Monte-Carlo inference

Alexandra Carpentier A.CARPENTIER@STATSLAB.CAM.AC.UK

Editor:

1. Random number generation

A first challenge is to be able to generate random numbers. Indeed, computers cannot at this stage generate truly random numbers.

The sequences computer generate are called pseudo-random numbers. They are not random, but they have the same kind of properties as truly random numbers.

1.1 Generation of uniform random numbers

We are going to consider more in details the generation of uniform random variables on $[0,1]$ ($\mathcal{U}_{[0,1]}$). Indeed, as we will see later, we can somehow use the randomness in uniform random variables to simulate any other real valued random variables.

1.1.1 Uniform random number generators

A pseudo-random number generator for uniform random variables on $[0,1]$ is aimed at generating random variables $u_0, u_1, u_2 \dots$ that seem uniform.

Congruential generator. A congruential generator with module $M \in \mathbb{N}$, multiplier a, shift c and seed $x_0 \in \{0, 1, \dots, M-1\}$ is defined setting

$$x_i \equiv (ax_{i-1} + c) \mathbf{mod} M, \qquad i \in \leq M-1,$$

and then considering $u_i = x_i/M$.

Note that this sequence has period at most M. M should then be choosen very large. Note that a should also be chosen quite large, and such that the sequence has maximal period.

For instance the NAG Fortran generator G05CAF uses $M = 2^{59}$ and $a = 13^{10}$ (and $c = 0$).

Note on other generators. There are other generators as e.g. the feedback shift generator. Most of them are of the form

$$x_n = f_n(x_{n-1}, \dots, x_0),$$

and geometric-like sequenecs such as for the congruential generator provide quite good results.

1.1.2 Considerations on such generators

Representation issues.

- The real numbers are not all represented by a computer (each floating number is encoded on a given number of bits).
- The density of floating numbers represented on a computer is not uniform.

Goodness of fit. It is possible to test the quality of a uniform random number generator by performing a goodness of fit test (with respect to $\mathcal{U}_{[0,1]}$) on the generated data. We recall a well known goodness of fit statistics, the **chi-squared goodness of fit test** (although many other exist).

Consider a set of pseudo-random numbers of size n. We start by dividing the interval $[0, 1]$ in k bins of equal size, and set

$$\hat{\chi}_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

where o_i is the number of pseudo-random numbers in bin i and $e_i = n/k$. The critical values for this statistic correspond to the quantiles of a χ^2 distribution with $k - 1$ degrees of freedom.

1.2 Random variable generation

In the last subsection, we discussed the simulation of uniform random variables. However, most of the time, one is interested by generating random variables according to other distributions. We are now going to discuss some methods for simulating such random variables.

We assume now that we are able to simulate uniform random variables on $[0, 1]$.

1.2.1 Method of inversion

We first present the method in the discrete case since it is very intiutive in this case. We then present more formally the method for any real valued random variable.

Intuition in the discrete case Simulation of a Bernouilli random variable. A very simple random variable is a Bernouilli random variable of parameter $1 - p$.

- Simulate $U \sim \mathcal{U}_{[0,1]}$.
- Set $X = 1$ if $U \geq p$, and $X = 0$ otherwise.

Then as wished, we have $\mathbb{P}(X = 1) = 1 - p$ and $\mathbb{P}(X = 0) = p$, and X is indeed a Bernouilli random variable.

Figure 1: Illustration of the inversion method for Bernouilli random variables.

General discrete case. We consider now a discrete random variable X with M mass points $\{m_1, \dots m_M\}$ with probability $\{p_1, \dots, p_M\}$ (such that $\sum_{i=1}^{M} p_i = 1$ and $p_i \geq 0$ for any $i \in \{1, \dots, M\}$), i.e. a random variable such that for any $i \in \{1, \dots, M\}$

$$\mathbb{P}(X = m_i) = p_i.$$

We assume without loss of generality that $m_1 < m_2 < \dots < m_M$.

The CDF of such a random variable is

$$F(x) = \mathbb{P}(X \leq x) = \sum_{i=1}^{\lfloor x \rfloor} p_i.$$

Let us write $F(m_i) = F_i$ since these are the points of interest. Note that $p_i = F_i - F_{i-1}$.

An intuitive way of simulating according to F is as follows:

- Simulate $U \sim \mathcal{U}_{[0,1]}$.
- Set $X = m_i$ if $F_{i-1} \leq U \leq F_i$.

Similarly as what happens in the Bernouilli case, we have for any $i \in \{1, \dots, M\}$ that

$$\mathbb{P}(X = m_i) = \mathbb{P}(U \in [F_{i-1}, F_i]) = F_i - F_{i-1} = p_i.$$

Figure 2: Method of inversion.

General method We want to simulate a scalar random variable with CDF F that is continuous (no point mass) and strictly increasing.

An important classic lemma is as follows.

Lemma 1 *Let $X \sim F$ be a scalar random variable defined on $\mathcal{X} \subset \mathbb{R}$. Assume that F is strictly increasing and continuous. Then*

$$U = F(X) \sim \mathcal{U}_{[0,1]}.$$

Proof Let $U = F(X)$. Since $F : \mathcal{X} \to [0,1]$, we know that U takes values in $[0,1]$. We have for any $u \in [0,1]$ that

$$\begin{aligned} \mathbb{P}(U \leq u) &= \mathbb{P}(F(X) \leq u) \\ &= \mathbb{P}(X \leq F^{-1}(u)) \\ &= F(F^{-1}(u))) = u. \end{aligned}$$

where F^{-1} is the inverse of F (which exists since F is strictly increasing), and that is such that $F(F^{-1}(u))) = u$ for any u (F is continous).

Since $\mathbb{P}(U \leq u) = u$, then U is an uniform random variable. ■

This simple lemma contains actually the central idea of the so-called *inversion method* for generating random variables. Indeed, Lemma 1 implies that if U is an uniform ramdom

variable, then

$$X = F^{-1}(U) \sim F.$$

Indeed, in the same way, for any $x \in \mathcal{X}$, we have

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) = F(x). \end{aligned} \tag{1}$$

Exercise 1: Explain how to implement a generator for $Exp(\lambda)$ random variables.

Exercise 2: Explain how to implement a generator from a mixture density $\sum_{i=1}^{k} w_i f_i$ where the w_i are positive weights of sum 1, and f_i are densities.

Remark 1: We made for simplicity the assumption that F is strictly increasing, but this method can be extended to any distributions F that have support in $\mathbb{R}$ by just considering a pseudo-inverse of F instead of the inverse.

Remark 2: We however need that F is continuous so that Lemma 1 is verified. However, even when Lemma 1 is not verified, the argument in Equation (1) still holds (verify this).

