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

Chapter 3: Variance Reduction.

1 Introduction

Variance reduction is the search for alternative and more accurate estimators of a given quantity. The possibility of variance reduction is what separates Monte Carlo from direct simulation. Simple variance reduction methods often are remarkably effective and easy to implement. It is good to think about them as you wait for a long Monte Carlo computation to finish. In some applications, such as rare event simulation and quantum chemistry, they make practicle what would be impossible otherwise. Most advanced Monte Carlo is some kind of variance reduction.

Among the many variance reduction techniques, which may be used in combination, are *control variates*, *partial integration*, *systematic sampling*, and *importance sampling*. The method of control variates is useful when a crude version of the problem can be solved explicitly. This is often the case in simple problems (possibly the definition of "simple") such pricing problems in quantitative finance where the crude solvable version could be Black Scholes. Partial integration, also called *Rao Blackwellization* lowers variance by replacing integrals over come variables or over parts of space by their averages. Systematic sampling methods range from the simplest, *antithetic variates*, to the slightly more sophisticated *stratified sampling*, to *quasi Monte Carlo* integration. Importance sampling has appeared already as sampling with a weight function. It also is the basis of reweighting and score function strategies for sensitivity analysis. Methods for *rare event sampling* mostly use importance functions, often suggested by the mathematical theory of *large deviations*.

2 Control variates

Suppose X is a random variable and that we want to evaluate

$$A = E[V(X)] \; .$$

We may estimate A by generating L independent samples of X and taking

$$\widehat{A} = \frac{1}{L} \sum_{k=1}^{L} V(X_k) \; . \tag{1}$$

The error is of the order of

$$\widehat{A} - A \sim \frac{\sigma_V}{\sqrt{L}} \;\; , \quad \sigma_V^2 = E\left[\left(V(X) - A \right)^2 \right] \; .$$

Thus, the number of samples and the run time needed to achieve a given accuracy is inversely proportional to the variance.

A *control variate* is an easily evaluated random variable, $W(X)$, so that $B = E[W(X)]$ is known. If $W(X)$ is correlated with $V(X)$ with covariance

$$C_{VW} = \mathrm{cov}(V, W) = E\left[(V - A)(W - B)\right] \ ,$$

then the random variable,

$$Z = V(X) - \alpha\left(W(X) - B\right) \ , \tag{2}$$

can have less variance than $V(X)$. This will make the control variate estimator

$$\widehat{A} = \frac{1}{L}\sum_{k=1}^{L}\left(V(X_k) - \alpha W(X_k)\right) + \alpha B \tag{3}$$

more accurate than the simple one (1), often dramatically so.

We choose α to minimize the variance of Z in (2). The variance is

$$\sigma_Z^2 = \sigma_V^2 - 2\alpha C_{VW} + \alpha^2\sigma_W^2 \ .$$

The optimal α is

$$\alpha^* = \frac{C_{VW}}{\sigma_W^2} \ , \tag{4}$$

and the corresponding optimal variance is

$$\sigma_Z^2 = \sigma_V^2 - \frac{C_{CW}^2}{\sigma_W^2} = \sigma_V^2\left(1 - \rho_{VW}^2\right) \ , \tag{5}$$

in terms of the correlation coefficient

$$\rho_{VW} = \mathrm{corr}(V, W) = \frac{C_{VW}}{\sigma_V\sigma_W} \ .$$

Thus, the quality of W as a control variate depends on the correlation between V and W.

