Monte Carlo Methods for Absolute Beginners

Christophe Andrieu

School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK
c.andrieu@bristol.ac.uk
http://www.stats.bris.ac.uk/~maxca

1 Motivation and Basic Principles of the Monte Carlo Method

The modern history of Monte Carlo techniques dates back from the 1940's and the Manhattan project. There are earlier descriptions of Monte Carlo experiments, Buffon's famous needle experiment is one them, but examples have been traced back to Babylonian and old testament times [13]. As we shall see these techniques are particularly useful in scenarios where it is of interest to perform calculations that involve - explicitly or implicitly - a probability distribution π on a space X (typically $\mathsf{X} \subset \mathbb{R}^{n_x}$ for some integer n_x), for which closed-form calculations cannot be carried out due to the algebraic complexity of the problem. As we shall see the main principle of Monte Carlo techniques consists of replacing the algebraic representation of π, e.g. $1/\sqrt{2\pi}\exp(\frac{-1}{2}x^2)$ with a *sample* or *population* representation of π, e.g. a set of samples $X_1, X_2, \ldots, X_N \overset{iid}{\sim} \pi(x) = 1/\sqrt{2\pi}\exp(\frac{-1}{2}x^2)$. This proves in practice to be extremely powerful as difficult - if not impossible - *exact* algebraic calculations are typically replaced with simple calculations in the sample domain. One should however bear in mind that these are *random approximations* of the true quantity of interest. An important scenario where Monte Carlo methods can be of great help is when one is interested in evaluating expectations of functions, say f, of the type $\mathbb{E}_\pi(f(X))$ where π is the probability distributions that defines the expectation. The nature of the approach, where algebraic quantities are approximated by random quantities, requires one to quantify the random fluctuations around the true desired value. As we shall see, the power of Monte Carlo techniques lies in the fact that the *rate* at which the approximation converges towards the true value of interest is immune to the dimension n_x of the space X where π is defined. This is the second interest of Monte Carlo techniques.

These numerical techniques have been widely used in physics over the last 50 years, but their interest in the context of Bayesian statistics and more generally statistics was only fully realized in the late eighties early nineties. Although we will here mostly focus on their application in statistics, one should bear in mind that the material presented in this introduction to the topic has applications far beyond statistics.

The prerequisites for this introduction are a basic first year undergraduate background in probability and statistics. Keywords include random variable,

O. Bousquet et al. (Eds.): Machine Learning 2003, LNAI 3176, pp. 113–145, 2004.

law of large numbers, estimators, central limit theorem and basic notions about Markov chains.

1.1 Motivating Example

In this section we motivate and illustrate the use of Monte Carlo methods with a toy example. We then point out the power of the approach on a "real" example.

Calculating π with the Help of Rain and the Law of Large Numbers

A Physical Experiment Consider the 2×2 square, say $\mathcal{S} \subset \mathbb{R}^2$, with inscribed disc $\mathcal{D}$ of radius 1 as in Figure 1.

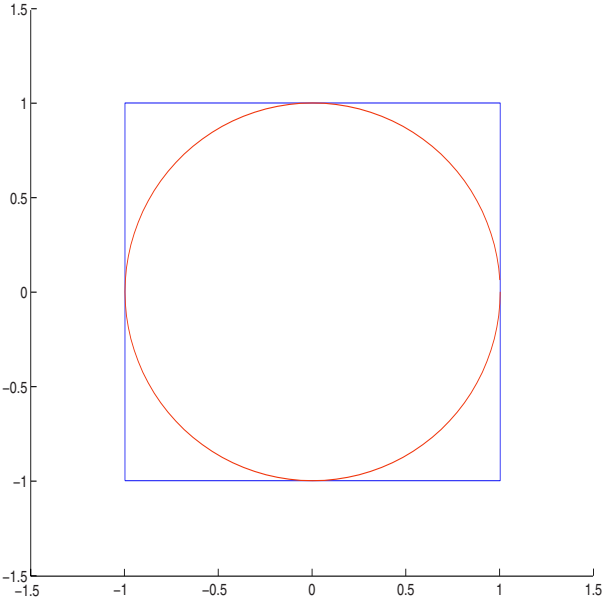

Fig. 1. A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1

Imagine that an "idealized" rain falls uniformly on the square $\mathcal{S}$, i.e. the probability for a drop to fall in a region $\mathcal{A}$ is proportional to the area of $\mathcal{A}$. More precisely, let D be the random variable defined on $\mathsf{X} = \mathcal{S}$ representing the location of a drop and $\mathcal{A}$ a region of the square, then

$$\mathbb{P}(D \in \mathcal{A}) = \frac{\int_{\mathcal{A}} dxdy}{\int_{\mathcal{S}} dxdy}. \tag{1}$$

where x and y are the Cartesian coordinates. Now assume that we have observed N such *independent* drops, say $\{D_i, i = 1, \ldots, N\}$ as in Figure 2.

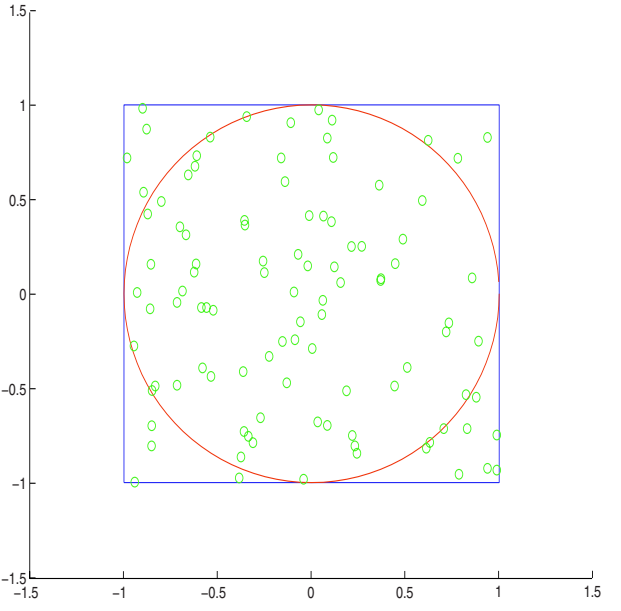

Fig. 2. A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1

Intuitively, without any knowledge of elementary statistics, a sensible technique to estimate the probability $\mathbb{P}(D \in \mathcal{A})$ of falling in a given region $\mathcal{A} \subset \mathcal{S}$ (and think for example of $\mathcal{A} = \mathcal{D}$) would consist of using the following formula

$$\mathbb{P}(D \in \mathcal{A}) \simeq \frac{\text{number of drops that fell in } \mathcal{A}}{N}.$$

This formula certainly makes sense, but we would like to be more rigorous and give a statistical justification to it.

$\mathbb{P}(D \in \mathcal{A})$ *as an Expectation.* Let us first introduce the indicator function of a set $\mathcal{A}$, defined as follows,

$$\mathbb{I}_{\mathcal{A}}(x, y) = \begin{cases} 1 \text{ if point } D = (x, y) \in \mathcal{A}, \\ 0 \text{ otherwise,} \end{cases}.$$

We define the random variable $V(D) := \mathbb{I}_{\mathcal{A}}(D) := \mathbb{I}_{\mathcal{A}}(X, Y)$, where X, Y are the random variables that represent the Cartesian coordinates of a uniformly distributed point on $\mathcal{S}$, denoted $D \sim \mathcal{U}_{\mathcal{S}}$. Using V, it is not hard to show that

$$\mathbb{P}(D \in \mathcal{A}) = \int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x, y) \frac{1}{4} dx dy = \mathbb{E}_{\mathcal{U}_{\mathcal{S}}}(V),$$

where for a probability distribution π we will denote $\mathbb{E}_{\pi}$ the expectation with respect to π.

The Law of Large Numbers. Now, similarly, let us introduce $\{V_i := V(D_i), i = 1, \ldots, N\}$ the random variables associated to the drops $\{D_i, i = 1, \ldots, N\}$ and consider the sum

$$S_N = \frac{\sum_{i=1}^{N} V_i}{N}. \tag{2}$$

We notice that an alternative expression for S_N is

$$S_N = \frac{\text{number of drops that fell in } \mathcal{A}}{N},$$

which corresponds precisely to the formula which we intuitively suggested to approximate $\mathbb{P}(D \in \mathcal{A})$. However Eq. (2) is statistically more explicit, in the sense that it tells us that our suggested approximation of $\mathbb{P}(D \in \mathcal{A})$ is the empirical average of independent and identically distributed random variables, $\{V_i, i = 1, \ldots, N\}$. Assuming that the rain lasts forever and therefore that $N \to +\infty$, then one can apply the *law of large numbers* (since $\mathbb{E}_{\mathcal{U}_\mathcal{S}}(|V|) < +\infty$ here) and deduce that

$$\lim_{N \to +\infty} S_N = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V), \text{ (almost surely).}$$

As we have already proved that $\mathbb{P}(D \in \mathcal{A}) = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V)$, the law of large numbers mathematically justifies our intuitive method of estimating $\mathbb{P}(D \in \mathcal{A})$, provided that N is large enough.

A Method of Approximating π. We note that as a special case we have defined a method of calculating π. Indeed,

$$\mathbb{P}(D \in \mathcal{D}) = \int_{\mathcal{D}} \frac{1}{4} dxdy = \frac{\pi}{4}.$$

S_N as defined in Eq. (2) with $\mathcal{A} = \mathcal{D}$ is an unbiased estimator of $\pi/4$, which is also ensured to converge towards $\pi/4$ for N very large. The quantity $S_N - \pi/4$ for a day of rain as a function of the number of drops for one rainfall is presented in Figure 3. However in practice one is interested in obtaining a result in finite time, i.e. for N finite. S_N is a random variable which can be rewritten as $S_N = \pi/4 + E_N$ where E_N is a random error term. It is naturally of interest to characterize the precision of our estimator, i.e. characterize the average magnitude of the fluctuations of the random error E_N, as illustrated in Figure 4. A simple measure of the average magnitude of E_N is its variance,

$$var(E_N) = var(S_N) = \frac{1}{N} var(V_1),$$

as the $\{V_i, i = 1, \ldots, N\}$ are independent. It is worth remembering that since S_N is unbiased,

$$\sqrt{var(S_N)} = \sqrt{\mathbb{E}\left[(S_N - \mathbb{P}(D \in \mathcal{D}))^2\right]},$$

which using the result above implies that the *mean square error* between S_N and $\mathbb{P}(D \in \mathcal{D})$ decreases as $1/\sqrt{N}$. This is illustrated in Figure 5 where the dotted