Remark 3: If you can generate random variables according to F and want to generate random variables according to G, then for reasons similar as above

$$G^{-1}(F(X))$$

will be distributed according to G.

Remark 4: Even if no close form exists for the inverse of the CDF of the distribution you want to simulate, you can approximate it by finding the function $x(u)$ such that $F(x(u)) = u$ for any u.

1.2.2 Rejection sampling

We present now another method for generating according to a density f if we can compute f at any point but if simulating according to it is somehow complicated.

First intuition Assume that we can generate an uniform[1] random variable U on some domain $\mathcal{U}$. We write μ the probability associated to this uniform random variable. Assume that now we wish to simulate a random variable X uniformly on some domain $\mathcal{X} \subset \mathcal{U}$ (i.e. according to $\mu(.|\mathcal{X})$).

A very natural way to proceed is as follows.

1. Simulate U uniformly on $\mathcal{U}$.
2. If $U \in \mathcal{X}$, set $X = U$, otherwise reject U and go to step 1.

Lemma 2 *Assume that $\mu(\mathcal{X}) = p > 0$. Then X is distributed uniformly on $\mathcal{X}$.*

1. Uniform is really not important here, we just write it this way to make the concept simpler. Actually, this assumption is not going to appear in the proof.

Figure 3: First intuition of rejection sampling.

Proof Let $U_1, \dots, U_n, \dots$ be i.i.d. $\mathcal{U}$. Let N be the first index i such that $U_i \in \mathcal{X}$ (set $N = \infty$ if no such index exists).

By definition, $X = U_N$.

Let us first study the distribution of the $U_i | U_i \in \mathcal{X}$. By definition, it is distributed according to $\mu(.|\mathcal{X})$.

Now let $k > 0$. The random variable $X|N = k$ is actually egal to $U_k | U_1 \notin \mathcal{X}, U_2 \notin \mathcal{X}, \dots, U_{k-1} \notin \mathcal{X}, U_k \in \mathcal{X}$. Since the U_i are independent, it actually has the same distribution as $U_k | U_k \in \mathcal{X}$, i.e. $\mu(.|\mathcal{X})$. This holds for any k, therefore $X|N < \infty$ is distributed according to $\mu(.|\mathcal{X})$.

Now, we have

$$\begin{aligned}
\mathbb{P}(N = \infty) &= \mathbb{P}(U_1 \notin \mathcal{X}, U_2 \notin \mathcal{X}, \dots, U_n \notin \mathcal{X}, \dots) \\
&= \lim_{n \to \infty} \mathbb{P}(U_1 \notin \mathcal{X}, U_2 \notin \mathcal{X}, \dots, U_n \notin \mathcal{X}) \\
&= \lim_{n \to \infty} (1 - p)^n = 0.
\end{aligned}$$

This implies that the distribution of $X|N < \infty$ is equal to the distribution of X (since $\{N = \infty\}$ is negligible), and this concludes the proof. ■

Remark: The condition $\mu(\mathcal{X}) = p > 0$ is actually very important, since if it is not verified, then no samples U will be sampled in $\mathcal{X}$. It is actually important that p is not too small, so that the even $\{U \in \mathcal{X}\}$ has not a too small probability and so that some samples

are actually generated. As a matter of fact, it is possible to prove that if n samples are generated uniformly on $\mathcal{U}$, then asymptotically we have pn samples generated uniformly on $\mathcal{X}$.

This idea can actually be generalized to arbitrary densities, if one thinks about *weighting* the reject.

Generalisation to arbitrary densities Suppose that we have two densities f and g defined on the same domain $\mathcal{X} \subset \mathbb{R}$.

Suppose that sampling according to f is difficult, but that it is easy to sample according to some other density g such that there exists a constant $M \in [1, \infty[$ such that

$$f(x) \leq Mg(x), \forall x \in \mathcal{X}. \tag{2}$$

g is called a majorising or envelope density, and it dominates f.

Similarly to what we described above, we propose the following procedure

1. Simulate Y according to g, and some independent $U \sim \mathcal{U}_{[0,1]}$.
2. If $U \leq \frac{f(Y)}{Mg(Y)}$, set $X = Y$, otherwise go to step 1.

Figure 4: Rejection sampling.

Theorem 3 *If the condition in Equation* (2) *is verified, then the output X from the rejection algorithm has density f.*

Proof We have for any $x \in \mathcal{X}$

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}) \\ &= \frac{\mathbb{P}(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)})}{\mathbb{P}(U \leq \frac{f(Y)}{Mg(Y)})}. \end{aligned} \tag{3}$$

Now since Y and U are independent

$$\begin{aligned}\mathbb{P}(U \leq \frac{f(Y)}{Mg(Y)}) &= \int_{-\infty}^{+\infty} \mathbb{P}(U \leq \frac{f(y)}{Mg(y)}, Y = y)dy \\ &= \int_{-\infty}^{+\infty} \mathbb{P}(U \leq \frac{f(y)}{Mg(y)})\mathbb{P}(Y = y)dy \\ &= \int_{-\infty}^{+\infty} \int_0^{\frac{f(y)}{Mg(y)}} dug(y)dy = \int_{-\infty}^{+\infty} \frac{f(y)}{Mg(y)} g(y)dy \\ &= \frac{1}{M}\int_{-\infty}^{+\infty} f(y)dy = \frac{1}{M}, \end{aligned} \tag{4}$$

since $\int_{-\infty}^{+\infty} f(y)dy = 1$ (f is a density). In a similar way, since Y and U are independent

$$\begin{aligned}\mathbb{P}(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}) &= \int_{-\infty}^{x} \mathbb{P}(U \leq \frac{f(y)}{Mg(y)})g(y)dy \\ &= \frac{1}{M}\int_{-\infty}^{x} f(y)dy. \end{aligned} \tag{5}$$

By bringing together Equations (3), (4), (5), we have

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)dy,$$

which concludes the proof. ■

Remark: Similarly to what we discussed about the condition $\mu(\mathcal{X}) = p > 0$, it is important to choose a density g such that there exists a constant M has small as possible such that Equation (2) is verified.

1.2.3 The method of composition (optional, see exemple sheet)

The objective is to generate

$$f(x) = \int_\Theta f(x, \theta)p(\theta)d\theta,$$

where $x \in \mathcal{X} \subset R \to f(x, \theta)$ and $\theta \in \Theta \to p(\theta)$ are densities. We assume that we can sample according to these densities.

