3.3 Variance Reduction Techniques

- We have already seen that there can be different Monte Carlo methods for the same problem.
- Different method can differ in efficiency in a number of respects.
  - The amount of analytical work required to found the Monte Carlo estimator.
  - The programming complexity of the algorithm.
  - The computational complexity of the algorithm.
  - The variability of the Monte Carlo estimate.
- The most common way to improve the efficiency of the Monte Carlo experiment is to control the last item of these.

3.3 Variance Reduction Techniques

There are many techniques for variance reduction, we are going to analyze three of them:

- Control variable.
- Antithetic variables.
- Importance sampling.

3.3.1 Control Variable

- Let $\hat{\theta}$ be a Monte Carlo estimate of a parameter θ based on a sample of size n with $\mathrm{E}(\hat{\theta}) = \theta$.
- Suppose that we can find another estimator C such that C and $\hat{\theta}$ are correlated and $\mathrm{E}(C) = \mu$.
- The simplest way to guarantee that C and $\hat{\theta}$ are correlated is to have them based on the same sequence of random numbers.
- C is called a **control variable.**

The control variable estimator is:

$$\boxed{\hat{\theta}_C = \hat{\theta} - \beta(C - \mu)}$$

for some known value of β.

Properties of the new estimator:

- $\mathrm{E}(\hat{\theta}_C) = \theta$.
- $\mathrm{Var}(\hat{\theta}_C) = \mathrm{Var}(\hat{\theta}) + \beta^2 \mathrm{Var}(C) - 2\beta \mathrm{Cov}(\hat{\theta}, C)$.

Theorem 3.3.1 *For a Monte Carlo estimate $\hat{\theta}$ and known control variable C, the minimum variance of $\hat{\theta}_C = \hat{\theta} - \beta(C - \mu)$ is achieved when*

$$\beta = \frac{Cov(\hat{\theta}, C)}{Var(C)}$$

and that minimum variance is

$$Var(\hat{\theta}_C) \geq (1 - \rho^2) Var(\hat{\theta}),$$

where ρ is the correlation coefficient between $\hat{\theta}$ and C, that is

$$\rho = \frac{Cov(\hat{\theta}, C)}{\sqrt{Var(\hat{\theta})}\sqrt{Var(C)}}.$$

The most efficient estimator with control variable is then given by,

$$\boxed{\hat{\theta}_C = \hat{\theta} - \frac{\mathrm{Cov}(\hat{\theta}, C)}{\mathrm{Var}(C)}(C - \mu),}$$

and its variance is given by,

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_C) &= \mathrm{Var}(\hat{\theta}) + \left(\frac{\mathrm{Cov}(\hat{\theta}, C)}{\mathrm{Var}(C)}\right)^2 \mathrm{Var}(C) - 2\frac{\mathrm{Cov}(\hat{\theta}, C)}{\mathrm{Var}(C)}\mathrm{Cov}(\hat{\theta}, C) \\ &= \mathrm{Var}(\hat{\theta}) - \frac{\mathrm{Cov}^2(\hat{\theta}, C)}{\mathrm{Var}(C)} = (1 - \rho^2)\mathrm{Var}(\hat{\theta}).\end{aligned}$$

In general we do not have an analyical value for $\mathrm{Cov}(\hat{\theta}, C)$ then in practice we also use a Monte Carlo estimator to evaluate β.

Exercise: Use the Monte Carlo estimator to estimate π and apply a variance reduction technique with the control variable, $C = \overline{U}$. Compute with R, and present the results in a data.frame as follows.

N	$\hat{\pi}$	$\text{SE}(\hat{\pi})$	$\hat{\pi}_C$	$\text{SE}(\hat{\pi}_C)$
10				
50				
100				
500		...		
1000				
5000				
10000				

3.3.2 Antithetic Variables

Suppose that $\widehat{\theta}$ and $\widehat{\theta}'$ are two unbiased Monte Carlo estimators of the parameter θ. Then

$$\boxed{\widehat{\theta}_A = \frac{1}{2}(\widehat{\theta} + \widehat{\theta}')}$$

is also an unbiased estimator of θ, and

$$\begin{aligned} Var(\widehat{\theta}_A) &= \frac{1}{4}\mathrm{Var}(\widehat{\theta}) + \frac{1}{4}\mathrm{Var}(\widehat{\theta}') + \frac{1}{2}\mathrm{Cov}(\widehat{\theta}, \widehat{\theta}') \\ &\simeq \frac{1}{2}Var(\widehat{\theta}) + \frac{1}{2}Cov(\widehat{\theta}, \widehat{\theta}'). \end{aligned}$$

(Assuming that the two original estimators were roughly equally variable.)

This estimator is the more efficient the more negative be the covariance between $\widehat{\theta}$ and $\widehat{\theta}'$.

- Since we are doing twice as much computational work to estimate θ, there will only be a gain in efficiency when the two estimates are negatively correlated.

- If the random variables that we used are as negatively correlated as possible, then we expect that the resulting estimators will share similar properties.

- When the the random numbers are derived from Uniform random variables this is easy to get. Indeed, if U is a uniform random variable $1 - U$ is also a uniform random variable and the correlation coefficient between these two is -1 so they are maximally negatively correlated.

