Advanced Simulation Methods

Chapter 1 - Introduction

1 Introduction

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 $f : \mathbb{X} \to \mathbb{R}$.

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

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

When f is differentiable and $\sup_{x \in [0,1]} |f'(x)| < M < \infty$ then the approximation error is $\mathcal{O}\left(n^{-1}\right)$; see Figure 1.

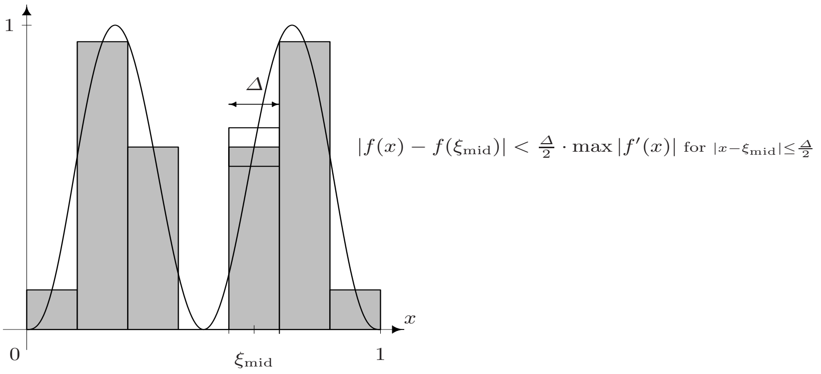
1
Δ
$|f(x) - f(\xi_{\text{mid}})| < \frac{\Delta}{2} \cdot \max |f'(x)|$ for $|x - \xi_{\text{mid}}| \leq \frac{\Delta}{2}$
x
0
ξ_{mid}
1

Figure 1: Numerical Integration by Riemman sums

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

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

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.

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 we will also discuss a few other applications, as examplified below.

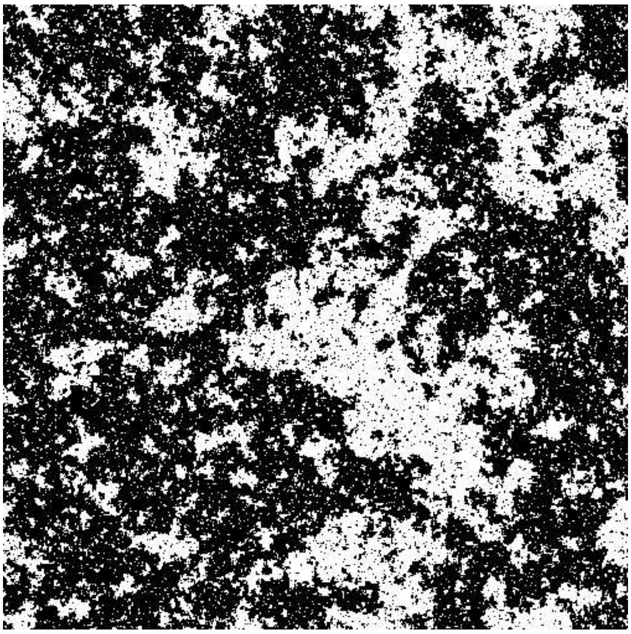

Figure 2: Sample from an Ising model

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

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

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

2.2 Statistical Mechanics

The Ising model serves 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 r.v. X_σ. For physical reasons, the probability distribution of $X = \{X_\sigma\}_{\sigma\in\mathcal{G}}$ on $\{-1, 1\}^{m^2}$ is given by the so-called Gibbs distribution

$$\pi_\beta\,(x) = \frac{\exp\left(-\beta U\,(x)\right)}{Z_\beta}$$

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

$$U\,(x) = J \sum_{\sigma\sim\sigma'} x_\sigma x_{\sigma'}$$

If $x_\sigma = x_{\sigma'}$ and $\sigma \sim \sigma'$ where '$\sim$' denotes a pre-defined neighbourhood structure then the probability $\pi\,(x)$ includes a term $\exp\,(-J)$ and $\exp\,(J)$ otherwise. Hence the sign of J tells us whether there is a preference for equal or opposite spins at sites σ and σ'.

