



STK4051/STK9051

Computational Statistics

Riccardo De Bin

`debin@math.uio.no`

Outline of the lecture

- Variance Reduction Techniques
 - Importance Sampling
 - Antithetic sampling
 - Control variates
 - Rao-Blackwellization

Variance Reduction Techniques: Introduction

Can we do better than the simple Monte Carlo estimator of

$$\int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i)$$

where the variables $\mathbf{X}_1, \dots, \mathbf{X}_n$ are randomly sampled from f ?

Yes! Integral estimates with lower variance:

1. Importance sampling
2. Antithetic sampling
3. Control variates
4. Rao-Blackwellization

Importance Sampling: motivating example

Goal: estimate the probability that a die roll will yield a 1.

Simple Monte Carlo:

- roll the die n times;
- expect to see about $n/6$ ones;
- the point estimate is the proportion of ones in the sample;
- the variance of this estimator is $\frac{1}{n^2} n p (1 - p) = \frac{5}{36n}$;
- to achieve an estimate with a $CV = \sqrt{\text{var}[X]}/E[X]$ of 5%, we should expect to have to roll the die 2,000 times.

Importance Sampling: motivating example

Alternative strategy:

- bias the die by replacing the 2 and 3 faces with 1;
- increases the probability of rolling a one to 0.5;
- no longer sampling from the target distribution (fair die);
- to correct for this, we should weight each roll of a one by $1/3$,
 - ▶ let $Y_i = 1/3$ if the roll is a one and $Y_i = 0$ otherwise;
- the expectation of the sample mean of the Y_i is $1/6$;
- the variance of the sample mean is $1/36n$;
- to achieve a CV of 5%, we need only 400 rolls.

Importance Sampling: introduction

Main idea: force the event of interest to occur more frequently than in naive Monte Carlo to improve accuracy,

- for rare events, large reductions in variance are possible.

Some importance sampling terminology:

- The die rolling example is successful because an importance sampling distribution (corresponding to rolling the die with three ones) is used to over-sample a portion of the state space that receives lower probability under the target distribution (for the outcome of a fair die);
- An importance weighting corrects for this bias and can provide an improved estimator.

Importance Sampling: example

Importance sampling is based upon the principle that $E_X[h(\mathbf{X})]$, where X has density $f(x)$, can be rewritten as

$$\mu = \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}$$

with g being another density function,

- called the **importance sampling function** or **envelope**.

To estimate $E\{h(\mathbf{X})\}$, draw $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. from g and use

$$\hat{\mu}_{\text{IS}}^* = \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i) w^*(\mathbf{X}_i)$$

where $w^*(\mathbf{X}_i) = f(\mathbf{X}_i)/g(\mathbf{X}_i)$ are unstandardized weights,

- also called *importance ratios*.

Important: it must be easy to sample from g and to evaluate f , even when it is not easy to sample from f .

Importance Sampling: example

Equivalently, we can write

$$\mu = \frac{\int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}}{\int f(\mathbf{x}) d\mathbf{x}} = \frac{\int h(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}}$$

This suggests drawing $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. from g and using

$$\hat{\mu}_{\text{IS}} = \sum_{i=1}^n h(\mathbf{X}_i) w(\mathbf{X}_i)$$

where $w(\mathbf{X}_i) = w^*(\mathbf{X}_i) / \sum_{i=1}^n w^*(\mathbf{X}_i)$ are standardized weights,

- or *normalised importance ratios*.

Here f is known only up to a proportionality constant,

- common when f is a posterior density in a Bayesian analyses.

Importance Sampling: Convergence

Both estimators converge by the strong law of large numbers,

- when the support of the envelope includes the support of f .

Choice between standardized and unstandardized weights:

- standardized weights provides a better estimator when $w^*(\mathbf{X})$ and $h(\mathbf{X})w^*(\mathbf{X})$ are strongly correlated;
- standardized weights do not require knowing the proportionality constant for f ;
- unstandardised weights are superior when estimating small probabilities.

