3 Variance Reduction Methods, I

We return to the problem of Monte-Carlo *integration*, namely, estimating the expected value

$$\ell = E(Z)\,, \quad Z = H(X)$$

where H is a known function, and X a random vector with pdf $f = f_X$.

The basic Monte-Carlo estimator is

$$\hat{\ell}_N = \frac{1}{N}\sum_{i=1}^{N} H(X_i)$$

where $(X_1, \ldots, X_N)$ is a *random sample* from f_X, namely an iid sequence with each $X_i \sim f_X$.

Clearly, $\hat{\ell}_N$ is an *unbiased* estimator:

$$E(\hat{\ell}_N) = \frac{1}{N}\sum_{i=1}^{N} EH(X_i) = \ell\,.$$

The mean square error (MSE) of the estimator is given by

$$\mathrm{MSE}(\hat{\ell}_N) \triangleq E(\hat{\ell}_N - \ell)^2 = \mathrm{Var}(\hat{\ell}_N) = \cdots = \frac{1}{N}\mathrm{Var}(H(X))\,.$$

(Note that the MSE and variance are the same here since the estimator is unbiased.) The variance of is the central quantity by which we measure the quality of the estimator.

As $\mathrm{Var}(H(X))$ may be large, a whole field of Monte-Carlo methods is dedicated to reducing the variance of the sampled RV. The basic idea is to use a modified estimator

$$\hat{\ell}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

where (Y_i) is a sequence of RVs with $E(Y_i) = E(H(X))$, and $\mathrm{Var}(Y_i) \leq \mathrm{Var}(H(X))$.

The most important method in this class is *Importance Sampling*, which is discussed in the next lecture. Here we discuss several other elementary approaches.

Monte Carlo Methods – Lecture Notes, N. Shimkin, Spring 2015

3.1 Monitoring the Estimation Error

Before going into variance reduction, let us discuss briefly the topic of regulating the estimation error.

Consider again the basic MC estimator

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

were (Z_i) are iid and $V = \mathrm{Var}(Z) < \infty$. As we have seen, $\mathrm{Var}(\hat{\ell}) = \frac{1}{N}\mathrm{Var}(Z)$.

Empirical variance estimation: In general $\mathrm{Var}(Z)$ is not known, and we can estimate it empirically using

$$\hat{V} \equiv S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \hat{\ell})^2$$

We know that $\hat{V}$ is an unbiased estimator of $\mathrm{Var}(Z)$, and converges to it w.p. 1 as $N \to \infty$.

Confidence Intervals: Using the variance $V = \mathrm{Var}(Z)$ (or its estimate $\hat{V}$), we can form approximate bounds on the estimation error. Note that, by the CLT,

$$\sqrt{N}(\hat{\ell} - \ell) \Rightarrow N(0, V)$$

(convergence in distribution to a normal RV). Therefore, an approximate $1-\alpha$ confidence interval for ℓ is given by

$$I_{1-\alpha} = (\hat{\ell} - \beta\frac{\sqrt{V}}{\sqrt{N}}, \hat{\ell} + \beta\frac{\sqrt{V}}{\sqrt{N}}) \, ,$$

where β is the $(1 - \frac{\alpha}{2})$-quantile of the normal distribution $\phi(x)$, namely

$$\int_{-\beta}^{\beta} \phi(x)dx = 1 - \alpha, \quad \text{or} \quad \beta = \Phi^{-1}(1 - \frac{\alpha}{2})$$

For example, for $\alpha = 0.05$ (95% confidence) we get $\beta = 1.96$.

Coefficient of Variation: It is often the case that the *relative* estimation error $\frac{\hat{\ell}-\ell}{\ell}$ is important, rather than the absolute one. for that purpose the following measure is used:

$$\kappa_N = \frac{\sqrt{\mathrm{Var}(\hat{\ell})}}{\ell} = \frac{1}{\sqrt{N}} \frac{\sqrt{\mathrm{Var}(Z)}}{\ell} = \frac{1}{\sqrt{N}} \kappa_1 \, .$$

κ_N is appropriately called the *relative error* of $\hat{\ell}$, and its square κ_N^2 is the *squared coefficient of variation* of $\hat{\ell}$.

