Simulation - Lecture 1 - Introduction and Monte Carlo

Lecture version: Monday 20th January, 2020, 11:21

Robert Davies - (adapted from slides from Julien Berestycki and others)

Part A Simulation and Statistical Programming

Hilary Term 2020

Simulation and Statistical Programming

- **Lectures on Simulation** (Prof. R. Davies): Tuesdays 2-3pm Weeks 1-8. LG.02, the IT suite
- **Computer Lab on Statistical Programming** (Prof. R. Davies): Friday 9-11am Weeks 3-8 LG.02, the IT suite
- **Departmental problem classes**: Weeks 3, 5, 7. Wednesday 9am, 4-5am, Thursday 10-11am, 11am-12pm. Various locations
- Hand in problem sheet solutions by Monday noon of same week for all classes
- Webpage: http://www.stats.ox.ac.uk/~rdavies/teaching/PartASSP/2020/index.htm
- This course builds upon the notes and slides of Julien Berestycki, Geoff Nicholls, Arnaud Doucet, Yee Whye Teh and Matti Vihola.

Outline

Introduction

Monte Carlo integration

Monte Carlo Simulation Methods

- Computational tools for the simulation of random variables and the approximation of integrals/expectations.
- These simulation methods, aka Monte Carlo methods, are used in many fields including statistical physics, computational chemistry, statistical inference, genetics, finance etc.
- The Metropolis algorithm was named the top algorithm of the 20th century by a committee of mathematicians, computer scientists & physicists.
- With the dramatic increase of computational power, Monte Carlo methods are increasingly used.

Objectives of the Course

- Introduce the main tools for the simulation of random variables and the approximation of multidimensional integrals:
  - Integration by Monte Carlo,
  - inversion method,
  - transformation method,
  - rejection sampling,
  - importance sampling,
  - Markov chain Monte Carlo including Metropolis-Hastings.
- Understand the theoretical foundations and convergence properties of these methods.
- Learn to derive and implement specific algorithms for given random variables.

Computing Expectations

- Let X be either
  - a discrete random variable (r.v.) taking values in a countable or finite set Ω, with p.m.f. f_X
  - or a continuous r.v. taking values in $\Omega = \mathbb{R}^d$, with p.d.f. f_X
- Assume you are interested in computing

$$\begin{aligned}\theta &= \mathbb{E}\left(\phi(X)\right) \\ &= \begin{cases} \sum_{x \in \Omega} \phi(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_{\Omega} \phi(x) f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}\end{aligned}$$

  where $\phi : \Omega \to \mathbb{R}$.
- It is impossible to compute θ exactly in most realistic applications.
- Even if it is possible (for Ω finite) the number of elements may be so huge that it is practically impossible
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(\sum_{k=1}^{d} x_k^2 \geq \alpha\right).$
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(x_1 < 0, ..., x_d < 0\right).$

Example: Queuing Systems

- Customers arrive at a shop and queue to be served. Their requests require varying amount of time.
- The manager cares about customer satisfaction and not excessively exceeding the 9am-5pm working day of his employees.
- Mathematically we could set up stochastic models for the arrival process of customers and for the service time based on past experience.
- **Question**: If the shop assistants continue to deal with all customers in the shop at 5pm, what is the probability that they will have served all the customers by 5.30pm?
- If we call $X \in \mathbb{N}$ the number of customers in the shop at 5.30pm then the probability of interest is

$$\mathbb{P}\left(X = 0\right) = \mathbb{E}\left(\mathbb{I}(X = 0)\right).$$

- For realistic models, we typically do not know analytically the distribution of X.

Example: Particle in a Random Medium

- A particle $(X_t)_{t=1,2,\ldots}$ evolves according to a stochastic model on $\Omega = \mathbb{R}^d$.
- At each time step t, it is absorbed with probability $1 - G(X_t)$ where $G : \Omega \to [0, 1]$.
- **Question**: What is the probability that the particle has not yet been absorbed at time T?
- The probability of interest is

$$\mathbb{P}\left(\text{not absorbed at time } T\right) = \mathbb{E}\left[G(X_1)G(X_2)\cdots G(X_T)\right].$$

- For realistic models, we cannot compute this probability.