Importance Sampling: Convergence

First consider the unstandardised weight estimator,

$$\hat{\mu}_{IS}^* = \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i) w^*(\mathbf{X}_i), \text{ with } w^*(\mathbf{X}_i) = \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}$$

Note that:

- $E[w^*(\mathbf{X}_i)] = \int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) d\mathbf{x} = 1;$
- $E[\hat{\mu}_{IS}^*] = \int h(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mu;$
- $\text{Var}[\hat{\mu}_{IS}^*] = \frac{1}{n} \text{Var}[h(\mathbf{X}) w^*(\mathbf{X})].$
- so $\hat{\mu}_{IS}^*$ is unbiased;
- an estimator of its Monte Carlo standard error is $\frac{1}{n} \text{Var}[h(\mathbf{X}) w^*(\mathbf{X})]$, based on $g(\mathbf{x})$.

Importance Sampling: Convergence

Consider now the standardised version,

$$\hat{\mu}_{IS} = \sum_{i=1}^n h(\mathbf{X}_i) w(\mathbf{X}_i), \text{ with } w(\mathbf{X}_i) = \frac{w^*(\mathbf{X}_i)}{\sum_{j=1}^n w^*(\mathbf{X}_j)}$$

and define $\hat{\mu}_{IS}^* = \bar{t}/\bar{w}^*(\mathbf{X})$.

Based on Taylor approximation of $\frac{1}{\bar{w}^*}$ around 1,

$$\frac{1}{\bar{w}^*} \approx 1 - (\bar{w}^* - 1) + (\bar{w}^* - 1)^2$$

we get

$$\begin{aligned} E[\hat{\mu}_{IS}] &\approx E\{\bar{t}[1 - (\bar{w}^* - 1) + (\bar{w}^* - 1)^2]\} + \mathcal{O}(n^{-2}) \\ &= E\{\bar{t} - (\bar{t} - \mu)(\bar{w}^* - 1) - \mu(\bar{w}^* - 1) + \bar{t}(\bar{w}^* - 1)^2\} + \mathcal{O}(n^{-2}) \\ &= \mu - \frac{1}{n} \text{cov}[t(\mathbf{X}), w^*(\mathbf{X})] + \frac{\mu}{n} \text{var}[w^*(\mathbf{X})] + \mathcal{O}(n^{-2}) \end{aligned}$$

- standardizing the importance weights introduces a slight bias.

Importance Sampling: Convergence

Similar computations lead to

$$\text{var}[\hat{\mu}_{IS}] = \frac{1}{n} [\text{var}[t(\mathbf{X})] + \mu^2 \text{var}[w^*(\mathbf{X})] - 2\mu \text{cov}[t(\mathbf{X}), w^*(\mathbf{X})] + \mathcal{O}(n^{-2})]$$

Combining the bias and variance estimates, we find

$$\text{MSE}[\hat{\mu}_{IS}] - \text{MSE}[\hat{\mu}_{IS}^*] = \frac{1}{n} (\mu^2 \text{var}[w^*(\mathbf{X})] - 2\mu \text{cov}[t(\mathbf{X}), w^*(\mathbf{X})]) + \mathcal{O}(n^{-2})$$

We have a gain in using the standardised version if

$$\begin{aligned} \text{cov}[t(\mathbf{X}), w^*(\mathbf{X})] &> \frac{\mu \text{var}[w^*(\mathbf{X})]}{2} \\ &\Updownarrow \\ \text{cor}[t(\mathbf{X}), w^*(\mathbf{X})] &> \frac{\sqrt{\text{var}[w^*(\mathbf{X})]}}{2\sqrt{\text{var}[t(\mathbf{X})]}/\mu} = \frac{\text{cv}[w^*(\mathbf{X})]}{2\text{cv}[t(\mathbf{X})]} \end{aligned}$$

