Advanced Simulation Methods

Chapter 1 - Introduction to Monte Carlo methods

These notes constitute some supporting material for the Advanced Simulation course, Hilary Term 2015, University of Oxford. I have inherited them from Prof. Arnaud Doucet who was giving the course until 2013, modified them with Dr Rémi Bardenet in 2014 when we jointly gave the course, and modified them again on my own in 2015. I am entirely responsible for any typo and omission. Note that some of the material covered during the lectures might not be mentioned in these notes, and the material mentioned in these notes might not be covered during the lectures.

1 Approximation of integrals

In many scientific problems of interest including finance, operations research, statistical physics and statistics, it is required to numerically compute integrals, i.e.,

$$I = \int_{\mathbb{X}} f(x)\, dx$$

where $f : \mathbb{X} \to \mathbb{R}$. For simple choices of functions f and spaces $\mathbb{X}$, the integral can be computed exactly, but in general one has to resort to numerical approximations of I.

When $\mathbb{X} = [0, 1]$, then we can simply approximate I through

$$\widehat{I}_n = \sum_{i=0}^{n-1} \frac{1}{n} f\left(\frac{(i + 1/2)}{n}\right),$$

which is called the Riemann sum approximation. This corresponds to the approximation of the area under the curve $y = f(x)$ by the sum of the areas of the rectangles pictured in Figure 1. When f is differentiable and $M = \sup_{x\in[0,1]} |f'(x)| < \infty$ then the approximation error is $\mathcal{O}\left(n^{-1}\right)$. Indeed the error of the k-th rectangle, for $k \in \{0, ..., n-1\}$ is

$$\begin{aligned}\varepsilon_k &= \left| \int_{k/n}^{(k+1)/n} f(x)dx - \frac{1}{n} f\left(\frac{(k + 1/2)}{n}\right) \right| \\ &= \left| \int_{k/n}^{(k+1)/n} \left(f(x) - f\left(\frac{(k + 1/2)}{n}\right)\right) dx \right|.\end{aligned}$$

Now we use the fact that for all $x, y \in [a, b]$, there exists $c \in [a, b]$ such that

$$f(x) - f(y) = (x - y)f'(c).$$

Using the bound M on f', we obtain

$$\begin{aligned}\varepsilon_k &\leq \int_{k/n}^{(k+1)/n} \left| f(x) - f\left(\frac{(k + 1/2)}{n}\right)\right| dx \\ &\leq M \int_{k/n}^{(k+1)/n} \left| x - \frac{(k + 1/2)}{n}\right| dx \\ &\leq M \frac{1}{2n^2}.\end{aligned}$$

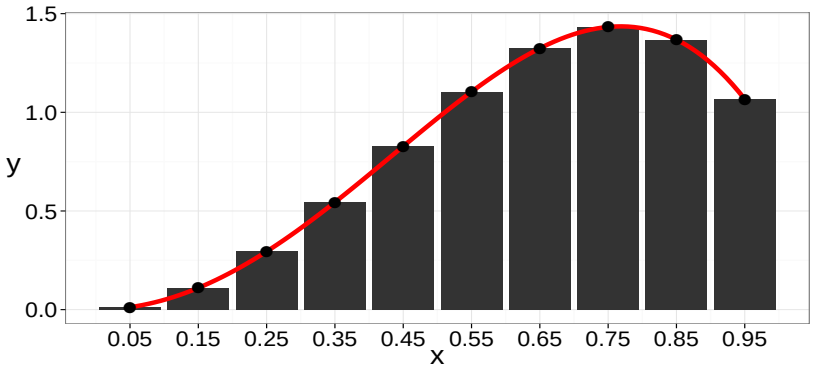

Figure 1: Numerical Integration of $f : x \mapsto 10(\cos(x)(1+x^2)-1)$, shown in red, by 10 rectangles shown in grey.

Summing these errors over the n rectangles yield a total error in $\mathcal{O}\left(n^{-1}\right)$.