Example: Ising Model

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

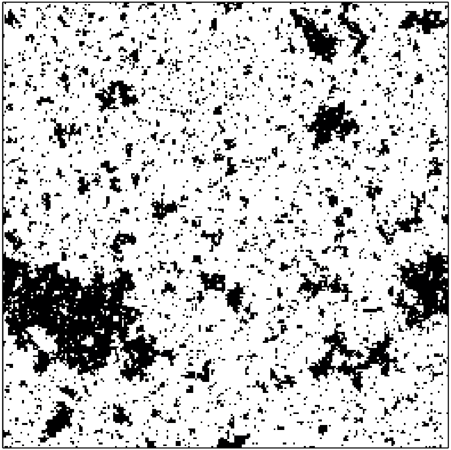
$$\pi(x) = \frac{\exp(-\beta U(x))}{Z_\beta}$$

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 $\mathbb{E}\left[U(X)\right]$ and Z_β.

Example: Ising Model



Sample from an Ising model for $m = 250$.

Example: Statistical Genetics

- At variable sites in the genome in a population, we can represent represent one chromosome as a haplotype as a vector of binary 0/1s. We humans are diploid so have two copies of each chromosome
- We often acquire data as "reads", observing those 0/1s along the genome
- We may be interested in trying to determine the haplotypes of an individual given some set of observed sequencing reads where we observe some of the underlying haplotypes, from one of an individuals two haplotypes.
- Let $L_r \in \{1, 2\}$ represent whether a read came from the maternal or paternal haplotype
- Then we might be interested in $P(H_i, H_j|O) \propto P(O|H_i, H_j) = \sum_{L_1, L_2, ...} P(O|H_i, H_j, L_1, L_2, ...)P(L_1, L_2, ...)$
- Naively, for M sequencing reads, this has computational cost 2^M, which is unfeasible for realistic M
- Monte Carlo methods allow us to estimate $P(H_i, H_j|O)$ and similar calculations, and are used frequently in genetics

Bayesian Inference

- Suppose (X, Y) are both continuous r.v. with a joint density $f_{X,Y}(x, y)$.
- Think of Y as data, and X as unknown parameters of interest
- We have

$$f_{X,Y}(x, y) = f_X(x)\ f_{Y|X}(y|x)$$

  where, in many statistics problems, $f_X(x)$ can be thought of as a prior and $f_{Y|X}(y|x)$ as a likelihood function for a given $Y = y$.
- Using Bayes' rule, we have

$$f_{X|Y}(x|y) = \frac{f_X(x)\ f_{Y|X}(y|x)}{f_Y(y)}.$$

- For most problems of interest, $f_{X|Y}(x|y)$ does not admit an analytic expression and we cannot compute

$$\mathbb{E}\left(\phi(X)|Y = y\right) = \int \phi(x) f_{X|Y}(x|y) dx.$$

Outline

Introduction

Monte Carlo integration

Monte Carlo Integration

Definition (Monte Carlo method)

Let X be either a discrete r.v. taking values in a countable or finite set Ω, with p.m.f. f_X, or a continuous r.v. taking values in $\Omega = \mathbb{R}^d$, with p.d.f. f_X. Consider

$$\theta = \mathbb{E}\left(\phi(X)\right) = \begin{cases} \sum_{x \in \Omega} \phi(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_{\Omega} \phi(x) f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

where $\phi : \Omega \to \mathbb{R}$. Let $X_1, ..., X_n$ be i.i.d. r.v. with p.d.f. (or p.m.f.) f_X. Then

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i),$$

is called the Monte Carlo estimator of the expectation θ.

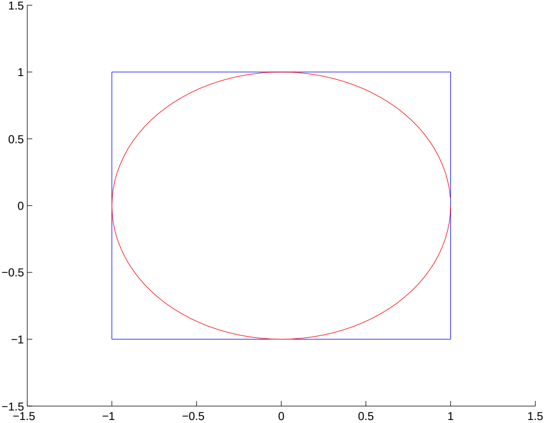
- Monte Carlo methods can be thought of as a stochastic way to approximate integrals.

Monte Carlo Integration

```
Algorithm 1 Monte Carlo Algorithm
  ▶ Simulate independent $X_1, ..., X_n$ with p.m.f. or p.d.f. $f_X$
  ▶ Return $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$.
```

Computing Pi with Monte Carlo Methods

- Consider the 2×2 square, say $\mathcal{S} \subseteq \mathbb{R}^2$ with inscribed disk $\mathcal{D}$ of radius 1.



