Independent Monte Carlo

Interested in

$$E\left[f(X)\right] = \int f(x)\,d\nu(x) = \mu_f$$

Approximate μ_f with

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

where $x_1,\ldots,x_n$ is sampled from the probability measure $\nu(X)$.

Under certain regularity conditions,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to E\left[f(X)\right]$$

If $x_1,\ldots,x_n$ are an iid sample from $\nu(X)$, then

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to E\left[f(X)\right]$$

converges by the law of large numbers, assuming that $E\left[\left|f(X)\right|\right] < \infty$.

In addition, if

$$E\left[f(X)^2\right] = \int f(x)^2 \, d\nu(x) < \infty$$

then by the CLT

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

is approximately normally distributed, with mean $E\left[f(X)\right]$ and variance $\mathrm{Var}\left(f(X)\right)/n$.

The usual estimate of $\sigma_f^2 = \mathrm{Var}\left(f(X)\right)$ is

$$s_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i) - \bar{f}\right)^2,$$

the usual unbiased estimate of $\mathrm{Var}\left(f(X)\right)$.

Example: Confidence interval properties

Want to look at properties of the normal theory 95% CI for μ

$$\bar{x} \pm t_{0.025}\frac{s}{\sqrt{m}}$$

1) Coverage probability (assuming $\mu = 0$)

$$C = E\left[I\left(\bar{x} - t_{0.025}\frac{s}{\sqrt{m}} \le 0 \le \bar{x} - t_{0.025}\frac{s}{\sqrt{m}} \right) \right]$$

$$= E\left[I\left(\frac{|\bar{x}|\sqrt{m}}{s} < t_{0.025} \right) \right]$$

2) Mean interval width

$$E[w] = E\left[2t_{0.025}\frac{s}{\sqrt{m}} \right] = \frac{2t_{0.025}}{\sqrt{m}} E[s]$$

when $m = 10$ for the following distributions

1) $N(0,1)$

2) $\text{Cauchy}(0,1)$

3) t_3

4) $U(-1,1)$

For the $N(0,1)$, it is known that the true coverage rate is 95% and the mean interval width is

$$2t_{0.025}\frac{\Gamma\left(m/2\right)}{\Gamma\left(m-1/2\right)}\frac{\sqrt{2}}{\sqrt{m(m-1)}}$$

For m = 10, the mean width is 1.391597.

Also for the $\text{Cauchy}(0,1)$, the mean interval width is ∞.

In the other cases, determining the exact values is difficult since the distributions of $\bar{x}$ and s are not tractable.

In each case, m = 1000 imputations will be generated

Estimates:

1) Coverage probability

$$\hat{C}=\frac{1}{n}\sum_{i=1}^{n}I\left(\frac{|\bar{x}_i|\sqrt{m}}{s_i}<t_{0.025}\right)$$

2) Mean interval width

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$$
$$= \frac{2t_{0.025}}{\sqrt{m}}\frac{1}{n}\sum_{i=1}^{n} s_i$$
$$= \frac{2t_{0.025}}{\sqrt{m}}\bar{s}$$

Coverage Rate:

Distribution	$\hat{C}$	$SE_{\hat{C}}$
$N(0,1)$	0.954	0.0066245
Cauchy$(0,1)$	0.981	0.0043173
t_3	0.958	0.0063432
$U(-1,1)$	0.967	0.0056490

where

$$SE_{\hat{C}} = \sqrt{\frac{\hat{C}\left(1-\hat{C}\right)}{n}}$$

Mean Interval Width:

Distribution	$\bar{w}$	$SE_{\bar{w}}$
$N(0,1)$	1.372	0.010122
$\text{Cauchy}(0,1)$	34.339	9.631
t_3	2.149	0.032455
$U(-1,1)$	0.824	0.004204

The estimation errors for the $N(0,1)$ cases are

Coverage: 0.004

Mean width: 0.0193

both of which are within 2 SE's.

Cauchy example:

The assumption that $E\left[\left|f(X)\right|\right] < \infty$ is important. Since the Cauchy has no finite moments, $E[s] = \infty$, and thus the mean interval width is ∞. Thus the reported sample average and standard error is not meaningful.

However even though the mean width is not defined, the coverage rate is well defined. For every dataset, the indicator function is well defined and the integral of any indicator function of probable interest is finite.

Sample size:

When designing a Monte Carlo study, the sample size *m* needs to be determined.

Usual approach is by bounding the SE.

Want

$$SE \le \frac{\sigma_f}{\sqrt{n}}$$

which gives

$$n \ge \frac{\sigma_f^2}{SE^2}$$

where SE is the desired standard error and $\sigma_f^2 = \mathrm{Var}\left(f\left(X\right)\right)$.