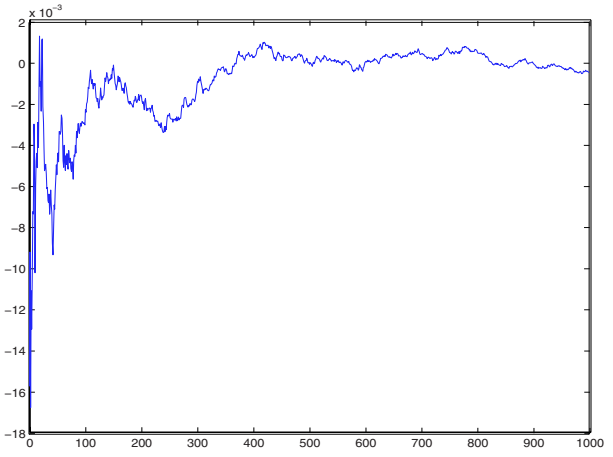

Fig. 3. Convergence of $S_N - \pi/4$ as a function of the number of samples, for one realization (or rainfall)

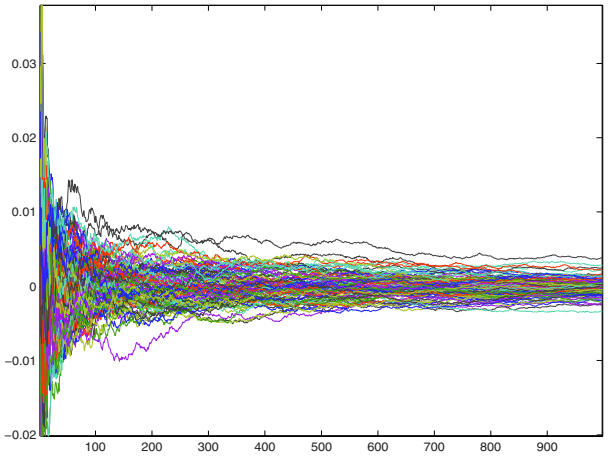

Fig. 4. Convergence of $S_N - \pi/4$ for 100 realizations of the rain

lines represent $\pm\sqrt{var(V)/N}$ and the dashed lines represent the empirical mean square error $S_N - \pi/4$ estimated from the 100 realizations in Figure 4. One can be slightly more precise and first invoke here an asymptotic result, the *central limit theorem* (which can be applied here as $var(V) < +\infty$). As $N \to +\infty$,

$$\sqrt{N} S_N \to_d \mathcal{N}(\pi/4, var(V)),$$

which implies that for N large enough the probability of the error being larger than $2\sqrt{var(V)/N}$ is

$$\mathbb{P}\left(|S_N - \pi/4| > 2\sqrt{var(V)/N}\right) \simeq 0.05,$$

with $2\sqrt{var(V)} = 0.8211$. In the present case (we are sampling here from a Bernoulli distribution) one can be much more precise and use a non-asymptotic result. Indeed, using a Bernstein type inequality, one can prove [22, p. 69] that for any integer $N \geq 1$ and $\varepsilon > 0$,

$$\mathbb{P}\left(|S_N - \pi/4| > \varepsilon\right) \leq 2\exp\left(-2N\varepsilon^2\right)$$

which tells us that for any $\alpha \in (0, 1]$,

$$\mathbb{P}\left(|S_N - \pi/4| > \varepsilon\right) < \alpha$$

which on the one hand provides us with a minimum number of samples in order to achieve a given precision of α,

$$N = \left[\frac{\log(2/\alpha)}{2\varepsilon^2}\right],$$

where for a real x the quantity $[x]$ denotes the integer part of x, or alternatively tells us that for any $N \geq 1$,

$$\mathbb{P}\left(|S_N - \pi/4| > \sqrt{\frac{\log(40)}{2N}}\right) \leq 0.05$$

with $\sqrt{\log(40)/2} = 1.3541$.

Both results tell us that in some sense the approximation error is inversely proportional to $\sqrt{N}$.

A General and Powerful Method. Now consider the case where $\mathsf{X} = \mathbb{R}^{n_x}$ for any integer n_x, and in particular large values of n_x. Replace now $\mathcal{S}$ and $\mathcal{D}$ above with a hypercube $\mathcal{S}^{n_x}$ and an inscribed hyperball $\mathcal{D}^{n_x}$ in X. If we could observe a hyper-rain, then it would not be difficult to see that the method described earlier to estimate the area of $\mathcal{D}$ could be used to estimate the volume of $\mathcal{D}^{n_x}$. The only requirement is that one should be able to tell if a drop fell in $\mathcal{D}^{n_x}$ or not: in other words one should be able to calculate $\mathbb{I}_{\mathcal{D}^{n_x}}(D)$ point-wise. Now a very important result is that the arguments that lead earlier to the formal validation of the Monte Carlo approach to estimate $\pi/4$ remain identical here (check it to convince yourself!). In particular the rate of convergence of the estimator in the mean square sense is again *independent of the dimension* n_x.

This would not be the case if we were using a deterministic method on a grid of regularly spaced points. Typically, the rate of convergence of such deterministic methods is of the form $1/N^{r/n_x}$ where r is related to the smoothness of the contours of region $\mathcal{A}$, and is N the number of function (here $\mathbb{I}_{\mathcal{A}}$) evaluations. Monte Carlo methods are thus extremely attractive when n_x is large.

A More General Context. In the previous subsection, we have seen that a simple experience involving the rain can help us to evaluate an *expectation* in an extremely simple way. In this subsection we generalist the ideas developed earlier in order to tackle the generic problem of estimating

$$\mathbb{E}_\pi(f(x)) \triangleq \int_{\mathsf{X}} f(x)\pi(x)dx,$$

where $f : \mathsf{X} \to \mathbb{R}^{n_f}$ and π is a probability distribution on $\mathsf{X} \subset \mathbb{R}^{n_x}$. We will assume that $\mathbb{E}_\pi(|f(x)|) < +\infty$ but that it is difficult to obtain an analytical expression for $\mathbb{E}_\pi(f(x))$.

1.2 Generalization of the Rain Experiment

In the light of the square/circle example, assume that $N >> 1$ *i.i.d.* samples $X^{(i)} \sim \pi$ $(i = 1, \ldots, N)$ are available to us (since it is unlikely that rain can generate samples from any distribution π, we will address the problem of sample generation in the next section). Now consider any set $\mathcal{A} \subset \mathsf{X}$ and assume that we are interested in calculating $\pi(\mathcal{A}) = \mathbb{P}(X \in \mathcal{A})$ for $X \sim \pi$. We naturally choose the following estimator

$$\pi(\mathcal{A}) \simeq \frac{\text{number of samples in } \mathcal{A}}{\text{total number of samples}},$$

which by the law of large numbers is a consistent estimator of $\pi(\mathcal{A})$ since

$$\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\mathcal{A}}(X_i) = \mathbb{E}_\pi(\mathbb{I}_{\mathcal{A}}(X)) = \pi(\mathcal{A}).$$

A way of generalizing this in order to evaluate $\mathbb{E}_\pi(f(x))$ consists of considering the estimator

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

which is unbiased. From the law of large numbers $S_N(f)$ will converge and

$$\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} f(X_i) = \mathbb{E}_\pi(f(X)) \ a.s.$$

Here again a good measure of the approximation is the variance of $S_N(f)$,

$$var_\pi\left[S_N(f)\right] = var_\pi\left[\frac{1}{N}\sum_{i=1}^{N} f(X^{(i)})\right] = \frac{var_\pi\left[f(X)\right]}{N}.$$

Now the central limit theorem applies if $var_\pi\left[f(X)\right] < \infty$ and tells us that

$$S_N(f) \overset{N\to+\infty}{\to}_d \mathcal{N}\left(\sqrt{N}\mathbb{E}_\pi(f(X)), var_\pi\left[f(X)\right]\right),$$

and the conclusions drawn in the rain example are still valid here:

1. The rate of convergence is immune to the dimension of X.
2. It is easy to take complex integration domains into account.
3. It is easily implementable and general. The requirements are
   (a) to be able to evaluate $f(x)$ for any $x \in \mathsf{X}$,
   (b) to be able to produce samples distributed according to π.

1.3 From the Algebraic to the Sample Representation

In this subsection we make explicit the - approximate - sample representation of π. Let us first introduce the delta-Dirac function δ_{x_0} for $x_0 \in \mathsf{X}$, defined as follows

$$\int_{\mathsf{X}} f(x)\delta_{x_0}(x)dx = f(x_0),$$

for any $f : \mathsf{X} \to \mathbb{R}^{n_f}$. Note that this implies in particular that for $\mathcal{A} \subset \mathsf{X}$,

$$\int_{\mathsf{X}} \mathbb{I}_{\mathcal{A}}(x)\delta_{x_0}(x)dx = \int_{\mathcal{A}} \delta_{x_0}(x)dx = \mathbb{I}_{\mathcal{A}}(x_0).$$

Now, for $X_i \sim \pi$ for $i = 1, \ldots, N$, we can introduce the following mixture of delta-Dirac functions

$$\widehat{\pi}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}(x),$$

which is the *empirical measure* of the sample, and consider for any $\mathcal{A} \subset \mathsf{X}$

$$\widehat{\pi}_N(\mathcal{A}) \triangleq \int_{\mathcal{A}} \widehat{\pi}_N(x)\, dx = \sum_{i=1}^{N} \int_{\mathcal{A}} \frac{1}{N}\delta_{X_i}(x) = \sum_{i=1}^{N} \frac{1}{N}\mathbb{I}_{\mathcal{A}}(x).$$

which is precisely $S_N(\mathbb{I}_{\mathcal{A}})$. What we have touched upon here is simply the sample representation of π, of which an illustration can be found in Figure 6 for a Gaussian distribution. **The concentration of points in a given region of the space represents** π. Note that this approach is in contrast with what is usually done in parametric statistics, i.e. start with samples and then introduce a distribution with an algebraic representation for the underlying population. Note that here each sample X_i has a weight of $1/N$, but that it is also possible to consider weighted sample representations of π: the approach is called *importance sampling* and will be covered later on.

Now consider the problem of estimating $\mathbb{E}_\pi(f)$. We simply replace π with its sample representation $\widehat{\pi}_N$ and obtain

$$\mathbb{E}_\pi(f) \simeq \int_{\mathsf{X}} f(x) \sum_{i=1}^{N} \frac{1}{N}\delta_{X_i}(x)\, dx = \sum_{i=1}^{N} \frac{1}{N} \int_{\mathsf{X}} f(x)\, \delta_{X_i}(x)\, dx = \frac{1}{N}\sum_{i=1}^{N} f(X_i),$$

which is precisely $S_N(f)$, the Monte Carlo estimator suggested earlier. The interest of this approximating representation of π will become clearer later, in particular in the context of importance sampling.