In practice it is not likely that one would know the optimal α (4) in advance, but it can be estimated from Monte Carlo data. From the samples X_k we can evaluate $V_k = V(X_k)$ and $W_k = W(X_k)$, then

$$\begin{aligned}
\widehat{\sigma_W^2} &= \frac{1}{L}\sum_{k=1}^{L}\left(W_k - B\right)^2 \ , \\
\widehat{A}^{(1)} &= \frac{1}{L}\sum_{k=1}^{L} V_k \qquad \text{(simple estimator of } A\text{)}, \\
\widehat{C_{VW}} &= \frac{1}{L}\sum_{k=1}^{L}\left(V_k - \widehat{A}^{(1)}\right)\left(W_k - B\right) \ ,
\end{aligned}$$

$$\widehat{\alpha^*} = \frac{\widehat{C_{VW}}}{\widehat{\sigma^2_W}} , \tag{6}$$

$$\widehat{A} = \widehat{A^{(1)}} - \widehat{\alpha^*} \frac{1}{L} \sum_{k=1}^{L} (W_k - B) \ .$$

The estimate (6) may not be a very accurate estimate of (4), but the perfomance does not depend strongly on α when α is close α^*, where the derivative is zero.

One can use more than one control variate. Given $W_1(X)$, ..., $W_n(X)$, we can form

$$Z = V(X) - \sum_{l=1}^{n} \alpha_l W_l(X) \ . \tag{7}$$

The optimal coefficients, the α_l that minimize $\mathrm{var}(Z)$, are found by solving the system of linear equations

$$\mathrm{cov}(V, W_l) = \sum_{l=1}^{n} \mathrm{cov}(W_l, W_m)\alpha_m \ . \tag{8}$$

Should the coefficients in (8) be unknown, we can estimate them from Monte Carlo data as above.

Let $V_{\mathcal{S}}$ denote the control variate sum on the right of (7) so that $V = Z + V_{\mathcal{S}}$. The optimality conditions for the coefficients α_l is that $V_{\mathcal{S}}$ be uncorrelated with Z. If this were not so, we could use $W = V_{\mathcal{S}}$ as an additional control variate and further reduce the variance. Because they are uncorrelated, $\mathrm{var}(V) = \mathrm{var}(Z) +$. In statisticians' terminology, the total variance of V is the sum of the *explained* part, $\mathrm{var}(V_{\mathcal{S}})$, and the *unexplained* part, $\mathrm{var}(Z)$.

Linear algebra has a geometrical way to express this. Given a random variable, X, there is a vector space consisting of mean zero functions of X with finite variance. If $V(X) - A$ and $W(X) - B$ are two such, their *inner product* is $\langle V - A, W - B \rangle = \mathrm{cov}(V, W)$. The corresponding length is $\|V\|^2 = \langle V - A, V - A \rangle = \mathrm{var}(V)$. In the vector space is the subspace, $\mathcal{S}$, spanned by the vectors $W_l(X) - B_l$. Minimizing $\mathrm{var}(Z)$ in (7) is the same as finding the $V_{\mathcal{S}} \in \mathcal{S}$ that minimizes $\|V - V_{\mathcal{S}}\|^2$. This is the element of $V_{\mathcal{S}}$ closest to $V - A$. In this way we write $V = Z + V_{\mathcal{S}}$ with $V_{\mathcal{S}}$ perpendicular to Z.

Example: From the introduction. Let $B \subset R^3$ be the *unit ball* of points with $|x| \leq 1$. Suppose X and Y are independent and uniformly distibuted in B and try to evaluate

$$E\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] \ .$$

Since the functional $V(X,Y) = \frac{e^{-\lambda|X-Y|}}{|X-Y|}$ depends on $|X - Y|$, we seek control variates that have this dependence, the difficulty being finding functionals whose expected value is known. One possibility is $W_1(X,Y) = |X - Y|^2$ with

$$E[W_1] = E\left[|X|^2\right] + 2E[\langle X, Y \rangle] + E\left[|Y|^2\right] \ .$$

The middle term on the right vanishes because X and Y are independent. The other two each are equal to $\frac{3}{5}$, so $E[W_1] = \frac{6}{5}$. With $\lambda = .2$, the improvement takes us from $\text{var}(V) \approx .99$ to $\text{var}(Z) \approx .72$, about 26% lower. Another possibility is $W_2 = |X - Y|^4$ with $W[W_2] = \frac{6}{7} + \frac{6}{5}$. Using these two control variates together gives $\text{var}(Z) \approx .58$, an almost 50% reduction. The Matlab program that does this, CV1.m, is posted.