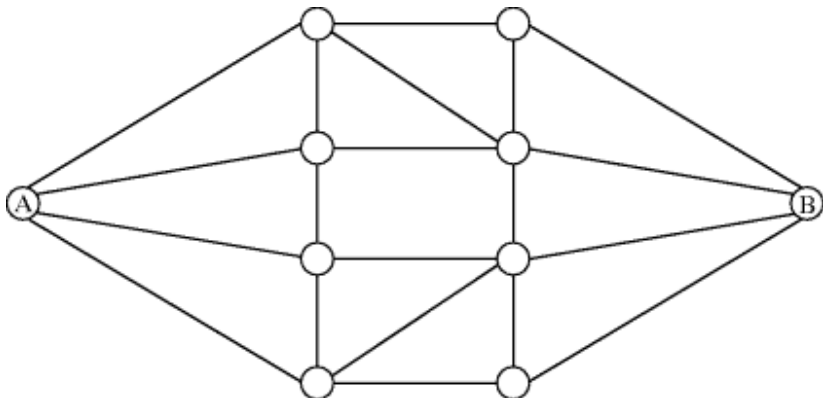
Importance Sampling: Convergence

Choice of g :

- to avoid excessive variability, $f(\mathbf{x})/g(\mathbf{x})$,
 - ▶ must be bounded;
 - ▶ g must have heavier tails than f .
- otherwise some standardized importance weights will be huge,
 - ▶ a rare draw from g with much higher density under f than under g receives huge weight and inflates the variance;
- As $E[f(\mathbf{x})/g(\mathbf{x})] = 1$, the ratio must sometimes be large to counterbalance the predominance of values between 0 and 1;
- to avoid high variance for $h(\mathbf{x})f(\mathbf{x})/g(\mathbf{x})$, choose the function g so that $f(\mathbf{x})/g(\mathbf{x})$ is large only when $h(\mathbf{x})$ is very small.

Importance Sampling: network failure example

Many systems can be represented by connected graphs



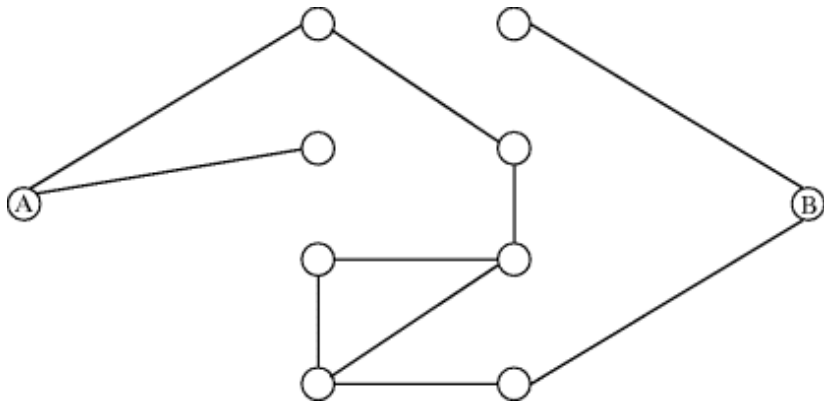
(Givens & Hoeting, 2013, figure 6.10)

Importance Sampling: network failure example

A graph

- A graph is comprised of nodes (circles) and edges (lines);
- A signal sent from A to B must follow a path along any available edges;
- In order for the signal to successfully reach B , a connected path from A to B must exist,
 - ▶ there may be degraded network where only a few routes remain from A to B ;
 - ▶ next figure shows one with 10 broken edges
 - ▶ if the lowest horizontal edge in the figure above were broken, the network would fail.

Importance Sampling: network failure example



(Givens & Hoeting, 2013, figure 6.11)

Importance Sampling: network failure example

Network graphs can be used to model many systems:

- transmission of signals,
 - ▶ analog voice transmission;
 - ▶ electro-magnetic digital signals;
 - ▶ optical transmission of digital data;
- conceptual models,
 - ▶ edges represent different machines or people whose participation may be needed to achieve some outcome.

Typical quantity of interest: probability of network failure given specific probabilities for the failure of each edge

Importance Sampling: network failure example

Simple example:

- Network in the first figure with 20 potential edges;
- each edge fails independently with the same $p = 0.05$.

Let

- $\mathbf{X} = (X_1, \dots, X_{20})$ is the random outcome for the edges;
- $b(\mathbf{X})$ is the number of broken edges in $\mathbf{X}$;
- $h(\mathbf{X})$ is the network failure indicator, so

$$h(\mathbf{X}) = \begin{cases} 1 & \text{if } A \text{ and } B \text{ are not connected} \\ 0 & \text{if } A \text{ and } B \text{ are connected} \end{cases}$$

The probability of network failure is

$$\mu = E\{h(\mathbf{X})\}.$$

Computing μ for a network of any realistic size can be a very difficult combinatorial problem.

