Lecture Notes for STAT546: Computational Statistics

—Lecture 1: Monte Carlo

Faming Liang

Purdue University

August 21, 2024

What are Monte Carlo Methods?

The subject of Monte Carlo methods can be viewed as a branch of "experimental mathematics" in which one uses random numbers to conduct experiments. Typically the experiments are done on a computer using anywhere from hundreds to billions of random numbers.

Two categories of Monte Carlo experiments

(1) Direct simulation of a random system.

(2) Addition of artificial randomness to a system.

Direct simulation of a random system

We study a real system that involves some randomness, and we wish to learn what behavior can be expected without actually watching the real system. We first formulate a mathematical model by identifying the key random (and nonrandom) variables which describe the given system. Then we run a random copy of this model on a computer, typically many times over different values of random variables. Finally we collect data from these runs and analyze the data. Here are some applications:

(a) Operations research: e.g., hospital management.

(b) Reliability theory.

(c) Physical, biological, and social sciences: models with complex nondeterministic time evolution, including particle motion, population growth, epidemics, and social phenomena.

A recent concept: Digital Twin

Addition of artificial randomness to a system One represents the underlying problem (which may be completely deterministic) as part of a different random system and then performs a direct simulation of this new system.

Example: estimating π.

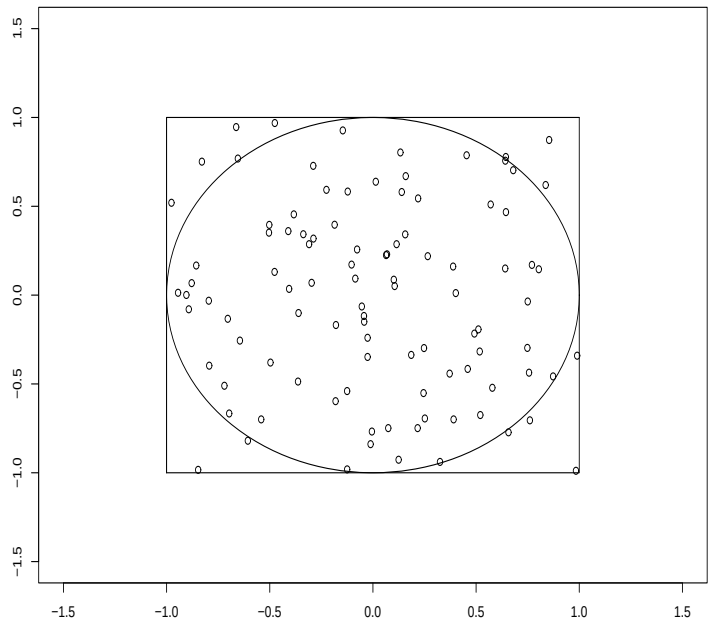


Figure 1: Monte Carlo estimation of π.

(a) Markov chain Monte Carlo: for problems in statistical physics (e.g., proteins, quantum systems) and in Bayesian statistics (for complicated posterior distributions).

(b) Optimization: e.g. solving the traveling salesman and knapsack problems by simulated annealing or genetic algorithms.

Basic problems of Monte Carlo

(a) Draw random variables

$$\boldsymbol{X} \sim \pi(\boldsymbol{x})$$

Sometimes with unknown normalizing constant

(b) Estimate the integral

$$I = \int h(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x} = E_\pi h(\boldsymbol{X})$$

Example: Estimate the integral $I = \int_0^1 h(x)dx$.

- Draw $u_1, \ldots, u_n$ iid from Uniform(0,1).
- Law of large numbers:

$$\hat{I} = \frac{1}{n}[h(u_1) + \cdots + h(u_n)] \to I, \quad \text{as } n \to \infty$$

- Error Computation:
  - Unbiased $E(\hat{I}) = I$.
  - Variance $\text{Var}(\hat{I}) = \frac{\text{Var}(h)}{n} = \frac{1}{n}\int_0^1 (h(x) - I)^2 dx$.

Pseudo-random number generator

- A sequence of pseudo-random number (U_i) is a deterministic sequence of numbers in [0,1] having the same relevant statistical properties as a sequence of random numbers.
- von Neumann's "middle square" method: start with 8653, square it and make the middle 4 digits:

$$8653, 8744, 4575, 9306, 6016, 1922, \ldots$$

- Congruential generator:

$$X_i = aX_{i-1} \bmod M; \quad \text{and} U_i = X_i/M.$$

  say, $a = 7^5 = 16807$, and $M = 2^{31} - 1$.
- Other methods and statistical testing (see Ripley 1987; Marsaglia and Zaman, 1993)

Generating random variables from simple distributions

(a) The inversion method.

