Variance Reduction

Zdravko Botev
University of New South Wales
Sydney, Australia

Ad Ridder
Vrije University
Amsterdam, Netherlands

Abstract

Increased computer speed and memory have encouraged simulation analysts to develop ever more realistic stochastic models. Despite these advancements in computing hardware, the most significant gains in the speed of stochastic simulation are still the result of methodological advances and clever algorithmic design. Here we survey a few of the most important methods for variance reduction and speedup that will benefit any simulation, no matter what the capabilities of the computing hardware.

Keywords: control variables; stratified sampling; importance sampling; quasi Monte Carlo; multilevel Monte Carlo

1. Preliminaries

In this exposition we assume that we wish to compute an expectation $\ell = \mathbb{E}[H(\mathbf{X})]$, in which $\mathbf{X} : \Omega \mapsto \mathcal{D} \subset \mathbb{R}^d$ is the input random d-dimensional vector, $H : \mathcal{D} \mapsto \mathbb{R}$ is the model of interest, and $Y = H(\mathbf{X})$ is the output. Note that we allow $\mathbf{X}$ to be a finite (random) sample path of a stochastic process. In the crude Monte Carlo method (**stat03619**) one draws i.i.d. replications $\mathbf{X}_i$, $i = 1, \ldots, n$, of $\mathbf{X}$ and computes the sample average of the outputs,

$$\widehat{Y}_n \stackrel{\text{def}}{=} \frac{H(\mathbf{X}_1) + \cdots + H(\mathbf{X}_n)}{n}, \tag{1}$$

as an unbiased estimator of ℓ. The performance of the Monte Carlo estimator is expressed statistically by its variance $\mathrm{Var}[\widehat{Y}_n] = \frac{1}{n}\mathrm{Var}\big[H(\mathbf{X})\big]$, and computationally by its computational cost $t_1 n$, where t_1 is the cost of evaluating $H(\cdot)$.

We are interested in constructing an unbiased estimator $\widehat{\ell}$ of ℓ, that performs better than the crude Monte Carlo estimator. This means that, when both estimators require the same computational cost, then the variance of $\widehat{\ell}$ is smaller, $\mathrm{Var}[\widehat{\ell}] \leq \mathrm{Var}\big[\widehat{Y}_n\big]$. Variance reduction is typically needed in simulations of highly complex models with high computation cost per replication, and in simulations of rare events (**stat07823**) with too many replications in crude Monte Carlo. Many applications of these phenomena in a variety of fields, to name a few, finance, statistics, reliability, systems biology, power networks, can be found in reference[1].

Classic variance reduction techniques are the methods using antithetic variates (**stat05006**), control variates, conditional Monte Carlo, stratified sampling (**stat05999**), and importance sampling[2,3,5,4,6] (**stat05402**). Various variations and combinations of these methods have been developed, such as Latin hypercube sampling (**stat03803**), weighted Monte Carlo, model reduction, and moment matching. Many new ideas from engineering, statistics and physics, have resulted recently in new algorithms. In this short survey we will describe three of these classic techniques, and then we will conclude with a quick glance of modern methods.

2. Classic Variance Reduction Techniques

2.1. Control Variate Method

Recall the Monte Carlo estimator $\widehat{Y}_n$ for estimating the expected model output $\ell = \mathbb{E}[Y]$, where $Y = H(\mathbf{X})$. A random variable C is called a control variate for Y if (a) its expectation $\mu \stackrel{\text{def}}{=} \mathbb{E}[C]$ is known or readily available, (b) Y and C are correlated, and (c) C can be easily simulated. We then define the *control variate estimator*:

$$\widehat{\ell} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \alpha(C_i - \mu)\big) = \widehat{Y}_n - \alpha(\overline{C}_n - \mu), \tag{2}$$

where α is a scalar parameter to be determined. It is easy to prove that the variance of the control variate estimator is minimized by $\alpha^* \overset{\text{def}}{=} \frac{\text{Cov}(Y,C)}{\text{Var}[C]}$. The resulting minimal variance is

$$\text{Var}[\widehat{\ell}] = (1-\rho^2)\text{Var}[\widehat{Y}_n] < \text{Var}[\widehat{Y}_n], \tag{3}$$

where ρ denotes the correlation coefficient between Y and C. Since $\text{Cov}(Y,C)$ is usually not available, the optimal control coefficient α^* must be estimated by executing a pilot run. Estimation of α^* implies that the variance reduction becomes smaller than equation (3) suggests, and that the estimator may become biased.