Antithetic Variables - Example

If

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(u_i)$$

where $u_1, u_2, ..., u_n$, is a sample of independent random variables $U \sim U(0,1)$, then we can consider

$$\hat{\theta}' = \frac{1}{n}\sum_{i=1}^{N} g(1-u_i).$$

The random variables U and $1-U$ are negatively correlated, $cov(U, 1-U) = -\frac{1}{12}$, then in general the antithetic variable estimator

$$\widehat{\theta}_A = \frac{1}{2}(\widehat{\theta} + \widehat{\theta'}) = \frac{1}{2n}\sum_{i=1}^{n}(g(u_i) + g(1-u_i))$$

has lower variance. The reduction is guaranteed when the function g satisfies some conditions that we are going to show next.

Proposition

Let

$$\widehat{\theta}_A = \frac{1}{2}(\widehat{\theta} + \widehat{\theta'}) = \frac{1}{2n}\sum_{i=1}^{n}\left(g(u_i) + g(1-u_i)\right).$$

where $u_1, u_2, ..., u_n$, *is a sample of independent random variables* $U \sim U(0,1)$. *If the function* g *is continuous, strictly monotonic and its first order derivative is also continuous, then,*

$$Var(\hat{\theta}_A) \leq \frac{1}{2} Var(\hat{\theta}).$$

Outline of the proof:

Consider the auxiliary function

$$\phi(x) = \int_0^x g(1-u)du - \theta x, \qquad \phi(0) = \phi(1) = 0$$

Show that if $g'(x) \geq 0$, then $\phi(x) \geq 0$.

Finally,

$$\int_0^1 \phi(x)g'(x)dx \geq 0 \Leftrightarrow \int_0^1 g(u)g(1-u)du \leq \theta^2.$$

In general:
If X_i is a random variable with distribution F, then $X_i = \mathrm{F}^{-1}(U_i)$ and $Y_i = \mathrm{F}^{-1}(1 - U_i)$ are iid and negatively correlated.

Example: $U \sim \mathrm{U}(0, 1)$ then $X = -log(U)$ and $Y = -log(1 - U)$ are both exponentially distributed and have negative correlation.

Exercise: Show that $cov(X, Y) \leq 0$.

3.3.3 Importance Sampling

If we consider the parameter,
$\theta = \mathrm{E}(h(x)) = \int_D g(x)p(x)dx$. The Monte Carlo estimator would be

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

where $x_1, x_2, ..., x_n$ is a sample with independent elements of a random variable with density $p(x)$. If $f(x)$ is another density function with support D, then the following equality is true:

$$\int_D g(x)p(x)dx = \int_D \frac{g(x)p(x)}{f(x)} f(x)dx.$$

Then the estimator,

$$\boxed{\hat{\theta}_{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(y_i)p(y_i)}{f(y_i)}}$$

where $y_1, y_2, ..., y_n$ is a sample with independent elements of a random variable with density $f(x)$, is an alternative estimaor of θ.

Theorem 3.3.3 *The minimal variance of the estimator $\widehat{\theta}_{IS}$ occurs for*

$$f(x) = \frac{|g(x)|p(x)}{\int_D |g(x)|p(x)dx} \text{ and the minimal variance is}$$

$$Var(\widehat{\theta}_{IS}) \geq \frac{1}{N}\Big\{\Big[\int_D |g(x)|p(x)dx\Big]^2 - \theta^2\Big\}.$$

Proof: Use the Cauchy-Schwartz inequality:

$$\Big(\int_D |a(x)b(x)|dx\Big)^2 \leq \Big(\int_D |a(x)|^2 dx\Big)\Big(\int_D |b(x)|^2 dx\Big).$$

Remarks:

- It follows from the previous theorem that if $g(x)$ has constant sign then the variance can be reduced to 0.
- Unfortunately this is never possible in practice since the optimal density depends on an integral which is of the same type as the unknown θ.

However we can observe tha the Importance sampling estimator is more eficiente if the function $\frac{g(x)p(x)}{f(x)}$ is approximately constant, since this way, when the sample values vary, the values $\frac{g(y_i)p(y_i)}{f(y_i)}$ remain constant, which implies a smaller variability of the estimator.

This observation allows us to conclude that:

$\rightarrow$ Since we cannot find the optimal density, we can look for a density which has a shape close to that of $g(x)p(x)$ and this should reduce the variance of the estimator.

- Often we will decide to sample from a specific type of density and then try to find parameters which make the density take on the required shape.

- In some cases the best parameters can be found but usually we will rely on plotting the density for different parameters and comparing them to the required shape.

Importance Sampling

An alternative density function $f(x)$ which is commonly used is,

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

where $M(t) = E_p(e^{tx}) = \int e^{tx}\,(x)\, p(x)dx$. The parameter t should be chosen by taking into account the criterion described above.

- If $p(x) = \lambda e^{-\lambda x}$, then $f_t(x) = (\lambda - t)e^{-(\lambda - t)x}$.
- If $p(x) = p^x(1-p)^{1-x}, x = 0, 1, ..,$ (Bernoulli with parameter p) then

  $f_t(x) = \left(\frac{pe^t}{pe^t + 1 - p}\right)^x \left(\frac{1-p}{pe^t + 1 - p}\right)^{1-x}$ is a Bernoulli random variable with parameter $p_t = \frac{pe^t}{pe^t + 1 - p}$.
- If $p(x)$ is the density of a Normal random variable $\mathrm{N}(\mu, \sigma^2)$ then $f_t(x)$ is also the density of a Normal but with adjusted mean, $\mathrm{N}(\mu + \sigma^2 t, \sigma^2)$.

Standard deviation estimated by Monte Carlo of the importance sampling estimator:

$$\mathrm{SE}(\widehat{\theta}_{\mathrm{IS}}) = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{g(x_i)p(x_i)}{f(x_i)} - \widehat{\theta}_{\mathrm{IS}}\right)^2}.$$