JHU 553.633/433: Monte Carlo Methods
J. C. Spall
24 October 2018

Variance Reduction for Monte Carlo Simulation

(Chapter 5 in Textbook)

- Common random numbers
- Antithetic method
- Importance Sampling

Introduction

- **Motivation for Chapter 5:** One great advantage of simulation experiments—versus real-world experiments—is possibility for variance reduction of output
  - Possible because analyst "controls" nature with simulation (e.g., choice of random number generator and seed)
- Aim is to use information on model to transform output of simulation model such that estimated quantities such as $\ell = E[H(\boldsymbol{X})]$ (Chap. 4) are more accurate
  - Estimated quantities other than ℓ may also be of interest (e.g., differences of two simulation outputs with different inputs)
- Improved accuracy is obtained by reducing the variance of the estimate
- Variance reduction is achievable in output from simulation runs in a way not possible with comparable output from nature

2

Common Random Numbers

- Common random numbers (CRNs) useful when considering **differences** of simulation outputs
- In simulation (vs. real world), CRNs often feasible
- Differences, say $X - Y$, come up in many applications:
  - Comparison of two scenarios: which is better?
    - Airport configurations
    - Investment strategies
    - Environmental policies
    - Etc.
  - Finite difference and simultaneous perturbation gradient approximations (used in optimization and numerical analysis; see Spall, 2003)
  - Which of two algorithms is better to use on given problem?
  - Statistical two-sample tests
  - Etc.
- "Despite its simplicity, CRNs is the most useful and popular variance reduction technique" (Law, 2014, Chap. 11)

3

Common Random Numbers (cont'd)

- Recall CRNs useful when considering differences of simulation outputs
- Intuition: Two simulation runs contributing to difference quantity of interest should rely as much as possible on same underlying randomness (i.e., generated random variables)

  $\Rightarrow$ observed differences are due to differences in system configurations rather than fluctuations of experimental conditions
- CRNs based on famous formula for random variables X, Y:

  $$\boxed{\mathrm{var}(X - Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X, Y)}$$
- Maximizing covariance minimizes variance of difference
- **Practical implementation of CRNs usually achieved by simply running the two simulations with same seed**

4

Common Random Numbers (cont'd)

- Note: CRNs do ***not*** require that inverse transform for cdf exists (in contrast to implication of textbook)
  - But, some nice results follow if inverse transform exists….
- Motivational example: Let X and Y be random variables with known cdfs F_1 and F_2; wish to estimate $E(X - Y)$
  - Crude Monte Carlo method: generate ***independent*** random samples of X and Y and use $X - Y$ as unbiased estimate
- CRN estimate for $E(X - Y)$ using inverse transform:
  - CRN uses same ("common") U to generate both X and Y
  - Using inverse transform: $X - Y = F_1^{-1}(U) - F_2^{-1}(U)$
  - By monotonicity of F^{-1}: $\operatorname{cov}\left(F_1^{-1}(U), F_2^{-1}(U)\right) \geq 0$
  - CRN maximizes covariance $\Rightarrow$ minimizes var($X - Y$)
- More general result:
  - Estimate $E[H_1(X) - H_2(Y)]$, where H_1 and H_2 are *monotonic in same direction* (e.g. $E(X^2 - Y^2)$, $X \geq 0, Y \geq 0$)
  - CRN $H_1\left(F_1^{-1}(U)\right) - H_2\left(F_2^{-1}(U)\right)$ minimizes variance of $H_1(X) - H_2(Y)$

5

Common Random Numbers: Multivariate Case

- Most practical Monte Carlo simulations involve **many** random effects; motivates need for considering vectors
- Vector case: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are n-dimensional random vectors with component distributions F_i and G_i, respectively
  - Both $\boldsymbol{X}$ and $\boldsymbol{Y}$ have independent components
  - H_1 and H_2 are monotonic in same direction in each component; e.g., sign($\partial H_1/\partial X_i$) = sign($\partial H_2/\partial Y_i$) for all i if differentiable
  - Can be shown that $H_1\left(F_1^{-1}(U_1), \ldots, F_n^{-1}(U_n)\right) - H_2\left(G_1^{-1}(U_1), \ldots, G_n^{-1}(U_n)\right)$ minimizes variance **when using inverse transform**
  - Only permits dependence between *like components* (i.e., X_i and Y_i, not across i
- Implementation Issues:
  - Component-wise independence might be difficult to achieve
  - *Synchronization* problem: *matching up* like components can be difficult

