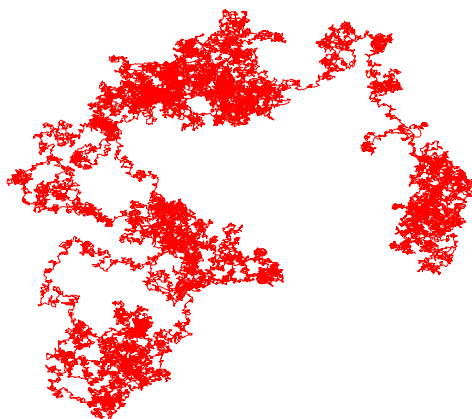




Master de Mathématiques et Applications

Spécialité *Mathématiques de la Modélisation*

Méthodes Numériques Probabilistes



Julien Reygner

Version du 24 novembre 2025

Contents

I	Lectures and exercises	3
1	Random variable simulation	5
1.1	Random variable simulation	5
1.2	Random vector simulation	14
1.3	Complements	19
2	The Monte Carlo method	23
2.1	The Monte Carlo method	23
2.2	Variance reduction	26
2.3	Complements	29
3	Markov chains and ergodic theorems	35
4	Convergence to equilibrium of Markov chains	37
5	The Markov Chain Monte Carlo Method	39
6	Stochastic processes, Brownian motion and Itô calculus	41
7	Stochastic differential equations	43
8	Long time behavior of diffusion processes	45
II	Topics in stochastic simulation algorithms	47
1	Splitting algorithm for rare event estimation	49
2	Asymptotic efficiency of importance sampling through large deviation theory	51
	Bibliography	53

Foreword

The central focus of this course is the Monte Carlo method. We first present its basic principle in the context of independent and identically distributed samples, which requires reviewing methods for simulating random variables, as well as limit theorems that allow us to quantify the accuracy of this approach. In particular, several variance reduction techniques are introduced.

The goal of the second part of the course is to introduce the Markov Chain Monte Carlo method. We present the notion of discrete-time Markov chains, followed by the main results concerning long-time behaviour. We then describe several stochastic sampling algorithms (Metropolis–Hastings and Gibbs).

The third part of the course is devoted to the connections between diffusion processes and partial differential equations. After a brief review of stochastic calculus, we present in detail the Feynman–Kac formula and discretisation methods for stochastic differential equations, which make it possible to implement the Monte Carlo method to solve parabolic or elliptic partial differential equations. The link with the Markov Chain Monte Carlo method is finally established through the study of the long-time behavior of diffusion processes.

The reader is assumed to be familiar with basic measure theory, including Lebesgue integration, and basic probability theory: random variables and limit theorems. We refer to [2] or [3] for comprehensive textbooks in this direction.

★ ★ ★

These lecture notes contain two parts: Part **I** contains the contents covered during the lectures, complemented with exercises and further developments, while Part **II** contains problems which are adapted from past exams.

In Part **I**, each Lecture should roughly correspond to one session of the course. Some of the exercises will be discussed during the lectures, but definitely not all, so you should have enough material to practice between the sessions. Some exercises contain a numerical part to implement by yourself. There is no written correction for these exercises.

This document will be regularly updated during the trimester, on the course’s webpage¹. Some parts will not be seen in class, they will be marked with a star *, and they will not be examinable.

For any questions or comment, please send me an email at julien.reygner@enpc.fr.

¹<https://cermics.enpc.fr/~reygnerj/anedp.html>

Part I

Lectures and exercises

Lecture 1

Random variable simulation

The goal of this Lecture is to present algorithms for random variable generation: given a probability distribution P on some measurable space $(E, \mathcal{E})$, how to generate independent and identically distributed variables $X_1, \dots, X_n$ with law P ? This will be used already in Lecture 2 to compute integrals with the *Monte Carlo method*.

Throughout this Lecture, we work on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and, for any $p \in [1, +\infty)$, we denote by $\mathbf{L}^p(\mathbb{P})$ the space of real-valued random variables X such that $\mathbb{E}[|X|^p] < +\infty$.

1.1 Random variable simulation

1.1.1 Uniform distribution and basic applications

Uniform random variables on $[0, 1]$

It is an obvious fact that a *deterministic* algorithm cannot generate a *truly random* sequence, as was written by von Neumann: ‘Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.’¹. Hence, *pseudo-random number generators* are deterministic algorithms which, starting from a *seed* x_0 , return a sequence $x_1, x_2, \dots$ of numbers which exhibits the same statistical properties as a sequence of *independent and identically distributed* random numbers.

Because of the finiteness of the memory of a computer, a pseudo-random number generator is necessarily *ultimately periodic*, that is to say that there exists $t \geq 0$, which may depend on x_0 , such that for n large enough, $x_{n+t} = x_n$. In the sequel we call *maximal period* the largest value of t over all possible values of x_0 . Since ‘truly random’ sequences should not be periodic, it is an intuitive statement that a ‘good’ pseudo-random number generator should have a large maximal period.

We first present a class of pseudo-random generators which are relatively easy to describe. *Linear congruential generators* were introduced in 1948 and depend on the following integer parameters:

- a *modulus* $m > 0$;
- a *multiplier* $0 < a < m$;
- an *increment* $0 \leq c < m$.

¹Quoted in D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd edition, Addison-Wesley, 1998.

The seed is an integer $x_0 \in \{0, \dots, m-1\}$. The sequence $(x_n)_{n \geq 1}$ is then computed according to the recurrence relation

$$x_{n+1} = ax_n + c \pmod{m},$$

which produces integer numbers in $\{0, \dots, m-1\}$. Typically, taking $m = 2^{32}$ allows to get integers encoded on 32 bits.

In general, the maximal period of linear congruential generators (which is at most m) can be computed. Yet, their quality remains very sensitive to the choice of a and m . More complex pseudo-random generators have thus been elaborated. The most widely used generator in current scientific computing languages is called *Mersenne Twister*. It was developed in 1997², it is based on the arithmetic properties of Mersenne numbers and its period is $2^{19937} - 1 \simeq 4.3 \cdot 10^{6001}$.

Whatever the chosen pseudo-random number generator, let us take as granted that given a seed $x_0 \in \{0, \dots, m-1\}$, it returns a sequence $(x_n)_{n \geq 1}$ of numbers in $\{0, \dots, m-1\}$, which has the following statistical properties:

- (i) they look independent;
- (ii) they look uniformly distributed in $\{0, \dots, m-1\}$ in the sense that each integer $x \in \{0, \dots, m-1\}$ appears in the sequence $(x_n)_{n \geq 1}$ with equal frequency $1/m$.

Defining $U_n = x_n/m \in [0, 1)$, we thus obtain a sequence of pseudo-random independent variables such that, for any $n \geq 1$, for any interval $C \subset [0, 1]$,

$$\mathbb{P}(U_n \in C) = \frac{1}{m} \sum_{x=0}^{m-1} \mathbb{1}_{\{x/m \in C\}} \simeq \int_{u=0}^1 \mathbb{1}_{\{u \in C\}} du.$$

This motivates the following definition.

Definition 1.1.1 (Uniform distribution). *A random variable U in $[0, 1]$ is called uniformly distributed on $[0, 1]$ if it has the density*

$$p(u) = \mathbb{1}_{\{u \in [0, 1]\}}.$$

We denote $U \sim \mathcal{U}[0, 1]$.

Exercise 1.1.2. *Let $U \sim \mathcal{U}[0, 1]$. Show that the random variable $1 - U$ has the same distribution as U .*

From now on, we shall thus work under the assumption that our computer is able to generate independent variables $(U_n)_{n \geq 1}$ which are uniformly distributed on $[0, 1]$. In the sequel of this section, we study how to use this sequence in order to sample a random variable X with a given distribution.

Remark 1.1.3. *Most scientific computing languages allow you to fix the seed of your pseudo-random number generator. This makes your code no longer random but this may prove very helpful for reproducibility, comparison of your code and experimental results with others, or simply debugging.*

²Matsumoto, M. and Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Transactions on Modeling and Computer Simulations* (1998).

Elementary discrete distributions

We first introduce several discrete distributions.

Definition 1.1.4 (Bernoulli, binomial and geometric distributions). *Let $p \in [0, 1]$.*

- (i) *A random variable X in $\{0, 1\}$ such that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ is called a Bernoulli random variable with parameter p . We denote $X \sim \mathcal{B}(p)$.*
- (ii) *Let $n \geq 1$ and $X_1, \dots, X_n$ be independent Bernoulli random variables with parameter p . The random variable $S := X_1 + \dots + X_n$ is called a binomial random variable with parameters n and p . We denote $S \sim \mathcal{B}(n, p)$.*
- (iii) *Assume that $p \in (0, 1]$ and let $(X_i)_{i \geq 1}$ be a sequence of independent Bernoulli random variables with parameter p . The random variable $T := \min\{i \geq 1 : X_i = 1\}$ is called a geometric random variable with parameter p . We denote $T \sim \text{Geo}(p)$.*

The numerical sampling of the Bernoulli, binomial and geometric distributions is addressed in the next exercise.

Exercise 1.1.5. *Let $(U_n)_{n \geq 1}$ be a sequence of independent uniform variables on $[0, 1]$.*

1. *Using an if test, how to draw a random variable $X \sim \mathcal{B}(p)$?*
2. *Using a for loop, how to draw a random variable $S \sim \mathcal{B}(n, p)$?*
3. *Using a while loop, how to draw a random variable $T \sim \text{Geo}(p)$?*

1.1.2 The inverse CDF method

Discrete distributions

Let X be a random variable taking its values in some finite set E with cardinality m , and let $(p_x)_{x \in E}$ be its probability mass function (that is to say, $p_x = \mathbb{P}(X = x)$). An intuitive algorithm allowing to sample X from a uniform random variable $U \in [0, 1]$ is the following:

1. label the elements of E in some arbitrary order $x_1, \dots, x_m$;
2. select the unique index $i \in \{1, \dots, m\}$ such that $p_{x_1} + \dots + p_{x_{i-1}} < U \leq p_{x_1} + \dots + p_{x_i}$;
3. return $X = x_i$.

It is clear that we have

$$\mathbb{P}(X = x_i) = \mathbb{P}(p_{x_1} + \dots + p_{x_{i-1}} < U \leq p_{x_1} + \dots + p_{x_i}) = p_{x_i},$$

so that X has the correct law.

CDF and inverse CDF

The generalisation of this approach to arbitrary, real-valued random variables, is based on the introduction of the *Cumulative Distribution Function* of such variables.

Definition 1.1.6 (Cumulative Distribution Function). *Let X be a real-valued random variable. The Cumulative Distribution Function (CDF) of X is the function $F_X : \mathbb{R} \rightarrow [0, 1]$ defined by*

$$\forall x \in \mathbb{R}, \quad F_X(x) := \mathbb{P}(X \leq x).$$

Remark 1.1.7. Since the Borel σ -field on $\mathbb{R}$ is generated by the intervals of the form $(-\infty, x]$, by Dynkin's Lemma, two random variables have the same CDF if and only if they have the same law.

