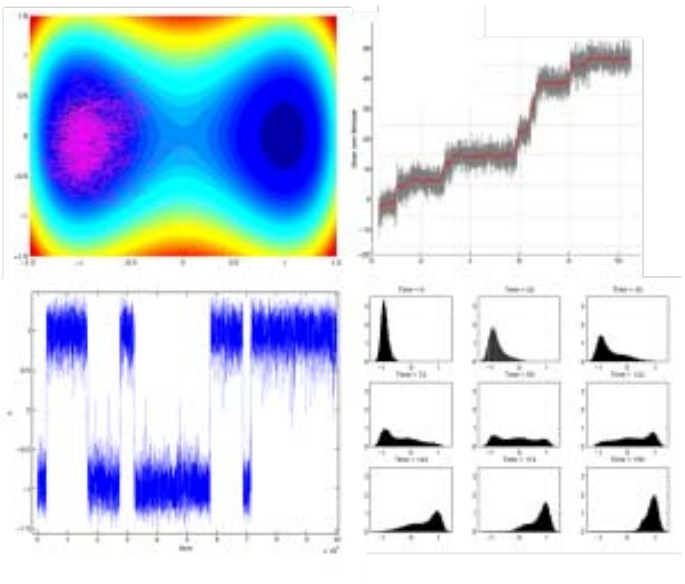
Strategies for Improving the Efficiency of Monte-Carlo Methods

Paul J. Atzberger



Comments should be sent to:
atzberg@math.ucsb.edu

Introduction

The Monte-Carlo method is a useful approach is estimating numerically the value of integral expressions, especially in high dimensions. For the integral

$$I = \int_\Omega f(\mathbf{x}) d\mathbf{x}$$

in which Ω is the domain of integration, the integral I can be related to an expectation of a random variable with respect to some probability measure. For probability measures of a random variable X that have a density $\rho(\mathbf{x})$ the expectation can be expressed as:

$$E(f(\mathbf{X})) = \int_\Omega f(\mathbf{x})\rho(\mathbf{x}) d\mathbf{x}.$$

The integral I can be expressed in terms of an expectation in a number of different ways. One rather general approach is to use a density having the feature that $\rho(\mathbf{x}) > 0$ whenever $f(\mathbf{x}) \neq 0$. This gives that:

$$\begin{aligned} I &= \int_\Omega f(\mathbf{x}) d\mathbf{x} = \int_\Omega \frac{f(\mathbf{x})}{\rho(\mathbf{x})}\rho(\mathbf{x}) d\mathbf{x} \\ &= E\left(\frac{f(\mathbf{X})}{\rho(\mathbf{X})}\right) \\ &= E\left(g(\mathbf{X})\right). \end{aligned}$$

where $g(\mathbf{x}) = \frac{f(\mathbf{x})}{\rho(\mathbf{x})}$. In the case of a domain of integration Ω which is finite, we can always use the random variable $\mathbf{X}$ uniformly distributed on Ω with density $\rho(\mathbf{x}) = \frac{1}{|\Omega|}$ to obtain:

$$\begin{aligned} I &= \int_\Omega f(\mathbf{x}) d\mathbf{x} = \int_\Omega \frac{f(\mathbf{x})}{\frac{1}{|\Omega|}} \frac{1}{|\Omega|} d\mathbf{x} \\ &= |\Omega| E(f(\mathbf{X})). \end{aligned}$$

The utility of expressing the integral in terms of an expectation derives from the *Law of Large Numbers*, which states that for a collection of independent identically distributed random variables $\{\mathbf{X}_i\}_{i=1}^\infty$:

$$E\left(g(\mathbf{X})\right) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N g(\mathbf{X}_i).$$

This offers a way to estimate the numerical value of I, in particular:

- generate N random variates $\{\mathbf{X}_i\}_{i=1}^N$ with distribution $\rho(\mathbf{x})$ on Ω.
- approximate the expectation using the *Law of Large Numbers* $I \approx \frac{1}{N}\sum_{i=1}^N g(\mathbf{X}_i)$.

This gives a probabilistic approach to estimating the quantity I.