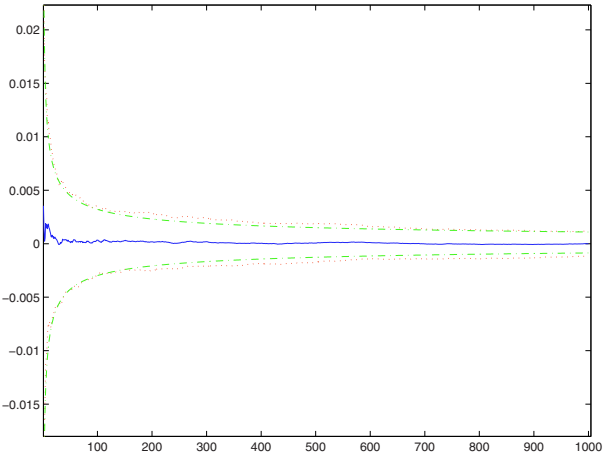

Fig. 5. Variance of $S_N - \pi/4$ across 100 realizations as a function of the number of samples and the theoretical variance

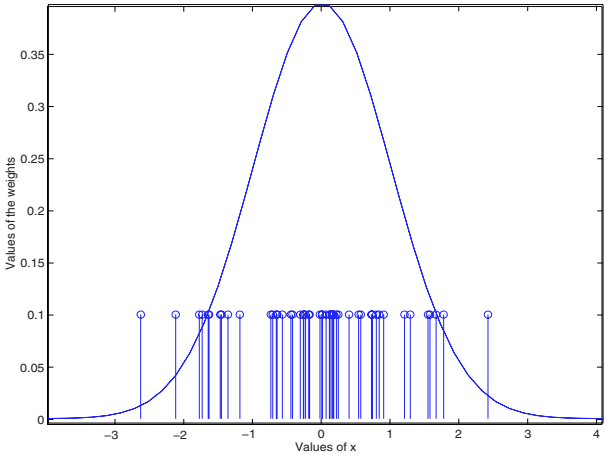

Fig. 6. Sample representation of a Gaussian distribution

1.4 Expectations in Statistics

The aim of this subsection is to illustrate why it is important to compute expectations in statistics, in particular in the Bayesian context.

Assume that we are given a Bayesian model, i.e. a likelihood $p(y|\theta)$ and a prior distribution $p(\theta)$. We observe some data y and wish to estimate θ. In a Bayesian framework, all the available information about θ is summarized by the posterior distribution, given by Bayes' rule,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}.$$

The expression looks simple, but the bottom of the fraction is an integral, and more precisely an expectation

$$\mathbb{E}_{p(\theta)}(p(y|\theta)) = \int_\Theta p(y|\theta)p(\theta)d\theta.$$

In many situations this integral typically does not admit a closed-form expression.

Example 1. We observe $y=(y_1, y_2, \ldots, y_T)$ which are *iid* such that $y_i \sim \mathcal{N}(\mu_j, \sigma_j^2)$ with probability p_j for $j = 1, 2$. Here $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1)$. The likelihood in this case is

$$p(y|\theta) = \prod_{i=1}^{T} \left[p_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}} + (1-p_1)\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}} \right].$$

The normalizing constant of the posterior can be complicated, e.g. impose *a priori* constraints on the parameters $\sigma_1^2 < 10\sigma_2^2 + \sqrt{\mu_1\mu_2}$ and $\mu_2 < \pi$.

Other important examples include the evaluation of the posterior mean square estimate of θ,

$$\hat{\theta}_{MSE} := \mathbb{E}_{p(\theta|y)}(\theta) = \int_\Theta \theta p(\theta|y)d\theta,$$

the median, i.e. the solution $\hat{\theta}_{median}$ of

$$\mathbb{E}_{p(\theta|y)}(\mathbb{I}(\theta \leq \hat{\theta}_{median})) = \int_{-\infty}^{+\infty} \mathbb{I}(\theta \leq \hat{\theta}_{median})p(\theta|y)d\theta = 1/2.$$

but also the evaluation of the marginal posterior distribution $p(\theta_1|y)$ of $p(\theta_1, \theta_2|y)$,

$$\begin{aligned} p(\theta_1|y) &= \int_\Theta p(\theta_1, \theta_2|y)d\theta_2 \\ &= \int_\Theta p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2 \\ &= \mathbb{E}_{p(\theta_2|y)}(p(\theta_1|\theta_2, y)) \ldots \end{aligned}$$

Similar problems are encountered when computing, marginal posterior means, posterior variances, posterior credibility regions.

1.5 A Simple Application

In 1786 Laplace was interested in determining if the probability θ of a male birth in Paris over a certain period of time was above 0.5 or not. The official figures gave $y_1 = 251,527$ males birth for $y_2 = 241,945$ female births. The observed proportion was therefore 0.509,. We choose a uniform distribution as prior distribution for θ the proportion of male births. The posterior distribution is

$$p(\theta|y) = \mathcal{B}e(\theta; 251528, 241946).$$

Imagine that we have no table and are interested in the posterior mean of this posterior distribution. Furthermore, imagine that we can sample (using a computer) a large number N of independent samples $(\theta_i, i = 1, \ldots, N)$ from this distribution. One could propose the following estimator

$$\frac{1}{N}\sum_{i=1}^{N}\theta_i$$

as from the law of large numbers,

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\theta_i = \mathbb{E}_{p(\theta|y)}(\theta).$$

We could also estimate the posterior variance as

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\theta_i^2 = \mathbb{E}_{p(\theta|y)}(\theta^2).$$

Now consider the following more challenging problems: we want to find estimates of the median of this posterior distribution, as well as a 95% credibility interval. We start with the median, and assume that we have ordered the samples, that is for any $i < j$, $\theta_i < \theta_j$ and for simplicity that N is an even number. Let $\bar{\theta}$ be the median of the posterior distribution. Then we know that

$$\mathbb{P}(\theta_i \geq \bar{\theta}) = \int_{-\infty}^{+\infty}\mathbb{I}(\bar{\theta} < \theta)p(\theta|y)d\theta = 1/2$$
$$\mathbb{P}(\theta_i \leq \bar{\theta}) = \int_{-\infty}^{+\infty}\mathbb{I}(\bar{\theta} < \theta)p(\theta|y)d\theta = 1/2$$

so that (assuming for simplicity that N is even and that we have ordered $(\theta_i, i = 1, \ldots, N)$), it is sensible to chose an estimate for $\bar{\theta}$ between $\theta_{N/2}$ and $\theta_{N/2+1}$. Now assume that we are looking for θ^- and θ^+ such that

$$\mathbb{P}(\theta^- \leq \theta \leq \theta^+) = \int_{-\infty}^{+\infty} I(\theta^- \leq \theta \leq \theta^+)p(\theta|y)d\theta = 0.95$$

or

$$\mathbb{P}(0 \leq \theta \leq \theta^-) = 0.025 \text{ and } \mathbb{P}(\theta^+ \leq \theta \leq 1) = 0.025$$

and assuming again for simplicity that $N = 1000$ and that the samples have been ordered. We find that a reasonable estimate of θ^- is between θ_{25} and θ_{26} and an estimate of θ^+ between θ_{975} and θ_{976}. Finally we might be interested in calculating

$$\mathbb{P}(\theta < 0.5) = \int_0^{0.5} p(\theta|y)d\theta = \int_0^1 I(\theta \leq 0.5)p(\theta|y)d\theta$$

which suggests the following estimator of this probability

$$\mathbb{P}(\theta < 0.5) \simeq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\theta_i \leq 0.5).$$

(one can in fact find that $\mathbb{P}(\theta \leq 0.5|y_1, y_2) = 1.146058490255674 \times 10^{-42}$).

1.6 Further Topic: Importance Sampling

In this subsection we explore the important method of importance sampling.[1] This method is of interest either in the case where samples from the desired distribution π are not available, but samples from a distribution q are, or as a way of possibly reducing the variance of an estimator.

Importance Sampling. Consider a probability distribution q such that $\pi(x) > 0 \Rightarrow q(x) > 0$. Then one can write

$$\mathbb{E}_\pi(f(x)) = \int_{\mathsf{X}} f(x)\pi(x)dx = \int_{\mathsf{X}} f(x)\underbrace{\frac{\pi(x)}{q(x)}}_{w(x)}q(x)dx = \mathbb{E}_q(w(x)f(x))$$

We are now integrating the function $w(x)f(x)$ with respect to the distribution q. Now provided that we can produce N *i.i.d.* samples $X_1, \ldots, X_N$ from q, then one can suggest the following estimator

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi(X_i)}{q(X_i)}f(X_i) = \int_{\mathsf{X}} f(x)\frac{1}{N}\sum_{i=1}^{N}\frac{\pi(X_i)}{q(X_i)}\delta_{X_i}(x)dx.$$

It is customary to call $w_i = \frac{\pi(X_i)}{q(X_i)}$ the *importance weight* and q the importance distribution. Now it is natural to introduce a delta-Dirac approximation of π is of the form

$$\widehat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} w_i\delta_{X_i}(dx)$$

The interpretation of this weighted empirical measure is rather simple. Large w_i's indicate an underrepresentation of π by samples from q around X_i. Small w_i's indicate an overrepresentation of π by samples from q around X_i. This phenomenon is illustrated in Figure 1 where the importance weights required to represent a double exponential with samples from either a Gaussian or a t-Student are presented. Note that in the case where $q = \pi$ then $w_i = 1/N$ and we recover the representation presented earlier.

It is also worth noticing that if the normalizing constants of π and/or q are not known, then it is possible to define (with $\pi^*(x) \propto \pi(x)$ and $q^*(x) \propto q(x)$)

$$w_i = \frac{\pi^*(X_i)/q^*(X_i)}{\sum_{j=1}^{N}\pi^*(X_j)/q^*(X_j)}.$$

[1] This material can be skipped at first.

