4 Simulation design and analysis

Often in simulations one is interested in determining

$$\theta = \mathbf{E}\phi(X) = \int \psi(x)dx.$$

Here, X has density f, and $\psi(x) = \phi(x)f(x)$. We think of θ as a parameter connected with some stochastic model. To estimate θ, the model is simulated to obtain the output $X_1, \ldots, X_n$ which are such that $\theta = \mathbf{E}\phi(X_i), i = 1, \ldots, n$. Thus we can estimate θ by the so-called *raw estimate* or *crude Monte Carlo estimate*

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i).$$

From the law of large numbers, this is an unbiased estimate of θ, and

$$Var\hat{\theta}_0 = \frac{1}{n} Var\phi(X).$$

Here we will analyze estimates $\hat{\theta}$ of θ with respect to their variance. In particular, the aim of variance reduction is to produce an alternative estimator of θ having hopefully a much smaller variance than $\hat{\theta}_0$. Note that the order of magnitude cannot be improved in general.

Example 1 *Suppose we want to estimate*

$$\theta = \int_0^1 \sqrt{1-x^2}dx.$$

(We know that $\theta = \frac{\pi}{4}$). Then $nVar(\hat{\theta}_0)$ can be calculated to be $\frac{2}{3} - \frac{\pi^2}{16} = .0498$. This example is from Morgan (1984) and will recur.

Remark: Hit-or-miss Monte Carlo. If ψ is zero outside a finite interval (a, b) and $0 \leq \psi(x) \leq c$ for some constant c, and for all x, one could think of estimating $\theta = \int_a^b \psi(x)dx$ by simulating $(X_i, Y_i), i = 1, \ldots, n$ uniformly from the box $[a, b] \times [0, c]$ and count the number of observations that fall under the curve ψ, that is,

$$\hat{\theta}_1 = \frac{c(b-a)}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq \psi(X_i)).$$

This yields again an unbiased estimate of θ. It can easily be seen, see, for example, Ripley (1987), p. 121, that this *hit-or-miss Monte Carlo* method is less efficient than $\hat{\theta}_0$. In particular,

$$\begin{aligned}
\mathbf{P}(Y_i \le \psi(X_i)) &= \frac{1}{(b-a)c} \int_a^b \int_0^c \mathbf{1}(y_i \le \psi(x_i)) dy_i dx_i \\
&= \frac{1}{(b-a)c} \int_a^b \psi(x_i) dx_i \\
&= \frac{\theta}{(b-a)c},
\end{aligned}$$

so that

$$\begin{aligned}
Var(\hat{\theta}_1) &= \frac{\theta(c(b-a)-\theta)}{n} \\
&= \frac{c}{n} \int_a^b \psi(x)(b-a) dx - \frac{\theta^2}{n} \\
&\ge \frac{1}{n} \int_a^b \psi^2(x)(b-a) dx - \frac{\theta^2}{n} \\
&= Var\hat{\theta}_0,
\end{aligned}$$

where we took $f(x)$ to be the uniform density on $[a, b]$. Note that equality holds only if $\psi \equiv c$, in which case both variances vanish. *Hit-or-miss is always worse than crude Monte-Carlo.*

In Example 1, it can be shown that $nVar(\hat{\theta}_1) = \frac{\pi(4-\pi)}{16} \approx .1685$.

Note that hit-or-miss and crude Monte Carlo differ in replacing the indicator $\mathbf{1}(Y_i \le \psi(X_i))$ by its conditional expectation given X_i, namely $\psi(X_i)/c$. This illustrates a general principle for variance reduction: If, at any point of a Monte Carlo simulation, we can replace an estimate by an exact value, we shall reduce the sampling error in the final result.