Theorem 1

(Probability Integral Transformation Theorem) If $U \sim Uniform[0,1]$ *and* F *is a cdf, then* $X = F^{-1}(U) \equiv inf\{x : F(x) \geq u\}$ *follows* F*.*

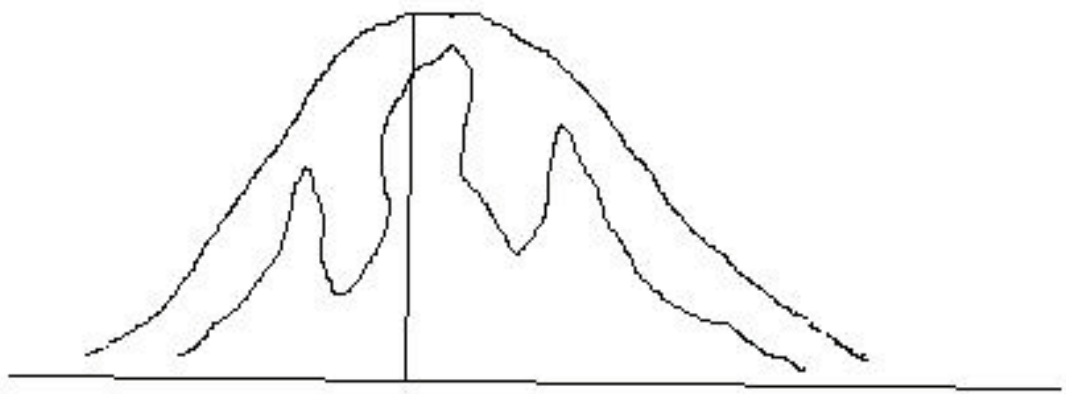
Example: Generation of exponential random variables. $F(x) = 1 - \exp(-\lambda x)$. Set $F(x) = U$, and derive that $x = -\frac{1}{\lambda}\log(1-U)$.

- Generate $U \sim Unif[0,1]$.
- Set $x = -\frac{1}{\lambda}\log(U)$.
- $s \sim \exp(\lambda)$.

(b) The rejection method.

- Generate x from $g(x)$, where $g(x)$ is called the envelope distribution.
- Draw u from unif[0,1]
- Accept x if $u < \pi(x)/cg(x)$.
- The accepted x follows $\pi(x)$.

Efficiency: The acceptance rate$=\int \pi(x)dx/[c\int g(x)dx] = 1/c$.



(c) Specialized methods.

- Generating standard Gaussian variables. (polar method)
- Generating Beta or Dirichlet variables. (Gentle, J.E., 1998)

Variance Reduction Methods

- **Control variables**—Suppose of interest is $\mu = E(X)$.
  - Let C be an r.v. with $E(C) = \mu_c$ and be correlated with X.
  - Let $X(b) = X + b(C - \mu_C)$.

$$\text{Var}(X(b)) = \text{Var}(X) + b^2\text{Var}(C) + 2b\text{Cov}(X, C).$$

    Choose b so as to have a reduced variance $\text{Var}(X(b)) = (1 - \rho_{XC}^2)\text{Var}(X)$.
  - Another scenario: $E(C) = \mu$. Then we can form $X(b) = bX + (1 - b)C$.

- **Antithetic Variates**—a way to produce negatively correlated Monte Carlo samples.

$$X = F^{-1}(U); \quad X' = F^{-1}(1 - U)$$

Then $\text{Cov}(X, X') \leq 0$.
Exercise: prove that X and X' are negatively correlated.
(J.S. Liu, 2001)

- **Rao-Blackwellization** This method reflects a basic principle in Monte Carlo computation: One should carry out analytical computation as much as possible.
  A straightforward estimator of $I = E_\pi h(\boldsymbol{x})$ is

$$\hat{I} = \frac{1}{m}\{h(\boldsymbol{x}^{(1)} + \cdots + h(\boldsymbol{x}^{(m)}\}.$$

  Suppose that $\boldsymbol{x}$ can be decomposed into two parts (x_1, x_2) and that the conditional expectation $E[h(\boldsymbol{x})|x_2]$ can be carried out analytically (integrating out x_1). An alternative estimator of I is

$$\tilde{I} = \frac{1}{m}\{E[h(\boldsymbol{x})|x_2^{(1)}] + \cdots + E[h(\boldsymbol{x})|x_2^{(m)}\}.$$

