Stochastic Simulation and Monte Carlo Methods

Olivier Wintenberger

2021–2022

Contents

Part I

Preliminaries

Chapter 1

Conditional expectation

These lecture notes are preceded by some preliminaries on the important notion of conditional expectation, useful for the rest of the course.

1.1 Conditioning as a projection

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triplet. The sample space Ω is the space of any outcomes, the event space $\mathcal{F}$ constitutes a σ-algebra of events A (closed under complement and countable union). The probability function $\mathbb{P}$ that assigns to each event $A \in \mathcal{F}$ a probability $0 \leq \mathbb{P}(A) \leq 1$.

The probability function must satisfies the countable additivity assumption meaning that for every countable union of disjoint sets $A_n \in \mathcal{F}$, $n \geq 1$, we have

$$\mathbb{P}(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n). \tag{1.1}$$

We move to the definition of the random variable (rv).

Definition 1. *A random variable is a function $X\colon \Omega \mapsto \mathbb{R}$ such that is measurable:*

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega\,;\ X(\omega) \leq x\} \in \mathcal{F}\,.$$

We have

$$\mathbb{P}(X^{-1}((-\infty, x])) = \mathbb{P}(\{\omega \in \Omega\,;\ X(\omega) \leq x\}) = \mathbb{P}(X \leq x) = F_X(x).$$

The function F_X is called the cumulative distribution function of X. In these notes, it will be convenient to work on F_X rather than $\mathbb{P}$ as much as possible. It is possible since

Proposition. *The distribution of the rv X is characterized by F_X.*

The cdf F_X is a cadlag (continue à droite, limite à gauche) function.

We define $F_X(a,b] = F_X(b) - F_X(a)$ and extend the measure $F_X(A)$ to any Borel set $A \subset \mathbb{R}$ using the relation (1.1).

The expectation $\mathbb{E}[h(X)]$ is the Lebesgue integral

$$\int_{\mathbb{R}} h(x) dF_X(x) = \lim_{n\to\infty} \int_{\mathbb{R}} h_n(x) dF_X(x) = \lim_{n\to\infty} \sum_{j=1}^{n} c_j F_X(A_n)$$

where h is any positive measurable function approximated by $h_n = \sum_{j=1}^n c_j\, \mathbb{1}_{A_n}$ with Borel sets A_n and $c_j > 0$.

We can extend the notion of expectation to any integral function h such that $\mathbb{E}[|h(X)|] < \infty$ by considering the positive and negative parts of h. In particular from F_X one can compute the moments such as

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 dF_X(x)\,.$$

This second order moments can be infinite!

The space $\mathbb{L}^2(\mathbb{R})$ gather the rv with finite second order moments.

Definition 2. *The space $\mathbb{L}^2(\mathbb{R})$ is defined as the set of square integrable rv:*

$$\mathbb{L}^2(\mathbb{R}) = \{X \ rv \ on \ \mathbb{R} : \ \mathbb{E}[X^2] < \infty\}\,.$$

We say that $X = Y$ a.s. iff $\mathbb{P}(X = Y) = 1$

The right notion for considering $\mathbb{L}^2$ is the one of Hilbert space:

Definition 3. *An Hilbert space is a complete vector space equipped with the scalar product $< \cdot, \cdot >$.*

Any finite dimensional vector space $\mathbb{R}^d$ is an Hilbert space. The scalar product is

$$< x, y >= \sum_{i=1}^{d} x_i y_i\,, \qquad x, y \in \mathbb{R}^d.$$

An Hilbert space might have infinite dimension. Indeed $\mathbb{L}^2$ is an example

Proposition. *The space $\mathbb{L}^2(\mathbb{R})$ is an Hilbert space equipped with $< X, Y >= \mathbb{E}[XY]$.*

On any Hilbert space H we define the notion of orthogonal projection. Let $\|x\|^2 =< x, x >$, $x \in H$, H an Hilbert space.

Proposition. *Let L be any closed sub-vector space of H. There exists a unique $\pi_L(x) \in L$ such that $\|x - z\| \geq \|x - \pi_L(x)\|$ for any $z \in H$. We have $< x - \pi_L(x), z >= 0$ for any $z \in L$ and $\pi_L(x)$ is the orthogonal projection of x on L. Finally the Pythagorean theorem applies*

$$\|z - x\|^2 = \|z - \pi_L(x)\|^2 + \|\pi_L(x) - x\|^2\,, \qquad z \in L\,.$$

The conditional expectation can be seen as a projection. Let $H = \mathbb{L}^2(\mathbb{R})$ and $L = \{h(Y) : h \text{ Borel function and } \mathbb{E}[h(Y)^2] < \infty\}$ for some rv Y. Then

Proposition. *The space L is a closed sub-vector space of $\mathbb{L}^2(\mathbb{R})$ and the orthogonal projection $\pi_L(X)$ is called the conditional expectation of X on Y and it is denoted $\mathbb{E}[X \mid Y]$.*

Remark. *There exists a measurable function h^* such that $\mathbb{E}[X \mid Y] = h^*(Y)$ a.s.*

We have $\mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] \leq \mathbb{E}[(X - h(Y))^2]$ for any h measurable.

The properties of the projection are inherited by the conditional expectation:

Proposition. *The conditional expectation $\mathbb{E}[X \mid Y]$ is a rv such that $\mathbb{E}[(\mathbb{E}[X \mid Y])^2] < \infty$.*

We have the tower property $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ and $Var(\mathbb{E}[X \mid Y]) \leq Var(X)$.

For any measurable function h we have $\mathbb{E}[h(Y)(X - \mathbb{E}[X \mid Y])] = 0$.

Note that one can extend the previous notions to $\mathbb{R}^d$: $X \in \mathbb{L}^2(\mathbb{R}^d)$ is the space of square integrable random vectors such that $\mathbb{E}[\|X\|^2] < \infty$, the conditional expectation $\mathbb{E}[X \mid Y]$ is a random vector in $\mathbb{L}^2(\mathbb{R}^d)$ such that $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ and

$$\mathbb{E}[\|\mathbb{E}[X \mid Y] - \mathbb{E}[X]\|^2] \le \mathbb{E}[\|X - \mathbb{E}[X]\|^2].$$

However the notion of norm in $\mathbb{L}^2(\mathbb{R}^d)$ is of different nature than the notion of variance of X that is a $d \times d$ matrix

$$\mathrm{Var}\,(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T].$$

1.2 Conditioning as an integration

Let X be a rv with cdf F_X. We say that X is continuous (absolutely continuous wrt the Lebesgue measure) if F_X admits a density f_X so that

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)dF_x(x) = \int_{\mathbb{R}} h(x)f_X(x)dx\,,$$

for any positive measurable function h.

Remark. *We will extend the notion of density to cdf absolutely continuous wrt other measures: X is a discrete rv if she takes value in $\{x_i\}_{i\in\mathbb{N}}$, then it admits a density f_X with respect to the counting measure $\nu = \sum_{i\in\mathbb{N}} \varepsilon_{\{x_i\}}$ where $\varepsilon_{\{x\}}(A) = 1$ if $x \in A$, $= 0$ else. We have $f_X(x_i) = \mathbb{P}(X_i = x_i)$, $i \in \mathbb{N}$ and*

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)dF_X(x) = \sum_{x\in\{x_i, i\in\mathbb{N}\}} h(x)f_X(x) = \int_{\mathbb{R}} h(x)f_X(x)d\nu(x)\,.$$

The density satisfies $f_X(x) > 0$ whenever $x \in \mathrm{Supp}(X)$ where the support $\mathrm{Supp}(X)$ is the Borel set such that

$$\mathbb{P}(X \in \mathrm{Supp}(X)) = 1, \qquad \mathbb{P}(X \notin \mathrm{Supp}(X)) = 0\,.$$

Moreover $\int_{\mathrm{Supp}(X)} f_X(x)d\nu(x) = 1$. By convention $f_X(x) = 0$ for $x \notin \mathrm{Supp}(X)$.

Proposition. *Any positive function on a Borel set S that sum up to 1 is a density (extended on $\mathbb{R}$ by being $= 0$ elsewhere) of a rv X so that Supp$(X) = S$.*

Let X and Y admitting densities f_X and f_Y (wrt to 2 measures ν_1 and ν_2 on $\mathbb{R}$). Assume that $(X, Y) \in \mathbb{R}^2$ admits a density (wrt the product measure $\nu_1\nu_2$).

Definition 4. *The conditional density of X given $Y = y$, $y \in$ Supp(Y), is defined as*

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}\,.$$

The function $f_{X|Y=y}$ is a density since $\int_{\mathbb{R}} f_{X,Y}(x, y)d\nu_1(x) = f_Y(y)$ and its support is included in $\mathrm{Supp}(X)$.

It is not a density wrt y!

Definition 5. *The rv $X \mid Y = y$ is defined by its density $f_{X|Y=y}$.*

Assuming that it exists, we define $\mathbb{E}[X \mid Y = y]$ as the expectation of $X \mid Y = y$:

$$\mathbb{E}[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y=y}(x) d\nu_1(x) \,.$$

The conditional expectation $\mathbb{E}[X \mid Y = y]$ is not a rv!
It is deterministic.

For any Borel set A such that $\mathbb{P}(Y \in A) \neq 0$ and $\mathbb{E}[X \, \mathbb{I}_{Y\in A}] < \infty$, we extend the previous notions to

$$f_{X|Y\in A}(x) = \frac{\int_A f_{X,Y}(x,y) d\nu_2(y)}{\int_A f_Y(y) d\nu_2(y)}$$

and

$$\mathbb{E}[X \mid Y \in A] = \int_{\mathbb{R}} x f_{X|Y\in A}(x) d\nu_1(x) \,.$$

We have

$$\mathbb{E}[X \mid Y \in A] = \frac{\mathbb{E}[X \, \mathbb{I}_{Y\in A}]}{\mathbb{P}(Y \in A)} \,.$$

If $\mathbb{E}[|X|] < \infty$ we denote $h^*(y) = \mathbb{E}[X \mid Y = y]$ for all $y \in Y$. Then $\mathbb{E}[X \mid Y] = h^*(Y)$ is called the conditional expectation and satisfies the tower property.

Theorem. *If $\mathbb{E}[X^2] < \infty$ ($X \in \mathbb{L}^2(\mathbb{R})$) then $\mathbb{E}[X \mid Y]$ coincides with the orthogonal projection of X on $L = \{h(Y) : h$ Borel function and $\mathbb{E}[h(Y)^2] < \infty\}$.*

Proof. We check the orthogonal property $\mathbb{E}[(X - \mathbb{E}[X \mid Y])h(Y)] = 0$ for any measurable function h and $\mathbb{E}[X \mid Y] = \int_{\mathbb{R}} x f_{X|Y}(x) d\nu_1(x)$. We have

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]h(Y)] &= \int_{\mathbb{R}} \mathbb{E}[X \mid Y = y] h(y) f_Y(y) d\nu_2(y) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x f_{X|Y=y}(x) d\nu_1(x) h(y) f_y(y) d\nu_2(y) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x h(y) f_{X|Y=y}(x) f_y(y) d\nu_1(x) d\nu_2(y) \\
&= \int_{\mathbb{R}^2} x h(y) f_{X,Y}(x,y) d(\nu_1\nu_2)(x,y) \\
&= \mathbb{E}[X h(Y)] \,.
\end{aligned}$$

□

The properties of the integral are inherited by the conditional expectation.

Proposition. *We have*

1. *The support of $\mathbb{E}[X \mid Y]$ is included in the one of X,*

2. *Positivity: If $X \leq Z$ a.s. then $\mathbb{E}[X \mid Y] \leq \mathbb{E}[Z \mid Y]$ a.s.,*

3. *Jensen's inequality: if h is a convex function then*

$$h(\mathbb{E}[X \mid Y]) \leq \mathbb{E}[h(X) \mid Y] \,, \qquad a.s.$$

4. *Linearity: for any measurable function h we have*

$$\mathbb{E}[h(Y)X \mid Y] = h(Y)\mathbb{E}[X \mid Y] \,, \qquad a.s.$$

5. *Hölder inequality: for any* $p, q > 1$ *such that* $p^{-1} + q^{-1} = 1$ *we have*

$$\mathbb{E}[|XZ| \mid Y] \le (\mathbb{E}[|X|^p \mid Y])^{1/p} (\mathbb{E}[|Z|^q \mid Y])^{1/q}\,, \qquad a.s.$$

Example 1 (Linear regression)**.** *Assume the linear model with random design* $Y_t = X^T\theta + \epsilon$ *such that* θ, $X \in \mathbb{R}^d$, Y, $\epsilon \in \mathbb{R}$, X *and* ϵ *independent of* X *and* $\mathbb{E}[\epsilon] = 0$. *Then* $\mathbb{E}[Y \mid X] = X^T\theta$ *and it is the best* $\mathbb{L}^2(\mathbb{R})$*-approximation of* Y *given* X *when* $Var(X)$ *and* $Var(\epsilon)$ *are finite.*

Example 2 (Logistic regression)**.** *Consider the binomial variable* $Y \in \{0,1\}$ *and* X *any continuous rv. Then* (Y, X) *admits a density wrt* $(\varepsilon_0 + \varepsilon_1)\times$ *Lebesgue and*

$$f_{Y|X=x}(y) = \frac{f(y,x)}{f(x)}\,, \qquad y = 0, 1\,.$$

It is the density of a Bernoulli variable of parameter $\mathbb{E}[Y \mid X = x] = h^*(x)$. *The function* h^* *with value in* $[0,1]$ *cannot be linear (logistic for instance).*

Recall that the conditional probability of $A \in \mathcal{F}$ given $B \in \mathcal{F}$ with $\mathbb{P}(B \neq 0)$ is defined as

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

Let $X = \mathbb{I}_A$ for some event $A \in \mathcal{F}$. Then X follows a Binomial distribution such that $\mathbb{P}(X = 1) = \mathbb{P}(A)$.