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

Computing Pi with Monte Carlo Methods

- We have

$$\frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} = \frac{\pi}{4}.$$

- How could you estimate this quantity through simulation?

$$\begin{aligned} \frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} &= \int\int_{\mathcal{S}} \mathbb{I}\left((x_1, x_2) \in \mathcal{D}\right)\ \frac{1}{4} dx_1 dx_2 \\ &= \mathbb{E}\left[\phi(X_1, X_2)\right] = \theta \end{aligned}$$

  where the expectation is w.r.t. the uniform distribution on $\mathcal{S}$ and

$$\phi(X_1, X_2) = \mathbb{I}\left((X_1, X_2) \in \mathcal{D}\right).$$

- To sample uniformly on $\mathcal{S} = (-1, 1) \times (-1, 1)$ then simply use

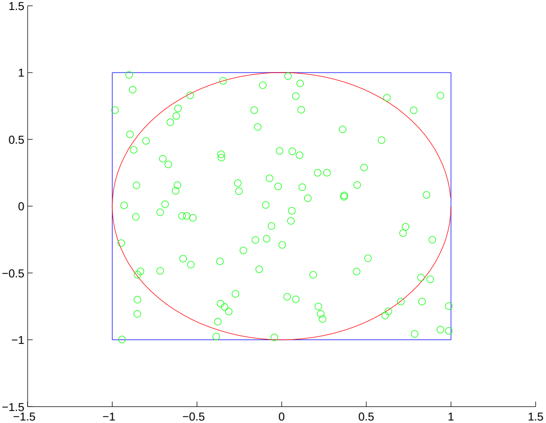
$$X_1 = 2U_1 - 1,\ X_2 = 2U_2 - 1$$

  where $U_1, U_2 \sim \mathcal{U}(0, 1)$.

Computing Pi with Monte Carlo Methods

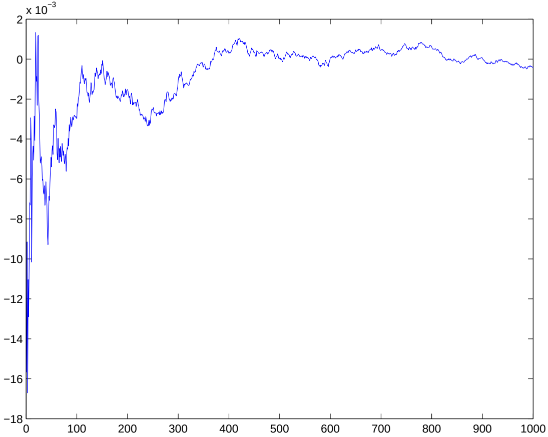
```
n <- 1000
x <- array(0, c(2,1000))
t <- array(0, c(1,1000))

for (i in 1:1000) {
  # generate point in square
  x[1,i] <- 2*runif(1)-1
  x[2,i] <- 2*runif(1)-1

  # compute phi(x); test whether in disk
  if (x[1,i]*x[1,i] + x[2,i]*x[2,i] <= 1) {
    t[i] <- 1
  } else {
    t[i] <- 0
  }
}
print(sum(t)/n*4)
```

Computing Pi with Monte Carlo Methods



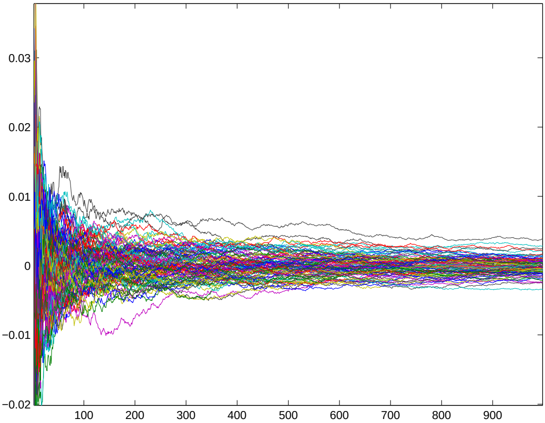
A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1 and Monte Carlo samples.

Computing Pi with Monte Carlo Methods



$\hat{\theta}_n - \theta$ as a function of the number of samples n.

Computing Pi with Monte Carlo Methods



$\hat{\theta}_n - \theta$ as a function of the number of samples n, 100 independent realizations.

Applications

- *Toy example*: simulate a large number n of independent r.v. $X_i \sim \mathcal{N}(\mu, \Sigma)$ and

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\sum_{k=1}^{d} X_{k,i}^2 \geq \alpha \right).$$

- *Queuing*: simulate a large number n of days using your stochastic models for the arrival process of customers and for the service time and compute

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i = 0 \right)$$

  where X_i is the number of customers in the shop at 5.30pm for ith sample.

