Lecture Notes on Monte Carlo Methods
Fall Semester, 2005
Courant Institute of Mathematical Sciences, NYU
Jonathan Goodman, goodman@cims.nyu.edu

Introduction

created October 15, 2005

This is a set of lecture notes for a graduate class on Monte Carlo methods given at the Courant Institute of Mathematical Sciences at NYU in the fall of 2005. I am motivated by seeing many new people begin to use Monte Carlo. Many new techniques and applications are emerging. I hope the class will cover the basic principles quickly but with enough depth to be helpful to practitioners. Then we will cover a few of the many advanced areas, hopefully in a way that people from diverse fields can understand. Given the variety of people using Monte Carlo and the range of application areas, it is surprising how much common ground can be found.

Roughly speaking, *Monte Carlo*[1] means computing using random numbers. It is helpful to refine this definition and distinguish between true Monte Carlo and *simulation*. We take simulation to mean generating individual random objects faithfully according to some model. For example, we might want to see what shape clouds come from a specific model of cloud formation that involves randomness. The point of simulation might not be gather detailed statistics, but just to see what a few random objects look like.

By contrast, Monte Carlo uses random numbers as a means to evaluate quantities that themselves are not random. For example, suppose $f(x)$ is the probability density for a one dimensional random variable, X. One way to evaluate[2] $A = E[X]$ is to generate many *samples* (random variables X_k with probability density f) and average them. More generally, we could generate a number of random objects (e.g. clouds) and collect interesting statistics about them. The difference is that the expected value of X or of some more complex statistic is a property of a random variable but is not itself random. Therefore, it may be possible to evaluate A without generating random samples with probability density $f(x)$. For example, we could estimate $A = \int xf(x)dx$ by numerical quadrature. Practitioners often find clever methods that are better (faster or more accurate) than *plain simulation*. Simulation is mostly programming but Monte Carlo is all about devising and understanding new computational strategies. Each piece of creativity will improve your results.

When choosing a Monte Carlo method for a given problem, one should have a strong bias against Monte Carlo at all. A practical deterministic method almost always is better than Monte Carlo. If you have an integral in fewer than, say, four variables, do it by deterministic quadrature. If you have a

[1] The name comes from Monte Carlo, a traditional gambling center in Europe, another place where random numbers are important.

[2] See the next section for basic definitions and notation.

Markov chain with less than a thousand states or a diffusion process in less than four dimensions, use a backward equation. The statistical error in Monte Carlo is at least $O(1/\sqrt{L})$ (which isn't very small), where L is the number of samples[3] It is a rare deterministic method that is worse than this.

Keep this in mind when reading Monte Carlo books. We authors often use one dimensional examples to illustrate specific principles although we know that this is not the fastest or most accurate way to compute a one dimensional integral. The real world is full of problems where Monte Carlo is the method of choice.

1 Overview

An example will illustrate the themes discussed in these notes. One part of a physics problem reduces to calculating the integral

$$A(\lambda) = \int_{|x|<1} \int_{|y|<1} \frac{e^{-\lambda|x-y|}}{|x-y|} dxdy \;, \tag{1}$$

Here, the x and y variables both are in R^3. This is a six dimensional integral (maybe reducible to five or even four dimensions), which makes direct quadrature difficult. The problem is first to calculate $A(\lambda)$ for various λ values and next to find λ_* that gives $A(\lambda_*) = \frac{1}{2}$.

The first step in evaluating $A(\lambda)$ is to express in terms of the expected value of a random variable. Here we will have

$$A(\lambda) = Const \cdot B(\lambda) \;\;, \quad B(\lambda) = E_\lambda[Z] \;. \tag{2}$$

The $E_\lambda[\cdot]$ on the right means that the probability distribution of Z depends on λ. For example, we could let X and Y be independent random variables with uniform probability density in $B = \left\{ x \in R^3 \text{ with } |x| < 1 \right\}$. This probability density is (recall that the volume of B is $\frac{4}{3}\pi$)

$$f(x) = \begin{cases} \frac{3}{4\pi} & \text{if } x \in B \\ 0 & \text{otherwise.} \end{cases}$$