6

Synchronization Issue

- To gain benefit of CRNs, want randomness in two simulations used in same way as much as possible: ***synchronization***
- Typically comparing two simulations under two different system configurations
  - Conservative investment strategy vs. aggressive investment strategy
  - Current signal light timings vs. new signal light timings
  - Reliability analysis with minimal parts replacement strategy vs. frequent (proactive) parts replacement strategy
  - Etc.
- Full benefit of CRNs requires synchronization of random number streams for both simulations:
  1. Same total quantity of random numbers in both simulations
  2. Each random number used for specific purpose in one configuration is used for exactly same purpose in other configuration

7

Example: Lack of Synchronization

- Consider two simulation outputs $H_1(\boldsymbol{X})$ and $H_2(\boldsymbol{Y})$ to be compared: Simulations have different "policies" (settings)
- Example: outputs represent total vehicle wait time in urban transportation network under two traffic control policies
  - Table below shows how things can go wrong
  - Consider random number stream $\{U_1, U_2, \text{etc.}\}$

Input	$X_i = F_i^{-1}(U_i)$	$Y_i = G_i^{-1}(U_i)$	Synchronized?
U_1	Arrival 1 into network	Arrival 1 into network	**Yes**
U_2	Arrival 1 turns left at intersection 1	Arrival 1 hits red light at intersection 1 (blocked)	**Yes (randomness for behavior at intersection 1)**
U_3	Random behavior of arrival 1 after intersection 1	Not useful for arrival 1; use for arrival 2 into network	**No (randomness applies to different entities)**

- Randomness from U_3 onward not synchronized; lose benefits of CRNs

8

Partial Synchronization

- Full synchronization often not possible
  - Two simulations may inherently require different number of random variables and/or different usage
- Partial synchronization, leading to partial CRNs, can provide some benefit of CRNs
  - Partial CRNs still makes two simulations more alike than two independent simulations
- Aim is to make as many of random variables as possible be used in same way
- Three common ways to achieve partial CRNs:
  1. Run two simulations with same seed and "hope" that two stochastic processes stay matched in random number usage as long as possible
  2. Use separate random number streams for separate process; e.g., one stream for server, one stream for inter-arrival times in queue
  3. Use inverse transform (vs. A-R method) in order to know a priori how many numbers are needed

9

Example 1 of CRNs: Queuing Models

- Comparison of *M/M/1* and *M/M/2* queues:
  - Configuration 1 (*M/M/1*): arrival rate 1 customer/minute; mean service time 0.9 minutes.
  - Configuration 2 (*M/M/2*): arrival rate 1 customer/minute; mean service time 1.8 minutes on both servers.
  - Objective: select the configuration with smaller expected average delay (averaged over 100 delays). (Configuration 2 is superior based on analytical solutions: 4.13 for configuration 1; 3.70 for configuration 2; ability to compute delay not part of 553.633)
  - Note: Running two configurations using one string of random numbers does not work due to lack of synchronization
  - Synchronization possible by using separate common random number streams for arrival and service random variables; for one run of the two configurations:

    Use $U_1^a, U_2^a, \ldots$ to drive arrival processes for both configurations;

    Use $U_1^s, U_2^s, \ldots$ to drive service processes for both configurations

10

Example 2 of CRNs: Derivative Estimation

- Suppose that performance measure ℓ (Chap. 4) depends on scalar parameter θ, i.e., $\ell = \ell(\theta) = E[H(\boldsymbol{X},\theta)]$
- Interested in approximation of $d\ell(\theta)/d\theta$ from only two simulation runs
  - Derivative widely used in sensitivity studies and optimization
- Finite difference approximation is $FD \equiv [\ell(\theta+\delta) - \ell(\theta-\delta)]/(2\delta)$, where $\delta > 0$ is "small"
- But ℓ not available; use H instead (H is simulation output)
- Estimates of $d\ell(\theta)/d\theta$: Non-CRN has independent $\boldsymbol{X}$ and $\boldsymbol{Y}$; CRN has common $\boldsymbol{X}$:

$$\widehat{\nabla\ell}_{\text{non-CRN}} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{H(\boldsymbol{X}_i,\theta+\delta)-H(\boldsymbol{Y}_i,\theta-\delta)}{2\delta};\ \widehat{\nabla\ell}_{\text{CRN}} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{H(\boldsymbol{X}_i,\theta+\delta)-H(\boldsymbol{X}_i,\theta-\delta)}{2\delta}$$

