Variance Reduction Techniques

Ruming Liu

October 23, 2021

Contents

1 Control Variates

1.1 Methodology

Suppose that the Y_i are independent and identically distributed and that our goal is to estimate $E[Y_i]$. The usual estimator is the sample mean $\bar{Y} = (Y_1 + ... + Y_n)/n$. This estimator is unbiased and consistent.

But we still can use some techniques to make our estimator more efficient. The method is calculate another output X_i along with Y_i. Suppose that the pairs $(X_i, Y_i), i = 1, ..., n$, are i.i.d. and that the expectation $E[X]$ is known. Then for any fixed b we can calculate

$$Y_i(b) = Y_i - b(X_i - E[X]). \tag{1.1}$$

And then compute the sample mean by

$$\bar{Y}(b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b(X_i - E[X])). \tag{1.2}$$

This is a control variate estimator; the observed error $\bar{X} - E[X]$ serves as a control in estimating $E[Y]$.

As an estimator of $E[Y]$, it is apparently unbiased. And each $Y_i(b)$ has variance

$$Var[Y_i(b)] = Var[Y_i - b(X_i - E[X])] = \sigma_Y^2 - 2b\sigma_X\sigma_Y\rho_{XY} + b^2\sigma_X^2 \equiv \sigma^2(b), \tag{1.3}$$

then the control variate estimator $\bar{Y}(b)$ has variance $\sigma^2(b)/n$ and the ordinary sample mean $\bar{Y}$ has variance σ_Y^2/n. Hence the control variate estimator is more efficient if $-2b\sigma_X\sigma_Y\rho_{XY} + b^2\sigma_X^2 < 0$. And from (1.3), the optimal coefficient b^* is given by:

$$b^* = \frac{\sigma_Y}{\sigma_X}\rho_{XY} = \frac{Cov[X, Y]}{Var[X]}. \tag{1.4}$$

We can also find the ratio of the variance of the optimally controlled estimator to that of the uncontrolled estimator is

$$\frac{Var[\bar{Y}_i - b^*(\bar{X}_i - E[X])]}{Var[\bar{Y}]} = 1 - \rho_{XY}^2. \tag{1.5}$$

The variance reduction factor increases very sharply as $|\rho_{XY}|$ approaches 1.

These remarks and equation (1.4) apply if the optimal coefficient b^* is known. But sometimes σ_Y, ρ_{XY} may be unknown, we can use the population parameters to estimate

$$\hat{b}_n = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{1.6}$$

1.2 Output Analysis

In analyzing variance reduction techniques, along with the effectiveness of a technique it is important to consider how the technique affects the statistical interpretation of simulation outputs. But we will see that some variance reduction techniques complicate interval estimation by introducing dependence across replications. This issue arises with control variates if we use

the estimated coefficient $\hat{b}_n$. It turns out that in the case of control variates the dependence can be ignored in large samples; a more careful consideration of small-sample issues will be given in section 1.4.

For any fixed b, the control variate estimator $\bar{Y}(b)$ is the sample mean of independent replications $Y_i(b), i = 1, ..., n$. Accordingly, an asymptotically valid $1 - \delta$ confidence interval for $E[Y]$ is given by

$$\bar{Y}(b) \pm z_{\delta/2} \frac{\sigma(b)}{\sqrt{n}}. \tag{1.7}$$

In practice, $\sigma(b)$ is typically unknown but can be estimated using the sample standard deviation

$$s(b) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i(b) - \bar{Y}(b))^2}. \tag{1.8}$$

The confidence interval (4.10) remains asymptotically valid if we replace $\sigma(b)$ with $s(b)$.

If we use the estimated coefficient $\hat{b}_n$, then the estimator

$$\bar{Y}(\hat{b}_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{b}_n (X_i - E[X])) \tag{1.9}$$

When n is large enough, we still can use the central limit theorem applies in the form

$$\frac{\bar{Y}(\hat{b}_n) - E[Y]}{s(\hat{b}_n)/\sqrt{n}} \Rightarrow N(0, 1).$$

We may summarize this discussion as follows. It is a simple matter to estimate asymptotically valid confidence intervals for control variate estimators. Moreover, for large n, we get all the benefit of the optimal coefficient b^* by using the estimate $\hat{b}_n$. However, for finite n, there may still be costs to using an estimated rather than a fixed coefficient; we return to this point in Section 1.4.

1.3 Control Variates and Weighted Monte Carlo

In introducing the idea of a control variate in Section 1.1, we explained that the observed error in estimating a known quantity $(X - E[X])$ can be used to adjust an estimator of an unknown quantity $(E[Y])$. But the technique has an alternative interpretation as a method for assigning weights to replications. This alternative perspective is sometimes useful, particularly in relating control variates to other methods.