However, for $\mathbb{X} = [0,1] \times [0,1]$ assuming

$$\widehat{I}_n = \frac{1}{n} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} f\left(\frac{(i+1/2)}{m}, \frac{(j+1/2)}{m}\right)$$

and $n = m^2$ then the approximation error is $\mathcal{O}\left(n^{-1/2}\right)$ and generally for $\mathbb{X} = [0,1]^d$ we have an approximation error in $\mathcal{O}\left(n^{-1/d}\right)$. This suggests that this type of deterministic approximations is inappropriate to compute high-dimensional integrals. Note that there are more sophisticated deterministic approximations, such as the trapezoidal rule or Simpson's rule, but they all suffer from the same degeneracy when the dimension increases.

The aim of this course is to introduce stochastic simulation methods, which are the most common tools used to perform numerical integration in high-dimensional scenarios. These methods, also known as Monte Carlo methods, were introduced in the 1940s and have become extremely popular in statistics over the past 20 years, as they allow to perform inference for complex statistical models. This course will be primarily focused on applications of Monte Carlo methods to Bayesian statistics, although the same methods are extensively used in other applications, as examplified below.

2 Examples of Applications

2.1 Volume of a Convex Body

Let $S \subset [0,1]^d$ be a convex body. In numerous applications, we are interested in computing the volume of this body which is simply given by

$$\text{vol}(S) = \int_{[0,1]^d} \mathbb{I}_S(x)\, dx$$

where $\mathbb{I}_S(x) = 1$ if $x \in S$ and 0 otherwise.

2.2 Statistical Mechanics

The Ising model is used to model the behavior of a magnet and is the best known/most researched model in statistical physics. The magnetism of a material is modelled by the collective contribution of dipole moments of many atomic spins.

Consider a simple 2D-Ising model on a finite lattice $\mathcal{G} = \{1,2,...,m\} \times \{1,2,...,m\}$ where each site $\sigma = (i,j)$ hosts a particle with a +1 or -1 spin modeled as a random variable X_σ. For physical reasons, the probability distribution of $X = \{X_\sigma\}_{\sigma \in \mathcal{G}}$ on $\mathbb{X} = \{-1,1\}^{m^2}$ is given by the so-called Gibbs distribution

$$\forall x \in \mathbb{X} \quad \pi_\beta(x) = \frac{\exp(-\beta U(x))}{Z_\beta}$$

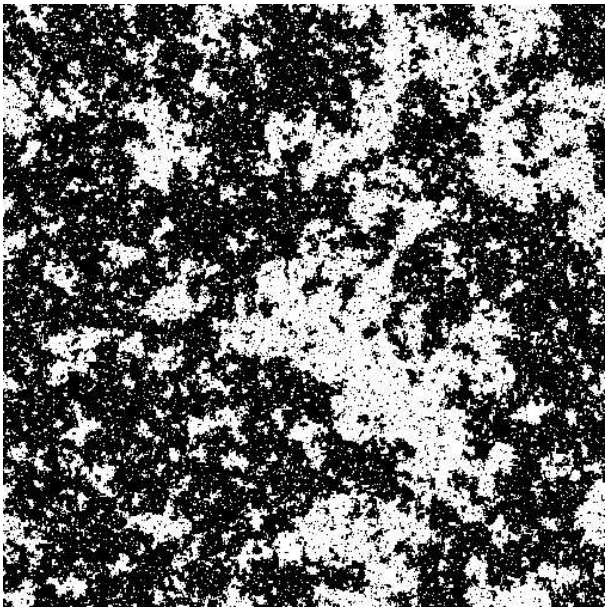

Figure 2: Sample from an two-dimensional Ising model.

where $\beta > 0$ is the inverse temperature and the potential energy U is

$$\forall x \in \mathbb{X} \quad U(x) = J \sum_{\sigma \sim \sigma'} x_\sigma x_{\sigma'},$$

for some $J \in \mathbb{R}$, where $\sigma \sim \sigma'$ refers to the set of pairs of sites that are "neighbors" in some pre-defined sense. For instance we can define that two sites $\sigma = (i, j)$ and $\sigma' = (i', j')$ are neighbors if and only if $|i - i'| \leq 1$ and $|j - j'| \leq 1$. According to this form of potential energy, if $x_\sigma = x_{\sigma'}$ and $\sigma \sim \sigma'$ then the probability $\pi_\beta(x)$ includes a term $\exp(-J)$, otherwise it includes a term $\exp(J)$. Hence the sign of J tells us whether there is a preference for equal or opposite spins at sites σ and σ'. The normalizing constant Z_β ensures that π_β is a probability distribution, that is, $\sum_{x \in \mathbb{X}} \pi_\beta(x) = 1$. Thus it is defined as