Exercise 1.1.8 (Properties of CDFs). Let F_X be the CDF of a random variable X . Show that:

1. F_X is nondecreasing;
2. $\lim_{x \rightarrow -\infty} F_X(x) = 0$, $\lim_{x \rightarrow +\infty} F_X(x) = 1$;
3. F_X is right continuous and has left limits.

When X has a density p , Definition 1.1.6 yields the identity

$$\forall x \in \mathbb{R}, \quad F_X(x) = \int_{y=-\infty}^x p(y)dy,$$

which shows that F_X is continuous and dx -almost everywhere differentiable, with $F'_X = p$.

Definition 1.1.9. Let F_X be the CDF of a random variable X . The pseudo-inverse of F_X is the function $F_X^{-1} : [0, 1] \rightarrow [-\infty, +\infty]$ defined by

$$\forall u \in [0, 1], \quad F_X^{-1}(u) := \inf\{x \in \mathbb{R} : F_X(x) \geq u\},$$

with the conventions that $\inf \mathbb{R} = -\infty$ and $\inf \emptyset = +\infty$.

The pseudo-inverse of a CDF is nondecreasing, left continuous with right limits. When F_X is continuous and increasing, then F_X^{-1} is the usual inverse bijection of F_X . In general, it need not hold that $F_X(F_X^{-1}(u)) = u$ or $F_X^{-1}(F_X(x)) = x$, but the following weaker statement remains true.

Lemma 1.1.10 (CDF and pseudo-inverse). Let F_X be the CDF of a random variable X . For all $x \in \mathbb{R}$, $u \in (0, 1)$, we have $F_X^{-1}(u) \leq x$ if and only if $u \leq F_X(x)$.

Proof. Since F_X is right continuous, for any $u \in (0, 1)$ the set $\{x \in \mathbb{R} : F_X(x) \geq u\}$ is closed, therefore $F_X(F_X^{-1}(u)) \geq u$. Since F_X is nondecreasing, we deduce that if $F_X^{-1}(u) \leq x$ then $u \leq F_X(x)$. Conversely, if $u \leq F_X(x)$, then by the definition of F_X^{-1} , $F_X^{-1}(u) \leq x$. $\square$

Corollary 1.1.11 (The inverse CDF method). Let F_X be the CDF of a random variable X , and let $U \sim \mathcal{U}[0, 1]$. The random variables X and $F_X^{-1}(U)$ have the same distribution.

Proof. By Lemma 1.1.10 and Definition 1.1.1, for all $x \in \mathbb{R}$,

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = \int_{u=0}^{F_X(x)} du = F_X(x),$$

so that the random variables X and $F_X^{-1}(U)$ have the same CDF. From Remark 1.1.7 we conclude that they have the same distribution. $\square$

We illustrate this method on the exponential distribution.

Definition 1.1.12 (Exponential distribution). Let $\lambda > 0$. A random variable X in $[0, +\infty)$ is called exponential with parameter λ if it has the density

$$p(x) = \mathbb{1}_{\{x>0\}} \lambda e^{-\lambda x}.$$

We denote $X \sim \mathcal{E}(\lambda)$.

An immediate computation shows that the CDF of X writes

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\lambda x} & \text{otherwise.} \end{cases}$$

As a consequence, for all $u \in [0, 1]$,

$$F_X^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u),$$

with the obvious convention that $\ln 0 = -\infty$. Therefore, to draw a random variable $X \sim \mathcal{E}(\lambda)$, one may take a uniform variable U on $[0, 1]$ and return $-\frac{1}{\lambda} \ln(1 - U)$. Notice that, by Exercise 1.1.2, it is also equivalent to return $-\frac{1}{\lambda} \ln(U)$.

1.1.3 The Box–Muller method for Gaussian random variables

We recall that the *Gauss integral* is equal to³

$$\int_{x \in \mathbb{R}} \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi}.$$

Definition 1.1.13 (Standard Gaussian variables). *A random variable G in $\mathbb{R}$ is a standard Gaussian variable if it has the density*

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Exercise 1.1.14. *If G is a standard Gaussian variable, show that $G \in \mathbf{L}^p(\mathbb{P})$ for any $p \in [1, +\infty)$ and compute $\mathbb{E}[G]$ and $\text{Var}(G)$.*

It follows from this exercise that for any $\mu, \sigma \in \mathbb{R}$, the random variable $X = \mu + \sigma G$ satisfies $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$. This remark is used in the next definition.

Definition 1.1.15 (Gaussian variable). *If G is a standard Gaussian variable, then for any $\mu, \sigma \in \mathbb{R}$, the random variable*

$$X = \mu + \sigma G$$

is called a Gaussian random variable with mean μ and variance σ^2 . Its law is denoted by $\mathcal{N}(\mu, \sigma^2)$.

Gaussian variables are also called *normal*. The fact that the law of X only depends on σ through σ^2 is justified by the following result.

Exercise 1.1.16. *Show that if $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$, then X has density*

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

We insist on the fact that the definition of Gaussian random variables also includes the case where $\sigma = 0$, in which case X is the almost surely constant random variable equal to μ . In this case, the law of X is the Dirac measure δ_μ and therefore it does not have a density.

By definition, the problem of sampling from Gaussian distributions reduces to the case of the standard Gaussian distribution. Let $\Phi : \mathbb{R} \rightarrow [0, 1]$ denote its CDF, given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^x \exp\left(-\frac{y^2}{2}\right) dy.$$

³Do not hesitate to redo the computation just to be sure that you still know how to!

It is known that Φ cannot be expressed in terms of usual functions, such as polynomials, exponentials or logarithms. Hence the inverse CDF method cannot be applied in the present case. We shall present an *ad hoc* approach, called the *Box–Muller method*⁴.

Proposition 1.1.17 (Box–Muller method). *Let $R \sim \mathcal{E}(1/2)$ and $\Theta \sim \mathcal{U}[0, 2\pi]$ be independent random variables. The random variables*

$$X := \sqrt{R} \cos \Theta, \quad Y := \sqrt{R} \sin \Theta,$$

are independent and follow the standard Gaussian distribution.

Proof. We use the dummy function method and let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be measurable and bounded. Since R and Θ are independent, the law of the pair (R, Θ) is the product of the marginal densities, and therefore

$$\begin{aligned} \mathbb{E}[f(X, Y)] &= \mathbb{E}\left[f\left(\sqrt{R} \cos \Theta, \sqrt{R} \sin \Theta\right)\right] \\ &= \int_{r=0}^{+\infty} \int_{\theta=0}^{2\pi} f(\sqrt{r} \cos \theta, \sqrt{r} \sin \theta) \frac{d\theta}{2\pi} \frac{1}{2} e^{-r/2} dr. \end{aligned}$$

Using the polar change of coordinates $x = \sqrt{r} \cos \theta$, $y = \sqrt{r} \sin \theta$ in the right-hand side, we get

$$\mathbb{E}[f(X, Y)] = \int_{x, y \in \mathbb{R}} f(x, y) \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) dx dy,$$

which shows that the pair (X, Y) has density

$$\frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right),$$

which implies that X and Y are independent standard Gaussian variables. $\square$

Since both R and Θ can be sampled using the inverse CDF method, Proposition 1.1.17 provides a method to sample X and Y from two independent uniform random variables on $[0, 1]$.

Remark 1.1.18 (What does my computer really do?⁵). *The Box–Muller method is used by NumPy's `random.standard_normal` function to generate Gaussian variables. Its newer random number generator class, called `Generator`, uses another method called the Ziggurat algorithm, which is presented in the next Section. In contrast, the statistical software *R* uses the inverse CDF method to generate Gaussian samples, with a numerical approximation of the function Φ^{-1} .*

1.1.4 Rejection sampling

Sampling from a conditional probability

Let Q be a probability distribution on some measurable space $(E, \mathcal{E})$ and $D \in \mathcal{E}$ such that $Q(D) > 0$. The conditional probability $Q(\cdot|D)$ is the probability measure on D defined by

$$Q(C|D) = \frac{Q(C)}{Q(D)}$$

⁴Box, G. E. P. and Muller, M. E. A Note on the Generation of Random Normal Deviates, *The Annals of Mathematical Statistics* (1958).

⁵According to the blog post <https://medium.com/mti-technology/how-to-generate-gaussian-samples-3951f2203ab0>.

for any measurable subset $C \subset D$; in other words, it is the measure with density $\frac{\mathbb{1}_{\{x \in D\}}}{Q(D)}$ with respect to Q .

Assume that you are able to draw iid samples $X_1, \dots, X_n$ from Q , and that you want to draw a random variable X with distribution $Q(\cdot|D)$. A somewhat obvious algorithm is then to draw samples from Q and to only keep those which fall into D . In other words, set $N := \inf\{n \geq 1 : X_n \in D\}$ and return $X := X_N$. The next statement shows that this algorithm is correct.

Proposition 1.1.19 (Rejection sampling from a conditional probability). *With the notation introduced above:*

- (i) $N \sim \mathcal{G}_{\text{eo}}(Q(D))$;
- (ii) X_N has distribution $Q(\cdot|D)$;
- (iii) N and X_N are independent.

Proof. We fix $n \geq 1$, a measurable subset C of D , and compute

$$\begin{aligned} \mathbb{P}(N = n, X_N \in C) &= \mathbb{P}(X_1 \notin D, X_2 \notin D, \dots, X_{n-1} \notin D, X_n \in C) \\ &= (1 - Q(D))^{n-1} Q(C). \end{aligned}$$

Taking $C = D$ shows the first point of the Proposition. Summing over n yields the second point, and allows to deduce that the right-hand side above rewrites $\mathbb{P}(N = n)\mathbb{P}(X_N \in C)$, which leads to the third point. $\square$

Generalisation

The rejection algorithm exposed above can be cleverly employed to sample from probability distributions which are not directly conditional probabilities as given by Proposition 1.1.19.

Lemma 1.1.20 (Uniform density on the graph of p). *Let $p : \mathbb{R}^d \rightarrow [0, +\infty)$ be a probability density. Let*

$$D := \{(x, y) \in \mathbb{R}^d \times [0, +\infty) : 0 \leq y \leq p(x)\}$$

be the graph of p .

- (i) D has $(d + 1)$ -dimensional Lebesgue measure 1.
- (ii) If (X, Y) is distributed according to the uniform density on D , then X has density p .