We start with the case of a single control, thus (X_i, Y_i) are i.i.d. The control variate estimator with estimated optimal coefficient $\hat{b}_n$ is $\bar{Y}(\hat{b}_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{b}_n (X_i - E[X]))$, where $\hat{b}_n$ is given by

$$\bar{Y}(b) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - b(X_i - E[X])).$$

By substitution, we arrive at

$$\bar{Y}(\hat{b}_n) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(\bar{X} - X_i)(\bar{X} - E[X])}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right) Y_i \equiv \sum_{i=1}^{n} \omega_i Y_i \tag{1.10}$$

In other words, the control variate estimator is a weighted average of Y_i. The weights ω_i are completely determined by the observations X_i of the control.

1.4 Small-Sample Issues

[1] In our discussion of output analysis with the method of control variates, we consider the situation when $n \to \infty$. In this section, we supplement these large-sample properties with a discussion of statistical issues that arise in analyzing control variate estimators based on a finite number of samples. We note that stronger distributional assumptions on the simulation output lead to confidence intervals valid for all n. Moreover, it becomes possible to quantify the loss in efficiency due to estimating b^*. This offers some guidance in deciding how many control variates to use in a simulation.

For any fixed b, the control variate estimator $\bar{Y}(b)$ is unbiased. But if using $\hat{b}_n$, the bias is

$$Bias(\bar{Y}(\hat{b}_n)) = -E[\hat{b}_n(\bar{X} - E[X])] \tag{1.11}$$

which may not be zero because $\hat{b}_n$ and $\bar{X}$ are not independent. But if the regression of Y on X is linear, the control variate estimator is still unbiased. More precisely, if

$$E[Y|X] = c_0 + c_1 X^{(1)} + ... + c_d X^{(d)}, \tag{1.12}$$

then $E[\bar{Y}(\hat{b}_n)] = E[Y]$.

1.5 Nonlinear Controls

Our discussion of control variates has thus far focused exclusively on linear controls, meaning estimators of the form

$$\bar{Y} - b^T(\bar{X} - E[X]), \tag{1.13}$$

with the vector b either known or estimated. There are, however, other ways one might use the discrepancy between $\bar{X}$ and $E[X]$ to try to improve the estimator $\bar{Y}$ in estimating $E[Y]$. For example, in the case of scalar X, the estimator

$$\bar{Y}\frac{E[X]}{\bar{X}}$$

and thus may be attractive if X_i and Y_i are positively correlated. Other estimators of this type include

$$\bar{Y}exp(\bar{X} - E[X])$$

$$\bar{Y}^{\bar{X}/E[X]}$$

Although the introduction of nonlinear controls would appear to substantially enlarge the class of candidate estimators, it turns out that in large samples, a nonlinear control variate estimator based on a smooth h is equivalent to an ordinary linear control variate estimator.

[1] As for more details about multiple controls situation, please check MCMF page 201.

2 Antithetic Variates

The method of antithetic variates attempts to reduce variance by introducing negative dependence between pairs of replications. The method can take various forms; the most broadly applicable is based on the observation that if U is uniformly distributed over [0, 1], then 1 − U is too. Hence, if we generate a path using as inputs $U_1, ..., U_n$, we can generate a second path using $1-U_1, ..., 1-U_n$ without changing the law of the simulated process. The variables U_i and $1-U_i$ form an antithetic pair in the sense that a large value of one is accompanied by a small value of the other. This suggests that an unusually large or small output computed from the first path may be balanced by the value computed from the antithetic path, resulting in a reduction in variance.

These observations extend to other distributions through the inverse trans- form method: $F^{-1}(U)$ and $F^{-1}(1-U)$ both have distribution F but are antithetic to each other because F^{-1} is monotone. For a distribution symmetric about the origin, $F^{-1}(1-u)$ and $F^{-1}(u)$ have the same magnitudes but opposite signs.

To analyze this approach more precisely, suppose our objective is to estimate an expectation $E[Y]$ and that using some implementation of antithetic sampling produces a sequence of pairs of observations $(Y_1, \widetilde{Y_1}), ..., (Y_n, \widetilde{Y_n})$. The key features of the antithetic variates method are the following:

- $(Y_1, \widetilde{Y_1}), ..., (Y_n, \widetilde{Y_n})$ are i.i.d.
- for each i, Y_i and $\widetilde{Y_i}$ have the same distribution, though they are not independent

The antithetic variates estimator is simply the average of all 2n observations,

$$\hat{Y_{AV}} = \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i + \widetilde{Y_i}}{2}). \tag{2.1}$$