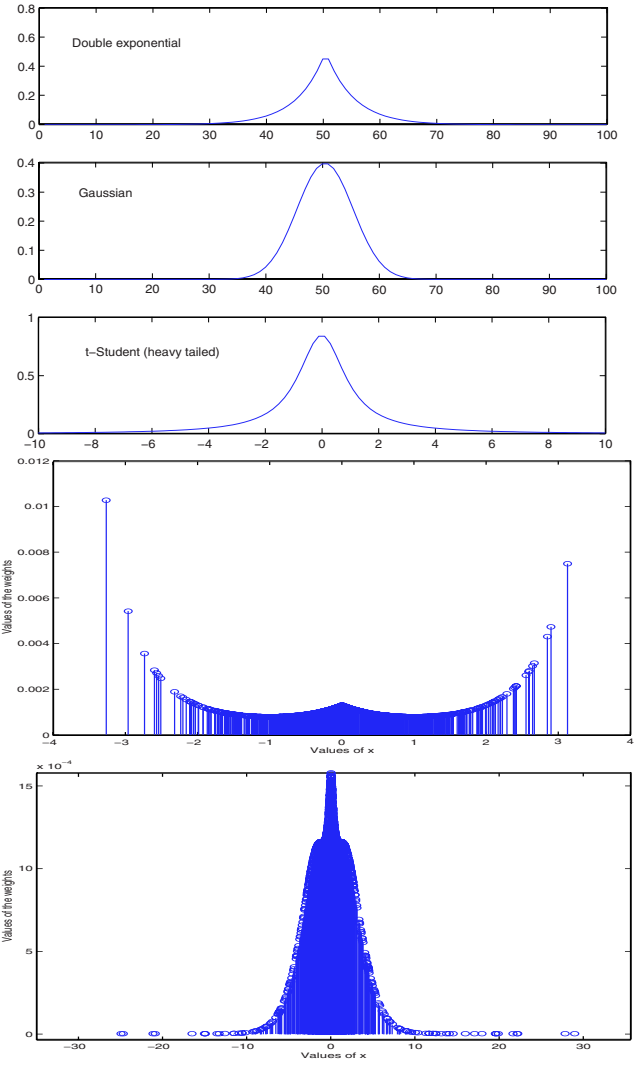

Fig. 7. Top: The three distributions. Middle: importance weights to represent a double exponential with samples from a Gaussian. Bottom: importance weights to represent a double exponential with samples from a t-Student

And consider the following estimator

$$I_N(f) = \sum_{i=1}^{N} w_i f(X_i) = \sum_{i=1}^{N} \frac{\pi^*(X_i)/q^*(X_i)}{\sum_{j=1}^{N} \pi^*(X_j)/q^*(X_j)} f(X_i).$$

This estimator is *biased*, but

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\pi^*(X_i)/q^*(X_i)}{\sum_{j=1}^{N}\pi^*(X_j)/q^*(X_j)}f(X_i)=\frac{\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\pi^*(X_i)/q^*(X_i)f(X_i)}{\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\pi^*(X_j)/q^*(X_j)}$$
$$=\frac{\int_{\mathsf{X}}f(x)\,w(x)\,q(x)\,dx}{\int_{\mathsf{X}}w(x)\,q(x)\,dx}$$

as the unknown normalizing constants cancel.

Example 2 (Naive). In a Bayesian framework the target distribution is $\pi(\theta)\triangleq p(\theta|y)$, the posterior distribution. One can suggest (and this is not necessarily a good choice) $q(\theta)\triangleq p(\theta)$. In this case the weights will be proportional to the likelihood since

$$w(\theta)=p(\theta|y)/p(\theta)\propto\frac{p(y|\theta)\,p(\theta)}{p(\theta)}\propto p(y|\theta).$$

Unfortunately this technique is not as general as it might seem. Let us consider the variance of the importance sampling estimator in the simple case where the normalizing constants are known and where $f=C$, i.e. is a constant. In this case

$$var_q(I_N(f))=\frac{C}{N}\left[\mathbb{E}_q\left(w_1^2\right)-\mathbb{E}_q\left(w_1\right)^2\right]$$

which suggests that even in the simplest case the variance of the weights should be finite and as small as possible for the variance of $I_N(f)$ to be small. The examples provided earlier in Figure 1, where π was a double exponential and q either a normal or t-Student distribution, illustrates the possibly large variations of the weights.

Zero Variance Estimator. Here we illustrate a possible interest of importance sampling, which is however specialized. We start with the trivial remark that the variance of a constant function is null, i.e. $var_\pi[f]=0$ if f is a constant. We seek here to exploit this property in the context of Monte Carlo integration, although this might seem of little interest at first sight since no numerical method is needed to evaluate $\mathbb{E}_\pi(f)$ for a constant function f. However we are going to use this as a motivation to describe a method of reducing the variance of a Monte Carlo estimator for a fixed number of samples. Now assume that $\mathbb{E}_\pi(f)$ and that for simplicity $f\geq 0$. Using the convention $0/0=1$ we can rewrite $\mathbb{E}_\pi(f)$ as

$$\mathbb{E}_\pi(f)=\int_{\mathsf{X}}f(x)\pi(x)dx=\int_{\mathsf{X}}\frac{f(x)}{f(x)}\pi(x)f(x)dx$$
$$=\int_{\mathsf{X}}1\times\int_{\mathsf{X}}\pi(x'')f(x'')dx''\frac{\pi(x)f(x)}{\int_{\mathsf{X}}\pi(x')f(x')dx'}dx,$$

that is

$$\mathbb{E}_\pi(f)=\mathbb{E}_q(\mathbb{E}_\pi(f))$$

where

$$q(x) := \frac{\pi(x)f(x)}{\int_{\mathsf{X}} \pi(x')f(x')dx'}$$

can be thought of as being a probability density. If we could sample from q then we could integrate the constant function $\int_{\mathsf{X}} \pi(x'')f(x'')dx''$ and obtain a zero variance estimator. Naturally we have not solved the problem since the constant is precisely the integral that we are seeking to calculate!

The calculations above can be generalized to functions f that are not everywhere positive, with in this case,

$$q\left(x\right) = \frac{\left|f\left(x\right)\right|\pi\left(x\right)}{\int_{\mathsf{X}} \pi(x')|f(x')|dx'}.$$

Despite our disappointing/absurd results, the strategy however suggests ways to improve the constant $var_\pi(f)$, by trying to sample from a distribution close to q. Note however that q depends on f, and that as a consequence such a method is therefore very specialized.

Conclusions. To summaries, the pros and cons of importance sampling are as follows:

- **Advantages.** Easy to implement, parallelizable, sequential version are possible (particle filter etc.). If q is a clever approximation of π, then we typically expect good results. It can be used a specialized way of reducing the variance of estimators.
- **Drawbacks.** If we do not have $var_\pi(w(x)) < +\infty$, then typically $\widehat{I}_N\left(f\right)$ can be a poor estimator since its variance is large. This poses the problem of the choice of $q\left(x\right)$? Where are the modes of $\pi\left(x\right)$? Importance sampling is typically limited to small dimensions for the parameter space, say $n_x = 10-50$ depending on the application.

Despite the possible drawbacks, importance sampling has proved to be extremely useful in the context of sequential importance sampling.

2 Classical "Exact" Simulation Methods

In this section we review some classical simulation techniques. We call those techniques "exact" as they allow one to generate samples in a *finite number of iterations* of a procedure. Note that the instant when a sample from the distribution of interest is produced is identifiable, that is we can stop the procedure and be sure that we have generated a sample from the distribution of interest. As we shall see in the next section this is not always the case. Unfortunately the simulation techniques presented in this chapter cannot typically be used in order to sample from complex distributions as they tend not to scale well with the dimension n_x and cases where little is known about π. However these techniques can be thought of as being building blocks of more complex algorithms that will be presented in the next chapter.

From now on we will assume that a computer can generate *independent uniformly distributed random variables*, or at least that it can generate a good approximation of such random variables (indeed computers should usually follow a deterministic behavior, and one must find ways around in order to produce something that looks random).

2.1 The cdf Inversion Method

We present here this method in the case where $\mathsf{X} = \mathbb{R}$ for simplicity. The multivariate generalization is not difficult. First we consider a simple discrete example where $X \in \mathsf{X} = \{1, 2, 3\}$ and such that

$$\mathbb{P}(X = 1) = \frac{1}{6},\ \mathbb{P}(X = 2) = \frac{2}{6},\ \mathbb{P}(X = 3) = \frac{1}{2}.$$

Define the cumulative probability distribution (cdf) of X as

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i=1}^{3} \mathbb{P}(X = i)\mathbb{I}(i \leq x)$$

for $x \in [0, 3]$ and its inverse

$$F_X^{-1}(u) = \inf\{x \in \mathsf{X}; F_X(x) \geq u\},$$

for $u \in [0, 1]$. The cdf corresponding to our example is represented in Figure 8. A method of sampling from this distribution consists of sampling $u \sim \mathcal{U}(0, 1)$ and find $x = F_X^{-1}(u)$. The probability of u falling in the vertical interval i is precisely equal to the probability $\mathbb{P}(X = i)$. The method indeed produces samples from the distribution of interest.

Now in the continuous case, and assuming that the distribution has a density the cdf takes the form

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{+\infty} \pi(u)\, I(u \leq x) du = \int_{-\infty}^{x} \pi(u)\, du.$$

A normal distribution and its cdf are presented in Figure 9. Intuitively the algorithm suggested in the discrete case should be valid here, since modes of π mean large variations of F_X and therefore a large probability for a uniform distribution to fall in these regions.

More rigorously, consider the algorithm

$$\text{Sample } u \sim \mathcal{U}(0, 1) \text{ and set } Y = F_\pi^{-1}(u).$$

We prove that this algorithm produces samples from π. We calculate the cdf of X produced by the algorithm above. For any $y \in \mathsf{X}$ we have

$$\begin{aligned}
\mathbb{P}(Y \leq y) &= \mathbb{P}\left(Y = F_X^{-1}(u) \leq y\right) \\
&= \mathbb{P}(u \leq F_X(y)) \text{ since } F_X \text{ is non decreasing} \\
&= \int_0^1 I(u \leq F_X(y)) \times 1 du = \int_0^{F_X(y)} du = F_X(y),
\end{aligned}$$

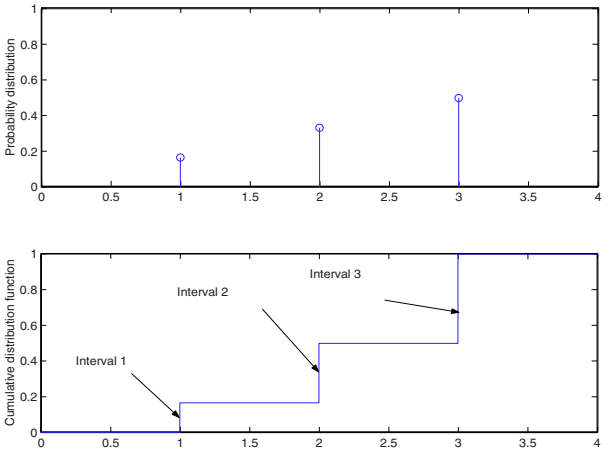

Fig. 8. The distribution and cdf of a discrete random variable

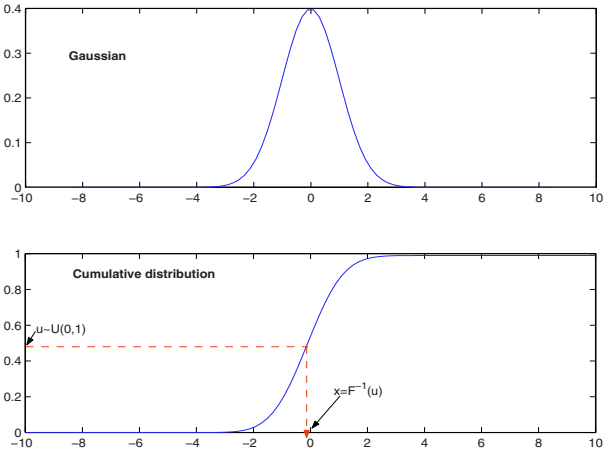

Fig. 9. The distribution and cdf of a normal distribution

which shows that the cdf of Y produced by the algorithm above is precisely the cdf of $X \sim \pi$.

