http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo20/

Class notes: Monte Carlo methods
Week 12, SDE, control variates
tentative Jonathan Goodman
Tentative December 2, 2020

1 Control variates

Control variates is (are?) a variance reduction method that applies when you have an approximate solution to your problem. Suppose X and Y are scalar random variables that represent the "quantity of interest " or *QOI*, and an approximation to X that is better understood. The important things are that X and Y are correlated, the stronger the correlation the better, and that the expected value of Y is known.

We seek

$$A = \mathrm{E}[X] \ .$$

There is a way to generate pairs (X, Y), either by direct simulation or MCMC. We generate n sample (X_k, Y_k) pairs.

$$B = \mathrm{E}[Y] \ \text{ is known.}$$

The *control variate* estimator is

$$\widehat{A}_\beta = \left[\frac{1}{n} \sum_{k=1}^{n} (X_k - \beta Y_k) \right] - \beta B \tag{1}$$

The choice $\beta = 0$ is the direct estimate without the control variate. You can check that $\widehat{A}_\beta$ is unbiased in the sense that, for any β,

$$\mathrm{E}\left[\widehat{A}_\beta \right] = A \ .$$

We use the control variate Y and the multiplier β to reduce the variance.

Suppose at first that the samples are independent and we know the covariance

$$\sigma_{XX} = \mathrm{var}(X) \ , \ \ \sigma_{YY} = \mathrm{var}(Y) \ , \ \ \sigma_{XY} = \mathrm{cov}(X, Y) \ , \ \ \rho_{XY} = \mathrm{corr}(X, Y) = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\, \sigma_{YY}}} \ .$$

Then we can optimize the estimator variance by miniizing

$$\begin{aligned} \mathrm{var}\left(\widehat{A}_\beta \right) &= \frac{1}{n} \mathrm{var}(\, X - \beta Y) \\ &= \frac{1}{n} \left[\sigma_{XX} - 2\beta\sigma_{XY} + \beta^2 \sigma_{YY} \right] \ . \end{aligned}$$

The answer is

$$\beta_* = \frac{\sigma_{XY}}{\sigma_{YY}} \ . \tag{2}$$

The resulting variance may be written as

$$\begin{aligned} \mathrm{var}(\, X - \beta_* Y) &= \sigma_{XX} \left(1 - \frac{\sigma^2_{XY}}{\sigma_{XX}\, \sigma_{YY}} \right) \\ \mathrm{var}(\, X - \beta_* Y) &= \mathrm{var}(X) \left(1 - \mathrm{corr}(X,Y)^2 \right) \end{aligned}$$

The control variate estimator variance is smaller than the direct estimator variance by a factor that depends on the correlation between X and Y. Note that perfect correlation $\rho_{XY} = \pm 1$, leads to a zero variance estimator. People often seen good variance reduction methods by starting with impossible zero variance estimators and seeing how close they can come with a practical algorithm.

In practice, it is unlikely that the optimal multiplier (2) would be known. If we don't know the mean of X, it is unlikely that we know the covariance between X and Y. Instead, we can estimate them from the same MC samples

$$\widehat{\beta}_* = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}^2_{XY}} \; . \tag{3}$$

We can use the standard estimators

$$\begin{aligned} \widehat{\sigma}_{XY} &= \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \overline{X} \right) \left(Y_k - \overline{Y} \right) \\ \widehat{\sigma}_{YY} &= \frac{1}{n-1} \sum_{k=1}^{n} \left(Y_k - \overline{Y} \right)^2 \\ \overline{X} &= \frac{1}{n-1} \sum_{k=1}^{n} X_k \\ \overline{Y} &= \frac{1}{n-1} \sum_{k=1}^{n} Y_k \end{aligned}$$

The resulting estimator is $\widehat{A}_{\widehat{\beta}_*}$. This has variance slightly larger than if the optimal multiplier were known $\widehat{A}_{\beta_*}$. It also is slightly biased.

I want to talk about errors in estimating β_* and how they effect the ultimate estimate of A. We have done estimates like this before. The new thing here is that we use inexact estimated quantities in formulas. Estimation errors might be amplified as they propagate through a string of formulas. From a technical point of view, what I'm doing amounts to repeated applications of the central limit theorem and what is called *Slutsky's theorem*. To start, the central limit theorem for $\overline{X}$ may be expressed as

$$\overline{X} \approx A + \frac{C_1}{\sqrt{n}} Z_X \; , \quad Z_X \sim \mathcal{N}(0,1) \; . \tag{4}$$

The central limit theorem says that the error is Gaussian and a calculation says the variance is proportional to $\frac{1}{n}$. The right side of (4) includes a random variable with that distribution. The expression (4) is useful because you can plug it into other equations and quickly see which error terms are the largest. For example,

$$\begin{aligned}
\widehat{\sigma}_{XY} &= \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - A + \frac{C_1}{\sqrt{n}}Z_X\right)\left(Y_k - B + \frac{C_2}{\sqrt{n}}Z_Y\right) \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left[(X_k - A)(Y_k - B) + \frac{C_1}{\sqrt{n}}Z_X(Y_k - B) + (X_k - A)\frac{C_2}{\sqrt{n}}Z_Y + \frac{C_1C_2}{n}Z_XZ_Y\right] \\
&= \left[\frac{1}{n-1}\sum_{k=1}^{n}(X_k - A)(Y_k - B)\right] + \frac{3C_1C_2}{n}Z_XZ_Y \ .
\end{aligned}$$