Physicists are often interested in computing $\mathbb{E}_{\pi_\beta}\,[U\,(X)]$ and Z_β. However, analytical results for the Ising model are very difficult to obtain and physicists often use simulation methods in order to perform these calculations.

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 $\{S(t)\}_{t\in\mathbb{N}}$

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

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.

2.4 Bayesian Statistics

Let us consider a random variable Y taking values in a (measurable) space $\mathcal{Y}$. Given $\theta \in \Theta$, we assume that Y follows a probability density function $p_Y(y;\theta)$ (w.r.t. to a dominating measure, say Lebesgue if $\mathcal{Y} = \mathbb{R}^p$). Having observed $Y = y$, we are interested in performing inference about θ.

In the frequentist approach, θ is an unknown but fixed value and inference is performed based on the log-likelihood function $\ell(\theta) = \log p_Y(y;\theta)$. 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 $\Theta = \mathbb{R}^d$). Bayesian inference relies on the posterior density

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

where

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

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 any interval $I(y)$ such that

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

Another use of the posterior is for prediction of new observations. Assume that Z is independent of Y given $\vartheta = \theta$, but admits the same distribution $p_Y(z;\theta)$. Then the predictive density of Z having observed $Y = y$ is

$$p_{Z|Y}(z|\,y) = \int_\Theta p_Y(z;\theta)\,p_{\vartheta|Y}(\theta|\,y)\,d\theta \tag{5}$$

In contrast to a simple plug-in rule $p_Y\left(z;\widehat{\theta}\right)$ where $\widehat{\theta}$ is a point estimate of θ (e.g. the MLE), the above predictive density takes into account the uncertainty about the parameter θ.

Important Notational Remark: The above expressions are notationally 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

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

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 $p_{\vartheta|Y}\left(\theta|\,y\right)$ and $p\left(z|\,y\right)$ means $p_{Z|Y}\left(z|\,y\right)$. 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 of using different letter for the densities, i.e., $\mu\left(\theta\right) = p_{\vartheta}\left(\theta\right)$, $g\left(y|\,\theta\right) = p_Y\left(y;\theta\right)$, $p\left(\theta|\,y\right) = p_{\vartheta|Y}\left(\theta|\,y\right)$ and $f\left(z|\,y\right) = p_{Z|Y}\left(z|\,y\right)$.

Example 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, $\vartheta \sim \mathcal{N}\left(\mu, \kappa^2\right)$, 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 $\mathbb{V}\left(\vartheta|y\right) = \mathbb{E}\left(\vartheta^2|y\right) - \mathbb{E}^2\left(\vartheta|y\right) = \omega^2$.

If we set $I\left(y\right) = \left(\nu - \Phi^{-1}\left(1-\alpha/2\right)\omega, \nu + \Phi^{-1}\left(1-\alpha/2\right)\omega\right)$, then $\mathbb{P}\left(\vartheta \in I\left(y\right)|\,y\right) = 1-\alpha$.

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, we can do all the calculations analytically. However for general Bayesian models, this is not the case and numerical integration is necessary. In most cases, ϑ is an high dimensional parameter and so Monte Carlo methods are necessary.

Example 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 say $p\left(\theta\right)$ on ϑ and Bayesian inference relies on

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

which is not a standard distribution. The denominator cannot be computed analytically.

3 Basics of Monte Carlo Methods

Consider for the time being the following generic problem. We are interested in computing

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

where $\pi(x)$ is a probability density (w.r.t. to a dominating measure dx) on $\mathbb{X}$ and $\phi : \mathbb{X} \to \mathbb{R}$. The basic Monte Carlo method proceeds as follows.

Monte Carlo method

- Simulate independent $X_1, ..., X_n$ from π.
- Return $\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$.

It is trivial to check that $\widehat{I}_n$ is unbiased. More importantly, this estimate is consistent.

