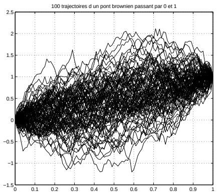
Randal Douc

Numerical methods in mathematical finance



April 20, 2018

John Von Neumann Institute

Contents

Plan of action

This is a 24 hours course schedule. The plan of action is the following.

1. Central limit theorem, Confidence intervals: lecture, 1H30.
2. Exact sampling: The inverse cumulative distribution function, the rejection sampling: lecture: 1H30, tutorials, **computer sessions:** 1H30.
3. Importance sampling: lecture, 1H30, tutorials, 1H30.
4. **Computer sessions:** *sampling from a given distribution*: 1H30.
5. Sampling from the Brownian motion: lecture, 1H30, tutorials, 1H30
6. The Euler scheme: lecture, 1H30, tutorials, 1H30.
7. **Computer sessions:** *discretization of SDE*: 1H30
8. Variance reduction, antithetic variates, control variates, conditioning, stratified sampling: lecture, 3H00. Td, 1H30
9. **Computer sessions:** *variance reduction*: 1H30
10. Weak discrepancy sequence: lecture, 1H30.

Chapter 1
Preliminaries

Contents

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. My email: `randal.douc@it-sudparis.eu`

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 $r \leq s$, we write $[r:s] = \{r, r+1, \ldots, s\}$,
- $X \perp\!\!\!\perp Y$ means X and Y are independent random variables,
- $X \overset{\mathcal{L}}{\equiv} Y$ means X and Y have the same law.
- Let $(\mathsf{Z}, \mathscr{Z})$ and $(\mathsf{X}, \mathscr{X})$ two measurable spaces. Assume that for all $w \in \mathsf{Z}$, the two random vectors X and $Y(w)$ take values in X and assume the existence of a third random variable W taking values in Z, then the notation: $X|_{W=w} \overset{\mathcal{L}}{\equiv} Y(w)$ means that for all $A \in \mathscr{X}$,
$$\mathbb{P}(X \in A|W)|_{W=w} = \mathbb{P}(Y(w) \in A)$$
In words, the distribution of X conditionally on W taken on $W = w$ is the same as the unconditonal distribution of $Y(w)$.
- $\liminf_n a_n = \lim_{n\to\infty} (\inf_{k\geq n} a_k)$ and similarly, $\limsup_n a_n = \lim_{n\to\infty} \left(\sup_{k\geq n} a_k\right)$. 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 $\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$.
(d) for all $u \in \mathbb{R}$, we have $\lim_n \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_n}\right] = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right]$

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 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 $u \geq 0$, $\mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right] = \mathrm{e}^{-u^2/2}$. Moreover, $X \sim \mathrm{N}(\mu, \sigma^2)$ iif for all $u \geq 0$, $\mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right] = \mathrm{e}^{-u^2 \mathbb{V}\mathrm{ar}(X)/2 + \mathrm{i}u\mathbb{E}(X)}$ and in that case, $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$ and $\mu = \mathbb{E}(X)$.

1.1 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)	$\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$	No explicit expression	$\mathbb{E}[X] = \mu$, $\mathbb{V}\mathrm{ar}X = \sigma^2$
Exponential	$\exp(\lambda)$	$\lambda > 0$	$\lambda \mathrm{e}^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$	$\left(1 - \mathrm{e}^{-\lambda x}\right) \mathbb{1}_{\mathbb{R}^+}(x)$	$\mathbb{E}[X] = 1/\lambda$, $\mathbb{V}\mathrm{ar}X = 1/\lambda^2$
Gamma	$\Gamma(k, \theta)$	$(k, \theta) \in \left(\mathbb{R}_+^*\right)^2$	$\frac{x^{k-1}\mathrm{e}^{-x/\theta}}{\Gamma(k)\theta^k}$	$\frac{\Gamma_{x/\theta}(k)}{\Gamma(k)}$	

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)

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

$$\Gamma_x(k) = \int_0^x t^{k-1}\mathrm{e}^{-t}\mathrm{d}t \; (\blacktriangleright\text{INCOMPLETE GAMMA FUNCTION})$$

Chapter 2
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 financial applications boil down to the calculation of expectations. For example, in finance, the prices of financial derivatives can be written as expectations. 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 (except in the particular case of the Black and Scholes formula).

2.1 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 2.1. (►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*

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

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

Remark 2.2 *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 2.3 A preliminary result: *Let (Y_i) be iid random variables such that $\mathbb{E}\left[|Y_1|\right] < \infty$ and $\mathbb{E}\left[Y_1\right] > 0$, 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 $\theta(y_1, y_2, \ldots,) = (y_2, y_3, \ldots)$ be the shift operator. Then, a.s.,

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

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 2.4. (►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 2.5 *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*

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

then

$$\frac{\bar{X}_N - \mathbb{E}[X]}{\sqrt{\frac{\tilde{\sigma}_N^2}{N}}} \overset{w}{\Rightarrow} \mathscr{N}(0.1)\,, \qquad \frac{\bar{X}_N - \mathbb{E}[X]}{\sqrt{\frac{\hat{\sigma}_N^2}{N}}} \overset{w}{\Rightarrow} \mathscr{N}(0.1) \tag{2.1}$$

Lemma 2.6 *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.*

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

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

$$\hat{\sigma}_N^2 = \frac{1}{N-1}\left(\sum_{i=1}^N (X_i - \mathbb{E}[X])^2 - 2\underbrace{\sum_{i=1}^N (X_i - \mathbb{E}[X])}_{N(\bar{X}_N - \mathbb{E}[X])}(\bar{X}_N - \mathbb{E}[X]) + N(\bar{X}_N - \mathbb{E}[X])^2\right) = \frac{\sum_{i=1}^N (X_i - \mathbb{E}[X])^2}{N-1} - \frac{N}{N-1}(\bar{X}_N - \mathbb{E}[X])^2 \tag{2.2}$$

The LGN shows that $\bar{X}_N \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}[X]$ and we deduce by applying again the LGN on the expression of $\hat{\sigma}_N^2$ obtained in (2.2), that $\hat{\sigma}_N^2 \xrightarrow{\mathbb{P}\text{-a.s.}} \sigma^2$. Finally, by (2.2),

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

■

Remark 2.7 *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 2.5 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 2.8. (►SLUTSKI'S LEMMA) *If*

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

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

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

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 2.9. ($\blacktriangleright\Delta$-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

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

■

2.2 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 2.5,

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

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 $a \approx 1.96$.