Proof. The $(d + 1)$ -dimensional Lebesgue measure of D is

$$|D| = \int_{x \in \mathbb{R}^d} \int_{y=0}^{+\infty} \mathbb{1}_{\{y \leq p(x)\}} dy dx = \int_{x \in \mathbb{R}^d} p(x) dx = 1.$$

If (X, Y) is uniformly distributed in D , the marginal density of X is

$$\int_{y=0}^{+\infty} \mathbb{1}_{\{y \leq p(x)\}} dy = p(x). \quad \square$$

As a consequence of Lemma 1.1.20, to draw $X \in \mathbb{R}^d$ with density p , it suffices to draw (X, Y) uniformly in D . This is where the method described in the previous Subsection becomes relevant: if one is able to find $R \subset \mathbb{R}^{d+1}$ such that $D \subset R$ and one may draw uniform samples $(X_n, Y_n)_{n \geq 1}$ in R , then letting $N := \inf\{n \geq 1 : (X_n, Y_n) \in R\}$ one deduces that X_N has density p . In this

perspective, let us assume that there exists $f : \mathbb{R}^d \rightarrow [0, +\infty)$ such that $p(x) \leq f(x)$ for all $x \in \mathbb{R}$, so that

$$D \subset R := \{(x, y) \in \mathbb{R}^d \times [0, +\infty) : 0 \leq y \leq f(x)\},$$

and that:

- (i) $|R| = \int_{x \in \mathbb{R}^d} f(x) dx < +\infty$;
- (ii) one is able to draw random variables with density $q(x) = f(x)/|R|$.

The next Lemma shows how to draw samples $(X_n, Y_n)_{n \geq 1}$ which are uniformly distributed in R .

Lemma 1.1.21 (Uniform density on the graph of f). *Let $q : \mathbb{R}^d \rightarrow [0, +\infty)$ be the probability density defined by $q(x) = f(x)/|R|$. Let X with density q and $U \sim \mathcal{U}[0, 1]$ be independent from X . Then the pair $(X, Uf(X))$ is uniformly distributed in R .*

Proof. For any measurable and bounded function $g : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}[g(X, Uf(X))] = \int_{x \in \mathbb{R}^d} \int_{u=0}^1 g(x, uf(x)) q(x) dx du.$$

Replacing $q(x)$ with $f(x)/|R|$ and then setting $y = uf(x)$ we get

$$\mathbb{E}[g(X, Uf(X))] = \int_{x \in \mathbb{R}^d} \int_{y=0}^{f(x)} g(x, y) \frac{1}{|R|} dx dy,$$

which shows that $(X, Uf(X))$ has density $\frac{1}{|R|} \mathbb{1}_{\{0 \leq y \leq f(x)\}}$. $\square$

The overall rejection procedure is summarised in the next statement, where we write $k = |R|$.

Theorem 1.1.22 (Rejection sampling). *Let $p : \mathbb{R}^d \rightarrow [0, +\infty)$ be a probability density. Assume that there exist a probability density $q : \mathbb{R}^d \rightarrow [0, +\infty)$ and $k \geq 1$ such that, dx -almost everywhere, $p(x) \leq kq(x)$. Let $(X_n)_{n \geq 1}$ be a sequence of independent random variables in $\mathbb{R}^d$ with density q , and $(U_n)_{n \geq 1}$ be a sequence of independent random variables uniformly distributed in $[0, 1]$, independent from $(X_n)_{n \geq 1}$. Let $N := \inf\{n \geq 1 : kq(X_n)U_n \leq p(X_n)\}$.*

- (i) $N \sim \text{Geo}(1/k)$.
- (ii) X_N has density p .
- (iii) N and X_N are independent.

Proof. Let $f(x) = kq(x)$, and $R = \{(x, y) \in \mathbb{R}^d \times [0, +\infty) : 0 \leq y \leq f(x)\}$. By Lemma 1.1.21, the pairs $(X_n, U_nf(X_n))$ are independent and uniformly distributed on R : moreover, N rewrites as $\inf\{n \geq 1 : (X_n, U_nf(X_n)) \in D\}$, where $D = \{(x, y) \in \mathbb{R}^d \times [0, +\infty) : 0 \leq y \leq p(x)\}$. Therefore, by Proposition 1.1.19, we have: N is a geometric random variable with parameter $|D|/|R| = 1/k$; $(X_N, U_Nf(X_N))$ is uniformly distributed on D — which by Lemma 1.1.20 implies that X_N has density p ; N and $(X_N, U_Nf(X_N))$ are independent. $\square$

Remark 1.1.23. Theorem 1.1.22 can easily be generalised to the case where one wants to draw X from a probability measure P on some abstract space E , and has access to samples under $Q \gg P$, with $\frac{dP}{dQ} \leq k$, Q -almost everywhere. Then the statement of Theorem 1.1.22 remains in force, with N defined as the first index for which $kU_n \leq \frac{dP}{dQ}(X_n)$.

Rejection sampling is useful when one is not able to sample directly from p , but can find q such that $p \leq kq$ and sampling from q is easier. Clearly, the smaller k , the faster the algorithm, therefore from a computational point of view it is of interest to take q as a ‘good approximation’ of p .

The Ziggurat algorithm

Let $p : [0, +\infty) \rightarrow [0, +\infty)$ be a continuous probability density, which is assumed to be nonincreasing on $[0, +\infty)$. The typical example that we have in mind is

$$p(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right). \quad (1.1)$$

Exercise 1.1.24. Let X have density p given by (1.1), and ϵ be independent from X and such that $\mathbb{P}(\epsilon = -1) = \mathbb{P}(\epsilon = 1) = 1/2$. Show that ϵX is a standard Gaussian variable.

The basis of the Ziggurat algorithm⁶ consists in covering the graph of f with a number L of horizontal layers defined as follows.

1. Fix $x_1 > 0$, set $y_1 = p(x_1)$, and define the layer 0 as

$$([0, x_1] \times [0, y_1]) \cup \{(x, y) \in [0, +\infty) \times [0, +\infty) : x > x_1, y \leq p(x)\}.$$

Denote by

$$A := x_1 y_1 + \int_{x=x_1}^{+\infty} p(x) dx$$

the area of the layer 0.

2. On top of the layer 0, add a rectangular layer of width x_1 and height A/x_1 , so it also has area A . The top of this layer is at height $y_2 = y_1 + A/x_1$, and intersects the density function at a point (x_2, y_2) , where $y_2 = p(x_2)$.
3. Further rectangular layers of area A are then stacked on top, until $y_L \leq p(0)$ in which case we set $x_L = 0$.

We obtain a covering of the graph of p with L layers of equal area A , see Figure 1.1. This covering is the graph R of a function $f : [0, +\infty) \rightarrow [0, +\infty)$ which is such that

$$f(x) = \begin{cases} y_k & \text{if } x \in [x_k, x_{k-1}), k \in \{2, \dots, L\}, \\ p(x) & \text{if } x \geq x_1, \end{cases}$$

and

$$|R| = \int_{x=0}^{+\infty} f(x) dx = LA.$$

Following Proposition 1.1.19 and Lemma 1.1.20, the goal is now to draw pairs (X, Y) uniformly distributed on R , and to return X if $Y \leq p(X)$. However, the Ziggurat algorithm uses a different approach from Lemma 1.1.21 to draw such pairs: first, it picks one of the layers uniformly at random; then, it draws (X, Y) uniformly in this layer. Since all layers have the same area, this indeed returns a pair (X, Y) which is uniformly distributed on R . Moreover, if the chosen layer has index $k \in \{1, \dots, L-1\}$, then it is a rectangle, so drawing (X, Y) uniformly in the layer is easy:

- first, draw X uniformly in $[0, x_k]$;

⁶Presented for instance in Marsaglia, G. and Tsang, W. W. The Ziggurat Method for Generating Random Variables, *Journal of Statistical Software* (2000).

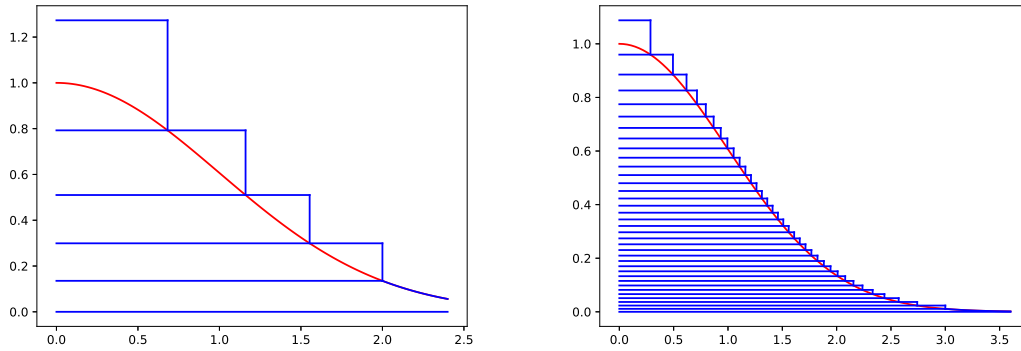


Figure 1.1: Construction of the layers, starting from $x_1 = 2$ on the left-hand picture (then $L = 5$), and from $x_1 = 3$ on the right-hand picture (then $L = 36$).

- if $X \leq x_{k+1}$ then whatever the draw of Y it will always satisfy the condition that $Y \leq p(X)$, so return X ;
- if $X > x_{k+1}$ then draw Y uniformly in $[y_k, y_{k+1}]$ and return X if $Y \leq p(X)$, otherwise restart with a new layer.

If the chosen layer is 0, then the algorithm needs to have a *fallback procedure*, which is able to generate samples from the density $\mathbb{1}_{\{x > x_1\}} p(x) / \int_{x'=x_1}^{+\infty} p(x') dx'$. Then it works as follows:

- set $x_0 = A/y_1 > x_1$ and draw X uniformly on $[0, x_0]$;
- if $X \leq x_1$ then return X ;
- if $X > x_1$ then draw X' according to the fallback procedure and return X' .

Exercise 1.1.25. Check that this algorithm indeed returns a random variable X with density p .

Exercise 1.1.26. Show that, to sample from the density p given by (1.1), a possible fallback procedure is to draw $T_1 \sim \mathcal{E}(x_1)$ and $T_2 \sim \mathcal{E}(1)$ independent, and return $X' = T_1 + x_1$ if $2T_1 > T_2^2$, otherwise restart.

The reason why this algorithm is efficient is that, except for the fallback procedure, it does not require the evaluation of complicated functions, such as sine or cosine, which may be costly. Moreover, if the number of layers is large enough, samples are almost never drawn in the layer 0, and almost often land in the interval $[0, x_{k+1}]$ of the layer k , so they are rarely rejected and drawing one sample only costs the draw of the random index of the layer k , and of the uniform variable $X \in [0, x_k]$.

1.2 Random vector simulation

In this Section, we consider the issue of simulating random vectors, that is to say random variables with values in $\mathbb{R}^d$. For any $p \geq 1$, we denote by $\mathbf{L}^p(\mathbb{P}; \mathbb{R}^d)$ the set of random vectors whose coordinates are random variables in $\mathbf{L}^p(\mathbb{P})$. If $X = (X_1, \dots, X_n) \in \mathbf{L}^1(\mathbb{P}; \mathbb{R}^d)$, we denote by $\mathbb{E}[X]$ the d -dimensional vector $(\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])$. If $X = (X_1, \dots, X_n) \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$, we denote by $\text{Cov}[X]$ the $d \times d$ matrix with coefficients $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$.