$$Z_\beta = \sum_{x \in \mathbb{X}} \exp(-\beta U(x)).$$

Physicists are often interested in computing $\mathbb{E}_{\pi_\beta}[U(X)]$ and Z_β. However, analytic results for the Ising model are very difficult to obtain and physicists often use simulation methods in order to perform these calculations. Note that the problem of computing sums is equivalent to the problem of computing integrals and is formally unified by measure theory.

2.3 Financial Mathematics

Let $S(t)$ denote the price of a stock at time t. We consider a call option granting the holder the right to buy the stock at a fixed price K at a fixed time T in the future; the current time being $t = 0$. This is a so-called European option. If at time T the stock price $S(T)$ exceeds the strike price K, the holder exercises the option for a profit of $S(T) - K$. If $S(T) \leq K$, the option expires worthless. The payoff to the holder at time T is thus

$$\max(0, S(T) - K)$$

and to get the present value of this payoff we need to multiply it by a discount factor $\exp(-rT)$ where r is a compounded interest rate. The expected present value is thus

$$\exp(-rT)\,\mathbb{E}[\max(0, S(T) - K)]$$

where the expectation is with respect to the distribution of the random variable $S(T)$.

If we knew explicitly the distribution of $S(T)$, then computing $\mathbb{E}[\max(0, S(T) - K)]$ would be a low-dimensional integration problem. However, this distribution is typically not available and we only have access to a stochastic model for $\{S(t)\}_{t\in\mathbb{N}}$ of the form

$$\begin{aligned} S(t+1) &= g(S(t), W(t+1)) \\ &= g(g(S(t-1), W(t)), W(t+1)) \\ &= g^2(S(t-1), W(t), W(t+1)) \\ &= g^n(S(0), W(1), ..., W(t+1)) \end{aligned}$$

where $\{W(t)\}_{t\in\mathbb{N}}$ is a sequence of i.i.d. random variables of probability density functions $\{p_W\}_{t\in\mathbb{N}}$ and g is a known nonlinear mapping. We can thus rewrite

$$\mathbb{E}[\max(0, S(T) - K)] = \int \max[0, g^n(s(0), w(1), ..., w(T)) - K] \left\{ \prod_{t=1}^{T} p_W(w(t)) \right\} dw(1) \cdots dw(T)$$

which is a high-dimensional integral whenever T is large.

3 Bayesian Statistics

In this course we will primarily use examples from Bayesian statistics, although numerical integration problems also arise in classical statistics as well as in other fields of science, as illustrated by the above examples.

In statistics the data is usually a collection of n values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$ in some space $\mathcal{Y}$, typically $\mathbb{R}^{d_y}$ for some d_y. A statistical model consider the data to be realisations of random variables $(Y_1, \ldots, Y_n)$ defined on the same space. Let us denote $Y_1, \ldots, Y_n$ by Y and $y_1, \ldots, y_n$ by y. The distribution of these random variables, which is specified by the model, has a density written $p_Y(y; \theta)$ with respect to some dominating measure, where θ is the parameter of the model living in some space Θ. The density of the observations, seen as a function of the parameter, is called the likelihood and denoted by $\mathcal{L}_n$:

$$\mathcal{L}_n : \theta \in \Theta \mapsto p_Y(y; \theta).$$

In the frequentist approach, θ is an unknown fixed value and inference is performed based on the likelihood function. The standard estimator is the maximum likelihood estimator $\hat{\theta}_n$, that is the parameter θ maximizing $\mathcal{L}_n(\theta)$ for the dataset $(y_1, \ldots, y_n)$. Note that because θ is not random, we write a semi-colon ";" in $p_Y(y; \theta)$ instead of a vertical bar "|" to emphasize that this is not a conditional distribution. On the contrary, in the Bayesian approach, the unknown parameter is regarded as a random variable ϑ and we assign a prior probability distribution to it, of density $p_\vartheta(\theta)$ (w.r.t. to a dominating measure denoted $d\theta$, say Lebesgue if $\Theta = \mathbb{R}^{d_\theta}$ for some d_θ). The distribution of Y given $\vartheta = \theta$ can now be interpreted as a proper conditional distribution and we thus denote it by $p_{Y|\vartheta}(y \mid \theta)$. Bayesian inference relies on the posterior density

$$p_{\vartheta|Y}(\theta|\, y) = \frac{p_{Y|\vartheta}(y \mid \theta)\, p_\vartheta(\theta)}{p_Y(y)} \tag{1}$$

obtained using Bayes formula, where

$$p_Y(y) = \int_\Theta p_{Y|\vartheta}(y \mid \theta)\, p_\vartheta(\theta)\, d\theta \tag{2}$$

is the so-called marginal likelihood or evidence.

Based on this posterior distribution, we can compute various point estimates such as the posterior mean of ϑ

$$\mathbb{E}(\vartheta|y) = \int_\Theta \theta\; p_{\vartheta|Y}(\theta|\, y)\, d\theta \tag{3}$$

or the posterior variance. We can also compute credible intervals, that is a interval C such that

$$\mathbb{P}(\vartheta \in C|\, y) = 1 - \alpha. \tag{4}$$

The posterior distribution can be used prediction of new observations. Assume that we want to predict the next observation y_{n+1} given that we already have $y = (y_1, \ldots, y_n)$. Then the predictive density of Y_{n+1} having observed $Y = y$ is

$$p_{Y_{n+1}|Y}\left(y_{n+1}|\, y\right) = \int_{\Theta} p_{Y_{n+1}|Y,\vartheta}\left(y_{n+1} \mid y, \theta\right) p_{\vartheta|Y}\left(\theta|\, y\right) d\theta. \tag{5}$$

The above predictive density takes into account the uncertainty about the parameter θ. By contrast, if we had first estimated the parameter, say by some $\hat{\theta}$, and then plugged the value into a predictive distribution of Y_{n+1} using $\hat{\theta}$, then we would not have taken parameter uncertainty into account.

Notation Remark: The above notation is precise but heavy. It is standard in the Bayesian literature not to use subscripts to index the densities of interest and to use a simpler notation; i.e. (1)-(2)-(3)-(5) will be written in most of the literature as

$$\begin{aligned} p\left(\theta|\, y\right) &= \frac{p\left(y|\, \theta\right) p\left(\theta\right)}{p\left(y\right)}, \\ p\left(y\right) &= \int_{\Theta} p\left(y|\, \theta\right) p\left(\theta\right) d\theta. \\ \mathbb{E}\left(\vartheta|y\right) &= \int_{\Theta} \theta\ p\left(\theta|\, y\right) d\theta, \\ p\left(y_{n+1}|\, y\right) &= \int_{\Theta} p\left(y_{n+1}|\, y, \theta\right) p\left(\theta|\, y\right) d\theta. \end{aligned}$$

This is imprecise as arguments of the densities should only be dummy variables whereas in this notation they define the densities we consider; i.e. $p\left(\theta\right)$ means $p_{\vartheta}\left(\theta\right)$ and $p\left(y\right)$ means $p_Y\left(y\right)$, $p\left(\theta|\, y\right)$ means $p_{\vartheta|Y}\left(\theta|\, y\right)$, etc. However this is standard and will be used here whenever it does not lead to any confusion. Note that another way to improve this imprecise notation consists in using different letters for the densities, i.e. $\mu\left(\theta\right) = p_{\vartheta}\left(\theta\right)$, $g\left(y|\, \theta\right) = p_{Y|\vartheta}\left(y \mid \theta\right)$, $p\left(\theta|\, y\right) = p_{\vartheta|Y}\left(\theta|\, y\right)$, etc.