2.3 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.

2.4 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 3
Exact or approximate sampling

Contents

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

3.1 Exact Sampling

3.1.1 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 3.1. (▶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.

Proposition 3.2 *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:

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

Indeed if $u \leq F_Y(v)$, then by definition of F_Y^{-1}, we have $F_Y^{-1}(u) \leq v$. Assume now that $F_Y^{-1}(u) \leq v$. Applying the (nondecreasing) function F_Y on both sides yields $F_Y(F_Y^{-1}(u)) \leq F_Y(v)$. We thus only need to prove that $u \leq F_Y(F_Y^{-1}(u))$. By definition of the generalized inverse (and of the infimum) there exists a sequence (y_n) such that $y_n \to F_Y^{-1}(u)$, $y_n \geq F_Y^{-1}(u)$ and $u \leq F_Y(y_n)$, we then appeal to the right continuity of F_Y to get $u \leq F_Y(F_Y^{-1}(u))$. 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 3.3 *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

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

Example 3.4 (►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 3.5 *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 $F^{-1}(1-U) = -\frac{1}{\lambda}\ln(U) \sim \exp(\lambda)$.*

3.1.2 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 3.6 *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 $(u,x) \mapsto \mathbb{1}_{[0,1]}(u)g(x)$. This allows to write, knowing that the vectors $\begin{pmatrix} U_k \\ X_k \end{pmatrix}$ are i.i.d.,

$$\begin{aligned}
\mathbb{P}(Y \in A,\ \nu = k) &= \mathbb{P}(X_k \in A,\ \nu = k) \\
&= \mathbb{P}\left(U_1 > \frac{f(X_1)}{Mg(X_1)}, \ldots, U_{k-1} > \frac{f(X_{k-1})}{Mg(X_{k-1})},\ U_k \leq \frac{f(X_k)}{Mg(X_k)},\ X_k \in A\right) \\
&= \mathbb{P}\left(U_1 > \frac{f(X_1)}{Mg(X_1)}\right)^{k-1} \mathbb{P}\left(U_k \leq \frac{f(X_k)}{Mg(X_k)},\ X_k \in A\right) \\
&= \left(\iint_{\{u > \frac{f(x)}{Mg(x)}\}} \mathbb{1}_{[0,1]}(u)g(x)\ \mathrm{d}u\mathrm{d}x\right)^{k-1} \iint_{\{u \leq \frac{f(x)}{Mg(x)}\} \cap \{x \in A\}} \mathbb{1}_{[0,1]}(u)g(x)\ \mathrm{d}u\mathrm{d}x \\
&= \left(\int \left(1 - \frac{f(x)}{Mg(x)}\right) g(x)\ \mathrm{d}x\right)^{k-1} \int_{\{x \in A\}} \frac{f(x)}{Mg(x)} g(x)\ \mathrm{d}x \\
&= \left(1 - \frac{1}{M}\right)^{k-1} \frac{1}{M} \int_{\{x \in A\}} f(x)\mathrm{d}x
\end{aligned} \tag{3.1}$$

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, (3.1) shows that ν and Y are independent random variables. ■

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

Corollary 3.7 (►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

```
draw $X \sim g$, $U \sim \mathscr{U}[0.1]$ independently
while $U \geq \frac{f(X)}{Mg(X)}$ do draw $X \sim g$, $U \sim \mathscr{U}[0.1]$ independently
end while
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 $\mathbb{G}\mathrm{eom}(M^{-1})$. 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!

3.1.3 Sampling from a conditional distribution

3.1.3.1 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 $[a,b]$.

3.1.3.2 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

```
draw $U \sim \mathscr{U}[0,1]$
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 $F_X(x) = F_X(a) + u\,[F_X(b) - F_X(a)]$, we deduce that G has a generalized inverse which is:

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

This completes the proof.

3.1.4 Other sampling methods

3.1.4.1 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 3.8 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 3.9 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.

3.1.4.2 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 $\boxed{\mathscr{N}(\mu, \underbrace{\Sigma}_{AA^T})}$.

3.1.4.3 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*.

3.2 Approximate sampling

3.2.1 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

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

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!

3.2.1.1 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 3.10

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

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;

$$\left(\int f(x)|h(x)|\mathrm{d}x\right)^2 = \left(\int \frac{f(x)}{\sqrt{g(x)}}|h(x)|\ \sqrt{g(x)}\mathrm{d}x\right)^2 \le \left(\int \frac{f^2(x)}{g(x)}|h(x)|^2\ \mathrm{d}x\right)\left(\underbrace{\int(\sqrt{g(x)})^2\mathrm{d}x}_{1}\right) = \int \frac{f(x)}{g(x)}h^2(x)f(x)\mathrm{d}x \tag{3.2}$$

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

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

This shows (3.2). 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

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

■

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.

3.2.2 *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.

3.3 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 4
Discretization of SDE

Contents

Keywords: *Euler scheme, discretization of SDE, Brownian bridge, sampling of the maximum of a Brownian bridge.*

So far, we have studied sampling methods for evaluating expectations of random variables using Monte Carlo methods. This is only possible when one is able to sample according to the distribution of these random variables. The quantities that appear in finance are often expectations of the form $\mathbb{E}(h(X_t))$ or $\mathbb{E}(h(X_t, 0 \leq t \leq T))$ where $(X_t)_{t \geq 0}$ is a solution of the SDE

$$\mathrm{d}X_t = \mu(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W_t, \quad X_0 = x.$$

As we do not always know the distribution of (X_t) (except in some very specific cases as for the Black and Scholes model), we will see how to approximate this process by a series of r.v through a discretization procedure. To estimate $\mathbb{E}(h(X_T))$ or $\mathbb{E}(h(X_t, 0 \leq t \leq T))$, the practitioner will therefore be confronted with two types of error: a Monte Carlo error and a SDE discretization error. First, we will look at the sampling from the trajectory of a Brownian motion over a time interval $[0,T]$ for a given T.