Proposition 1 ***(Strong Law of large numbers)****: Assume* $\mathbb{E}[|\phi(X)|] < \infty$ *then* $\widehat{I}_n$ *is a strongly consistent estimator of* I*.*

Proof. This follows from a direct application of the strong law of large numbers

$$\lim_{n\to\infty} \widehat{I}_n = \mathbb{E}[\phi(X_1)] = I \text{ almost surely}$$

Proposition 2 ***(Central Limit Theorem)****: Assume* I *and* $\sigma^2 = \mathbb{V}(\phi(X)) = \int_{\mathbb{X}} [\phi(x) - I]^2\, \pi(x)\, dx$ *are finite then*

$$\mathbb{E}\left(\left(\widehat{I}_n - I\right)^2\right) = \mathbb{V}\left(\widehat{I}_n\right) = \frac{\sigma^2}{n}$$

and

$$\frac{\sqrt{n}}{\sigma}\left(\widehat{I}_n - I\right) \overset{D}{\to} \mathcal{N}(0,1).$$

Moreover, if

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\phi(X_i) - \widehat{I}_n\right)^2$$

denotes the sample variance, then the probability that the interval

$$\widehat{I}_n \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}$$

contains I *converges to* $1 - \alpha$ *as* $n \to \infty$*.*

Proof. We have $\mathbb{E}\left(\left(\widehat{I}_n - I\right)^2\right) = \mathbb{V}\left(\widehat{I}_n\right)$ as $\mathbb{E}\left(\widehat{I}_n\right) = I$ and

$$\mathbb{V}\left(\widehat{I}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(\phi(X_i)) = \frac{\sigma^2}{n}.$$

Asymptotic normality follows from the standard central limit theorem.

Now introduce $E_n = \sqrt{n}\left|\widehat{I}_n - I\right|$ and let $z = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Then we have for any δ

$$\left|\mathbb{P}(E_n \leq zS_n) - \mathbb{P}\left(E_n \leq zS_n, \left|\frac{S_n}{\sigma} - 1\right| < \delta\right)\right| \leq \mathbb{P}\left(\left|\frac{S_n}{\sigma} - 1\right| \geq \delta\right)$$

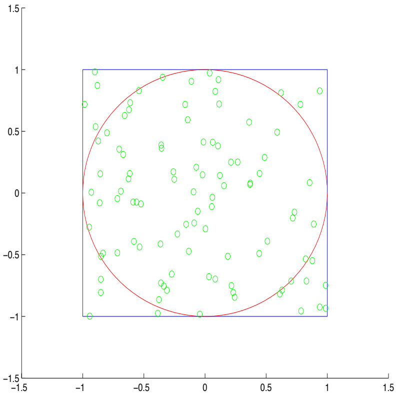

Figure 3: A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1 and Monte Carlo samples

where $\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{S_n}{\sigma} - 1\right| \geq \delta\right) = 0$ by the law of large numbers. Now $\left|\frac{S_n}{\sigma} - 1\right| < \delta \Leftrightarrow \sigma(1-\delta) < S_n < \sigma(1+\delta)$ so $E_n \leq zS_n \Rightarrow E_n \leq z\sigma(1+\delta)$ and $E_n \leq z\sigma(1-\delta) \Rightarrow E_n \leq zS_n$ so

$$2\Phi(z\sigma(1-\delta)) - 1 \leq \liminf \mathbb{P}(E_n \leq zS_n) \leq \limsup \mathbb{P}(E_n \leq zS_n) \leq 2\Phi(z\sigma(1+\delta)) - 1$$

As $\delta \to 0$, the two bounds converge to $1 - \alpha$.

Whatever being $\mathbb{X}$; e.g. $\mathbb{X} = \mathbb{R}$ or $\mathbb{X} = \mathbb{R}^{1000}$, the error is still in $\sigma/\sqrt{n}$. This is in contrast with deterministic methods where the rate of convergence of the approximation error towards zero is dimension dependent; e.g. $\mathcal{O}\left(n^{-1/d}\right)$ for Riemannian sums. It is sometimes said that Monte Carlo beats the curse of dimensionality but this is not quite true as σ^2 typically depends of $\dim(\mathbb{X})$. The main advantage of Monte Carlo methods lies in the fact that they are extremely flexible and are in some applications the only viable option.

