Monte Carlo Integration

* Many problems in statistics involve integration.

* This is particularly true in *Bayesian Statistics* where the basic methodology gives a distribution.

* It is often of interest to find moments of this distribution and (for continuous distributions) this involves integration.

* The Monte Carlo method is a technique to use simulation to approximate the value of an integral.

Monte Carlo Integration

* Suppose we wish to evaluate

$$I = \int_a^b h(x)\,dx$$

where a and b are constants but may be infinite.

* Now suppose that we can write

$$h(x) = g(x)f(x)$$

where $f(x)$ is a probability density function with support (a, b).

* We can then rewrite the integral as

$$I = \int_a^b g(x)f(x)\,dx = \mathsf{E}\left[g(X)\right]$$

where X is a random variable having pdf f.

Monte Carlo Integration

* Now suppose that we can generate values having pdf f by one of the methods described earlier.

* We can then generate N independent random variates $X_1, \ldots, X_N$ each having pdf f.

* Thus we can evaluate the Monte Carlo estimator

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

Monte Carlo Integration

Theorem 7

Suppose that $X_1, \ldots, X_N$ *are* iid *with pdf* $f(x)$ *on support* $\mathcal{X}$ *and* $\hat{I}_N = N^{-1} \sum g(X_i)$ *is a Monte Carlo estimator of* $I = \int_{\mathcal{X}} g(x) f(x)\, dx$*. Then*

$$\begin{aligned} \mathsf{E}(\hat{I}_N) &= I \\ \mathsf{Var}(\hat{I}_N) &= \frac{1}{N} \left[\int_{\mathcal{X}} g^2(x) f(x)\, dx - I^2 \right] \end{aligned}$$

* A similar result holds if the X_i are discrete with pmf f except that integrals are replaced by sums.

Monte Carlo Integration

* Since Monte Carlo produces an estimate of the required value, it is important to give a measure of its variability.

* We see from the previous theorem that the variance of the estimator reduces as $N \to \infty$ so increasing the simulation size will improve the accuracy of the estimator.

* Unfortunately the variance of the estimator is generally a function of the unknown I.

* Instead of $\text{Var}(\widehat{I}_N)$ we therefore return an estimate of the square root of this variance. This is called the standard error

$$\text{se}(\widehat{I}_N) = \sqrt{\frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N} g^2(X_i) - \widehat{I}_N^2\right]}.$$

Variance Reduction Techniques

* There can often be different Monte Carlo methods for the same problem.

* Different methods can differ in efficiency in a number of respects.
  - The amount of analytical work required by the user prior to using Monte Carlo.
  - The programming complexity of the algorithm.
  - The computational complexity of the algorithm.
  - The variability of the Monte Carlo estimate.

* Controlling the last of these is the most common way to improve efficiency of an Monte Carlo experiment.

Control Variables

* Suppose that we have a Monte Carlo estimate $\widehat{I}$ based on a sample of size N with $\mathsf{E}(\widehat{I}) = I$.

* Now let us suppose that we can find another quantity C which is correlated with $\widehat{I}$ and whose mean we know to be μ.

* The simplest way to guarantee that C and $\widehat{I}$ are correlated is to have them based on the same sequence of random numbers.

* C is called a control variable.

Monte Carlo Estimator Based on a Control Variable

* Now consider the new estimator

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

for some known value of β.

* The mean and variance of this new Monte Carlo estimator are

$$\begin{aligned} \mathrm{E}(\hat{I}_C) &= I \\ \mathrm{Var}(\hat{I}_C) &= \mathrm{Var}(\hat{I}) + \beta^2 \mathrm{Var}(C) - 2\beta \mathrm{Cov}(\hat{I}, C) \end{aligned}$$

* For an appropriate choice of C and β we can have

$$\mathrm{Var}(\hat{I}_C) < \mathrm{Var}(\hat{I})$$

Variance of Control Variable Monte Carlo Estimator

Theorem 8

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

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

and that minimum variance is

$$\mathrm{Var}(\hat{I}_C) = (1 - \rho^2)\mathrm{Var}(\hat{I})$$

where ρ is the correlation coefficient between $\hat{I}$ and C.

