Variance Reduction Methods

Introduction

In this lecture we revisit the problem of reducing the variance of the samples that are generated to estimate the mean value λ of some random variable X. Recall that this has the effect of reducing the standard error and increasing the confidence in the accuracy of the estimate. In this lecture we look at four other variance-reduction techniques that are described below.

Antithetic variables Instead of a single sample sequence $X_1, \ldots, X_n$ sampled from X, this method instead uses an additional variable Y for which $E[Y] = E[X] = \lambda$, but is negatively correlated with X. The i th sample now becomes $\frac{X_i+Y_i}{2}$.

Control variables A **control variable** is a random variable Y for which $E[Y] = \mu$ is known, and for which sampling variable X is replaced with $X + c(Y - \mu)$, where c is chosen so as to minimize $\text{Var}(X + c(Y - \mu))$.

Conditional sampling Conditional sampling selects a suitable variable Y for which $E[X|Y]$ has a smaller variance than X. Y is chosen so that, when it is observed, the uncertainty (i.e. entropy) of X gets reduced.

Stratified (Regional) sampling Stratified sampling is a divide-and-conquer strategy that partitions the domaiin of X into subdomains, and then a sampling plan is devised for each subdomain. Moreover, common sense suggests that the smaller the subdomain, the less variance that should exist in that region, and so a variance reduction may be achievable in each of the subdomains that comprise the original domain.

Antithetic Variables

Proposition 1. If X and Y are two random variables, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

And in general, for $n \geq 1$ random variables $X_1, \ldots, X_n$,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i \neq j} \mathrm{Cov}(X_i, X_j).$$

Proof of Proposition 1.

$$\mathrm{Var}(X_1 + \cdots + X_n) = E[(X_1 + \cdots + X_n)^2] - E^2[X_1 + \cdots + X_n] =$$

$$E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i \neq j} X_i X_j\right] - \left(\sum_{i=1}^{n} E[X_i]\right)^2 =$$

$$\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i \neq j} E[X_i X_j] - \sum_{i=1}^{n} E^2[X_i] - 2\sum_{i \neq j} E[X_i]E[X_j] =$$

$$\sum_{i=1}^{n} E[X_i^2] - \sum_{i=1}^{n} E^2[X_i] + 2\sum_{i \neq j} E[X_i X_j] - 2\sum_{i \neq j} E[X_i]E[X_j] =$$

$$\sum_{i=1}^{n} \left(E[X_i^2] - E^2[X_i]\right) + 2\sum_{i \neq j} (E[X_i X_j] - E[X_i]E[X_j]) =$$

$$\sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i \neq j} \mathrm{Cov}(X_i, X_j).$$

Now suppose X and Y are random variables for which $E[Y] = E[X] = \lambda$. Then

$$E[\frac{X+Y}{2}] = \frac{1}{2}E[X] + \frac{1}{2}E[Y] = \lambda/2 + \lambda/2 = \lambda.$$

Moreover,

$$\mathrm{Var}(\frac{X+Y}{2}) = \frac{1}{4}[\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)].$$

Notice that, if X and Y are dependent and have negative correlation, then $\mathrm{Cov}(X,Y) < 0$, which causes a reduction in variance (compared to the case where X and Y are uncorrelated). Note: by definition, the correlation between X and Y is defined as

$$\mathrm{cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Example 1. Given $U \sim U(0,1)$, show that U and $1-U$ are negatively correlated.

From past examples (e.g. recall the Network Reliability problem) we see that the random variable X that is sampled in order to estimate $E[X] = \lambda$, is often a function of one or more $U(0,1)$ random variables $U_1, \ldots, U_n$. Thus, if we express X as $X = h(U_1, \ldots, U_n)$, and if $h(U_1, \ldots, U_n)$ is negatively correlated with $h(1-U_1, \ldots, 1-U_n)$, then a reduction in variance occurs when two independent samples of X are replaced with $\frac{1}{2}(h(U_1, \ldots, U_n) + h(1-U_1, \ldots, 1-U_n)$. This turns out to be true for the case when h is monotone (either increasing or decreasing) with respect to each of its domain variables. The proof of the following proposition may be found in the Appendix of Chapter 9 of Sheldon Ross's "Simulation".