As before, we can estimate κ_N from the obtained sampled using

$$\hat{\kappa} = \frac{1}{\sqrt{N}} \frac{\sqrt{\hat{V}}}{\hat{\ell}} .$$

In terms of κ, the $(1-\alpha)$ confidence interval for ℓ is

$$I_{1-\alpha} = \hat{\ell}(1 \mp \beta\kappa_N)$$

Example: Rare-event Probability Estimation [RK, example 1.14]. Let

$$\ell = \mathbb{P}(X \geq \gamma)$$

where γ is a large number, so that $\ell << 1$.

Consider the crude MC estimator,

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} Z_i , \quad Z_i = 1_{\{X_i \geq \gamma\}} .$$

The relative error here is

$$\kappa_N = \frac{1}{\sqrt{N}} \frac{\sqrt{\mathrm{Var}(Z)}}{\ell} = \frac{1}{\sqrt{N}} \frac{\sqrt{\ell(1-\ell)}}{\ell} \approx \frac{1}{\sqrt{N\ell}} .$$

Suppose $\ell = 10^{-6}$, then for relative error $\kappa = 0.01$ we need

$$N \approx \frac{1}{\kappa^2 \ell} = 10^{10} .$$

This demonstrates that crude MC is practically useless for estimating rare event probabilities.

3.2 Control Variates

We wish to estimate $\ell = E(Z)$, $Z = H(X)$.

Let W be an RV *correlated* with Z, and with known mean $m_W = E(W)$. Then

$$Z_\alpha = Z - \alpha(W - m_W)$$

has the same mean as Z (for any α), and variance

$$\mathrm{Var}(Z_\alpha) = \mathrm{Var}(Z) - 2\alpha\mathrm{Cov}(Z,W) + \alpha^2\mathrm{Var}(W)\,.$$

This variance is minimized by

$$\alpha^* = \frac{\mathrm{Cov}(Z,W)}{\mathrm{Var}(W)}\,,$$

which gives

$$\mathrm{Var}(Z_{\alpha^*}) = (1-\rho_{ZW}^2)\mathrm{Var}(Z)\,,$$

where $\rho = \frac{\mathrm{Cov}(Z,W)}{\sigma_Z\sigma_W}$ is the correlation coefficient.

Usually, α^* is estimated during the simulation run by estimating the required covariances from empirical data.

Example 1. For illustration, consider the following estimator of $\pi/4$:

$$Z = 1_{\{U_1^2+U_2^2\le 1\}}$$

The indicator $W_1 = 1_{\{U_1+U_2\le 1\}}$ of the lower triangle is positively correlated with Z, and its mean is 0.5. We have (e.g., by simulation) that $1-\rho^2 \approx 0.727$.

Similarly, the indicator $W_2 = 1_{\{U_1+U_2\ge\sqrt{2}\}}$ is *negatively* correlated with Z, has mean $(2-\sqrt{2})^2/2$, and $1-\rho^2 \approx 0.25$.

Note that no new RVs need to be sampled at each step, only U_1 and U_2.

Example 2: Stochastic shortest path. Suppose we wish to estimate the expected length of the shortest path in a network with stochastic link weights $X = (X_e,\, e\in E)$. That is,

$$H(X) = \min_{\pi\in\Pi} L_\pi(X)\,, \quad L_\pi(X) = \sum_{e\in\pi} X_e\,.$$

Here π is the set of paths available in the network.

The length of *any* path is positively correlated with $H(X)$ and can be used as a control variate.

3.3 Common Random Numbers

Suppose we wish to estimate the difference

$$\ell = E(X-Y)\,,$$

using

$$\hat{\ell} = \frac{1}{N}\sum_{1}^{N}(X_i - Y_i)\,.$$

One way would be to sample independently from f_X and f_Y, using $X_i = F_X^{-1}(U_i)$, $Y_i = F_Y^{-1}(\tilde{U}_i)$.

A better option is to use the same random numbers for X_i and Y_i:

$$X_i = F_X^{-1}(U_i)\,, \quad Y_i = F_Y^{-1}(U_i)\,.$$

Since F_X and F_Y are increasing functions, we obtain *positive* correlation between X_i and Y_i. Therefore,

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y) < \mathrm{Var}(X) + \mathrm{Var}(Y)\,.$$

(In fact, it is know that this construction maximizes the correlation between RVs X and Y with given marginals.)