Importance Sampling: network failure example

Naive Monte Carlo estimate of μ

- draw $\mathbf{X}_1, \dots, \mathbf{X}_n$ independently and uniformly at random from the set of all possible network configurations whose edges fail independently with probability p ,

$$\hat{\mu}_{\text{MC}} = \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i).$$

- the estimator has variance $\mu(1 - \mu)/n$;

For $n = 100,000$ and $p = 0.05$, simulation yields

- Monte Carlo mean estimate: $\hat{\mu}_{\text{MC}} = 2.00 \times 10^{-5}$;
- Monte Carlo standard error estimate: 1.41×10^{-5} .

Importance Sampling: network failure example

Note:

- $h(\mathbf{X})$ is very rarely 1 unless p is unrealistically large;
- huge number of networks simulations are required in order to estimate μ with sufficient precision;

Importance sampling strategy:

- focus on simulation of $\mathbf{X}$ for which $h(\mathbf{X}) = 1$, compensating for this bias through the assignment of importance weights;
- here using the non-standardized importance weights.

Importance Sampling: network failure example

Simulate $\mathbf{X}_1^*, \dots, \mathbf{X}_n^*$ by

- generating network configurations formed by breaking edges;
- assuming independent edge failure with probability $p^* > p$.

The importance weight for $\mathbf{X}_i^*$ can be written as

$$w^*(\mathbf{X}_i^*) = \left(\frac{1-p}{1-p^*} \right)^{20} \left(\frac{p(1-p^*)}{p^*(1-p)} \right)^{b(\mathbf{X}_i^*)}$$

and the importance sampling estimator of μ is

$$\hat{\mu}_{\text{IS}}^* = \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i^*) w^*(\mathbf{X}_i^*).$$

Importance Sampling: network failure example

Denoting with $\mathcal{F}$ the subset of configurations for which A and B are not connected,

$$\begin{aligned}\text{var}\{\hat{\mu}_{\text{IS}}^*\} &= \frac{1}{n} \text{var}\{h(\mathbf{X}_i^*)w^*(\mathbf{X}_i^*)\} \\ &= \frac{1}{n} \left(\mathbb{E}\{[h(\mathbf{X}_i^*)w^*(\mathbf{X}_i^*)]^2\} - [\mathbb{E}\{h(\mathbf{X}_i^*)w^*(\mathbf{X}_i^*)\}]^2 \right) \\ &= \frac{1}{n} \left(\sum_{\mathbf{x} \in \mathcal{F}} \left(w^*(\mathbf{x})p^b(\mathbf{x})(1-p)^{20-b(\mathbf{x})} \right) - \mu^2 \right)\end{aligned}$$

For our network, failures only occur when $b(X) \geq 4$. Therefore,

$$w^*(\mathbf{X}) = \left(\frac{1-p}{1-p^*} \right)^{20} \left(\frac{p(1-p^*)}{p^*(1-p)} \right)^4.$$

Importance Sampling: network failure example

Using $p^* = 0.25$ and $p = 0.05$, $w^*(\mathbf{X}) \leq 0.07$, and

$$\begin{aligned}\text{var}\{\hat{\mu}_{\text{IS}}^*\} &\leq \frac{1}{n} \left[0.07 \sum_{\mathbf{x} \in \mathcal{F}} p^b(\mathbf{x})(1-p)^{20-b(\mathbf{x})} - \mu^2 \right] \\ &= \frac{1}{n} \left[0.07 \sum_{\mathbf{x} \in \mathcal{C}} h(\mathbf{x}) p^b(\mathbf{x})(1-p)^{20-b(\mathbf{x})} - \mu^2 \right] \\ &= \frac{0.07\mu - \mu^2}{n}\end{aligned}$$

where $\mathcal{C}$ denotes the set of all possible network configurations.