The Monte-Carlo method has an accuracy which can be estimated as:

$$\begin{aligned} \text{error} &= \left| \frac{1}{N} \sum_{i=1}^N g(\mathbf{X}_i) - I \right| \\ &= \left| \frac{\sigma_g}{\sqrt{N}} \left(\frac{\sum_{i=1}^N g(\mathbf{X}_i) - NI}{\sigma_g \sqrt{N}} \right) \right| \\ &\approx \left| \frac{\sigma_g}{\sqrt{N}} \eta(0,1) \right| \end{aligned}$$

where

$$\sigma_g^2 = \int_\Omega \left(g(\mathbf{x}) - I\right)^2 \rho(\mathbf{x}) d\mathbf{x}$$

and $\eta(0,1)$ denotes a standard normal random variable (Gaussian random variable) with mean zero and variance 1. The last approximation was obtained by using the *Central Limit Theorem*, which states that for a sum of i.i.d random variables Y_i with mean μ and finite variance σ^2:

$$\frac{\sum_{i=1}^N Y_i - N\mu}{\sigma\sqrt{N}} \to \eta(0,1), \text{ as } N \to \infty.$$

This shows that asymptotically the error converges at a rate $O(\frac{1}{\sqrt{N}})$, independent of the dimensionality of the problem considered. Furthermore, the convergence rate in the Monte-Carlo method is strongly influenced by the prefactor σ_g which depends on the function $f(\mathbf{x})$ and the sampling distribution with density $\rho(x)$ that is used. The prefactor σ_g presents the primary avenue by which the convergence rate can be improved. We shall discuss a number of approaches by which one can attempt to reduce the size of σ_g.

Variance Reduction Techniques

An important consideration in designing effective Monte-Carlo methods is to formulate the estimate for the integrals in terms of the expectation of random variables that have variances as small as possible. The Monte-Carlo method gives the estimate:

$$I = E\left(g(x)\right) \approx \frac{1}{N}\sum_{i=1}^N g(X_i).$$

As discussed in the section on the accuracy of the method the error can be estimated by:

$$\text{error} = \frac{\sigma_g}{\sqrt{N}}$$

where

$$\sigma_g^2 = \int_\Omega \left(g(\mathbf{x}) - I\right)^2 \rho(\mathbf{x}) d\mathbf{x}.$$

In the Monte-Carlo estimate recall the random variables X_i were generated having probability density $\rho(x)$.

This suggests one approach by which the Monte-Carlo method rate of convergence can be improved. In particular, to use a probability density $\rho(x)$ for which generation of variates X_i is not too difficult while making σ_g^2 small.

Antithetic Variates

In the Monte-Carlo method a sequence of independent identically distributed variates $\mathbf{X}_i$ were generated. One strategy to reduce the variance of the Monte-Carlo estimate is to attempt to develop a corresponding estimate based instead on a sequence of variates $\mathbf{X}_i$ which have desirable correlations resulting in cancellations in the sum which yield to a smaller effective variance for the estimate. The justification of such an estimate of course would requiring confirming a generalization of the *Law of Large Numbers* in which random variables are correlated, which is possible in many cases, see (1; 3).

To make the central ideas behind this strategy more concrete consider the following Monte-Carlo estimate when the number of samples is even, $N = 2n$:

$$I \quad = \quad \frac{1}{N}\sum_{i=1}^N g(\mathbf{X}_i)$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N} G_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{G_{2i-1}+G_{2i}}{2} \\
&= \frac{1}{n}\sum_{i=1}^{n} H_i
\end{aligned}
$$

where we let $G_i = g(\mathbf{X}_i)$ and $H_i = \frac{G_{2i-1}+G_{2i}}{2}$.