4.1 Stratified sampling

If ψ was piecewise constant, then we could easily estimate θ by sampling one observation each from of the intervals where ψ is constant. The idea of stratified sampling for

$$\theta = \mathbf{E}\phi(X) = \int_a^b \psi(x) dx$$

on a finite interval (a,b) is to break the interval (a,b) into pieces where ψ is approximately constant. This idea is related to stratified sampling from populations. Say, we partition

$$a = \alpha_0 < \alpha_1 < \cdots < \alpha_k = b$$

and sample n_j observations from $(\alpha_{j-1}, \alpha_j), j = 1, \ldots, k$. Let $X_{1j}, X_{2j}, \ldots, X_{n_j j}$ be i.i.d. uniform on (α_{j-1}, α_j). Then we use the crude Monte Carlo estimate

$$\hat{\theta}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \psi(X_{ij})$$

on each of the intervals, and we combine them to give

$$\hat{\theta}_s = \sum_{j=1}^{k} (\alpha_{j-1} - \alpha_j) \hat{\theta}_j.$$

Due to the independence of the components, we obtain

$$\begin{aligned} Var(\hat{\theta}_s) &= \sum_{j=1}^{k} \frac{1}{n_j} \left((\alpha_j - \alpha_{j-1}) \int_{\alpha_{j-1}}^{\alpha_j} \psi^2(x)dx - \left\{ \int_{\alpha_{j-1}}^{\alpha_j} \psi(x)dx \right\}^2 \right) \\ &= \sum_{j=1}^{k} \frac{a_j}{n_j}, \end{aligned}$$

say. If $A = \sum_{j=1}^{k} \sqrt{a_j}$, then a Lagrange multiplier argument shows that the optimal allocation is to have $n_j = \frac{n}{A}\sqrt{a_j}, j = 1, \ldots, n$. Unfortunately the a_j's are typically not available.

In Example 1, when splitting the range of integration at the point $x = \frac{1}{\sqrt{2}}$, the minimum variance obtainable is $\frac{.01169}{n}$, and this is achieved with the sample sizes in the ratio 1 : 1.249.

4.2 Importance sampling

The idea here pushes the stratified sampling approach further: sample more frequently from those parts of the curve that display more variability. Now, for

$$\theta = \int \psi(x)dx = \int \phi(x)f(x)dx = \mathbf{E}\phi(X),$$

we would choose f non-uniform. Ideally, we would like to choose $f(x) = \frac{\psi(x)}{\theta}$, but of course θ is not available. In general, if g is another density, we can write

$$\int \psi(x)dx \quad = \quad \int \frac{\psi(x)}{g(x)} g(x)dx = \mathbf{E}\rho(Y),$$

where Y has density g, and $\rho(x) = \frac{\psi(x)}{g(x)}$. Suppose that Y_n are i.i.d. with density g, then put

$$\theta_g = \frac{1}{n} \sum_{j=1}^{n} \frac{\psi(Y_i)}{g(Y_i)}.$$

If a density function g can be chosen so that the random variable $\frac{\psi(X)}{g(X)}$ has a small variance, then this approach can result in a more efficient estimator of θ. Thus a good choice of g would be one that mimics the shape of ψ.

In Example 1, $g(x) = 2(1-x)$ gives $nVar(\hat{\theta}) = .1331$, whereas $g(x) = \frac{2}{3}(2-x)$ gives $nVar(\hat{\theta}) = .01339$.

Example: Exponential tilting. (See Ross (1996), p.170 ff.) If θ is very small, then exponential tilting might be useful. Let $M(t) = \int e^{tx} f(x)dx$ be the moment-generating function corresponding to the density f. Then the density

$$f_t(x) = \frac{e^{tx}}{M(t)} f(x)$$

is called a *tilted* density of f, $-\infty < t < \infty$. Similarly a tilted probability mass function can be defined. For a Bernoulli(p)-variable, for example, we have $f(x) = p^x(1-p)^{1-x}$, $M(t) = pe^t + 1 - p$ and

$$f_t(x) = \frac{1}{M(t)}(pe^t)^x(1-p)^{1-x} = \left(\frac{pe^t}{pe^t+1-p}\right)^x \left(\frac{1-p}{pe^t+1-p}\right)^{1-x}.$$