Definition 6. *We denote*

$$\mathbb{P}(A \mid Y) = \mathbb{E}[X \mid Y] = h^*(Y)$$

which is a rv on $[0,1]$ *such that* $\mathbb{E}[\mathbb{P}(A \mid Y)] = \mathbb{P}(A)$.

Part II

Stochastic Simulation

Chapter 2

Pseudo Random Number Generator

In order to simulate any rv one needs to introduce some randomness in a computer that is, by definition of a machine, deterministic.

The basic continuous distribution is the uniform one on $[0,1]$.

Definition 7. *A rv U is uniform on $[0,1]$, $U \sim \mathcal{U}[0,1]$, if $F_U(x) = x$ for $x \in [0,1]$.*

The rv U is continuous since its cdf is absolutely continuous wrt the Lebesgue measure $F_U(\{x_0\}) = F_U(x_0) - \lim_{x \uparrow x_0} F_U(x) = 0$ (by continuity of F_U).

It admits a density $f_U(x) = 1$ over its support $\mathrm{Supp}(U) = [0,1]$.

Definition 8. *A (simple) recursive algorithm is repeating the same iterations at each step returning an output. The iterations at each step $1 \leq k \leq n$ (n is called the epoch) may depend on the preceding step $k-1$.*

By definition, a recursive algorithm is scalable in the epoch, i.e. its complexity is proportional to n (and not n^a with $a > 1$). We write its complexity $O(n)$ where O means up to some deterministic constant. For a recursive algorithm, the factor of n in the complexity depends on each iteration.

Definition 9. *The efficiency of a recursive algorithm is the number of iteration of each step.*

A recursive algorithm that generates a sample of uniform rv is called a PRNG.

Definition 10. *A Pseudo Random Number Generator (PRNG) is a deterministic algorithm $\mathcal{A}$ such that given n returns $U_1, \ldots, U_n$ close to iid $\mathcal{U}[0,1]$ rv.*

Remark. *Since a PRNG is a deterministic recursive algorithm, it has:*

- *a seed for $n = 1$ determining U_1 which fix the first value as arbitrary as possible,*
- *a recursive procedure $\mathcal{A}(U_n) = U_{n+1}$ to design U_{n+1} as independent as possible as $U_1, \ldots, U_n$,*
- *a periodicity, the first n such that $U_n = U_1$ whatever is the seed.*

The seed is usually a deterministic function of the internal clock of the processor.

If one fixes the seed (seed$(\cdot)$ in R), any run $\mathcal{A}$ is the same.

In order to ensure that U_{n+1} is as independent as possible of the past value (but deterministic due to the restriction of the machine) one uses number theory

Definition 11. *A linear congruential PRNG is sampling U_i recursively from the seed U_1 thanks to the recursion*

$$\begin{aligned} X_i &\leftarrow aX_{i-1} + b \qquad mod(m) \\ U_i &\leftarrow X_i/m \end{aligned}$$

for (a, b, m) well-chosen.

A linear congruential PRNG is a recursive algorithm with efficiency $O(1)$, i.e. each iteration is a multiplication and a division with remainder. Such PRNG generates elements of the grid $\{0, \ldots, (m-1)/m\}$, $\{1, \ldots, (m-1)/m\}$ in the multiplicative case.

Proposition. *The best possible period of such PRNG is m, $m-1$ if the PRNG is multiplicative, i.e. $b = 0$.*

Number theory provides optimal period conditions for such PRNG.

Theorem. *A necessary condition for a multiplicative PRNG to have period $m-1$ is that m is a prime number.*

A sufficient condition for a multiplicative PRNG to have period $m-1$ is that a is a primitive root of m for the multiplicative group

$$a^{(m-1)/p} \neq 1\,, \qquad mod(m)$$

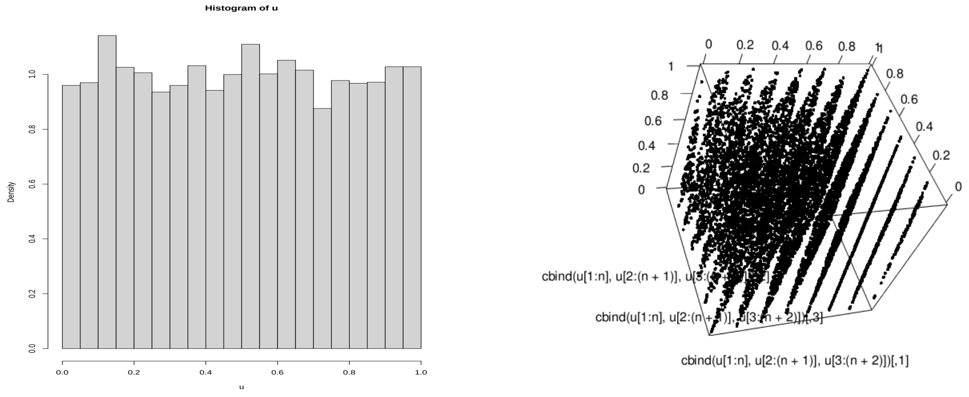
for any prime factor p of $m-1$.

The strategy is then to find a prime number m that is large but such that $m-1$ does not have many prime factors in order to check easily the sufficient condition. Such prime numbers are of the Mersenne type $2^M - 1$ where M is an auxiliary prime number.

Example 3.

- rand*(·) in C, $m = 2^{31} - 1$,*
- drand48*(·) in C, $m = 2^{48} - 1$,*
- RANDU*(·) by IBM, $m = 2^{31} - 1$, $b = 0$,*
- rand*(·) by Mapple, $m = 10^{12} - 11$, $b = 0$.*

R uses a Mersenne-Twister PRNG such that $m = 2^{19937} - 1$!!



Histogram for RandU that shows uniform distributions of the marginal and 3d-plot of consecutive samples (U_i, U_{i+1}, U_{i+2}). Visualization of a spurious deterministic relation.

Chapter 3

Simulation of a random variable

3.1 The inverse transform sampling

From now we assume that a PRNG can sample $(U_1, \dots, U_n)$ iid $\mathcal{U}nif(0,1)$.

3.1.1 The generalized inverse

Let X be a rv with cumulative distribution function F_X.

Definition 12. *The generalized inverse of the cadlag non-decreasing function $F_X : \mathbb{R} \mapsto [0,1]$, denoted $F_X^{\leftarrow} : [0,1] \mapsto \mathbb{R}$, is defined as*

$$F_X^{\leftarrow}(y) = \inf\{x \in \mathbb{R};\, F_X(x) \geq y\}\,, \qquad y \in (0,1]\,.$$

Here we use the convention $\inf \emptyset = +\infty$. The generalized inverse $F_X^{\leftarrow}$ is well defined and coincides with the inverse F_X^{-1} when F_X is invertible, which is equivalent of being increasing and continuous. Note that

$$\{(u,x) \in (0,1] \times \mathbb{R};\, u \leq F_X(x)\} = \{(u,x) \in (0,1] \times \mathbb{R};\, F_X^{\leftarrow}(u) \leq x\}\,. \tag{3.1}$$

Example 4. *Let $X \sim \mathcal{E}xp(\lambda)$, $\lambda > 0$, then $F_X(x) = 0$, $x < 0$, $F_X(x) = 1 - \exp(-\lambda x)$, $x > 0$ then $F_X^{\leftarrow}(y) = \lambda^{-1}\log(1/(1-y))$ for $0 < y \leq 1$.*

Let $X \sim \mathcal{B}inom(p)$, $0 < p < 1$, then $F_X(x) = 0$, $x < 0$, $F_X(x) = 1 - p$, $0 \leq x < 1$, $F_X(x) = 1$, $x \geq 1$ then $F_X^{\leftarrow}(y) = 0$ for $0 < y \leq 1 - p$ and $F_X^{\leftarrow}(y) = 1$ for $1 - p < y \leq 1$.

3.1.2 The inverse transform sampling

We have the useful proposition

Proposition. *If $U \sim \mathcal{U}(0,1)$ then $F_X^{\leftarrow}(U)$ is distributed as F_X.*

Proof. Using (3.1), we have $\mathbb{P}(F_X^{\leftarrow}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = F_U(F_X(x)) = F_X(x)\,.$ □

Example 5. *Let $X \sim \mathcal{B}inom(p)$, $0 < p < 1$ then $F_X^{\leftarrow}(U) = 1\!\!1_{(1-p,1]}(U)$ is $\mathcal{B}inom(p)$ distributed when $U \sim \mathcal{U}(0,1)$.*

```
Algorithm 1: The inverse transform sampling
  Parameters: $n$ the number of samples, $F_X$ the target distribution.
  Do
    1. Sample $U_1, \dots U_n$ iid $\mathcal{U}(0,1)$ (runif(n))
    2. Apply the inverse transform $F_X^{\leftarrow}$ so that $X_i = F_X^{\leftarrow}(U_i)$, $1 \le i \le n$.
  Return $(X_1, \dots, X_n)$.
```

When the generalized inverse $F^{\leftarrow}$ is explicit, the inverse transform is a recursive algorithm with efficiency $O(1)$. However it has some serious limitation.

Remark (Limitation of the inverse transform)**.**

1. *When no explicit formula on the distribution F_X is known (as $X \sim \mathcal{N}(0,1)$ with distribution Φ only known through its density φ),*
2. *When F_X involves an infinite number of steps and is intractable in practice. For instance X discrete with countably many $\{x_k\}$ with $\mathbb{P}(X = x_k) > 0$ for all $k \ge 0$.*

3.1.3 Inverse transform adapted to discrete rv

Let X be discrete with countably many $\{x_k\}_{k\in\mathbb{N}}$ with $\mathbb{P}(X = x_k) > 0$ for all $k \ge 0$.

Sort x_k such that $x_{(k)}$ satisfies

$$\mathbb{P}(X = x_{(k-1)}) > \mathbb{P}(X = x_{(k)}) > \mathbb{P}(X = x_{(k+1)})$$

for all $k \in \mathbb{N}$.

```
Algorithm 2: Recursive test procedure
  Sample $U_i$ and do the recursive test procedure from $k = 0$
    • If $U_i \le \mathbb{P}(X = x_{(1)}) + \cdots + \mathbb{P}(X = x_{(k)})$ then Return $X = x_{(k)}$
    • Else $k \leftarrow k + 1$.
```

There is k tests with probability $\mathbb{P}(X = x_{(k)})$ on average thus the total number of tests for sampling one rv is

$$\sum_{k\in\mathbb{N}} k\mathbb{P}(X = x_{(k)}).$$

This number is potentially infinite... If it is finite one says that the efficiency of the recursive algorithm is

$$O_{\mathbb{P}}\Big(\sum_{k\in\mathbb{N}} k\mathbb{P}(X = x_{(k)})\Big),$$

Where $O_{\mathbb{P}}$ means that each step requires a random number of iterations with finite mean.

3.2 The reject method

3.2.1 Proposal distribution

Let X be distributed as f_X (determined by its density). Let Y be distributed as a proposal, i.e. Y is easily sampled (uniform or inverse transform).

Definition 13. *The proposal dominates the target distribution when $\exists M > 0$ such that $f_X(x) \le M f_Y(y)$.*

Note that necessarily $M \ge 1$ and X is absolutely continuous wrt Y.

3.2.2 The rejection sampling

```
Algorithm 3: The rejection sampling
  Parameters: $n$ the number of samples, $f_X$ the target density and a dominating
    proposal $f_Y$.
  While $k \leq n$ Do
    1. Sample $U \sim \mathcal{U}(0,1)$ (runif(1)) and $Y \sim f_Y$ independently,
    2. • If $U \leq f_X(Y)/(M f_Y(Y))$ then $X_k = Y$ and $k \leftarrow k+1$,
       • Else reject and Return to 1.
  Return $(X_1, \dots, X_n)$.
```

Proposition. *For each k the rejection sampling returns X_k after T_k iterations where T_k is a rv Geom$(1/M)$ and $X_k \sim F_X$.*

Corollary. *The rejection sampling returns $(X_1, \dots, X_n)$ iid F_X using $\sum_{k=1}^n T_k$ sampling of uniform and proposal rv.*

Proposition. *The efficiency of the rejection sampling is $O_{\mathbb{P}}(M)$.*

Example 6. *Consider the truncated normal distribution*

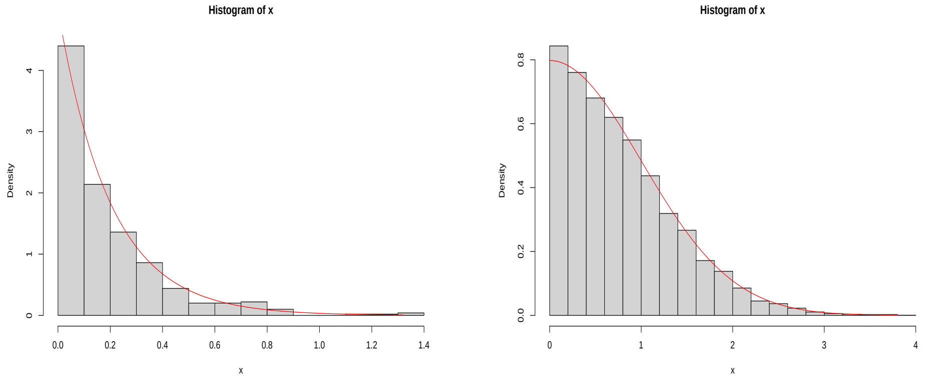
$$f_X(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp(-\frac{x^2}{2}), & \textit{if } x > 0, \\ 0, & \textit{else.} \end{cases}$$

Consider the proposal $Y \sim \mathcal{E}xp(1)$ so that

$$\frac{f_X(x)}{f_Y(x)} \leq M \approx 1,316.$$

The efficiency of the rejection sampling is $O_{\mathbb{P}}(1,316)$.