Using antithetic variate reduces variance if

$$Var[\hat{Y}_{AV}] < Var[\frac{1}{2n}\sum_{i=1}^{2n} Y_i] \tag{2.2}$$

i.e., if

$$Var[Y_i + \widetilde{Y_i}] < 2Var[Y_i].$$

Because Y_i and $\widetilde{Y_i}$ have a same distribution, thus the variance on the left can be written as

$$Var[Y_i + \widetilde{Y_i}] = Var[Y_i] + Var[\widetilde{Y_i}] + 2Cov[Y_i, \widetilde{Y_i}] = 2Var[Y_i] + 2Cov[Y_i, \widetilde{Y_i}]. \tag{2.3}$$

Thus the condition for antithetic sampling to reduce variance becomes

$$Cov[Y_i, \widetilde{Y_i}] < 0. \tag{2.4}$$

For example, the antithetic pairs $(U, 1-U)$ with $U \sim Unif[0, 1]$ and $(Z, -Z)$ with $Z \sim N(0, 1)$. The negative covariance leads our estimator more efficient.

3 Stratified Sampling

3.1 Methodology

Stratified sampling refers broadly to any sampling mechanism that constrains the fraction of observations drawn from specific subsets (or strata) of the sample space. Suppose, more specifically, that our goal is to estimate $E[Y]$ with Y real-valued, and let $A_1, ..., A_K$ be disjoint subsets of the real line for which $P(Y \in \cup_i A_i) = 1$. Then

$$E[Y] = \sum_{i=1}^{K} P(Y \in A_i) E[Y|Y \in A_i] = \sum_{i=1}^{K} p_i E[Y|Y \in A_i] \tag{3.1}$$

The simplest case is proportional sampling, in which we ensure that the fraction of observations drawn from stratum A_i matches the theoretical probability $p_i = P(Y \in A_i)$. If the total sample size is n, this entails generating $n_i = np_i$ samples from A_i. For each $i = 1, ..., K$, let $Y_{ij}, j = 1, ..., n_i$ be independent draws from the conditional distribution of Y given $Y \in A_i$. An unbiased estimator of $E[Y|Y \in A_i]$ is provided by the sample mean $(Y_{i1} + \cdots + Y_{in_i})/n_i$ of observations from the ith stratum. It follows from (3.1) that an unbiased estimator of $E[Y]$ is provided by

$$\hat{Y} = \sum_{i=1}^{K} p_i \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij}. \tag{3.2}$$

We generalize this formulation in two simple but important ways. First, we allow the strata to be defined in terms of a second variable X. This stratification variable could take values in an arbitrary set; to be concrete we assume it is R^d-valued and thus take the strata A_i to be disjoint subsets of R^d with $P(X \in \cup_i A_i) = 1$. Then (3.1) generalizes to

$$E[Y] = \sum_{i=1}^{K} P(X \in A_i) E[Y|X \in A_i] = \sum_{i=1}^{K} p_i E[Y|X \in A_i] \tag{3.3}$$

In some applications, Y is a function of X (for example, X may be a discrete path of asset prices and Y the discounted payoff of a derivative security), but more generally they may be dependent without either completely determining the other. To use (3.3) for stratified sampling, we need to generate pairs (X_{ij}, Y_{ij}), $j = 1, ..., n_i$, having the conditional distribution of (X, Y) given $X \in A_i$.

As a second extension of the method, we allow the stratum allocations $n_1, ..., n_K$ to be arbitrary (while summing to n) rather than proportional to $p_1, ..., p_K$. In this case, the first representation in (3.2) remains valid but the second does not. If we let $q_i = n_i/n$ be the fraction of observations drawn from stratum $i, i = 1, ..., K$, we can write

$$\hat{Y} = \sum_{i=1}^{K} p_i \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{K} \frac{p_i}{q_i} \sum_{j=1}^{n_i} Y_{ij}. \tag{3.4}$$

By minimizing the variance of this estimator over the q_i, we can find an allocation rule that is at least as effective as a proportional allocation. We return to this point in the next section.

3.2 Optimal Allocation

The variance of $\hat{Y}$ is given by

$$Var[\hat{Y}] = \sum_{i=1}^{K} p_i^2 Var[\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}] = \sum_{i=1}^{K} p_i^2 \frac{\sigma_i^2}{n_i} = \frac{\sigma^2(q)}{n}, \tag{3.5}$$

where

$$\sigma^2(q) = \sum_{i=1}^{K} \frac{p_i^2}{q_i}\sigma_i^2. \tag{3.6}$$

$$\sum q_i = 1 \tag{3.7}$$

Optimizing the allocation can produce further variance reduction. For any fixed n, minimizing $\sigma^2(q)$ subject to the constraint that $(q_1, ..., q_K)$ be a probability vector yields the optimal allocation

$$q_i^* = \frac{p_i\sigma_i}{\sum_{j=1}^{K} p_j\sigma_j}, i = 1, ..., K. \tag{3.8}$$