Now the variance appearing in the error expression when using the random variates H_i is

$$
\begin{aligned}
\sigma_h^2 &= \mathrm{Var}\left(H_i\right) \\
&= \frac{1}{4}\left(\sigma_g^2 + \sigma_g^2 + 2\mathrm{Cov}(H_{2i-1}, H_{2i})\right) \\
&= \frac{1}{2}\left(\sigma_g^2 + \mathrm{Cov}(G_{2i-1}, G_{2i})\right).
\end{aligned}
$$

Now it is important to note that we had to generate two variates G_{2i-1}, G_{2i} to obtain one variate H_i so real progress has only been made in the amount of computational effort to estimate I only if $\sigma_h^2 < \frac{1}{2}\sigma_g^2$. By the last term we see that the estimate is improved if we introduce negative pairwise correlations for G_i so that $\mathrm{Cov}(G_{2i-1}, G_{2i}) < 0$. Furthermore, we see that if $\mathrm{Cov}(G_{2i-1}, G_{2i}) > 0$ the performance of the estimate is actually worse. Thus some care must be taken to ensure that negative correlations are generated. The variate G_{2i} is often referred to as the "antithetic" variate of G_{2i-1}.

One strategy to generate a sequence of G_i with negative correlations is to make use of the transformation theorems. In the case that we generate by using the transformation from the uniform variate $G_{2i-1} = F^{-1}(U_{2i-1})$ we can let $G_{2i} = F^{-1}(1 - U_{2i-1})$. This will yield the desired negative correlations. In the case that $g(\mathbf{X})$ is monotonic we could also apply this to $\mathbf{X}_{2i-1}$ and $\mathbf{X}_{2i}$ to obtain the desired correlations. In the case that $\mathbf{X}_{2i-1} \in \mathbb{R}^n$ and the probability is symmetric about 0 the antithetic variate can sometimes be generated by letting $\mathbf{X}_{2i} = -\mathbf{X}_{2i-1}$, but again care must be taken to ensure this in fact yields negative correlations. In general it is not important how the correlations are obtained algorithmically only that they be negative and that the random variates G_i be identically distributed with the appropriate probability.

Partial Averaging

The Monte-Carlo method estimates the integral I by computing an expectation of a random variable with probability density $\rho(\mathbf{x})$ which can be expressed as:

$$
I = \int g(\mathbf{x})\rho(\mathbf{x})d\mathbf{x}.
$$

In the case that the integrals appearing above can be partially evaluated, the variance of the estimate can be significantly reduced. In effect, if the integrals can be integrated analytically over a subspace of Ω this allows us to replace the integrand $g(\mathbf{x})$ with another integrand $\bar{g}(\mathbf{x})$ yielding the same value I but with a smaller variance $\sigma_{\bar{g}}^2 < \sigma_g^2$.

To make this more mathematically clear suppose we can compute the conditional expectation of $g(\mathbf{X})$ for some sigma algebra $\mathcal{F}$:

$$
\bar{g} = \mathrm{E}\left(g(\mathbf{X})|\mathcal{F}\right).
$$

By the definition of the conditional expectation:

$$
\mathrm{E}\left(\bar{g}(\mathbf{X})\right) = \mathrm{E}\left(\mathrm{E}\left(g(\mathbf{X})|\mathcal{F}\right)\right) = \mathrm{E}\left(g(\mathbf{X})\right) = I.
$$

This gives another approach to approximating I, we can generate variates $\mathbf{X}_i$ and estimate using:

$$I \approx \frac{1}{N}\sum_{i=1}^{N}\bar{g}(\mathbf{X}_i).$$

It can be seen that this estimate has a smaller variance than the original estimate as follows:

$$\begin{aligned}
\sigma_g^2 &= \int \left(g(\mathbf{x})-I\right)^2 \rho(\mathbf{x})dx \\
&= \int g^2(\mathbf{x})\rho(\mathbf{x})dx - I^2 \\
\sigma_{\bar{g}}^2 &= \int \left(\bar{g}(\mathbf{x})-I\right)^2 \rho(\mathbf{x})dx \\
&= \int \bar{g}^2(\mathbf{x})\rho(\mathbf{x})dx - I^2 \\
&\leq \int g^2(\mathbf{x})\rho(\mathbf{x})dx - I^2 = \sigma_g^2
\end{aligned}$$

where we obtain the last line using Jensen's Inequality $\bar{g}^2 = \left(\mathrm{E}\left(g(\mathbf{X})|\mathcal{F}\right)\right)^2 \leq \mathrm{E}\left(g^2(\mathbf{X})|\mathcal{F}\right)$. Therefore, any averaging procedure yields:

$$\sigma_{\bar{g}}^2 \leq \sigma_g^2.$$

Example: We now demonstrate more concretely how these results can be used in practice. Consider the integral on the unit square:

$$I = \int_0^1\int_0^1 e^{-\frac{(x-0.5)^2+(y-0.5)^2}{2}}dxdy.$$

If the integral were not evaluated on the unit square but on a circle we could find the value exactly. In particular,

$$\begin{aligned}
\tilde{I}(R) &= \int_{x^2+y^2<R^2} e^{-\frac{x^2+y^2}{2}}dxdy \\
&= \int_0^{2\pi}\int_0^R re^{-\frac{r^2}{2}}drd\theta \\
&= 2\pi\left(1-e^{-\frac{R^2}{2}}\right).
\end{aligned}$$

The integrand can be integrated explicitly for $(x-0.5)^2+(y-0.5)^2 < 0.5^2$. Thus we can obtain an integral which evaluates to I by replace $g(x,y)$ by the "partially averaged" integrand with

$$I = \int_0^1\int_0^1 \bar{g}(x,y)dxdy$$

where

$$\bar{g}(x,y) = \begin{cases} \frac{4}{2\pi}2\pi\left(1-e^{-\frac{0.5^2}{2}}\right) & \text{if } (x-0.5)^2+(y-0.5)^2<0.5^2 \\ e^{-\frac{x^2+y^2}{2}} & \text{otherwise} \end{cases}.$$

Note we obtained the first expression by computing the average $\frac{1}{2\pi 0.5^2}\int_{(x-0.5)^2+(y-0.5)^2<0.5^2} g(x,y)dxdy$. *Exercise:* Compute the reduction in variance that occurs by using the method of partial averaging in this example.

Importance Sampling

Importance sampling is concerned with the choosing $\rho(x)$ for the random variates X_i so that regions which contribute significantly to the expectation of $g(X)$ are sampled with greater frequency. Thus regions where $f(x)$ is large should be sampled more frequently than those regions where $f(x)$ is comparatively very small. A practical consideration in choosing $\rho(x)$ is that the distribution be amenable to efficient generation of the pseudo-random variates.

In the case that $f(x) > 0$ a probability density always exists which gives a Monte-Carlo method having zero error, $\sigma_g = 0$. This is obtained by considering the definition of g:

$$g(x) = \frac{f(x)}{\rho(x)}.$$

From this it follows immediately that $\rho(x) = \frac{f(x)}{I}$ gives a $\sigma_g^2 = 0$. In general, efficiently generating variates with such a density requires I, which if already known would preclude performing the Monte-Carlo estimate in the first place. However, this result suggests a strategy to reduce the variance. In particular, one should try to choose a density which approximates $f(x)/I$ as close as possible.

Example: Let us consider estimating the integral

$$I = \int_{-4}^{4} \frac{e^{-x^2}}{(|x^2 - 1| + 0.01)^{1/2}} dx.$$

To efficiently estimate the integral using the Monte-Carlo method we see that one approach is to try to sample as closely as possible with a probability density approximating the optimal density $\rho_0(x) = f(x)/I$. We shall now discuss two Monte-Carlo estimates of the integral. In the first we use a standard Gaussian to sample. In the second we use a mixture probability density consisting of a linear combination of Gaussian densities.

Sampling with a Standard Gaussian:
If the Monte-Carlo method uses the density $\rho_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, the prefactor is $\sigma_g = 2.2853$. See Figure 1 for a plot of the density $rho_1(x)$ and the optimal density $\rho_0(x)$. From the plot we see that the Monte-Carlo estimate should be improved if the samples are distributed with higher frequency at the two peaks and in the region where $f(x)$ is non-negligible.

Sampling with a Gaussian Mixture (Stratified Sampling):
We now discuss a probability distribution which generates samples at the two peaks and in the regions where $f(x)$ is non-negligible more frequently. We can readily generate any random variates which have a density which is given as a linear combination of probability densities of form:

$$\rho_{\text{mix}}(x) = \sum_j \alpha_i \rho_j(x).$$

where $\sum_j \alpha_j = 1$. This is done by dividing the interval into subintervals of size α_j and generating the uniform variate $U \in [0, 1]$. If U falls in the j^{th} interval we generate Y_j with density ρ_j and let $X = Y_j$. Thus at the expense of generating an additional uniform variate we can easily generate variates with multi-modal probability distributions.