Note that one can recover the gaussian rv by multiplying $N = \mathbb{1}_{U>1/2} X + \mathbb{1}_{U \leq 1/2} X$ for $U \sim \mathcal{U}(0,1)$ independent of X.



On the left, sampling of the exponential distribution by the inverse transform method. On the right sampling of the truncated normal distribution using the rejection method with exponential distributed proposal.

3.3 Sampling specific distributions

3.3.1 Gaussian rv

For 2 standard gaussian independent rvs X_1 and X_2, the gaussian vector (X_1, X_2) is isotropic, i.e. its distribution

$$f_{(X_1,X_2)}(x_1,x_2) = \frac{1}{2\pi}\exp(-(x_1^2+x_2^2)/2)\,, \qquad (x_1,x_2)\in\mathbb{R}^2\,,$$

solely depend on the square of the radius $\|(x_1,x_2)\|^2 = x_1^2 + x_2^2$ and not on the angle $\arctan(x_1/x_2)$

Proposition. *Let $R \sim \mathcal{E}xp(1/2)$ and $\theta \sim \mathcal{U}(0,2\pi)$ independent. Then $X_1 = \sqrt{R}\cos(\theta)$ and $X_2 = \sqrt{R}\sin(\theta)$ are 2 independent rvs $\mathcal{N}(0,1)$.*

Remark. *A χ^2_n-distribution is the distribution of $\sum_{i=1}^n X_i^2$ for $X_1,\dots,X_n$ iid $\mathcal{N}(0,1)$.*

It is also a $\Gamma(n/2,1/2)$ distribution where a $\Gamma(n,\lambda)$ distribution is given by its density

$$f_\Gamma(x) = \frac{\lambda^k}{\Gamma(k)}x^{k-1}\exp(-\lambda x)\,\mathbb{1}_{(0,\infty)}(x)\,.$$

Then a χ^2_2-distribution is also a $\mathcal{E}xp(1/2)$-distribution.

Algorithm 4: Box-Muller sampling

```
The aim is to sample standard gaussian rv $\mathcal{N}(0,1)$ as efficiently as we can.
Parameters: $n$ the (even) number of samples.
Do
    1. Sample $U_1,\dots,U_n$ iid $\mathcal{U}(0,1)$ (runif(n)),
    2. Inverse transform $U_1,\dots U_{n/2}$ into $R_1,\dots,R_{n/2}$ iid $\mathcal{E}xp(1/2)$, $U_{n/2+1},\dots U_n$ into
       $\theta_1,\dots,\theta_{n/2}$ iid $\mathcal{U}(0,2\pi)$
    3. Compute $X_{2k} = \sqrt{R_k}\cos(\theta_k)$ and $X_{2k+1} = \sqrt{R_k}\sin(\theta_k)$.
Return $(X_1,\dots,X_n)$.
```

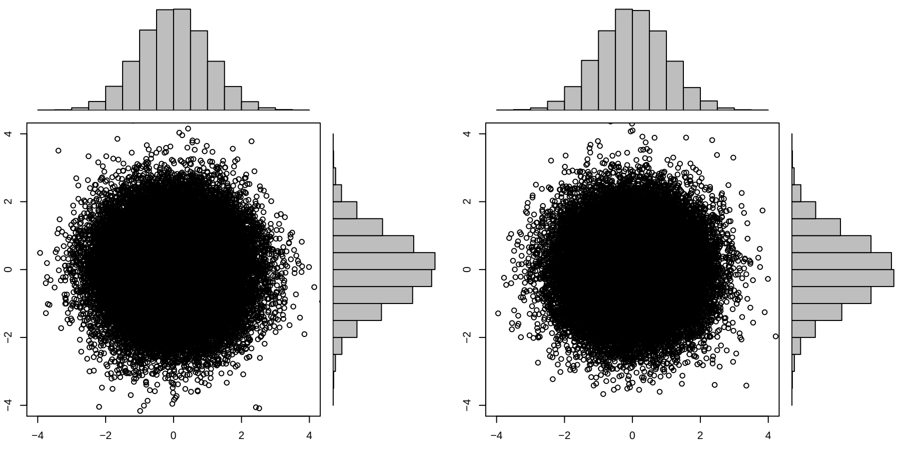
Proposition. *The efficiency of the Box-Muller sampling is $O(1)$!*

A gaussian rv $X = \mathcal{N}(\mu,\sigma^{\in})$ is given by its distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\,, \qquad x\in\mathbb{R}\,.$$

From $(Y_1,\dots,Y_n)$ obtained by the Box-Muller sampling we get $(X_1,\dots,X_n)$ applying

$$X_k = \mu + \sigma Y_k\,, \qquad 1\le k\le n\,.$$



Box-Muller algorithm sampling a couple of centered gaussian random variables. On the left, the original BM algorithm, on the right the cosinus and sinus functions are replaced with a rejection step. The two distributions are identically isotropic gaussian random vectors.

3.3.2 Poisson distribution

The discrete rv X is $\mathcal{P}ois(\lambda)$, $\lambda > 0$, distributed if

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}\exp(-\lambda), \qquad k = 0, 1, \ldots$$

Proposition. *Let $(E_k)_{k\geq 1}$ be iid $\mathcal{E}xp(\lambda)$, $\lambda > 0$ then*

$$\mathbb{P}(S_n \leq 1 \leq S_{n+1}) = \frac{\lambda^n}{n!}\exp(-\lambda)\,, \qquad S_n = \sum_{k=1}^{n} E_k, \qquad n \geq 0\,.$$

Algorithm 5: Poisson sampling

```
Parameters: $n$ the number of samples,.
While $k \leq n$ Do
    1. $X_k \leftarrow 0$, $S \sim \mathcal{E}xp(\lambda)$,
    2. While $S < 1$ Do
        • Sample $E \sim\sim \mathcal{E}xp(\lambda)$,
        • Compute $S \leftarrow S + E$ and $X_k \leftarrow X_k + 1$,
Return $(X_1, \ldots, X_n)$.
```

Its efficiency is $O_{\mathbb{P}}(1 + \lambda)$.

3.3.3 Mixture distributions

Definition 14 (Mixture distribution)**.** *A mixture distribution consists in two steps of randomness;*

- *The mixture components which are conditional densities $f_{X|Y=y}$,*
- *Mixtures weights which correspond to the mixing distribution $f_Y(y)$ with support $\mathcal{Y}$.*

Then $X \sim \int_{\mathcal{Y}} f_{X|Y=y} f_Y(y) d\nu(y)$ is following a mixture distribution.

Algorithm 6: Mixture sampling

```
Parameters: $n$ the number of samples.
For $1 \leq i \leq n$ Do
    1. $Y_i \sim f_Y$,
    2. $X_i \sim f_{X|Y_i}$
Return $(X_1, \dots, X_n)$.
```

Example 7 (Student's distribution as a mixture distribution)**.** *The Student's distribution with df $k > 0$ is given by the distribution*

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \Big(1 + \frac{t^2}{k}\Big)^{-(k+1)/2}, \qquad t \in \mathbb{R}.$$

We also have $T \sim \mathcal{N}(0, k/Y)$ where $Y \sim \chi^2_k$. At each step $1 \leq i \leq n$ we simulate

1. $(N_j)_{1 \leq j \leq k}$ *iid* $\mathcal{N}(0,1)$
2. $Y_i = \sum_{j=1}^k N_j^2$,
3. $T_i \sim \mathcal{N}(0, k/Y_i)$, *i.e.* $T_i = N' \sqrt{k/Y_i}$ *for* $N' \sim \mathcal{N}(0,1)$*independent of* $(N_j)_{1 \leq j \leq k}$.

Part III

Monte Carlo methods

Chapter 4

Crude Monte Carlo approximation and accelerations

4.1 Monte-Carlo approximation error

Let h be an integrable function, we want to estimate

$$I = \int h(x)dx,$$

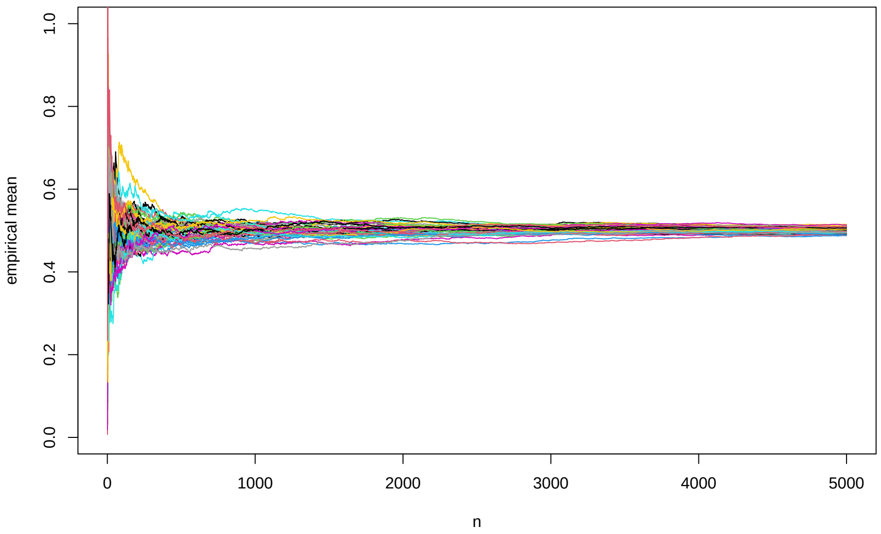
and the primitive of h is unknown. One can use

1. Numerical (deterministic) approximation discretizing $\mathrm{Supp}(h)$ and approximating I by $n^{-1}\sum_{i=1}^{n} h(x_i)$,
2. Monte-Carlo (random) approximation drawing $Y_1, \dots, Y_N \sim \mathcal{U}(\mathrm{Supp}(h))$ and approximating I by $|\mathrm{Supp}(h)|N^{-1}\sum_{i=1}^{N} h(Y_i)$ where $|\mathrm{Supp}(h)|$ is the Lebesgue measure of $\mathrm{Supp}(h)$.

The Monte Carlo approximation is based on to fundamental asymptotic theorem for iid sequences. The first one is the Strong Law of Large Number (SLLN) that justifies the expression of the Monte Carlo approximation:

Theorem (SLLN)**.** *Consider $Y_1, \dots, Y_N$ iid and h such that $\mathbb{E}[|h(Y)|] < \infty$ then*

$$\frac{1}{N}\sum_{i=1}^{N} h(Y_i) \to \int h(y) f_Y(y) d\nu(y)\,, \qquad N \to \infty, \qquad a.s.$$



Visualization of the convergence in the SLLN toward the expectation 1/2 of the sample means of $\mathcal{E}xp(2)$ distribution.

For Monte Carlo method we get as $N \to \infty$

$$\frac{1}{N}\sum_{i=1}^{N} h(Y_i) \to \int h(y) f_Y(y) dy = \int h(y) \frac{1}{|\text{Supp}(h)|} dy\,, \qquad a.s.$$

Numerical and (uniform) Monte Carlo methods are very close converging methods but the second one is much more flexible.

Definition 15. *Consider f_Y the uniform distribution on $|Supp(h)| < \infty$ such that $g = h/f_Y$ is integrable, sample $Y_1, \ldots, Y_N$ iid f_Y then*

$$\hat{I}_N^{(MC)} = \frac{1}{N}\sum_{i=1}^{N} g(Y_i) \to I\,, \qquad a.s.$$

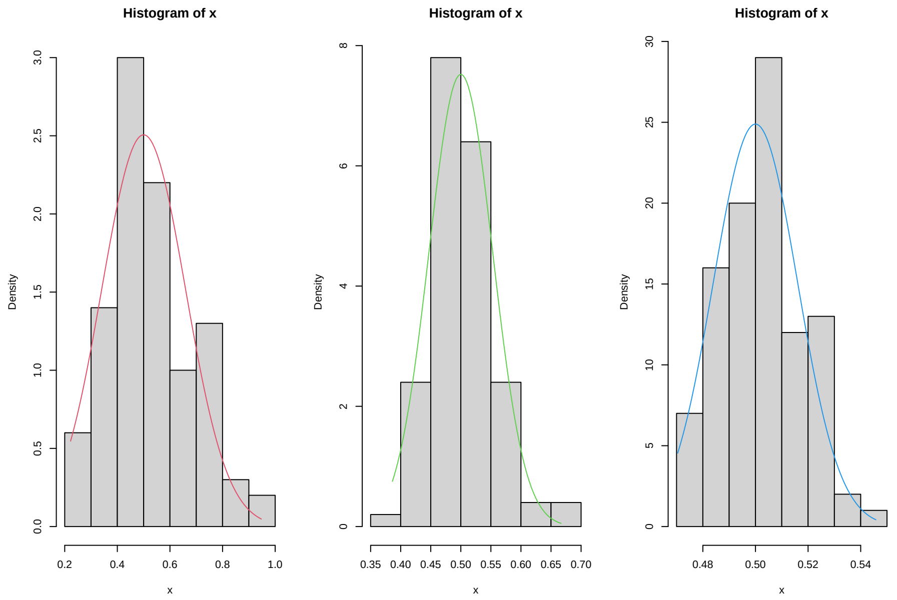
The estimator $\hat{I}_N^{(MC)}$ is called the crude Monte Carlo estimator of I.