Example 3. Consider the exponential distribution with parameter 1, i.e. $X \sim \pi(x) = \exp(-x)\,\mathbb{I}_{[0,+\infty)}(x)$. The cdf of X is $F_X(x) = 1 - \exp(-x)$. Now the inverse cdf is $F_X^{-1}(u) = -\log(1-u)$, and for $u \sim \mathcal{U}(0,1)$ then $-\log(1-u) \sim \pi$.

This example is interesting as it illustrates one of the fundamental idea of most simulation methods: sample from a distribution from which it is easy to sample (here the uniform distribution) and then transform this random variable

(here through F_X^{-1}). However this method is only applicable to a limited number of cases as it requires a closed form expression of the inverse of the cdf, which is not explicit even for a distribution as simple and common as the normal distribution.

2.2 The Rejection Method

The rejection method allows one to sample according to a distribution π which is only known up to a proportionality constant, say $\pi^* \propto \pi$. It relies again on the assumption that samples can be generated from a so-called *proposal* distribution q defined on X, which might as well be known only up to a normalizing constant, say $q^* \propto q$. Then, instead of being transformed by a deterministic function as in the inverse cdf method, the samples produced from π are either rejected or accepted. More precisely, assume that for any $x \in \mathsf{X}$, $C = \sup_{x \in \mathsf{X}} \frac{\pi^*(x)}{q^*(x)} < +\infty$ (note that this imposes that for any $x \in \mathsf{X}$, $\pi^*(x) > 0 \Rightarrow q^*(x) > 0$) and consider $C' \geq C$. Then the accept/reject procedure proceeds as follows:

Accept/Reject procedure

1. Sample $Y \sim q$ and $u \sim \mathcal{U}(0,1)$.
2. If $u < \frac{\pi^*(Y)}{C' q^*(Y)}$ then return Y; otherwise return to step 1.

The intuition behind the method can be understood from Figure 10.

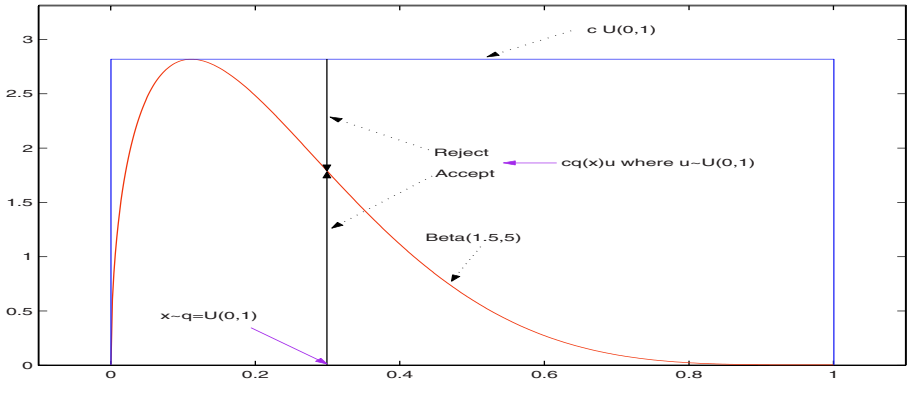

Fig. 10. The idea behind the rejection method

Now we prove that $\mathbb{P}(Y \leq x | Y \text{ accepted}) = \mathbb{P}(X \leq x)$. We will extensively use the trivial identity

$$q(x) = \frac{q^*(x)}{\int_{\mathsf{X}} q^*(y)dy}.$$

For any $x \in \mathsf{X}$, consider the joint distribution

$$\begin{aligned}\mathbb{P}\left(Y \leq x \text{ and } Y \text{ accepted}\right) &= \int_0^1 \int_{-\infty}^{x} I(u \leq \frac{\pi^*(y)}{C'q^*(y)})q(y) \times 1 dy du \\ &= \int_{-\infty}^{x} \frac{\pi^*(y)}{C'q^*(y)} q(y) dy \\ &= \frac{\int_{-\infty}^{x} \pi^*(y) dy}{C' \int_{\mathsf{X}} q^*(y) dy},\end{aligned}$$

and the probability of being accepted is the marginal of $\mathbb{P}\left(Y \leq x \text{ and } Y \text{accepted}\right)$, that is

$$\mathbb{P}\left(Y \text{ accepted}\right) = \int_{\mathsf{X}} \frac{\pi^*\left(y\right)}{C'q^*\left(y\right)} q\left(y\right) dy = \frac{\int_{\mathsf{X}} \pi^*\left(y\right) dy}{C' \int_{\mathsf{X}} q^*\left(y\right) dy}. \tag{3}$$

Consequently

$$\mathbb{P}(Y \leq x | Y \text{ accepted}) = \frac{\int_{-\infty}^{x} \pi^*\left(y\right) dy}{\int_{\mathsf{X}} \pi^*\left(y\right) dy} = \int_{-\infty}^{x} \pi\left(y\right) dy.$$

The expression for the probability of being accepted in Eq. (3) tells us that in order to design an efficient algorithm, C' should be chosen as small as possible, and that the optimal choice corresponds to C. However this constant might be very large, in particular for large n_x and C might not even be known. In the most favorable scenarios, at best an upper bound might be known.

Example 4. We want to sample from a $\mathcal{B}e(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$ distribution. We can generate samples from $\mathcal{U}(0,1)$. One can find $\sup_{x \in [0,1]} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{1}$ analytically for $\alpha, \beta > 1$! Note that we do not assume known the normalizing constant!

Example 5. Let us assume that one wants to simulate samples from $\pi\left(\theta\right) \triangleq p(\theta|y) \propto p(y|\theta)p\left(\theta\right)$. We assume that $p(y|\theta)$ is known analytically and $p(y|\theta) \leq C$ for any θ, where C is known. We also assume that we are able to simulate from $p\left(\theta\right)$. Thus one can choose $q\left(\theta\right) = p\left(\theta\right)$ and use the accept/reject procedure to sample from $p(\theta|y)$. Indeed

$$\frac{p(\theta|y)}{p\left(\theta\right)} = \frac{p(y|\theta)}{p(y)} \leq \frac{C}{p(y)} = M \tag{4}$$

is bounded and

$$\frac{\pi\left(\theta\right)}{Mq\left(\theta\right)} = \frac{p(\theta|y)}{\frac{C}{p(y)}p\left(\theta\right)} = \frac{p(y|\theta)}{C} \tag{5}$$

can be evaluated analytically. However, the acceptance rate $1/M$ is usually unknown as it involves $p(y)$ which is itself usually unknown.

We can summaries the pros and cons of the accept/reject procedure:

– **Advantages:**
  1. seems rather universal, and compared to the inverse cdf method requires less algebraic properties.
  2. in principle neither the normalization constant of π nor that of q are needed.
– **Drawbacks:**
  1. how to construct the proposal $q(x)$ to minimize C?
  2. typically C increases exponentially with n_x.

2.3 Deterministic Transformations

These methods rely on clever changes of variables, which transform one distribution to another. A typical setup is the following: consider $Y \sim q$ from which it is easy to sample, and consider $g : \mathsf{X} \to \mathsf{X}$ a differentiable and one-to-one transformation. Now define the transformed random variable

$$X = g(Y).$$

We know that the density, say π, of X can be expressed in terms of q and the Jacobian $\left|\frac{\partial g^{-1}(x)}{\partial x}\right|$ of the transformation g as follows

$$\pi(x) = q\left(g^{-1}(x)\right)\left|\frac{\partial g^{-1}(x)}{\partial x}\right|.$$

Naturally for a predefined π it is not always obvious to find proper g and q, but we present here a celebrated example. The Box-Muller transformation is a method of transforming two *i.i.d.* uniformly distributed random variables Y_1 and Y_2 on $[0,1]$ into two *i.i.d.* normally distributed random variables X_1 and X_2 with distribution $\mathcal{N}(0,1)$. The transformation is as follows

$$\begin{aligned} X_1 &= \sqrt{-2\log(Y_1)}\cos(2\pi Y_2) \\ X_2 &= \sqrt{-2\log(Y_1)}\sin(2\pi Y_2). \end{aligned} \tag{6}$$

We compute the inverse transformation and find that

$$\begin{aligned} Y_1 &= \exp\left(-(X_1^2 + X_2^2)/2\right) \\ Y_2 &= \frac{1}{4} + \frac{1}{2\pi}\arctan\left(\frac{X_2}{X_1}\right) \end{aligned}$$

Now one can check that the Jacobian of the transformation is

$$\frac{1}{\left(\sqrt{2\pi}\right)^2}\exp\left(-(x_1^2 + x_2^2)/2\right).$$

Consequently

$$\pi(x_1, x_2) = \frac{1}{\left(\sqrt{2\pi}\right)^2}\exp\left(-(x_1^2 + x_2^2)/2\right) \times 1,$$

which proves the result. This method is simple to implement on a computer, and is to a certain extent efficient in the sense that two uniformly distributed random variables Y_1 and Y_2 give two normally distributed random variables X_1 and X_2 through the deterministic transformation in Eq. (6). In this sense no computation is wasted in producing samples that are ultimately rejected. Note however that this transformation requires the evaluation of log and cos which can be costly in terms of computer time, and even more efficient alternatives have been proposed in the literature.

Although apparently limited, this type of transformation can be very useful in practice to sample from simple distributions that are then fed into more complex algorithms. Most of the efficient algorithms to sample from gamma's, beta's etc. are a mixture of such deterministic transformations and the accept/rejection method.

3 MCMC Methods

3.1 Motivation

So far we have seen methods of sampling from relatively low dimensional distributions, which in fact collapse for even modest dimensions. For example consider the following -over-used- Bayesian example, the nuclear pump data example (Gaver and O'Muircheartaigh, 1987). This example describes multiple failures in a nuclear plant with the data, say y, given in the following table:

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Times	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

The modeling is based on the assumption that the failures of the $i-$th pump follow a Poisson process with parameter λ_i $(1 \leq i \leq 10)$. For an observed time t_i, the number of failures p_i is thus a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable. The unknowns here consist therefore of $\theta := (\lambda_1, \ldots, \lambda_{10}, \beta)$ and the aim here is to estimate quantities related to $p(\theta|y)$. For reasons invoked by the authors one chooses the following prior distributions,

$$\lambda_i \overset{iid}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$. Note that this introduces a *hierarchical* parameterization of the problem, as the hyperparameter β is considered unknown here. A prior distribution is therefore ascribed to this hyperparameter, therefore robustifying the inference. The posterior distribution is proportional to

$$\begin{aligned}
&\prod_{i=1}^{10} \{(\lambda_i t_i)^{p_i} \exp(-\lambda_i t_i) \lambda_i^{\alpha-1} \exp(-\beta\lambda_i)\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta) \\
&\propto \prod_{i=1}^{10} \{\lambda_i^{p_i+\alpha-1} \exp(-(t_i+\beta)\lambda_i)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta).
\end{aligned}$$

This multidimensional distribution is rather complex, and it is not obvious how the inverse cdf method, the rejection method or importance sampling could be used in this context. However one notices that the following conditionals have a familiar form,

$$\lambda_i|(\beta, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta) \text{ for } 1 \leq i \leq 10$$
$$\beta|(\lambda_1, \ldots, \lambda_{10}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i), \tag{7}$$

and instead of directly sampling the vector $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$ at once, one could suggest sampling it progressively and iteratively, starting for example with the λ_i's for a given guess of β, followed by an update of β given the new samples $\lambda_1, \ldots, \lambda_{10}$. More precisely, given a sample, at iteration t, $\theta^t := (\lambda_1^t, \ldots, \lambda_{10}^t, \beta^t)$ one could proceed as follows at iteration $t + 1$,

1. $\lambda_i^{t+1}|(\beta^t, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta^t)$ for $1 \leq i \leq 10$,
2. $\beta^{t+1}|(\lambda_1^{t+1}, \ldots, \lambda_{10}^{t+1}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i^{t+1})$.