Choice of Control Variables

* We thus need to find a variable C which is strongly correlated with $\hat{I}$ and then choose β according to the formula in the theorem.

* Although variables correlated with $\hat{I}$ are not hard to find, it is often the case that we cannot evaluate the covariance between them and so we cannot find the best β.

* One strategy which is sometimes employed is to use the Monte Carlo results themselves to estimate the correct β and so achieve variance reduction in that way.

* If possible, it is preferable to do as much analytic calculation as possible prior to the Monte Carlo run and so reduce the error and variability in finding β.

Antithetic Variables

* Another variance reduction scheme relies on making two Monte Carlo estimates of the same quantity.

* Suppose that $\widehat{I}_1$ and $\widehat{I}_2$ are both unbiased Monte Carlo estimators of I. Then so is $\widehat{I}_A = \frac{1}{2}\left(\widehat{I}_1 + \widehat{I}_2\right)$

* The variance of $\widehat{I}_A$ is

$$\begin{aligned} \mathrm{Var}(\widehat{I}_A) &= \frac{\mathrm{Var}(\widehat{I}_1)}{4} + \frac{\mathrm{Var}(\widehat{I}_2)}{4} + \frac{\mathrm{Cov}(\widehat{I}_1, \widehat{I}_2)}{2} \\ &\approx \frac{\mathrm{Var}(\widehat{I}_1)}{2} + \frac{\mathrm{Cov}(\widehat{I}_1, \widehat{I}_2)}{2} \end{aligned}$$

if $\mathrm{Var}(\widehat{I}_1) = \mathrm{Var}(\widehat{I}_2)$.

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

Antithetic Variables

* Typically both original estimators will be means of some functions of a sequence of random variables.

* If we can make the random variables used in each estimator as negatively correlated as possible, then we can hope that the resulting estimators will share similar properties.

* Since all random numbers are derived from Uniform random variates this is not hard to do.

* It is clear that if U is a uniform random variable then so is $1-U$ and the correlation coefficient between these two is -1 so they are maximally negatively correlated.

* The concept of antithetic variables uses this to try to improve on the variability of Monte Carlo estimators.

Antithetic Variables

* Consider estimation of $I = \int_0^1 g(x)\, dx$.

* Let $U_i, \ldots, U_N$ be iid uniform$(0,1)$ random variates and define

$$\widehat{I}_1 = \frac{1}{N}\sum_{i=1}^{N} g(U_i) \qquad \widehat{I}_2 = \frac{1}{N}\sum_{i=1}^{N} g(1-U_i)$$

  for the same random variables $U_1, \ldots, U_N$.

* Both $\widehat{I}_1$ and $\widehat{I}_2$ are unbiased estimators of I with the same variance but are negatively correlated for monotone g.

* Hence the antithetic variable estimator

$$\widehat{I}_A = \frac{1}{2N}\sum_{i=1}^{N}\Big[g(U_i) - g(1-U_i)\Big]$$

  may have reduced variability.

Theorem 9

Suppose that we are interested in estimation of the integral $I = \int_0^1 g(x)\,dx$ where g is a continuous, monotonic function with continuous first derivatives. Let $U_1, \ldots, U_N$ be iid $U(0,1)$ *random variables and define*

$$\widehat{I} = \frac{1}{N} \sum g(U_i) \quad \textit{and} \quad \widehat{I}_A = \frac{1}{2N} \sum \Big[g(U_i) - g(1 - U_i) \Big]$$

to be two Monte Carlo estimates of I. Then

$$\mathrm{Var}(\widehat{I}_A) \leqslant \frac{\mathrm{Var}(\widehat{I})}{2}$$

Antithetic Variables

* In general we are not trying to integrate over the interval $(0, 1)$.

* In that case we have estimators

$$\widehat{I}_1 = N^{-1} \sum_{i=1}^{N} g(X_i) \qquad \widehat{I}_2 = N^{-1} \sum_{i=1}^{N} g(Y_i)$$

where X_i and Y_i have the same distribution.

