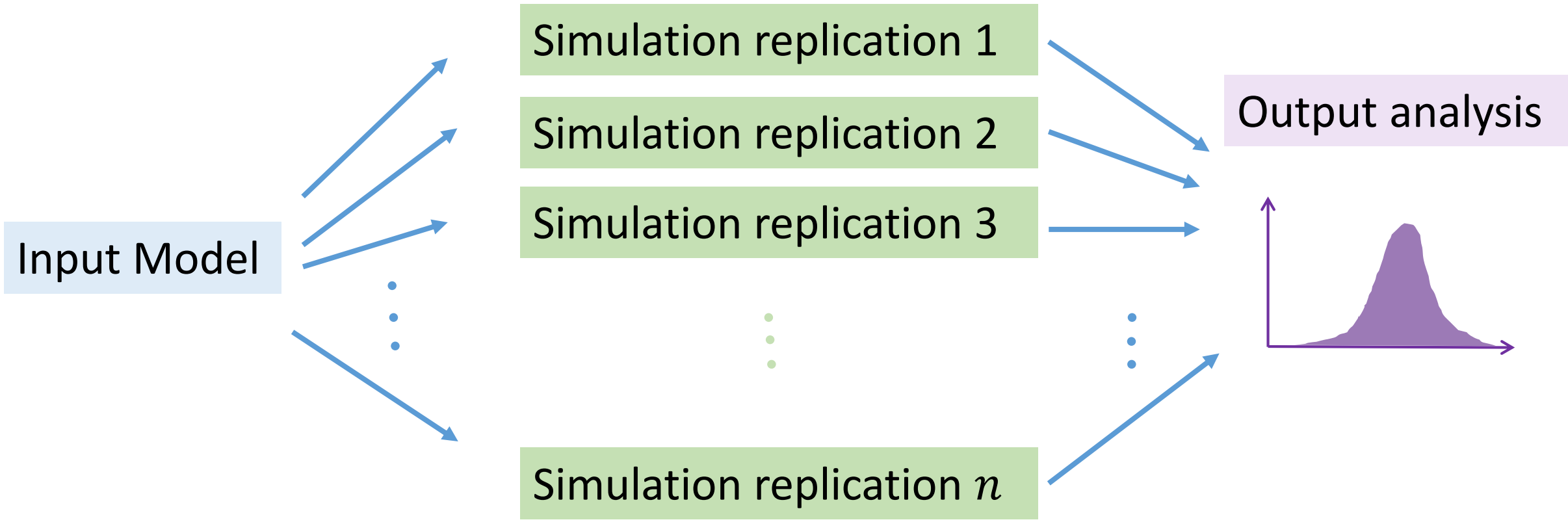
Variance Reduction and Rare-Event Simulation

