Antithetic Variates

Generate: $u_1, \ldots, u_n \sim U(0,1)$

Let $x_i = G^{-1}(u_i), x_i^* = G^{-1}(1-u_i)$

$$\begin{aligned}\hat{\mu}_f &= \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} f(x_i) + \frac{1}{n}\sum_{i=1}^{n} f\left(x_i^*\right)\right\} \\ &= \frac{1}{2n}\left\{\sum_{i=1}^{n} f(x_i) + \sum_{i=1}^{n} f\left(x_i^*\right)\right\}\end{aligned}$$

Example:

Let $X \sim Exp(2)$ and we want to find

$$\begin{aligned}E\left[\sqrt{X}\right] &= \int 0.5\sqrt{x}e^{-x/2}dx \\ &= \sqrt{2}\Gamma(1.5) = 1.253314\end{aligned}$$

Generate n = 1000 values from $Exp(2)$ and use Antithetic variates

Sampler	Estimate	SE
U sample	1.239673	0.0199
1 – U sample	1.258362	0.0205
Antithetic	1.249017	0.0032

The error with antithetic estimate is -0.0043.

If a single sample of n = 2000 was taken, the standard error would be approximately 0.0143.

The gain in efficiency due to antithetic variates is approximately 20.25 (the square of the ratio of the standard errors).

To get the same efficiency out of a single sample, almost 40,000 samples would be needed.

Antithetic variate generation

If x_i is contained in the sample, then the corresponding sample that needs to be added is

$$x_i^* = G^{-1}\left(1 - G\left(x_i\right)\right)$$

This approach is reasonable when $G(x)$ and $G^{-1}(x)$ are nice functions.

In fact you only need $G^{-1}(x)$ to be nice as you can use the procedure described at the start of the class based on the uniform distribution to generate the samples needed.

Symmetric distributions

If $X \sim G$ has a symmetric distribution around a mean μ (e.g. Normal, Logistic, etc), the antithetic variates approach is easy since

$$x_i^* = 2\mu - x_i$$

This idea can also be expanded if $h(X)$ is symmetric for some monotonic function h.

For example if X is lognormal, then $\log X$ is a symmetric random with mean μ. Then the antithetic variate is

$$x_i^* = \frac{e^{2\mu}}{x_i}$$

In general the antithetic variate satisfies

$$x_i^* = h^{-1}\left(2\mu - h\left(x_i\right)\right)$$

when there is a symmetrizing function $h(x)$.

Note for the lognormal example, I wouldn't implement it in this fashion. Instead I would generate $z_1, \ldots, z_n \sim N(0,1)$ and set

$$x_i = e^{\mu + \sigma z_i}, x_i^* = e^{\mu - \sigma z_i}$$

since most lognormal generators start with normal random variables in the first place.

Control Variates

Similar to antithetic variates where you want to use correlation to reduce variability

The underlying idea is to look at

$$E\left[f(X)\right] = E\left[f(X) - g(X)\right] + E\left[g(X)\right]$$

where $E\left[g(X)\right]$ is known analytically and the random variables $f(X)$ and $g(X)$ are positively correlated.

$$\begin{aligned}&\operatorname{Var}\left(f(X) - g(X)\right)\\&\quad = \operatorname{Var}\left(f(X)\right) - 2\operatorname{Cov}\left(f(X), g(X)\right) + \operatorname{Var}\left(g(X)\right)\end{aligned}$$

If $f(X)$ and $g(X)$ are highly enough correlated, this will have a smaller variance than $\text{Var}(f(X))$.

This implies that the estimate of $E[f(X)]$

$$\hat{\mu}_{f,C} = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \left(\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mu_g\right)$$

will have a smaller variance than

$$\hat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

This approach has ties to regression.

Let $\mu_g = E[g(X)]$. Then the original formulation can be though of as looking at

$$f(X) - (g(X) - \mu_g)$$

Instead of this, lets look at

$$f_b(X) = f(X) - b(g(X) - \mu_g)$$

For all *b*, $E[f_b(X)] = E[f(X)]$.