Then multiplying and dividing by the normalizing $\frac{4}{3}\pi$ gives

$$\begin{aligned} A(\lambda) &= \left(\frac{4}{3\pi}\right)^2 \int\int \frac{e^{-\lambda|x-y|}}{|x-y|} f(x)f(y)dxdy \\ &= \left(\frac{4}{3\pi}\right)^2 E\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] \\ &= \left(\frac{4}{3\pi}\right)^2 E[Z] \;, \end{aligned}$$

[3]We use L for the number of samples, the *run length* so that n always can be the number of components of a multivariate random variable.

where

$$Z = \frac{e^{-\lambda|X-Y|}}{|X-Y|} \ . \tag{3}$$

We do not have a formula for the probability density of Z, but we can generate a random *sample* by choosing X and Y independently in B then applying (3).

To evaluate $B(\lambda)$, we generate L independent Z samples, Z_1, ..., Z_L, and use the *estimator*[4]

$$E[Z] = B(\lambda) \approx \widehat{B}(\lambda) = \frac{1}{L}\sum_{k=1}^{L} Z_k \ . \tag{4}$$

The *law of large numbers* states that $\widehat{B} \to B$ as $L \to \infty$. It may take quite a large L, and lots of computer time, to get $\widehat{B}$ close enough to B to satisfy us. Computational *error bars* tell us how far $\widehat{B}$ is likely to be from B. We will follow the statistician's habit of expressing this with a *confidence interval*. For example, we might say that the probability that B is not between $\widehat{B} - r$ and $\widehat{B} + r$ is about 5%. A Monte Carlo practitioner who neglects error bars deserves to get the wrong answer, and will. It is not always necessary to present the error bars to the consumers of your results, just as you don't show them how you debugged and tested your computer code.

Applications demand more than just $\widehat{B}(\lambda_j)$ for a few λ_j. We want to know the function $B(\lambda)$, the shape of its graph, its derivatives, the inverse function, etc. The statistical noise in the estimators $\widehat{B}(\lambda_j)$ may be amplified in finding this other properties of $B(\lambda)$. There often are better ways[5]. *Sensitivity analysis* provides estimates for derivatives of $B(\lambda)$ having less noise than the naive finite difference (see below). *Stochastic approximation* provides more sophisticated ways to find solutions of equations such as $B(\lambda) = \frac{1}{2}\left(\frac{pi}{4}\right)^2$ (solve for λ so that $A(\lambda) = \frac{1}{2}$).

There are other estimators of $B(\lambda)$ besides (3),(4). *Variance reduction* means searching for alternatives with (hopefully) less statistical error. For example, if we could sample (X, Y) pairs from the 6 dimensional probability density

$$g(x, y) = Const \cdot \frac{f(x)f(y)}{|x-y|} \ , \tag{5}$$

then $B(\lambda) = E_g[e^{-\lambda|X-Y|}]$ would give more accurate estimates of B for the same number of samples. This is an example of *importance sampling*. It is harder to create samples from the density (5) than to sample X and Y from f independently. Powerful sampling methods such as *rejection* and *Markov chain Monte Carlo* (MCMC) will be very handy.

The simple estimator (3),(4) is *unbiased*, which means that $B(\lambda) = E[\widehat{B}(\lambda)]$. Many estimators have bias as well as statistical error. For example, since $B(\lambda)$ is

[4]Statisticians put a "hat" on a quantity to indicate a statistical estimate of it. In this way, the statistical estimate of B is $\widehat{B}$.