We already saw a sub-case of this problem in the discrete case, as an exercise in Subsubsection 1.2.1. The idea is the following:

1. Sample $T \sim p$.
2. Sample $X \sim f(., T)$.

Theorem 4 *The output X from the composition algorithm has density f.*

Proof For any $x \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{P}(X \leq x) &= \int_\Theta \mathbb{P}(X \leq x, T = \theta) d\theta \\ &= \int_\Theta \mathbb{P}(X \leq x | T = \theta) \mathbb{P}(T = \theta) d\theta \\ &= \int_\Theta \int_{-\infty}^{x} f(x, \theta) p(\theta) dx d\theta.\end{aligned}$$

This concludes the proof. ■

1.3 Generation of a Gaussian random variable

The simplest way for generating a Gaussian random variable of mean 0 and variance 1 ($\mathcal{N}(0, 1)$ is called the *Box-Muller* method. It is based on a polar transformation in dimension 2, i.e. from the following observation.

If one generates two independent random variables (X_1, X_2) that are each distributed according to $\mathcal{N}(0, 1)$, one can rewrite them in the polar system as (θ, R) where

1. θ is the angle between the vectors $(X_1, 0)$ and (X_1, X_2).
2. $R^2 = \|(X_1, 0) + (0, X_2)\|_2^2 = X_1^2 + X_2^2$.

Then we have by definition

$$X_1 = R\cos(\theta) \quad \text{and} \quad X_2 = R\sin(\theta).$$

A property of the $\mathcal{N}\big((0, 0), I_2\big)$ distribution is that θ and R^2 are independent (isotropic distribution). We can thus generate them separately.

Generation of θ By definition of the normal distribution, θ is uniform between 0 and 2π. We can then

1. Generate $U \sim \mathcal{U}([0, 1])$.
2. Set $\Theta = 2\pi U$.

Generation of R^2 By definition, $R^2 \sim \chi_2^2$. By definition of the χ_2^2 distribution, we can generate it as

1. Generate $V \sim \mathcal{U}([0, 1])$.
2. Set $R^2 = -2\log(V)$.

Exercise: Prove this assertion (see Subsection 1.2.1).

Figure 5: Re parametrization for simulation of a Gaussian random variable.

Conclusion We saw how to generate (θ, R^2). We can thus generate $(X_1, X_2) \sim \mathcal{N}\big((0,0), I_2\big)$ (one should draw U, V independent of each other).

Remark 1/Exercise 1: Another way of generating a normal random variable relies on the *ratio of uniforms* method. You can check it online.

Remark 2: This method is not very fast since it requires the computation of a square root and two trigonometric functions. In general, the method used is a combination of a Box Muller scheme, and a rejection algorithm, as

1. Generate (U, V) i.i.d. according to $\mathcal{U}([-1,1])$ and set $W = U^2 + V^2$.
2. If $W \geq 1$, go to step 1 (reject).
3. Otherwise, set

$$X_1 = U\sqrt{-2\frac{\log(W)}{W}} \quad \text{and} \quad X_2 = V\sqrt{-2\frac{\log(W)}{W}}.$$

Exercise 2: Prove this variant of the Box Muller method.

2. Monte-Carlo method and non-parametric inference

Monte-Carlo methods are mostly used to evaluate quantities of the form:

$$\theta(F) = \mathbb{E}_F \phi(X) = \int_{\mathcal{X}} \phi(x) f(x) dx,$$

where X is a random variable of distribution F and density f.

It is mostly useful for integrating functions whose integral does not have an analytic form. Another application is Monte-Carlo tests, that we will see later.

Examples:

- The mean $\theta(F) = \mathbb{E}_F X$ of a random variable.
- The probability $\theta(F) = \mathbb{E}_F \mathbf{1}\{X \in A\}$ that X belongs to A.
- The median $\theta(F) = F^{-1}(1/2)$.

2.1 The plug-in principle

Consider some unknown distribution F, and some functional $\theta(F)$. Consider $X_1, \ldots, X_n$ n i.i.d. data generated at random with distribution F.

A common estimator $\hat{F}_n$ of F, called the empirical distribution function, is defined by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}.$$

$\hat{F}_n$ is called a non-parametric estimator because the number of degrees of freedom grows with n (it can be equal to n).

Figure 6: Empirical distribution.

The plug-in principle recommends to consider, instead of $\theta(F)$, the plug-in functional $\theta(\hat{F}_n)$. This functional is often easier to compute than $\theta(F)$.

Examples:

- Integral: If

$$\theta(F) = \mathbb{E}_F \phi(X),$$

then

$$\theta(\hat{F}_n) = \mathbb{E}_{\hat{F}_n} \phi(X) = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i).$$

This estimate is the estimator of the integral of F by simulation, i.e. the Monte-Carlo integral estimator.

- Quantile: If

$$\theta(F) = F^{-1}(\alpha),$$

then

$$\theta(\hat{F}_n) = \hat{F}_n^{-1}(\alpha) = X_{(\lfloor n\alpha \rfloor)},$$

where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ is the ordered dataset.

We are now going to see several popular variants of the plug-in principle: Monte-Carlo integration, quantile approximation by Monte-Carlo methods and its application to Monte-Carlo tests, and finally the Bootstrap.

2.2 Monte-Carlo integration

We consider the setting of Monte-Carlo integration, i.e.

$$\theta(F) = \mathbb{E}_F \phi(X) = \int_{\mathcal{X}} \phi(x) f(x) dx,$$

and the plug-in estimator,

$$\theta(\hat{F}_n) = \mathbb{E}_{\hat{F}_n} \phi(X) = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i).$$

Lemma 5 *Assume that $\int_{\mathcal{X}} \phi(x)^2 f(x) dx < \infty$. Then $\theta(\hat{F}_n)$ is such that*

$$\mathbb{E}_{F^n}[\theta(\hat{F}_n)] = \theta(F), \text{ and } \mathbb{V}_{F^n}[\theta(\hat{F}_n)] = \frac{1}{n}\Big(\int_{\mathcal{X}} \big(\phi(x) - \theta(F)\big)^2 f(x) dx \Big).$$

Proof We have for any $i \leq n$

$$\mathbb{E}_F \phi(X_i) = \int_{\mathcal{X}} \phi(x) f(x) dx = \theta(F).$$

This implies unbiasedness. Also

$$\begin{aligned}
\mathbb{V}_F \phi(X_i) &= \mathbb{E}_F \big(\phi(X_i) - \mathbb{E}_F \phi(X_i)\big)^2 \\
&= \int_{\mathcal{X}} \big(\phi(x) - \mathbb{E}_F \phi(X_i)\big)^2 f(x) dx \\
&= \int_{\mathcal{X}} \phi(x)^2 f(x) dx - \theta(F)^2,
\end{aligned}$$

since f is a density and thus its integral equals 1. This concludes the proof for the variance since the samples are i.i.d.. ■

Remark: In order for this estimate to have finite variance, it is necessary and sufficient ti impose that ϕ is of integrable square according to g.