This suggestion is of great interest: indeed instead of directly sampling in a space with 11 dimensions one samples in spaces of dimension 1, which can be achieved using either of the methods reviewed in previous sections. However the structure of the algorithm calls for many questions: by sampling from these conditional distributions are we sampling from the desired joint distribution? If yes, how many times should the iteration above be repeated? In fact the validity of the approach described here stems from the fact that the sequence $\{\theta^t\}$ defined above is a Markov chain and, as we shall see, some Markov chains have very nice properties.

3.2 Intuitive Approach to MCMC

Basic Concepts. Assume that we wish to sample from a distribution π. The idea of MCMC consists of running an ergodic Markov chain. In order to illustrate this intuitively, consider Figure 11. The target distribution corresponds to the continuous line. It is a normal distribution. We consider here 1000 Markov chains run in parallel, and independent. We assume that the initial distribution of these Markov chains is a uniform distribution on $[0, 20]$. We then apply a (specially designed) Markov transition probability to all of the 1000 samples, in an independent manner. Observe how the histograms of these samples evolve with the iterations. Obviously the normal distribution seems to "attract" the distribution of the samples and even to be a fixed point of the algorithm. This is is what we wanted to achieve, i.e. it seems that we have produced 1000 independent samples from the normal distribution. The numbers $1, 2, 3, 4$ and 5 correspond to the location of samples $1, 2, 3, 4$ and 5 along the iterations. In fact one can show that in many situations of interest it is not necessary to run N Markov chains in parallel in order to obtain 1000 samples, but that one can consider a unique Markov chain, and build the histogram from this single Markov chain by forming

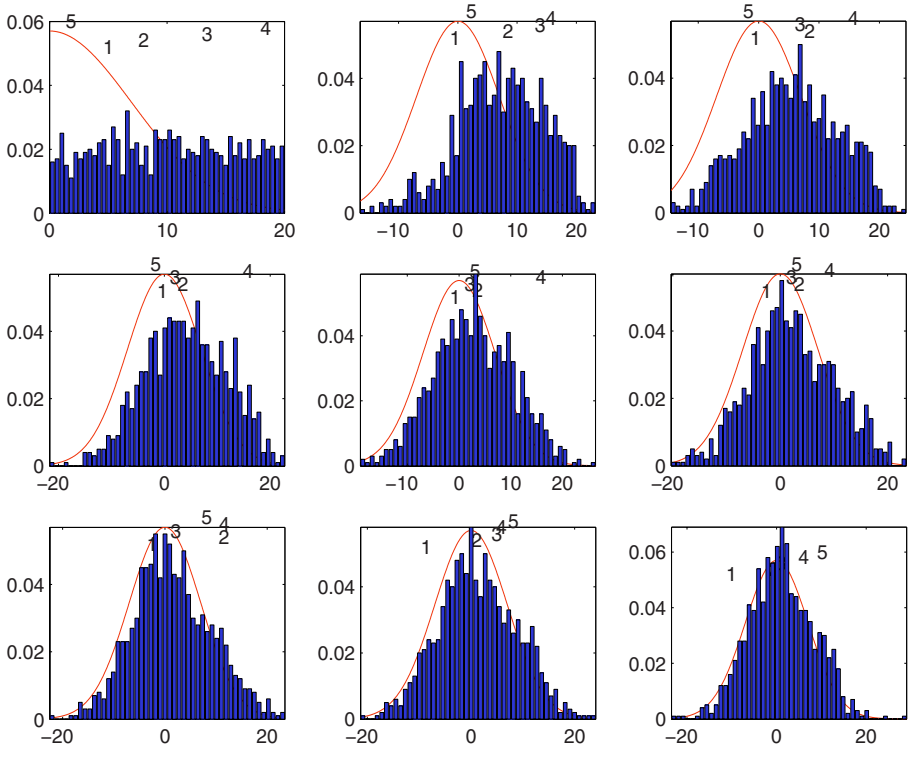

Fig. 11. From top left to bottom right: histograms of 1000 independent Markov chains with a normal distribution as target distribution

histograms from one trajectory. This idea is illustrated in Figure 12. The target distribution is here a mixture of normal distributions. Notice that the estimate of the target distribution, through the series of histograms, improves with the number of iterations. Assume that we have stored $\{X_i, 1 \leq i \leq N\}$ for N large and wish to estimate $\int_{\mathsf{X}} f(x)\pi(x)dx$. In the light of the numerical experiments above, one can suggest the estimator

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i),$$

which is exactly the estimator that we would use if $\{X_i, 1 \leq i \leq N\}$ were independent. In fact, it can be proved, under relatively mild conditions, that such an estimator is consistent *despite the fact that the samples are NOT independent!* Under additional conditions, a central limit theorem also holds for this estimator, and the rate of convergence is again $1/\sqrt{N}$. Note however that the constant involved in the CLT will be different from the constant in the independent case, as it will take into account the fact that the samples are not independent.

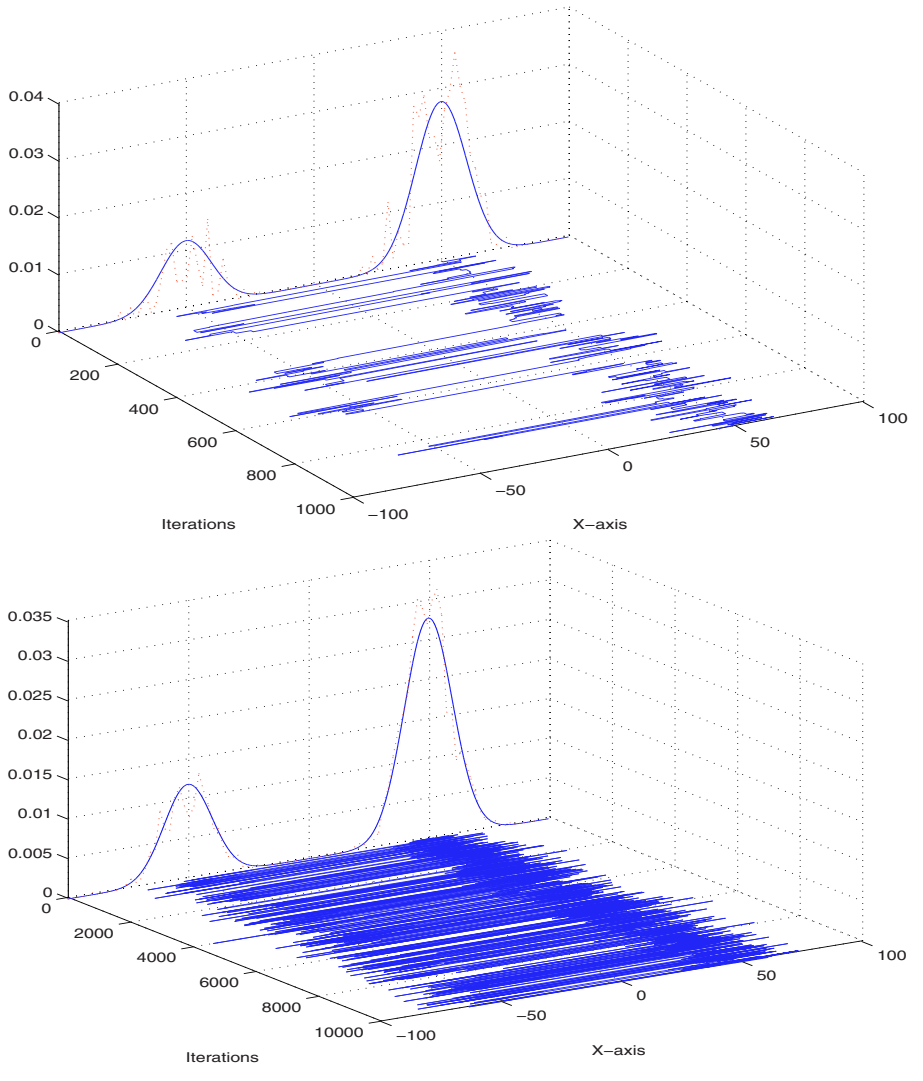

Fig. 12. Sampling from a mixture of normal distributions following the path of a single Markov chain. Full line: the target distribution - Dashed line: histogram of the path. Top: 1000 iterations only. Bottom: 10000 iterations

Unfortunately not all Markov chains, with transition probability say P, will have the following three important properties observed above:

1. The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, i.e.

$$\int_{\mathsf{X}} \pi(x)P(x,y) = \pi(y).$$

2. The successive distributions of the Markov chains are "attracted" by π, or converge towards π.
3. The estimator

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

is consistent, and converges towards $\mathbb{E}_\pi(f(X))$.