The method can be easily extended to a *multiple control variable estimator*, where a single replication of the estimator is: $Y - \sum_{k=1}^m \alpha_k(C_k - \mu_k)$, where the optimal control coefficients α_k follow from a linear regression optimization[7,8] (**stat03209**).

2.2. *Stratified and Conditional Sampling*

Suppose now that there is some finite random variable Z taking values from $\{z_1,\dots,z_m\}$, such that (a) the probabilities $p_k \overset{\text{def}}{=} \mathbb{P}(Z = z_k)$ are known, and (b) for any $k = 1,\dots,m$, it is easy to sample from the conditional distribution of $\mathbf{X}$ given $Z = z_k$. We can then write

$$\ell = \mathbb{E}\big[\mathbb{E}[H(\mathbf{X})]\big|Z\big] = \sum_{k=1}^m p_k \mathbb{E}[H(\mathbf{X})|Z = z_k]. \tag{4}$$

The sets $\{Z = z_k\}$, $k = 1,\dots,m$ are called strata, and they form a partition of the sample space Ω. Let us define the conditional strata estimators as sample averages of the conditional outputs:

$$\widehat{Y}^{(k)}_{n_k} \overset{\text{def}}{=} \frac{1}{n_k}\sum_{i=1}^{n_k} H(\mathbf{X}_{ki}), \quad k = 1,\dots,m,$$

where $\mathbf{X}_{k1},\dots,\mathbf{X}_{kn_k}$ are i.i.d. samples simulated from the conditional distribution of $\mathbf{X}$ given $Z = z_k$. Now we set the strata sample sizes $n_1 + \cdots + n_m = n$, and define the *stratified estimator*:

$$\widehat{\ell} \overset{\text{def}}{=} \sum_{k=1}^m p_k \widehat{Y}^{(k)}_{n_k}.$$

From equation (4) it readily follows that the estimator is unbiased. Assuming proportional assignment of the strata sample sizes, that is, $n_k = p_k n$, the variance is computed as follows

$$\begin{aligned}\text{Var}[\widehat{\ell}] &= \sum_{k=1}^m p_k^2 \frac{1}{n_k}\text{Var}[H(\mathbf{X})|Z = z_k] = \frac{1}{n}\sum_{k=1}^m p_k \text{Var}[H(\mathbf{X})|Z = z_k] \\ &= \frac{1}{n}\mathbb{E}\text{Var}[H(\mathbf{X})|Z] \le \frac{1}{n}\text{Var}[H(\mathbf{X})] = \text{Var}[\widehat{Y}_n].\end{aligned}$$

The inequality follows from a well-known identity in Probability Theory—for any two random variables X, Z, it holds:

$$\text{Var}[X] = \mathbb{E}\big[\text{Var}[X|Z]\big] + \text{Var}\big[\mathbb{E}[X|Z]\big] \ge \max\{\text{Var}\big[\mathbb{E}[X|Z]\big],\ \mathbb{E}\big[\text{Var}[X|Z]\big]\}. \tag{5}$$

It can be shown that the strata sample sizes n_k that minimize the variance are[3,4]

$$n_k = n\frac{p_k\sigma_k}{\sum_{j=1}^m p_j\sigma_j},$$

where $\sigma_k^2 \overset{\text{def}}{=} \text{Var}[H(\mathbf{X})|Z = z_k]$. A practical problem is that the standard deviations σ_k are usually not available, so these variances are estimated by pilot runs. Finally, we also observe that (5) can be leveraged for variance reduction. In particular, if $\mathbf{Z}$ is any random vector with distribution G, ideally strongly dependent on $\mathbf{X}$, then the *conditional estimator*:

$$\widehat{\ell} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[H(\mathbf{X})|\mathbf{Z}_i], \qquad \mathbf{Z}_1,\dots,\mathbf{Z}_n \text{ are iid draws from } G,$$

will always have smaller variance than the crude Monte Carlo estimator. The key difficulty here is finding a vector (or scalar) $\mathbf{Z}$ such that the conditional expectation $\mathbb{E}[H(\mathbf{X})|\mathbf{Z} = \mathbf{z}]$ is available as a simple-to-evaluate formula.