- Note that $E\left(\widehat{\nabla\ell}_{\text{non-CRN}}\right)=E\left(\widehat{\nabla\ell}_{\text{CRN}}\right) = FD$
- Most importantly, CRN leads to variance reduction for small δ:

$$\boxed{\operatorname{var}\left(\widehat{\nabla\ell}_{\text{non-CRN}}\right) \propto \frac{1}{n\delta^2} \ \text{ and } \ \operatorname{var}\left(\widehat{\nabla\ell}_{\text{CRN}}\right) \propto \frac{1}{n}}$$

11

Antithetic Method

- Antithetic random variable (ARV) estimate for $E(X+Y)$
  - $F_1^{-1}(U)+F_2^{-1}(1-U)$ minimizes variance of $X+Y$
- More general result
  - Estimate $E[H_1(X) - H_2(Y)]$
  - H_1 and H_2 are *monotonic in opposite direction*
  - ARV estimate $H_1\left(F_1^{-1}(U)\right)-H_2\left(F_2^{-1}(1-U)\right)$ minimizes var$[H_1(X) - H_2(Y)]$
- *Example 1:* Estimate $\ell = E[H(X)]$ based on N simulation runs $H(X_1),\ldots, H(X_N)$
  - Crude Monte Carlo (independent summands): $N^{-1}\sum_{i=1}^{N}H\left(F^{-1}(U_i)\right)$
  - ARV estimate: $N^{-1}\sum_{i=1}^{N/2}\left[H\left(F^{-1}(U_i)\right)+H\left(F^{-1}(1-U_i)\right)\right]$
- *Example 2:* Used in Spall (2005)* for variance reduction for sums of random Hessian matrix estimates to calculate Fisher information matrix. (*Spall, J. C. (2005), "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings," *J. Comp. Graphical Stat.,* vol. 14(4), pp. 889–909)

12

Importance Sampling

- "The most fundamental variance reduction technique is *importance sampling*" (Rubinstein and Kroese, 2008)
- Let $\boldsymbol{X}$ be random vector with density function $f(\boldsymbol{x})$ and $H(\boldsymbol{X})$ a quantity of interest. The crude Monte Carlo estimate of $E[H(\boldsymbol{X})]$ is $\frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{X}_i),\ \boldsymbol{X}_1,\ldots,\boldsymbol{X}_N \sim f(\boldsymbol{x})$
- Let $g(\boldsymbol{x})$ be density function that dominates $f(\boldsymbol{x})$ ($g(\boldsymbol{x}) = 0 \Rightarrow H(\boldsymbol{x})f(\boldsymbol{x}) = 0$) (*note:* "dominates" not same as "majorizes" for accept–reject method of Chap. 2). Then,

$$E\left[H(\boldsymbol{X})\right] = \int H(\boldsymbol{x})\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x} = E_g\left[H(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right]$$

- Importance sampling estimate of $\ell = E[H(X)]$ by simulating from g: $\boxed{\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{X}_i)\frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)},\ \boldsymbol{X}_1,\ldots,\boldsymbol{X}_N \sim g(\boldsymbol{x})}$
- Note:
  - (i) $g(\boldsymbol{x})$ is *instrumental density*; few restrictions on g. Note that "dominates" slightly weaker than supp(f) $\subseteq$ supp(g)
  - (ii) $f(\boldsymbol{x})/g(\boldsymbol{x})$ is called ***likelihood ratio***

13

Importance Sampling— Choice of *g*

- Poor choice of g can lead to unbounded variance.
  - Avoid choices of g with $E_f\left[\frac{H^2(\boldsymbol{X})f(\boldsymbol{X})}{g(\boldsymbol{X})}\right] = \int \frac{H^2(\boldsymbol{x})f^2(\boldsymbol{x})}{g(\boldsymbol{x})}d\boldsymbol{x} = \infty$
  - Avoid g having "light tail" relative to f
- Optimal instrumental density $g^*(z)$ can be obtained by minimizing the variance of the importance estimate

$$g^*(\boldsymbol{x}) = \frac{|H(\boldsymbol{x})|f(\boldsymbol{x})}{\int |H(\boldsymbol{z})|f(\boldsymbol{z})d\boldsymbol{z}}$$

  - Note that if $H(\boldsymbol{x}) \geq 0$, then $g^*(\boldsymbol{x}) = H(\boldsymbol{x})f(\boldsymbol{x})/E[H(\boldsymbol{X})] = H(\boldsymbol{x})f(\boldsymbol{x})/\ell$; hence optimal choice is not practical
- In practice, we should choose g such that $g(\boldsymbol{x})$ is nearly proportional to $|H(\boldsymbol{x})|f(\boldsymbol{x})$
  - That is, $|H(\boldsymbol{x})|f(\boldsymbol{x})/g(\boldsymbol{x})$ is almost constant.