The first point is easily solved: the Metropolis-Hastings algorithm provides us with a *generic* mechanism of building Markov chains that admit a given distribution π as invariant distribution, whose density is known *only up to a normalizing constant*. Note that this later property is very convenient in a Bayesian framework! The reason for which the Metropolis-Hastings algorithm admits any desired distribution π as invariant distributions stems from the fact that it is *reversible* with respect to π, i.e. for any $x, y \in \mathsf{X}$,

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

and therefore automatically admits π as invariant distribution (indeed integrate the equality above with respect to x over X). In order to answer the second and third points one needs to introduce two notions: *irreducibility* and *aperiodicity*. The notion of reducibility (i.e. non-irreducibility) is illustrated in Figure 13: the Markov chain cannot reach a region of the space X where the distribution π has positive mass. Therefore irreducibility means that two arbitrarily chosen points in X with positive densities, can always communicate in a finite number of iterations. It is quit remarkable that under this simple condition, provided that π is an invariant distribution of the Markov chain and $\mathbb{E}_\pi(|f(x)|) < +\infty$, then $N^{-1}\sum_{i=1}^{N} f(x_i)$ is consistent (see [24]). In order to ensure that the series of distributions of the Markov chain converges it is furthermore necessary to ensure aperiodicity. To illustrate this, consider the following toy example. $\mathsf{X} = \{1, 2\}$ and $P(1,2) = 1$ and $P(2,1) = 1$. One easily checks that

$$\pi^{\mathsf{T}} P = \pi^{\mathsf{T}} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \pi^{\mathsf{T}},$$

admits the solution $\pi = (1/2, 1/2)^{\mathsf{T}}$, i.e. π is an invariant distribution of the Markov chain. Clearly this chain has a periodic behavior, with period 2, so that if at iteration $i = 0$ the chain always starts in 1, i.e. $\mu = (1, 0)^{\mathsf{T}}$, then the distributions of the Markov chain are

$$\begin{aligned} \mu^{\mathsf{T}} P^{2k} &= \mu^{\mathsf{T}} \\ \mu^{\mathsf{T}} P^{2k+1} &= (0,1)^{\mathsf{T}} \; k \geq 0, \end{aligned}$$

that is the distributions do not converge. On the other hand the proportions of time spent in state 1 and 2 converge to $1/2, 1/2$ and we expect $N^{-1}\sum_{i=1}^{N} f(X_i)$ to be consistent.

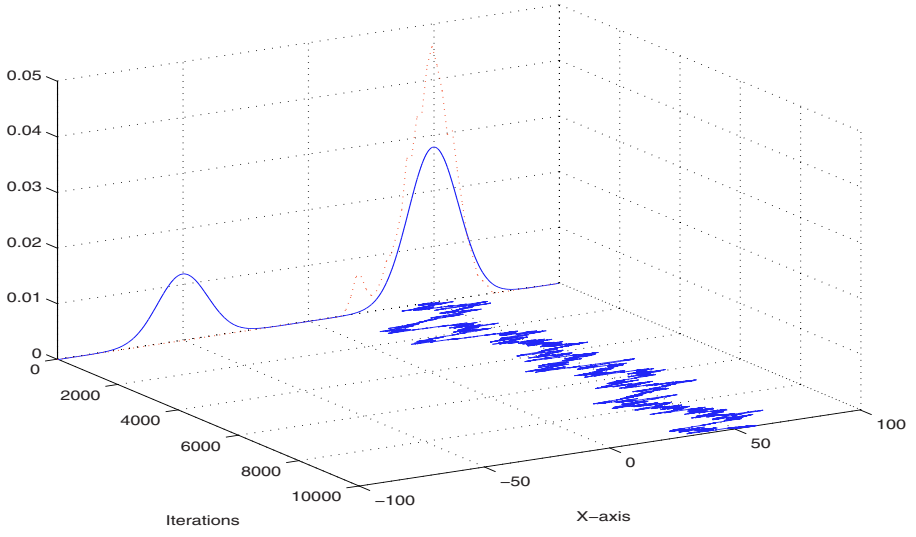

Fig. 13. In this case the Markov chain cannot explore the complete distribution: this is an illustration of reducibility (or in fact here quasi-reducibility)

The Gibbs Sampler. In the light of the appendix on Markov chains, one can ask if the following algorithm is likely to produce samples from the required posterior distribution,

$$\lambda_i|(\beta, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta) \text{ for } 1 \leq i \leq 10$$
$$\beta|(\lambda_1, \ldots, \lambda_{10}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i).$$

There are many ways of sampling from these unidimensional distribution (including rejection sampling, but there are even much more efficient ways). The idea of the Gibbs sampler consists of replacing a difficult global update of θ, with successive updates of the components of θ (or in fact in general groups of components of θ). Given the simple and familiar expressions of the conditional distributions above, one can suggest the following algorithm

1. $\lambda_i^{t+1}|(\beta^t, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta^t)$ for $1 \leq i \leq 10$,
2. $\beta^{t+1}|(\lambda_1^{t+1}, \ldots, \lambda_{10}^{t+1}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i^{t+1})$.

Maybe surprisingly, this algorithm produces samples from the posterior distribution $p(\theta|y)$, provided that the required distribution is invariant and the Markov chain irreducibility and aperiodicity are satisfied. We start with a result, in a simple case for simplicity. The generalization is trivial.

Proposition 1. *Let $p(a, b)$ be a probability density. Consider the Gibbs sampler which updates (a, b) using the conditional distributions $p(a|b)$ and $p(b|a)$. The*

Markov chain generated by this algorithm admits $p(a, b)$ as invariant distribution.

Proof. From the definition of invariance, we want to prove that for any a', b',

$$\int_{\mathsf{X}} p(a,b)p(a'|b)p(b'|a')dadb \stackrel{?}{=} p(a',b').$$

We start from the left hand side, and apply basic probability rules

$$\begin{aligned}\int_{\mathsf{X}} p(a,b)p(a'|b)p(b'|a')dadb &= \int_{\mathsf{X}} p(b)p(a'|b)p(b'|a')db \\ &= \int_{\mathsf{X}} p(a',b)p(b'|a')db \\ &= \int_{\mathsf{X}} p(b|a')p(a')p(b'|a')db \\ &= p(a',b') \times 1.\end{aligned}$$

Now, in order to ensure the convergence of estimators of the type $N^{-1}\sum_{i=1}^{N} f(X_i)$, it is sufficient to ensure irreducibility. This is not automatically verified for a Gibbs sampler, as illustrated in Figure 14 with a simple example. However in the nuclear pumps failure data, irreducibility is automatic: all the conditional distributions are strictly positive on the domain of definition of the parameters ($(0, +\infty)$ for each of them). One can therefore reach any set A from any starting point x with positive probability in one iteration of the Gibbs sampler.

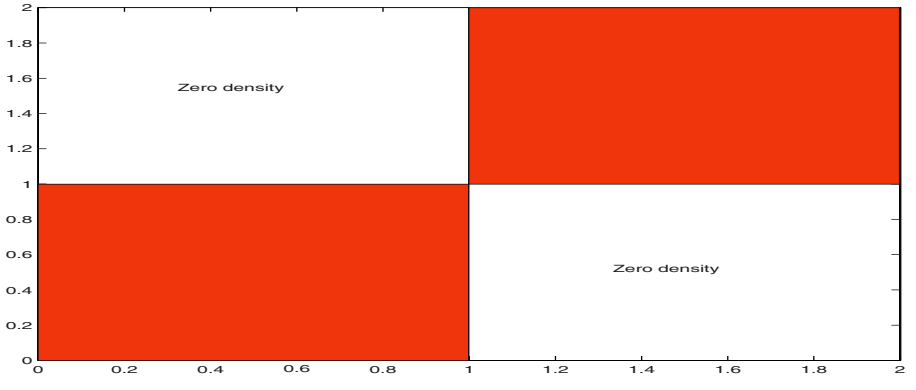

Fig. 14. A distribution that can lead to a reducible Gibbs sampler

It is relatively easy to prove aperiodicity as well, but we will not stress on this here, as we are in practice mostly interested in estimators of the type $N^{-1}\sum_{i=1}^{N} f(X_i)$.

Although natural, generally easy to implement, the Gibbs sampler does not come without problems. First it is clear that it requires one to be able to identify conditional distributions in the model, from which it is routine to sample. This is in fact rarely the case with realistic models. It is however generally the case when distributions from an exponential family are involved in the modeling. Another problem of the Gibbs sampler, is that its speed of convergence is directly influenced by the correlation properties of the target distribution π. Indeed, consider the toy two-dimensional example in Figure 15. This is a bidimensional normal distribution with strong correlation between x and y. A Gibbs sampler along the x and y axis will require many iterations to go from one point to another point that is far apart, and is somehow strongly constrained by the properties (both in terms of shape and algebraic properties) of π.

In contrast the Metropolis-Hastings algorithm which is presented in the next subsection possesses an extra degree of freedom, its proposal distribution which will determine how π is explored. This is illustrated in Figure 16, where for a good choice of the proposal distribution, the distribution π is better explored than in Figure 15, for the same number of iterations.

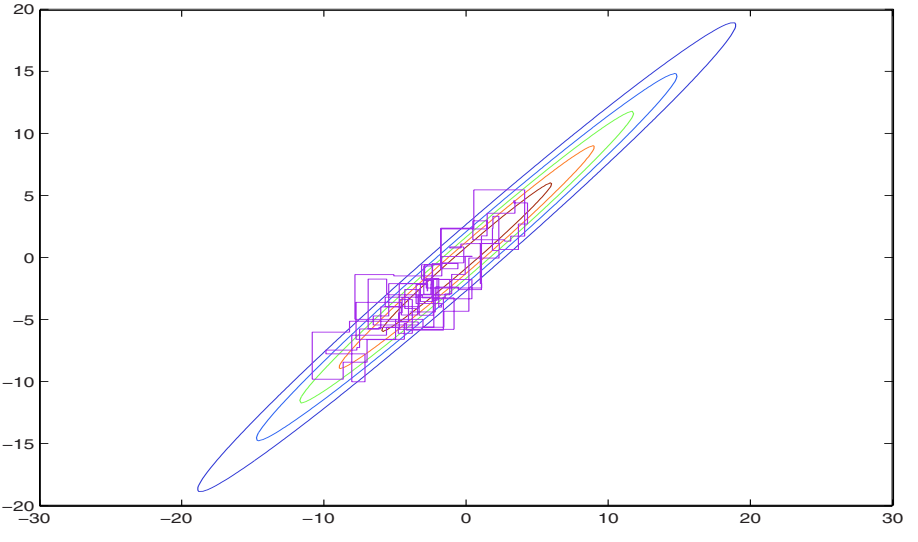

Fig. 15. A distribution for which the Gibbs sampler along the x and y axis might be very slow

The Metropolis-Hastings Algorithm. Let π be the density of a probability distribution on X and let $\{\theta \in \mathsf{X} : q(\theta, \cdot)\}$ be a family of probability densities from which it is possible to sample. The Metropolis-Hastings algorithm proceeds as follows.