Lemma 6 *Assume that $\int_{\mathcal{X}} \phi(x)^2 f(x) dx < \infty$. Then the two following statements hold.*

1. *$\theta(\hat{F}_n)$ converges almost surely to $\theta(F)$.*
2. *$\sqrt{n}(\theta(\hat{F}_n) - \theta(F))$ converges in distribution to a $\mathcal{N}\Big(0, \int_{\mathcal{X}} \big(\phi(x) - \theta(F)\big)^2 f(x) dx\Big)$.*

Proof The first part of the lemma is a direct application of the strong law of large numbers (since $\mathbb{E}_F |\phi(X)| \leq \sqrt{\mathbb{E}_F \phi(X)^2}$).

The second part is a direct application of the central limit theorem to

$$\theta(\hat{F}_n) = \mathbb{E}_{\hat{F}_n} \phi(X) = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i),$$

which is a sum of the i.i.d. random variables $\phi(X_i)$ of finite mean $\theta(F)$ and variance $\int_{\mathcal{X}} \big(\phi(x) - \theta(F)\big)^2 f(x) dx = \int_{\mathcal{X}} \phi(x)^2 f(x) dx - \theta(F)^2$. ■

The deviations of this estimate can be large, in particular if the function ϕ is highly variable. We are now going to discuss how to modify this plug-in estimate in order to make its variance smaller, by using side information that we have at our disposal. In the rest of this Subsection, we discuss variants of the Monte-Carlo integration method that have smaller variance.

2.2.1 Importance sampling

Let g be a density (and G the associated distribution) that is strictly positive whenever $f\phi$ is non zero (this condition is actually equivalent to ϕ dominating $f\mathbf{1}\{\phi \neq 0\}$).

Let $X_1, \dots, X_n$, be i.i.d. samples sampled according to g. We define the following estimate of $\theta(F)$ as

$$\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} w_i \phi(X_i), \quad \text{where} \quad w_i = \frac{f(X_i)}{g(X_i)}.$$

The $(w_i)_i$ are called importance weights.

Lemma 7 *$\hat{\theta}_g$ is such that*

$$\mathbb{E}_{G^n}[\hat{\theta}_g] = \theta(F), \text{ and } \mathbb{V}_{G^n}[\hat{\theta}_g] = \frac{1}{n}\Big(\int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx - \theta(F)^2\Big).$$

Proof We have for any $i \leq n$

$$\mathbb{E}_G[\phi(X_i)\frac{f(X_i)}{g(X_i)}] = \int_{\mathcal{X}} \phi(x)\frac{f(x)}{g(x)} g(x) dx = \theta(F).$$

This implies unbiasedness. Also

$$\begin{aligned}
\mathbb{V}_G[\phi(X_i)\frac{f(X_i)}{g(X_i)}] &= \mathbb{E}_G\big(\phi(X_i)\frac{f(X_i)}{g(X_i)} - \mathbb{E}_G[\phi(X_i)\frac{f(X_i)}{g(X_i)}]\big)^2 \\
&= \int_{\mathcal{X}} \big(\phi(x)\frac{f(x)}{g(x)} - \theta(F)\big)^2 g(x) dx \\
&= \int_{\mathcal{X}} \phi(x)^2 \frac{f(x)^2}{g(x)} dx - 2\int_{\mathcal{X}} \phi(x) f(x) dx \theta(F) + \theta(F)^2 \\
&= \int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx - \theta(F)^2.
\end{aligned}$$

This concludes the proof for the variance since the samples are i.i.d.. ■

As seen in Lemma 7, the variance depends crucially on the function $\psi(x) = \frac{\phi^2(x) f^2(x)}{g(x)}$, and more specifically, on its integral: the smaller this integral, the smaller the variance.

Intuition of importance sampling: The objective of importance sampling is to sample more where the integral is higher, i.e. where $|\phi| f$ is large. Indeed, it is in those points that the largest contribution to the integral is. It is formally proved in next theorem that it is a such density that is optimal.

Theorem 8 *The density g^* that minimises the variance of $\hat{\theta}_g$ is*

$$g^* = \frac{f|\phi|}{\int_{\mathcal{X}} f|\phi|}$$

Proof By Lemma 7 (and by just developing the square in the variance), we know that $\mathbb{V}_{G^n}$ is minimised according to g if and only if

$$\int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx,$$

is minimised. We have for any density g

$$\begin{aligned}\int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx &= \int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)^2} g(x) dx \\ &= \mathbb{E}_G\Big[\big(\frac{\phi(X) f(X)}{g(X)}\big)^2\Big] \\ &\geq \Big(\mathbb{E}_G\Big[\frac{|\phi(X)| f(X)}{g(X)}\Big]\Big)^2,\end{aligned}$$

by Jensen inequality. This implies

$$\begin{aligned}\int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx &= \int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)^2} g(x) dx \\ &\geq \Big(\mathbb{E}_G\Big[\frac{|\phi(X)| f(X)}{g(X)}\Big]\Big)^2 \\ &= \Big(\int_{\mathcal{X}} |\phi(x)| f(x) dx\Big)^2,\end{aligned}$$

which implies that for any density g

$$\mathbb{V}_{G^n}[\hat{\theta}_g] \geq \frac{1}{n}\Big(\Big(\int_{\mathcal{X}} |\phi(x)| f(x) dx\Big)^2 - \theta(F)^2\Big),$$

Also, by plugging g^* in the variance in Lemma 7, we get

$$\mathbb{V}_{(G^*)^n}\hat{\theta}_{g^*} = \frac{1}{n}\Big(\Big(\int_{\mathcal{X}} |\phi(x)| f(x) dx\Big)^2 - \theta(F)^2\Big).$$

This implies that g^* minimises the variance, and this concludes the proof. ■

Remark 1: In theory, it is clear - we should sample from g^*. But in practice, we are probably not able to sample from g^*: indeed the problem of sampling directly from g^* seems harder than the problem of integrating ϕ.

Exercise: To insists on this point, apply the importance sampling technique to

$$f = \mathbf{1}\{[0,1]\}, \quad \text{and} \quad \phi = \mathbf{1}\{[0,u]\},$$

where $0 \leq u \leq 1$.

Remark 2: Serious difficulties arise in importance sampling if the proposal distribution g gets small much faster than f out in the tails of the distribution.

2.2.2 Control variates (optional)

Let $\hat{\theta}$ be an unbiased estimate of θ.

Idea: Use random variables that are correlated to $\hat{\theta}$ but better "known" than θ.

A random variable C is called *control variate* if it is correlated with $\hat{\theta}$, and its mean μ_C is known. We the define the following modified estimate

$$\hat{\theta}_\beta = \hat{\theta} - \beta(C - \mu_C),$$

which depends of the real parameter β.

