How Much Can a Control Variate Reduce Variance?

Due: Wednesday, September 9 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with distribution P, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and identically distributed samples from P. A scientific simulation computes an expensive output $F(\mathbf{X}_i)$ and a surrogate $G(\mathbf{X}_i)$ from each sample, where $F, G : \mathbb{R}^d \to \mathbb{R}$. The quantity of interest is $\mu = \mathbb{E}[F(\mathbf{X})]$, while $\mathbb{E}[G(\mathbf{X})]$ is known exactly from the model. Assume $F(\mathbf{X})$ and $G(\mathbf{X})$ have finite positive variances. For a chosen constant $c \in \mathbb{R}$, the control-variate estimator is

$$\widehat{\mu}_{n,c} = \frac{1}{n} \sum_{i=1}^{n} \left[F(\mathbf{X}_i) - c \left(G(\mathbf{X}_i) - \mathbb{E}[G(\mathbf{X})] \right) \right].$$

Additional simulation runs are costly, so we want to choose the member of this family that extracts as much information as possible from G.

1. The correction changes the variance of each simulation output. What is $\mathrm{Var}(\widehat{\mu}_{n,c})$?
2. We can choose the strength of the correction. Which value of c globally minimizes this variance?
3. Ordinary Monte Carlo uses the same n simulations without the surrogate. What multiple of its variance $\mathrm{Var}(F(\mathbf{X}))/n$ is attained at this optimum?

Hint: The estimator above is unbiased for every c because $c(G(\mathbf{X}_i) - \mathbb{E}[G(\mathbf{X})])$ has mean zero. For independent, identically distributed variables with finite variance, the variance of their average is the one-sample variance divided by n. You may also use $\mathrm{Var}(U + aV) = \mathrm{Var}(U) + 2a\,\mathrm{Cov}(U, V) + a^2\,\mathrm{Var}(V)$.

Takeaway: *A second measurement with a known mean can make an estimate more precise without running more simulations. It helps most when it is strongly correlated with the quantity we want to estimate.*