Note:

- When estimating the sum $E(X + Y)$, we can similarly use $X_i = F_X^{-1}(U_i)$, $Y_i = F_Y^{-1}(1 - U_i)$.
- The same trick can be used to estimate $\ell = E(X)$, by using the partition

$$\ell = \frac{1}{2}E(X^a + X^b)\,, \quad X^a, X^b \sim f_X\,.$$

Example: Symmetric Distributions. Suppose that the distribution f_X of X is symmetric about its mean μ. Then

$$\frac{1}{2}(F^{-1}(U) + F^{-1}(1 - U)) = \mu\,,$$

That is, we get μ precisely with one sample.

This not that very surprising, since we could deterministically compute $\mu = F^{-1}(0.5)$. However, if f is *nearly* symmetric this can be useful.

Further comments:

- The efficiency improvements accomplished by these methods are seldom dramatic.
- Common RVs are often used in comparative simulations, which compare system performance under different parameters or controls.

3.4 Stratified Sampling

Here the sample space Ω is decomposed into disjoint region, $\Omega_1, \ldots, \Omega_K$ (called *strata*) of known probabilities, and the mean is estimated on each separately. The choice of the strata should be so that $Z = H(X)$ is as homogeneous as possible (low variance) on each stratum.

Let $I(\omega) = k$ if $\omega \in \Omega_k$. By conditioning we get

$$E(Z) = \sum_k p(I = k) E(Z|I = k) \triangleq \sum_k p_k Z^{(k)}$$

Here (p_k) are assumed known, and each $Z^{(k)}$ can be estimated by

$$\hat{Z}^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} Z_i^{(k)}$$

where $Z_i^{(k)}$ is sampled from $p(Z|I = k)$. We then use

$$\hat{\ell} = \sum_k p_k \hat{Z}^{(k)} .$$

Example 1: Suppose $Z = H(X)$, $X = g(U, Y)$ where $U \in [0, 1]$ is a uniform RV independent of Y. We can divide the interval $[0, 1]$ into disjoint interval of lengths (probabilities) (p_k). To sample $Z^{(k)}$ we simply sample $U^{(k)}$ from the k-th interval. Note that the choice of interval to sample from (i.e,, of N_k) is deterministic and not stochastic.

Example 2: Suppose X is generated by a *mixture distribution*,

$$f_X(x) = \sum w_k f_k(x) , \quad w_k > 0, \sum w_k = 1$$

(which corresponds to choosing $X_k \sim f_k$ w.p. w_k). Here we can naturally identify $p_k = w_k$, $Z^{(k)} = H(X_k)$ with $X_k \sim f_k$.

The resulting variance of the stratified estimator is

$$\sigma^2_{\text{str}} \triangleq \text{Var}(\hat{\ell}) = \sum_k p_k^2 \frac{\sigma_k^2}{N_k} , \quad \sigma_k^2 \triangleq \text{Var}(Z|I = k)$$

It remains to choose the sample sizes (N_k). The simple choice of $N_k = N p_k$ already guarantees

$$\sigma^2_{\text{str}} = \frac{1}{N} \sum_k p_k \sigma_k^2 \leq \frac{1}{N} \text{Var}(Z)$$

An optimal choice of (N_k) depends of course also on the σ_k's (how?), which can be estimated online.

3.5 Conditional Monte Carlo

Here we also use a conditional decomposition of the problem, which can be viewed as complementary to the strata decomposition. Suppose there exists an RV Y (on the same sample space as X) so that

$$h(y) \overset{\triangle}{=} E(H(X)|Y = y)$$

can be computed analytically. A typical case is when $H(X) = H(X_1, Y)$.

Then

$$\ell = E(H(X)) = E(E(H(X)|Y)) = E(h(Y)) \,.$$

Hence $h(Y)$ is an unbiased estimator of ℓ, and

$$\mathrm{Var}(h(Y)) \leq \mathrm{Var}(H(X))$$

(show that).