[5]Ask A.C. for a joke with this punch line.

a nonlinear function of λ, an estimator, $\widehat{\lambda_*}$, of λ_* with $B(\lambda_*) = \frac{1}{2}\left(\frac{p_i}{4}\right)^2$ probably has *bias*

$$\text{bias} = E[\lambda_*] - E[\widehat{\lambda_*}] = O\left(\frac{1}{L}\right) .$$

There may be a tradeoff between statistical error and bias. For example, (3),(4) has statistical error $B(\lambda) - \widehat{B}(\lambda) = O\left(\frac{1}{\sqrt{L}}\right)$. The finite difference estimator

$$\widehat{B'}(\lambda) = \frac{\widehat{B}(\lambda + \Delta\lambda) - \widehat{B}(\lambda)}{\Delta\lambda} \tag{6}$$

has bias

$$E\left[\widehat{B'}\right] - B' = \frac{B(\lambda + \Delta\lambda) - B(\lambda)}{\Delta\lambda} - B'(\lambda) = O(\Delta\lambda) .$$

Assuming $\widehat{B}(\lambda+\Delta\lambda)$ and $\widehat{B}(\lambda)$ have independent statistical error, the statistical error of $\widehat{B'}$ is on the order of $1/\Delta\lambda\sqrt{L}$. Altogether, we minimize the total error by (setting all constants equal to one)

$$\min_{\Delta\lambda}(\text{bias} + \text{noise}) = \min_{\Delta\lambda}\left(\Delta\lambda + \frac{1}{\Delta\lambda\sqrt{L}}\right) = \frac{1}{L^{1/4}} .$$

Sensitivity analysis and variance reduction will improve this.

2 Background

The prerequisites for the class are a course on stochastic processes (Stochastic Calculus at the Courant Institute) and some experience with scientific computing at the level of our Scientific Computing class. We will use linear algebra and multivariate calculus at the level of the Courant Institute beginning classes. The discussion will be as informal mathematically as possible.

Random variables generally are denoted by capitol letters, X, Y, T, etc., with specific values denoted by lower case: x, y, t, etc. A random element of R^n would be called a random vector or a multivariate random variable and could be written $X = (X_1, \ldots, X_n)$. We have a scalar random variable when $n = 1$. The probability density for X might be called $f(x)$, so that

$$P(A) = P(X \in A) = \int_A f(x)dx . \tag{7}$$

Here $P(A)$ is the probability of the event $A \subseteq R^n$. The *expected value* of X with the law f is

$$E_f[X] = \int_{R^n} xf(x)dx .$$

We write $E[\cdot]$ for $E_f[\cdot]$ when the f is clear. In one dimension it may be clearer to use old fashion differential notation

$$P(x \leq X \leq x + dx) = f(x)dx . \tag{8}$$

Here the event A is the small interval $(x, x + dx)$. The numbers $P(A)$ form a probability measure even if they are not given in terms of a density as in (1.1). The *law* of a random variable either is its probability density or its probability measure. We write $X \sim f$ or $X \sim P$ to indicate that X has the law given by f or P. The vector notation above notwithstanding, we often write $X_k \sim f$ to indicate a sequence random variables each with the law f. If $X_k \in R^n$, we can indicate vector components by $X_k = (X_{k1}, \ldots, X_{kn})$. We say the X_k are *samples* of the law f or samples of the random variable X. Monte Carlo computations involve thousands or millions of samples.

A *standard uniform* random variable is a scalar, U, with probability density $f(u) = 1$ for $0 \leq u \leq 1$ and $f(u) = 0$ otherwise. A *pseudo random number generator* is a piece of computer code that produces a sequence U_k, $k = 1, 2, \ldots$, that resembles a sequence of independent samples of a standard uniform. This is not to be confused with *quasi random numbers* which are deliberately not random, in an attempt to be more uniform than actual random numbers. The term "pseudo random" indicates that the output of the pseudo random number generator is not actually random. If you run the pseudo random number generator twice, with the same seed (see a later lecture for a more technical discussion), you will get the same U_k.

The central limit theorem underlies much Monte Carlo error analysis. The simplest case is Y_1, ..., Y_L, independent samples of a scalar random variable, Y with $E[Y] = 0$ and $E[Y^2] = \sigma_Y^2$. The theorem is that the law of $Z_L = L^{-1/2} \sum_{k=1}^{L} Y_k$ converges to a Gaussian law with mean zero and variance σ_Y^2 as $L \to \infty$. The next case is a multivariate random variable, Y with mean zero and covariance $C_Y = E[YY^t]$. The law of Z_L converges to a multivariate normal with mean zero and covariance C_Y.