Any Monte Carlo approximation has an error. Its analysis is based on the second fundamental asymptotic theorem for iid sequences, namely the Central Limit Theorem (CLT):

Theorem (CLT)**.** *Consider $Y_1, \ldots, Y_N$ iid and g such that $\mathbb{E}[g(Y)^2] < \infty$ then*

$$\sqrt{\frac{N}{Var(g(Y))}}\Big(\frac{1}{N}\sum_{i=1}^{N} g(Y_i) - \int g(y) f_Y(y) d\nu(y)\Big) \to \mathcal{N}(0,1)\,,$$

in distribution.



Visualization of the gaussian approximation in the CLT of the sample mean distributions of $\mathcal{E}xp(2)$ samples.

Given a level of approximation $\epsilon > 0$, if $\hat{I}_N^{(MC)} = N^{-1}\sum_{i=1}^N g(Y_i)$ with $\mathrm{Var}\,(g(Y)) < \infty$ then we have

$$\begin{aligned}\mathbb{P}(|\hat{I}_N^{(MC)} - I| > \epsilon) &= \mathbb{P}(\sqrt{N/\mathrm{Var}\,(g(Y))}|\hat{I}_N^{(MC)} - I| > \epsilon\sqrt{N/\mathrm{Var}\,(g(Y))})\\ &\approx 2(1-\Phi(\epsilon\sqrt{N/\mathrm{Var}\,(g(Y))}))\\ &\approx 2\phi(\epsilon\sqrt{N/\mathrm{Var}\,(g(Y))})\sqrt{\mathrm{Var}\,(g(Y))/(N\epsilon^2)}\\ &\approx o(\exp(-\epsilon^2 N/\mathrm{Var}\,(g(Y)))).\end{aligned}$$

Thus the approximation is controlled when $\epsilon^2 \approx \mathrm{Var}\,(g(Y)))/N$. We denote $|\hat{I}_N^{(MC)} - I| = O_{\mathbb{P}}(\sqrt{\mathrm{Var}\,(g(Y))/N})$

Recipe: Given a satisfactory approximation level $\epsilon > 0$, simulate N of the order $1/\sqrt{\epsilon}$ providing that $\mathrm{Var}\,(g(Y))$ stays small.

Thus the only requirements in the crude Monte Carlo methods are to sample easily the proposal f_Y and to keep $\mathrm{Var}\,(g(Y))$, i.e. $\int h(y)^2 f_Y(y)dy$ small. The first requirement is deeply link with sampling methods from the previous chapter, the second one yields to seek variance reduction.

4.2 Link between sampling and crude Monte Carlo approximation

Consider $I = |\Delta| = \int \mathbb{I}_\Delta(x)dx$ then considering a rectangle $\Delta \subset R$, we have the Monte Carlo method

Algorithm 7: Monte Carlo approximation based on acceptation

```
Parameters $N$ the number of iterations.
For $1 \leq i \leq N$ Do
    1. Sample $Y_i \sim \mathcal{U}(R)$,
    2. If $Y_i \in \Delta$ then $X_i = Y_i$.
Return $I_N^{(MC)} = |R|n/N$ where $n = \#\{X_i\}$.
```

The Monte Carlo algorithm shares some similarities with the reject sampling since the different steps are the same. However it is also different since now the number of runs N is deterministic and the number of acceptances n is random. That $X_i \sim \mathcal{U}(\Delta)$ as from the rejection sampling allows to estimate the normalized constant $|\Delta|$ of the density of the uniform distribution thanks to the acceptance rate n/N.

Proposition. *The reject sampling generates a (random size) n sample $X_i \sim \mathcal{U}(\Delta)$ such that*

$$\hat{I}_N^{(MC)} = |R|n/N = \frac{1}{N}\sum_{i=1}^{N} |R|\, \mathbb{1}_\Delta(Y_i) \to |R| \int \mathbb{1}_\Delta(y)dy/|R| = I$$

is a $O_{\mathbb{P}}(\sqrt{|R||\Delta|(1-|\Delta|/|R|)/N})$ approximation.

Note that luckily the error of approximation depends only on N, the number of iterations in the Monte Carlo approximation, and not on the random acceptance number n

Based on this reject principle one can approximate $\int h(x)dx$ for any $h \geq 0$ thanks to the ratio sampling:

Theorem. *Assume that $h \geq 0$ is integrable then $C_h = \{(u,v) \in \mathbb{R}^2;\ 0 \leq u \leq \sqrt{h(v/u)}\}$ is finite, i.e. $|C_h| < \infty$.*

In practice one has to assume that $x^2h(x) < \infty$ so that the constants

$$\begin{aligned} a &= \sqrt{\sup\{h(x);\ x \in \mathbb{R}\}} \\ b_+ &= \sqrt{\sup\{x^2h(x);\ x \geq 0\}} \\ b_- &= -\sqrt{\sup\{x^2h(x);\ x \leq 0\}} \end{aligned}$$

are well defined and then $C_h \subset [0,a] \times [b_-, b_+]$.

Algorithm 8: Ratio MC approximation

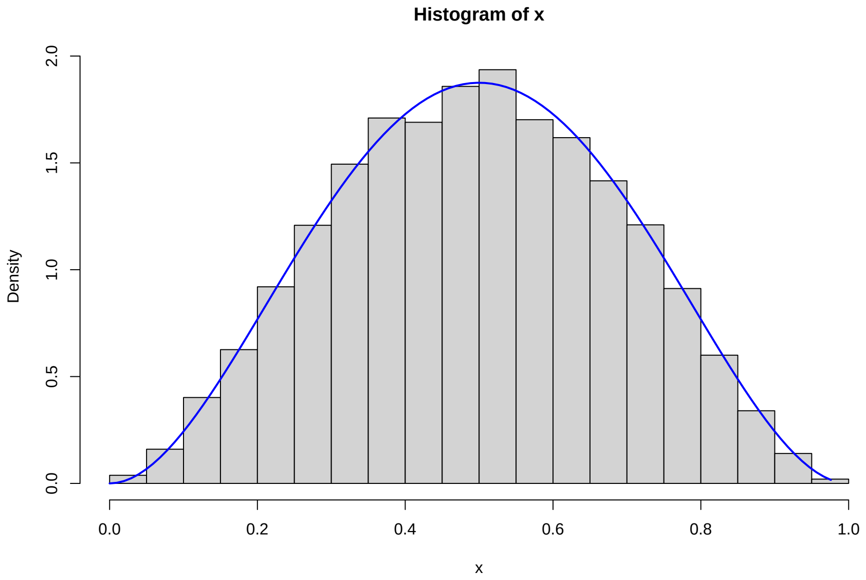
```
Parameters: $N$ the number of iterations, $h$ the integrand.
For $1 \leq i \leq N$ Do
    1. Sample $U_i$, $V_i$ two independent rv $\mathcal{U}(0,1)$,
    2. Transform $U = aU_1$, $V = b_- + (b_+ - b_-)U_2$
    3. If $U, V \notin C_h$ then $X_i = U/V$.
Return $\hat{I}_N^{(MC)} = 2a(b_+ - b_-)n/N$ where $n = \#\{X_i\}$.
```

Proposition. *The ratio sampling generates $X_1, \ldots, X_n$ iid such that $f_X \sim h/\int h$.*

Moreover we have that

$$\hat{I}_N^{(MC)} = 2a(b_+ - b_-)n/N$$

is a $O_{\mathbb{P}}(\sqrt{2a(b_+ - b_-)\int h(x)dx/N})$ approximation.



The use of the ratio method for sampling a β-distribution. No need of the knowledge of the normalizing constant!

4.3 Variance reduction

4.3.1 Antithetic variance reduction method

Consider $\hat I_N^{(1)}$ and $\hat I_N^{(2)}$ 2 Monte Carlo approximation of I then

$$\mathrm{Var}\,((\hat I_N^{(1)}+\hat I_N^{(2)})/2)=\frac{1}{4}(\mathrm{Var}\,(\hat I_N^{(1)})+\mathrm{Var}\,(\hat I_N^{(2)})+2\mathrm{Cov}(\hat I_N^{(1)},\hat I_N^{(2)})).$$

Thus if $\mathrm{Cov}(\hat I_N^{(1)},\hat I_N^{(2)}))\le 0$ then one accelerates

$$\mathrm{Var}\,((\hat I_N^{(1)}+\hat I_N^{(2)})/2)\le(\mathrm{Var}\,(\hat I_N^{(1)})+\mathrm{Var}\,(\hat I_N^{(2)}))/4$$

Theorem (Antithetic variables)**.** *Consider g monotonic and $U\sim\mathcal{U}(0,1)$ then*

$$Cov(g(U),g(1-U))\le 0\,.$$

We say that U and $1-U$ are antithetic.

We say that U and $1-U$ are antithetic.

Proof. Consider U' iid to U. Then $\mathrm{Cov}(g(U)-g(U'),g(1-U)-g(1-U'))=2\mathrm{Cov}(g(U),g(1-U))$ and

$$\begin{aligned}\mathrm{Cov}(g(U)-g(U'),g(1-U)-g(1-U'))=&\mathbb{E}[(g(U)-g(U'))(g(1-U)-g(1-U'))]\\=&\mathbb{E}[(g(U)-g(U'))(g(1-U)-g(1-U'))\,\mathbb{1}_{U>U'}]\\&+\mathbb{E}[(g(U)-g(U'))(g(1-U)-g(1-U'))\,\mathbb{1}_{U'\ge U}]\,.\end{aligned}$$

□

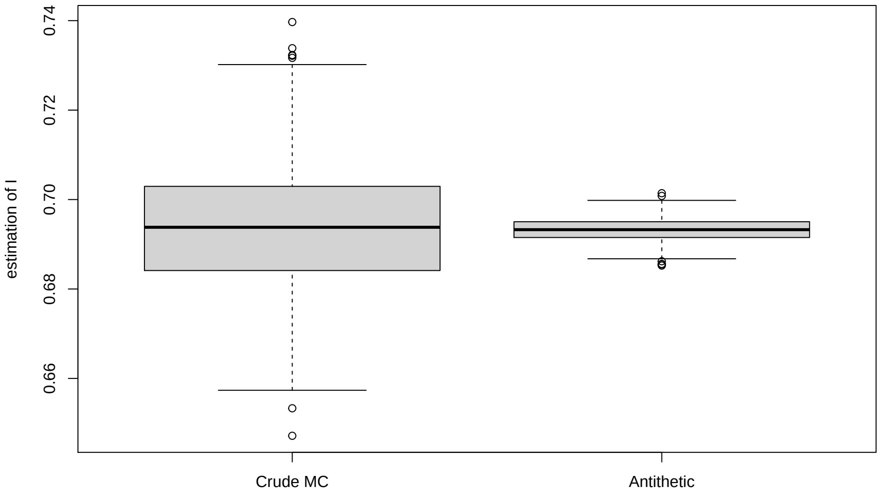
Remark. *The more linear g the smaller $Cov(g(U),g(1-U))$; we have $Cov(g(U),g(1-U))=-1$.*

We have Φ that is increasing as Φ^{-1}, $X=\Phi^{-1}(U)$ and $-X=\Phi^{-1}(1-U)$ are antithetic.

Example 8. *Consider* $I = \int_0^1 \frac{1}{1+x}dx = \log 2$, $\hat{I}_N^{(MC)} = \frac{1}{N}\sum_{i=1}^N \frac{1}{1+U_i}$ *and*

$$\hat{I}_N^{(Anti)} = \frac{\sum_{i=1}^N \frac{1}{1+U_i} + \sum_{i=1}^N \frac{1}{2-U_i}}{2N}.$$

Note that the relative gain of the antithetic method is at least 1/2. Indeed the antithetic approach allows us to consider 2 crude Monte Carlo approximations $\hat{I}_N^{(1)}$ and $\hat{I}_N^{(2)}$ using only a N-sample $U_1, \dots, U_N$.



Acceleration of the crude MC method thanks to the use of antithetic variables. The distribution of the error is more concentrated around the value of interest. In particular the variance is reduced.

4.4 Control variate acceleration

Consider $I = \mathbb{E}[g(Y)] = \mathbb{E}[g(Y) + \ell(Y)] - \mathbb{E}[\ell(Y)]$. If ℓ explain a part of g and if $\mathbb{E}[\ell(Y)]$, it is easier to approximate $\mathbb{E}[g(Y)] = \mathbb{E}[g(Y) + \ell(Y)]$ rather than I.

Definition 16. *A control variate* $\ell(Y)$ *is such that* $\mu = \mathbb{E}[\ell(Y)]$ *is known. Then*

$$\hat{I}_N^{(Cont)} = \frac{1}{N}\sum_{i=1}^N (g(Y_i) + \ell(Y_i)) - \mu.$$

Note that if $\ell(Y)$ then $c\ell(Y)$, $c \in \mathbb{R}$, is also a control variate. We can compute the optimal control variate and the optimal gain:

Proposition. *Given* ℓ, *the optimal control variate is* $c^*\ell$ *with*

$$c^* = -\frac{Cov(g(Y), \ell(Y))}{Var(\ell(Y))}.$$

Then the relative gain on the variance is $1 - Cor(g(Y), \ell(Y))^2$.

In practice c^* is not know. It indicates that g and ℓ should be negatively correlated so that $\mathbb{E}[g(Y)+\ell(Y)]$ is easier than I to approximate.