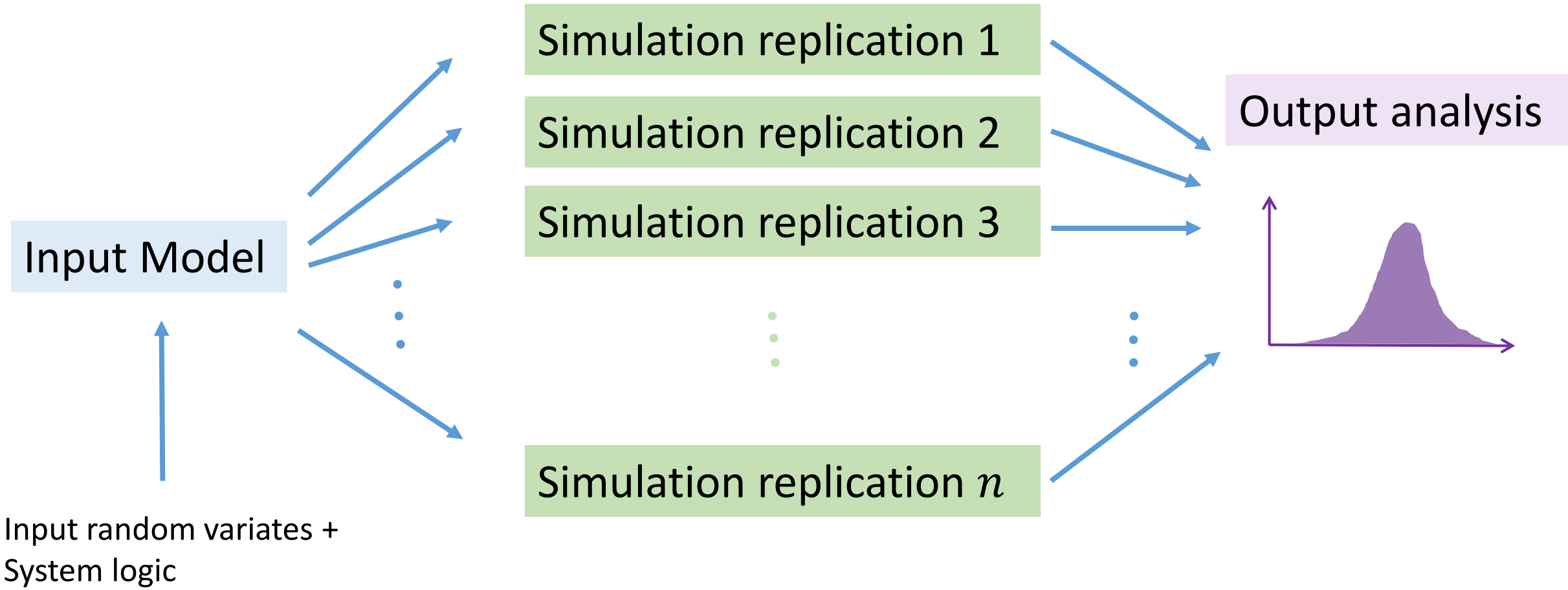
Swiss Doctoral School
Sept 12-15, 2021

Henry Lam
Columbia University

Basics of Monte Carlo



Basics of Monte Carlo



Basics of Monte Carlo

Example – Discrete-event / queueing model:

- Input variates: $X = (X^1, X^2, \dots)$ = interarrival/service time sequences
- System logic: $h(\cdot)$ = a map that "transforms" the sequences into the average waiting time over some time horizon
- Target output quantity: $E[h(X)]$ = expected average waiting time

In a single-server system,

$h(X^1, \dots, X^T)$ = waiting time of the T-th customer can be expressed as a recursion that outputs W^T, with $W^1 = 0$, $W^t = (W^{t-1} + X^t)_+$ for $t = 1, \dots, T$ (Lindley's recursion)

Goals of Monte Carlo Simulation

Prediction / decision-making using complex stochastic models that are otherwise intractable:

- Optimization: Faster service reduces waiting time but needs higher cost (more staffing); decision variable = staffing rule
- Sensitivity analysis: Impact on waiting time due to increase in arrival rate
- Feasibility analysis: Test if a staffing rule achieves a performance standard

These tasks lead to simulation-based optimization, stochastic gradient estimation, stochastic programming…

Other Examples

Operations Research:

- Financial option pricing: Estimate the expected payoff of a process governed by stochastic differential equation
- Inventory management: Estimate the expected revenue/cost of inventory policy under uncertain demand
- Emergency response system: Estimate the response times and availability of EMS ambulances

Machine learning:

- Model-based reinforcement learning

Statistics:

- Bayesian computation, e.g., MCMC

Others: Multi-agent systems, physical and biological simulation…

Some Basics

Interested in predicting: $\mu = E[h(\boldsymbol{X})]$

- Monte Carlo uses computer to repeatedly generate i.i.d. copies of $\boldsymbol{X}_i$ (plugged into $h(\boldsymbol{X}_i)$), $i = 1, \dots, n$
- Report an estimate