There is the same problem here as with determining the sample size necessary to bound the size of a confidence interval: What is σ_f^2?

Sometimes you can guess on what σ_f^2 might be.

For example, in the coverage rate case

$$\sigma_f^2 = C(1-C)$$

Since for the examples, C will be around 0.95, use that to pick n.

It can be tougher for other problems. For the width example, the question comes down to determining $\text{Var}(s_i)$. While this could be done for the normal (and the Cauchy), it is tougher for the other distributions.

One approach is to do a small test sample to get a guess of σ_f^2 and use this to figure out how many more samples need to be added.

Single sample – Multiple questions

In the example, a single sample was used to answer both questions (i.e. $\bar{x}_i$ and s_i are the same in averages). I could have used these same samples to answer many more questions (e.g. $\text{Var}(w)$, $E\left[|\bar{x}|\right]$, $E\left[s^2 + 4.2\sqrt{|m\bar{x}|}\right]$, etc)

When dealing with multiple quantities to be studied, you need to pick a sample size that meets the requirements for all quantities (at least the important ones).

Implementation in S-Plus/R & Matlab

When possible use vectorized calculations, not loops, particularly with S-Plus.

Vectorized	Loop

rnorm.vec <- function(n, mu=0, sigma=1) {
 	
ndat<-matrix(rnorm(10*n, mu, sigma), ncol=10)
xbar <- apply(ndat, 1, mean)
s <- sqrt(apply(ndat, 1, var))
 	
cover <- abs(xbar) * sqrt(10) / s <= qt(0.975, 9)
width <- 2 * qt(0.975, 9) * s / sqrt(10)
 	
C <- mean(cover)
wbar <- mean(width)
 	
list(cover=cover, C=C, width=width, wbar=wbar) }
	rnorm.loop <- function(n, mu=0, sigma=1) {
	
	xbar<-rep(0,n)
	s <- rep(0,n)
	cover <- rep(0,n)
	width <- rep(0,n)
	
	for(i in 1:n) {
	
	  x <- rnorm(10, mu, sigma)
	  xbar[i] <- mean(x)
	  s[i] <- sqrt(var(x))
	
	  cover[i] <- abs(xbar[i]) * sqrt(10) / s[i] <= qt(0.975, 9)
	  width <- 2 * qt(0.975, 9) * s[i] / sqrt(10)
	}
	
	C <- mean(cover)
	wbar <- mean(width)
	
	list(cover=cover, C=C, width=width, wbar=wbar) }

Run times when n = 10,000

	R	S-Plus
Vectorized	1.5 sec	35 sec
Loop	2.5 sec	48 sec
Loop/Vector	1.67	1.37

Tests done on 1.6GHz Pentium 4 running Windows XP

R version: 1.8.1

S-Plus version: 6.0 Release 2

General consensus about S-Plus would have suggested that the loop/vector ratio should have been higher with S-Plus than with R.

While I'm not sure how to quantify it with this setup, the memory use for loops is usually worse than for vectorized setups, particularly with S-Plus.

In Matlab, looping isn't as bad, though if a procedure can be done with vectorized calculations its, usually preferable.

Getting more precise estimates

1) increase n

2) different sampling scheme

Stratified Sampling

Break the sample space S into disjoint regions $S_1, \dots, S_K$

Sample points $x_{k1}, \dots, x_{kn_k}$ in region k

Within each region get sample average

$$\bar{f}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} f(x_{ki})$$

Then estimate μ_f by

$$\hat{\mu}_f = \sum_{k=1}^{K} P[S_k] \bar{f}_k$$

This estimator is based on the idea

$$E\left[f(X)\right] = E\left[E\left[f(X) \middle| S_k\right]\right]$$

The variance of this estimator is

$$\mathrm{Var}\left(\hat{\mu}_f\right) = \sum_{k=1}^{K} \left(P[S_k]\right)^2 \frac{\mathrm{Var}\left(f(X) \middle| X \in S_k\right)}{n_k}$$

If the regions are picked reasonably, this will have a smaller variance than

$$\frac{\mathrm{Var}\left(f\left(X\right)\right)}{n}$$

If $n_k = nP\left[S_k\right]$ (proportional sampling), the variance of the stratified estimator reduces to

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_f\right) &= \frac{1}{n}\sum_{k=1}^{K} P\left[S_k\right]\mathrm{Var}\left(f\left(X\right)\middle|X \in S_k\right) \\ &= \frac{1}{n}E\left[\mathrm{Var}\left(f\left(X\right)\middle|Z\right)\right]\end{aligned}$$

where Z is a random variable satisfying $Z = k$ when the single random point drawn falls in S_k.