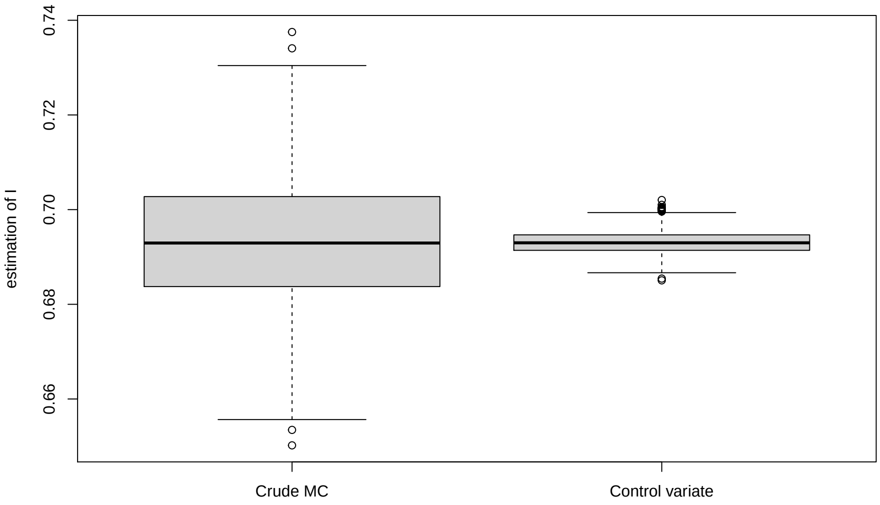
Example 9. *Consider* $I = \int_0^1 \frac{1}{1+x}dx = \log 2$, *consider* $g(y) = 1/(1+y)$ *and* $\ell(y) = 1+y$ *so that* $\mu = 3/2$. *Then*

$$c^* \approx 0.4773.$$

Indeed g is a decreasing function of x and ℓ is increasing. We also

$$\hat{I}_N^{(Cont)} = \frac{\sum_{i=1}^N \frac{1}{1+U_i} + c^* \sum_{i=1}^N (1+U_i)}{N} - c^*\mu.$$

The relative gain is usually less than for the antithetic acceleration.



Acceleration of the crude MC method thanks to the use of control variates. The distribution of the error is more concentrated around the value of interest. In particular the variance is reduced.

Chapter 5

Importance Sampling

5.1 Importance sampling

Monte Carlo approximations achieve their full generality and flexibility when f_Y is not the uniform distribution. Note that given f_Y there is no need for $|\mathrm{Supp}(h)| < \infty$ and the SLLN and CLT apply until $\mathbb{E}[g^2(Y)] < \infty$ for $g = h/f_Y$. The variance term of the Monte Carlo approximation is then equal to

$$\mathrm{Var}\left(g(Y)\right) = \int \frac{h(y)^2}{f_Y(y)} dy - I^2 \, ,$$

On the contrary to its expectation $\mathbb{E}[g(Y)] = I$, the variance depends on the properties of the function h^2/f_Y thus on f_Y. The optimal f_Y is $\propto h$ so that

$$\mathrm{Var}\left(g(Y)\right) = 0.$$

But sampling $f_Y \propto h$ is more complicated than approximating I!

One has to make a compromise between the difficulty of sampling Y (simplicity of f_Y) and the difficulty of approximate I from Y (small variance $\mathrm{Var}\left(g(Y)\right)$ or small ratio h^2/f_Y).

5.2 Importance weights

We consider f_Y simplistic (uniform or normal) and an intermediate density f (also called importance sampling density) and

$$\frac{h(y)}{f_Y(y)} = g(y)\frac{f(y)}{f_Y(y)}$$

where now $g(y) = h(y)/f(y)$ so that h^2/f is small, meaning that f is large when h^2 is large. The density f puts mass on large values of h^2, where it is important to sample for having a good approximation. We interpret $w(y) = \dfrac{f(y)}{f_Y(y)}$ so that $\mathbb{E}[w(Y)] = 1$ as importance weights.

Definition 17. *The Importance Sampling approximation of $I = \int g(y)f(y)dy$ from the proposal f_Y is*

$$\hat{I}_N^{(IS)} = \frac{1}{N}\sum_{i=1}^{N} g(Y_i)w(Y_i)$$

where $g = h/f$ is as small as possible and $w(y) = f(y)/f_Y(y)$.

The weights put mass where f is large, emphasizes samples Y_i that are important, close to where h^2 is large. The density f is an intermediate step for realizing a good approximation. Given that $g = h/f$ is bounded we have

$$\begin{aligned}\operatorname{Var}(g(Y)w(Y)) &= \mathbb{E}[g(Y)^2 w(Y)^2] - I^2\\ &= \mathbb{E}[g(Y)^2(w(Y)^2-1)] + \mathbb{E}[g(Y)^2] - I^2\\ &\le \sup|g|^2 \operatorname{Var}(w(Y)) + \operatorname{Var}(g(Y)).\end{aligned}$$

The variance is always larger than without importance weights and the intermediate function f should be used as a target density for the proposal f_Y. Then the next step is to choose f_Y as close as possible to f, for instance the gaussian distribution with the same mean and same variance than f.

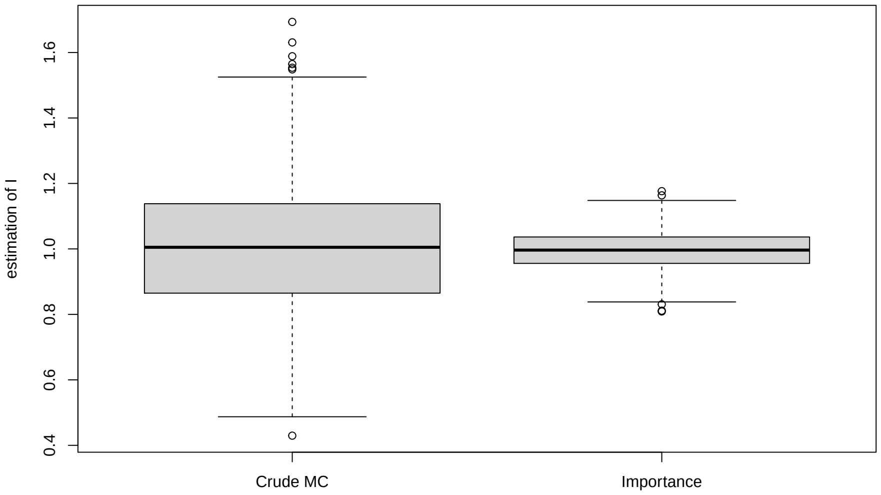
Example 10. *Consider $I = \int_0^{10} \exp(-2|x-5|)dx$. The crude Monte Carlo would be, for U_i iid $\mathcal{U}(0,10)$*

$$\hat{I}_N^{(MC)} = \frac{1}{N}\sum_{i=1}^{N} \exp(-2|U_i - 5|).$$

The IS approximation would exploit the fact that $h(x) = \exp(-2|x-5|)$ puts a lot of importance around 5. We should consider the proposal $f_{\sigma^2} = \mathcal{N}(5,\sigma^2)$ and

$$\hat{I}_N^{(IS)} = \frac{1}{N} g_{\sigma^2}(Y_i) w_{\sigma^2}(Y_i)$$

where $g_{\sigma^2}(y) = \sqrt{2\pi\sigma^2}\exp(-2|y-5|+(y-5)^2/(2\sigma^2))\,\mathbb{1}_{[0,10]}(y)$ and $w_{\sigma^2}(y) = 10/\sqrt{2\pi\sigma^2}\exp(-(y-5)^2/(2\sigma^2))$. The good compromise for finding a small g_{σ^2} depends on the hyperparameter σ^2. The smaller the more concentrated around 5 the sample and the larger away from 5, the larger the less concentrated around 5 and the smaller around 5. Given a good compromise g_{σ^2} it is then possible to do the IS approximation from $\mathcal{N}(5,\sigma^2)$. Note that the variance of w_{σ^2} should also be small and thus σ^2 not too big.



Acceleration of the crude MC method thanks to the use of Importance sampling. The choice of the intermediate distribution is crucial.

Chapter 6

General Importance Sampling

6.1 Motivation

Importance Sampling can be used for approximating integrals of the form

$$I = \int h(x)dx = \int g(x)f(x)dx$$

where the intermediate f is a density known up to a constant. For instance it might be that the density involves some complicated normalizing constants (Beta, Gamma coefficients...). The problem is well-posed in the sense that I is unique since given any non-negative function function f then $f/\int f$ is the unique density equals to f up to a constant. The idea is then to estimate $\int gf = \int g(f/f_Y)f_Y$ thanks to modified weights.

Algorithm 9: General Monte Carlo approximation

```
Parameters: $N$ the number of iterations, $g = h/f$ where $f$ is a non-negative
  function, the proposal $f_Y$.
For $1 \le i \le N$ Do
    1. Sample $Y_i$ iid $f_Y$,
    2. Compute $W(Y_i) = f(Y_i)/f_Y(Y_i)$.
Compute weights $\hat{w}_N(Y_i) = W(Y_i)/(N^{-1}\sum_{i=1}^N W(Y_i))$.
Return $\hat{I}_N^{(IS)} = \frac{1}{N}\sum_{i=1}^N g(Y_i)\hat{w}_N(Y_i)$.
```

As an immediate corollary of the SLLN, the approximation $\hat{I}_N^{(IS)}$ is consistent.

Proposition. *If* $\mathbb{E}[W(Y)] < \infty$ *and* $\mathbb{E}[|g(Y)|W(Y)] < \infty$ *then, as* $N \to \infty$

$$\frac{1}{N}\sum_{i=1}^N W(Y_i) \to \int f, \qquad \frac{1}{N}\sum_{i=1}^N g(Y_i)W(Y_i) \to \int g(y)f(y)dy, \qquad a.s.$$

so that $\hat{I}_N^{(IS)} \to I$ *a.s.*

Note that by construction

$$\frac{1}{N}\sum_{i=1}^N \hat{w}_N(Y_i) = 1\,.$$

The analysis of the variance is more complicated. However from the CLT and Slutsky Lemma we get

Proposition. *If $\mathbb{E}[W(Y)] < \infty$ and $Var(g(Y)W(Y)) < \infty$ we have as $N \to \infty$*

$$\frac{1}{N}\sum_{i=1}^{N} W(Y_i) \to \int f, \qquad a.s.$$

$$\sqrt{\frac{N}{Var(g(Y)W(Y))}}\Big(\frac{1}{N}\sum_{i=1}^{N} g(Y_i)W(Y_i) - \int g(y)f(y)dy\Big) \to \mathcal{N}(0,1)$$

in distribution so that $|\hat{I}_N^{(IS)} - I| = O_{\mathbb{P}}(\sqrt{Var(g(Y)W(Y))/N})$.

Given that g is small one has to control the variance of the pseudo importance weights $\mathrm{Var}\,(W(Y))$.

6.2 Application to Bayesian inference

One observes $X_1, \dots, X_n \in \mathcal{X}$ iid $\sim f_{\theta^*}$ where $\theta \in \Theta$ for some known set $\Theta \in \mathbb{R}^d$ and some unknown parameter θ^*, $\mathcal{X}$ being the observation space. Statistical inference consists in estimating θ^* from an estimator $\hat{\theta}_n$ based on the observations $X_1, \dots, X_n$:

$$\hat{\theta}_n = T(X_1, \dots, X_n)$$

where T is some measurable function from $\mathcal{X}^n \mapsto \Theta$. There exist two different approaches

1. Frequentist approach: we assume θ^* deterministic and one estimates it thank to the maximum likelihood principle

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f_\theta(X_i) =: \arg\max_{\theta\in\Theta} L_n(\theta)\,,$$

   It is the Maximum Likelihood Estimator (MLE).

2. Bayesian approach: we assume θ random $\sim \pi$ an a known priori distribution on Θ. The posterior distribution is given by the Bayes formula

$$f(\theta \mid X_1, \dots, X_n) = \frac{L_n(\theta)\pi(\theta)}{\int_\Theta L_n(\theta)\pi(\theta)d\nu(\theta)}\,.$$

   The Bayes estimator is the posterior mean

$$\hat{\theta}_n^B = \mathbb{E}[\theta \mid X_1, \dots, X_n]\,.$$

Proposition. *Given the prior distribution π the Bayes estimator is minimizing the quadratic risk among any square integrable estimator*

$$\hat{\theta}_n^B = \arg \min_{T(X_1,\dots,X_n)\in\mathbb{L}^2} \mathbb{E}[(\theta - T(X_1, \dots, X_n))^2].$$

Example 11. *Let $X_1, \dots, X_n$ iid $\mathcal{N}(\theta, 1)$ with $\Theta \in (0,1)$. Then $\hat{\theta}_n = \min(\max(n^{-1}\sum_{i=1}^n X_i, 0), 1)$. Consider $\pi = \mathcal{U}(0,1)$ then we have*

$$\begin{aligned}\hat{\theta}_n^B &= \mathbb{E}[\theta \mid X_1, \dots, X_n]\\ &= \frac{\int_\Theta \theta L_n(\theta)\pi(\theta)d\nu(\theta)}{\int_\Theta L_n(\theta)\pi(\theta)d\nu(\theta)}\\ &\propto \int_0^1 \theta L_n(\theta)d\theta \in (0,1)\,.\end{aligned}$$

Instead of computing the maximizer and then projecting on Θ we incorporate the constraint into the a priori. In practice we approximate by Importance Sampling

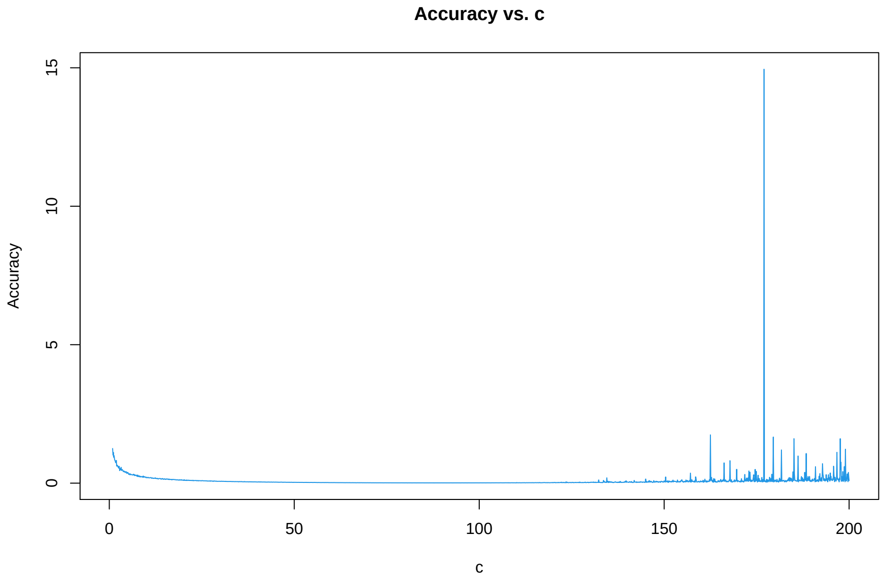
$$\hat{\theta}_n^{B(IS)} = \frac{1}{N}\sum_{i=1}^{N} \theta_i \hat{w}_N(\theta_i)$$

where

$$\hat{w}_N(\theta_i) = \frac{L_n(\theta_i)/f_Y(\theta_i)\, 1\!\!1_{(0,1)}(\theta_i)}{N^{-1}\sum_{i=1}^{N} L_n(\theta_i)/f_Y(\theta_i)}$$

for f_Y the proposal (on θ). One can choose $f_Y = \pi$, the non-informative proposal from the Bayesian approach. One can also take $f_Y = f_{\sigma^2} = \mathcal{N}(\hat{\theta}_n, \sigma^2)$, a proposal concentrated around the MLE with an hyper-parameter σ^2. The obtained approximation is a general Importance Sampling method and the hyper-parameter σ^2 should be tuned using importance weights.