2.3. Importance Sampling

Recall the law $\mathbb{P}$ on the range space $\mathcal{D} \subset \mathbb{R}^d$ of the random vector $\mathbf{X}$. Let $F : \mathbb{R}^d \mapsto [0, 1]$ be the distribution function of $\mathbf{X}$, defined by

$$F(\mathbf{x}) \overset{\text{def}}{=} \mathbb{P}\big(\mathbf{X} \in \prod\nolimits_{j=1}^{d} (-\infty, x_j]\big),$$

for any $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Then, we compute probabilities and expectations as integrals with respect to the Lebesque-Stieltjes measure associated with the distribution function F, for instance,

$$\ell = \mathbb{E}[H(\mathbf{X})] = \int_{\mathbb{R}^d} H(\mathbf{x})\, dF(\mathbf{x}).$$

Note that probabilities, say $\mathbb{P}(A), A \subset \mathcal{D}$, are computed in this way by setting $H(\mathbf{x}) = \mathbb{I}_A(\mathbf{x})$, the indicator function of the set A. The importance sampling method of this section is specifically applicable in settings where A is a rare event (**stat04405**), that is, having very small probability, say $\mathbb{P}(A) \approx 10^{-9}$.

Now, suppose that G is a proper distribution function on $\mathbb{R}^d$, with associated probability measure $\mathbb{Q}$ defined by $\mathbb{Q}(A) = \int_A dG(\mathbf{x})$, such that (a) $\mathbb{Q}(\mathcal{D}) = 1$, and (b) $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, which means that $\mathbb{Q}(A) > 0$ whenever $\mathbb{P}(A) > 0$. According to the Radon-Nikodym Theorem (**stat02331**), it holds that there is a function $L : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ such that for any Borel set A,

$$\int_A dF(\mathbf{x}) = \int_A L(\mathbf{x})\, dG(\mathbf{x}). \tag{6}$$

The function L is called the likelihood ratio and usually written as $L(\mathbf{x}) = dF(\mathbf{x})/dG(\mathbf{x})$; the alternative probability measure $\mathbb{Q}$ is said to be the importance sampling probability measure, or the change of measure (**stat04560**). We then define the *importance sampling estimator*:

$$\widehat{\ell} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} H(\mathbf{X}_i) L(\mathbf{X}_i),$$

where i.i.d. $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are generated from the importance sampling distribution G.

Using equation (6) it easy to see that the estimator is unbiased. More importantly, variance reduction is obtained when the change of measure has been chosen properly, as will be explained below.

2.3.1. Variance Analysis and Reduction.

We denote expectations and variances with respect to the importance sampling distribution by the subscript G. Thus, the variance of the importance sampling estimator satisfies

$$\mathrm{Var}_G[\widehat{\ell}] = \frac{1}{n}\mathrm{Var}_G[H(\mathbf{X})L(\mathbf{X})] = \frac{1}{n}\Big(\mathbb{E}_G\big[H^2(\mathbf{X})L^2(\mathbf{X})\big] - \ell^2\Big),$$

and noting that the variance of the Monte Carlo estimator equals

$$\mathrm{Var}[\widehat{Y}_n] = \frac{1}{n}\mathrm{Var}[H(\mathbf{X})] = \frac{1}{n}\Big(\mathbb{E}\big[H^2(\mathbf{X})\big] - \ell^2\Big),$$

we see that variance reduction is obtained if and only if the second moment satisfies $\mathbb{E}_G\big[H^2(\mathbf{X})L^2(\mathbf{X})\big] \leq \mathbb{E}\big[H^2(\mathbf{X})\big]$. Applying again equation (6), we achieve the necessary and sufficient condition

$$\mathbb{E}\big[H^2(\mathbf{X})L(\mathbf{X})\big] \leq \mathbb{E}\big[H^2(\mathbf{X})\big]. \tag{7}$$

The interpretation of this condition leads to the heuristic of obtaining variance reduction by importance sampling: wherever $H^2(\mathbf{x})$ is large, $L(\mathbf{x})$ should be small. Specifically, for rare event probability estimation, that is, $H(\mathbf{x}) = \mathbb{I}_A(\mathbf{x})$, this heuristic says to bias the set A with more probability mass. However, careless shifting of probability mass may lead to overbiasing resulting in bad underestimation[9].

The best one can could do is to minimize the left-hand side of equation (7) with respect to suitable distributions G. The optimal solution is[10]

$$dG^*(\mathbf{x}) \overset{\text{def}}{=} \frac{|H(\mathbf{x})| dF(\mathbf{x})}{\mathbb{E}\big[|H(\mathbf{X})|\big]}. \tag{8}$$

When $H \geq 0$, for instance in case of probability estimation, it follows that $\mathbb{E}\big[H^2(\mathbf{X})L(\mathbf{X})\big] = \ell^2$, and thus $\mathrm{Var}_G[\widehat{\ell}] = 0$: a single sample suffices! However, the optimal distribution displayed in equation (8) is not implementable because it requires knowledge of the unknown quantity.

