Advanced Simulation - Lecture 1

Patrick Rebeschini

January 15th, 2018

Administrivia

- www.stats.ox.ac.uk/∼rebeschi/teaching/1718/AdvSim
- Email: patrick.rebeschini@stats.ox.ac.uk
- **Lectures**: Mondays 9-10 & Wednesdays 9-10, weeks 1-8.
- Class tutors / Teaching Assistant:
  - Patrick Rebeschini / Sebastian Schmon (sebastian.schmon@magd.ox.ac.uk), Tuesdays 9:00-10:30, weeks 3, 5, 6, 8, LG.04.
  - Sebastian Schmon / Paul Vanetti (paul.vanetti@spc.ox.ac.uk), Tuesdays 10:30-12:00, weeks 3, 5, 6, 8, LG.04.
  - MSc: Patrick Rebeschini Tuesdays 11:00-12:00, weeks 3, 5, 6, 8, LG.02.
- Hand in of solutions by Friday 13:00 in the Adv. Simulation tray.

Objectives of the Course

- Many scientific problems involve intractable integrals.
- Monte Carlo methods are numerical methods to approximate high-dimensional integrals.
- Based on the simulation of random variables.
- Main application in this course: Bayesian statistics.
- Monte Carlo methods are increasingly used in econometrics, ecology, environmentrics, epidemiology, finance, signal processing, weather forecasting...
- More than $1,000,000$ results for "Monte Carlo" in Google Scholar, restricted to articles post 2000.

Computing Integrals

- For $f : \mathbb{X} \to \mathbb{R}$, let

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

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

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

- If $\sup_{x \in [0,1]} |f'(x)| < M < \infty$ then the approximation error is

$$\mathcal{O}\left(n^{-1}\right).$$

Riemann Sums

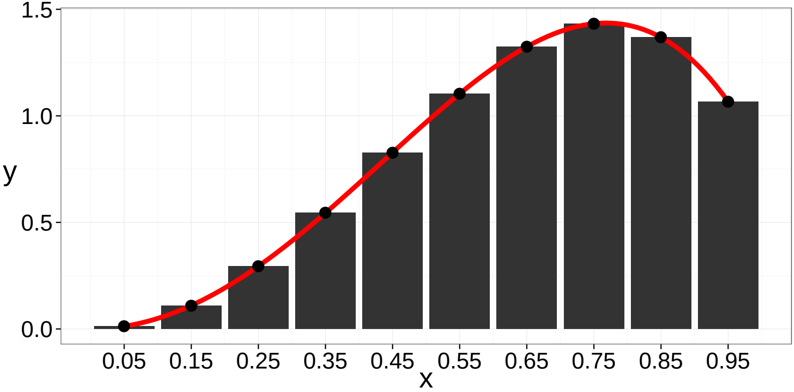

Figure: Riemann sum approximation (black rectangles) of the integral of f (red curve).

Computing High-Dimensional Integrals

- For $\mathbb{X} = [0,1] \times [0,1]$ assuming

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

  and $n = m^2$ then the approximation error is $\mathcal{O}\left(n^{-1/2}\right)$.

- Generally for $\mathbb{X} = [0,1]^d$ we have an approximation error in

$$\mathcal{O}\left(n^{-1/d}\right).$$

- So-called "curse of dimensionality".

- Simpson's rule also degrades as d increases.

Computing High-Dimensional Integrals

- For $f : \mathbb{X} \to \mathbb{R}$, write

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

where π is a probability density function on $\mathbb{X}$ and

$$\varphi : x \mapsto f(x)/\pi(x).$$

- Monte Carlo method:
  - sample n independent copies $X_1, \ldots, X_n$ of $X \sim \pi$,
  - compute

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

- Then $\hat{I}_n \to I$ almost surely and the approximation error is

$$\mathcal{O}(n^{-1/2})$$

whatever the dimension of the state space $\mathbb{X}$ (use CLT).

Computing High-Dimensional Integrals

- Non-asymptotically, we can prove this result using the mean-square error. We have:

$$
\begin{aligned}
(I - \widehat{I}_n)^2 &= I^2 - 2I\widehat{I}_n + \widehat{I}_n^2 \\
&= I^2 - \frac{2I}{n}\sum_{i=1}^{n}\varphi(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\varphi(X_i)^2 + \frac{1}{n^2}\sum_{i\neq j}\varphi(X_i)\varphi(X_j).
\end{aligned}
$$

As the samples are i.i.d. and $I = \mathbb{E}_\pi\left[\varphi(X)\right]$, we have

$$
\begin{aligned}
\mathbb{E}_\pi[(I - \widehat{I}_n)^2] &= I^2 - 2I^2 + \frac{1}{n}\mathbb{E}_\pi[\varphi(X_1)^2] + \frac{1}{n^2}n(n-1)I^2 \\
&= \frac{\mathbb{E}_\pi[\varphi(X_1)^2] - I^2}{n} = \frac{\mathbb{V}_\pi(\varphi(X_1))}{n}
\end{aligned}
$$

and $\sqrt{\mathbb{E}_\pi[(I - \widehat{I}_n)^2]} = \frac{\sqrt{\mathbb{V}_\pi(\varphi(X_1))}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ if $|\varphi(x)| \leq 1 \ \forall x$.