Note that in this setting $h(y) = y$ and the intermediate function $f = L_n$; the importance is given to points θ that are likely given the observations $X_1, \ldots, X_n$. As n tends to ∞ then the unknown parameter θ^* becomes more and more likely. Thus $\hat{\theta}_n^B$ will be closer and closer to $\hat{\theta}_n$. Asymptotically, the two frequentist and Bayesian approaches are then equivalent.



The accuracy with respect to the hyperpameter $c = 1/\sigma^2$, where σ^2 is the variance of the proposal. For small variance the accuracy is large because of a large variance of g and for large one the accuracy may be large due to instability of the weights and their large variance.

Part IV

Markov Chain and Monte Carlo methods

Chapter 7

Markov chain

7.1 Transition kernel

We consider discrete time process (X_t) for $t = 0, 1, \ldots$.

Definition 18. *A Markov chain (X_t) is a sequence of random element of $\mathcal{X}$ ($= \mathbb{R}^d$ or $\{x_i\}$)such that*

$$F_{X_n|X_{n-1},\ldots,X_0} = F_{X_n|X_{n-1}}.$$

The Markov chain is homogeneous when $F_{X_n|X_{n-1}}$ is independent of $n \geq 1$.

Definition 19. *The kernel of an homogeneous Markov chain is the function K such that $F_{X_n|X_{n-1}=x}(dx) = K(x, dx)$. We have for any $x \in \mathcal{X}$ that $K(x, \cdot)$ is a probability measure and for any $A \in \mathcal{B}(\mathcal{X})$ (Borel set of $\mathcal{X}$) $K(\cdot, A)$ is a non-negative function on $\mathcal{X}$.*

When $F_{(X_0,X_1)}$ admits a density wrt the measure ν, we recognize

$$K(x, dy) = f_{X_1|X_0=x}(y)\nu(dy)\,.$$

Example 12. *Consider the random walk $X_n = X_{n-1} + Z_n$ where (Z_n) is iid f_Z. Then it is a Markov chain with kernel*

$$K(x, A) = \mathbb{P}(x + Z \in A) = \int_A f_Z(y - x)\nu(dy).$$

We denote $K(x, dy) = f_Z(y - x)\nu(dy)$ called the convolution kernel.

Remark. *Consider the finite discrete case $\mathcal{X} = \{x_i\}_{1\leq i\leq k}$ then*

$$P_{i,j} := K(x_i, \{x_j\}) = \mathbb{P}(X_n = x_j|X_{n-1} = x_i), \qquad 1 \leq i, j \leq k$$

constitutes a $k \times k$ matrix P that characterizes the kernel. It is called the transition matrix.

Definition 20. *A $k \times k$ matrix P is stochastic iff $P_{i,j\geq 0}$ and $\sum_{j=1}^k P_{i,j} = 1$.*

Proposition. *A transition matrix is a stochastic matrix and given a transition matrix there exists a Markov chain such as it is its transition matrix.*

Example 13. *The matrix*

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

is a transition matrix.

Definition 21. *We compose a kernel K to the right with an integrable mesurable function f*

$$Kf(x) = \int_{\mathcal{X}} f(y)K(x,dy)\,.$$

We compose to the left with any measure μ

$$\mu K(dy) = \int_{\mathcal{X}} K(x,dy)\mu(dx).$$

We compose it with itself recursively

$$K^{(n)}(x,dy) = \int_{\mathcal{X}} K^{(n-1)}(x,dz)K(z,dy)\,, \qquad n \geq 1\,.$$

Remark. *In the finite discrete case $\mathcal{X} = \{x_i\}_{1\leq i\leq k}$ we identify the compositions of the kernel with usual matrix multiplication on P; to the right with a vector*

$$(Pf)_i = \sum_{j=1}^{k} P_{i,j} f_j$$

for any $f \in \mathbb{R}^d$, to the right with a raw vector $\mu \in \mathbb{R}^d$

$$(\mu P)_j = \sum_{i=1}^{k} \mu_i P_{i,j}$$

and with itself

$$(P^n)_{i,j} = \sum_{\ell=1}^{k} (P^{n-1})_{i,\ell} P_{\ell,j}\,, \qquad n \geq 1\,.$$

The composition of a kernel has some probabilistic interpretation. We have

$$Kg(x) = \int_{\mathcal{X}} g(y)K(x,dy) = \int_{\mathcal{X}} g(y) f_{X_1|X_0=x}(y)\nu(dy) = \mathbb{E}[g(X_1) \mid X_0 = x]\,.$$

Moreover

$$\mu K(dy) = \int_{\mathcal{X}} K(x,dy)\mu(dx) = \int_{\mathcal{X}} f_{X_1|X_0=x}(y)\nu(dy)\mu(dx) = \int \mathbb{P}(X_1 \in dy \mid X_0 = x)\mu(dx).$$

It is the distribution of X_1 knowing that X_0 is μ-distributed. Note that μ does not have to be a probability measure but μK is always a probability measure and we denote

$$\mu K = \mathbb{P}_\mu(X_1 \in \cdot)\,, \qquad \mathbb{P}_x(X_1 \in \cdot) = \mathbb{P}_{\delta_{\{x\}}}(X_1 \in \cdot) = \mathbb{P}(X_1 \in \cdot \mid X_0 = x)\,.$$

Finally we have

$$K^{(n)}(x,dy) = \mathbb{P}(X_n \in dy \mid X_0 = x)\,.$$

We prove this identity for $n = 2$, the general case $n \geq 2$ follows by induction

$$\begin{aligned}\mathbb{P}(X_2 \in dy \mid X_0 = x) &= \mathbb{P}(X_2 \in dy, X_1 \in \mathcal{X} \mid X_0 = x)\\ &= \int \mathbb{P}(X_2 \in dy, X_1 = z \mid X_0 = x)\nu(dz)\end{aligned}$$

But, admitting the existence of densities, we have

$$\begin{aligned}\mathbb{P}(X_2 \in dy, X_1 = z \mid X_0 = x) &= \frac{f_{(X_2,X_1,X_0)}(y,z,x)}{f_{X_0}(x)}\nu(dy)\\ &= \frac{f_{(X_2,X_1,X_0)}(y,z,x)/f_{(X_1,X_0)}(z,x)}{f_{X_0}(x)/f_{(X_1,X_0)}(z,x)}\nu(dy)\\ &= f_{X_2|(X_1,X_0)=(z,x)}(y)f_{X_1|X_0=x}(z)\nu(dy)\\ &= f_{X_2|X_1=z}(y)f_{X_1|X_0=x}(z)\nu(dy)\end{aligned}$$

where the last identity comes from the Markov property. Combining these identities we obtain

$$\mathbb{P}(X_2 \in dy \mid X_0 = x) = \int f_{X_2|X_1=z}(y)\nu(dy)f_{X_1|X_0=x}(z)\nu(dz) = \int K(z,dy)K(x,dz)$$

and one recognizes the composition of the kernel.

Example 14.

1. *In the finite discrete setting with transition matrix P we identify $K^{(n)}$ and P^n.*
2. *In the iid setting, $K(x,dy) = F_X(dy)$ and $K^{(n)}(x,dy) = F_X(dy)f_X(x)$.*
3. *For the random walk $X_n = X_{n-1} + Z_n$ where (Z_n) is iid f_Z we have*

$$\begin{aligned}K^{(n)}(x,A) &= \mathbb{P}(X_n \in A \mid X_0 = x)\\ &= \mathbb{P}(x + S_n \in A \mid X_0 = x)\\ &= \int_A f_{S_n}(y-x)\nu(dy)\end{aligned}$$

where $S_n = \sum_{i=1}^n Z_i$ and

$$\begin{aligned}f_{S_n}(s) &= \int\cdots\int_{z_1+\cdots+z_n=s} f_Z(z_1)\cdots f_Z(z_n)\nu(dz_1)\cdots\nu(dz_n)\\ &= \int\cdots\int_{z_1,\ldots,z_{n-1}} f_Z(z_1)\cdots f_Z(z_{n-1})\\ &\quad f_Z(s-z_1-\cdots-z_{n-1})\nu(dz_1)\cdots\nu(dz_{n-1})\\ &= \int_{s_{n-1}}\cdots\int_{s_1} f_Z(s_1)f_Z(s_2-s_1)\nu(ds_1)\cdots f_Z(s-s_{n-1})\nu(ds_{n-1}).\end{aligned}$$

Thus we check the relation $K^{(n)}(x,dy) = \int_{\mathcal{X}} K^{(n-1)}(x,dz)K(z,dy)$ and the composition of the kernel is also a convolution.

Lemma (Chapman-Kolmogorov)**.** *Let K be a transition kernel then*

$$K^{(n)}(x,dy) = \int K^{(n-k)}(x,dz)K^{(k)}(z,dy),, \qquad 1 \le k \le n\,.$$

Proof.

$$\mathbb{P}(X_n \in dy \mid X_0 = x) = \int \mathbb{P}(X_n \in dy \mid X_{n-k} = z)\mathbb{P}(X_{n-k} \in dz \mid X_0 = x)\nu(dz)\,.$$

□

7.2 Stationarity and irreducibility

We denote $f\nu$ the probability measure $f\nu(dy) = f(y)\nu(dy)$ with density f wrt the reference measure ν.

Definition 22 (Invariant distribution)**.** *The distribution $f\nu$ is invariant iff $f\nu K = f\nu$.*

Then $f\nu K^{(n)} = f\nu$ and if $f\nu$ is a probability measure and $X_0 \sim f\nu$ then $X_t \sim f\nu$ for all $t \geq 0$.

In that case (X_t) is said to be stationary because for any $h \geq 0$ the trajectory $(X_t, \dots, X_{t+h})$ has the same distribution than $(X_0, \dots, X_h)$.

The distribution of $(X_t, \dots, X_{t+h})$ is characterized by $f\nu$ and K.

More precisely

$$F_{(X_t,\dots,X_{t+h})}(dx_t, \dots, dx_{t+h}) = f(x_t)\nu(dx_t)K(x_t, dx_{t+1}) \cdots K(x_{t+h-1}, dx_{t+h})$$

that we denote

$$F_{(X_t,\dots,X_{t+h})} = f\nu K \otimes \cdots \otimes K.$$

Definition 23 (Hitting time)**.** *For any Borel set A we denote τ_A the hitting time at A*

$$\tau_A = \inf\{t \geq 1 : \; X_t \in A\} ,$$

with the convention $\inf \emptyset = +\infty$.

A stopping time $\tau \in \mathbb{N} \cup \{\infty\}$ is a random element so that $\{\tau = t\} \in \sigma(X_0, \dots, X_t)$ for any $t \geq 1$.

Any hitting time is a stopping time. Indeed

$$\{\tau_A = t\} = \{X_1 \notin A, \dots, X_{t-1} \notin A, X_t \in A\} .$$

Example 15 (Example: coin tossing)**.** *Consider the game with two players with two fortunes $A, B \geq 1$.*

At each round t they toss a coin $Z_t = \pm 1$ so that if -1 then A gives to B one euro.

Let $X_t = X_{t-1} + Z_t$ be the gain of A starting from $X_0 = 0$.

It is a random walk and a Markov chain.

We have two hitting times $\tau_{(-A,-\infty)} = \inf\{t \geq 1 : \; X_t < -A\}$ and $\tau_{(B,\infty)} = \inf\{t \geq 1 : X_t > B\}$ that stop the game.

The ruin of A is the event $\{\tau_A < \tau_B\}$ w.p. $\mathbb{P}(\tau_A < \tau_B)$.

Proposition. *A Markov chain satisfies the strong Markov property*

$$F_{X_{\tau+1}|X_\tau,\dots,X_0} = F_{X_{\tau+1}|X_\tau}$$

for any topping time τ.

