measurables.

Exercise 7.5. Set for any $x \geq 1$: $f(x) = \sum_{n=1}^{\infty} n\mathrm{e}^{-nx}$, what is the value of $\int_1^\infty f(x)\mathrm{d}x$?

Exercise 7.6. Calculate the limit as n goes to the infinity of the quantity $\int_0^\infty \arctan(\mathrm{n}x)\mathrm{e}^{-x^n}\mathrm{d}x$.

Index