This example shows a relatively modest variance reduction from two not very insightful control variates. Variance reduction methods that seem impressive in one dimensional examples may become less effective in higher dimensional problems, as this relatively modest six dimensional problem illustrates.

3 Partial averaging

Partial averaging, or Rao-Blackwellization, reduces variance by averaging ofver some of the variables or over part of the integration domain. for example, suppose (X, Y) is a random variable wtih probability density $f(x, y)$. Let $V(X, Y)$ be a random variable and

$$\widetilde{V}(x) = E[V(X, Y) \mid x] = \frac{\int V(x, y) f(x, y) dy}{\int f(x, y) dy} \tag{9}$$

A simple inequality shows that except in the trivial case where V already was independent of y,

$$\text{var}(\widetilde{V}) < \text{var}(V) \ . \tag{10}$$

In fact, the reader can check that

$$\text{var}(V) = \text{var}(\widetilde{V}) + E\left[\left(V - \widetilde{V}\right)^2\right] \ . \tag{11}$$

The conclusion is that if a problem can be solved partially, if some of the integrals (9) can be computed explicitly, the remaining problem is easier.

A more abstract and general version of of the partial averaging method is that if $\mathcal{G}$ is a sub $\sigma-$algebra and

$$\widetilde{V} = E[V \mid \mathcal{G}] \ ,$$

then we again have (11) and the variance reduction property (10). Of course, the method still depends on being able to evaluate $\widetilde{V}$ efficiently.

Subset averaging is another concrete realization of the partial averaging principle. Suppose B is a subset (i.e., an event) and that $E[V \mid B]$ is known. If

$$\widetilde{V}(x) = \begin{cases} E[V \mid B] & \text{if } x \in B, \\ V(x) & \text{if } x \notin B, \end{cases}$$

then again $\text{var}(V) = \text{var}(\widetilde{V})$ except in trivial situations. For example, we might take B to be the largest set for which $E[V \mid B]$ can be evaluated by symmetry.

Example. Consider just the Y integration in the previous example.:

$$E_Y\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] = \frac{3}{4\pi}\int_{|y|\leq 1}\frac{e^{-\lambda|x-y|}}{|x-y|}dy\ .$$

For each x, define $B_x = \{y \mid |x-y| \leq 1-|x|\}$. This is the largest round ball about x contained in the integration domain $|y| \leq 1$. The conditional expectation

$$E_Y\left[V(x,Y) \mid B_x\right] = \frac{\frac{3}{4\pi}\int_{y\in B_x}\frac{e^{-\lambda|x-y|}}{|x-y|}dy}{P(Y\in B_x)}$$

may be evaluated using radial symmetry. The numerator is

$$\begin{aligned}\frac{3}{4\pi}\int_{r=0}^{1-|x|}\frac{e^{-\lambda r}}{r}4\pi r^2 dr &= 3\int_{r=0}^{1-|x|}e^{-\lambda r}r\,dr \\ &= \frac{3}{\lambda^2}\left(1-e^{-\lambda(1-|x|)}\left(1+\lambda(1-|x|)\right)\right)\ ,\end{aligned}$$

And $P(Y\in B_x) = (1-|x|))^3$, so that

$$\begin{aligned}&E_Y\left[V(x,Y) \mid B_x\right] \\ &= u(x) = (1-|x|))^{-3}\,\frac{3}{\lambda^2}\left(1-e^{-\lambda(1-|x|)}\left(1+\lambda(1-|x|)\right)\right)\ . \qquad (12)\end{aligned}$$

Therefore

$$A = E_{(X,Y)}\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] = E_{(X,Y)}\left[\widetilde{V}(X,Y)\right]\ ,$$

where

$$\widetilde{V}(X,Y) = \begin{cases}\dfrac{e^{-\lambda|X-Y|}}{|X-Y|} & \text{if } |X-Y| \geq 1-|X|\ , \\ u(X) & \text{if } |X-Y| < 1-|X|\ .\end{cases}$$