Before going into the details of this chapter, we recall some properties concerning the conditional expectation (and consequently concerning the conditional distribution). Let U, V and W be three random vectors taking values in $\mathbb{R}^k$, $\mathbb{R}^\ell$ and $\mathbb{R}^n$ respectively. Assume that (V,W) have the following diametrically opposed behavior with respect of U:

- $V = f(U)$ where f is measurable ”deterministic” function,
- W et U are independent, i.e. $W \perp\!\!\!\perp U$.

Then,

$$\mathbb{E}\left[h(V,W) \,|\, U\right] = \int h(V,w)\mu_W(\mathrm{d}w) \tag{4.1}$$

where μ_W is the (unconditional) distribution of W. The message of this equation is that V is unchanged since it is a deterministic function of U, while W has a distribution conditionally on U which is the ”unconditional” distribution of W. Many calculations on the conditional expectation boil down to (4.1).

4.1 Sampling from a Brownian motion

4.1.1 Exact sampling with a fixed grid

Let $(W_t)_{0\leq t\leq 1}$ be a standard Brownian motion on $[0,T]$. Lt $0=t_0<t_1<\ldots<t_n=T$ be a grid on the interval $[0,T]$. We wish to sample a trajectory of the Brownian motion on each point of the subdivision that is, we wish to determine the law of the discrete time process $(W_{t_i})_{i=0,\ldots,n}$.

Algorithm 4 Brownian motion algorithm

```
Draw $G_i \sim \mathscr{N}(0,1)$ for $i=1,\ldots,n$ independently.
$X_0 = 0$.
for $i=1$ to $n$ do $X_i = X_{i-1} + \sqrt{t_i - t_{i-1}} G_i$
end for
Output: $X_0, \ldots X_n$
```

Proposition 4.1 *Let $(G_i)_{i=1,\ldots,n}$ be i.i.d. according to $\mathscr{N}(0,1)$. Define*

$$X_0 = 0, \quad X_i = \sum_{j=1}^{i} \sqrt{t_j - t_{j-1}} G_j \quad \textit{for } i > 0.$$

The vectors $(W_{t_0},\ldots,W_{t_n})$ and $(X_0,\ldots,X_n)$ have the same distribution:

$$(W_{t_0},\ldots,W_{t_n}) \overset{\mathscr{L}}{\equiv} (X_0,\ldots,X_n)$$

Proof. To prove the proposition, it is sufficient to note that $(X_1-X_0,\ldots,X_n-X_{n-1})$ is a gaussian vector with independent entries and such that for all $i\geq 1$, $X_i - X_{i-1} \sim \mathscr{N}(0, t_i - t_{i-1})$. As $X_0 = W_{t_0} = 0$, we deduce that the vectors $(X_0, X_1-X_0,\ldots,X_n-X_{n-1})$ and $(W_{t_0}, W_{t_1}-W_{t_0},\ldots,W_{t_n}-W_{t_{n-1}})$ have the same distribution and by deterministic mapping, the vectors $(W_{t_0},\ldots,W_{t_n})$ and $(X_0,\ldots,X_n)$ also have the same distribution. ■

The previous proposition allows to simulate exactly the Brownian motion on a fixed grid $0=t_0<t_1<\ldots<t_n=T$, in other words, **there is no discretization error for the Brownian sampling**, on the contrary to what will be seen for other processes.

4.1.2 Exact sampling conditionally on a grid

Now, assume we have already sampled $W_{t_1},\ldots,W_{t_n}$ and we are trying to refine the result by adding an extra term to the subdivision, i.e. we seek to simulate W_u (with $0<u<T$) conditionally on $(W_{t_1},\ldots,W_{t_n})$. In this case, we can show, if $t_i<u<t_{i+1}$, that the law of W_u conditionally on $(W_{t_1},\ldots,W_{t_n})$ only depends on (W_{t_i}, W_{t_i+1}), i.e.

$$W_u|_{(W_{t_1},\ldots,W_{t_n})} \overset{\mathscr{L}}{\equiv} W_u|_{(W_{t_i},W_{t_{i+1}})}$$

We start with a technical lemma on Gaussian variables which is extremely used (in this course or elsewhere!). It is a relatively simple lemma but one whose subtleties must be fully understood for the future.

Lemma 4.2 *For square integrable random variables, we denote by $<\cdot,\cdot>$ the scalar product defined by $<U,V>=\mathbb{E}(UV)$. We consider a centered gaussian vector (X,Y,Z). We write*

$$\mathscr{P}^Z_\perp[X,Y] = \hat{\alpha}X + \hat{\beta}Y\textit{: the orthogonal projection of } Z \textit{ on } \mathscr{H} = \{\alpha X + \beta Y; \alpha,\beta \ \in \mathbb{R}\}$$

Then, the distribution of Z conditionally on (X,Y) is given by:

$$\boxed{Z|_{(X,Y)} \sim \mathscr{N}\left(\ \mathscr{P}^Z_\perp[X,Y],\ \|Z-\mathscr{P}^Z_\perp[X,Y]\|^2\ \right)} \quad \textit{where } \|U\|^2 = <U,U> = \mathbb{E}(U^2).$$

Remark 4.3 *i) The expectation of this conditional distribution is $\mathscr{P}^Z_\perp[X,Y]$; it is thus a function of (X,Y) and therefore a* **random variable.**

ii) The variance of this conditional distribution could be (in general) a random variable but here, as stated by the lemma, it is equal to $\|Z-\mathscr{P}^Z_\perp[X,Y]\|^2$; it is therefore **deterministic** *it represents the error in the orthogonal projection of Z on $\mathscr{H}$. Finally, as all the r.v of this lemma are centered, we also have*

$$\|Z-\mathscr{P}^Z_\perp[X,Y]\|^2 = \mathbb{E}\left[(Z-\mathscr{P}^Z_\perp[X,Y])^2\right] - 0^2 = \mathbb{V}ar(Z-\mathscr{P}^Z_\perp[X,Y])$$

iii) These two remarks being said, we can rewrite the result of the lemma in a what that highlights what depends on the <u>values taken</u> by the r.v. (X,Y) et what depends on the <u>distribution</u> of (X,Y): denote by (x,y) the value taken by the couple of r.v. (X,Y), then (please, be careful to fully understand why we use the lowercase in this expression)

$$\boxed{Z|_{(X,Y)=(x,y)} \ \sim\ \mathscr{N}\left(\mathscr{P}^Z_\perp[x,y], \mathbb{V}ar\left(Z-\mathscr{P}^Z_\perp[X,Y]\right)\right)} \tag{4.2}$$

PROOF.