Proposition 2. If $h(x_1, \ldots, x_n)$ is a monotone function with respect to each variable x_i, $i = 1, \ldots, n$, then $h(U_1, \ldots, U_n)$ and $h(1-U_1, \ldots, 1-U_n)$ are negatively correlated.

Example 2. Verify that $h(x,y) = x(1-y)$ is monotone increasing with respect to x and monotone decreasing with respect to y. Verify that $U_1(1-U_2)$ and $(1-U_1)U_2$ are negatively correlated.

Example 3. Consider a connected graph G with m edges $e_1, \ldots, e_m$, and for which e_i has a probability equal to p_i of being removed from G. Moreover, assume that, for all $i \neq j$, the event of e_i being removed is independent of e_j being removed. If D denotes the event of G becoming disconnected (as the result of randomly removing some of its edges), then express D as a function $h(U_1, \ldots, U_m)$, where h is montone and $U_1, \ldots, U_m \sim U(0, 1)$.

Example 4. Consider a G/G/1 queueing system with a FIFO queue discipline, and let W_i^Q, $i = 1, \ldots, n$, denote the time that customer i spends in the queue. If the goal is to estimate $X = W_1^Q + \cdots + W_n^Q$, then prove that X can be expressed as a function $h(U_1, \ldots, U_m)$, where h is montone and $U_1, \ldots, U_m \sim U(0, 1)$, assuming that all customer arrival and service times are independent, and can be sampled using the inverse transform technique applied to continuous CDF's F_A and F_S, where, e.g., F_A is used to sample arrival times.

Example 5. Consider a squence of random numbers $U_1, U_2, \ldots$ and let random variable $N \geq 2$ be observed as the first index n for which $U_n > U_{n-1}$. Prove that $E[N] = e$, and $\mathrm{Var}(N) = 3e - e^2 \approx 0.7658$. Hint: argue that $P(N > n) = 1/n!$, and use the fact that $P(N = n) = P(N > n-1) - P(N > n)$.

In the above example, notice that a similar argument shows that $E[M] = e$, where $M \geq 2$ be observed as the first index n for which $U_n \leq U_{n-1}$. Moreover, if we can show that M and N are negatively correlated, then a sampling procedure that observes both M and N for a sequence $U_1, U_2, \ldots$ may use $(M + N)/2$, and have a lower variance. To compute the variance of $M + N$, notice that we always observe either $M = 2$ or $N = 2$. In other words, either M or N will be observed when $n = 2$. Now let, T_N (respectively T_M) denote the number of additional U samples (beyond 2) that must be observed until a U sample exceeds (respectively, is less than or equal to) its predecessor. Then

$$\mathrm{Var}(M + N) = \mathrm{Var}(M + N|M = 2)P(M = 2) + \mathrm{Var}(M + N|N = 2)P(N = 2) =$$

$$\mathrm{Var}(4 + T_N)(1/2) + \mathrm{Var}(4 + T_M)(1/2) = \mathrm{Var}(T_N),$$

since T_M and T_N have the same distribution. To get $\mathrm{Var}(T_N)$, we use the fact that

$$E[N] = e = E[N|N = 2]P(N = 2) + E[N|N > 2]P(N > 2) = (2)(1/2) + (2 + E[T_N])(1/2) =$$

$$2 + \frac{1}{2}E[T_N] \Rightarrow E[T_N] = 2e - 4.$$

Also,

$$E[N^2] = 3e = E[N^2|N = 2]P(N = 2) + E[N|N > 2]P(N > 2) = (4)(1/2) + E[(2 + T_N)^2](1/2) =$$

$$2 + 2 + 2E[T_N] + \frac{1}{2}E[T_N^2] \Rightarrow E[T_N^2] = 8 - 2e.$$

Thus, $\mathrm{Var}(T_N) = (8 - 2e) - (2e - 4)^2 = 14e - 4e^2 - 8 \approx 0.5$. Therefore, if N_1 and N_2 are two independent samples of N, then $\mathrm{Var}(N_1 + N_2) \approx 1.5$, whereas