Computational experiments (Matlab script CV3.m posted) with $\lambda = .2$ show that $\mathrm{var}(\widetilde{V}) \approx .61$. We may further reduce the variance using the earlier control variates $W_1 = |X-Y|^2$ and $W_2 = |X-Y|^4$. Using only W_1 gives $\mathrm{var}(Z) \approx .35$. Using W_1 and W_2 together gives $\mathrm{var}(Z) \approx .24$. Thus, the combined effects of not very sophisticated partial averaging and two simple control variates reduces the variance, and the work needed to achieve a given accuracy, by a factor of 4 (from .99 to .24).

4 Importance sampling and rare events

Evaluating $A = E_f\left[V(X)\right]$ need not mean sampling f. In many situations, particularly those governed by $f-$rare[1] events, much better results come from

[1] This means that the f probability of B is small.

sampling a well chosen different distrubution, g. To get an unbiased estimator,

$$\begin{aligned} A &= \int V(x)f(x)dx \\ &= \int V(x)\frac{f(x)}{g(x)}g(x)dx \\ A &= E_g\left[V(X)L(X)\right] \ . \end{aligned} \tag{13}$$

The ratio $L(x) = f(x)/g(x)$ may be called the *likelihood ratio* or the *importance function* or the *score function*, or the *Radon Nikodym derivative*. Of course we must have $g(x) > 0$ whenever $f(x) > 0$. Estimators using (13) are unbiased for any g. Our task is to find distributions g that we can sample so that

$$\mathrm{var}_g\left[V(X)L(X)\right]$$

is small. Of course, the variance is zero if $V(x)L(x)$ is constant, i.e. $g(x) = V(x)f(x)/A$. Even when V is positive, this requires knowing A in advance. Nevertheless, we take the intuition that a good g will be large where Vf is large although f may not be large there.

Rare event sampling is a special case that illustrates many of the ideas. If B is some event and $V(x) = I_B(x)$ is the *indicator function*[2] (or *characteristic function*), then

$$p = P(B) = E_f\left[I_B(X)\right] = E_g\left[I_B(X)L(X)\right] \ .$$

The standard estimator using f samples would be

$$\widehat{p} = \frac{N}{L} = \frac{1}{L}\sum_{k=1}^{L} I_B(X_k) \ . \tag{14}$$

Here N is the number of *hits*, samples landing in B. The variance of $\widehat{p}$ is $p(1-p)/L$. for small p (rare events), this is nearly the same as p/L.

We should distinguish between absolute and relative accuracy, particularly when estimating small p. The absolute error is $\Delta p = \widehat{p} - p \sim \sigma(\widehat{p}) \approx \sqrt{p/L}$. The relative error is $\Delta p/p$, which could be much larger. For example, if $p = 10^{-5}$ and $\widehat{p} = 10^{-4}$, then the absolute error is about 10^{-4}, which may seem small, but the relative error is about 10. The estimator is off by a factor of ten, not a good result. Note that the relative error is of the order of

$$\frac{\sigma(\widehat{p})}{p} \approx \frac{1}{\sqrt{pL}} = \frac{1}{\sqrt{E\left[N\right]}} \ . \tag{15}$$

That is, the relative error is governed by the expected number of hits. If I want 10% accuracy in $\widehat{p}$, I need to generate about 100 hits, which could mean very many mostly fruitless trials.

[2]This means $I_B(x) = 1$ if $x \in B$ and $I_B(x) = 0$ if $x \notin B$.

The first goal of importance sampling is to make hits more likely. But this is not enough. Often the conditional probability withing B, $f_B(x) = f(x)/P(B)$ for $x \in B$, is sharply peaked within B. Informally, we say that rare events happen in predictable ways. If g is peaked in the wrong parts of B, the resulting variance could be larger than for the naive estimator. To be a good importance sampler, g samples must hit B often, and in roughly the same way f samples that hit B do.