By definition, $\mathscr{P}^Z_\perp[X,Y]$ is measurable with respect to the σ-field generated by X,Y while $Z-\mathscr{P}^Z_\perp[X,Y]$ being orthogonal to $\mathscr{H}$ and X,Y being centered, we have $\mathbb{C}\text{ov}(Z-\mathscr{P}^Z_\perp[X,Y],X) = \mathbb{C}\text{ov}(Z-\mathscr{P}^Z_\perp[X,Y],Y) = 0$, the vector (X,Y,Z) being gaussian, it is also the case for $(X,Y,Z-\mathscr{P}^Z_\perp[X,Y])$. This finally shows that $Z-\mathscr{P}^Z_\perp[X,Y]$ is independent of (X,Y). We deduce

$$Z = \underbrace{\mathscr{P}^Z_\perp[X,Y]}_{\text{function of }(X,Y)} + \underbrace{(Z-\mathscr{P}^Z_\perp[X,Y])}_{\text{independent from }(X,Y)} \tag{4.3}$$

Furthermore, $Z-\mathscr{P}^Z_\perp[X,Y]$ is clearly a gaussian and centered r.v. Finally,

$$Z-\mathscr{P}^Z_\perp[X,Y] \sim \mathscr{N}(0,\underbrace{\mathbb{V}\text{ar}(Z-\mathscr{P}^Z_\perp[X,Y])}_{\|Z-\mathscr{P}^Z_\perp[X,Y])\|^2})$$

And (4.3) then implies $Z|_{(X,Y)} \sim \mathscr{N}\left(\ \mathscr{P}^Z_\perp[X,Y],\ \|Z-\mathscr{P}^Z_\perp[X,Y]\|^2\ \right)$ ■

Proposition 4.4 *If $t \le u \le T$, then denoting by (a,b) the value taken by the couple (W_t,W_T),*

$$W_u|_{(W_t,W_T)=(a,b)} \sim \mathscr{N}\left(\frac{T-u}{T-t}a + \frac{u-t}{T-t}b\ ,\ \frac{(T-u)(u-t)}{T-t}\right)$$

Sampling according this conditional distribution is given by the following algorithm 5.

Algorithm 5 Sampling of $W_u|_{(W_t,\ W_T)}$

```
Input: $W_t, W_T$
Sample $G \sim \mathscr{N}(0,1)$.
Set $W_u = \frac{T-u}{T-t}W_t + \frac{u-t}{T-t}W_T + \sqrt{\frac{(T-u)(u-t)}{T-t}}G$.
Output: $W_u$
```

PROOF. To simplify the notations, let us simply note $\hat{W}_u = \mathscr{P}^{W_u}_\perp[W_t,W_T]$ the orthogonal projection (in the sense of the scalar product $<\cdot,\cdot>$) of W_u on $\mathscr{H} = \{\alpha W_t + \beta W_T; \alpha,\beta \in \mathbb{R}\}$. Given the Lemma 4.2, we only need to identify the expression of $\hat{W}_u$ and the projection error $\|W_u - \hat{W}_u\|^2 = \mathbb{V}\text{ar}(W_u - \hat{W}_u)$.

We immediately note that $\left\{ \frac{W_t}{\|W_t\|}, \frac{W_T - W_t}{\|W_T - W_t\|} \right\}$ is an orthonormal base of $\mathscr{H}$, and the projection of W_u on this orthonormal basis can be written as:

$$\begin{aligned}\hat{W}_u &= < W_u, \frac{W_t}{\|W_t\|} > \frac{W_t}{\|W_t\|} + < W_u, \frac{W_T - W_t}{\|W_T - W_t\|} > \frac{W_T - W_t}{\|W_T - W_t\|} \\ &= \frac{< W_u, W_t >}{< W_t, W_t >} W_t + \frac{< W_u, W_T - W_t >}{< W_T - W_t, W_T - W_t >} (W_T - W_t) \\ &= W_t + \frac{u-t}{T-t}(W_T - W_t) = \frac{T-u}{T-t} W_t + \frac{u-t}{T-t} W_T\end{aligned}$$

where the penultimate line comes from $< W_a, W_b >= \mathbb{E}(W_a W_b) = \mathbb{C}\text{ov}(W_a, W_b) = a \wedge b$. Finally, the error associated to the orthogonal projection writes

$$\mathbb{E}\left[(W_u - \hat{W}_u)^2\right] = \mathbb{E}\left[\left(\frac{T-u}{T-t}(W_u - W_t) + \frac{u-t}{T-t}(W_u - W_T)\right)^2\right] = \left(\frac{T-u}{T-t}\right)^2 (u-t) + \left(\frac{u-t}{T-t}\right)^2 (T-u) = \frac{(T-u)(u-t)}{T-t}$$

This concludes the proof. ■

4.2 The Euler scheme

4.2.1 Principle of the discretization scheme

We intend here to approximate the solution of a Stochastic Differential Equation (SDE). Let $(X_t)_{t\geq 0}$ be a process taking values in $\mathbb{R}^d$, which is solution to the SDE:

$$X_t = X_0 + \int_0^t \mu(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}W_s \tag{4.4}$$

where $(W_t)_{t\geq 0}$ is a standard r-brownian vector-valued motion. On the contrary to the sampling from the brownian motion, which is an *exact* sampling, we will need here to use *approximate* sampling.