Wick's theorem is a recipe that evaluates any moment of a multivariate normal. Suppose $l_1(Y)$, ..., $l_{2m}(Y)$ are an even number of *linear functionals*.[6] A *pairing* is a collection of m pairs $\{\{j_1, k_1\}, \cdots, \{j_m, k_m\}\}$, so that each of the numbers $1, \ldots, 2m$ appears exactly once. Reordering the pairs or exchanging the numbers in a pair gives the same pairing. The pairings of $\{1, 2, 3, 4\}$ are $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 3\}, \{2, 4\}\}$, and $\{\{1, 4\}, \{2, 3\}\}$. The number of pairings of $2m$ numbers is[7] $(2m - 1)(2m - 3) \cdots 3$. Wick's theorem states that if Y is a mean zero multivariate normal, then

$$E\left[l_1(Y) \cdot \cdots \cdot l_{2m}(Y)\right] = \sum_{\text{pairings}} E\left[l_{j_1}(Y) l_{k_1}(Y)\right] \cdot \cdots \cdot E\left[l_{j_m}(Y) l_{k_m}(Y)\right] \quad (9)$$

An example illustrates many features of this formula. Let (Y_1, Y_2) be a mean zero bivariate normal with $\text{var}(Y_1) = 2$, $\text{var}(Y_2) = 5$, and $\text{cov}(Y_1, Y_2) = 3$, then

[6] A linear functional is a scalar linear function of Y. If Y is a column vector, $l(Y)$ can be represented as $l \cdot Y$, for some row vector, also called l. Scalar functions, particularly of high or infinite dimensional variables, often are called *functionals*.

[7] Number one chooses a partner from among the $2m - 1$ other numbers. This is the factor $2m - 1$. Then the lowest as yet unpaired number chooses a partner from the $2m - 3$ remaining numbers, and so on.

corresponding to the three pairings above, we have the three terms

$$\begin{aligned} E\left[Y_1^2 Y_2^2\right] &= E\left[Y_1 \cdot Y_1 \cdot Y_2 \cdot Y_2\right] \\ &= E\left[Y_1 Y_1\right] E\left[Y_2 Y_2\right] + E\left[Y_1 Y_2\right] E\left[Y_1 Y_2\right] + E\left[Y_1 Y_2\right] E\left[Y_1 Y_2\right] \\ &= 2 \cdot 5 + 3 \cdot 3 + 3 \cdot 3 \\ &= 33 \, . \end{aligned}$$

The last two terms in the sum are the same but correspond to the distinct pairings $\{\{1,3\},\{2,4\}\}$ and $\{\{1,4\},\{2,3\}\}$. All the terms on the right are given by covariances, which illustrates the general fact that a Gaussian is determined by its mean and covariance matrix. This example used row vectors $l_1 = l_2 = (1,0)$ and $l_3 = l_4 = (0,1)$. For more complicated row vectors, we have the formula (which the reader should verify)

$$E[l_1(Y) l_2(Y)] = l_1 C l_2^t \, ,$$

where C is the covariance matrix of Y. For a scalar mean zero Gaussian, all pairings give the same contribution, so

$$E[X^4] = 3\sigma_X^4 \, , \quad E[X^6] = 15\sigma_X^6 \, , \quad E[X^8] = 105\sigma_X^8 \, , \text{etc.}$$

One of the proofs of the central limit theorem also explains Wick's theorem. If a mean zero random variable satisfies Wick's theorem, then all its moments are the same as the Gaussian moments so it is Gaussian[8]. We take a concrete case:

$$E[l_1(Z_l) \cdot \cdots \cdot l_4(Z_L)] \rightarrow \text{ (Wick formula) } \quad \text{as } L \rightarrow \infty.$$