The mathematical study of rare events is the theory of *large deviations*[3]. A typical large deviations theorem says $P_n(B) \sim e^{-Rn}$, where n is some measure of the problem size. The proof usually involves a change of measure of the kind we have been discussing, a g that knows how rare f events that hit B do so.

4.1 Cramer's theorem

Cramer's theorem illustrates several ideas about large deviations. Let Y be a scalar mean zero random variable with density $h(y)$ that decays rapidly for large y. Let $S = \frac{1}{n}(Y_1 + \cdots + Y_n)$. We expect S to be within $O(\frac{1}{\sqrt{n}})$ of zero, but (for positive b) what is $P(S > b)$? Under suitable hypotheses below, we will find that

$$P(S \geq b) = e^{-R(b)n}\left(C(b)n^{-1/2} + O\left(n^{-3/2}\right)\right) . \tag{16}$$

Let us verify this when Y is Gaussian with variance $\sigma_y^2 = 1$. then S is Gaussian with variance $\sigma_S^2 = 1/n$ and the corresponding Gaussian probability density for f gives

$$P(S \geq b) = \sqrt{\frac{n}{2\pi}} \int_{s=b}^{\infty} e^{-ns^2/2} ds .$$

Using the new integration variable[4] $u = \sqrt{n}s$ gives

$$P(S \geq b) = \frac{1}{\sqrt{2\pi}} \int_{u=\sqrt{n}b}^{\infty} e^{-u^2/2} du .$$

The maximum value of the integrand, $e^{-nb^2/2}$, gives a rough idea how small the probability will be. Moreover, most of the mass is concentrated near the left endpoint: when u is large $e^{-u^2/2}$ is a very rapidly decreasing function of u. This motivates the change of variable $u = \sqrt{n}b + v$, so $u^2/2 = nb^2/2 + \sqrt{n}bv + v^2/2$, and

$$P(S \geq b) = \frac{1}{\sqrt{2\pi}} e^{-nb^2/2} \int_{v=0}^{\infty} e^{-\sqrt{n}bv} e^{-v^2/2} dv .$$

The first factor decays so rapidly that by the time $e^{-v^2/2}$ is significantly different from one, the product is essentially zero. This suggests the approximation

$$\int_{v=0}^{\infty} e^{-\sqrt{n}bv} e^{-v^2/2} dv \approx \int_{v=0}^{\infty} e^{-\sqrt{n}bv} dv = \frac{1}{\sqrt{n}b} .$$

[3] Amir Dembo has a nice book on large deviations that discusses their use in importance sampling.

[4] This is a general way, called *Watson's lemma*, to estimate integrals like this one, see the book by Erdely or the book by Bender and Orszag.

The interested reader will be able to show that the error in this approximation is $O(n^{-3/2})$ so that

$$P\left(S \geq b\right) = e^{-nb^2/2}\left(\frac{1}{\sqrt{2\pi}b}\frac{1}{\sqrt{n}} + O\left(n^{-3/2}\right)\right) \ .$$

This is of the general form (16) with $R = b^2/2$ and $C = 1/\sqrt{2\pi}b$.

This formula illustrates the predictability of rare events. If B is the event $S \geq b$, then most of the hits in B have S only slightly above b. In fact, $P(S \geq b+\epsilon \mid S \geq b) \sim e^{-\epsilon bn}$. Moreover, we explore the mechanism for samples $S \geq b$ by finding $h_b(y)$, the conditional probability density of, say, Y_1, given that $S \geq b$. In this Gaussian world, the density of Y_1 when $S = b + \epsilon$ is Gaussian and (after some thought) $h_b \sim \mathcal{N}(b + \epsilon, 1)$. Since ϵ is small for large n, this gives $f_b(y) \to \mathcal{N}(b, 1)$ as $n \to \infty$. This suggests that we can sample more effectively by drawing the Y_k from $\mathcal{N}(b, 1)$ than from $\mathcal{N}(0, 1)$.