- *Particle in Random Medium*: simulate a large number n of particle paths $(X_{1,i}, X_{2,i}, ..., X_{T,i})$ where $i = 1, ..., n$ and compute

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} G(X_{1,i}) G(X_{2,i}) \cdots G(X_{T,i})$$

Monte Carlo Integration: Properties

- **Proposition**: Assume $\theta = \mathbb{E}\left(\phi(X)\right)$ exists. Then the Monte Carlo estimator $\hat{\theta}_n$ has the following properties
  - Unbiasedness
    $$\mathbb{E}\left(\hat{\theta}_n\right) = \theta$$
  - Strong consistency
    $$\hat{\theta}_n \to \theta \text{ almost surely as } n \to \infty$$
- *Proof*: We have
  $$\mathbb{E}\left(\hat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\phi(X_i)\right) = \theta.$$

  Strong consistency is a consequence of the strong law of large numbers applied to $Y_i = \phi(X_i)$ which is applicable as $\theta = \mathbb{E}\left(\phi(X)\right)$ is assumed to exist.

Monte Carlo Integration: Central Limit Theorem

- **Proposition**: Assume $\theta = \mathbb{E}\left(\phi(X)\right)$ and $\sigma^2 = \mathbb{V}\left(\phi(X)\right)$ exist then

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

  and

  $$\frac{\sqrt{n}}{\sigma}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathsf{d}} \mathcal{N}(0,1).$$

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

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

  The CLT applied to $Y_i = \phi(X_i)$ tells us that

  $$\frac{Y_1 + \cdots + Y_n - n\theta}{\sigma\sqrt{n}} \xrightarrow{\mathsf{d}} \mathcal{N}(0,1)$$

  so the result follows as $\hat{\theta}_n = \frac{1}{n}\left(Y_1 + \cdots + Y_n\right)$.

Monte Carlo Integration: Variance Estimation

- **Proposition**: Assume $\sigma^2 = \mathbb{V}\left(\phi(X)\right)$ exists then

$$S^2_{\phi(X)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi(X_i) - \hat{\theta}_n\right)^2$$

is an unbiased sample variance estimator of σ^2.

- Proof. Let $Y_i = \phi(X_i)$ then we have

$$\begin{aligned}
\mathbb{E}\left(S^2_{\phi(X)}\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left(Y_i - \overline{Y}\right)^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} Y_i^2 \quad - n\overline{Y}^2\right) \\
&= \frac{n\left(\mathbb{V}\left(Y\right) + \theta^2\right) - n\left(\mathbb{V}\left(\overline{Y}\right) + \theta^2\right)}{n-1} \\
&= \mathbb{V}\left(Y\right) = \mathbb{V}\left(\phi(X)\right).
\end{aligned}$$

where $Y = \phi(X)$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

How Good is The Estimator?

- Chebyshev's inequality yields the bound

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\mathbb{V}\left(\hat{\theta}_n\right)}{c^2\sigma^2/n} = \frac{1}{c^2}.$$

- Another estimate follows from the CLT for large n

$$\frac{\sqrt{n}}{\sigma}\left(\hat{\theta}_n - \theta\right) \overset{d}{\approx} \mathcal{N}(0,1) \Rightarrow \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \approx 2\left(1 - \Phi(c)\right).$$

- Hence by choosing $c = c_\alpha$ s.t. $2\left(1 - \Phi(c_\alpha)\right) = \alpha$, an approximate $(1-\alpha)100\%$-CI for θ is

$$\left(\hat{\theta}_n \pm c_\alpha \frac{\sigma}{\sqrt{n}}\right) \approx \left(\hat{\theta}_n \pm c_\alpha \frac{S_{\phi(X)}}{\sqrt{n}}\right).$$

Monte Carlo Integration

- Whatever being Ω; e.g. $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^{1000}$, the error is still in $\sigma/\sqrt{n}$.
- This is in contrast with deterministic methods. The error in a product trapezoidal rule in d dimensions is $\mathcal{O}(n^{-2/d})$ for twice continuously differentiable integrands.
- It is sometimes said erroneously that it beats the curse of dimensionality but this is generally not true as σ^2 typically depends of $\dim(\Omega)$.

Recap

- Monte Carlo is a method to evaluate an integral / sum
- Widely used in high dimensional statistical problems
- It is computationally straightforward
- It has desirable limit properties
- Hard part is often sampling of X
- Some art required for tough X, but beyond scope of this course