Thus the original problem can be modified by choosing the b to minimize $\mathrm{Var}\left(f_b\left(X\right)\right)$, which can be done with

$$b = \frac{\mathrm{Cov}\left(f\left(X\right), g\left(X\right)\right)}{\mathrm{Var}\left(g\left(X\right)\right)} = \rho\frac{\sigma_f}{\sigma_g}$$

The idea behind this method is that by using the control variate $g\left(X\right)$, we can see how likely the estimate of $E\left[f\left(X\right)\right]$ just based on the sampled $f\left(x_1\right),\ldots,f\left(x_n\right)$ is off.

With this adjustment, the estimate of $E\left[f\left(X\right)\right]$ is

$$\hat{\mu}_{f,C,b} = \frac{1}{n}\sum_{i=1}^{n} f\left(x_i\right) - b\left(\frac{1}{n}\sum_{i=1}^{n} g\left(x_i\right) - \mu_g\right)$$

The variance of this estimator is

$$\mathrm{Var}\left(\hat{\mu}_{f,C,b}\right) = \frac{1}{n}\left(\sigma_f^2 - 2b\sigma_{fg} + b^2\sigma_g^2\right)$$

The various variance and covariance terms can be estimated using the standard unbiased estimators.

Example:

Let $X \sim Exp(2)$ and we want to find

$$E\left[\sqrt{X}\right] = \int 0.5\sqrt{x}e^{-x/2}dx$$
$$= \sqrt{2}\Gamma(1.5) = 1.253314$$

Let

$$f(x) = \sqrt{x};\ g(x) = x$$

We know that $E[X] = 2$ and it can be shown that

$$\text{Cov}\left(\sqrt{X}, X\right) = 2^{1.5}\left(\Gamma(2.5) - \Gamma(1.5)\right)$$
$$= 1.253314$$

This gives the optimum b of

$b = 0.3133285$

(Note in most problems we can’t figure this covariance out exactly)

Generate n = 1000 values from $Exp(2)$ and use optimum b for $g(x) = x$. For the simulated data we get.

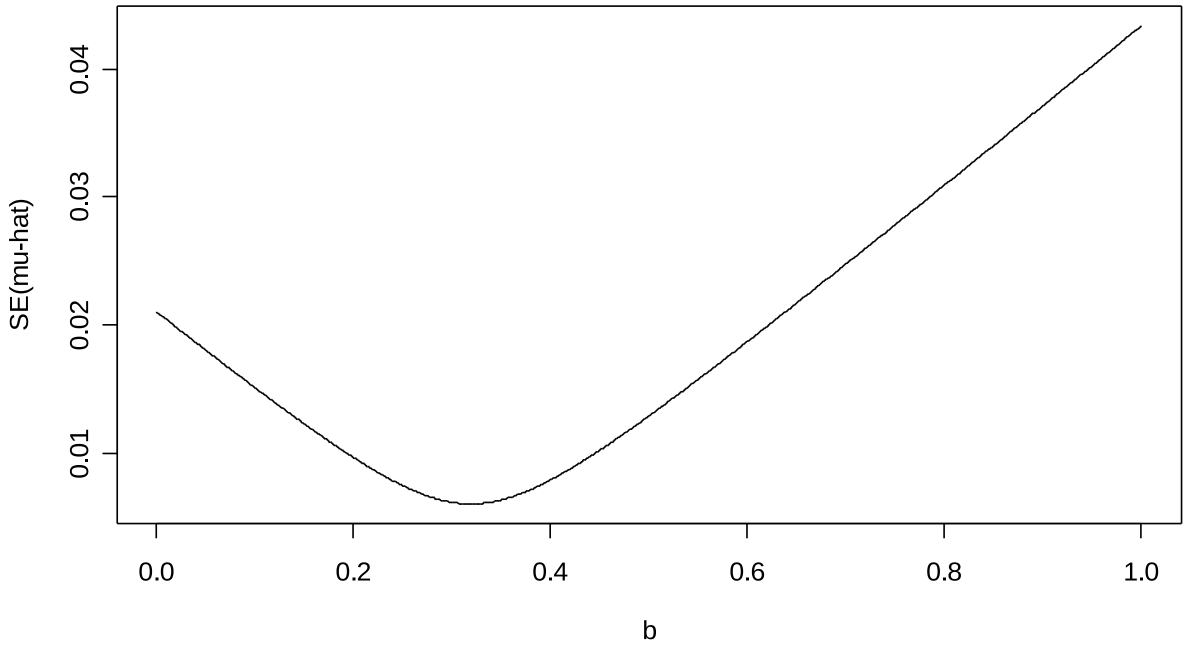
$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i} = 1.256252$$