Since

$$\begin{aligned}\mathrm{Var}\left(f\left(X\right)\right) = E\left[\mathrm{Var}\left(f\left(X\right)\middle|Z\right)\right] \\ + \mathrm{Var}\left(E\left(f\left(X\right)\middle|Z\right)\right)\end{aligned}$$

The stratified estimator has a smaller variance that the sample average estimator.

Nonproportional sampling can give even more efficiency

The optimal sample size choices, subject to $\sum n_k = n$ is

$$n_k = n \frac{P[S_k]\sqrt{\mathrm{Var}\left(f(X)\middle|X \in S_k\right)}}{\sum_{j=1}^{K} P\left[S_j\right]\sqrt{\mathrm{Var}\left(f(X)\middle|X \in S_j\right)}}$$

This implies that regions with high variability should get more samples than regions of small variability.

Antithetic Variates

Combines 2 correlated estimators to achieve a more precise estimator.

Based on the idea

$$\mathrm{Var}\left(\frac{V+W}{2}\right) = \frac{1}{4}\mathrm{Var}(V) + \frac{1}{4}\mathrm{Var}(W) + \frac{1}{2}\mathrm{Cov}(V,W)$$

If V and W are negatively correlated, then you get a more precise estimate than if they were uncorrelated (or positively correlated).

So we need to generate coupled, negatively correlated random variables.

The following proposition gives us an approach for doing this

Proposition 21.4.1. Suppose X is a random variable and the functions $f(x)$ and $g(x)$ are both increasing or both decreasing. If the random variables $f(X)$ and $g(X)$ have finite second moments, then

$$\operatorname{Cov}\left(f(X), g(X)\right) \geq 0$$

If $f(x)$ is increasing and $g(x)$ is decreasing (or vice-versa), then the covariance ≤ 0.

Proof (See Lange, page 291)

Suppose we wish to calculate

$$\int f(x) g(x) dx$$

where $f(x)$ is an increasing function and the density $g(x)$ has CDF $G(x)$. Then the function $f\left(G^{-1}(u)\right)$ is increasing and the function $f\left(G^{-1}(1-u)\right)$ is decreasing when $u \in [0,1]$.

If $U_1, \ldots, U_n$ is and iid sample from $U(0,1)$, then

$$E\left[f\left(G^{-1}(U)\right)\right] = E\left[f\left(G^{-1}(1-U)\right)\right]$$
$$= \int f(x)\, g(x)\, dx$$

and $f\left(G^{-1}(U)\right)$ and $f\left(G^{-1}(1-U)\right)$ are negatively correlated.

Thus the estimator

$$\frac{1}{2n}\sum_{i=1}^{n}\left\{f\left(G^{-1}(U_i)\right) + f\left(G^{-1}(1-U_i)\right)\right\}$$

has a smaller variance than

$$\frac{1}{2n}\sum_{i=1}^{2n} f\left(G^{-1}(U_i)\right)$$

The idea behind this estimator is that if $U_1,\ldots,U_n$ are uniform, so are $1-U_1,\ldots,1-U_n$. Then this implies that $G^{-1}(U_1),\ldots,G^{-1}(U_n)$ and $G^{-1}(1-U_1),\ldots,G^{-1}(1-U_n)$ are both sets of draws from $X \sim G$.

What this estimator is doing is making sure that if the p^{th} quantile is in the sample of X, so is the $1 - p^{th}$ quantile.

In a sense, its giving a more balance sample.

Note that this also works if $f(x)$ is a decreasing function.

Example:

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

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

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 $G\left(x\right)$ and $G^{-1}\left(x\right)$ are nice functions.

In fact you only need $G^{-1}\left(x\right)$ 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 $\mathrm{Var}(f(X))$.

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 $\operatorname{Var}\left(f_b\left(X\right)\right)$, which can be done with

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

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

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

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

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

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.