The most simple discretization is the *Euler scheme*. Choose a discretization step $\boxed{h = \frac{T}{N}}$ of the time interval $[0, T]$. The exact solution satisfies

$$X_h = X_0 + \int_0^h \mu(X_s)\mathrm{d}s + \int_0^h \sigma(X_s)\mathrm{d}W_s$$

A ”natural” approximation of X_h is the following

$$X_h \approx X_0 + \mu(X_0)h + \sigma(X_0)(W_h - W_0).$$

By induction, the Euler scheme associated to the SDE (4.4) is

$$X_0^N = X_0, \tag{4.5}$$

$$X_{(p+1)h}^N = X_{ph}^N + \mu(X_{ph}^N)h + \sigma(X_{ph}^N)(W_{(p+1)h} - W_{ph}) \tag{4.6}$$

To get the ideas right, let’s translate this into an algorithm.
The generalization of the Euler scheme to SDE where the drift volatility terms may depend on time

$$X_t = X_0 + \int_0^t \mu(s, X_s)\mathrm{d}s + \int_0^t \sigma(s, X_s)\mathrm{d}W_s$$

is immediate and is left to the reader.

Algorithm 6 Euler scheme

```
Input: $N, T, X_0$
Set $X(0) = X_0$ and $h = T/N$.
for $i = 1$ to $N$ do
    Draw $G \sim \mathcal{N}(0,1)$
    $X(i) = X(i-1) + \mu(X(i-1))h + \sigma(X(i-1))\sqrt{h}G$
end for
Output: $X(0), \ldots, X(N)$
```

4.2.2 Convergence properties

The following theorem, which we will admit, gives the convergence of the discretization scheme as well as an evaluation of the convergence speed of the approximation:

Theorem 4.5. *Assume that μ et σ are two lipshitz functions. Let $(W_t)_{t\geq 0}$ be a standard r-vector valued brownian motion. Denote by $(X_t)_{t\geq 0}$ the unique solution to the SDE:*

$$X_t = X_0 + \int_0^t \mu(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}W_s$$

and denote by $(X_{kh}^N)_{k\geq 0}$ the sequence of r.v. defined by (4.5). *We then have the following results:*

STRONG CONVERGENCE- *There exists β such that for all $q \geq 1$,*

$$\left(\mathbb{E}\left(\sup_{0\leq k\leq N} |X_{kh}^N - X_{kh}|^{2q}\right)\right)^{\frac{1}{2q}} \leq \beta\sqrt{h}.$$

Moreover, for all $\alpha < 1/2$,

$$\frac{1}{h^\alpha} \sup_{0\leq k\leq N} |X_{kh}^N - X_{kh}| \xrightarrow[h\to 0]{\mathbb{P}\text{-}a.s.} 0$$

WEAK CONVERGENCE- *If μ and σ are C^4-function with bounded derivatives (up to the order 4) and if f is a C^4-function with derivatives (up to the order* 4*) are bounded by a polynomial, then if $\boxed{h = T/N}$, there exists a constant C_T such that*

$$\left|\mathbb{E}(f(X_T^N)) - \mathbb{E}(f(X_T))\right| \leq C_T(f)h$$

n

Remark 4.6 *This theorem shows that the convergence speed is $h^{1/2}$ (take $q = 1$ in the* Strong Convergence *section) and the almost sure convergence speed is $h^{1/2-\varepsilon}$ for any $\varepsilon > 0$. Moreover, for very regular functions, the speed of low convergence (which is mainly of interest to practitioners because it shows the convergence of the "discretized" price towards the real price) is in the order of h.*

4.2.2.1 Other discretization methods

Euler's scheme is obtained by a first-order approximation in the Taylor expansion. There are other schemes like the Milshtein schema, whose very crude principle is to push development to a higher order. Its efficiency is real in dimension 1, i.e. when (X_t) is a process with real values. Nevertheless in larger dimension, its effectiveness compared to Euler's scheme becomes more questionable. The principle and certain properties of this type of scheme can easily be found in the literature, but we will not study it in this course.

Another very simple method to improve approximation performance is known as the *Romberg* method. It is based on a very simple observation: let us suppose that U_N is a sequence admitting the following expansion:

$$U_N = a + \frac{b}{N} + \frac{c}{N^2} + \dots$$

Then

$$2U_N - U_{N/2} = a + \frac{d}{N^2} + \dots$$

and the $1/N$ term disappeared... we have an improvement by one order... The Romberg method therefore consists in replacing $f(X_t^N)$ by $2f(X_t^N) - f(X_t^{N/2})$ and we can then show that it allows to discard an order in the estimation of $\mathbb{E}(f(X_t))$.

4.3 Sampling a maximum conditionally on an Euler scheme

In financial applications, Euler's discretization will approximate payoff of a quantity $f(X_T)$ where T is the final instant and $(X_t)_{t\geq 0}$ is the price process of an asset satisfying an SDE like (4.4). We can also be interested in options on the maximum as $f(X_T, M_T)$ where $M_T = \sup_{0\leq t\leq T} X_t$. In this case, to approximate M_T, the naïve approach of taking the maximum on the discretization grid in an Euler scheme may prove to be relatively ineffective. However, we have seen that we could easily calculate the law of W_u conditionally to extreme values of the Brownian. We will express this result in terms of a process distribution and not in terms of a marginal distribution like that of W_u; it is the "celebrated" *Brownian bridge* that we will define first. Then we will see that the *distribution of the maximum* of a Brownian bridge is perfectly known and easy to sample from. Finally, we will see how these properties will allow us to simulate the distribution of the *maximum of a process conditional on Euler's discretization scheme*.

In what follows, we will place ourselves in a space of dimension 1, i.e., the brownian motions will be considered in dimension 1, and the process (X_t) will be a real process.