- The constant on the r.h.s. of the bound is 1, hence independent of the dimension of the state space $\mathbb{X}$.

Computing High-Dimensional Integrals

- In many cases the integrals of interest will directly be expressed as

$$I = \int_{\mathbb{X}} \varphi(x)\pi(x)dx = \mathbb{E}_\pi \left[\varphi(X)\right],$$

  for a specific function φ and distribution π.
- The distribution π is often called the “target distribution”.
- Monte Carlo approach relies on independent copies of

$$X \sim \pi.$$

- Hence the following relationship between integrals and sampling:

$$\begin{aligned} &\text{Monte Carlo method to approximate } \mathbb{E}_\pi \left[\varphi(X)\right] \\ &\Leftrightarrow \text{simulation method to sample } \pi \end{aligned}$$

- Thus Monte Carlo sometimes refer to simulation methods.

Ising Model

- 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.
- The distribution of $X = \{X_\sigma\}_{\sigma \in \mathcal{G}}$ on $\{-1, 1\}^{m^2}$ is given by

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

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

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

- Physicists are interested in computing $\mathbb{E}_{\pi_\beta}[U(X)]$ and Z_β.
- The dimension is m^2, where m can easily be 10^3.

Ising Model

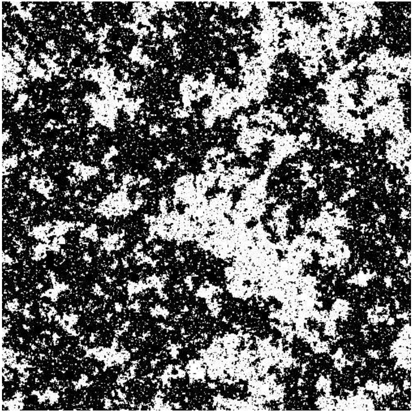

Figure: One draw from the Ising model on a 500×500 lattice.

Option Pricing

- Let $S(t)$ denote the price of a stock at time t.
- European option: grants 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 $t = 0$.
- If at time T the price $S(T)$ exceeds the strike 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

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

- To get the expected value at $t = 0$, we need to multiply it by a discount factor $\exp(-rT)$ where r is a compounded interest rate:

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

Option Pricing

- If we knew explicitly the distribution of $S(T)$ then $\mathbb{E}[\max(0, S(T) - K)]$ is a low-dimensional integral.

- **Problem**: We only have access to a complex stochastic model for $\{S(t)\}_{t\in\mathbb{N}}$

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

where $\{W(t)\}_{t\in\mathbb{N}}$ is a sequence of random variables and g is a known function.

Option Pricing

- The price of the option involves an integral over the T latent variables
$$\{W(t)\}_{t=1}^{T}.$$

- Assume these are independent with probability density function p_W.

- We can write

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

Bayesian Inference

- Given $\theta \in \Theta$, we assume that Y follows a probability density function $p_Y(y; \theta)$.

- Having observed $Y = y$, we want to perform inference about θ.

- In the frequentist approach θ is unknown but fixed; inference in this context can be performed based on

$$\ell(\theta) = \log p_Y(y; \theta).$$

- In the Bayesian approach, the unknown parameter is regarded as a random variable ϑ and assigned a prior $p_\vartheta(\theta)$.

Frequentist vs Bayesian

- Probabilities refer to limiting relative frequencies. They are (supposed to be) objective properties of the real world.
- Parameter are fixed unknown constants. Because they are not random, we cannot make any probability statements about parameters.
- Statistical procedures should have well-defined long-run properties. For example, a 95% confidence interval should include the true value of the parameter with limiting frequency at least 95%.

Frequentist vs Bayesian

- Probability describes degrees of subjective belief, not limiting frequency. Thus we can make probability statements about things other than data that can recur from some source; e.g. the probability that there will be an earthquake in Tokyo on September 27th, 2018.

- We can make probability statements about parameters, e.g.

$$\mathbb{P}\left(\theta \in [-1, 1] \mid Y = y\right)$$

- We make inference about a parameter by producing a probability distribution for it. Point estimates and interval estimates may then be extracted from this distribution.

Bayesian Inference

- Bayesian inference relies on the *posterior*

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

where

$$p_Y\left(y\right) = \int_\Theta p_Y\left(y;\theta\right) p_\vartheta\left(\theta\right) d\theta$$

is the so-called *marginal likelihood* or *evidence*.

- Point estimates such as posterior mean of ϑ

$$\mathbb{E}\left(\vartheta|y\right) = \int_\Theta \theta p_{\vartheta|Y}\left(\theta|\, y\right) d\theta$$

can be computed.

Bayesian Inference

- Credible intervals: any interval C such that

$$\mathbb{P}\left(\vartheta \in C \middle| y\right) = 1 - \alpha.$$

- Assume the observations are independent given $\vartheta = \theta$ then the predictive density of a new observation Y_{new} having observed $Y = y$ is

$$p_{Y_{new}|Y}\left(y_{new} \middle| y\right) = \int_{\Theta} p_Y\left(y_{new}; \theta\right) p_{\vartheta|Y}\left(\theta \middle| y\right) d\theta$$

- In contrast to a simple plug-in rule $p_Y\left(y_{new}; \widehat{\theta}\right)$ 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.