$$\mathrm{Var}(M + N) = \mathrm{Var}(4 + T_N) = \mathrm{Var}(T_N) \approx 0.5,$$

and so we acheive a variance reduction of 1/3.

Method of Control Variables

Suppose we are interested in estimating $E[X] = \lambda$. Then instead of exclusively sampling X, we may also introduce another variable Y, called the **control variable**, that is correlated with X, and sample $Z = X + c(Y - \mu)$, where c is a constant, and $\mu = E[Y]$. Notice that

$$E[Z] = E[X] + cE[(X - \mu)] = \lambda + c(\mu - \mu) = \lambda + c0 = \lambda,$$

and so sampling Z will lead to an unbiased estimator. Moreover,

$$\text{Var}(X + c(Y - \mu)) = \text{Var}(X) + c^2\text{Var}(Y) + 2c\text{Cov}(X, Y),$$

which is minimized when

$$c = -\frac{\text{Cov}(X, Y)}{\text{Var}(Y)},$$

and yields a variance equal to

$$\text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}.$$

Therefore, the method of control variables will always reduce variance, so long as X and Y have nonzero correlation.

One issue faced by this method is that usually neither $\text{Cov}(X, Y)$ nor $\text{Var}(X)$ are known beforehand. However, as a preprocessing step, we may obtain the estimators

$$\hat{\text{Cov}}(X, Y) = \sum_{i=1}^{n}(X_i - \hat{\lambda})(Y_i - \mu)/(n - 1)$$

and

$$\hat{\sigma_Y^2} = \sum_{i=1}^{n}(Y_i - \mu)^2/(n - 1)$$

to obtain the estimate

$$\hat{c} = \frac{\sum_{i=1}^{n}(X_i - \hat{\lambda})(Y_i - \mu)}{\sum_{i=1}^{n}(Y_i - \mu)^2}$$

for c.

Example 6. Recall from Example 3 that the event D of G becoming disconnected may be expressed as $D = h(E_1, \ldots, E_m)$, where E_i is the event that edge e_i is removed, $i = 1, \ldots, m$, and h is a monotone increasing function. Notice also that

$$S = \sum_{i=1}^{m} E_i$$

is also montone increasing with respect to $E_1, \ldots, E_m$, and represents the number of removed edges. Thus, D and S are positively correleated, and so

$$Z = D + c(S - \sum_{i=1}^{m} p_i),$$

has less variance than D, where

$$c = -\frac{\text{Cov}(D,S)}{\text{Var}(S)} = -\frac{\text{Cov}(D,S)}{\sum_{i=1}^{m} p_i(1-p_i)}.$$

Example 7. Consider a queueing system for which customers arrive according to some interarrival distribution, and are served according to service distribution G. Note that service times are independent of arrival times. Then, letting $N(t)$ denote the number of customers that arrive during a fixed interval $[0,t]$, if we desire to estimate the average total time cutomers spend int the system, denoted by,

$$X = \sum_{i=1}^{N(t)} W_i,$$

where W_i denotes the time customer i spent in the system, then this is positively correlated with

$$Y = \sum_{i=1}^{N(t)} S_i,$$

where S_i denotes the time customer i spent in service. Moreover, since service times are independent of arrival times,

$$E[Y] = E[G]E[N(t)],$$

and thus

$$Z = X + c(Y - E[G]E[N(t)]),$$

has lower variance than X, where

$$c = -\frac{\text{Cov}(X,Y)}{\text{Var}(Y)}.$$

Example 8. Suppose the goal is to estimate $\lambda = E[e^U]$, where $U \sim U(0,1)$. Determine the reduction in variance that is obtained when U is used as a control variable.