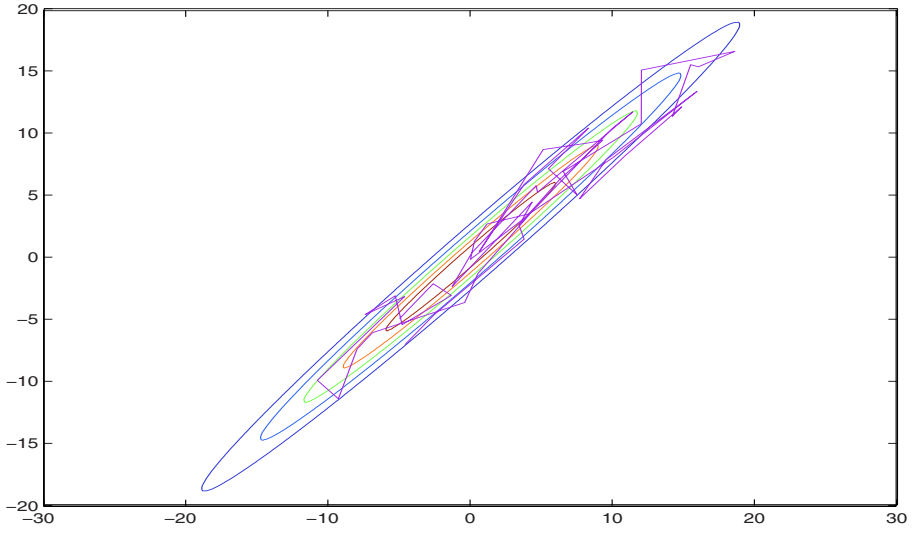

Fig. 16. A distribution for which the Gibbs sampler might be very slow, but here explored with an appropriate Metropolis-Hastings algorithm

---

Metropolis-Hastings Algorithm

1. Initialization, $i = 0$. Set randomly or deterministically θ_0.
2. Iteration i, $i \geq 1$.
   - Propose a candidate $\theta \sim q(\theta_{i-1}, \cdot)$.
   - Evaluate the acceptance probability

$$\alpha(\theta_{i-1}, \theta) = \min\left\{1, \frac{\pi(\theta)/q(\theta_{i-1}, \theta)}{\pi(\theta_{i-1})/q(\theta, \theta_{i-1})}\right\} \tag{8}$$

   - Then $\theta_i = \theta$ with probability $\alpha(\theta_{i-1}, \theta)$ otherwise $\theta_i = \theta_{i-1}$.

Example 6. Let us assume that we want to simulate a set of samples from $p(\theta|y)$. Using Bayes' theorem we have $p(\theta|y) \propto p(y|\theta)p(\theta)$. A MH procedure consists of simulating some candidates θ' according to $q(\theta,\theta')$, evaluating some quantities $\alpha(\theta,\theta') = \min\left\{1, \frac{p(y|\theta')p(\theta')q(\theta',\theta)}{p(y|\theta)p(\theta)q(\theta,\theta')}\right\}$, and accepting these candidates with probability $\alpha(\theta,\theta')$.

As pointed out earlier, q is to a certain extent an extra degree of freedom compared to the Gibbs sampler and an infinite number of possible choices for q is possible. We here briefly review two classical choices.

Random Walk: A simple choice consists of proposing as candidate a perturbation of the current state, i.e. $\theta' = \theta + z$ where z is a random increment of density $\varphi(z)$.

- This algorithm corresponds to the particular case $q(\theta, \theta') = \varphi(\theta' - \theta)$. We obtain the following acceptance probability:

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')\varphi(\theta - \theta')}{\pi(\theta)\varphi(\theta' - \theta)}\right\} \tag{9}$$

- If $q(\theta, \theta') = \varphi(\theta - \theta') = \varphi(\theta' - \theta)$ then we obtain

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\} \tag{10}$$

  This algorithm is called the Metropolis algorithm [15].

Independent Metropolis-Hastings: In this case, we select the candidate independently of the current state according to a distribution $\varphi(\theta')$. Thus $q(\theta, \theta') = \varphi(\theta')$ and we obtain the following acceptance probability:

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')\varphi(\theta)}{\pi(\theta)\varphi(\theta')}\right\} \tag{11}$$

In the case where $\pi(\theta)/\varphi(\theta)$ is bounded, i.e. we could also apply the accept/reject procedure, this procedure shows (fortunately) better asymptotic performance in terms of variance of ergodic averages.

Example 7. In a Bayesian framework, if we want to sample from $p(\theta|y) \propto p(y|\theta)p(\theta)$ then one can take $p(\theta)$ as candidate distribution. Then the acceptance reduces to

$$\alpha(\theta, \theta') = \min\left\{\frac{p(y|\theta')}{p(y|\theta)}, 1\right\} \tag{12}$$

There are many possible variations on this theme, see [24] and [2].

Metropolis-Hastings One-at-a-Time. It should not be surprising if the problems encountered with classical sampling techniques are also problems with the plain MH algorithm. In particular, when θ is high-dimensional, it typically becomes very difficult to select a good proposal distribution: either the acceptance probability is very low or very large and the chain does not explore π very rapidly, or the chain explores only one mode of the distribution. To solve this problem one can use the strategy adopted by the Gibbs sampler. Define a partition of $\theta := (\theta_1, \ldots, \theta_p)$. Then each component θ_k can be updated according to a MH update with proposal distribution, say q_k which admits the conditional distribution $\pi(\theta_k|\theta_{-k})$ (where $\theta_{-k} := (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_p)$) as invariant distribution.

MH One-at-a-Time

1. Initialization, $i = 0$. Set randomly or deterministically $\theta^{(0)} = \theta_0$.
2. Iteration i, $i \geq 1$.

– For $k = 1$ to p

– Sample $\theta_k^{(i)}$ according to a MH step with proposal distribution

$$q_k((\theta_{-k}^{(i)}, \theta_k^{(i-1)}), \theta_k) \tag{13}$$

and invariant distribution $\pi(\theta_k|\theta_{-k}^{(i)})$.
End For.

■

This algorithm includes the Gibbs sampler as a special case. Indeed, this latter corresponds to the particular case where the proposal distributions of the MH steps are equal to the full conditional distributions, *i.e.* $q_k((\theta_{-k}^{(i)}, \theta_k^{(i-1)}), \theta_k) = \pi(\theta_k|\theta_{-k}^{(i)})$, so that the acceptance probabilities are equal to 1 and no candidate is rejected.

Theoretical Aspects of the MH Algorithm. In this subsection we establish that the MH transition probability admits π as invariant distribution, and then briefly discuss the irreducibility and aperiodicity issues. The transition probability of the Metropolis-Hastings algorithm is for $x, A \in \mathsf{X}, \mathcal{B}(\mathsf{X})$

$$\begin{aligned} P(x, A) &= \int_A \alpha(x, y) q(x, y) dy + \mathbb{I}_A(x) \int_{\mathsf{X}} (1 - \alpha(x, y)) q(x, y) dy \\ &= \int_A \alpha(x, y) q(x, y) dy + \mathbb{I}_A(x)[1 - \int_{\mathsf{X}} \alpha(x, y) q(x, y) dy]. \end{aligned}$$

We now prove that P is reversible with respect to π. First notice that

$$\begin{aligned} \alpha(x, y)\pi(x)q(x, y) &= \min\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\}\pi(x)q(x, y) \\ &= \min\{\pi(x)q(x, y), \pi(y)q(y, x)\} \\ &= \pi(y)q(y, x) \min\{\frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}, 1\} \\ &= \pi(y)q(y, x)\alpha(y, x). \end{aligned}$$

Consequently for any $A, B \in \mathcal{B}(\mathsf{X})$,

$$\begin{aligned}\int_B \pi(x)P(x,A)dx &= \int_B \int_A \pi(x)\alpha(x,y)q(x,y)dxdy \\ &\quad + \int_B I_A(x)\pi(x)[1 - \int_{\mathsf{X}} \alpha(x,y)q(x,y)dy]dx \\ &= \int_A \int_B \pi(y)q(y,x)\alpha(y,x)dxdy \\ &\quad + \int_{\mathsf{X}} \mathbb{I}_{A\cap B}(x)\pi(x)[1 - \int_{\mathsf{X}} \alpha(x,y)q(x,y)dy]dx \\ &= \int_A \int_B \pi(y)q(y,x)\alpha(y,x)dxdy \\ &\quad + \int_A \mathbb{I}_B(x)\pi(x)[1 - \int_{\mathsf{X}} \alpha(x,y)q(x,y)dy]dx \\ &= \int_A \pi(y)P(y,B)dy.\end{aligned}$$

A simple condition which ensures the irreducibility and the aperiodicity of the MH algorithm is that $q\,(x,y)$ is continuous and strictly positive on the support of π for any x [20].

References

1. D. Gammerman, *Markov Chain Monte Carlo*, Chapman and Hall, London, 1997.
2. Monte Carlo Statistical Methods, Springer-Verlag, 2000.
3. **The MCMC preprint service provides papers in the MCMC field:** http://www.statslab.cam.ac.uk/˜mcmc/
4. J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, John Wiley & Sons, 1995.
5. S. Brooks, "Markov Chain Monte Carlo Method and its Application", *The Statistician*, vol. 47, 69-100, 1998.
6. S. Chib and E. Greenberg, "Markov Chain Monte Carlo Simulation Methods in Econometrics", *Econometric Theory*, 12, 409-431, 1996.
7. L. Devroye, *Non-Uniform Random Variate Generation*, Springer, New-York, 1986.
8. A.E. Gelfand and A.F.M. Smith, "Sampling-based Approaches to Calculating Marginal Densities", *J. Am. Statis. Assoc.*, vol. 85, no. 410, 398-409, 1990.
9. A. Gelman, G.O. Roberts and W.R. Gilks, "Efficient Metropolis Jumping Rules", in *Bayesian Statistics V*, Clarendon Press, Oxford, 599-608, 1996.
10. A. Gelman and D.B. Rubin, "Markov Chain Monte Carlo Methods in Biostatistics", *Stat. Meth. Med. Res.*, 339-355, 1996.
11. J. Geweke, "Bayesian Inference in Econometrics Models using Monte Carlo Integration", *Econometrica*, vol. 57, 1317-1339, 1989.
12. W.R. Gilks, S. Richardson and D.J. Spiegelhalter (editors), *Markov Chain Monte Carlo in Practice*, Chapman&Hall, 1996.
13. J.H. Halton, "A retrospective and prospective survey of the Monte Carlo method," SIAM Review, vol. 12, no. 1, January 1970.

14. W.K. Hastings, "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika* 57, 97-109, 1970.
15. N. Metropolis, N. Rosenblutt, A.W. Rosenblutt, M.N. Teller, A.H. Teller,"Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics,* 21, 1087-1092, 1953.
16. S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, 1993.
17. B.D. Ripley, *Stochastic Simulation*, Wiley, New York, 1987.
18. C.P. Robert, *The Bayesian Choice*, Springer-Verlag, 1996.
19. C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, 1999.
20. A.F.M. Smith and A.E. Gelfand, "Bayesian Statistics without Tears: a Sampling-Resampling Perspective", *American Statistician*, vol. 46, no. 2, 84-88, 1992.
21. A.F.M. Smith and G.O. Roberts, "Bayesian Computation via the Gibbs sampler and Related Markov Chain Monte Carlo Methods", *J. Roy. Stat. Soc. B*, vol. 55, 3-23, 1993.
22. A.N. Shiryaev, Probability, 2^{nd} edition, Springer.
23. M.A. Tanner, *Tools for statistical inference : methods for the exploration of posterior distributions and likelihood functions*, Springer-Verlag, New York, 1993.
24. L. Tierney, "Markov Chains for Exploring Posterior Distributions", *The Annals of Statistics*, vol. 22, 1701-1762, 1994.