$\mathbb{E}[X_j]) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$. It is called the *covariance matrix* of X , it is symmetric, and by Exercise 1.2.1 below, it is nonnegative.

Exercise 1.2.1. Let $X \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$ with covariance matrix K .

1. Show that, for any $u \in \mathbb{R}^d$, $\text{Var}(\langle u, X \rangle) = \langle u, Ku \rangle$.
2. Show that, for any $b \in \mathbb{R}^k$, $A \in \mathbb{R}^{k \times d}$, $\text{Cov}[b + AX] = AK A^\top$.

1.2.1 Gaussian vectors

We recall that the *characteristic function* $\Psi_X : \mathbb{R}^d \rightarrow \mathbb{C}$ of a random vector $X \in \mathbb{R}^d$ is defined by

$$\forall u \in \mathbb{R}^d, \quad \Psi_X(u) = \mathbb{E} \left[e^{i\langle u, X \rangle} \right] = \mathbb{E} [\cos(\langle u, X \rangle)] + i \mathbb{E} [\sin(\langle u, X \rangle)].$$

Two random vectors have the same law if and only if their characteristic functions coincide.

Proposition 1.2.2 (Characteristic function of Gaussian variables). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, for any $u \in \mathbb{R}$,*

$$\Psi_X(u) = \exp \left(i\mu u - \frac{\sigma^2}{2} u^2 \right).$$

The proof of Proposition 1.2.2 is postponed to Exercise 1.3.9.

Definition 1.2.3 (Gaussian vector). *A random vector $X \in \mathbb{R}^d$ is Gaussian if, for any $u \in \mathbb{R}^d$, there exist $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$ such that $\langle u, X \rangle \sim \mathcal{N}(\mu, \sigma^2)$.*

Let $X \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$. Set $m = \mathbb{E}[X] \in \mathbb{R}^d$ and $K = \text{Cov}[X] \in \mathbb{R}^{d \times d}$. For any $u \in \mathbb{R}^d$, it is immediate that $\mathbb{E}[\langle u, X \rangle] = \langle u, m \rangle$, and by Exercise 1.2.1, $\text{Var}(\langle u, X \rangle) = \langle u, Ku \rangle$. Therefore, if X is Gaussian, then necessarily, $\langle u, X \rangle \sim \mathcal{N}(\langle u, m \rangle, \langle u, Ku \rangle)$, and thus by Proposition 1.2.2,

$$\Psi_X(u) = \mathbb{E} \left[e^{i\langle u, X \rangle} \right] = \exp \left(i\langle u, m \rangle - \frac{1}{2} \langle u, Ku \rangle \right).$$

We deduce the following statement.

Proposition 1.2.4 (Characteristic function of Gaussian vectors). *The random vector X is Gaussian if and only if there exist $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d \times d}$ such that, for any $u \in \mathbb{R}^d$,*

$$\Psi_X(u) = \exp \left(i\langle u, m \rangle - \frac{1}{2} \langle u, Ku \rangle \right).$$

In this case, we have $m = \mathbb{E}[X]$ and $K = \text{Cov}[X]$, and we denote by $\mathcal{N}_d(m, K)$ the law of X .

We now address the question of how to simulate a random vector drawn from the Gaussian measure $\mathcal{N}_d(m, K)$ for some given $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d \times d}$. To proceed, we first remark that the Box–Muller method described in Proposition 1.1.17 allows to simulate independent realisations $G_1, \dots, G_d$ of the standard Gaussian distribution. We next recall that, by the Spectral Theorem, for any symmetric nonnegative matrix $K \in \mathbb{R}^{d \times d}$, there exists $\lambda_1, \dots, \lambda_d \geq 0$ and an orthonormal basis $(e_1, \dots, e_d)$ of $\mathbb{R}^d$ such that for any i , $Ke_i = \lambda_i e_i$.

Proposition 1.2.5 (Simulation of Gaussian vectors). *Let $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d \times d}$ be a symmetric and nonnegative matrix, with associated eigenvalues $\lambda_1, \dots, \lambda_d \geq 0$ and eigenvectors $(e_1, \dots, e_d)$. Let $G_1, \dots, G_d$ be independent standard Gaussian variables. Then*

$$X = m + \sum_{i=1}^d G_i \sqrt{\lambda_i} e_i \sim \mathcal{N}_d(m, K).$$

Proof. For any $u \in \mathbb{R}^d$,

$$\langle u, X \rangle = \langle u, m \rangle + \sum_{i=1}^d G_i \sqrt{\lambda_i} \langle u, e_i \rangle$$

is a sum of independent Gaussian variables, therefore by Exercise 1.3.9, it is a Gaussian variable. Hence, X is a Gaussian vector. Besides, it is immediate that $\mathbb{E}[\langle u, X \rangle] = \langle u, m \rangle$, and by independence,

$$\text{Var}(\langle u, X \rangle) = \sum_{i=1}^d \lambda_i \langle u, e_i \rangle^2 = \langle u, K u \rangle,$$

which shows that $\mathbb{E}[X] = m$ and $\text{Cov}[X] = K$. $\square$

Proposition 1.2.5 has the practical interest to show that, up to diagonalising the covariance matrix, it is possible to sample from the Gaussian measure $\mathcal{N}_d(m, K)$ as soon as independent standard Gaussian random variables are available. It may also be useful for theoretical purposes, as in the next exercise.

Exercise 1.2.6. Show that, if K is invertible, $X \sim \mathcal{N}_d(m, K)$ has density

$$\frac{1}{\sqrt{(2\pi)^d \det(K)}} \exp\left(-\frac{\langle x - m, K^{-1}(x - m) \rangle}{2}\right)$$

with respect to the Lebesgue measure on $\mathbb{R}^d$. If K is not invertible, can you find a similar density with respect to another measure?

1.2.2 Copulas

Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$. In general, the collection of the *marginal* laws of $X_1, \dots, X_d$ does not characterise the *joint* law of the vector, and a supplementary information is needed to describe how these variables depend on each other. For Gaussian vectors, this information is contained in the covariance matrix. Beyond the case of Gaussian vectors, the notion of *copula* allows to characterise the dependency between the coordinates of a random vector.

Definition 1.2.7 (Copula). A function $C : [0, 1]^d \rightarrow [0, 1]$ is called a copula if there exists a random vector $(U_1, \dots, U_d) \in [0, 1]^d$ such that:

- (i) for any $i \in \{1, \dots, d\}$, $U_i \sim \mathcal{U}[0, 1]$;
- (ii) for any $(u_1, \dots, u_d) \in [0, 1]^d$, $C(u_1, \dots, u_d) = \mathbb{P}(U_1 \leq u_1, \dots, U_d \leq u_d)$.

As a consequence of Definition 1.2.7, a copula has the following properties:

- it is nondecreasing in each coordinate;
- for any $u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d$, $C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_d) = 0$;
- for any u_i , $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$.

Some elementary examples of copulas are given by the *independent* copula

$$C(u_1, \dots, u_d) = u_1 \cdots u_d,$$

and the *comonotonic* copula

$$C(u_1, \dots, u_d) = \min(u_1, \dots, u_d).$$

Exercise 1.2.8. Describe the law of the random vectors $(U_1, \dots, U_d)$ respectively associated with the independent and comonotonic copulas.

The main result about copulas is the following statement, in which we generalise Definition 1.1.6 to random vectors by letting $F_X(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$. It remains true that the CDF of X characterises its law.

Theorem 1.2.9 (Sklar's Theorem⁷). Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ be a random vector with CDF F_X .

(i) There exists a copula C_X such that for any $(x_1, \dots, x_d) \in \mathbb{R}^d$,

$$F_X(x_1, \dots, x_d) = C_X(F_{X_1}(x_1), \dots, F_{X_d}(x_d)).$$

(ii) If the marginal CDFs $F_{X_1}, \dots, F_{X_d}$ are continuous, then the copula is unique and given by, for any $(u_1, \dots, u_d) \in [0, 1]^d$,

$$C_X(u_1, \dots, u_d) = F_X(F_{X_1}^{-1}(u_1), \dots, F_{X_d}^{-1}(u_d)).$$

The copula of a random vector therefore allows to isolate the dependency structure of its components, apart from their marginal distributions.

Exercise 1.2.10 (The Gaussian copula). Let $X \sim \mathcal{N}_d(m, K)$ and R the associated correlation matrix, with coefficients

$$R_{ij} = \rho(X_i, X_j) := \begin{cases} \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}} & \text{if } K_{ii}K_{jj} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that the copula of X is given by

$$C_X(u_1, \dots, u_d) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)),$$

where Φ is the CDF of the standard Gaussian distribution on $\mathbb{R}$, and Φ_R is the CDF of the Gaussian measure $\mathcal{N}_d(0, R)$.

Given the system of marginal distributions and the copula of a random vector X , we now ask how to generate samples of X . This is done with the following two-step procedure.

Lemma 1.2.11 (Sampling vectors with given marginal distributions and copulas). Let C be a copula and $F_1, \dots, F_d$ be CDFs on $\mathbb{R}$. Consider the following algorithm:

1. Generate $(U_1, \dots, U_d)$ with CDF C .
2. Return $X = (F_1^{-1}(U_1), \dots, F_d^{-1}(U_d))$.

The vector X has copula C and each component X_i has CDF F_i .

⁷Sklar, A. Fonctions de répartition à n dimensions et leurs marges, *Publications de l'Institut de Statistiques de l'Université de Paris* (1959).

The proof of Lemma 1.2.11 is straightforward. We now focus on the first step, namely: given a copula C , how to sample $(U_1, \dots, U_d) \in [0, 1]^d$ with CDF C ? To proceed, we assume that $(U_1, \dots, U_d)$ has a density $c(u_1, \dots, u_d)$, related with C by the identity

$$\frac{\partial^d C}{\partial u_1 \dots \partial u_d}(u_1, \dots, u_d) = c(u_1, \dots, u_d).$$

For any $k \in \{1, \dots, d\}$, the marginal density of $(U_1, \dots, U_k)$ is given by

$$\begin{aligned} c_k(u_1, \dots, u_k) &= \int_{v_{k+1}=0}^1 \dots \int_{v_d=0}^1 c(u_1, \dots, u_k, v_{k+1}, \dots, v_d) dv_d \dots dv_{k+1} \\ &= \frac{\partial^k C}{\partial u_1 \dots \partial u_k}(u_1, \dots, u_k, 1, \dots, 1), \end{aligned}$$

and, for $k \in \{1, \dots, d-1\}$, it satisfies the identity

$$c_k(u_1, \dots, u_k) = \int_{v_{k+1}=0}^1 c_{k+1}(u_1, \dots, u_k, v_{k+1}) dv_{k+1}.$$

As a consequence, for any $(u_1, \dots, u_k) \in [0, 1]^k$, the function $F_{k+1}(\cdot | u_1, \dots, u_k)$ defined by

$$\forall u_{k+1} \in [0, 1], \quad F_{k+1}(u_{k+1} | u_1, \dots, u_k) := \frac{\int_{v_{k+1}=0}^{u_{k+1}} c_{k+1}(u_1, \dots, u_k, v_{k+1}) dv_{k+1}}{\int_{v_{k+1}=0}^1 c_{k+1}(u_1, \dots, u_k, v_{k+1}) dv_{k+1}}$$

is a CDF, which can be interpreted as the conditional CDF of U_{k+1} given $(U_1, \dots, U_k)$ ⁸.