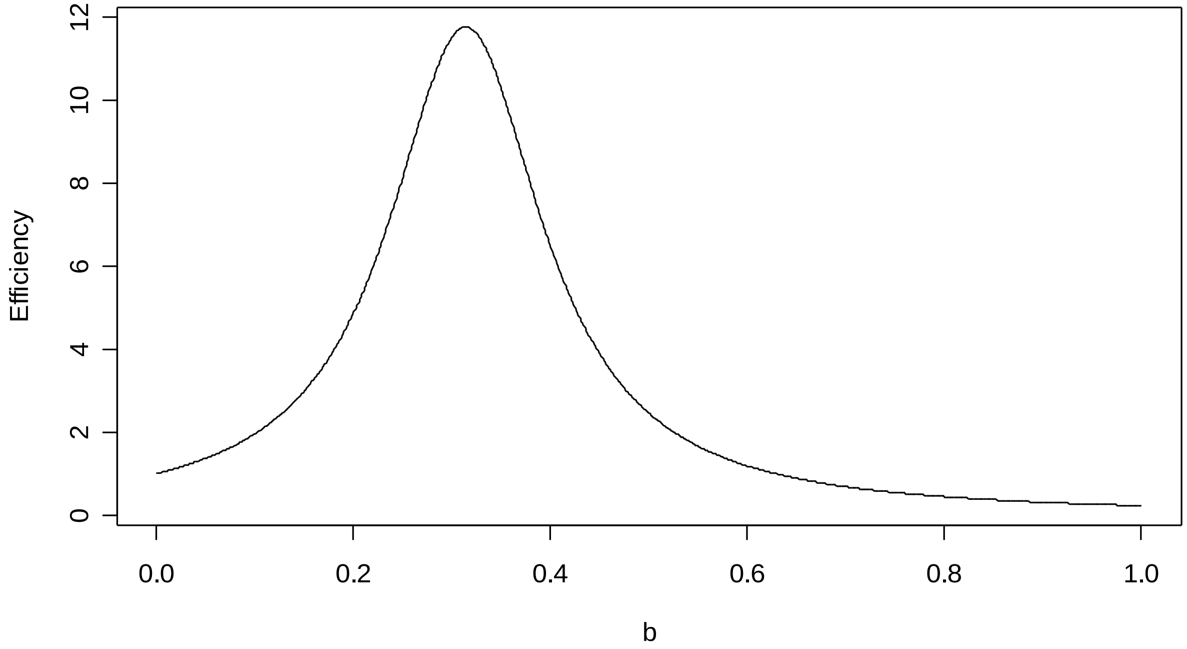
$$\frac{1}{n}\sum_{i=1}^{n}x_i = 2.01843$$

Sampler	Estimate	SE
Crude MC	1.256252	0.020993
Control	1.250476	0.006026

In this example, the control variates approach is almost 12 times more efficient that the standard approach.

We can also look the efficiency with other choices of b.


SE(mu-hat)
0.01
0.02
0.03
0.04
0.0
0.2
0.4
0.6
0.8
1.0
b


Efficiency
0
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
1.0
b

Note that you usually don't know the optimum b since getting $\operatorname{Cov}\left(f(X), g(X)\right)$ is usually intractable analytically. However you can use your sample to estimate it (and $\operatorname{Var}\left(g(X)\right)$ if necessary).

Another equivalent approach (assuming that you estimate $\operatorname{Var}\left(g(X)\right)$) is to run the linear regression of $f(X)$ on $g(X)$ (i.e. fit model

$$f(X) = a + bg(X) + \varepsilon$$

with the observed $\left\{f(x_i)\right\}$ and $\left\{g(x_i)\right\}$).

Note that estimating the optimal b will introduce a slight bias in the estimate of $E\left[f(X)\right]$ and a slightly overoptimistic SE.

However these problems usually aren't enough to worry about and asymptotically it gives the correct answer.