Proof. We implicitly work under $\tau < \infty$ and we have

$$\begin{aligned}
&\mathbb{P}(X_{\tau+1} \in A_{\tau+1}, \tau < \infty \mid X_\tau \in A_\tau, \dots, X_0 \in A_0)\\
&\qquad = \sum_{t=1}^{\infty} \mathbb{P}(X_{\tau+1} \in A_{\tau+1} | X_\tau \in A_\tau, \dots, X_0 \in A_0)\\
&\qquad = \sum_{t=1}^{\infty} \frac{\mathbb{P}(X_{t+1} \in A_{t+1}, \tau = t, X_t \in A_t, \dots, X_0 \in A_0)}{\mathbb{P}(X_\tau \in A_\tau, \dots, X_0 \in A_0)}\\
&\qquad = \sum_{t=1}^{\infty} \mathbb{P}(X_{t+1} \in A_{t+1} \mid \tau = t, X_t \in A_t, \dots, X_0 \in A_0)\\
&\qquad\quad \mathbb{P}(\tau = t \mid X_\tau \in A_\tau, \dots, X_0 \in A_0)\\
&\qquad = \sum_{t=1}^{\infty} \mathbb{P}(X_{\tau+1} \in A_{\tau+1} \mid \tau = t, X_\tau \in A_\tau)\mathbb{P}(\tau = t \mid X_\tau \in A_\tau, \dots, X_0 \in A_0)\\
&\qquad = \mathbb{P}(X_{\tau+1} \in A_{\tau+1} \mid X_\tau \in A_\tau) \sum_{t=1}^{\infty} \mathbb{P}(\tau = t \mid X_\tau \in A_\tau, \dots, X_0 \in A_0)\\
&\qquad = \mathbb{P}(X_{\tau+1} \in A_{\tau+1} \mid X_\tau \in A_\tau).
\end{aligned}$$

where we used that $\{\tau = t\} \in \sigma(X_0, \dots, X_t)$, the Markov property and the homogeneity. The desired result follows by dividing by $\mathbb{P}(\tau < \infty)$. □

7.3 Accessibility and irreducibility

Definition 24. *A Borel set B is accessible from a Borel set A iff $\mathbb{P}_A(\tau_B < \infty) = \mathbb{P}(\tau_B < \infty \mid X_0 \in A) > 0$.*

They communicate if A is accessible from B.

Note that $\{x\}$ is not accessible when X_t is continuous.

Definition 25. *A Markov chain is ν-irreducible on the state space $\mathcal{X}$ (dicrete or continuous) if any Borel sets A and B of $\mathcal{X}$ communicate when $\nu(A)\nu(B) > 0$.*

Definition 26 (The period, discrete case)**.** *The period $d(x)$ of $\{x\}$ is the greatest common divisor of $\tau_{\{x\}} \geq 1$ a.s. starting from x.*

Note that $\tau_{\{x\}}$ from x is called the return time to $\{x\}$

Proposition. *A discrete finite Markov chain (X_t) is such that $d(x) = d$ is independent of $x \in \mathcal{X}$.*

Then d is the period of the Markov chain and it is aperiodic iff $d = 1$

Proof. It is enough to show that if x_i and x_j communicate then $d(x_i) = d(x_j)$. If x_j is accessible from x_i it means that there exists $n_i \geq 1$ such that $P_{i,j}^{n_i} > 0$. Similarly there exists $n_j \geq 1$ such that $P_{j,i}^{n_j} > 0$. Let k be such that $P_{i,i}^k > 0$ such that $d(x_i)$ divides k. Then

$$P_{j,j}^{n_j+k+n_i} = P_{j,i}^{n_j} P_{i,i}^k P_{i,j}^{n_i} > 0$$

so that $P_{i,i}^{2k} > 0$ and thus $P_{j,j}^{n_j+2k+n_i}$. One deduces that $d(j)$ divides both $n_j + k + n_j$ and $n_j + 2k + n_j$ thus it divides the difference k. It also divides the greatest common divisor of such k which is $d(i)$. But as the roles of i and j are the same we also have that $d(i)$ divides $d(j)$ and the result follows. □

Example 16.

1. *(X_t) iid is irreducible on $\mathcal{X} = Supp(X)$ and it is aperiodic,*

2. *Consider the stochastic matrix*
$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$
Then (X_t) is irreducible on $\mathcal{X} = \{x_1, x_2, x_3\}$. It is aperiodic and even strongly aperiodic on $\{x_1\}$ and $\{x_3\}$.

3. *Consider the stochastic matrix*
$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$
Then (X_t) is reducible to $\mathcal{X} = \{x_2, x_3\}$.

4. *Consider the stochastic matrix*
$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}.$$
Then (X_t) is irreducible to $\mathcal{X} = \{x_1, x_2, x_3\}$. Its period is 2.

Example 17 (Example: coin tossing)**.** *In the coin tossing game with infinite fortune $A = B = \infty$ and $p = \mathbb{P}(toss = head) \in (0, 1)$ then the gain (X_t) of the player A is an irreducible Markov chain on $\mathbb{Z}$ and its periodicity is 2.*

In the continuous case, one has to adapt the notion of aperiodicity since the return time to any point x is infinite. We define

Definition 27. *The Markov chain (X_t) satisfy the minorization condition iff there exist a probability measure μ, a small set C that is accessible, a time $t \geq 1$ and $\epsilon > 0$ such that*
$$K^{(t)}(x, dy) > \epsilon\mu(dy), \qquad x \in C.$$

A Markov chain is Harris if it is ν-irreducible and satisfies the minorization condition with $t = 1$.

Note that in the discrete finite irreducible case $C = \{x\}$ is a small set for any $x \in \mathcal{X}$ and $d(x)$ is the greatest common divisor of the times t satisfying the minorization condition.

Definition 28. *The Markov chain is strongly aperiodic iff it satisfies the minorization condition with $t = 1$ and $\mu(C) > 0$.*

Note that a finite discrete Markov chain is strongly aperiodic if it has a non null diagonal element $P_{i,i} > 0$ for some i.

7.4 Atom and regeneration

Definition 29 (Atom)**.** *A set A that is accessible from any $x \in \mathcal{X}$ is an atom iff $\mathbb{P}_x(X_1 \in dy) = \mu$ is independent of $x \in A$.*

1. In the finite discrete case any $\{x\}$ is an atom if the Markov chain is irreducible,
2. In the continuous case there is no atom.

Proposition (Regeneration)**.** *Let A be an atom and τ_A its associated hitting time. Then*

$$(X_0, \dots, X_{\tau_A}) \textit{ is independent of } (X_{\tau_A+1}, \dots).$$

Proof. Use the strong Markov property so that

$$\begin{aligned} F_{(X_{\tau_A+h},\dots,X_{\tau_A+1})|X_{\tau_A},\dots,X_0} &= F_{(X_{\tau_A+h},\dots,X_{\tau_A+1})|X_{\tau_A}} \\ &= \mathbb{P}_{X_{\tau_A}} K \otimes \cdots \otimes K \\ &= \mu K \otimes \cdots \otimes K \end{aligned}$$

since $X_{\tau_A} \in A$ an atom. □

The problem is that the atom notion suits only to the discrete case. In the continuous case, the solution consists in enlarging the Markov chain with a discrete iid sequence. For simplicity we consider only the strongly aperiodic case.

Theorem (Nummelin)**.** *Assume a Markov chain (X_t) is Harris. Consider the chain (X_t, δ_t) with δ_t iid $\sim \mathcal{B}ern(\epsilon)$ and $X_{t+1} \sim K(X_t, \cdot)$ if $X_t \notin C$, else*

$$X_{t+1} \sim \begin{cases} \mu & if \qquad \delta_t = 1 \\ \dfrac{K(X_t, \cdot) - \epsilon\mu}{1-\epsilon} & else. \end{cases}$$

Then the Markov chain (X_t, δ_t) admits an atom $A = C \times \{1\}$ and the marginal (X_t) is unchanged.

Note that (X_t, δ_t) has a well defined transition kernel since $(K(X_t, \cdot) - \epsilon\mu)/(1-\epsilon)$ is also a transition kernel (non-negative and summing up to 1).

Proof. It is obvious that when $X_{t+1} \sim \mu$ the Markov chain (X_t, δ_t) regenerates independently of its own past. We also have

$$\mathbb{P}_x(X_1 \in dy) = \epsilon\mu(dy) + (1-\epsilon)\frac{K(x, dy) - \epsilon\mu(dy)}{1-\epsilon} = K(x, dy)$$

□

Note that the Markov chain (X_t, δ_t) regenerates and it is also the case of the marginal

$$(X_0, \dots, X_{\tau_A}) \text{ is independent of } (X_{\tau_A+1}, \dots).$$

However A is not an atom for (X_t) itself, we call it a pseudo-atom.

7.5 Recurrence and ergodicity

Let (X_t) be a Harris Markov chain starting at $t = 1$ from $X_1 \sim \mu$.

Define $\tau_A(k) = \inf\{t \geq \tau_A(k-1)+1 \ : X_t \in A\}$ the successive hitting times to the atom $(\tau_A(0) = 0)$.

Proposition. *The cycles*

$$(X_{\tau_A(k-1)+1}, \dots, X_{\tau_A(k)})_{k\geq 1}$$

are iid. Thus (X_t) is stationary, its marginal density f is satisfying $f\nu K = f\nu$.

Proof. We start y showing that the lengths of the cycles $R_A(k) = \tau_A(k) - \tau_A(k+1)$ are iid. Indeed

$$\mathbb{P}(R_A(k) > t) = \int_{A^c} \cdots \int_{A^c} \mu \underbrace{K \otimes \cdots \otimes K}_{t}$$

is independent of k and of the other hitting times $\tau_A(k)$. Then given $R_A(k) = t$ the distribution of the k-th cycle is

$$\mu K \otimes \cdots \otimes K\,,$$

independent of k and the other cycles. □

Note that the cycles have a specific case of (discrete) mixture distribution, where the mixing distribution variable is their lengths.

Define the number of passage

$$\eta_A(N) = \sum_{t=1}^{N} \delta_A(X_t) = \#\{1 \leq k \leq N;\, \tau_A(k) \leq N\}.$$

We have

Theorem (Renewal theorem)**.** *If (X_t) is a Harris Markov chain such that $\mathbb{E}[\tau_A(1)] < \infty$ then, as $N \to \infty$ we have*

$$\frac{\eta_A(N)}{N} \to \frac{1}{\mathbb{E}[\tau_A(1)]}\,, \qquad a.s.$$

We say that (X_t) is positive recurrent and we notice that $\mathbb{E}[\tau_A(1)] = \mathbb{E}_\mu[\tau_A] = \mathbb{E}[R_A(1)]$. We then say that the Markov chain is Harris positive recurrent.

On the opposite, it is negative recurrent if $\mathbb{E}[\tau_A(1)] = \infty$ and then $\eta_A(N)/N \to 0$.

Proof. We have $\{\eta(N) = j\} = \{\tau_A(j) \leq N < \tau_A(j+1)\}$ so that

$$\frac{\tau_A(j)}{j} \leq \frac{N}{\eta(N)} < \frac{\tau_A(j+1)}{j}.$$

As N goes to ∞ we have that j also goes to ∞ otherwise it contradicts $R_A(1) < \infty$ which in turns contradicts $\mathbb{E}[R_A(1)] < \infty$. Thus we conclude by a sandwich argument since the SLLN applies

$$\frac{\tau_A(j)}{j} = \frac{\sum_{t=1}^{j} R_A(t)}{j} \to \mathbb{E}[R_A(1)] \qquad a.s.\,, \qquad j \to \infty\,.$$

□

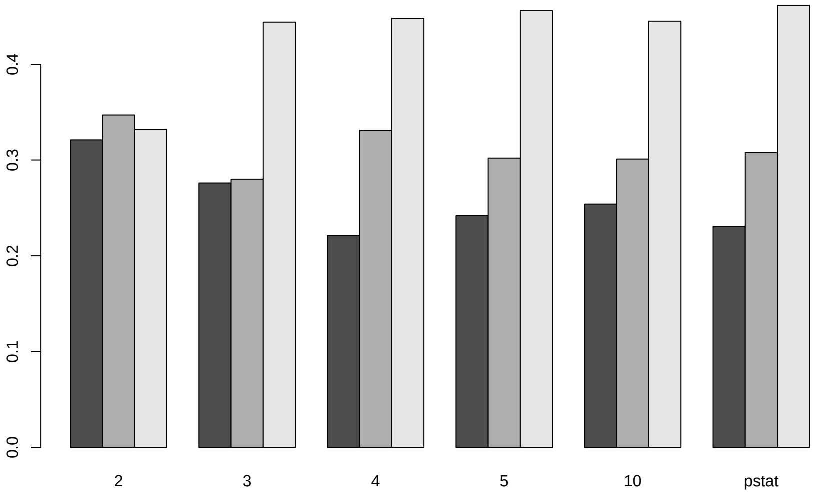
Example 18.

1. *In the iid case, the chain is positive recurrent as $\tau_A = 1$.*
2. *Consider the stochastic matrix*

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix} .$$

Then (X_t) is irreducible, any state $\{x_1\}$, $\{x_2\}$ and $\{x_3\}$ is an atom.

There exists a raw vector f invariant such that $fP = f$. Moreover $f(x_i) = 1/\mathbb{E}_{x_1}[\tau_{x_1}] > 0$ and the Markov chain is positive recurrent.



The evolution of the distribution of a 3-state irreducible Markov chain. Starting from a uniform initial distribution, the empirical distribution converges rapidly to the unique invariant distribution.

We are now ready to state our main result on Markov chain.