Lemma 9 *We have*

$$\mathbb{E}\hat{\theta}_\beta = \theta, \quad \text{and} \quad \mathbb{V}\hat{\theta}_\beta = \mathbb{V}\hat{\theta} - 2\beta\text{Cov}(C, \hat{\theta}) + \beta^2\mathbb{V}C.$$

This is minimised when β is equal to the correlation coefficient between Y and C, i.e.

$$\rho = \frac{\text{Cov}(C, \hat{\theta})}{\mathbb{V}C},$$

and

$$\mathbb{V}\hat{\theta}_\rho = (1 - \rho^2)\mathbb{V}\hat{\theta}.$$

Remark: With this choice of $\beta = \rho$, we will have that $\mathbb{V}\hat{\theta}_\rho < \mathbb{V}\hat{\theta}$ as long as $\text{Cov}(\hat{\theta}, C) \neq 0$.

Remark 2: These ideas can obviously be extended to more than one control variate.

Figure 7: Control variates.

2.2.3 Stratified sampling

Consider the general integration problem where one wants to compute

$$\mu = \int_{\mathcal{X}} \phi(x) f(x) dx.$$

Problem of random sampling: The points are chosen at random according to f - this creates a source of randomness and thus of error. The objective of stratified sampling is to reduce the error but reducing the amount of randomness in the picking of the points.

Stratified sampling:

- Divide the domain in K *strata* Ω_i that are measurable according to f, that form a partition of the domain, and that are such that we know exactly $w_i = \mathbb{P}_f(\Omega_i)$.
- Sample *exactly* T_i points in each stratum Ω_i, according to f restricted to Ω_i. The numbers T_i are *deterministic* and such that $\sum_i T_i = n$. Write the conditional empirical mean in stratum Ω_i as

$$\hat{\mu}_i = \frac{1}{T_i} \sum_{j=1}^{n} X_j \mathbf{1}\{X_j \in \Omega_i\}.$$

- Return the weighted estimate of the integral

$$\hat{\mu} = \sum_{i=1}^{K} w_i \hat{\mu}_i.$$

Figure 8: Stratified sampling.

Theorem 10 *We have*

$$\mathbf{E}\hat{\mu} = \mu, \quad \text{and} \quad \mathbf{V}\hat{\mu} = \sum_{i=1}^{K} \frac{w_i^2 \sigma_i^2}{T_i},$$

where $\mu_i = \frac{1}{w_i} \int_{\Omega_i} \phi(x) f(x) dx$ *is the conditional mean, and* $\sigma_i^2 = \frac{1}{w_i} \int_{\Omega_i} (\phi(x) - \mu_i)^2 f(x) dx$ *is the conditional variance in stratum* Ω_i*.*

Proof The proof employs similar tools and ideas as in Lemma 5 - we will detail this in class. ■

The entire problems consists now in the choice of the T_i.

Uniform stratified sampling A first, intuitive idea, consists in choosing

$$T_i^u = w_i n,$$

where $w_i = \mathbb{P}_f(\Omega_i)$. Doing this "consolidates" the random sampling while preserving the shape of the density f.

Figure 9: Stratified sampling proportional to the measures of the strata.

Corollary 11 *With this choice of T_i, we have*

$$\mathbf{E}\hat{\mu}_u = \mu, \quad \text{and} \quad \mathbf{V}\hat{\mu}_u = \sum_{i=1}^{K} \frac{w_i \sigma_i^2}{n}.$$

Remark: By Lemma 5, we can rewrite the variance of the "classic" Monte-Carlo estimate of the integral as

$$\mathbf{V}\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{K} w_i \sigma_i^2 + \frac{1}{n}\sum_{i=1}^{K} w_i (\mu_i - \mu)^2,$$

by a simple conditional variance decomposition. This implies that whenever the function is not constant over the strata,

$$\mathbf{V}\hat{\mu}_u < \mathbf{V}\hat{\mu}_{MC}.$$

Remark: There are no restriction whatsoever for applying this procedure on any domain $\mathcal{X}$ and with any density f - provided that one is able to compute analytically the w_i in an *exact* (or very, very precise) way.

Example of stratification: Assume $f = \mathbf{1}\{[0,1]\}$. For instance, one can stratify as $\Omega_i = [\frac{i}{K}, \frac{i+1}{K}]$. Then $w_i = \frac{1}{K}$.

Oracle stratified sampling A natural question is to find the allocation T_i that minimise the variance given a stratification. Solving the minimisation program

$$\min_{(T_i)_i} \mathbf{V}\hat{\mu} = \sum_{i=1}^{K} \frac{w_i^2 \sigma_i^2}{T_i} \quad s.t. \quad T_i \geq 0, \sum_i T_i = n, \tag{6}$$

implies the following corollary.

Corollary 12 *Choosing*

$$T_i^* = \frac{w_i \sigma_i}{\sum_j w_j \sigma_j} n$$

is the unique solution of the minimisation problem (6)*, and implies that for the resulting estimate $\hat{\mu}^*$*

$$\mathbf{E}\hat{\mu}^* = \mu, \quad \text{and} \quad \mathbf{V}\hat{\mu}_u^* = \frac{1}{n}\Big(\sum_{i=1}^{K} w_i \sigma_i\Big)^2.$$

Proof This is obtained by solving the Lagrangian associated to the minimisation problem (6), or by Jensen's inequality. ■

The idea of such an allocation is to allocate more points where ϕ varies more, and thus where there is more incertitude.

Remark: Note however that σ_i are in general not available. For this reason, the oracle allocation is not feasible. However, one could imagine spending n^ϵ of the budget in an uniform stratified sampling way (with $\epsilon < 1$) on estimating the σ_i, and then the rest of the samples according to these empirical estimate of the oracle proportions.

Figure 10: Oracle allocation.

Stratified sampling on smooth functions Consider an integration problem on $[0,1]$, where one wants to integrate a function ϕ over $[0,1]$. In this case, we have

$$f(x) = \mathbf{1}\{[0,1]\},$$

and f is the uniform measure on $[0,1]$, and we want to compute

$$\int_{[0,1]} \phi(x)dx = \int_{\mathbb{R}} \phi(x)f(x)dx.$$

Assume that ϕ is Lipschitz, i.e. there exists $C < 0$ such that for any u, v, $|\phi(u) - \phi(v)| \leq C|u - v|$.

Consider the stratification in n strata as $\Omega_i = [\frac{i}{n}, \frac{i+1}{K}]$. Then $w_i = \frac{1}{K}$.