We now consider the following algorithm:

1. draw $U_1 \sim \mathcal{U}[0, 1]$;
2. for $k = 1, \dots, d-1$, draw $U'_{k+1} \sim \mathcal{U}[0, 1]$ independently from $(U_1, \dots, U_k)$ and set $U_{k+1} = F_{k+1}^{-1}(U'_{k+1} | U_1, \dots, U_k)$.

Proposition 1.2.12 (Sampling from a copula). *The vector $(U_1, \dots, U_d) \in [0, 1]^d$ generated by the algorithm above has CDF C .*

Proof. For $k \in \{1, \dots, d\}$, we set $C_k(u_1, \dots, u_k) = C(u_1, \dots, u_k, 1, \dots, 1)$, and show by induction on k that C_k is the CDF of $(U_1, \dots, U_k)$ constructed by the algorithm above. For $k = 1$ this is straightforward since $C_1(u_1) = u_1$ by the basic properties of copulas. Let us now fix $k \in \{1, \dots, d-1\}$ such that $(U_1, \dots, U_k)$ has CDF C_k , and therefore density $c_k(u_1, \dots, u_k)$, and compute the CDF of $(U_1, \dots, U_{k+1})$. Since $U_{k+1} = F_{k+1}^{-1}(U'_{k+1} | U_1, \dots, U_k)$, with $U'_{k+1} \sim \mathcal{U}[0, 1]$ independent from $(U_1, \dots, U_k)$, we may write

$$\begin{aligned} &\mathbb{P}(U_1 \leq u_1, \dots, U_k \leq u_k, U_{k+1} \leq u_{k+1}) \\ &= \mathbb{P}(U_1 \leq u_1, \dots, U_k \leq u_k, F_{k+1}^{-1}(U'_{k+1} | U_1, \dots, U_k) \leq u_{k+1}) \\ &= \mathbb{P}(U_1 \leq u_1, \dots, U_k \leq u_k, U'_{k+1} \leq F_{k+1}(u_{k+1} | U_1, \dots, U_k)) \\ &= \int_{(v_1, \dots, v_k, v'_{k+1}) \in [0, 1]^{k+1}} \mathbb{1}_{\{v_1 \leq u_1, \dots, v_k \leq u_k, v'_{k+1} \leq F_{k+1}(u_{k+1} | v_1, \dots, v_k)\}} c_k(v_1, \dots, v_k) dv_1 \dots dv_k dv'_{k+1} \\ &= \int_{(v_1, \dots, v_k) \in [0, 1]^k} \mathbb{1}_{\{v_1 \leq u_1, \dots, v_k \leq u_k\}} F_{k+1}(u_{k+1} | v_1, \dots, v_k) c_k(v_1, \dots, v_k) dv_1 \dots dv_k. \end{aligned}$$

⁸because its derivative $c_k(u_1, \dots, u_k)/c_{k+1}(u_1, \dots, u_{k+1})$ is the conditional density of U_{k+1} given $(U_1, \dots, U_k)$.

By the definition of F_{k+1} , we have

$$F_{k+1}(u_{k+1}|v_1, \dots, v_k) c_k(v_1, \dots, v_k) = \int_{v_{k+1}=0}^{u_{k+1}} c_{k+1}(v_1, \dots, v_k, v_{k+1}) dv_{k+1},$$

which yields

$$\begin{aligned} & \mathbb{P}(U_1 \leq u_1, \dots, U_k \leq u_k, U_{k+1} \leq u_{k+1}) \\ &= \int_{(v_1, \dots, v_k, v_{k+1}) \in [0,1]^{k+1}} \mathbb{1}_{\{v_1 \leq u_1, \dots, v_k \leq u_k, v_{k+1} \leq u_{k+1}\}} c_{k+1}(v_1, \dots, v_{k+1}) dv_1 \cdots dv_{k+1} \\ &= C_{k+1}(u_1, \dots, u_{k+1}), \end{aligned}$$

and completes the proof. $\square$

1.3 Complements

1.3.1 Exercises

Exercise 1.3.1 (Inverse CDF for standard densities). *Apply the inverse CDF method to the following standard probability densities.*

1. The Pareto distribution with parameter $\alpha > 0$, with density $\mathbb{1}_{\{x>1\}} \alpha x^{-(\alpha+1)}$.
2. The Cauchy distribution with parameter $a > 0$, with density $\frac{a}{\pi} \frac{1}{a^2 + x^2}$.
3. The Weibull distribution with parameter $m > 0$, with density $\mathbb{1}_{\{x>0\}} m x^{m-1} \exp(-x^m)$.
4. The Rayleigh distribution with parameter $\sigma^2 > 0$, with density $\mathbb{1}_{\{x>0\}} \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2})$.

Exercise 1.3.2 (Geometric distribution with a single U). *Let $X \sim \mathcal{E}(\lambda)$.*

1. What is the law of $\lceil X \rceil$?⁹
2. Deduce an algorithm which returns a $\mathcal{Geo}(p)$ random variable with a single uniform random variable U .

Exercise 1.3.3 (Poisson distribution). *A random variable $N \in \mathbb{N}$ is distributed according to the Poisson distribution with parameter $\lambda > 0$ if, for any $k \in \mathbb{N}$,*

$$\mathbb{P}(N = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

We denote $N \sim \mathcal{P}(\lambda)$.

1. Let $(X_i)_{i \geq 1}$ be a sequence of independent exponential random variables with parameter λ . Show that $\inf\{n \geq 0 : X_1 + \dots + X_{n+1} \geq 1\} \sim \mathcal{P}(\lambda)$.
2. Deduce an algorithm to draw a random variable $N \sim \mathcal{P}(\lambda)$ using a sequence $(U_i)_{i \geq 1}$ of independent uniform variables on $[0, 1]$.

Exercise 1.3.4 (Inverse of the inverse CDF). *Show that if the CDF F_X of X is continuous, then $F_X(X) \sim \mathcal{U}[0, 1]$. What happens if F is discontinuous?*

⁹For any $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$.

Exercise 1.3.5 (Unbiasing a coin toss¹⁰). Assume that you have a random number generator which returns independent Bernoulli variables with an unknown parameter $p \in (0, 1)$. How to use it to draw a Bernoulli random variable with parameter $1/2$?

Exercise 1.3.6 (The Marsaglia polar method¹¹). Let $(U_n, V_n)_{n \geq 1}$ be a sequence of iid random pairs such that for any $n \geq 1$, U_n and V_n are independent and uniformly distributed on $[-1, 1]$. For any $n \geq 1$, we define $S_n = U_n^2 + V_n^2$ and set $N = \inf\{n \geq 1 : S_n < 1\}$.

1. What is the joint law of $(N, (U_N, V_N))$?
2. Compute the law of the random pair (X, Y) defined by

$$X = U_N \sqrt{\frac{-2 \log S_N}{S_N}}, \quad Y = V_N \sqrt{\frac{-2 \log S_N}{S_N}}.$$

Exercise 1.3.7 (Gamma distribution). The Gamma distribution with (shape) parameter $a > 0$ is the probability measure on $\mathbb{R}$ with density

$$p(x) = \mathbb{1}_{\{x > 0\}} \frac{1}{\Gamma(a)} x^{a-1} e^{-x},$$

where Γ is Euler's function

$$\Gamma(a) := \int_{x=0}^{+\infty} x^{a-1} e^{-x} dx.$$

We assume that $a > 1$ and want to implement the rejection sampling method with q the density of the exponential distribution with parameter λ .

1. Which value of λ should we take?
2. What will be the resulting value of k ?

Exercise 1.3.8. Implement both the Box–Muller method and the Ziggurat algorithm to generate large samples of independent standard Gaussian variables, and compare their efficiency in terms of computational time.

Exercise 1.3.9 (Characteristic function of Gaussian random variables). Let $G \sim \mathcal{N}(0, 1)$.

1. Show that Ψ_G is C^1 on $\mathbb{R}$, and that for all $u \in \mathbb{R}$, $\Psi'_G(u) + u\Psi_G(u) = 0$.
2. Deduce that $\Psi_G(u) = \exp(-u^2/2)$.
3. If $X \sim \mathcal{N}(\mu, \sigma^2)$, what is the expression of $\Psi_X(u)$?
4. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$ be independent. Compute the law of $X + Y$.

¹⁰This exercise is attributed to Von Neumann.

¹¹This method was introduced in Marsaglia, G. and Bray, T. A. (1964). A Convenient Method for Generating Normal Variables. *SIAM Review*.

1.3.2 Further comments

Beyond the fact that they are limited by their small period, linear congruential generators also suffer from other issues: you can check https://en.wikipedia.org/wiki/Linear_congruential_generator for further explanations.

A natural, slightly metaphysical question, is the following: given a probability measure P on some measurable space $(E, \mathcal{E})$, can we always construct a random variable X with distribution P ? The answer depends on the choice of the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Of course, by taking $(\Omega, \mathcal{A}, \mathbb{P}) = (E, \mathcal{E}, P)$, the *canonical* variable $X(\omega) = \omega$ has law P . So, given P on $(E, \mathcal{E})$, one may always find a probability space on which it is possible to construct a random variable with distribution P . However, if the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is given in advance, it may not be possible to construct such a random variable: for instance, if Ω is a finite set, then it is not possible to construct a random variable $X \in \mathbb{R}$ with a density. The overall idea is therefore that Ω must be “large enough”. In this perspective, the *Fundamental Principle of Simulation* [4, Section 1.2] states that, as soon as E is a Polish space¹², then one may take $\Omega = [0, 1]$ endowed with the Borel σ -field and the Lebesgue measure. Indeed, on this space, for any probability measure P , there exists a random variable X defined on this space with law P . It even holds that given a sequence of probability measures $(P_n)_{n \geq 1}$, such that each P_n is a probability measure on some Polish space E_n , there exists a sequence of independent random variables $(X_n)_{n \geq 1}$ defined on Ω such that each X_n has law P_n .

The notion of Gaussian vector introduced in Subsection 1.2.1 can be generalised to infinite-dimensional spaces, typically Hilbert or Banach spaces. This may be useful to model random fields for example. In this context, the simulation method presented in Proposition 1.2.5 can be extended, under the assumption that the covariance operator has good spectral properties. It is called the *Karhunen–Loève expansion*.

¹²that is to say that E a topological space which is separable and whose topology is induced by a distance which makes it complete, and $\mathcal{E}$ is the Borel σ -field induced by this topology.