Variance Reduction by Conditioning

There are times when knowledge of the value of a random variable Y can reduce the variance of what can be observed for X. In this case $E[X|Y]$ may have less variance than X. Note also that $E[E[X|Y]] = E[X]$, and so $E[X|Y]$ is an unbiased estimator of $E[X]$. Moreover, the amount of reduction in variance is determined by the following proposition.

Proposition 3. $\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$.

Proof of Proposition 3.

$$E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]) = E[E[X^2|Y] - E^2[X|Y]] + E[E^2[X|Y] - (E[E[X|Y]])^2] =$$
$$E[E[X^2|Y]] - E[E^2[X|Y]] + E[E^2[X|Y]] + E^2[X] = E[X^2] - E^2[X] = \text{Var}(X).$$

Corollay 1. $\text{Var}(X) - \text{Var}(E[X|Y]) = E[\text{Var}(X|Y)]$.

Example 9. Recall the method of approximating π that was presented in the Monte Carlo lecture.

That method is equivalent to defining an indicator variable I as

$$I = \begin{cases} 1 & \text{if } V_1^2 + V_2^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $V_1, V_2 \sim U(-1,1)$. Recall that $E[I] = \frac{\pi}{4}$, and $\text{Var}(I) = \frac{\pi}{4}(1 - \frac{\pi}{4})$.

We now show that $E[I|V_1]$ has a smaller variance. Indeed,

$$E[I|V_1 = v] = P(V_1^2 + V_2^2 \leq 1|V_1 = v) = P(V_2^2 \leq 1 - v^2|V_1 = v) =$$

$$P(V_2^2 \leq 1 - v^2) = P(-(1-v^2)^{1/2} \leq V_2 \leq (1-v^2)^{1/2}) =$$

$$\int_{-(1-v^2)^{1/2}}^{(1-v^2)^{1/2}} \frac{1}{2} dv_2 = (1-v^2)^{1/2}.$$

Thus, $E[I|V_1] = (1 - V_1^2)^{1/2}$. However, since V_1^2 has the same distribution as U^2, $U \sim U(0,1)$, it follows that

$$E[I|V_1] = (1 - U^2)^{1/2}.$$

Finally,

$$\text{Var}(1 - U^{1/2}) = E[(1 - U^2)] - (\frac{\pi}{4})^2 = \frac{2}{3} - (\frac{\pi}{4})^2 = 0.0498,$$

and

$$\frac{\text{Var}(I|V_1)}{\text{Var}(I)} = \frac{0.0498}{0.1685} = 0.295,$$

yielding a 70.5% reduction in variance.

Example 10. Consider a procedure for killing cancerous cells that has the side effect of also killing non-cancerous cells. Assume there are m cancerous cells, n non-cancerous cells, and the method kills cells one at a time. Moreover, associated with cell i, $i = 1, \ldots, m+n$, is a weight $w_i > 0$, such that, if cell i has yet to be killed, and S denotes the set of indices of current living cells, then the probability that cell $i \in S$ is killed next is equal to

$$\frac{w_i}{\sum_{j \in S} w_j}.$$

Let N denote the number of remaining non-cancerous cells after the final cancerous cell has been killed, and suppose the goal is to estimate $P(N \geq k)$ for some constant $k \geq 0$. Perhaps the most straightforward way of estimating N is to round by round randomly select a cell to kill, stop if there are no remaining cancer cells, and count the remaining non-cancerous cells. However, if S is the set of indices of currently living cells, then selecting a cell to kill in each round can be accomplished in $\mathrm{O}(|S|)$ steps. Adding this amount to each round, one gets $\mathrm{O}(n^2 + m^2)$ total steps. We now show how conditional sampling can reduce the cost to $\mathrm{O}(m+n)$ steps.