Importance Sampling: network failure example

Therefore:

- $\text{var}\{\hat{\mu}_{\text{IS}}^*\}$ is substantially smaller than $\text{var}\{\hat{\mu}_{\text{IS}}\}$;
- under the approximation that $c\mu - \mu^2 \approx c\mu$, for small μ and relatively larger c ,

$$\text{var}\{\hat{\mu}_{\text{IS}}\}/\text{var}\{\hat{\mu}_{\text{IS}}^*\} \approx 14.$$

Numerically:

- naive simulation strategy using $p = 0.05$,
 - ▶ 2 of 100,000 simulated networks failed;
- importance sampling strategy with $p^* = 0.25$,
 - ▶ 497 of 100,000 simulated networks failed.

Importance sampling estimates with $n = 100,000$ and $p^* = 0.25$,

- Monte Carlo mean estimate: $\hat{\mu}_{\text{IS}}^* = 1.01 \times 10^{-5}$;
- Monte Carlo standard error estimate: 1.56×10^{-6} .

Antithetic sampling: definition

Idea: find two i.i.d. unbiased Monte Carlo estimators, $\hat{\mu}_1$ and $\hat{\mu}_2$, that are negatively correlated, and average them,

$$\hat{\mu}_{AS} = (\hat{\mu}_1 + \hat{\mu}_2)/2$$

The antithetic sampling estimator has variance

$$\text{var}\{\hat{\mu}_{AS}\} = \frac{1}{4}(\text{var}\{\hat{\mu}_1\} + \text{var}\{\hat{\mu}_2\}) + \frac{1}{2}\text{cov}\{\hat{\mu}_1, \hat{\mu}_2\} = \frac{(1 + \rho)\sigma^2}{2n}$$

where

- ρ is the correlation between the two estimators;
- σ^2/n is the variance of either estimator with sample size n .

Note: allocating all $2n$ samples to one estimator would yield a variance of $\sigma^2/2n$, that is larger than $(1 + \rho)\sigma^2/2n$ because $\rho < 0$.

Antithetic sampling: construction

How to construct an identically distributed estimator $\hat{\mu}_2$ that is negatively correlated with a given $\hat{\mu}_1$?

- let $\mathbf{X}$ denote a set of i.i.d. random variables, $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$;
- suppose $\hat{\mu}_1(\mathbf{X}) = \sum_{i=1}^n h_1(\mathbf{X}_i)/n$,
 - ▶ h_1 is a real-valued function with $h_1(\mathbf{X}_i) = h_1(X_{i1}, \dots, X_{im})$;
- assume $E\{h_1(\mathbf{X}_i)\} = \mu$;
- similarly, let $\hat{\mu}_2(\mathbf{X}) = \sum_{i=1}^n h_2(\mathbf{X}_i)/n$.

If h_1 and h_2 are functions of m random variables, $U_1, \dots, U_m$, and if each function is monotone in each argument, one can show that

$$\text{cov}\{h_1(U_1, \dots, U_m), h_2(1 - U_1, \dots, 1 - U_m)\} \leq 0.$$

Antithetic sampling: construction

The Monte Carlo integral estimate $\hat{\mu}_1(\mathbf{X})$ can be written as

$$\hat{\mu}_1(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n h_1(F_1^{-1}(U_{i1}), \dots, F_m^{-1}(U_{im}))$$

where

- F_j is the cumulative distribution function of X_{ij} , $j = 1, \dots, m$
- the U_{ij} are independent $\text{Unif}(0, 1)$ random variables.

Since F_j is a cumulative distribution function,

- its inverse is nondecreasing;
- $h_1(F_1^{-1}(U_{i1}), \dots, F_m^{-1}(U_{im}))$ is monotone in each U_{ij} for $j = 1, \dots, m$ whenever h_1 is monotone in its arguments.

Antithetic sampling: construction

Now, if $U_{ij} \sim \text{Unif}(0, 1)$, then $1 - U_{ij} \sim \text{Unif}(0, 1)$. Hence,

$$h_2(1 - U_{i1}, \dots, 1 - U_{im}) = h_1(F_1^{-1}(1 - U_{i1}), \dots, F_m^{-1}(1 - U_{im}))$$

is monotone in each argument and has the same distribution as $h_1(F_1^{-1}(U_{i1}), \dots, F_m^{-1}(U_{im}))$. Therefore,

$$\hat{\mu}_2(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n h_1(F_1^{-1}(1 - U_{i1}), \dots, F_m^{-1}(1 - U_{im}))$$

is a second estimator of μ having the same distribution as $\hat{\mu}_1(\mathbf{X})$.