Lecture 2

The Monte Carlo method

2.1 The Monte Carlo method

The goal of the Monte Carlo method is to numerically approximate an integral which writes under the form

$$\mathcal{J} := \int_{x \in E} f(x) P(\mathrm{d}x), \quad (2.1)$$

where $(E, \mathcal{E})$ is a measurable space, P is a probability measure on E and $f \in \mathbf{L}^1(P)$.

2.1.1 Deterministic approach

Assume for simplicity that $E = [0, 1]^d$ and that $P(\mathrm{d}x) = \mathrm{d}x$ is the uniform distribution. Then fixing $N \geq 1$ and setting $x_{\vec{k}} = (k_1/N, \dots, k_d/N)$ for $\vec{k} = (k_1, \dots, k_d) \in \{0, \dots, N-1\}^d$, the basic deterministic approximation of $\mathcal{J}$ is given by

$$\mathcal{J}_N := \frac{1}{N^d} \sum_{\vec{k}} f(x_{\vec{k}}),$$

obtained by replacing f with the piecewise constant function which takes the value $f(x_{\vec{k}})$ on the cell $C_{\vec{k}} := [k_1/N, (k_1+1)/N) \times \dots \times [k_d/N, (k_d+1)/N)$.

The precision of this approximation is given by the fact that, if you assume that f is Lipschitz continuous, then

$$|\mathcal{J} - \mathcal{J}_N| = \left| \sum_{\vec{k}} \int_{C_{\vec{k}}} (f(x) - f(x_{\vec{k}})) \mathrm{d}x \right| \leq \sum_{\vec{k}} \int_{C_{\vec{k}}} |f(x) - f(x_{\vec{k}})| \mathrm{d}x \lesssim \frac{1}{N}.$$

As a consequence, to reach a precision $\epsilon \simeq 1/N$, one needs to evaluate f at $N^d \simeq (1/\epsilon)^d$ points. This quantity grows exponentially in d : this is the *curse of dimensionality*.

2.1.2 Stochastic approach

The formulation (2.1) of $\mathcal{J}$ allows us to rewrite it under the form

$$\mathcal{J} = \mathbb{E}[f(X)],$$

where X is a random variable in E with law P . Then, if $(X_n)_{n \geq 1}$ is a sequence of iid random variables with common distribution P , the (strong) Law of Large Numbers ensures that

$$\hat{\mathcal{J}}_n := \frac{1}{n} \sum_{i=1}^n f(X_i)$$

converges almost surely to $\mathcal{J}$. The precision of the approximation of $\mathcal{J}$ by $\hat{\mathcal{J}}_n$ is measured by the Central Limit Theorem, which ensures that if $\sigma^2 := \text{Var}(f(X)) < +\infty$, then

$$\lim_{n \rightarrow +\infty} \frac{\sqrt{n}}{\sigma} (\hat{\mathcal{J}}_n - \mathcal{J}) = \mathcal{N}(0, 1), \quad \text{in distribution.}$$

This result ensures in particular that, given $\alpha \in (0, 1/2)$ and denoting by $\phi_{1-\alpha/2}$ the quantile of order $1 - \alpha/2$ of $\mathcal{N}(0, 1)$ (see Figure 2.1), the interval

$$\left[\hat{\mathcal{J}}_n - \phi_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\mathcal{J}}_n + \phi_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (2.2)$$

contains $\mathcal{J}$ with probability converging to $1 - \alpha$ when $n \rightarrow +\infty$. Therefore, to reach a precision $\epsilon \simeq \sigma/\sqrt{n}$, one needs to evaluate f at $n \simeq \sigma^2/\epsilon^2$ points, which only depends on the underlying dimension of E through the prefactor σ^2 . So this method avoids the curse of dimensionality.

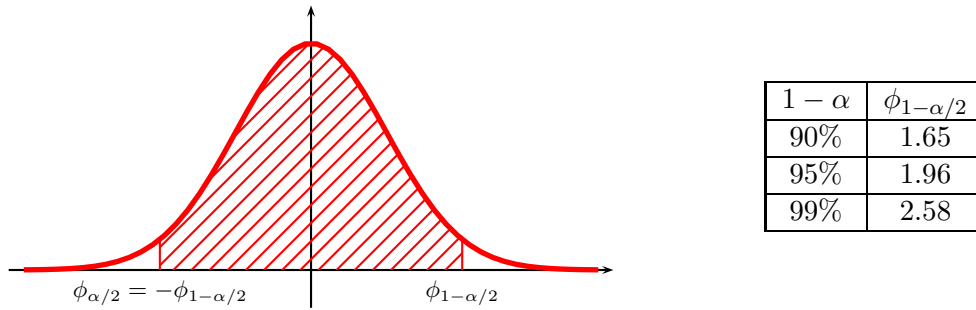


Figure 2.1: Quantiles of the standard Gaussian distribution. The hatched area on the figure is equal to $1 - \alpha$.

2.1.3 Confidence intervals

In general, σ^2 is not known either, so the interval (2.2) cannot be actually computed.

Asymptotic approach

If n is large, σ^2 may however be estimated by the empirical variance of the sample

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n \left(f(X_i) - \hat{\mathcal{J}}_n \right)^2,$$

whose computation does not require new samples from X nor new evaluation of the function f . By Slutsky's Lemma,

$$\lim_{n \rightarrow +\infty} \frac{\sqrt{n}}{\hat{\sigma}_n} (\hat{\mathcal{J}}_n - \mathcal{J}) = \mathcal{N}(0, 1), \quad \text{in distribution,}$$

and therefore the interval

$$\left[\hat{\mathcal{J}}_n - \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\mathcal{J}}_n + \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$$

also contains $\mathcal{J}$ with probability converging to $1 - \alpha$ when $n \rightarrow +\infty$; it is called an *asymptotic confidence interval*.

Nonasymptotic approach

If n is not large enough for the asymptotic result above to hold, the Bienaymé–Chebychev inequality yields, for any $r > 0$,

$$\mathbb{P} \left(\left| \hat{\mathcal{J}}_n - \mathcal{J} \right| \geq r \right) \leq \frac{1}{r^2} \text{Var}(\hat{\mathcal{J}}_n) = \frac{\sigma^2}{nr^2}.$$

Assuming that an upper-bound M_{σ^2} is available on σ^2 , as in Lemma 2.1.1 below, we deduce that the interval

$$\left[\hat{\mathcal{J}}_n - \sqrt{\frac{M_{\sigma^2}}{n\alpha}}, \hat{\mathcal{J}}_n + \sqrt{\frac{M_{\sigma^2}}{n\alpha}} \right]$$

contains $\mathcal{J}$ with probability *at least* $1 - \alpha$: this interval is called an *approximate confidence interval*.

This procedure requires to be able to derive upper bounds on the variance of $f(X_1)$. This is for example possible when f is bounded.

Lemma 2.1.1 (Universal bound on the variance). *Assume that $f(x) \in [a, b]$ for any $x \in E$, with $-\infty < a \leq b < +\infty$. Then*

$$\sigma^2 = \text{Var}(f(X)) \leq \frac{(b-a)^2}{4}.$$

Proof. The statement is obvious if $a = b$, otherwise we let $U = f(X)/(b-a) \in [0, 1]$. We have $U^2 \leq U$ and therefore

$$\text{Var}(U) = \mathbb{E}[U^2] - \mathbb{E}[U]^2 \leq \mathbb{E}[U] - \mathbb{E}[U]^2 \leq \sup_{u \in [0,1]} u - u^2 = \frac{1}{4},$$

and we complete the proof by noting that $\text{Var}(f(X)) = (b-a)^2 \text{Var}(U)$. □

The whole game of approximate confidence intervals is to find bounds as sharp as possible, because taking larger and larger confidence intervals increases the probability of $\mathcal{J}$ to belong to the interval, but makes the estimation less precise. In an extreme and caricatural case, $\mathbb{R}$ is an interval which contains $\mathcal{J}$ with probability 1, so larger than any $1 - \alpha$, but this is not informative on the value of $\mathcal{J}$ at all. In this perspective, when f is bounded, the *Hoeffding inequality*¹ provides sharper confidence intervals (as a function of α) as the Bienaymé–Chebychev inequality.

Proposition 2.1.2 (Hoeffding inequality). *Under the assumptions of Lemma 2.1.1, we have, for any $r \geq 0$ and $n \geq 1$,*

$$\mathbb{P} \left(\left| \hat{\mathcal{J}}_n - \mathcal{J} \right| \geq r \right) \leq 2 \exp \left(-\frac{2nr^2}{(b-a)^2} \right).$$

The proof of Proposition 2.1.2 is detailed in Exercise 2.3.1.

¹Hoeffding, W. Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* (1963).

2.2 Variance reduction

As is discussed in the previous Section, the precision of the Monte Carlo method essentially depends on the ratio $\sigma/\sqrt{n}$. There are two typical situations in which this ratio may be large:

- sampling from X , or evaluating f at the sample X , may be costly, so the computational budget n may be limited;
- the standard deviation σ of $f(X)$ may be large with respect to the expectation $\mathcal{J}$ of $f(X)$, so the size of the sample n required to have a good approximation $\sigma/\sqrt{n} \ll |\mathcal{J}|$ may be huge.

An instance of the second situation is the *rare event setting*: assume that $f(x) = \mathbb{1}_{\{x \in A\}}$ for some subset A such that $\mathcal{J} = \mathbb{P}(X \in A) \ll 1$: we are trying to estimate the probability of a rare event, which we prefer to denote by p rather than $\mathcal{J}$. Then $\sigma^2 = \text{Var}(\mathbb{1}_{\{X \in A\}}) = p(1-p) \simeq p$, so to reach a relative precision δ , that is to say to have $\sigma/\sqrt{n}$ of order δp , one needs $n \simeq 1/(p\delta^2)$ samples. If the probability that we aim to estimate is $p = 10^{-6}$ then, for a relative precision δ of 1%, this means that the sample must be of size $n = 10^{10}$.

Exercise 2.2.1. *In the rare event setting, what is the expected number of samples that you have to draw before observing a single realisation of the rare event?*

This discussion shows that there is an interest in reducing the variance σ^2 . In this Section, we present two approaches to this issue: the control variate method and importance sampling, which are respectively adapted to the first and second situations described above. In Subsection 2.3.1, two other methods are studied: the use of antithetic variables and stratified sampling, which are more concerned with the sampling of X . Last, the splitting algorithm for the estimation of the probability of rare events is studied in Problem 1.