Consider uniform stratified sampling on this stratification. Then we know by Corollary 11 that

$$\begin{aligned} V\hat{\mu}_u &= \sum_{i=1}^{n} \frac{w_i \sigma_i^2}{n} \\ &= \sum_{i=1}^{n} \frac{1}{n} \int_{i/n}^{(i+1)/n} (\phi(x) - \mu_i)^2 dx, \end{aligned}$$

and by the Lipschitz assumption, and since μ_i is the empirical mean restricted to stratum Ω_i, for any $x \in \Omega_i$, $|\phi(x) - \mu_i| \leq C/n$, and thus

$$\begin{aligned} V\hat{\mu}_u &\leq \sum_{i=1}^{n} \frac{1}{n} \int_{i/n}^{(i+1)/n} \frac{C^2}{n^2} dx \\ &= \sum_{i=1}^{n} \frac{C^2}{n^4} dx \\ &\leq \frac{C^2}{n^3}. \end{aligned}$$

Remark 1: The variance of Monte-Carlo estimate is of order $1/n$. Here the variance is much smaller, i.e. $1/n^3$! This comes from the fact that, on very small strata, Lipschitz functions are almost constant.

Figure 11: Stratification on Lipschitz functions.

Remark 2: This kind of ideas is linked to *quasi Monte-Carlo* methods - i.e.instead of sampling at random, create grids that are "more uniform" than uniform samples.

Remark 3: This can be generalized to other kind of smoothness classes - such as Hölder - and to higher dimension. But in high dimension, *curse of dimensionality* for such methods that rely on coverage.

2.3 Quantile approximation by Monte-Carlo methods and application to Monte-Carlo tests

We will first see how to approximate a quantile by Monte-Carlo approximation, and then how this can be applies to testing.

2.3.1 Quantile approximation.

Suppose that we want to approximate the c_α quantile associated to the tail probability α, of a distribution F.

A Monte Carlo approximate of this quantile is:

1. Choose $B \in \mathbb{N}$, as large as possible (as computational constraints allow)
2. Restrict the choice of α to that of $k \in \{1, ..., B\}$, implicitly defining $\alpha = \frac{k}{B+1}$.
3. Simulate random samples $T_1, \ldots, T_B$ according to F.
4. Let $\hat{c}_\alpha = T_{(k)}$, where $T_{(k)}$ is the $k-$th order statistic of $T_1, \ldots, T_B$.

Theorem 13 *Assume that F corresponds to a density f, then*

$$\mathbb{E}(F(\hat{c}_\alpha)) = \alpha.$$

Proof We have (here $\mathbb{P}_{F \times F^B}$ denotes the probability on the bootstrap samples, and also on the data samples, under F)

$$\begin{aligned}
\mathbb{E}(F(\hat{c}_\alpha)) &= \mathbb{P}_{Y, T_1 \ldots T_B \sim F \times F^B}\Big(Y \leq T_{(k)}\Big) \\
&= \int_{-\infty}^{\infty} \mathbb{P}_{F^n}\Big(t \leq T_{(k)} | Y = t\Big) f(t) dt \\
&= \int_{-\infty}^{\infty} \sum_{r=0}^{k-1} \mathbf{C}_B^r F(t)^r (1 - F(t))^{B-r} f(t) dt \\
&= \int_0^1 \sum_{r=0}^{k-1} \mathbf{C}_B^r u^r (1-u)^{B-r} du,
\end{aligned}$$

by a change of variable $u = F(t)$. Since $r\mathbf{C}_B^r u^{r-1}(1-u)^{B-r}$ is actually a $\text{Beta}(r, B-r+1)$ distribution with mean $\frac{r}{B+1}$, we have

$$\mathbb{P}_{F \times F^B}\Big(Y \leq T_{(k)}\Big) = \sum_{r=0}^{k-1} \frac{1}{r} \frac{r}{B+1} = \frac{k}{B+1} = \alpha,$$

and this concludes the proof. ■

2.3.2 Monte-Carlo test.

Let $X_1, ..., X_n$ be independent with distribution function F, and suppose we want to test

$$H_0 : F = F_0, \ \ against \ \ H_1 : F \neq F_0.$$

using a statistic $T = T(X_1, ..., X_n)$. If small values of T indicate departure from H_0 , we would like, for a test of size $\alpha \in [0, 1]$, to reject H_0 if $T < c_\alpha$, where c_α is the $\alpha-$th quantile of T.

However, if the null distribution of T is unknown, we might not be able to compute this quantile. The idea of a Monte-Carlo test is to approximate the quantile c_α by Monte-Carlo. It can be done in a similar way as in the previous part, i.e.

1. Choose $B \in \mathbb{N}$, as large as possible (as computational constraints allow)
2. Restrict the choice of α to that of $k \in \{1, ..., B\}$, implicitly defining $\alpha = \frac{k}{B+1}$.
3. Simulate random samples $X_{b,1}, \ldots, X_{b,n}$ for any $b \in \{1, \ldots, B\}$ according to F_0.
4. Let $\hat{c}_\alpha = T_{(k)}$, where $T_{(k)}$ is the $k-$th order statistic of $T_1, \ldots, T_B$.
5. Reject H_0 if $T < c_\alpha$.

Remark: As proved in last Theorem, such a Monte-Carlo test has indeed size exactly α. However, since the quantile c_α is an estimate of the true quantile, and not the true quantile itself, one will have a loss of power with respect to what would happen if one really considered the true quantile.

2.4 The Bootstrap

Let $X_1, \ldots, X_n$ be independent random variables with distribution F. It can happen that we are interested in the properties of a function of all the samples (as in the case of Monte-Carlo tests). In this case, since we only have n samples and can thus in theory compute an estimate of the samples only once, there is not much we can do.

To answer this problem, *re-sampling* methods were introduced. They consist in resampling from the sample in order to obtain many samples. The most popular of these methods is the Bootstrap.

The Bootstrap estimator. Suppose we are interested in the distribution $K_n(F)$ of a *root* or *pivot* $R_n(X, F)$ where $X = (X_1, \ldots, X_n)$ (think for example of the distribution of the statistic $T(X_1, \ldots, X_n)$ in Monte-Carlo tests).

The bootstrap estimator of $K_n(F)$ is $K_n(\hat{F})$ (still by the plug-in principle). In other words, we estimate the distribution of $R_n(X, F)$ by the distribution of $R_n(X^*, \hat{F})$ where $X^* = (X_1^*, \ldots, X_n^*)$ are sampled in a i.i.d. way from $\hat{F}$.

The approximation of the Bootstrap estimator by Monte-Carlo. It is possible to compute this bootstrap estimate of $K_n(F)$ given a sample X - but it can be computationally challenging. In order to cope with that, it is common to approximate this estimator by Monte-Carlo, as follows.

1. Draw B independent bootstrap samples $X_b^* = (X_{b,1}^*, \dots, X_{b,n}^*)$ from $\hat{F}^n$.

2. Approximate $K_n(\hat{F})$ by the empirical distribution function of $\big(R_n(X_b^*, \hat{F})\big)_b$.

Example: Bootstrap for confidence intervals.