The key idea is to use random exponential variables $T_1, \ldots, T_{m+n}$, where T_i has rate w_i. Then cell i is said to be killed at time T_i. We leave it as an exercise to show that, if S is the set of indices of cells that remain living at time t, then the probability that cell i, $i \in S$, is the next cell killed is equal to

$$\frac{w_i}{\sum_{j \in S} w_j}.$$

Thus, by observing the values of $T_1, \ldots, T_{m+n}$, we may count the number of indices $i > m$ for which $T_i > \max\{T_j | 1 \leq j \leq m\}$, as these will represent the non-cancerous cells that are remaining after the last cancerous cell is killed. Thus, we may observe the event $N \geq k$ in $\mathrm{O}(m+n)$ steps. For example, we may assign $X = 1$ if we observe that the k th largest value of $T_{m+1}, \ldots, T_n$ exceeds the largest value of $T_1, \ldots, T_m$, and assign $X = 0$ otherwise.

Now let Y be a random variable that is equal to the k th largest of $T_{m+1}, \ldots, T_{m+n}$. Then if X denotes the event $N \geq k$, then we see that

$$E[X|Y] = P(\max_{i=1}^{m}(T_i) < Y) = \prod_{i=1}^{m} P(T_i < Y) = \prod_{i=1}^{m}(1 - e^{-w_i Y}).$$

To see why $[E[X|Y]$ has a smaller variance than X, notice that, instead modulating between 0 and 1 like X, each sample of $[E[X|Y]$ consists of a very small number that does not change much in magnitude. Hence, its variance should be significantly smaller than that of X.

Stratfied Sampling

Stratified sampling represents another conditioning approach to reducing variance. Again, suppose the goal is to estimate $\lambda = E[X]$. Suppose Y is a finite random variable with $\mathrm{dom}(Y) = \{y_1, \ldots, y_k\}$ and $p_i = p(y_i)$ is known for each $i = 1, \ldots, k$. Then

$$E[X] = \sum_{i=1}^{k} E[X|Y = y_i] p(y_i).$$

Then λ can be estimated by estimating each of $E[X|Y = y_i]$ with $\hat{\lambda}_i$ and producing the estimate

$$\hat{\lambda} = \sum_{i=1}^{n} \hat{\lambda}_i p_i,$$

where $\hat{\lambda}_i$ is obtained using n_i samples, and $n_1 + \cdots + n_k = n$ and $n_i = np_i$. Notice that

$$\mathrm{Var}(\hat{\lambda}) = \sum_{i=1}^{n} p_i^2 \mathrm{Var}(\hat{\lambda}_i) = \sum_{i=1}^{n} p_i^2 \frac{\mathrm{Var}(X|Y = y_i)}{np_i} = \frac{1}{n} E[\mathrm{Var}(X|Y)].$$

Thus, by Proposition 3, we have

$$\frac{1}{n}\mathrm{Var}(X) - \frac{1}{n}E[\mathrm{Var}(X|Y)] = \frac{1}{n}\mathrm{Var}(E[X|Y]),$$

so that the reduction in variance depends on the amount of variance that occurs in $E[X|Y]$.

Example 10. Suppose on randomly selected weekdays we count the number of students who buy lunch at the Tube sandwich shop from 11:00AM to 1:00PM. Over time we discover that on Monday through Thursday the shop has an arrival rate of one student per minute, while on Fridays the arrival rate reduces to one per four minutes. If both arrival processes are Poisson, then compute $\mathrm{Var}(X)$, $E[\mathrm{Var}(X|Y)]$, and $\mathrm{Var}(E[X|Y])$, where X measures the number of students having lunch on a given day, and Y indicates the day of the week.