Note:

- $\text{cov}\{\hat{\mu}_1(\mathbf{X}), \hat{\mu}_2(\mathbf{X})\} \leq 0$.
- $\hat{\mu}_{\text{AS}} = (\hat{\mu}_1 + \hat{\mu}_2)/2$ has smaller variance than $\hat{\mu}_1$ with $2n$;
- only need to generate a single set of n random numbers, with the other n derived from the antithetic principle.

Antithetic sampling: network failure example

Back to the network failure example:

- let the i th simulated network, $\mathbf{X}_i$, be determined by standard uniform random variables $U_{i1}, \dots, U_{im}$ where $m = 20$;
- $\mathbf{X}_i$ is obtained by breaking the j th edge when $U_{ij} < p$;
- Network failure indicator: $h(\mathbf{X}_i) = h(U_{i1}, \dots, U_{im})$ where

$$h(\mathbf{X}_i) = \begin{cases} 1 & \text{if } A \text{ and } B \text{ are not connected} \\ 0 & \text{if } A \text{ and } B \text{ are connected} \end{cases}$$

Note: h is nondecreasing in each U_{ij}

- Obtain antithetic network draws $\mathbf{X}_i^*$ by breaking the j th edge when $U_{ij} > 1 - p$, for the same set of U_{ij} used to generate $\mathbf{X}_i$;
- $\frac{1}{2n} (\sum_{i=1}^n h(\mathbf{X}_i) + h(\mathbf{X}_i^*))$ is superior to $\frac{1}{2n} \sum_{i=1}^{2n} h(\mathbf{X}_i)$.

Antithetic sampling: normal expectation example

Suppose $X \sim N(0, 1)$ and $\mu = E[h(X)]$, with $h(X) = x/(2^x - 1)$.

Standard Monte Carlo estimator:

- sample $n = 100,000$ values from a standard Gaussian;
- $\hat{\mu}_{\text{MC}} = \sum_{i=1}^{100,000} h(X_i)/100,000$

An antithetic estimator:

- use the first $n = 50,000$ draws;
- the antithetic variate for X_i is simply $-X_i$;
- $\hat{\mu}_{\text{AS}} = \sum_{i=1}^{50,000} [h(X_i) + h(-X_i)]/100,000$.

Simulation results:

- $\widehat{\text{cor}}(h(X_i), h(-X_i)) = -0.95 \rightarrow \text{profitable};$
 - ▶ $\hat{\mu}_{\text{MC}} = 1.4993$ with MC standard error 0.0016;
 - ▶ $\hat{\mu}_{\text{AS}} = 1.4992$ with MC standard error 0.0003.

Control variates: introduction

Idea: improve estimation of an unknown integral by relating the estimate to a correlated estimator of an integral with known value.

Suppose:

- we wish to estimate the unknown quantity $\mu = E\{h(\mathbf{X})\}$
- we know of a related quantity, $\theta = E\{c(\mathbf{Y})\}$, whose value can be determined analytically.

Let $(\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$ denote pairs of random variables observed independently as simulation outcomes.

Control variates: introduction

The simple Monte Carlo estimators are

$$\hat{\mu}_{\text{MC}} = \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i) \quad \text{and} \quad \hat{\theta}_{\text{MC}} = \frac{1}{n} \sum_{i=1}^n c(\mathbf{Y}_i)$$

Note:

- $\hat{\theta}_{\text{MC}}$ is unnecessary since θ can be found analytically;
- $\hat{\mu}_{\text{MC}}$ is correlated with $\hat{\theta}_{\text{MC}}$ when $\text{cor}\{h(\mathbf{X}_i), c(\mathbf{Y}_i)\} \neq 0$;
- e.g., with correlation > 0 , an unusually high outcome for $\hat{\theta}_{\text{MC}}$ should be associated with an unusually high outcome for $\hat{\mu}_{\text{MC}}$.