14

Importance Sampling—Example

- Consider estimating $E(X)$ with the following density of X:

$$f(x) = \begin{cases} 2x & 0 \le x \le 1, \\ 0 & \text{otherwise} \end{cases}$$

- Crude Monte Carlo:

$$X_1, \ldots, X_N \sim f(x)$$

$$\frac{1}{N}\sum_{i=1}^{N} H(X_i) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

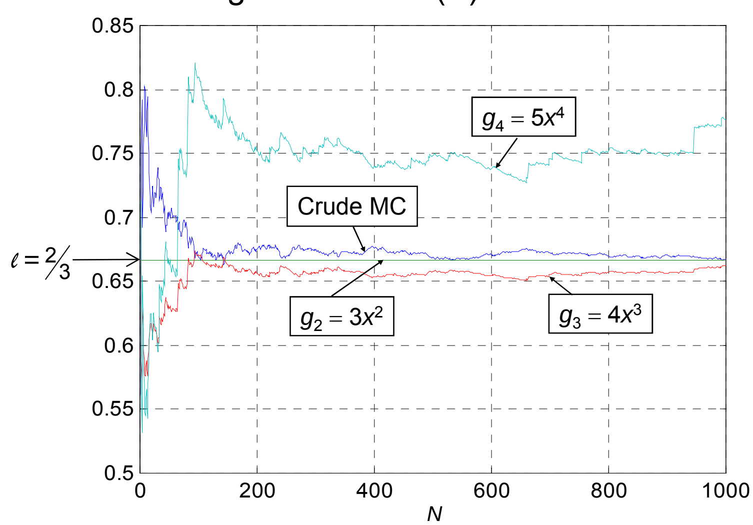
- Importance Sampling with the instrumental density:

$$g_n(x) = \begin{cases} (n+1)x^n & 0 \le x \le 1, \\ 0 & \text{otherwise} \end{cases}$$

- Note: $g^*(x) = H(x)f(x)/E[H(X)] = x(2x)/(2/3) = 3x^2$

Importance Sampling—Example, cont'd

- Plot shows $\hat{\ell}$ as a function of N for different pdfs $g(\boldsymbol{x})$
- Note that sampling with $g^*(x) = g_2(x) = 3x^2$ provides "instant" convergence to $\ell = E(X) = 2/3$



Importance Sampling—Special Case of g Having Same Form as f

- Often, we choose $g(\boldsymbol{x}) = f(\boldsymbol{x}|\theta')$, where θ' is specific value of parameter in f
  - θ' is called **reference parameter**
- Advantage is simplicity and general guarantee that dominance condition is satisfied
- Some theory exists for optimal (or at least reasonable) choice of θ'
  - Finding optimal sampling distribution reduces to finding optimal θ' ("only" a ***parameter*** optimization problem vs. a ***functional*** optimization problem in general)
- Estimate for $\ell = \ell(\theta)$ is

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{X}_i \mid \theta)\frac{f(\boldsymbol{X}_i|\theta)}{f(\boldsymbol{X}_i|\theta')}$$

17

References

- Glasserman, P. and D. D. Yao (1992), "Some Guidelines and Guarantees for Common Random Numbers," *Management Science*, vol. 38, No. 6.
- Glynn, P. W. and D. L. Iglehart (1989), "Important Sampling for Stochastic Simulations," *Management Science*, vol. 35, No. 11.
- Law, A. M. (2014), *Simulation Modeling and Analysis* (5th ed.), McGraw-Hill, New York.
- Rubinstein, R. Y. and Kroese, D. P. (2017), *Simulation and the Monte Carlo Method* (3rd ed.), Wiley, Hoboken, NJ.
- Rubinstein, R.Y. and B. Melamed (1998), *Modern Simulation and Modeling*, Chapter 4, Wiley.
- Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization*, Wiley, Hoboken, NJ.

18