This is the probability mass function of Bernoulli($pe^t/(pe^t+1-p)$). Note that

$$\frac{f(x)}{f_t(x)} \quad = \quad M(t)\left(\frac{p}{pe^t}\right)^x \left(\frac{1-p}{1-p}\right)^{1-x} = M(t)e^{-tx}.$$

In certain situations, the quantity of interest might be the sum of independent random variables $X_1, \ldots, X_n$ with density f each. In this case, the joint density f is the product of the one-dimensional densities. In this situation it may be useful to generate the X_i's according to their tilted densities. For instance, suppose we are interested in estimating the probability that a sum S_n of n independent Bernoulli-random variables, $X_i \sim Be(p_i), i = 1, \ldots, n$, exceed a large value a. Then

$$\theta = \mathbf{E1}(S_n \geq a).$$

Thus $\phi(x_1, \ldots, x_n) = \mathbf{1}(\sum_{i=1}^n x_i \geq a)$. Now simulate Y_i according to the t-tilted Bernoulli distribution with parameter

$$p_{t,i} = \frac{p_i e^t}{1 - p_i + p_i e^t}; \quad i = 1, \ldots, n.$$

Then the importance sampling estimator of θ is

$$\begin{aligned} \hat{\theta} &= \mathbf{1}(\sum_{i=1}^n Y_i \geq a) \prod_{i=1}^n \frac{f_i(Y_i)}{f_{i,t}(Y_i)} \\ &= \mathbf{1}(\sum_{i=1}^n Y_i \geq a) \prod_{i=1}^n M_i(t) e^{-tY_i} \\ &= \mathbf{1}(\sum_{i=1}^n Y_i \geq a) M(t) e^{-t\sum_{i=1}^n Y_i}, \end{aligned}$$

where $M(t) = \prod_{i=1}^n M_i(t)$. Since $t > 0$ it follows that $0 \leq \hat{\theta} \leq M(t)e^{-t\sum_{i=1}^n Y_i}$. To make the bound as small as possible, choose t to minimize $M(t)e^{-at}$. It can be shown that this minimizing t^* satisfies

$$\mathbf{E}\sum_{i=1}^n Y_i = \sum_{i=1}^n \frac{p_i e^t}{1 - p_i + p_i e^t} = a.$$

The optimal choice of t can be approximated numerically.

For example, if $n = 20, p_i = .4, a = 16$, then $\mathbf{E}(\sum_{i=1}^n Y_i) = 20\frac{.4e^t}{.4e^t + .6}$; this equals 16 is $t = \ln 6$. the importance sampling estimator is

$$\hat{\theta} = \mathbf{1}(\sum_{i=1}^n Y_i \geq 16) 6^{-\sum_{i=1}^n Y_i} 3^{20}.$$

It can be shown that $Var\hat{\theta} \leq 2.9131 \times 10^{-7}$, whereas $Var\hat{\theta}_0 = 3.160 \times 10^{-4}$. NOte that $\theta \approx 3.17 \times 10^{-4}$.

In general, variance reduction may or may not be obtained, depending on the choice of g.

4.3 Control variates

The idea behind control variates is formally related to regression. Suppose we have Y (perhaps $Y = \phi(X)$) and we want to estimate its mean

$$\theta = \mathbf{E}Y.$$

If Z is a related random variable with known mean μ, then put

$$W = Y - c(Z - \mu)$$

for some constant c. Then $\mathbf{E}W = \theta$ for any c. Thus, if $W_1, \ldots, W_n$ are i.i.d. with same distribution as W, then

$$\hat{\theta} = \bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$$

is unbiased for θ. Moreover,

$$\begin{aligned} nVar\hat{\theta} &= VarW \\ &= VarY + c^2VarZ - 2cCov(Y,Z) \\ &< VarY \quad \text{if and only if } Cov(Y,Z) > \frac{c}{2}VarZ. \end{aligned}$$

The above variance is minimized for

$$c^* = \frac{Cov(Y,Z)}{VarZ}.$$

Then we obtain

$$VarW = VarY - \frac{Cov^2(Y,Z)}{VarZ}.$$

Thus, variance reduction is always achievable by suitable choice of c whenever $Cov(Y,Z) \neq 0$.