* To make an efficiency gain we need $g(X_i)$ and $g(Y_i)$ to be negatively correlated.

* In general this is not easy to achieve.

Antithetic Variables

* One case in which we can try to make this happen is when we have

$$X_i \;=\; F^{-1}(U_i) \quad \text{for } U_i \sim \text{uniform}(0,1).$$

* In that case we can define

$$Y_i \;=\; F^{-1}(1-U_i)$$

for the same set of uniform random variables.

* Since F is monotone, so is F^{-1} so $g(X_1),\ldots,g(X_N)$ and $g(Y_1),\ldots,g(Y_N)$ will be negatively correlated if the function g is monotone.

Importance Sampling

* Consider estimation of the quantity

$$I = \mathsf{E}_f\Big[h(X)\Big] = \int h(x) f(x)\, dx$$

* The usual Monte Carlo method for this would be to simulate $X_1 \ldots, X_N$ with density $f(x)$ and use

$$\hat{I} = N^{-1} \sum h(X_i)$$

* In this process, the distribution from which we are simulating does not relate in any way to the function to be integrated.

* It is very possible for us to simulate many of the variables from areas where $h(x)$ is very small or even 0.

Importance Sampling

* The idea of importance sampling is that it would be better to simulate more variates in those areas which are *more important* for estimation of I.

* We would therefore simulate from a different density $g(x)$ which concentrates more mass in important areas.

* Note that we can write

$$\begin{aligned} I &= \int h(x) f(x)\, dx \\ &= \int h(x) \frac{f(x)}{g(x)} g(x)\, dx \\ &= \int h(x) w(x) g(x)\, dx \\ &= \mathsf{E}_g[h(X) w(X)] \end{aligned}$$

Importance Sampling

* If $X_1 \ldots, X_N$ are simulated to be *iid* with density g then a Monte Carlo estimator of I is

$$\hat{I}_{IS} = \frac{1}{N} \sum_{i=1}^{N} h(X_i) w(X_i)$$

* The quantity $w(X_i) = f(X_i)/g(X_i)$ is known as the importance sampling weight of the point X_i.

* The importance sampling weight adjusts for the fact that $X_1, \ldots, X_N$ were simulated from g rather than f.

Choosing the Importance Sampling Density

* We wish to choose the importance sampling density g in such a way that
  1. It is easy to simulate random variables having density g.
  2. $\text{Var}(\widehat{I}_{IS}) < \text{Var}(\widehat{I})$.

* Ideally we would like to make the variance of the resulting estimator as small as possible.

Optimal Importance Sampling Density

Theorem 10

The minimum variance of $\widehat{I}_{IS}$ is

$$\mathsf{Var}(\widehat{I}_{IS}) \geqslant \frac{1}{N}\left\{\left[\int |h(x)|f(x)\,dx\right]^2 - I^2\right\}$$

and this is achieved when we take

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

Optimal Importance Sampling Density

* Unfortunately the optimal importance sampling distribution is rarely available.

* However the previous theorem says we should look for density functions g which have shape close to $|h(x)|f(x)$ and the same support as f.

* Often we will decide on a particular family of distributions from which we can sample easily and then find parameter values θ such that $|h(x)|f(x)/g(x \mid \theta)$ is close to constant.

Standard Error of Importance Sampling Estimate

* As with all Monte Carlo estimates, we require a measure of the variability of the estimate. We will use the same random variates to estimate the variability.

* Recall that

$$\text{Var}(\hat{I}_{IS}) = \frac{1}{N}\left\{\int \frac{h^2(x)f^2(x)}{g(x)}\,dx - I^2\right\}$$

* We can use $X_1, \ldots, X_N$ to estimate this quantity and get the standard error

$$\text{se}(\hat{I}_{IS}) = \frac{1}{N}\sqrt{\sum \Big[h(X_i)w(X_i)\Big]^2 - N\hat{I}_{IS}^2}$$