  Both $\hat{I}$ and $\tilde{I}$ are unbiased because of the simple fact that

$$E_\pi h(\boldsymbol{x}) = E_\pi[E\{h(\boldsymbol{x})|x_2\}].$$

Following from the identity

$$\mathrm{Var}\{h(\boldsymbol{x})\} = \mathrm{Var}\{E[h(\boldsymbol{x})|x_2]\} + E\{\mathrm{Var}[h(\boldsymbol{x})|x_2]\},$$

we have

$$\mathrm{Var}(\hat{I}) = \frac{1}{m}\mathrm{Var}\{h(\boldsymbol{x})\} \geq \frac{1}{m}\mathrm{Var}\{E[h(\boldsymbol{x})|x_2]\} = \mathrm{Var}(\tilde{I}).$$

Importance Sampling and Weighted sample

- Of interest $\mu = \int h(x)\pi(x)dx$.
- We can (and sometimes prefer to) draw samples from $g(x)$, which concentrates on "region of importance" and has a larger support set than that of $\pi(x)$.
- Let $X_1, \ldots, X_N$ be samples from $g(x)$, then

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i h(X_i),$$

  where $w_i = \pi(X_i)/g(X_i)$.
- An alternative estimator is

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{n} w_i h(X_i),$$

  provided that the normalizing constants of g and π are known.

Importance Sampling

- For the simpler estimator: $E(\tilde{\mu}) = \mu$ and

$$\text{Var}(\tilde{\mu}) = \frac{1}{N}\text{Var}_g\{\frac{h(x)\pi(x)}{g(x)}\} = \frac{1}{N}\text{Var}_g\{Z\},$$

where $Z = h(x)w(x)$.

- For the more useful one: The normalizing constants of g and π are not required, and $E_g w(x) = 1$.

$$E(\hat{\mu}) = E\{\frac{\bar{Z}}{\bar{W}}\} \approx \mu - \text{Cov}(\bar{Z}, \bar{W}) + \mu\text{Var}(\bar{W}),$$

and

$$\text{Var}(\hat{\mu}) \approx \frac{1}{N}[\mu^2\text{Var}_g(W) + \text{Var}_g(Z) - 2\mu\text{Cov}_g(W, Z)].$$

Hence, the mean squared error (MSE) of $\tilde{\mu}$ is

$$MSE(\tilde{\mu}) = \frac{1}{N}\mathrm{Var}_g\{Z\},$$

and that for $\hat{\mu}$ is

$$\begin{aligned} MSE(\hat{\mu}) &= [E_g(\hat{\mu}) - \mu]^2 + \mathrm{Var}_g(\hat{\mu}) \\ &= \frac{1}{N}MSE(\tilde{\mu}) + \frac{1}{N}[\mu^2\mathrm{Var}_g(W) - 2\mu\mathrm{Cov}_g(W, Z)] + O(N^{-2}) \end{aligned}$$

Then MSE($\hat{\mu}$) is smaller in comparison with MSE($\tilde{\mu}$) when $\mu^2\mathrm{Var}_g(W) - 2\mu\mathrm{Cov}_g(W, Z) < 0$.

Properly weighted samples

The set $\{(X_i, W_i)\}_{i=1}^N$ is said "properly weighted with respect to π if for any $h(x)$, $\mu = E_\pi h(x)$ can be approximated as

$$\hat{\mu} = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i h(X_i).$$

Mathematically, this says that

$$\frac{E[Wh(X)]}{E(W)} = E_\pi h(x).$$

Rule of thumb in importance sampling:

$$n_{\text{eff}} \approx \frac{n}{1 + \text{Var}_g(W)}.$$

Summary of Importance Sampling

- Of interest $\mu = \int h(x)\pi(x)dx$.
- We can (and sometimes prefer to) draw samples from $g(x)$, which concentrates on "region of importance" and has a larger support set than that of $\pi(x)$.
- Let $X_1, \ldots, X_N$ be samples from $g(x)$, then

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i h(X_i),$$

where $w_i = \pi(X_i)/g(X_i)$.

Sequential Importance Sampling

- Suppose $\boldsymbol{x} = (x_1, \ldots, x_k)$; of interest is

$$\pi(\boldsymbol{x}) = \pi(x_1)\pi(x_2|x_1)\cdots\pi(x_k|x_1, \ldots, x_{k-1})$$

- The trial density can also be decomposed similarly

$$g(\boldsymbol{x}) = g_1(x_1)g_2(x_2|x_1)\cdots g_k(x_k|x_1, \ldots, x_{k-1})$$

$$w(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{g(\boldsymbol{x})} = \frac{\pi(x_1)\pi(x_2|x_1)\cdots\pi(x_k|x_1, \ldots, x_{k-1})}{g_1(x_1)g_2(x_2|x_1)\cdots g_k(x_k|x_1, \ldots, x_{k-1})}$$

  - The weight can be computed sequentially
  - In practice, we create a sequence of approximations, $\pi(x_1)$, $\pi(x_1, x_2)$, $\cdots$, to guide the IS at each stage.