4.3.1 Brownian bridge

Set

$$\boxed{Z_u^{a,b} = \left(\frac{T-u}{T-t}a + \frac{u-t}{T-t}b\right) + (W_u - \hat{W}_u)} \quad \text{where} \quad \boxed{\hat{W}_u := \frac{T-u}{T-t}W_t + \frac{u-t}{T-t}W_T.}$$

It is useful (and enlightening) to note that $Z_u^{a,b}$ and W_u can also be written as

$$Z_u^{a,b} = \mathscr{P}_\perp^{W_u}[a,b] + W_u - \mathscr{P}_\perp^{W_u}[W_t, W_T], \tag{4.7}$$

$$W_u = \mathscr{P}_\perp^{W_u}[W_t, W_T] + W_u - \mathscr{P}_\perp^{W_u}[W_t, W_T] \tag{4.8}$$

where $\mathscr{P}_\perp^{U}[W_t, W_T]$ is the orthogonal projection of U on the subspace of linear combinations of W_t and W_T and $\mathscr{P}_\perp^{U}[a,b]$ stands for the value taken by the r.v. $\mathscr{P}_\perp^{U}[W_t, W_T]$ when (W_t, W_T) takes the value (a,b): $\mathscr{P}_\perp^{U}[a,b] = \mathscr{P}_\perp^{U}[W_t, W_T]|_{(W_t,W_T)=(a,b)}$. To simplify the notation, we have discarded the dependence of t and T in the notation $(Z_u^{a,b})$.

Considering (4.7) and (4.8), the expression of the brownian bridge is "natural"!!! Indeed, clearly:

$$W_u|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} Z_u^{a,b}|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} Z_u^{a,b}$$

where the last equality follows from the fact that $W_u - \mathscr{P}_\perp^{W_u}[W_t, W_T]$ being orthogonal to (W_t, W_T) and these variables being centered, we have that $W_u - \mathscr{P}_\perp^{W_u}[W_t, W_T]$ and (W_t, W_T) are uncorrelated, which explains that $Z^{a,b}$ is independent from (W_t, W_T) since all these random variables form a gaussian vector... Now the little extra requirement is to show equality in law not only for the marginal W_u but for the whole process (W_u). Why searching for such a result? By anticipating, it is because we are interested in a maximum and $\sup_{u\in[t,T]} W_u$ involves the whole process (W_u) and not only a marginal part of the process...

Proposition 4.7 BROWNIAN BRIDGE- *The process $(Z_u^{a,b})_{t\le u\le T}$ is a gaussian process, that satisfies the terminal conditions*

$$Z_t^{a,b} = a, \quad Z_T^{a,b} = b$$

Moreover, the distribution of the process $(W_u)_{t\le u\le T}$ conditionally to $(W_t, W_T) = (a,b)$ and the distribution of $(Z_u^{a,b})_{t\le u\le T}$ are the same:

$$(W_u)_{t\le u\le T}|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} (Z_u^{a,b})_{t\le u\le T}$$

PROOF. It is absolutely clear from the definition of $(Z_u^{a,b})$ and from classical properties of the brownian motion $(W_u)_{u\ge 0}$ that $(Z_u^{a,b})_{t\le u\le T}$ is a gaussian process satisfying the terminal conditions of the proposition. To obtain the last part of the proposition, we need to show the equality between the characteristic functions of finite dimensional distributions, the continuity of the two processes will then allow to conclude with the equality in distribution between the two processes: for all grid $t_1 \le \ldots \le t_N$ de $[t,T]$, by setting $\mathbf{W} = \begin{pmatrix} W_{t_1} \\ \ldots \\ W_{t_n} \end{pmatrix}$ et $\mathbf{Z} = \begin{pmatrix} Z_{t_1}^{a,b} \\ \ldots \\ Z_{t_n}^{a,b} \end{pmatrix}$, let us show that for all vector $v \in \mathbb{R}^n$,

$$\mathbb{E}\left(\mathrm{e}^{\mathrm{i}v^T\mathbf{W}}\middle| W_t = a, W_T = b\right) = \mathbb{E}\left(\mathrm{e}^{\mathrm{i}v^T\mathbf{Z}}\right)$$

To this aim, let us show the following equality in distribution: $v^T\mathbf{W}|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} v^T\mathbf{Z}$. To see this, note that by (4.7), (4.8) and by linearity of the orthogonal projection, we have

$$\begin{aligned} v^T\mathbf{Z} &= \mathscr{P}_\perp^{v^T\mathbf{W}}[a,b] + v^T\mathbf{W} - \mathscr{P}_\perp^{v^T\mathbf{W}}[W_t, W_T] \\ v^T\mathbf{W} &= \mathscr{P}_\perp^{v^T\mathbf{W}}[W_t, W_T] + v^T\mathbf{W} - \mathscr{P}_\perp^{v^T\mathbf{W}}[W_t, W_T] \end{aligned}$$

Therefore $v^T\mathbf{W}|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} v^T\mathbf{Z}|_{(W_t,W_T)=(a,b)} \overset{\mathscr{L}}{\equiv} v^T\mathbf{Z}$. The proof is completed. ■

In what follows, we consider a brownian motion with terminal conditions $W_t = 0$ et $W_T = w$ and we assume $t = 0$. In such a case, $Z_u^{0,w}$ writes: $W_u - \frac{u}{T}(W_T - w)$.

Proposition 4.8 BROWNAN BRIDGE STARTING FROM 0- *For all $w \in \mathbb{R}$, set*

$$\boxed{Z_u^w = W_u - \frac{u}{T}(W_T - w)}$$

Then, (Z_u^0) is a gaussian process independent from W_T. Moreover,

$$\begin{aligned} &\forall u \in [0,T] \ , \ \mathbb{E}(Z_u^0) = 0 \\ &\forall u, v \in [0,T] \ , \ \mathbb{E}(Z_u^0 Z_v^0) = u \wedge v - \frac{uv}{T} \end{aligned}$$

The distribution of the process (W_s) conditionally on W_T taken on $W_T = w$ is the one of the process (Z_s^w):

$$(W_s)_{0\le s\le T}|_{W_T=w} \overset{\mathscr{L}}{\equiv} (Z_s^w)_{0\le s\le T}$$

We say that (Z_s^w) is the brownian bridge process starting from 0 and taking the value w at T.