This calculation uses the formula $\frac{1}{n-1} = \frac{1}{n}$. It shows that the error in estimating the covariance σ_{XY} is given by the central limit theorem as though $\overline{X} = A$ and $\overline{Y} = B$. That error is of order $\frac{1}{\sqrt{n}}$, while the error from misestimating A and B is order $\frac{1}{n}$. This calculation shows that the bias, at least to order $\frac{1}{n}$ depends on the covariance of Z_X with Z_Y, which is proportional to $\text{cov}(X, Y)$. The bias is order $\frac{1}{n}$.

To continue, write this result as

$$\widehat{\sigma}_{XY} = \sigma_{XY} + \frac{C_3}{\sqrt{n}}Z_{XY} + \frac{3C_1C_2}{n}Z_XZ_Y \ .$$

A similar calculation leads to

$$\widehat{\sigma}_{YY} = \sigma_{YY} + \frac{C_4}{\sqrt{n}}Z_{YY} + \frac{C_2^2}{n}Z_Y^2 \ .$$

The conclusions about $\widehat{\sigma}_{YY}$ are as above. These expressions lead to

$$\begin{aligned}
\widehat{\beta}_* &= \frac{\sigma_{XY} + \frac{C_3}{\sqrt{n}}Z_{XY} + \frac{3C_1C_2}{n}Z_XZ_Y}{\sigma_{YY} + \frac{C_4}{\sqrt{n}}Z_{YY} + \frac{C_2^2}{n}Z_Y^2} \\
&= \frac{\sigma_{XY}}{\sigma_{YY}} + \frac{1}{\sigma_{YY}}\left[\frac{C_3}{\sqrt{n}}Z_{XY} + \frac{3C_1C_2}{n}Z_XZ_Y\right] \\
&\qquad - \frac{\sigma_{XY}}{\sigma_{YY}^2}\left[\frac{C_4}{\sqrt{n}}Z_{YY} + \frac{3C_2^2}{n}Z_Y^2\right] \\
&\qquad + \frac{\sigma_{XY}^2}{\sigma_{YY}^3}\frac{C_4^2}{n}Z_{YY}^2 + O(n^{-\frac{3}{2}}) \ .
\end{aligned}$$

This calculation is long, but not hard. It shows that

$$\widehat{\beta}_* = \beta_* + \frac{1}{\sqrt{n}}(\text{ mean zero Gaussian }) + \frac{1}{n}(\text{ random variable, mean not zero }) \ .$$

This understanding of $\widehat{\beta}_*$ can be inserted into the control variate estimator and we can calculate the results as above. The result is that estimating β_*

increases the variance of the estimator a little and adds bias of order $\frac{1}{n}$. You could calculate the leading term in the bias.

1.1 Example, sampling with an approximate density

Control variates make a significant difference in many problems. It is a good idea as a Monte Carlo person always to look for cheap or useful approximations that can be made into control variates. This might involve an approximate analysis, a linearization, or just some intuitive approximation. The method comes with error estimates, so you can tell if your control variates are helping.

Here is an example control variate that is often used in practical Bayesian estimation problems. Suppose you have a posterior distribution (or a distribution you got another way) of the form

$$\rho(x) = \frac{1}{Z} e^{-\beta\phi(x)}$$

You want to estimate the posterior expectation of something

$$A = \mathrm{E}_\rho[V(X)] \ .$$

You could do this using MCMC samples of ρ. But there may be an alternative estimator in some cases that avoids MCMC. Depending on the problem, the extra variance of the alternative estimator may be acceptable because direct sampling is better (cheaper, independent samples) than MCMC. This estimator uses a Gaussian approximation to ρ.

Suppose the mode (maximum a-posteriori point, or MAP) is

$$x_* = \arg\min_x \phi(x) \ .$$

If β is large, or if the data strongly confines X to be near x_*, you can use a quadratic approximation of ϕ about x_*. The first derivative terms vanish because x_* is a minimizer. With quadratic terms, this is

$$\phi(x) \approx \phi(x_*) + \frac{1}{2}(x - x_*)^t H (x - x_*) \ .$$

Here, H is the Hessian matrix of second derivatives of ϕ evaluated at x_*. This gives a Gaussian approximation to the distribution

$$\rho(x) \approx \sigma(x) = \mathcal{N}(x_*, \beta^{-1}H^{-1}) \ .$$

The formula for σ is

$$\sigma(x) = \frac{\beta^{\frac{d}{2}}\sqrt{\det(H)}}{(2\pi)^{\frac{d}{2}}}\, e^{-\frac{\beta}{2}(x-x^*)^t H (x-x^*)} \ .$$

The importance sampling estimator using σ uses

$$A = \mathrm{E}_\sigma[\,V(X)L(X)] \ , \ \ L(x) = \frac{\rho(x)}{\sigma(x)} \ .$$

The estimator is

$$\widehat{A}_\sigma = \frac{1}{n}\sum_{k=1}^{n} V(X_k)L(X_k) \quad X_k \sim \mathcal{N}(x_*, \beta^{-1}H^{-1}) \; i.i.d. \; .$$

The variance is probably higher than the estimator using ρ. When we used importance sampling to reduce variance, we designed σ taking into account $V(x)$. We didn't do that here. However, there is a direct sampler for the Gaussian.