Example 2 *(See Ross (1996), p.144-145.) Suppose we want to use simulation to compute*

$$\theta = \mathbf{E}e^U$$

for $U \sim \mathcal{U}([0,1])$. Note that $nVar\hat{\theta}_0 = .2402$. A natural choice for a control variate is U. We then have

$$\begin{aligned} Cov(e^U, U) &= \mathbf{E}(Ue^U) - \mathbf{E}(U)\mathbf{E}(e^U) \\ &= \int_0^1 xe^x dx - \frac{e-1}{2} = .14086. \end{aligned}$$

Moreover, $VarU = \frac{1}{12}$. Hence, with $c^* = 12 \times .14086$ we have

$$\begin{aligned} nVar\hat{\theta} &= Var(e^U) - 12 \times (.14086)^2 \\ &\approx .2402 - .2380 = .0039. \end{aligned}$$

In Example 1, with $Z = 1 - X$ as control variate, $c^* = \frac{3}{2}\pi - 4$, and $nVar\theta \approx .00752$.

Of course this method can be generalized to

$$W = Y - \sum_{i=1}^{k} c_k(Z_k - \mu_k)$$

when such $Z_1, \ldots, Z_n$ are available.

In general, c^* will not be available. One could estimate c from the experiment, but then $\bar{W} = \frac{1}{n}\sum_{i=1}^n W_i$ will in general not be an unbiased estimator of θ. Instead, it is better to use a pilot simulation to estimate c^*, and then use this estimated c^* for the larger simulation.

It is appealing that even when this method is not very successful, the resulting variance is never increased.

Due to the relation to standard regression analysis, often also the term *regression-adjusted control variates* is used. The similarity is formal, though: regression analysis via least squares is based upon the assumption of linear dependence (and preferably normal errors) whereas nothing like this is needed for regression-adjusted control variates.

4.4 Antithetic variates

Here the idea is to generate two (or $2n$) correlated unbiased estimators Y_1, Y_2 of θ with the same marginal distribution, described by Y, say. We then put

$$\hat{\theta} = \frac{1}{2}(Y_1 + Y_2).$$

Thus

$$Var\hat{\theta} = \frac{1}{2} VarY(1 + corr(Y_1, Y_2)).$$

If

$$corr(Y_1, Y_2) < 0$$

then we obtain a smaller variance than with independent estimators.

A standard way of generating such correlated unbiased estimators from a distribution function F is to put

$$Y_1 = F^{-1}(U), \quad Y_2 = F^{-1}(1 - U),$$

where $U \sim \mathcal{U}([0, 1])$. Then the correlation is negative, following from (see Ripley (1987), p.129)

Proposition 1 *Suppose g is a monotonic function on* $(0, 1)$. *Then*

$$corr(g(U), g(1 - U)) < 0.$$

In Example 1, letting $Y_1 = \sqrt{1 - U^2}$ and $Y_2 = \sqrt{1 - (1 - U)^2}$ gives $Var\hat{\theta} = .0052$. In Example 2, with $Y_1 = e^U$ and $Y_2 = e^{1-U}$, we obtain $Var\hat{\theta} = \frac{1}{2}.2420(1 - .9677) = .0039$.