Figure 12: Bootstrap confidence interval.

Remark: A popular specific modification of the Bootstrap is the *Jacknife*. It consists in considering specific data-sets, so called *leave one out* datasets (which consist in the whole dataset minus one point). See exemple sheet for more informations on the Jacknife.

Very important remark: It is very important in all the Monte-Carlo methods to choose the number of Monte-Carlo samples wisely. A large number of Monte-Carlo samples is better approximation-wise, but is more costly in terms of computational time. In practice, one should adapt the number of samples to the number of parameters in the object one wants to measure - maybe around some hundreds for e.g. mean, variance, quantiles, and rather some thousands for the entire distribution.

3. Bayesian inference and associated methods

3.1 The concept of Bayesian inference

Let $\bar{X} = X_1, \dots, X_n$ be n i.i.d. samples generated by a likelihood model $L(\bar{X}, \theta)$. In classical inference, a common technique is to maximize the likelihood. Bayesians have a

different perspective about it, i.e. they put a prior $p(\theta)$ on θ which summarizes their a priori knowledge about θ. They then compute the a posteriori distribution of θ, i.e.

$$\pi(\theta|\bar{X}) = L(X,\theta)p(\theta).$$

The maximization of this posterior is an alternative to the maximization of the likelihood, and the resulting estimator is called the maximum a posteriori (MAP). Moreover, from a Bayesian perspective, this distribution π is of interest in itself, and Bayesians might want to estimate some characteristics of it such as expectation, variance, etc. An important remark about the expectation of the posterior is that, when n converges to infinity, it converges to θ. This quantity is often simpler to compute than the MAP.

For complicated likelihood, there might however be no closed forms for these quantities. By the plug-in principle, if we could generate samples of θ according to π, we could have good (e.g. Monte-Carlo) estimates of the quantities that interest us.

Objective: Generate samples according to π.

Problem: π can be horrible, in particular it can be *multi-dimensional*, and it is not always easy to use standard techniques to generate according to π.

Solution: Generate a Markov chain $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots)$ that is such that π is it's stationary distribution.

We saw many methods to simulate one dimensional random variables. There are however many problems where one wishes to simulate a multi-dimensional complex distribution. The topic of this chapter is to provide solutions for doing this.

3.2 Gibbs sampler

Simple Example Let $\bar{X} = X_1, \ldots, X_n$ be n data generated by the likelihood model $L(\bar{X}, \theta)$, where $\theta = (\mu, \sigma^2)$ is the parameter and L is the Gaussian density of parameter θ.

Let us put the following priors on θ:

$$p(\mu) = \frac{1}{\sqrt{2\pi}} \exp(-\mu^2/2),$$

$$p(\sigma^2) = \exp(-1/\sigma^2).$$

The posterior $\pi(\theta|\bar{X})$ is proportional to

$$\pi(\theta|\bar{X}) = \pi(\mu,\sigma^2) \simeq \frac{1}{\sqrt{2\pi}} \exp(-\mu^2/2) \exp(-1/\sigma^2) \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(X_i-\mu)^2/(2\sigma^2)).$$

The conditional posterior of μ is

$$\pi(\mu|\sigma^2, \bar{X}) = \frac{\frac{1}{\sqrt{2\pi}} \exp(-\mu^2/2) \exp(-1/\sigma^2) \prod_{i=1}^{n} \exp(-(X_i-\mu)^2/(2\sigma^2))}{\int_m \frac{1}{\sqrt{2\pi}} \exp(-m^2/2) \exp(-1/\sigma^2) \prod_{i=1}^{n} \exp(-(X_i-m)^2/(2\sigma^2)) dm},$$

i.e. the conditional posterior of μ is distributed according to a

$$\mu|\bar{X}, \sigma^2 \sim \mathcal{N}(\sum_i X_i/(1+\sigma^2), \frac{\sigma^2}{1+\sigma^2}).$$

The conditional posterior of σ^2 is

$$\pi(\sigma^2|\bar{X},\mu) = \frac{\frac{1}{\sqrt{2\pi}}\exp(-\mu^2/2)\exp(-1/\sigma^2)\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}}\exp(-(X_i-\mu)^2/(2\sigma^2))}{\int_{s^2}\frac{1}{\sqrt{2\pi}}\exp(-\mu^2/2)\exp(-1/s^2)\prod_{i=1}^n \frac{1}{s\sqrt{2\pi}}\exp(-(X_i-\mu)^2/(2s^2))ds^2},$$

i.e. the conditional posterior of σ^2 is distributed according to a

$$\sigma^2|\bar{X},\mu \sim IG(n/2+1, 1+\sum_i (X_i-\mu)^2/2),$$

where $IG(\alpha,\beta)$ has density $f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{-\alpha+1}\exp(-\beta/y)$.

Gibbs sampler works as follows.

1. Set initial values μ_0 and σ_0^2.
2. Given σ_0^2, simulate μ_1 according to $\pi(.|\sigma_0^2,\bar{X})$.
3. Given μ_1, simulate σ_1^2 according to $\pi(.|\mu_1^2,\bar{X})$.
4. Iterate and at time $t+1$...
5. Given σ_t^2, simulate μ_{t+1} according to $\pi(.|\sigma_t^2,\bar{X})$.
6. Given μ_{t+1}, simulate σ_{t+1}^2 according to $\pi(.|\mu_{t+1}^2,\bar{X})$.
7. At the end of the process...
8. Throw away all the beginning of the chain and consider only the last samples.

Since this algorithm produces a chain whose stationary distribution equals the posterior distribution, the samples will, on the long run, have the right property...

General method Let $\bar{\theta} = (\theta_1,\ldots,\theta_p)$ be the parameter of interest and $\pi(\theta_i|\bar{\theta}_{(-i)}) = \pi_i(\bar{\theta}_{(-i)})$ be the conditional posterior distributions. The Gibbs sampler works as follows

1. Set initial vector $\theta^{(0)}$.
2. Then at time $t+1$...
3. Set $\theta_1^{(t+1)} \sim \pi_1(\theta_2^{(t)},\theta_3^{(t)},\ldots,\theta_p^{(t)}) = \pi_1(\bar{\theta}_{(-1)}^{(t)})$.
4. Set $\theta_2^{(t+1)} \sim \pi_2(\theta_1^{(t+1)},\theta_3^{(t)},\ldots,\theta_p^{(t)})$.
5. ...
6. Set $\theta_p^{(t+1)} \sim \pi_p(\bar{\theta}_{(-p)}^{(t+1)})$.
7. Collect T samples like that.
8. At the end of the process...

Figure 13: Gibbs sampler.

9. Throw away all the b first samples and consider only the last samples (and also, in general, do some sub-sampling to diminish the coreelations)..