Rao-Blackwellization

The control variate approach used the idea to try to do some analytic computations to improve our estimator.

This next approach is based on the same idea, but focuses more on the function of interest

Suppose that X can be decomposed into two parts $\left(X^{(1)}, X^{(2)}\right)$ and that we are interested in estimating $E\left[f(X)\right] = E\left[f\left(X^{(1)}, X^{(2)}\right)\right]$.

One approach is to sample pairs $\left(X^{(1)}, X^{(2)}\right)$.

This can be done by

Sample $X^{(2)}$ from $g\left(X^{(2)}\right)$

Sample $X^{(1)}$ from $g\left(X^{(1)} \middle| X^{(2)}\right)$

Then estimate $E\left[f(X)\right]$ by

$$\hat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f\left(x_i^{(1)}, x_i^{(2)}\right)$$

Suppose however that $E\left[f(X)\middle|X^{(2)}=x_2\right]$ can be calculated analytically. Then the expectation can be estimated by

$$\hat{\mu}_{f,RB}=\frac{1}{n}\sum_{i=1}^{n}E\left[f(X)\middle|X^{(2)}=x_i^{(2)}\right]$$

Both of these estimators are unbiased.

However

$$\begin{aligned}\operatorname{Var}\left(\hat{\mu}_{f,RB}\right)&=\frac{1}{n}\operatorname{Var}\left(E\left[f(X)\middle|X^{(2)}\right]\right)\\&\leq\frac{1}{n}\operatorname{Var}\left(f(X)\right)=\operatorname{Var}\left(\hat{\mu}_f\right)\end{aligned}$$

This is based on

$$\begin{aligned}\operatorname{Var}\left(f(X)\right)=E&\left[\operatorname{Var}\left(f(X)\middle|X^{(2)}\right)\right]\\&+\operatorname{Var}\left(E\left[f(X)\middle|X^{(2)}\right]\right)\end{aligned}$$

This estimator suggests that wherever possible, do exact calculation over simulation.

Rao-Blackwellized estimators can be used in a wide range of settings, including importance sampling, SIS, or MCMC.

Importance Sampling

Used for a number of purposes:

- Variance reduction
- Allows for difficult distributions to be sampled from.
- Sensitivity analysis
- Reusing samples to reduce computational burden.

Idea is to sample from a different distribution that picks points in “important” regions of the sample space.

Want

$$E\left[f(X)\right] = \int f(x)g(x)dx$$

Instead of sampling from density (or probability mass function) $g(x)$, sample from a distribution with density (or pmf) $h(x)$.

Since we are sampling from the “wrong” distribution we have to make adjustments in our estimator.

$$\begin{aligned} E_g\left[f(X)\right] &= \int f(x)g(x)dx \\ &= \int f(x)\frac{g(x)}{h(x)}h(x)dx \\ &= E_h\left[f(X)\frac{g(X)}{h(X)}\right] \end{aligned}$$

This suggests the following estimation scheme

1) Sample $x_1,\ldots,x_n$ from $h(x)$.

2) Calculate weights

$$w_i = \frac{g(x_i)}{h(x_i)}$$

3) Use estimator

$$\hat{\mu}_{f,IS} = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i)$$

So instead of a regular average, this estimator is a weighted average.

So points that occur more often under $h(x)$ than $g(x)$ get downweighted and those that occur less often get upweighted.

Notice that $\hat{\mu}_{f,IS}$ is an unbiased estimate of $E_g\left[f(X)\right]$ regardless of which proposal distribution $h(x)$ as long as $h(x)$ has the same support as $g(x)$, i.e.

$$g(x) > 0 \text{ implies that } h(x) > 0$$

Note that $h(x) > 0$ can be allowed to occur when $g(x) = 0$, though doing this tends to be inefficient (but there are times you want to do this).

Since $\hat{\mu}_{f,IS}$ is unbiased, the main idea is to pick a distribution $h(x)$ that reduces the variance.

$$\mathrm{Var}_h\left(\frac{f(X)g(X)}{h(X)}\right) = E_h\left[\left(\frac{f(X)g(X)}{h(X)}\right)^2\right] - \mu_f^2$$

To do this, we want $h(x)$ to look like $f(x)g(x)$, i.e. make

$$\frac{f(x)g(x)}{h(x)}$$

look like a constant.