To evaluate the sum, we use a different summation index for each factor $j = 1,2,3,4$:

$$l_j(Z_L) = \frac{1}{\sqrt{L}} \sum_{k_j=1}^{L} l_j(Y_{k_j}) \, .$$

so that

$$\begin{aligned} & E\left[l_1(Z_L) l_2(Z_L) l_3(Z_L) l_4(Z_L)\right] \\ &= \frac{1}{L^2} \sum_{k_1=1}^{L} \sum_{k_2=1}^{L} \sum_{k_3=1}^{L} \sum_{k_4=1}^{L} E[l_1(Y_{k_1}) l_2(Y_{k_2}) l_3(Y_{k_3}) l_4(Y_{k_4})] \end{aligned}$$

Most of the expectations on the right side are zero. For example, if $k_1 \neq k_2$, $k_1 \neq k_3$, and $k_1 \neq k_4$ then $l_1(Y_{k_1})$ has mean zero and is independent of the other $l(Y)$ factors so $E[l_1(Y_{k_1}) l_2(Y_{k_2}) l_3(Y_{k_3}) l_4(Y_{k_4})] = 0$. To get a nonzero expectation, each term must be paired with at least one other. For example, the pairing $\{\{1,3\},\{2,4\}\}$ corresponds to possibly nonzero terms with $k_1 = k_3$ and $k_2 = k_4$: $E[l_1(Y_{k_1}) l_2(Y_{k_2}) l_3(Y_{k_1}) l_4(Y_{k_2})]$.

[8]The justification of this statement is an easy case of the *moment problem*, showing that two probability laws with the same moments are the same.

If $k_1 \neq k_2$ then Y_{k_1} and Y_{k_2} are independent, and we get a Wick type contribution:

$$\begin{aligned} E[l_1(Y_{k_1})l_2(Y_{k_2})l_3(Y_{k_1})l_4(Y_{k_2})] &= E[l_1(Y_{k_1})l_3(Y_{k_1})]E[l_2(Y_{k_2})l_4(Y_{k_2})] \\ &= \left(l_1 C l_3^t\right)\left(l_2 C l_4^t\right) . \end{aligned}$$

There are $L(L-1)$ such terms. The L remaining terms have $k_1 = k_2 = k_3 = k_4$. Altogether,

$$\begin{aligned} &E\left[l_1(Z_L)l_2(Z_L)l_3(Z_L)l_4(Z_L)\right] \\ &= \frac{L(L-1)}{L^2}\Big\{\left(l_1 C l_2^t\right)\left(l_3 C l_4^t\right) + \left(l_1 C l_3^t\right)\left(l_2 C l_4^t\right) + \left(l_1 C l_4^t\right)\left(l_2 C l_3^t\right)\Big\} \\ &\quad + \frac{L}{L^2} E[l_1(Y)l_2(Y)l_3(Y)l_4(Y)] . \end{aligned}$$

The first term on the right converges to the Wick formula as $L \to \infty$ while the second converges to zero.

This argument has drawbacks as a proof of the central limit theorem. It requires Y to have finite moments (e.g. so that the last term above is finite). It relies on the moment problem, which, at a minimum, is more sophisticated than this. It does not answer questions about convergence of the distribution of Z_L other than moments. However, if you believe the central limit theorem already and just want Wick's formula, take Y to be Gaussian.

Here is an example of the multivariate central limit theorem that comes up in the theory of time stepping methods for stochastic differential equations. Suppose $X \sim \mathcal{N}(0,1)$ and we define a three component random variable Y as $Y_k = (X^2 - 1, X^3, X^4 - 3)^t$. The Y variances (handily computed using Wick's theorem) are $C_{11} = 2$, $C_{22} = 15$ and $C_{33} = 96$. The off diagonal covariances all are zero except $C_{13} = 12$. Note that although Y_1 and Y_2 are uncorrelated ($C_{12} = 0$), they are far from independent. In fact, if we know Y_1, then $X = \pm\sqrt{Y_1 + 1}$: If we know Y_1 then we know everything about Y_2 except the sign. Nevertheless, if we have L independent X samples then the random variables

$$Z_{L,1} = \frac{1}{\sqrt{L}} \sum_{k=1}^{L} \left(X_k^2 - 1\right) \quad , \quad Z_{L,2} = \frac{1}{\sqrt{L}} \sum_{k=1}^{L} X_k^3$$

are nearly independent Gaussians for large L.