Theorem. *Assume that (X_t) is a Harris positive recurrent Markov chain starting at x arbitrary and that g satisfies $\mathbb{E}_\mu[(\sum_{t=1}^{\tau_A} g(X_t))^2] < \infty$. Then there exists some $V(g) > 0$ such that*

$$\sqrt{N}\Big(\frac{1}{N}\sum_{t=1}^{N} g(X_t) - \int g(x)f(x)d\nu(x)\Big) \to \mathcal{N}(0, V(g)) , \qquad N \to \infty ,$$

where f is the unique invariant density.

Proof. The proof starts by showing the existence of a unique invariant probability measure. The proof is constructive. Using cycles, we have

$$\frac{1}{N}\sum_{t=1}^{N} g(X_t) = \frac{1}{N}\sum_{k=1}^{\eta_A(N)} \sum_{t=\tau_A(k-1)+1}^{\tau_A(k)} g(X_t) = \frac{\eta_A(N)}{N}\frac{1}{\eta_A(N)}\sum_{k=0}^{\eta_A(N)-1} S_k(g) ,$$

where $(S_k(g))_{k\geq 1}$ constitutes an iid sequence. Since $\eta_A(N) \to \infty$ as $N \to \infty$ a.s. and $\mathbb{E}[|S_1(g)|] < \infty$ we can apply the SLLN so that a.s. as $N \to \infty$

$$\frac{1}{\eta_A(N)} \sum_{k=0}^{\eta_A(N)-1} S_k(g) = \frac{S_0(g)}{\eta_A(N)} + \frac{1}{\eta_A(N)} \sum_{k=1}^{\eta_A(N)-1} S_k(g) \to 0 + \mathbb{E}[S_1(g)]\,.$$

Combining this with the renewal theorem we get

$$\frac{1}{N}\sum_{t=1}^{N} g(X_t) \to \frac{\mathbb{E}_\mu[\sum_{t=0}^{\tau_A} g(X_t)]}{\mathbb{E}_\mu[\tau_A]}\,.$$

Moreover $(X_t)_{t\geq \tau_A+1}$ is stationary and its distribution is the invariant measure $f\nu$. Thus the limit can only be equal to

$$\mathbb{E}[g(X)] = \int g(y)f(y)\nu(dy)\,.$$

As the limit in the SLLN is unique it means that f is unique and that

$$\mathbb{E}_\mu\Big[\sum_{t=0}^{\tau_A} g(X_t)\Big] = \mathbb{E}_\mu[\tau_A]\mathbb{E}[g(X)]\,.$$

Similarly the CLT applies on $(C_k(g))_{k\geq 1}$ iid since $\mathbb{E}[C_1(g)^2] < \infty$ by assumption and we get

$$V(g) = \frac{1}{\mathbb{E}_\mu[\tau_A]^2}\mathbb{E}_\mu\Big[\Big(\sum_{t=1}^{\tau_A} g(X_t)\Big)^2\Big] - \mathbb{E}[g(X)]^2\,.$$

□

Note that $V(g)$ is small when g^2 is small but it depends also on the value of $\mathbb{E}_\mu[\tau_A^2]$.

Chapter 8

Metropolis-Hasting algorithm

8.1 The algorithm

Let $I = \int g(x)f(x)dx$ where f is a target density known up to a constant.

We would like to simulate following the target density f but it is more complicated than approximating h. Typically, if $h \geq 0$ we would like to simulate under the distribution $f = h/\int h$.

The MH algorithm will generate a Markov chain such that $f\nu K = f\nu$ where ν is the measure of reference and the target distribution F is proportional to $f\nu$. Note that the equation $f\nu K = f\nu$ is free of a multiplicative factor, i.e. $\alpha f\nu K = \alpha f\nu$ for any $\alpha > 0$. Thus the normalizing constant does not contribute to the determination of the kernel K.

It returns a sample X_t which is approximatively distributed as f but not iid.

Algorithm 10: The Metropolis-Hasting algorithm

```
Parameters: $g$, $f$ and a conditional density $f_{Y|X=x}$.
Initialization: $X_0 = x \in \mathrm{Supp}(f)$ arbitrary
For each step $t \geq 0$ Do
    • Sample $Y_{t+1} \sim f_{Y|X=X_t}$,
    • Choose $X_{t+1} = \begin{cases} Y_{t+1} & w.p. \quad \rho(X_t, Y_{t+1}) \\ X_t & w.p. \quad 1-\rho(X_t, Y_{t+1}) \end{cases}$
    where the MH acceptance probability is
        $$\rho(x,y) = \min\Big(\frac{f(y)f_{Y|X=y}(x)}{f(x)f_{Y|X=x}(y)}, 1\Big)$$
Return $\hat{I}_N^{(MH)} = \frac{1}{N}\sum_{t=1}^N g(X_t)$.
```

The acceptance probability requires the knowledge of f only up to a constant (if $f = h/\int h$ then $f(y)/f(x) = h(y)/h(x)$ and the knowledge of $I = \int h$ is not required!).

Note that $f_{Y|X=x} = f_Y$ is possible. Then we talk about the independence Metropolis algorithm.

A common choice is $f_{Y|X=x}(y) = \varphi_{\sigma^2}(y-x)$ where φ the density of $\mathcal{N}(0,\sigma^2)$, then the acceptance probability simplifies as

$$\rho(x,y) = \min\Big(\frac{f(y)}{f(x)}, 1\Big)$$

as for any symmetric proposal satisfying $f_{Y|X=x}(y) = f_{Y|X=y}(x)$. The proposal is accepted if it moves to a more important step $f(Y_{t+1}) \geq f(X_t)$. Otherwise, if the move is less important, there is still some chance to move but with low probability $f(Y_{t+1})/f(X_t) \leq 1$.

8.2 Detailed balance condition

In order to design the MH algorithm, one has to check that its kernel admits f as an invariant density. For that, we use the detailed balance condition

Definition 30. *A Markov chain satisfies the detailed balance condition wrt f iff*

$$K(x, dy)f(x)\nu(dx) = K(y, dx)f(y)\nu(dy)\,.$$

In some sense the role of x and y must be symmetric. We have

Proposition. *A Markov chain satisfying the detailed balance condition admits $f\nu$ as an invariant distribution.*

Proof. We have

$$\begin{aligned} f\nu K = \int K(x, dy)f(x)\nu(dx) &= \int f(x)K(x, dy)\nu(dx) \\ &= \int f(y)K(y, dx)\nu(dy) = f(y)\nu(dy)\int K(y, dx) = f\nu\,. \end{aligned}$$

□

It remains to show that MH satisfies the detailed balance condition. The difficulty is to identify the transition Kernel that is a mixture of discrete ($\delta_{\{X_t\}}$) and continuous ($f_{Y|X=X_t}$). More precisely

$$K^{(MH)}(x, dy) = \rho(x, y)f_{Y|X=x}(y)\nu(dy) + (1 - \rho(x, y))\delta_{\{x\}}(dy)\,.$$

We first check the symmetry property of the acceptance probability

$$f_{Y|X=x}(y)\rho(x, y)f(x) = \min(f(y)f_{Y|X=y}(x), f(x)f_{Y|X=x}(y)) = \rho(y, x)f_{Y|X=y}(x)f(y)\,.$$

Moreover, we have

$$(1 - \rho(x, y))\delta_{\{x\}}f(x)(dy)\nu(dx) = (1 - \rho(y, x))\delta_{\{y\}}(dx)f(y)\nu(dy)$$

since the Dirac measure vanishes iff $x \neq y$ so that the role of x and y are interchangeable since $y = x$. Thus we infer that

$$\begin{aligned} K^{(MH)}(x, dy)f(x)\nu(dx) &= \rho(y, x)f_{Y|X=y}(x)f(y)\nu(dy)\nu(dx) + f(y)(1 - \rho(y, x))\delta_{\{y\}}(dx)\nu(dy) \\ &= K^{(MH)}(y, dx)f(y)\nu(dy) \end{aligned}$$

and the desired result follows.

8.3 Convergence analysis

So far we design the MH algorithm so that f is an invariant distribution. Its uniqueness and invariance will follow from a minorization condition on the conditional proposal that will implies the minorization condition on K. Then the Markov chain will be Harris and it will be Harris positive recurrent because it admits an invariant density (the null recurrent case is excluded because the invariant measures then do not admit a density). More specifically:

Theorem (MH, continuous case ν Lebesgue). *Assume that $\mathcal{X} = Supp(f)$ is connected such that $\nu(\mathcal{X}) > 0$, that $\mathbf{m} < f < \mathbf{M}$ on $\mathcal{X}$ and that there exist $\mathbf{c}, \delta > 0$*

$$f_{Y|X=x}(y) > \mathbf{c}, \qquad |x-y| < \delta, x, y \in \mathcal{X}\,.$$

Then, if $\mathbb{E}_\mu[(\sum_{t=1}^{\tau_A} g(X_t))^2] < \infty$, we have

$$|\hat{I}_N^{(MH)} - I| = O_{\mathbb{P}}(\sqrt{V(g)/N})$$

where $V(g)$ is small when g^2 is small.

Proof. The proof is decomposed into two steps. We will first show that any neigborhood of any point $y \in \mathcal{X}$ is accessible from $X_0 = x \in \mathcal{X}$. Thus it would mean that (X_t) is $\nu_{|\mathcal{X}}$-irreducible where $\nu_{|\mathcal{X}}$ is the Lebesgue measure restricted to $\mathcal{X}$. By connected we meant that one can go from x to y in m steps smaller than $\delta/2$. For each steps $[x_i, x_{i+1}]$ where $\delta/2 < |x_i - x_{i+1}| < \delta$ we then have $f_{Y|X=x_i}(x_{i+1}) > \epsilon$,

$$\begin{aligned} K^{(MH)}(x_i, dx_{i+1}) &\geq \rho(x_i, x_{i+1}) f_{Y|X=x_i}(x_{i+1})\nu(dx_{i+1}) \\ &\geq \min(f(x_{i+1}) f_{Y|X=x_{i+1}}(x_i)/f(x_i), f_{Y|X=x_i}(x_{i+1}))\nu(dx_{i+1}) \\ &\geq \epsilon \frac{\mathbf{m}}{\mathbf{M}} \nu(dx_{i+1}) \end{aligned}$$

and thus

$$\begin{aligned} K^{(MH),(m)}(x, dy) &\geq \int \cdots \int K^{(MH)}(x, dx_1) \ldots K^{(MH)}(x_{m-1}, dy) \\ &\geq \Big(\epsilon \frac{\mathbf{m}}{\mathbf{M}}\Big)^m \int \cdots \int \nu(dx_1) \cdots \nu(dx_{m-1})\nu(dy) \\ &\geq \Big(\epsilon \frac{\mathbf{m}}{\mathbf{M}}\Big)^m \nu(B_1) \cdots \nu(B_m)\nu(dy) \end{aligned}$$

where $B_1, \ldots, B_m$ are the sets containing the consecutive steps $x_1, \ldots, x_m$.

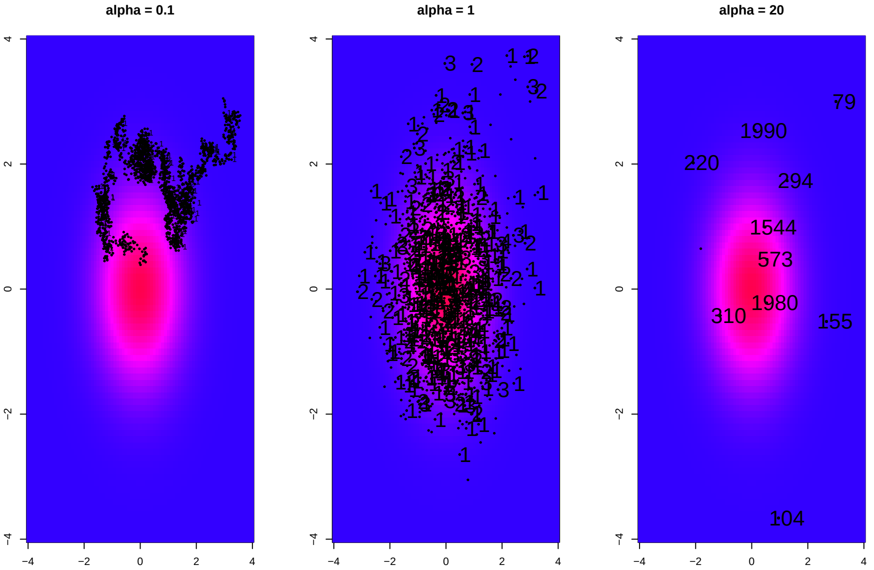
We get that any set $C = B(y, \delta) \cap \mathcal{X}$ is accessible. Moreover the minorization condition is satisfied with $m = 1$ as for any $y \in \mathcal{X}$ we have

$$K^{(MH)}(x, dy) \geq \Big(\epsilon \frac{\mathbf{m}}{\mathbf{M}}\Big)\nu(dy) = \varepsilon \nu(dy)\,, \qquad y \in C\,.$$

Thus it is Harris positive recurrent since it admits an invariant density. The CLT applies and we obtain the desired result. □

Note that $f = h$ is an optimal choice when $h \geq 0$, thus $g = 1$ and $\mathbb{E}_\mu[(\sum_{t=1}^{\tau_A} g(X_t))^2] = \mathbb{E}_\mu[\tau_A^2]$. Thus the CLT is implied by

$$\mathbb{E}_\mu[\tau_A^2] = \frac{\mathbb{E}_\mu[\tau_C^2]}{\varepsilon^2} = \frac{\mathbb{E}_\mu[\tau_C^2] M^2}{\epsilon^2 m^2}\,.$$



Visualization of MH samples of gaussian random vectors. Small steps in the proposal reduce the velocity of the algorithm that spends time to move from an initial guess. Large steps makes the rejection ratio large thus the algorithm get stuck in few different positions (number of rejections in digit numbers).