The optimal $h(x)$ satisfies

$$h(x) = \frac{|f(x)|g(x)}{\int |f(x)|g(x)dx}$$

Note that this usually can't be determined, due to the normalizing constant.

However this does give us a motivation for picking $h(x)$.

Example: Monte Carlo Evaluation of a Likelihood Ratio (Genetics Example)

Assume that you have a missing data model where $X = \left(X_{obs}, X_{mis}\right)$. Then the observed data likelihood ratio satistifies

$$l\left(\theta_1, \theta_0\right) = \frac{L\left(\theta_1\right)}{L\left(\theta_0\right)} = \frac{p_{\theta_1}\left(X_{obs}\right)}{p_{\theta_0}\left(X_{obs}\right)}$$

$$= E_{\theta_0}\left[\frac{p_{\theta_1}\left(X_{obs}, X_{mis}\right)}{p_{\theta_0}\left(X_{obs}, X_{mis}\right)} \middle| X_{obs}\right]$$

This can be estimated by sampling $z_1, \ldots, z_n$ from $p\left(X_{mis} \middle| X_{obs}\right)$ calculating

1) $f\left(z_i\right) = \dfrac{p_{\theta_1}\left(X_{obs}, z_i\right)}{p_{\theta_0}\left(X_{obs}, z_i\right)}$

2) $\hat{l}\left(\theta_1, \theta_0\right) = \dfrac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)$

Suppose that you are interested in getting $l\left(\theta_2, \theta_0\right)$, based on this Monte Carlo estimate.

This can be done with the importance sampling estimate

$$\hat{l}\left(\theta_2,\theta_0\right)=\frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)\frac{p_{\theta_2}\left(X_{obs},z_i\right)}{p_{\theta_1}\left(X_{obs},z_i\right)}$$

This can be shown to be an unbiased estimator of $l\left(\theta_2,\theta_0\right)$.

Genetics example:

Observed Data Model

$$\left(Y_1,Y_2,Y_3,Y_4\right)\sim \text{Multi}\left(197,\left(\frac{\lambda}{4},\frac{1-\lambda}{4},\frac{1-\lambda}{4},\frac{2+\lambda}{4}\right)\right)$$

$$g\left(Y|\lambda\right)=\left(\frac{\lambda}{4}\right)^{Y_1}\left(\frac{1-\lambda}{4}\right)^{Y_2+Y_3}\left(\frac{2+\lambda}{4}\right)^{Y_4}$$

Complete Data Model

$$\left(X_1,X_2,X_3,X_4,X_5\right)$$
$$\sim \text{Multi}\left(197,\left(\frac{\lambda}{4},\frac{1-\lambda}{4},\frac{1-\lambda}{4},\frac{\lambda}{4},\frac{1}{2}\right)\right)$$

$$g\left(X\middle|\lambda\right)=\left(\frac{\lambda}{4}\right)^{X_1+X_4}\left(\frac{1-\lambda}{4}\right)^{X_2+X_3}\left(\frac{1}{2}\right)^{X_5}$$

As seen before $X_4\left|Y_4\right.\sim \mathrm{Bin}\left(Y_4,\frac{\lambda}{2+\lambda}\right)$

The complete data likelihood ratio satisfies

$$\frac{g\left(Y,X\middle|\lambda_1\right)}{g\left(Y,X\middle|\lambda_0\right)}=\left(\frac{\lambda_1}{\lambda_0}\right)^{Y_1+X_4}\left(\frac{1-\lambda_1}{1-\lambda_0}\right)^{Y_2+Y_3}$$

Note that this implies the importance sample weight satisfies

$$w_i=c\left(Y_1,Y_2,Y_3,\theta_2,\theta_1\right)\left(\frac{\theta_2}{\theta_1}\right)^{z_i}$$

In this case $\hat{l}\left(\lambda_2,\lambda_0\right)$ has the form

$$\hat{l}\left(\lambda_2,\lambda_0\right)=\frac{c\left(Y_1,Y_2,Y_3,\lambda_2,\lambda_1\right)}{n}\sum_{i=1}^{n}f\left(z_i\right)\left(\frac{\lambda_2}{\lambda_1}\right)^{z_i}$$