Sequential Importance Sampling: weight caculation

- Generate $x_1 \sim g_1(x_1)$, set the weight

$$w_1(x_1) = \frac{\pi_1(x_1)}{g_1(x_1)}.$$

- Generate $x_2 \sim g_2(x_2|x_1)$, set the weight

$$w_2(x_1, x_2) = w_1(x_1)\frac{\pi(x_2|x_1)}{g_2(x_2|x_1)}.$$

- .....
- Generate $x_k \sim g_k(x_k|x_1, \ldots, x_{k-1})$, set the weight

$$w(x_1, \ldots, x_k) = w_{k-1}(x_1, \ldots, x_{k-1})\frac{\pi(x_k|x_1, \ldots, x_{k-1})}{g_t(x_k|x_1, \ldots, x_{k-1})}.$$

Variations of Sequential importance sampling

(Grassberger, 1997)
Set upper cutoff values C_t and lower cutoff values c_t for $t = 1, \ldots k$.

- **Enrichment:** If $w_t > C_t$, the chain $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ is split into r copies, each with weight w_t/r.
- **Pruning:** If $w_t < c_t$, one flips a fair coin to decide whether to keep it. If it is kept, its weight w_t is doubled to $2w_t$.

A smart way to enrich the good partial conformations, and prune the bad ones.

Similar idea: resampling.

Related works

- Sequential importance sampling with partial rejection control (Liu, Chen and Wong, JASA, 1998)
- Dynamically weighted importance sampling (Liang, JASA, 2002)

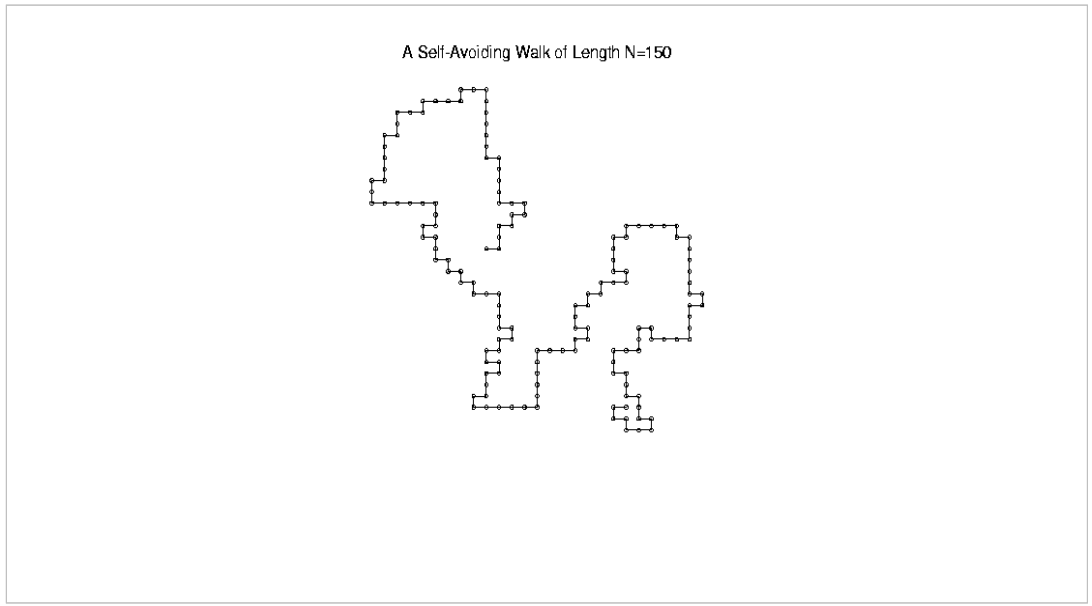
Example: Polymer Simulation

HP model: HHHHHHPPPPHHHPPHHHHHHH
H: hydrophobic amino acid (nonpolar);
P: hydrophilic amino acid (polar).
Self-avoid random walk: fold a protein on a 2D square lattice.

- Add one new position a time to one of the k ($k \leq 3$) available positions.
- Question: sampling distribution of the chain?

$$g(\text{a straight chain}) = \frac{1}{4} \times \frac{1}{3} \times \frac{1}{3}$$

Self-avoid random walk



85-Mer sequence