2.2.1 Control variate

In this Subsection, we assume that in addition to $X_1, \dots, X_n$, we are able to sample iid random variables $Y_1, \dots, Y_n$ whose common expectation $\mathbb{E}[Y]$ is known analytically. Then, for all $\beta \in \mathbb{R}$,

$$\mathcal{J} = \mathbb{E}[f(X)] = \mathbb{E}[f(X) - \beta Y] + \beta \mathbb{E}[Y],$$

which suggests to approximate $\mathcal{J}$ by the estimator

$$\hat{\mathcal{J}}_n^{\text{CV},\beta} := \frac{1}{n} \sum_{i=1}^n (f(X_i) - \beta Y_i) + \beta \mathbb{E}[Y].$$

The variance of this estimator is $(\sigma^{\text{CV},\beta})^2/n$, where

$$(\sigma^{\text{CV},\beta})^2 = \text{Var}(f(X) - \beta Y) = \sigma^2 - 2\beta \text{Cov}(f(X), Y) + \beta^2 \text{Var}(Y).$$

We may already remark that if $\text{Cov}(f(X), Y) = 0$ then $(\sigma^{\text{CV},\beta})^2$ is always larger than the variance σ^2 associated with the original Monte Carlo estimator: for the control variate method to be efficient, it is thus necessary that $f(X)$ and Y be correlated. The choice of β for which $(\sigma^{\text{CV},\beta})^2$ is minimal is then

$$\beta^* = \frac{\text{Cov}(f(X), Y)}{\text{Var}(Y)},$$

which yields the variance

$$(\sigma^{\text{CV},\beta^*})^2 = \sigma^2 (1 - \rho^2),$$

where

$$\rho = \frac{\text{Cov}(f(X), Y)}{\sqrt{\text{Var}(f(X)) \text{Var}(Y)}} \in [-1, 1]$$

is the correlation coefficient between $f(X)$ and Y . As a consequence, the more $f(X)$ and Y are correlated, the better the variance reduction. Typically, one may choose Y of the form $g(X)$, where the function g is close to f in regions where X has a high probability to take its values, while being ‘simpler’ than f , in the sense that $\mathbb{E}[g(X)]$ is easier to compute than $\mathbb{E}[f(X)]$ – see Exercise 2.3.4 for an illustration.

In practice, the optimal choice of β depends on the quantity $\text{Cov}(f(X), Y)$ which may need to be estimated. Let us introduce

$$\hat{C}_n = \frac{1}{n} \sum_{i=1}^n (f(X_i) - \hat{\mathcal{J}}_n)(Y_i - \bar{Y}_n).$$

The strong Law of Large Numbers shows that

$$\hat{\beta}_n^* := \frac{\hat{C}_n}{\text{Var}(Y)}$$

converges to β^* almost surely, and Slutsky’s Lemma then yields the following result.

Proposition 2.2.2 (Control variate method). *Let $(X_i, Y_i)_{1 \leq i \leq n}$ be a sequence of iid pairs such that $f(X_i), Y_i \in \mathbf{L}^2(\mathbb{P})$. For all $n \geq 1$, let*

$$\hat{\mathcal{J}}_n^{\text{CV}} := \frac{1}{n} \sum_{i=1}^n (f(X_i) - \hat{\beta}_n^* Y_i) + \hat{\beta}_n^* \mathbb{E}[Y],$$

with $\hat{\beta}_n^*$ defined above. The interval

$$\left[\hat{\mathcal{J}}_n^{\text{CV}} - \phi_{1-\alpha/2} \sqrt{\frac{(\hat{\sigma}_n^{\text{CV}})^2}{n}}, \hat{\mathcal{J}}_n^{\text{CV}} + \phi_{1-\alpha/2} \sqrt{\frac{(\hat{\sigma}_n^{\text{CV}})^2}{n}} \right],$$

where

$$(\hat{\sigma}_n^{\text{CV}})^2 = \hat{\sigma}_n^2 \left(1 - \frac{\hat{C}_n^2}{\hat{\sigma}_n^2 \text{Var}(Y)} \right) \rightarrow \sigma^2(1 - \rho^2),$$

is an asymptotic confidence interval.

The control variate method is typically suited for cases where the evaluation of f is costly: it may represent a high-precision numerical code. If a low-precision code g , sometimes called surrogate model, is available, then using $g(X)$ as a control variate generally provides a good variance reduction, while $\mathbb{E}[g(X)]$ can be estimated by direct Monte Carlo approach with a much larger sample size than the original one.

2.2.2 Importance sampling

Importance sampling is based on the remark that, for any probability measure Q on E such that $P \ll Q$,

$$\mathcal{J} = \int_{x \in E} f(x) dP(x) = \int_{x \in E} f(x) w(x) dQ(x),$$

where the function w is simply the density

$$w(x) = \frac{dP}{dQ}(x).$$

As a consequence, the quantity

$$\hat{\mathcal{J}}_n^{\text{IS}} := \frac{1}{n} \sum_{i=1}^n f(Y_i) w(Y_i),$$

where $Y_1, \dots, Y_n$ are iid with law Q , converges almost surely to $\mathcal{J}$. In fact, the existence of the density $w(x)$ is only necessary when $f(x) \neq 0$, so the actual condition on Q is that

$$\mathbb{1}_{\{f(x) \neq 0\}} dP(x) \ll \mathbb{1}_{\{f(x) \neq 0\}} dQ(x), \quad (2.3)$$

and we still denote by w the associated density.

Exercise 2.2.3. Show that if $P \ll Q$, then Q satisfies (2.3), but that the converse does not hold true in general.

The whole game of importance sampling then consists in choosing Q in order to make the asymptotic variance

$$(\sigma_Q^{\text{IS}})^2 := \text{Var}(f(Y)w(Y))$$

as small as possible.

Proposition 2.2.4 (Optimal choice of Q). Let $\bar{\mathcal{J}} = \mathbb{E}[|f(X)|]$, assume that this quantity is positive, and define the probability measure Q^* by

$$dQ^*(x) = \frac{|f(x)|}{\bar{\mathcal{J}}} dP(x).$$

(i) Q^* satisfies (2.3) and $(\sigma_{Q^*}^{\text{IS}})^2 = \bar{\mathcal{J}}^2 - \mathcal{J}^2$.

(ii) For any probability measure Q which also satisfies (2.3), $(\sigma_Q^{\text{IS}})^2 \leq (\sigma_{Q^*}^{\text{IS}})^2$.

(iii) If f has constant sign P -almost everywhere, then $(\sigma_{Q^*}^{\text{IS}})^2 = 0$.

Proof. As a preliminary remark, we note that for any Q satisfying (2.3),

$$(\sigma_Q^{\text{IS}})^2 = \mathbb{E} \left[(f(Y)w(Y))^2 \right] - \mathcal{J}^2, \quad Y \sim Q. \quad (2.4)$$

First, it is easily checked that $\mathbb{1}_{\{f(x) \neq 0\}} dP(x)$ has density

$$w^*(x) = \mathbb{1}_{\{f(x) \neq 0\}} \frac{\bar{\mathcal{J}}}{|f(x)|}$$

with respect to $\mathbb{1}_{\{f(x) \neq 0\}} dQ^*(x)$, therefore Q^* satisfies (2.3) and besides, if $Y^* \sim Q^*$, then

$$\begin{aligned} \mathbb{E} \left[(f(Y^*)w^*(Y^*))^2 \right] &= \int_{x \in E} \mathbb{1}_{\{f(x) \neq 0\}} |f(x)|^2 \left(\frac{\bar{\mathcal{J}}}{|f(x)|} \right)^2 dQ^*(x) \\ &= \bar{\mathcal{J}}^2 \int_{x \in E} \mathbb{1}_{\{f(x) \neq 0\}} dQ^*(x) \\ &= \bar{\mathcal{J}}^2, \end{aligned}$$

which, together with (2.4), proves (i). The point (iii) then immediately follows.

Second, let us fix Q which satisfies (2.3) and denote by w the associated density. By definition of $\bar{\mathcal{J}}$ and w , and the Cauchy–Schwarz inequality,

$$\begin{aligned}\bar{\mathcal{J}}^2 &= \left(\int_{x \in E} |f(x)| \mathbb{1}_{\{f(x) \neq 0\}} dP(x) \right)^2 \\ &= \left(\int_{x \in E} |f(x)| w(x) \mathbb{1}_{\{f(x) \neq 0\}} dQ(x) \right)^2 \\ &\leq \int_{x \in E} |f(x)|^2 w(x)^2 \mathbb{1}_{\{f(x) \neq 0\}} dQ(x) \\ &= \mathbb{E} \left[(f(Y)w(Y))^2 \right],\end{aligned}$$

with $Y \sim Q$. Combined with (2.4), this estimate completes the proof of (ii). $\square$

In practice it is impossible to implement the method with the optimal measure Q^* since the latter depends explicitly on the quantity $\bar{\mathcal{J}}$, which is likely to be unknown — and, in the case where f is nonnegative P -almost everywhere, is exactly the quantity $\mathcal{J}$ which we aim to estimate. Still, this lemma suggests that a ‘good’ choice of Q would be one which has a large mass under the measure $|f(x)|dP(x)$.

Importance sampling is particularly adapted for the estimation of rare event probabilities, an example of application is proposed in Exercise 2.3.5. In this context, the theory of *Large Deviations* is a good tool to study the efficiency of the method: this is presented in Problem 2.

2.3 Complements

2.3.1 Exercises

Exercise 2.3.1 (The Hoeffding inequality). *Throughout the exercise, we let $Y_1, \dots, Y_n$ be iid random variables which take their values in $[0, 1]$. We set $Z_i = Y_i - \mathbb{E}[Y_i]$ and, for any $\lambda \geq 0$, define*

$$F(\lambda) = \log \mathbb{E} [\exp(\lambda Z_1)].$$

1. Show that $F'(\lambda) = \mathbb{E}_\lambda[Z_1]$ and $F''(\lambda) = \text{Var}_\lambda(Z_1)$ for some probability measure $\mathbb{P}_\lambda$ to be defined.
2. Deduce that, for any $\lambda \geq 0$, $\mathbb{E}[\exp(\lambda Z_1)] \leq \exp(\lambda^2/8)$.
3. Deduce that, for any $r \geq 0$ and $n \geq 1$,

$$\mathbb{P} \left(\sum_{i=1}^n Z_i \geq r\sqrt{n} \right) \leq \exp \left(\frac{\lambda^2 n}{8} - \lambda r\sqrt{n} \right).$$

4. Optimising in $\lambda \geq 0$, conclude that

$$\mathbb{P} \left(\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \geq r\sqrt{n} \right) \leq \exp(-2r^2),$$

5. Complete the proof of Proposition 2.1.2.

6. Under the assumptions of Proposition 2.1.2, compute an approximate confidence interval for $\mathcal{J}$ based on the Hoeffding inequality, and compare the width of this interval with the width of the approximate confidence interval given by the Bienaymé–Chebychev inequality.

Exercise 2.3.2 (Antithetic variables). Let $f : [0, 1] \rightarrow \mathbb{R}$ be such that

$$\int_{u=0}^1 f(u)^2 du < +\infty.$$

We study a Monte Carlo method to approximate

$$\mathcal{J} := \int_{u=0}^1 f(u) du.$$

1. Let $U \sim \mathcal{U}[0, 1]$. Show that $\mathcal{J} = \frac{1}{2} (\mathbb{E}[f(U)] + \mathbb{E}[f(1 - U)])$.
2. Let $(U_n)_{n \geq 1}$ be a sequence of independent copies of U . Show that

$$\hat{\mathcal{J}}_{2n}^a := \frac{1}{2n} \sum_{i=1}^n (f(U_i) + f(1 - U_i))$$

converges almost surely to $\mathcal{J}$ and compute $\text{Var}(\hat{\mathcal{J}}_{2n}^a)$.