If comparison of $\hat{\theta}_{\text{MC}}$ to θ suggests such an outcome, then

- adjust $\hat{\mu}_{\text{MC}}$ downward accordingly;
- make opposite adjustment when the correlation is negative.

Control variates: introduction

Control variate estimator

$$\hat{\mu}_{CV} = \hat{\mu}_{MC} + \lambda(\hat{\theta}_{MC} - \theta)$$

where λ is a parameter to be chosen by the user.

Clearly,

$$\text{var}\{\hat{\mu}_{CV}\} = \text{var}\{\hat{\mu}_{MC}\} + \lambda^2 \text{var}\{\hat{\theta}_{MC}\} + 2\lambda \text{cov}\{\hat{\mu}_{MC}, \hat{\theta}_{MC}\}$$

and by minimizing this quantity with respect to λ we can obtain the minimal variance.

Control variates: introduction

Specifically, the minimal variance

$$\min_{\lambda} (\text{var}\{\hat{\mu}_{CV}\}) = \text{var}\{\hat{\mu}_{MC}\} - \frac{(\text{cov}\{\hat{\mu}_{MC}, \hat{\theta}_{MC}\})^2}{\text{var}\{\hat{\theta}_{MC}\}}.$$

is obtained by

$$\lambda = \frac{-\text{cov}\{\hat{\mu}_{MC}, \hat{\theta}_{MC}\}}{\text{var}\{\hat{\theta}_{MC}\}} : \quad (1)$$

- optimal λ depends on unknown moments of $h(\mathbf{X}_i)$ and $c(\mathbf{Y}_i)$
→ can be estimated using $(\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$;
- proportional reduction in variance obtained by using $\hat{\mu}_{CV}$ instead of $\hat{\mu}_{MC}$ is equal to the square of the correlation between $\hat{\mu}_{MC}$ and $\hat{\theta}_{MC}$.

Control variates: Control variates in practice

Consider the regression model

$$E\{h(\mathbf{X}_i)|\mathbf{Y}_i = \mathbf{y}_i\} = \beta_0 + \beta_1 c(\mathbf{y}_i)$$

with the usual regression assumptions and estimators. Then

$$\begin{aligned}\hat{\mu}_{CV} &= \hat{\mu}_{MC} + \hat{\lambda}(\hat{\theta}_{MC} - \theta) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \theta \quad \text{where } \hat{\lambda} = -\hat{\beta}_1\end{aligned}$$

Therefore:

- we can use linear regression software to estimate $\hat{\mu}_{CV}$;
- se of $\hat{\mu}_{CV}$ is the se for a regression fitted value;
- usual tools to obtain confidence interval too.

Control variates: importance sampling example

Importance sampling, i.e., sampling from an envelope that induces a correlation between $h(\mathbf{X})w^*(\mathbf{X})$ and $w^*(\mathbf{X})$:

- we know $E\{w^*(\mathbf{X})\} = 1$;
- use the control variate $\bar{w}^* = \sum_{i=1}^n w^*(\mathbf{X}_i)/n$.

If $\bar{w}^* > 1$, then

- average value of $h(\mathbf{X})w^*(\mathbf{X})$ is also probably unusually high ;
- $\hat{\mu}_{\text{IS}}^*$ probably differs from its expectation μ .

Importance sampling control variate estimator:

$$\hat{\mu}_{\text{ISCV}} = \hat{\mu}_{\text{IS}}^* + \lambda(\bar{w}^* - 1)$$

where $\hat{\mu}_{\text{IS}}^*$ is the estimator with unstandardized weights:

- estimate λ , $\text{se}(\hat{\mu}_{\text{ISCV}})$ from regressing $h(\mathbf{X})w^*(\mathbf{X})$ on $w^*(\mathbf{X})$;
- $\hat{\mu}_{\text{ISCV}}$ often has lower mean squared error than $\hat{\mu}_{\text{IS}}^*$.

Rao-Blackwellization: Introduction

Consider estimation of $\mu = E\{h(\mathbf{X})\}$ using a random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ drawn from f .

Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$ for some partitioning of the coordinates.

Recall:

- $E\{h(\mathbf{X}_i)\} = E\{E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\}$, where the outer expectation is taken with respect to the distribution of $\mathbf{X}_{i2}$;
- Rao-Blackwell theorem states that one can reduce the variance of an unbiased estimator by conditioning the estimator on the sufficient statistics.

Rao-Blackwellization: method

Idea: apply Rao-Blackwell to improve Monte Carlo estimates.

- Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$;
- assume that the conditional expectation $E\{h(\mathbf{X}_i)|\mathbf{x}_{i2}\}$ can be solved analytically.

The Rao-Blackwellized estimator can be defined as

$$\hat{\mu}_{\text{RB}} = \frac{1}{n} \sum_{i=1}^n E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}.$$

Note that:

- same mean as the ordinary Monte Carlo estimator, $\hat{\mu}_{\text{MC}}$;
- $\text{var}\{\hat{\mu}_{\text{MC}}\} = \frac{1}{n} \text{var}\{E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\} + \frac{1}{n} E\{\text{var}\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\} \geq \text{var}\{\hat{\mu}_{\text{RB}}\}$;
- therefore $\hat{\mu}_{\text{RB}}$ is superior to $\hat{\mu}_{\text{MC}}$ in terms of MSE.

Rao-Blackwellization: rejection sampling example

Goal: improve Monte Carlo estimate of μ by using all draws in rejection sampling instead of just the accepted draws.

Recall ordinary rejection sampling:

- candidates $Y_1, \dots, Y_M$ are generated sequentially;
- some candidates are rejected;
- the uniform r.v. $U_1, \dots, U_M$ provide the rejection decisions,
 - ▶ with Y_i being rejected if $U_i > w^*(Y_i)$;
 - ▶ $w^*(Y_i) = f(Y_i)/e(Y_i)$;
- rejection sampling stops at a random time, M , with the acceptance of the n th draw, yielding $X_1, \dots, X_n$.

Rao-Blackwellization: rejection sampling example

The ordinary Monte Carlo estimator of $\mu = E\{h(X)\}$ can then be re-expressed as

$$\hat{\mu}_{\text{MC}} = \frac{1}{n} \sum_{i=1}^M h(Y_i) 1_{\{U_i \leq w^*(Y_i)\}}, \quad (2)$$

Rao-Blackwellization of it yields the estimator

$$\hat{\mu}_{\text{RB}} = \frac{1}{n} \sum_{i=1}^M h(Y_i) t_i(\mathbf{Y})$$

where the $t_i(\mathbf{Y})$ are random quantities that depend on M and $\mathbf{Y} = (Y_1, \dots, Y_M)$ according to

$$\begin{aligned} t_i(\mathbf{Y}) &= E\{1_{\{U_i \leq w^*(Y_i)\}} | M, Y_1, \dots, Y_M\} \\ &= P[U_i < w^*(Y_i) | M, Y_1, \dots, Y_M]. \end{aligned} \quad (3)$$

Rao-Blackwellization: rejection sampling example

- $t_M(\mathbf{Y}) = 1$ since the final candidate was accepted;
- for other candidates, the probability can be found by averaging over permutations of subsets of the realized sample,

$$t_i(\mathbf{Y}) = \frac{w^*(Y_i) \sum_{A \in \mathcal{A}_i} \prod_{j \in A} w^*(Y_j) \prod_{j \notin A} (1 - w^*(Y_j))}{\sum_{B \in \mathcal{B}} \prod_{j \in B} w^*(Y_j) \prod_{j \notin B} (1 - w^*(Y_j))}.$$

where

- ▶ $\mathcal{A}_i$ is the set of all subsets of $\{1, \dots, i-1, i+1, \dots, M-1\}$ containing $n-2$ elements;
 - ▶ $\mathcal{B}$ is the set of all subsets of $\{1, \dots, M-1\}$ containing $n-1$ elements.
- $\hat{\mu}_{MC}$ is improved by using all the Y_i draws, suitably weighted.

References |

GIVENS, G. H. & HOETING, J. A. (2013). *Computational Statistics*. Wiley.