For the function above we use for example the density with $\sigma_1^2 = 1/100$, $\sigma_2^2 = 1$, $\sigma_3^2 = 1/100$

$$\rho_2(x) = \alpha_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-1)^2}{2\sigma_1^2}} + \alpha_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x+1)^2}{2\sigma_2^2}} + \alpha_3 \frac{1}{\sqrt{2\pi\sigma_3^2}} e^{-\frac{x^2}{2\sigma_3^2}}.$$

with $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\alpha_3 = 0.8$. Sampling with this density we have $\sigma_g = 1.2184$.

For a plot of the optimal density $\rho_0(x)$, Gaussian density $\rho_1(x)$ and mixture density $\rho_2(x)$, see figure 1. We find that the second method will converge with the prefactor σ_g about half that of the first method. Since the rate of convergence is the inverse to the square root of N we find that the second method requires only $1/2^2 = 1/4$ the number of samples as the first method for a comparable accuracy. We remark that even though the second method required generating the extra uniform random variate, this was more than made up for in the reduction in variance. We remark that the generation of the uniform variate could in principle we done away with in the calculation above by simply portioning a fraction of the samples according to the weights α_i. If this modification is made then the approach falls within the class of methods referred to as "stratified sampling".

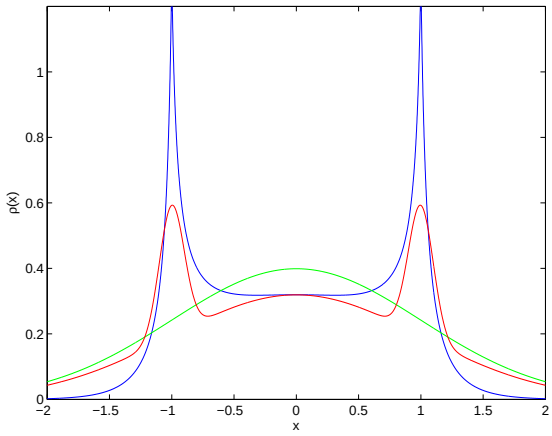


Figure 1: Importance Sampling

Control Variates

The Monte-Carlo method estimates the mean of a random variable. One approach to improving the effective variance that arises in the estimate is to leverage knowledge about the statistics of a random variable which shares similarities to the random variable being sampled. In particular, if the mean of random variable $H(\mathbf{X})$ is known, this can be used to reduce the variance of the estimate of $G(\mathbf{X})$. Let $\mu_H = \mathrm{E}\left(H(\mathbf{X})\right)$ and $I = \mu_G = \mathrm{E}\left(G(\mathbf{X})\right)$ then

$$I \quad = \quad \mathrm{E}\left(G(\mathbf{X}) - \alpha[H(\mathbf{X}) - \mu_H]\right)$$

where α is a parameter to be chosen for the method. This gives the estimate:

$$I \quad \approx \quad \alpha\mu_H + \frac{1}{N}\sum_{i=1}^{N}\left(G(\mathbf{X}_i) - \alpha H(\mathbf{X}_i)\right).$$

The variance of the estimate σ^2 can be computed as:

$$\sigma^2 \quad = \quad \sigma_g^2 + \alpha^2\sigma_H^2 - 2\alpha\mathrm{Cov}(G,H).$$

Since the variance governs the convergence of the estimate, the α parameter should ideally be set to make σ as small as possible. The optimal value α^* is:

$$\alpha^* \quad = \quad \frac{\mathrm{Cov}(G,H)}{\sigma_g^2}.$$

This gives for the minimum variance:

$$\sigma^2 \quad = \quad \sigma_g^2(1-\rho_{G,H}^2)$$

where $\rho_{G,H} = \mathrm{Cov}(G,H)/\sigma_G\sigma_H$ is the correlation between G, H. An important point is that the method of control variates will improve the estimate (reduce σ^2) provided $\rho_{G,H} \neq 0$. For example if H is independent of G, $\rho_{G,H} = 0$ and no improvement is made. Furthermore, we find that if H has correlation $\rho_{G,H} = \pm 1$ the estimate is exact. Intuitively, this suggests that if we choose an H which "behaves very similar" to G the variance will be substantially improved. In order to use the method in practice requires we address a few issues.