We conclude this section by a simple toy example.

Example 3 ***(Computing*** *π). Consider the case where we have a square say $\mathcal{S} \subseteq \mathbb{R}^2$, the sides being of length 2, with inscribed disk $\mathcal{D}$ of radius 1; see Figure 3. We are interesting in computing through Monte Carlo the area I of $\mathcal{D}$. We have*

$$\begin{aligned} I = \pi &= \int\int_{\mathcal{D}} dx_1 dx_2 \\ &= \int\int_{\mathcal{S}} \mathbb{I}_{\mathcal{D}}(x_1, x_2)\, dx_1 dx_2 \ \textit{as}\ \mathcal{D} \subset \mathcal{S} \\ &= 4\int\int_{\mathbb{R}^2} \mathbb{I}_{\mathcal{D}}(x_1, x_2)\, \pi(x_1, x_2)\, dx_1 dx_2 \end{aligned}$$

where $\mathcal{S} := [-1, 1] \times [-1, 1]$ and

$$\pi(x_1, x_2) = \frac{1}{4}\mathbb{I}_{\mathcal{S}}(x_1, x_2)$$

is the uniform density on the square $\mathcal{S}$. In this case, we have

$$\widehat{I}_n = 4\frac{n_{\mathcal{D}}}{n}$$

where $n_{\mathcal{D}}$ is the number of samples which fell within the disk; see Figure 3.

Remark. Practically we are not interested in obtaining Monte Carlo estimates which admits a small variance but in estimates admitting a small relative variance which is given by

$$\mathbb{V}\left(\frac{\widehat{I}_n}{I}\right) = \frac{\mathbb{V}\left(\widehat{I}_n\right)}{I^2}.$$

Example 4 ***(Computing the probability of an event).*** *Assume you are interested in estimating $\pi(A) = \int \mathbb{I}_A(x)\,\pi(x)\,dx$ then $\mathbb{I}_A(X_i)$ where $X_i \sim \pi$ is a Bernoulli random variable of success probability $I = \pi(A)$ and variance $\sigma^2 = \pi(A)(1-\pi(A))$. Hence the relative variance of the estimate is*

$$\mathbb{V}\left(\frac{\widehat{I}_n}{I}\right) = \frac{(1-\pi(A))}{n\ \pi(A)}$$

For this relative variance to be small if $\pi(A) \ll 1$, we need $n \gg 1$; i.e. if you want the error to be with probability 0.95 at most $0.1\pi(A)$ then

$$1.96 \times \sqrt{\frac{\pi(A)(1-\pi(A))}{n}} \leq 0.1 \times \pi(A)$$

which is equivalent to

$$n \geq 385 \times \frac{(1-\pi(A))}{\pi(A)}.$$

We will discuss more sophisticated methods to address this problem later.

Monte Carlo methods require being able to sample from the distribution π. Whenever π is a standard distribution, e.g. normal or exponential, we will see that they are simple methods to achieve this. We will then discuss how Markov chain Monte Carlo and Sequential Monte Carlo methods can be used to sample approximately from any distribution π. All these simulations methods theoretically rely on the availability of a sequence of independent random variables $(U_i, i \geq 1)$ that are uniformly distributed on $[0,1]$; i.e. $U_i \sim \mathcal{U}_{[0,1]}$. Practically we do not have access to such a sequence and rely on pseudo-random numbers.

4 Pseudo-Random Numbers

A pseudo-random (deterministic) number generator is an algorithm is an algorithm that generates numbers which "look" like independent random variables. In R, the command `u←runif(100)` return 100 realizations of (pseudo-random) uniform r.v. in $[0,1]$.

The behaviour of modern random number generators (basic ones are constructed on number theory $N_{i+1} = (aN_i + c) \bmod m$ for suitable a, c, m and $U_{i+1} = N_{i+1}/(m+1)$) resembles random numbers in many respects. Standard tests for uniformity, independence, etc. do not show significant deviations. Any reasonable programming language provide the user with a large collection of powerful random number generators.

The point worth remembering though is that computer generated random numbers are not random at all but that hopefully they look random enough for that not to matter.