In other words, the optimal allocation for each stratum is proportional to the product of the stratum probability and the stratum standard deviation. The optimal variance is thus

$$\sigma^2(q^*) = \sum_{i=1}^{K} \frac{p_i^2}{q_i^*}\sigma_i^2 = (\sum_{i=1}^{K} p_i\sigma_i)^2. \tag{3.9}$$

4 Importance Sampling

4.1 Methodology

Importance sampling attempts to reduce variance by changing the probability measure from which paths are generated. Changing measures is a standard tool in financial mathematics; When we switch from, say, the objective probability measure to the risk-neutral measure, our goal is usually to obtain a more convenient representation of an expected value. In importance sampling, we change measures to try to give more weight to "important" outcomes thereby increasing sampling efficiency.

To make this idea concrete, consider the problem of estimating

$$\alpha = E[h(X)] = \int h(x)f(x)dx$$

where X is a random element of R^d with probability density f, and h is a function from R^d to R. The ordinary Monte Carlo estimator is

$$\hat{\alpha} = \hat{\alpha}(n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \tag{4.1}$$

with $X_1, ..., X_n$ independent draws from f. Let g be any other probability density on R^d satisfying

$$f(x) > 0 \Rightarrow g(x) > 0 \tag{4.2}$$

for all $x \in R^d$. Then we can alternatively represent α as

$$\alpha = \int h(x)\frac{f(x)}{g(x)}g(x)dx. \tag{4.3}$$

This integral can be interpreted as an expectation with respect to the density g; we may therefore write

$$\alpha = \widetilde{E}[h(X)\frac{f(X)}{g(X)}], \tag{4.4}$$

$\widetilde{E}$ here indicating that the expectation is taken with X distributed according to g. If $X_1, ..., X_n$ are now independent draws from g, the importance sampling estimator associated with g is

$$\hat{\alpha}_g = \hat{\alpha}_g = \frac{1}{n}\sum_{i=1}^{n} h(X_i)\frac{f(X_i)}{g(X_i)}. \tag{4.5}$$

The weight $f(X_i)/g(X_i)$ is the likelihood ratio or Radon-Nikodym derivative evaluated at X_i.

It follows from (4.4) that $\widetilde{E}[\hat{\alpha}_g] = \alpha$ and thus $\hat{\alpha}_g$ is an unbiased estimator. To compare variances with and without importance sampling it therefore suffices to compare second moments. With importance sampling, we have

$$\widetilde{E}[(h(X)\frac{f(X)}{g(X)})^2] = E[h(X)^2\frac{f(X)}{g(X)}]. \tag{4.6}$$

This could be larger or smaller than the second moment $E[h(X)^2]$ without importance sampling; indeed, depending on the choice of g it might even be infinitely larger or smaller. Successful importance sampling lies in the art of selecting an effective importance sampling density g.

Consider the special case where h is non-negative. The product $h(x)f(x)$ is then also non-negative and may be normalized to a probability density. Suppose g is this density. Then

$$g(x) \propto h(x)f(x), \tag{4.7}$$

and $h(X_i)f(X_i)/g(X_i)$ equals the constant of proportionality in (4.7) regardless of the value of X_i. Thus the importance sampling estimator $\hat{\alpha}_g$ in

Nevertheless, this optimal choice of g does provide some useful guidance: in designing an effective importance sampling strategy, we should try to sample in proportion to the product of h and f. In option pricing applications, h is typically a discounted payoff and f is the risk-neutral density of a discrete path of underlying assets. In this case, the "importance" of a path is measured by the product of its discounted payoff and its probability density.

If h is the indicator function of a set, then the optimal importance sampling density is the original density conditioned on the set. In more detail, suppose $h(x) = 1\{x \in A\}$ for some $A \in R^d$. Then $\alpha = P(X \in A)$ and the zero-variance importance sampling density $h(x)f(x)/\alpha$ is precisely the conditional density of X given $X \in A$ (assuming $\alpha > 0$). Thus, in applying importance sampling to estimate a probability, we should look for an importance sampling density that approximates the conditional density. This means choosing g to make the event $\{X \in A\}$ more likely, especially if A is a rare set under f. [2]

[2] [Ruming's Note] The optimal choice for $g(x)$ is $g(x) = h(x)f(x)/P(X \in A)$, because only the sample on the support of $h(x)$ matters, so $g(x)$ puts all weights on this support. [End]

4.2 Likelihood Ratios

For more topics about this chapter:

Reference 1

Reference 2

Reference 3