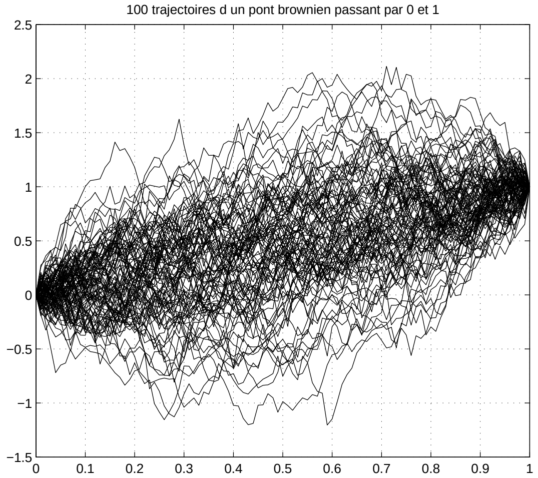


Fig. 4.1 Trajectories of a Brownian bridge starting from 0 at $t=0$ and taking the value 1 at $t=1$

PROOF. The proof follows directly from the previous proposition... It remains to show that $\mathbb{E}(Z_u^0 Z_v^0) = u \wedge v - \frac{uv}{T}$ but this is immediate by writing Z_t^0 as a function of the brownian motion W_t and by using$\mathbb{E}(W_s W_t) = s \wedge t$ (please check it carefully!). ■

Proposition 4.9 THE DISTRIBUTION OF THE MAXIMUM OF A BROWNIAN BRIDGE- *The cumulative distribution function of* $\sup_{t\in[0,T]} Z_t^w$ *writes: for all* $a > w$,

$$\mathbb{P}\left[\sup_{t\in[0,T]} Z_t^w \le a\right] = 1 - \mathrm{e}^{-\frac{2}{T}a(a-w)} = F_{w,T}(a)$$

One can draw $M = \sup_{t\in[0,T]} Z_t^w$ *by inverting the cumulative distribution function: Algorithm 7*

Algorithm 7 Maximum of a brownian bridge

```
Input: $T, w$
Draw $U \sim \mathscr{U}(0,1)$
Set $Y = F_{w,T}^{-1}(1-U) = \frac{1}{2}(w + \sqrt{w^2 - 2T\ln U})$
Output: $Y$
```

Exercise 4.10. Propose a way to draw $\sup_{t\in[0,T]} W_t$ by marginalisation.

4.3.2 Exact sampling from the distribution of a maximum conditionally on an Euler scheme

Recall that we consider $[0,T]$ and that the discretization step is $h = T/N$. Set $t_\ell = \ell h$ and consider $(\bar{X}_t^N)_{0\le t\le T}$ the process defined on the intervals $[t_k, t_{k+1}]$ by

$$\forall 0 \le u \le h, \quad \bar{X}^N_{t_k+u} = X^N_{t_k} + \mu(X^N_{t_k})\,u + \sigma(X^N_{t_k})\underbrace{(W_{t_k+u} - W_{t_k})}_{B^k_u}$$

where $(B^k_u)_{u\ge 0}$, defined by $B^k_u = W_{t_k+u} - W_{t_k}$, is a standard brownian motion. We can see clearly that the process $(\bar{X}^N_t)$ takes at t_k the same values as the Euler scheme, i.e. $\bar{X}^N_{t_k} = X^N_{t_k}$. We now want to draw the distribution of the maximum of $(\bar{X}^N_t)$ conditionally to $(X^N_{t_k})_{0\le k\le N}$.

The first remark is that we can only consider an interval $[t_k, t_{k+1}]$ by noting that

$$\sup_{t\in[0,T]} \bar{X}^N_t = \max_{k=0,\dots,N-1}\left[\sup_{t\in[t_k,t_{k+1}]} \bar{X}^N_t\right]$$

Finally, note that

$$\sup_{t\in[t_k,t_{k+1}]} \bar{X}^N_t|_{(\bar{X}^N_{t_\ell})_{0\le \ell\le N}} \overset{\mathscr{L}}{\equiv} \sup_{t\in[t_k,t_{k+1}]} \bar{X}^N_t|_{\bar{X}^N_{t_k},\bar{X}^N_{t_{k+1}}}$$

Lemma 4.11 *The distribution of $(\bar{X}^N_u)$ on the interval $[t_k, t_{k+1}]$ conditionally on the terminal values $(\bar{X}^N_{t_k}, \bar{X}^N_{t_{k+1}}) = (x_k, x_{k+1})$ is linked to a brownian bridge $Z^{\bar{w}}_u$ taking the value $\bar{w} = \frac{x_{k+1}-x_k}{\sigma(x_k)}$ at time h by the following equality in distribution:*

$$(\bar{X}^N_{t_k+u})_{0\le u\le h}|_{(\bar{X}^N_{t_k},\bar{X}^N_{t_{k+1}})=(x_k,x_{k+1})} \overset{\mathscr{L}}{\equiv} (x_k + \sigma(x_k)Z^{\bar{w}}_u)_{0\le u\le h}$$

PROOF. If $(\bar{X}^N_{t_k}, \bar{X}^N_{t_{k+1}}) = (x_k, x_{k+1})$, then B^k_h takes the value $w = \frac{x_{k+1}-x_k-\mu(x_k)h}{\sigma(x_k)}$. Therefore, by the crucial property of the brownian bridge,

$$\begin{aligned}(\bar{X}^N_{t_k+u})_{0\le u\le h}|_{(\bar{X}^N_{t_k},\bar{X}^N_{t_{k+1}})=(x_k,x_{k+1})} &\overset{\mathscr{L}}{\equiv} \left(x_k + u\mu(x_k) + \sigma(x_k)\left[B^k_u - \frac{u}{h}(B^k_h - w)\right]\right)_{0\le u\le h}\\ &\overset{\mathscr{L}}{\equiv} \left(x_k + \sigma(x_k)\left[B^k_u - \frac{u}{h}\left(B^k_h - \frac{x_{k+1}-x_k}{\sigma(x_k)}\right)\right]\right)_{0\le u\le h}\end{aligned}$$

where the last equality is obtained by plugging $w = \frac{x_{k+1}-x_k-\mu(x_k)h}{\sigma(x_k)}$ into the previous equality. The proof is completed. ■

Therefore, by Proposition 4.9, it is possible to sample $M_k = \sup_{0\le u\le h} \bar{X}^N_{t_k+u}$ conditionally on $(\bar{X}^N_{t_k}, \bar{X}^N_{t_{k+1}}) = (x_k, x_{k+1})$ in the following way:

i) Draw $U \sim \mathscr{U}]0,1[$
ii) Set $M_k = x_k + \sigma(x_k)F^{-1}_{\bar{w},h}(1-U)$ with $\bar{w} = \frac{x_{k+1}-x_k}{\sigma(x_k)}$ and $F^{-1}_{\bar{w},h}$ defined in the Proposition 4.9.