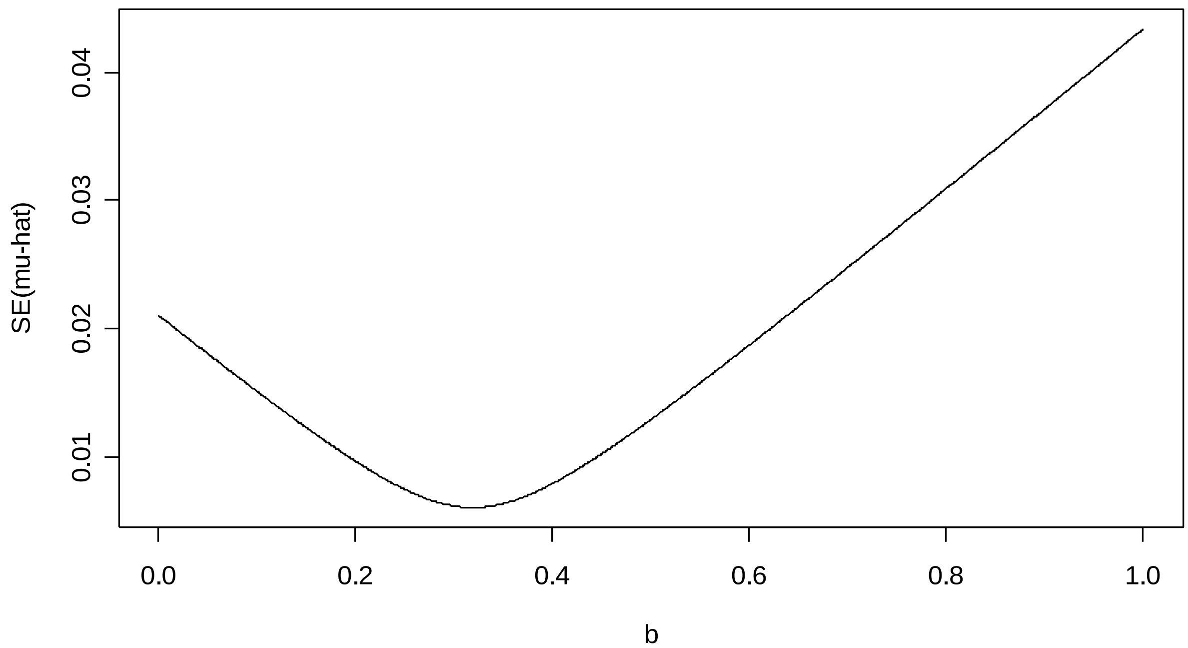
$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i} = 1.256252$$

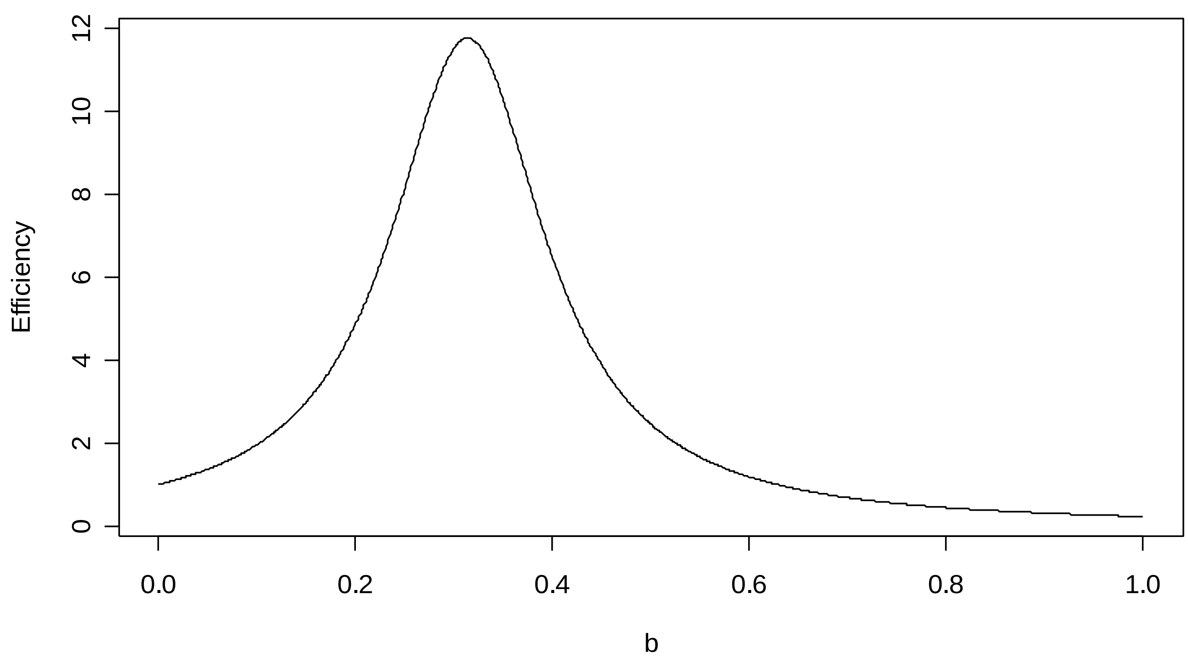
$$\frac{1}{n}\sum_{i=1}^{n}x_i = 2.01843$$

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 $\mathrm{Cov}\left(f(X), g(X)\right)$ is usually intractable analytically. However you can use your sample to estimate it (and $\mathrm{Var}\left(g(X)\right)$ if necessary).

Another equivalent approach (assuming that you estimate $\mathrm{Var}\left(g(X)\right)$) 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 $\{f(x_i)\}$ and $\{g(x_i)\}$).

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.