1 ***(Gaussian data).*** *Let $Y = (Y_1, ..., Y_n)$ be i.i.d. random variables with $Y_i \sim \mathcal{N}\left(\theta, \sigma^2\right)$ with σ^2 known and θ unknown. To perform Bayesian inference, we assign a prior on θ by introducing the random variable $\vartheta \sim \mathcal{N}\left(\mu, \kappa^2\right)$, then one can check that*

$$p\left(\theta|\, y\right) = \mathcal{N}\left(\theta; \nu, \omega^2\right)$$

where

$$\omega^2 = \frac{\kappa^2 \sigma^2}{n\kappa^2 + \sigma^2}$$

and

$$\begin{aligned} \nu &= \frac{\omega^2}{\kappa^2}\mu + \frac{n\omega^2}{\sigma^2}\overline{y} \\ &= \frac{\sigma^2}{n\kappa^2 + \sigma^2}\mu + \frac{n\kappa^2}{n\kappa^2 + \sigma^2}\overline{y} \end{aligned}$$

so that directly $\mathbb{E}\left(\vartheta|y\right) = \nu$ and $\mathbb{V}\left(\vartheta|y\right) = \mathbb{E}\left(\vartheta^2|y\right) - \mathbb{E}\left(\vartheta|y\right)^2 = \omega^2$.

If we set $C = \left(\nu - \Phi^{-1}\left(1 - \alpha/2\right)\omega, \nu + \Phi^{-1}\left(1 - \alpha/2\right)\omega\right)$, where Φ^{-1} denotes the inverse of the cumulative distribution function of the Normal distribution, then $\mathbb{P}\left(\vartheta \in C|\, y\right) = 1 - \alpha$.

If we are interested in $p\left(y_{n+1}|\, y\right)$ where $Y_{n+1} \sim \mathcal{N}\left(\theta, \sigma^2\right)$ then

$$\begin{aligned} p\left(y_{n+1}|\, y\right) &= \int_{\Theta} p\left(y_{n+1}|\, \theta\right) p\left(\theta|\, y\right) d\theta \\ &= \mathcal{N}\left(y_{n+1}; \nu, \omega^2 + \sigma^2\right). \end{aligned}$$

In this simple example, all the calculations can be done analytically. This is because the Normal prior is "conjugate" with the Normal model with unknown mean and known variance, i.e. the posterior distribution is in the same family of distributions as the prior distribution (here, the family of Normal distributions). In general, the calculation of posterior quantities cannot be performed exactly. Indeed, one might want to use another prior distribution than the conjugate one, or the model might not admit any conjugate prior distribution.

2 ***(Logistic Regression).*** *Let* $(x_i, Y_i) \in \mathbb{R}^d \times \{0,1\}$ *where* $x_i \in \mathbb{R}^d$ *is a given covariate and we assume that the data are independent with*

$$\mathbb{P}\left(Y_i = y_i | \, \theta\right) = \frac{\exp\left(-y_i x_i^T \theta\right)}{1 + \exp\left(-x_i^T \theta\right)}.$$

To perform Bayesian inference, we assign a prior $p(\theta)$ *on* θ *and Bayesian inference relies on*

$$p\left(\theta | \, y_1, ..., y_n\right) = \frac{p(\theta) \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i | \, \theta\right)}{\mathbb{P}\left(y_1, ..., y_n\right)}$$

which is not a standard distribution if $p(\theta)$ *is chosen to be a Normal distribution. There exists a conjugate prior distribution for* θ *but it is not standard itself. The denominator* $\mathbb{P}\left(y_1, ..., y_n\right)$ *cannot be computed analytically.*

In general, statistical models and the associated prior probability distributions should be chosen to represent a phenomenon and its uncertainties, and thus should not be chosen on the grounds of purely computational reasons, such as "to make the calculations easier". Thus in many situations we will encounter posterior distributions such that we cannot analytically compute the integrals listed above, e.g. the posterior mean and so on. Going back to the problem of computing integrals, in statistics the integrals will often be written

$$I = \int_{\Theta} \varphi(\theta) \, \pi(\theta) \, d\theta,$$

where π is a probability density function, φ is a "test" function and Θ a sample space; for instance with $\varphi : \theta \mapsto \theta$ and $\pi(\theta) = p(\theta \mid y)$, the integral corresponds to the posterior mean. In the context of approximating I, the distribution π is often called the "target distribution". The integral can also be written

$$I = \mathbb{E}_{\pi}\left[\varphi(\vartheta)\right]$$

where ϑ follows the distribution π. Monte Carlo methods generally consist in replacing such expectations by empirical averages.