Algorithm 8 Maximum conditionally to an Euler scheme

```
Input: $x_0, \dots, x_N$
for $k = 0$ to $N-1$ do
    Draw $U \sim \mathscr{U}(0,1)$
    Set $M(k) = \frac{1}{2}\left(x_{k+1} + x_k + \sqrt{(x_{k+1}-x_k)^2 - 2\sigma^2(x_k)h\ln(U)}\right)$
end for
Set $M = \max_{k=1,\dots,N} M(k)$
Output: $M$
```

By straightforward algebra,

$$x_k + \sigma(x_k)F^{-1}_{\bar{w},h}(1-U) = \frac{1}{2}\left(x_{k+1} + x_k + \sqrt{(x_{k+1}-x_k)^2 - 2\sigma^2(x_k)h\ln(U)}\right)$$

Finally, all these results show that we can draw the distribution of the maximum of $(\bar{X}^N_t)$ conditionally to $(\bar{X}^N_{t_k})_{0\le k\le N}$ by Algorithm 8.

4.4 Take-home message

You should
a) Know how to implement the Euler scheme and know the approximation results
b) Know how to add an intermediate point in an Euler scheme.
c) Understand the construction of a brownian bridge and the applications to sample the maximum of an Euler scheme.

4.5 Highlights

4.5.0.1 Leonhard Euler (15 april 1707 - 18 september 1783) Source: bibmath.

Leonhard Euler; 15 April 1707 – 18 September 1783) was a Swiss mathematician, physicist, astronomer, logician and engineer, who made important and influential discoveries in many branches of mathematics, such as infinitesimal calculus and graph theory, while also making pioneering contributions to several branches such as topology and analytic number theory. He also introduced much of the modern mathematical terminology and notation, particularly for mathematical analysis, such as the notion of a mathematical function. He is also known for his work in mechanics, fluid dynamics, optics, astronomy, and music theory.

Euler was one of the most eminent mathematicians of the 18th century and is held to be one of the greatest in history. He is also widely considered to be the most prolific mathematician of all time. His collected works fill 60 to 80 quarto volumes, more than anybody in the field. He spent most of his adult life in Saint Petersburg, Russia, and in Berlin, then the capital of Prussia.

A statement attributed to Pierre-Simon Laplace expresses Euler's influence on mathematics: ”Read Euler, read Euler, he is the master of us all.”

Chapter 5
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.

5.1 Importance Sampling

We refer the reader to Section (3.2.1) 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 3, we will focus on the non-asymptotic variance (instead of the asymptotic variance) of $\bar{S}_n$.

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

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

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

5.2 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

$$\bar{S}_n = \frac{1}{2n}\sum_{i=1}^{n}(h(U_i)+h(1-U_i)) \tag{5.1}$$

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

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

If $\mathbb{C}\text{ov}(h(U),h(1-U))$ is negative, then the speed of convergence of $\bar{S}_n$ 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 5.1 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 5.2 *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 5.3 *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 5.1 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 5.4. 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 $\frac{1}{n} \sum_{i=1}^{n} h(Y_i)$.

5.3 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 $\boxed{\mathbb{V}\mathrm{ar}(Y - \alpha X) \leq \mathbb{V}\mathrm{ar}(Y)}$ 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 $\bar{S}_n \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}(Y)$. Now, setting $S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$

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

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.

5.3.0.1 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 5.5 *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 5.6 *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* (5.1)*:*

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

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

Exercise 5.7. THE CALL-PUT PARITY- In a simplified manner, the payoff of a selling option (a Call) can be written as $(\lambda \mathrm{e}^{\sigma G} - K)_+$ where $G \sim \mathscr{N}(0,1)$. Noting that $(\lambda \mathrm{e}^{\sigma G} - K)_+ - (K - \lambda \mathrm{e}^{\sigma G})_+ = \lambda \mathrm{e}^{\sigma G} - K$, propose a control variate technique for the calculation of $\mathbb{E}(\lambda \mathrm{e}^{\sigma G} - K)_+$.

Exercise 5.8. 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 5.9. 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.

5.4 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:

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

Remark 5.10 *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 5.11. 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.

5.5 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) $Y_{i,j} \overset{\mathscr{L}}{\equiv} Y|_{Y\in S_i}$ 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

$$\mathbb{V}\mathrm{ar}(\bar{S}_n) = \sum_{i=1}^{p} \alpha_i^2 \frac{\sigma_i^2}{n_i}, \quad \mathbb{V}\mathrm{ar}(S_n) = \frac{1}{n}\left[\underbrace{\mathbb{E}[h^2(Y)]}_{\sum_{i=1}^{p}\alpha_i(\sigma_i^2+\mu_i^2)} - \left(\underbrace{\mathbb{E}[h(Y)]}_{\sum_{i=1}^{p}\alpha_i\mu_i}\right)^2\right] \tag{5.2}$$

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?

5.5.0.1 Proportional allocation

Chooose $\boxed{\frac{n_i}{n} = \alpha_i}$. Then, (5.2) writes

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

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 $\{1,\ldots,p\}$. The stratified sampling thus induces a lower variance in the case of proportional allocation.

5.5.0.2 Optimal allocation

We now aim at finding the optimal $(n_i^\star)$ such that (see (5.2)),

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

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 5.12 *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 3.10). 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.

5.6 Quasi Monte Carlo methods

5.6.1 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 5.13. "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 5.14 *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

$$\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbb{E}(h(U))\right| \le c(h)\frac{(\ln n)^d}{n} \tag{5.3}$$

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 $1/\sqrt{n}$.

5.6.1.1 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 $(a_{i,n})_{i\ge 0}$ 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 $(x_n^p)_{n\ge 0}$ 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.

5.6.1.2 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 9 Calcul de x_n^p

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

Remark 5.15 *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* (5.3) *gives an upper-bound for the error on the condition that we know exactly* $c(h)$*, which is not the case in general.*

5.7 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.

5.8 Highlights

5.8.0.1 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.

Index