Many successful approaches to approximate the optimal importance sampling change of measure have been studied. Notably, we mention large deviations[11] (**stat04568**); Lyapunov functions and bounds[12], approximation of the zero-variance distribution[13]; optimal exponential change of measure[14]; minimization of the Kullback-Leibler distance to the zero-variance distribution[15] (cross-entropy method).

3. Modern Variance Reduction Techniques

3.1. Quasi Monte Carlo

Frequently, it is possible to find a functional composition $\hbar : \mathcal{D} \mapsto \mathbb{R}$ that first maps the uniformly distributed vector $\mathbf{U} \sim \mathsf{U}(0,1)^d$ into $\mathbf{X} \sim F$ and then maps the resulting $\mathbf{X}$ into $H(\mathbf{X})$. With such a map we can write $\ell = \mathbb{E}[H(\mathbf{X})] = \mathbb{E}[\hbar(\mathbf{U})]$, and redefine the corresponding unbiased estimator (1) as $(\hbar(\mathbf{U}_1) + \cdots + \hbar(\mathbf{U}_n))/n$. As already seen, such an estimator has standard error that decays at the rate $\mathcal{O}(n^{-1/2})$. This convergence rate can frequently be improved to $\mathcal{O}(n^{-1/2-\delta}),\ \delta > 0$ by using a quasirandom or low-discrepancy pointset $\mathcal{V}_n = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ (**stat02329**) that fills the unit d-hypercube, $(0,1)^d$, in a much more regular and uniform fashion than n independent copies of the pseudorandom vector $\mathbf{U} \sim \mathsf{U}(0,1)^d$. The so-called *Halton* and *Sobol* pseudorandom pointsets[2,5] are the most widely used, and can be constructed readily in almost all software. Given such a pointset, we can then deliver the quasi Monte Carlo (QMC) approximation: $q(\mathcal{V}_n) = (\hbar(\mathbf{v}_1) + \cdots + \hbar(\mathbf{v}_n))/n$. Since the pointset $\mathcal{V}_n$ fills the cube $(0,1)^d$ in a regular and deterministic fashion, q is not a statistical estimator. However, it can be modified to yield an unbiased statistical estimator via m (usually $m \ll n$) independent random shifts of $\mathcal{V}_n$. This strategy gives the *randomized QMC* estimator:

$$\widehat{\ell} = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\hbar(\mathbf{V}_{j,k}), \quad \text{where } \mathbf{V}_{j,k} \stackrel{\text{def}}{=} \{(\mathbf{v}_j + \mathbf{U}_k) \mod 1\}.$$

The standard error of the randomized QMC is $\mathcal{O}(m^{-1/2}n^{-1/2-\delta})$, and frequently compares very favorably with the standard error of the crude Monte Carlo estimator (1) that uses $m \times n$ iid copies of $\mathbf{X}$ (instead of n). The size of the gain $\delta > 0$ depends very much on the structural properties of the map $\hbar$, with greater smoothness and regularity generally resulting in larger δ.

3.2. Multilevel Monte Carlo Estimator

Consider again the control variate estimator (2) with n_2 independent samples (instead of n). Suppose that condition (a) is relaxed, so that $\mu = \mathbb{E}[C]$ is no longer available in closed form, and condition (c) is strengthened so that a copy of C is $m > 1$ times cheaper to evaluate and simulate than a copy of Y. Then, we may still apply (2), but with the unknown μ necessarily replaced by an unbiased estimator $\widehat{\mu}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\widetilde{C}_i$, where $\widetilde{C}_1, \ldots, \widetilde{C}_{n_1}$ are iid copies of C, and independent from $(Y_1, C_1), \ldots, (Y_{n_2}, C_{n_2})$. This yields the *two-level Monte Carlo* estimator:

$$\widehat{\ell} = \frac{1}{n_2}\sum_{i=1}^{n_2}\big(Y_i - \alpha(C_i - \widehat{\mu}_{n_1})\big) = \frac{\alpha}{n_1}\sum_{i=1}^{n_1}\widetilde{C}_i + \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \alpha C_i),$$

where n_1 and n_2 are yet to be determined, given an overall computational budget $n = n_1 + n_2$. Since $n_2\mathrm{Var}[\widehat{\ell}] = \alpha^2\frac{n}{n_1}\mathrm{Var}[C] - 2\alpha\mathrm{Cov}(Y, C) + \mathrm{Var}[Y]$, minimizing for α yields:

$$\mathrm{Var}[\widehat{\ell}] = \left(\frac{n_1}{n_2}(1 - \rho^2) + 1\right)\mathrm{Var}[\widehat{Y}_n] \geq \mathrm{Var}[\widehat{Y}_n],$$

and initially it appears that we have lost efficiency. However, this is not the case if we account for both variance and computing cost. If the cost of each copy of C is one unit of time, then the cost of computing $\widehat{\ell}$ is $n_1 + (m+1)n_2 = n + mn_2$ units, and the cost of computing $\widehat{Y}_n$ is $m \times n$ units of time (or work). Therefore, the ratio of the work normalized variances is ($r = n_1/n_2$):

$$\frac{\mathrm{Var}[\widehat{\ell}] \times \mathrm{Cost}(\widehat{\ell})}{\mathrm{Var}[\widehat{Y}_n] \times \mathrm{Cost}(\widehat{Y}_n)} = \frac{\mathrm{Var}[\widehat{\ell}](n_1 + (m+1)n_2)}{\mathrm{Var}[\widehat{Y}_n]m(n_1 + n_2)} = \frac{(1 + r + m)(1 + r(1 - \rho^2))}{m(r+1)},$$

which (ignoring any negligible rounding errors due to n_1 and n_2 being integers) is minimized for $r^* = \sqrt{\rho^2 m/(1-\rho^2)} - 1$, giving the minimum ratio of work normalized variances:

$$\left(|\rho| + \sqrt{m(1-\rho^2)}\right)^2 \Big/ m$$

This is of the order $1/m$, when C and Y are highly correlated. Note how this efficiency gain is achieved by spending the bulk of the computing budget on simulating copies of the cheaper variable C to estimate $\alpha\mathbb{E}[C]$, with the remainder of the budget being expended on simulating the more costly variable Y, in order to estimate the correction $\mathbb{E}[Y - \alpha C]$. Thus, in allocating different computing budgets, n_1 and n_2, to two highly correlated variates, the two-level Monte Carlo estimator combines the idea of stratified sampling and control variable variance reduction.

Just like control variable estimation is generalized to multiple control variable estimation, the two-level estimation can be generalized to multilevel Monte Carlo[16] estimation.

References

[1] Rubino G. and Tuffin, B. (Eds.). (2009). *Rare event simulation using Monte Carlo methods*, Wiley, Chichester, UK.

[2] Asmussen, S. and Glynn, P.W. (2007). *Stochastic Simulation*, Springer-Verlag, New York.

[3] Fishman, G.S. (1996). *Monte Carlo: concepts, algorithms, and applications*, Springer-Verlag, New York.

[4] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York.

[5] Kroese, D.P., Taimre, T. and Botev, Z.I. (2011). *Handbook of Monte Carlo Methods*, Wiley, Hoboken, New Jersey.

[6] Rubinstein, R.Y. and Kroese, D.P. (2017). *Simulation and the Monte Carlo method*, 3rd ed., Wiley, Hoboken, New Jersey.

[7] Lavenberg, S.S. and Welch, P.D. (1981). A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science* **27**, 322-335.

[8] Rubinstein, R.Y. and Marcus, R. (1985). Efficiency of multivariate control variates in Monte Carlo simulation. *Operations Research* **33**, 661-677.

[9] Smith, P.J. (2001). Underestimation of rare event probabilities in importance sampling simulations. *Simulation* **76**(3), 140-150.

[10] Kahn, H. and Marshall, A. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America* **1**(5), 263-278.

[11] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations and differential games. *Stochastics and Stochastic Reports* **76**(6), 481-508.

[12] Blanchet, J., Glynn, P. and Leder, K. (2012). On Lyapunov inequalities and subsolutions for efficient importance sampling. *ACM Transactions on Modeling and Computer Simulation* **22**(3), 13:1-13:27.

[13] L'Ecuyer, P and Tuffin, B. (2011). Approximating zero-variance importance sampling in a reliability setting. *Annals of Operations Research* **189**(1), 277-297.

[14] Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics* **4**(4), 673-684.

[15] Kroese, D.P., Rubinstein, R.Y., Cohen, I., Porotsky, S. and Taimre, T. (2013). Cross-entropy method. In *Encyclopedia of Operations Research and Management Science*, S.I. Gass and M.C. Fu, eds, Springer, Boston, pp. 326-333.

[16] Giles, Michael B. (2008). Multilevel monte carlo path simulation, *Operations Research*, **56**(3), 607-617.