The proof of Cramer's theorem for general $h(y)$ builds on these observations. The probability density for S is[5]

$$\phi(s) = n\int\cdots\int h(y_1)\cdots h(y_n)\delta(y_1 + \cdots + y_n - ns)dy_1\cdots dy_n \ .$$

The trick is to use an exponential factor with unknown force parameter λ

$$e^{n\lambda s}\phi(s) = n\int\cdots\int e^{\lambda y_1}h(y_1)\cdots^{\lambda y_n} h(y_n)\delta(\cdots)dy_1\cdots dy_n \ ,$$

because $ns = y_1 + \cdots + y_n$ wherever the integral is different from zero. The factors $e^{\lambda y}h(y)$ can be normalized to be probability densities:

$$h_\lambda(y) = \frac{1}{Z(\lambda)}e^{\lambda y}h(y) \ , \tag{17}$$

with

$$Z(\lambda) = \int e^{\lambda y}h(y)dy = E\left[e^{\lambda Y}\right] \ . \tag{18}$$

The hypothesis of Cramer's theorem is that the *exponential moments* (18) are finite, at least for a suitable range of λ. This implies taht $h(y)$ decays exponentially in some average sense as $y \to \infty$. Clearly, the force λ, changes the expected value of Y. Denote this by

$$\mu(\lambda) = \ E_\lambda\left[Y\right] = \frac{1}{Z(\lambda)}\int ye^{\lambda y}h(y)dy \ . \tag{19}$$

Note that h_b in the Gaussian case has the form (17) with λ chosen so that $\mu(\lambda)$ b. In general, define $\lambda_*(s)$ by

$$\mu(\lambda_*(s)) = s \ . \tag{20}$$

[5]The factor of n in front insures that $\int \phi(s)ds = 1$ because $\int \delta(a - ns)ds = 1/n$.

It is easy to see that if $\lambda_*(s)$ exists for a given s, it is unique. Uning this λ_*, we have

$$\begin{aligned}&\frac{1}{n}e^{n\lambda_*(s)s}\cdot Z(\lambda_*(s))^n\phi(s)\\&\quad=\int\cdots\int h_{\lambda_*}(y_1)\cdots h_{\lambda_*}(y_n)\delta(y_1+\cdots+y_n-ns)dy_1\cdots dy_n\ .\end{aligned}$$

What is special about using $\lambda=\lambda_*(s)$ is that the right side is probability density at s for an average of iid random variables with mean s. This allows us to use the central limit theorem to approximate the right side. Let

$$\begin{aligned}\sigma^2_{\lambda_*}&=E_{\lambda_*}\left[(Y-s)^2\right]\\&=\frac{1}{Z(\lambda_*)}\int(y-s)^2e^{\lambda_* y}h(y)dy\ .\end{aligned}$$

Then the right side corresponds to the maximum of a $\mathcal{N}(s,n\sigma^2_{\lambda_*})$ density, which is

$$\frac{1}{\sqrt{n}\sigma_{\lambda_*}}\ .$$

Altogether[6]

$$\phi(s)=Z(\lambda_*(s))^{-n}e^{-n\lambda_*(s)s}\left(\frac{\sqrt{n}}{\sigma_{\lambda_*}}+O\left(n^{-1/2}\right)\right)\ .$$

This shows that

$$\phi(s)=e^{-R(s)n}\left(d(s)\sqrt{n}+O\left(n^{-1/2}\right)\right)\ ,$$

with

$$R(s)=\ln\left\{Z(\lambda_*(s))\right\}+s\lambda_*(s)\ , \tag{21}$$

and

$$d(s)=\frac{1}{\sigma_{\lambda_*}}\ .$$

As with the Gaussian case

$$p=\int_{s=b}^{\infty}\phi(s)ds\ ,$$

and most of the mass is near $s=b$. Again using the Watson lemma, we expand

$$R(s)=R(b)+R'(b)(x-b)+O\left((x-b)^2\right)\ ,$$

and get (16) with R given by (21) and

$$c=\frac{1}{R'(s)\sigma_{\lambda_*}}\ .$$

[6]The error term comes from the error term in the central limit theorem.