3. Let

$$\hat{\mathcal{J}}_{2n} := \frac{1}{2n} \sum_{i=1}^{2n} f(U_i)$$

be the standard Monte Carlo estimator of $\mathcal{J}$ which requires the same number of evaluations of the function f as $\hat{\mathcal{J}}_{2n}^a$ (but twice more random samples). Show that $\text{Var}(\hat{\mathcal{J}}_{2n}^a) \leq \text{Var}(\hat{\mathcal{J}}_{2n})$ if and only if $\text{Cov}(f(U), f(1 - U)) \leq 0$.

4. Assume that f is monotonic. Show that

$$\mathbb{E}[(f(U_1) - f(U_2))(f(1 - U_1) - f(1 - U_2))] \leq 0.$$

Deduce that in this case, $\text{Cov}(f(U), f(1 - U)) \leq 0$.

5. Conclude on the practical interest of the method.

Exercise 2.3.3 (Stratification). Let X be a random variable in $\mathbb{R}^d$ with law P and $f \in \mathbf{L}^2(P)$. Let

$$\mathcal{J} = \int_{x \in \mathbb{R}^d} f(x) dP(x) = \mathbb{E}[f(X)].$$

We assume that there is a finite partition of $\mathbb{R}^d$ into m measurable subsets $(A_k)_{1 \leq k \leq m}$, called strates, such that for any $k \in \{1, \dots, m\}$:

- $p_k := P(A_k) = \mathbb{P}(X \in A_k)$ is known (and positive);
- we know how to draw random samples $(X_n^k)_{n \geq 1}$ under the law $P(\cdot | A_k) = \mathbb{P}(X \in \cdot | X \in A_k)$.

For integers $n_1, \dots, n_m \geq 1$ such that $n_1 + \dots + n_m = n$, we set

$$\widehat{\mathcal{J}}_n^s := \sum_{k=1}^m p_k \widehat{\mathcal{J}}_{n_k}^k, \quad \widehat{\mathcal{J}}_{n_k}^k := \frac{1}{n_k} \sum_{i=1}^{n_k} f(X_i^k),$$

where the samples $(X_i^1)_{1 \leq i \leq n_1}, \dots, (X_i^m)_{1 \leq i \leq n_m}$ are independent from each other. Last, we define

$$\forall k \in \{1, \dots, m\}, \quad \mu_k := \mathbb{E}[f(X_1^k)], \quad \sigma_k^2 := \text{Var}(f(X_1^k)).$$

1. We first study generalities.

(a) Show that

$$\text{Var}(f(X)) = \sum_{k=1}^m p_k \sigma_k^2 + \sum_{k=1}^m p_k \left(\mu_k - \sum_{\ell=1}^m p_\ell \mu_\ell \right)^2.$$

Give an interpretation of this formula.

(b) Compute $\mathbb{E}[\widehat{\mathcal{J}}_n^s]$.

(c) How does $\widehat{\mathcal{J}}_n^s$ behave when $\min(n_1, \dots, n_m) \rightarrow +\infty$?

(d) Show that $\text{Var}(\widehat{\mathcal{J}}_n^s) = \sum_{k=1}^m \frac{p_k^2 \sigma_k^2}{n_k}$.

2. We now fix n and look for the optimal allocation of $(n_1, \dots, n_m)$.

(a) Show that, for any $n_1, \dots, n_m$,

$$\left(\sum_{k=1}^m p_k \sigma_k \right)^2 \leq n \sum_{k=1}^m \frac{p_k^2 \sigma_k^2}{n_k}.$$

(b) Deduce the optimal allocation $(n_1^*, \dots, n_m^*)$ in terms of variance (without taking into account the constraint that n_k must be an integer).

(c) What do you think of the practical use of this optimal allocation?

3. We finally study the proportional allocation $n_k = np_k$, assuming for simplicity that np_k is an integer.

(a) Show that in this case $n \text{Var}(\widehat{\mathcal{J}}_n^s) \leq \text{Var}(f(X))$. Interpret this result.

(b) State and prove a Central Limit Theorem for $\widehat{\mathcal{J}}_n^s$.

(c) How to choose the strates to reduce the statistical error?

Exercise 2.3.4 (An application of the control variate method). Let $X \sim \mathcal{N}(0, 1)$. For all $t > 0$, we define

$$f_t(x) = \frac{1}{1 + tx^2},$$

and set

$$\mathcal{J} = \mathbb{E}[f_t(X)] = \frac{1}{\sqrt{2\pi}} \int_{x \in \mathbb{R}} \frac{e^{-x^2/2}}{1 + tx^2} dx.$$

Let $X_1, \dots, X_n$ be independent $\mathcal{N}(0, 1)$ variables, and let $Y_i = 1 - tX_i^2$.

1. Compute $\mathbb{E}[Y_1]$.

2. Compare numerically the variances of the Monte Carlo estimator $\hat{\mathcal{J}}_n$ and of the control variate estimator $\hat{\mathcal{J}}_n^{\text{CV}}$.
3. How does this comparison vary with t ? What is your interpretation of this fact?

Exercise 2.3.5 (Application of importance sampling). Let $X \sim \mathcal{N}(0, 1)$, and $\mathcal{J} = \mathbb{P}(X \geq 20)$. Compute (numerically) the asymptotic variance of the importance sampling estimators of $\mathcal{J}$ obtained by taking:

- q the density of $20 + Y$, where $Y \sim \mathcal{E}(1)$;
- q the density of $\mathcal{N}(20, 1)$.

Exercise 2.3.6 (Importance sampling for Bernoulli distributions). Let P be the Bernoulli distribution with parameter p . Assume that you want to implement importance sampling to estimate p . What is the optimal distribution Q on $\{0, 1\}$?

Exercise 2.3.7 (Importance sampling with and without normalisation). In the setting of Subsection 2.2.2, the importance sampling estimator

$$\hat{\mathcal{J}}_n^{\text{IS}} := \frac{1}{n} \sum_{i=1}^n f(Y_i) w(Y_i),$$

where $Y_1, \dots, Y_n$ are iid under Q and $w(x) = \frac{dP}{dQ}(x)$, rewrites as the integral of f under the nonnegative measure

$$\hat{P}_n^{\text{IS}} := \frac{1}{n} \sum_{i=1}^n w(Y_i) \delta_{Y_i}.$$

In general, this measure is not a probability measure, because its total mass $\frac{1}{n} \sum_{i=1}^n w(Y_i)$ may be different from 1. One may therefore consider the normalised importance sampling estimator

$$\hat{\mathcal{J}}_n^{\text{NIS}} := \frac{\sum_{i=1}^n f(Y_i) w(Y_i)}{\sum_{i=1}^n w(Y_i)},$$

which is the integral of f under the probability measure

$$\hat{P}_n^{\text{NIS}} := \frac{\sum_{i=1}^n w(Y_i) \delta_{Y_i}}{\sum_{i=1}^n w(Y_i)}.$$

The goal of this exercise is to study the properties of this estimator. For simplicity, we assume that $f(x) \neq 0$ for any $x \in E$.

1. Show that $\hat{\mathcal{J}}_n^{\text{NIS}}$ converges to $\mathcal{J}$, almost surely.
2. Using the Delta method, show that

$$\lim_{n \rightarrow \infty} \sqrt{n} \left(\hat{\mathcal{J}}_n^{\text{NIS}} - \mathcal{J} \right) = \mathcal{N} \left(0, (\sigma_Q^{\text{NIS}})^2 \right),$$

with

$$(\sigma_Q^{\text{NIS}})^2 = \text{Var} (w(Y)(f(Y) - \mathcal{J})).$$

3. What is the optimal choice $Q^{*,\text{NIS}}$ of Q which minimises $(\sigma_Q^{\text{NIS}})^2$, and what is the associated value of $(\sigma_{Q^{*,\text{NIS}}}^{\text{NIS}})^2$?
4. We now want to compare $(\sigma_{Q^{*,\text{NIS}}}^{\text{NIS}})^2$ with the optimal asymptotic variance $\bar{\mathcal{J}}^2 - \mathcal{J}^2$ of the standard importance sampling estimator described in Proposition 2.2.4.
 - (a) What happens if f has constant sign, P -almost everywhere?
 - (b) Construct an exemple of distribution P and function f for which the normalised estimator has a strictly smaller optimal asymptotic variance than the standard estimator.

2.3.2 Further comments

The deterministic approach described in Subsection 2.1.1 is very basic. There are many more developed deterministic integration methods, see for instance [5, Chapter 9]. They however all suffer from the curse of dimensionality.

Hoeffding's inequality is a typical example of a *concentration inequality*, which is an important research topic in probability and statistics.

The variance reduction methods described in Section 2.2 only focus on decreasing σ . There are however alternative approaches. One may design sampling schemes which generate sequences $(X_n)_{n \geq 1}$ which are not iid, but more space-filling. This is the basis of *Quasi-Monte Carlo* methods. On the other hand, when the sample size is limited by the computational cost of the evaluation of the function f , *surrogate modelling* techniques seek an approximate model $\hat{f}$ which is cheaper to evaluate and therefore allows one to increase the sample size.

The development of Monte Carlo methods is closely linked to the rare event setting. A nice historical review with bibliographical references is available here: <https://perso.lpsm.paris/~aguyader/files/biblioEVT.pdf>. The book [1] also provides details on variance reduction methods, in particular in the rare event setting.

Lecture 3

Markov chains and ergodic theorems

Lecture 4

Convergence to equilibrium of Markov chains

Lecture 5

The Markov Chain Monte Carlo Method

Lecture 6

Stochastic processes, Brownian motion and Itô calculus

Lecture 7

Stochastic differential equations

Lecture 8

Long time behavior of diffusion processes

Part II

Topics in stochastic simulation algorithms

Problem 1

Splitting algorithm for rare event estimation

Problem 2

**Asymptotic efficiency of importance
sampling through large deviation theory**

Bibliography

- [1] Søren Asmussen and Philip W. Glynn. *Stochastic Simulation*. Springer, New York, NY (USA), 2007. English language.
- [2] Benjamin Jourdain. *Probabilités et statistiques*. Ellipses, Paris, 2016. In French; <http://cermics.enpc.fr/~jourdain/probastat/poly.pdf>.
- [3] Jean-François Le Gall. Intégration, probabilités et processus aléatoires. <https://www.math.u-psud.fr/~jflegall/IPPA2.pdf>, 2006. In French – Lecture given at École Normale Supérieure.
- [4] Gilles Pagès. *Numerical Probability: An introduction with applications to finance*. Springer, Cham (Switzerland), 2018. English language.
- [5] Terry J. Sullivan. *Introduction to Uncertainty Quantification*. Springer, Cham (Switzerland), 2015. English language.