The estimation procedure is as follows:

1. Generate samples $Y_1, \ldots, Y_N$ from f_Y
2. Compute $h_i = E(H(X)|Y = Y_i)$
3. Compute $\hat{\ell}_N = \frac{1}{N} \sum_{i=1}^{N} h_i$

We note that for this procedure to be effective, Y should be easy to generate, and $E(H(X)|Y = y)$ easy to compute.

The general approach of mixing analytical and numeric (simulation) computation in statistics is known as *Rao-Blackwellization.*

Example 1: Random Gaussian Sums. Let

$$\ell = \mathbb{P}(\sum_{k=1}^{Y} V_k \leq x) \,,$$

where (V_k) is an iid sequence of $N(0, 1)$ RV's, and $Y \geq 0$ is an independent, integer-valued RV.

Clearly,

$$\mathbb{P}(\sum_{k=1}^{Y} V_k \leq x | Y = K) = \mathbb{P}(W_K \leq x) = \Phi(\frac{x}{\sqrt{K}})$$

where $W_K \sim N(0, K)$, and $\Phi(x) = 0.5 + 0.5\text{erf}(x/\sqrt{2})$ is the cdf of the standard normal RV. We obtain the estimator

$$\hat{\ell}_N = \frac{1}{N} \sum_{i=1}^{N} \Phi(\frac{x}{\sqrt{Y_i}})$$

where $(Y_1, \dots, Y_N)$ are sampled from f_Y.

Example 2: Permutation Monte-Carlo for Reliability Models. [RK, 5.4.1] We are given a system of n components, where component (or link) j can fail with probability q_j (independently of the others), or remain functional with probability $p_j = 1 - q_j$. Let $X = (Y_1, \dots, Y_n) \in \{0, 1\}^n$ be the system state vector, with $\mathbb{P}(Y_j = 1) = p_j$. The system failure is captured by a binary function $H(X)$, where $H(X) = 0$ denotes failure. We are interested in the system failure probability $\mathbb{P}(H(X) = 0)$. As this probability is typically small, crude MC is not effective.

Reliability models such as this one are an important application area for MC methods. Here we describe a particular Conditional MC method, the Permutation method, which applies to our model.

Consider the system as a dynamic network, which starts at $t = 0$ with all links failed, and each link is repaired at time t_j, distributed as an exponential RV with rate $\mu_j = \ln(q_j^{-1})$. Then, at $t = 1$ the probability of $Y_j = 1$ is p_j. Thus, the distribution of the system state X coincides with that of state $X(1)$ of the dynamic network at time $t = 1$.

Let Π denote the *order* in which links are repaired, corresponding to ordering (t_j) in increasing order. Then Π is a random permutation. Conditioned on $\Pi = \pi$, let $b(\pi)$ denote the number of repairs required (in the order defined by π) to bring the system up. Then

$$g(\pi) \stackrel{\triangle}{=} \mathbb{P}(H(X(1)) = 0 | \Pi = \pi) = \mathbb{P}(t_{b(\pi)} > 1 | \pi)\,,$$

and

$$\mathbb{P}(H(X) = 0) = \mathbb{P}(H(X(1)) = 0) = E(g(\Pi))\,.$$

We observe that sampling a permutation π can be an be done by ordering $\{t_j \sim \text{Exp}(\mu_j)\}$ in increasing order, as above, or by sampling links without replacement with probabilities proportional to μ_i. Sampling can stop once $b(\pi)$ is reached.

It remains to compute $g(\pi)$. It can be shown that, conditioned on $\Pi = \pi$, $t_{b(\pi)}$ is distributed as the sum of independent RVs

$$\tau_1 + \cdots + \tau_{b(\pi)}$$

where $\tau_i \sim \text{Exp}(\lambda_j)$, $\lambda_j = \sum_{i=j}^{n} \mu_{\pi(i)}$. Therefore $g(\pi)$ can be computed using convolutx ion, or other methods (Laplace transform, Markov chain evolution).

We obtain the following Conditional MC algorithm. For $i = 1, \ldots, N$,

1. Sample a permutation π_i .
2. Determine $b(\pi_i)$.
3. Compute $g(\pi_i)$.

The estimator is $\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} g(\pi_i)$.