4.5 Conditional Monte Carlo

In general, we have for any random variables Y, W that

$$\begin{aligned} \mathbf{E}Y &- \mathbf{E}(\mathbf{E}(Y|W)) \\ VarY &- Var(\mathbf{E}(Y|W)) + \mathbf{E}(Var(Y|W)). \end{aligned}$$

Hence $Var(\mathbf{E}(Y|W)) \leq VarY$. If we can evaluate $\mathbf{E}(Y|W)$ analytically as a function h of W, we can estimate $\theta = \mathbf{E}Y$ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} h(W_i),$$

where $W_1, \ldots, W_n$ are i.i.d. copies of W. If the distribution of Y is built up by mixing over values of some W, this becomes an obvious target for the technique.

Example: Suppose $W \sim Poisson(\lambda)$, and, given $W = w$, we have that $Y \sim Beta(w, w^2+1)$. Then

$$\mathbf{E}(Y|W=w) = \frac{w}{w^2+w+1}.$$

Thus, to simulate $\theta = \mathbf{E}Y$, we simulate $W_1, \ldots, W_n \sim Poisson(\lambda)$ i.i.d. and put

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{W_i^2 + W_i + 1}.$$

Conditional Monte Carlo always provides variance reduction. The difficulty is to find W such that the conditional expectation is computable.

4.6 Isolating known components

In many cases, some parts of the expectation θ of $\phi(X)$ can be evaluated analytically. One may then attempt to organize the output analysis to that these known parts need not be simulated.

Example. Let $T_1, T_2, \ldots$ be i.i.d. and nonnegative with mean μ, and let $Z = \sup\{n : S_n \leq t\}$ be the number of renewals up to time t, where $S_n = T_1 + T_2 + \cdots + T_n$. Let $\theta = \mathbf{E}Z$. Letting $\tau = \inf\{n : S_n > t\}$, we then have $Z = \tau - 1$. By Wald's identity,

$$\mathbf{E}S_\tau = \mu\mathbf{E}\tau = \mu(\theta + 1).$$

This suggests the estimator

$$\hat{\theta}_1 = \frac{S_\tau}{\mu} - 1.$$

But we can write $S_\tau = t + \xi$, where $\xi = S_\tau - t$ is the overshoot. This yields

$$\theta = \frac{t + \mathbf{E}\xi}{\mu} - 1$$

and an alternative estimator is

$$\hat{\theta}_2 = \frac{t + \xi}{\mu} - 1.$$

For example, if the T_i are standard exponential and $t = 50$, then $Z \sim Poisson(50)$ so that $Var\hat{\theta}_1 = 50$. In contrast, since ξ is again standard exponential, $Var(\hat{\theta}_2) = 1$.

4.7 Experimental design

Many simulation experiments are designed to compare the effect of choosing different parameter values in the model. In such cases, ideas from the design of experiments can be used, see also Box *et al.* (1978). The analogue of a randomized block is a set of random numbers which can be re-used (*common random numbers*). Suppose we are simulating

$$\mathbf{E}\phi(X;\alpha),$$

where α is a parameter value, to be varied. To assess the variation between two different parameter values α_1 and α_2, we are interested in

$$\theta = \mathbf{E}\phi(X,\alpha_1) - \mathbf{E}\phi(X,\alpha_2).$$

If we use the same random numbers X for both parts, it is likely that $Cov(\phi(X,\alpha_1),\phi(X,\alpha_2) \neq 0$, and so we would obtain a variance reduction compared to choosing independent X's.

4.8 Discussion

Variance reduction techniques are typically most readily available for well structured problems. Typically, they involve a fair amount of both theoretical study of the problem and of additional programming effort. For this reason, variance reduction is most often only recommendable for large experiments.

Different variance reduction techniques used in combination may produce diminishing returns and may even conflict with each other to give adverse results.

It will generally be advantageous to break down a problem into components and to push the analytic treatment of the problem through as far as possible.

Further reading

1. S. Asmussen (1999). *Stochastic Simulation with a view towards stochastic processes.* MaPhySto, University of Aarhus. On the web at `http://www.maphysto.dk`

2. G.E.P. Box, W.G. Hunter, and J.S. Hunter (1978). *Statistics for Experimenters.* Wiley.