4 Basic Monte Carlo

The basic Monte Carlo method assumes that it is possible to obtain a collection of n independent draws from π, and that one can compute φ point-wise. We denote the draws from π by $\theta_1, \ldots, \theta_n$. Then the Monte Carlo estimator of I is defined as:

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(\theta_i).$$

The Monte Carlo estimator is unbiased, strongly consistent by the Law of Large Numbers (LLN), and satisfies the following Central Limit Theorem (CLT):

$$\sqrt{n}\left(\hat{I}_n - I\right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}_{\pi}\left[\varphi(\vartheta)\right]\right)$$

if $\mathbb{V}_{\pi}\left[\varphi(\vartheta)\right]$, the variance of $\varphi(\vartheta)$ when ϑ follows π, is finite. The asymptotic variance $\mathbb{V}_{\pi}\left[\varphi(\vartheta)\right]$ can be written

$$\mathbb{V}_{\pi}\left[\varphi(\vartheta)\right] = \int \left(\varphi(\theta) - I\right)^2 \pi(\theta) d\theta$$

and hence it can be itself approximated using the same sample $\theta_1, \ldots, \theta_n$ by

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\varphi(\theta_i) - \hat{I}_n\right)^2$$

which converges almost surely to $\mathbb{V}_{\pi}\left[\varphi(\vartheta)\right]$ by the LLN. Sometimes the "unbiased version" is preferred:

$$\tilde{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\varphi(\theta_i) - \hat{I}_n \right)^2,$$

which is unbiased in the sense that $\mathbb{E}\left[\tilde{\sigma}_n^2\right] = \mathbb{V}_\pi\left[\varphi(\vartheta)\right]$.

According to the rate $\sqrt{n}$ of the CLT, the variance of the estimator I_n is of order $\mathcal{O}(n^{-1})$, hence the standard deviation is of order $\mathcal{O}(n^{-1/2})$. This means that if one wants to divide the standard deviation by 10 (to obtain "10 times" more precision), one needs to sample 100 times more draws from π, which typically corresponds to 100 times more computational effort. This rate of convergence can seem to be very slow. Interestingly, the rate of convergence does not depend on the dimension d_θ of the sample space Θ; it is always $\sqrt{n}$.

Thus Monte Carlo methods converge slower than Riemann sums in one dimension, which error was shown to decrease in $\mathcal{O}(n^{-1})$; they are of the same accuracy as Riemann sums in two dimensions; and they are faster than Riemann sums for any dimension $d_\theta \geq 3$. Thus Monte Carlo methods have become standard tools to approximate integrals of moderate to high dimensions. In other words, Monte Carlo methods might seem slow, but they are still typically faster than alternative methods. Bakhvalov, Suldin and other mathematicians have proven results on the minimum error that can be obtained by algorithms using n pointwise evaluations of f to approximate $\int_{\mathbb{X}} f(x)dx$, and the rate $n^{-1/2}$ is found to be optimal when the dimension of $\mathbb{X}$ is large and/or the "smoothness" of f is low, in some sense. For instance the smoothness of f can be defined as the maximum integer k such that all k-th order partial derivatives of f are uniformly bounded on $\mathbb{X}$.

Note that the rate is $\sqrt{n}$ uniformly in d_θ, but high-dimensional integrals are still harder to approximate than low-dimensional integrals, as one would expect. Typically, the error associated with Monte Carlo methods is in $f(d_\theta)/\sqrt{n}$, where $f(d_\theta)$ is a polynomial in d_θ, or in the worst scenarios, exponential of d_θ. Thus the error might still be very large when d_θ is large, as one might not have enough computational power to scale n with $f(d_\theta)$.

In order to implement the above-described Monte Carlo method, one needs to obtain i.i.d. samples from π. The following lecture will describe ways to obtain i.i.d. samples from generic distributions π.