Example 11. Suppose that we knew in advance the value $\sigma_i^2 = \mathrm{Var}(X|Y = y_i)$, for all $i = 1, \ldots, k$, and we plan to estimate X via stratefied sampling with n samples, and using n_i samples for the i th stratum. Show that [Var($\hat{\lambda}$) is minimized when

$$n_i = \frac{np_i\sigma_i}{\sum_{i=1}^k p_i\sigma_i}.$$

Exercises

1. Suppose the goal is to estimate
$$\lambda = \int_0^1 e^{x^2} dx.$$
Show that the estimator $e^{U^2}(1+e^{1-2U})/2$ has less variance than $(e^{U_1^2}+e^{U_2^2})/2$, where $U, U_1, U_2 \sim U(0,1)$.

2. Show that if X and Y have the same distribution, then
$$\mathrm{Var}(X+Y)/2 \leq \mathrm{Var}(X).$$

3. Let $X_1, \ldots, X_5$ be independent exponential random variables, each having mean 1, and consider the quantity
$$\lambda = P\left(\sum_{i=1}^{5} iX_i \geq 21.6\right).$$
Explain how simulation can be used to estimate λ and provide the antithetic variable estimator.

4. Suppose X is a random variable whose mean is known. In order to estimate $P(X \leq a)$, we may perform an IMC simulation using
$$I = \begin{cases} 1 & \text{if } X \leq a \\ 0 & \text{if } X > a \end{cases}$$
Suppose X is used as a control variable. Determine the variance reduction obtained compared to only using I, when $X \sim U(0,1)$.

5. Repeat the previous exercise, but now assume $X \sim E(1)$.

6. Suppose that X is an exponential random variable with mean 1. Given another random variable that is negatively correlated with X and that is also exponential with mean 1.

7. Let $U_1, U_2, \ldots \sim U(0,1)$ be a sequence of uniform random variables, and let
$$N = \min_n (U_1 + \cdots + U_n > 1).$$
Prove that $E[N] = e$.

8. Prove that if $U \sim U(0,1)$ and $V \sim U(-1,1)$, then $U^2 = V^2$ have the same distribution.

9. Prove that if T_i, $i = 1, \ldots n$, are independent exponential random variables with rate λ_i, $i = 1, \ldots, n$, then $T = \min\{T_i | 1 \leq i \leq n\}$ is exponential with rate $\lambda_1 + \cdots + \lambda_n$. Hint: use the fact that, by independence,
$$P(T \geq x) = \prod_{i=1}^{n} P(T_i \geq x).$$

10. Suppose $S \sim E(\lambda)$ and $T \sim E(\mu)$ are two independent exponential random variables having resepective rates λ and μ. Prove that $P(S < T) = \frac{\lambda}{\lambda+\mu}$.

11. Use the previous two exercises to prove that if T_i, $i = 1, \dots n$, are independent exponential random variables with rates λ_i, $i = 1, \dots, n$, then

$$P(T_i = \min\{T_j | 1 \leq j \leq n\}) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}.$$

12. Let $X, Y \sim E(1)$ be independent exponential random variables. Suppose we want to estimate $\theta = P(X + Y \leq 2)$. The straightforward approach is to sample X and Y and set $I = 1$ if $X + Y \leq 2$, and $I = 0$ otherwise. Provide a new estimator for θ that uses conditioning. Show how the new estimator can be further improved witht the help of a control variable.

13. Consider the function

$$f(x) = \begin{cases} -x & \text{if } -1 \leq x \leq 0 \\ x^2 & \text{if } 0 \leq x \leq 1 \end{cases}$$

Compute the variance reduction that occurs when estimating $\int_{-1}^{1} f(x)dx$ by using stratefied sampling. In other words, the first approach uses $f(X)/\omega(X)$, where $\omega(x) = 1/2$ is the density function for $X \sim U(-1, 1)$, while the second approach first uses estimator $f_1(x) = -X$ using $X \sim U(-1, 0)$, and then estimator $f_2(x) = X^2$, where $X \sim U(0, 1)$.