This method generates a Markov chain whose stationary distribution is actually the posterior under not too strong assumptions... The proof of this fact is beyond the scope of this course

Remark: The samples $\bar{\theta}^{(t)}$ are *not* independent.

3.3 Metropolis Hastings algorithm

The Metropolis Hastings algorithm is a sequential form of rejection sampling. It is useful with respect to rejection sampling if you do not have a good upper enveloppe g for π (see Subsubsection 1.2.2). This is an MCMC method so you construct a Markov chain whose stationnary distribution is π.

The idea is the following. Assume that you already generated a chain $(\theta^{(1)}, \ldots, \theta^{(t)})$. When you are in a state $\theta^{(t)}$ of the Markov chain, the algorithm proposes you a new value X for your chain. You compare the probability of this value (according to π) with the probability of $\theta^{(t)}$, as in rejection sampling, and you randomly decide if you want $\theta^{(t+1)}$ to be x or not. And you iterate.

First intuition In order to understand this more clearly, we are going to consider a very simple mechanism for generating the proposition X.

Consider a distribution π defined on the set of integers $\{1, \ldots, M\}$ and the uniform measure μ on $\{1, \ldots, M\}$. Consider the following procedure.

1. Set initial vector $\theta^{(0)}$.
2. Then at time $t+1$...
3. Simulate $X \sim \mu$ and $U \sim \mathcal{U}_{[0,1]}$.
4. If $U \leq \frac{\pi(X)}{\pi(\theta^{(t)})}$, then $\theta^{(t+1)} = X$, otherwise $\theta^{(t+1)} = \theta^{(t)}$ item Collect T samples like that.
5. At the end of the process...
6. Throw away all the b first samples and consider only the last samples (and also, in general, do some sub-sampling to diminish the correlations).

Figure 14: Simplified version of Metropolis Hastings algorithm in a simple case.

Theorem 14 *Assume that for all $i \in \{1, \ldots, M\}$, we have $\pi(i) > 0$. Then the unique stationnary distribution of the markov chain $\theta^{(1)}, \ldots, \theta^{(t)}, \ldots$, is π.*

Proof First, let us write the transition Kernel K of this Markov Chain. I remind you that the transition kernel of a Markov Chain is simply the probability, given that you are in a state a, to go in a state b.

For our Markov Chain, we have for any $(a, b) \in \{1, \dots, M\}^2$

$$K(b, a) = \min(\frac{\pi(b)}{\pi(a)}, 1) + \Big(1 - \sum_{b'=1}^{M} \min(\frac{\pi(b')}{\pi(a)}, 1)\Big)\mathbf{1}\{a = b\}.$$

Since for all $i \in \{1, \dots, M\}$, we have $\pi(i) > 0$, it means that K is well defined, and that for any a, b, we have $K(a, b) > 0$. This implies in particular (since the number of state M is finite) that K admits one and only one stationnary measure, which is the single invariant point of K. Since for any $b \in \{1, \dots, M\}$

$$\begin{aligned}
(K\pi)(b) &= \sum_{a=1}^{M} K(b, a)\pi(a) \\
&= \sum_{a=1}^{M} \left(\min(\frac{\pi(b)}{\pi(a)}, 1) + \Big(1 - \sum_{b'=1}^{M} \min(\frac{\pi(b')}{\pi(a)}, 1)\Big)\mathbf{1}\{a = b\} \right)\pi(a) \\
&= \pi(b),
\end{aligned}$$

we know that $K\pi = \pi$, so π is the invariant point of K, and thus the stationnary measure. ■

Remark 1: This theorem is equivalent to the fact that for any measure ν on $\{1, \dots, M\}$

$$\lim_{n \to \infty} \sup_{b} |(K^n \nu)(b) - \pi(b)| = 0.$$

An interesting question as a practitionner is on the speed of convergence of this expression to 0. It is important both for the initialization, and also for choosing how many samples to discard.

Remark 2: Instead of the uniform measure μ, you could consider an arbitrary measure. What would then happen?

Extension in the continuous Consider now a distribution π on $[0, 1]$ that admits a density according to the uniform measure μ on $[0, 1]$. Consider the same procedure as above.

Remark 1: You should prefer this method over rejection sampling only if you do not have a good enveloppe for π.

Remark 2: This method produces, again, correlated samples. This is particularly important to remember this for continuous distributions...

Figure 15: Continuous case.

General method In the general case, the mechanism to choose the proposition X can depend on the current state $\theta^{(t)}$ of the chian.

Consider a distribution π defined on a domain $\mathcal{X}$. Assume that π is such taht for any atom $x \in \mathcal{X}$, $\{x\}$ is measurable according to π. For any $x \in \mathcal{X}$, define a transition measure $\mu(.|x)$ on $\mathcal{X}$.

Consider the following procedure.

1. Set initial vector $\theta^{(0)}$.
2. Then at time $t+1$...
3. Simulate $X \sim \mu(.|\theta^{(t)})$ and $U \sim \mathcal{U}_{[0,1]}$.
4. If $U \leq \frac{\pi(X)\mu(\theta^{(t)}|X)}{\pi(\theta^{(t)})\mu(X|\theta^{(t)})}$, then $\theta^{(t+1)} = X$, otherwise $\theta^{(t+1)} = \theta^{(t)}$ item Collect T samples like that.
5. At the end of the process...
6. Throw away all the b first samples and consider only the last samples (and also, in general, do some sub-sampling to diminish the correlations).

Figure 16: The Metropolis Hastings algorithm.

Remark 1: It is crucial to choose well the initial state of the chain $\theta^{(0)}$ in this case, in particular for unbounded distributions.

Remark 2: The transition probability μ should also be well chosen, in particular if one wants to have fast convergence of the chain to the stationnary measure.

Remark 2: The number of samples one wants to discard at the beginning depends very much on the problem. The same goes for the amount of correlation in the chain.

Remark 4: In many cases, it is an open research problem to solve these questions.

3.4 Effective sample size

As mentioned in this section, MCMC method produces a correlated chain. For this reason, t samples do not really provide the same information as t i.i.d. variables produced by π, but less. A crucial and very interesting question is on how much information is contained in a chain of length t, i.e. to what length T of i.i.d. samples distributed according to π is equivalent a MCMC chain of stationnary distribution π and length t (clearly, $T \leq t$).

A common notion for measuring this is *effective sample size*. Let, for any integer $l \geq 0$, us define $\gamma(l)$ as the correlation between two samples of lag l. Define $\rho(l) = \gamma(l)/\gamma(1)$. Then the effective sample size $\tilde{T}$ of a chain of length t is

$$\tilde{T} = \frac{t}{1 + 2\sum_{l=1}^{T-1} \rho(l)}.$$

This criter is often computed, using the empirical estimates of the autocorellations.