It might be that the Gaussian expectation of V is explicitly known. This would happen if $V(x)$ were a polynomial or exponential function. In that case we would know

$$B = \mathrm{E}_\sigma[\, V(X)]$$

We use the framework of control variates with random variables

$$\begin{aligned} U &= V(X)L(X) \; , \quad X \sim \mathcal{N}(x_*, \beta^{-1}H^{-1}) \\ W &= V(X) \; , \qquad\quad\; X \sim \mathcal{N}(x_*, \beta^{-1}H^{-1}) \; . \end{aligned}$$

We know $\mathrm{E}[W]$ and want $\mathrm{E}[U]$. If the Gaussian approximation is accurate, there should be a good correlation between U and W.

2 Diffusion processes

A *diffusion process* is a probability distribution for paths $X_t \in \mathbb{R}^n$. The components are $X_{j,t}$, for $j = 1, \cdots, n$. The dynamical model is a *stochastic differential equation*

$$dX_t = a(X_t)dt + b(X_t)dW_t \; . \tag{5}$$

The first term on the right is a deterministic *drift* term. Without the second term on the right, the equation would be written in a more familiar form

$$\frac{d}{dt}X_t = a(X_t) \; .$$

The second term on the right is the *noise term.* The process W_t is an $n-$dimensional Brownian motion. What is important here about Brownian motion is that *increments* $W_{t+s} - W_t$ are Gaussian with mean zero and covariance sI. Stochastic models of engineering or physical systems often take the form of stochastic differential equations.

The *Euler Maruyama* method for approximating the SDE involves a time step Δt and discrete times $t_k = k\Delta t$. The approximation to X_{t_k} will be $X_k^{\Delta t}$. The method is

$$X_{k+1}^{\Delta t} = X_k^{\Delta t} + \Delta t\, a(X_k^{\Delta t}) + \sqrt{\Delta t}\, b(X_k^{\Delta t})\, Z_k \; , \;\; Z_k \sim \mathcal{N}(0, I) \; .$$

A *path* is a sequence $X^{\Delta t} = (X_1^{\Delta t}, \cdots, X_N^{\Delta t})$. The number of components of $X^{\Delta t}$ is $d = nN$. We can find the probability density for the path $X^{\Delta t}$ using the fact that the conditional density of $X_{k+1}^{\Delta t}$ is normal. The probability density for a path is the product of these conditional densities

$$\rho(x) = \frac{1}{Z} \prod_{k=0}^{N-1} e^{(x_{k+1}-x_k-a(x_k)\Delta t)^t C(x_k)^{-1} (x_{k+1}-x_k-a(x_k)\Delta t)^t} \ , \ \ C(x) = \operatorname{cov}(\Delta x) = b(x)b^t(x) \ .$$

It's a clumsy formula but not hard to code or understand.

Suppose you know the starting point x_0 and you want to make a path up to time T. You can generate an approximate sample path by choosing Δt with an integer $N = T/\Delta t$ and using the Euler Maruyama formula N times for the N time steps. A harder problem is to sample paths that have been observed at more than one time. For example, suppose x_0 and $X_T = x_T$ are known and we want to generate sample paths from this conditional density. You could start by trying the single time Metropolis method. The dimension is $d = nN$ which is likely to be large, so the auto-correlation time may be a problem. Jonathan Weare (now "Professor Weare") wrote his PhD thesis on this sampling problem.

Now suppose you want

$$A = \mathrm{E}[\, V(X_T)] \ .$$

You could make M independent samples by direct Euler time stepping. This is a lot of work if Δt is small and N is large. There is a *multi-scale* control variate strategy due to Michael Giles for this. In the simplest two-scale version it involves *coupling* simulations with Δt and $2\Delta t$. The "coarse scale" simulations with $2\Delta t$ may be approximately like the "fine scale" simulations with Δt, but suppose the coarse simulations are not accurate enough. We use the notation

$$A_{\Delta t} = \mathrm{E}\big[\, V(X_T^{\Delta t})\big] \ .$$

The work to estimate $A_{2\Delta t}$ is half the work to estimate $A_{\Delta t}$. The strategy of Giles is to estimate the two quantities

$$A_{2\Delta t} \ , \ \ A_{\Delta t} - A_{2\Delta t} \ .$$

You add these to estimate $A_{\Delta t}$. The second quantity is small, so it may be estimated to good accuracy with fewer sample paths.