In practice, we will typically not be able to determine exactly the optimal parameter value α^*, since this requires knowledge of the variance of the random variable G whose statistics we are estimating in the first place. Also the exact value of $\rho_{G,H}$ will likely not be known. Nonetheless, the control variate strategy can be carried out in practice by estimating the unknown parameter α as we go along from the samples generated. Let us make the following definitions:

$$\begin{aligned}
\widetilde{\sigma}_H^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(H(\mathbf{X}_i)-\mu_H\right)^2 \\
\tilde{I}^{(1)} &= \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{X}_i) \\
\widetilde{C}_{G,H} &= \frac{1}{N}\sum_{i=1}^{N}(G(\mathbf{X}_i)-\tilde{I}^{(1)})(H(\mathbf{X}_i)-\mu_H).
\end{aligned}$$

We remark that all of the estimated quanta that μ_H is known by the assumptions above. This gives the estimate at the N^{th} stage:

$$\begin{aligned}
\tilde{\alpha}^* &= \frac{\widetilde{C}_{G,H}}{\sigma_G^2} \\
\tilde{I} &= \tilde{I}^{(1)} - \tilde{\alpha}^*\frac{1}{N}\sum_{i=1}^{N}(H(\mathbf{X}_i)-\mu_H).
\end{aligned}$$

Example: Let us consider a random variable S_T which models an economic asset having an the effective return over a period T given by a Gaussian with mean m_0 and variance σ_0^2. The value of the asset at time T is then:

$$S_T \quad = \quad S_0 e^{(\sigma_0 X + m_0)T}.$$

The asset has the expected final value:

$$\begin{aligned}
I &= \mathrm{E}\left(S_T\right) \\
&= \mathrm{E}\left(G(X)\right)
\end{aligned}$$

where $G(X) = S_0 e^{(\sigma_0 X + m_0)}$.

To apply the method of control variates we need to choose a random variable H for which we know the mean and which has positive correlation with G. One natural approach to obtaining such a random variable is to consider a Taylor expansion of $G(X)$ about a value X_0 occurring with high probability. Let H be the truncated Taylor expansion for $G(X)$:

$$H(X) \quad = \quad S_0\left(1+(\sigma_0 X+m_0)T+\frac{1}{2}(\sigma_0 X+m_0)^2T^2\right).$$

The mean can then be readily computed by:

$$\mu_H = \mathrm{E}\left(H(X)\right) = S_0\left(1 + m_0 T + \frac{1}{2}(\sigma_0^2 + m_0^2)^2 T^2\right).$$

The direct Monte-Carlo estimate of I when $S_0 = 1$, $\sigma_0 = 0.5$, $m_0 = 0.05$ and $T = 2$ is $\sigma_G^2 \approx 7.0048$. When using the method of control variates for the above choice of H we obtain $\sigma^2 \approx 0.9825$. This give a reduction in the variance of about 88%. This means that the estimate using the control variate will only require about 1/10 the number of samples to obtain a comparable accuracy as the direct Monte-Carlo method.

References

[1] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, 1950

[2] D. P. Landau and K. Binder, *A Guide to Monte-Carlo Simulations in Statistical Physics*, Cambridge University Press, 2000

[3] S. Ross, *Stochastic Processes*, Wiley Series in Probability and Mathematical Statistics, 1996

[4] W. H. Press and Saul A. Teukolsky and W. T. Vetterling and B. P. Flannery, *Numerical Recipes*, Cambridge University Press, 2002

[5] Cai, D., *Computational Finance Notes*, http://www.cims.nyu.edu, 2004.

[6] Goodman, , *Finance Class Notes*, http://www.cims.nyu.edu, 2003.