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i)$$

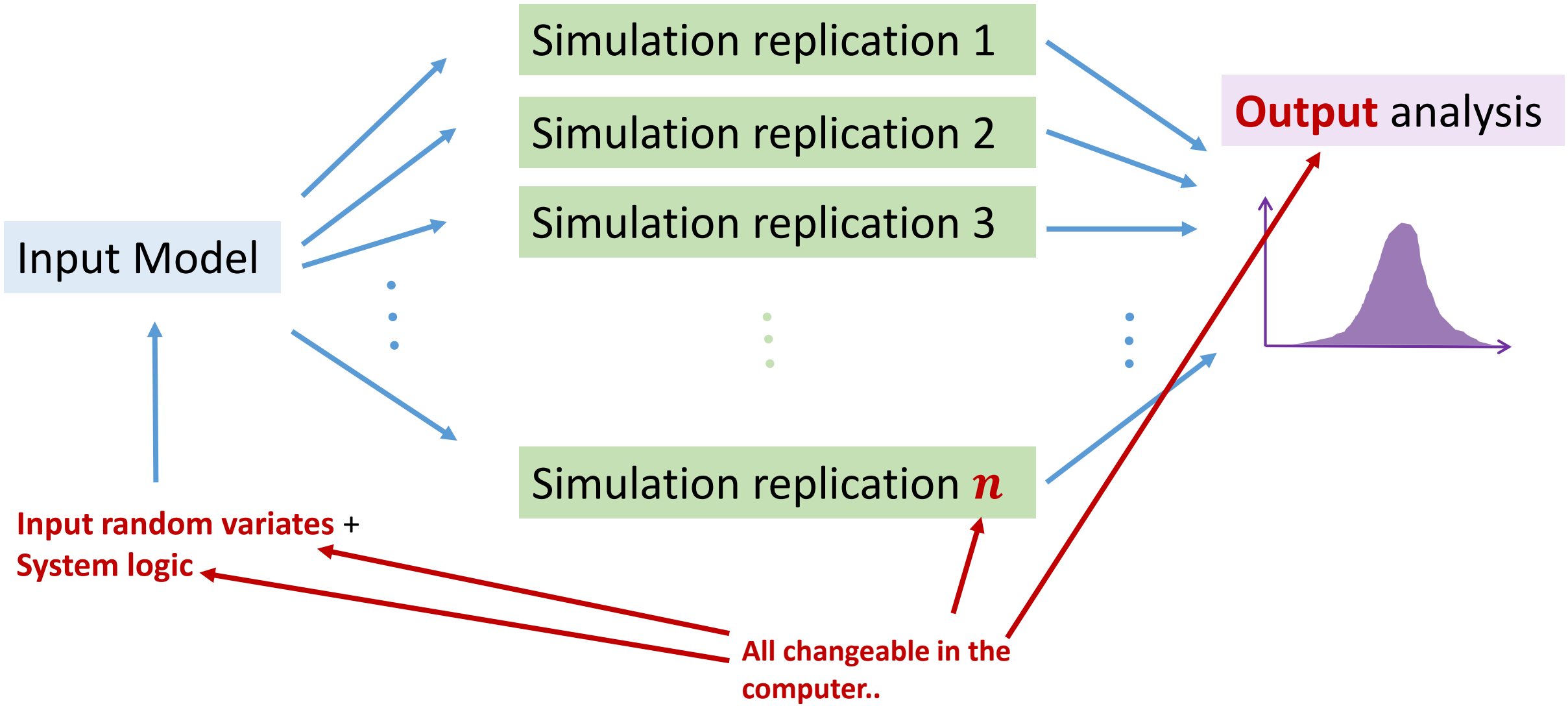
Law of large numbers: Given $E|h(\boldsymbol{X})| < \infty$, we have $\bar{Y} \to \mu$ almost surely

- This implies $\bar{Y}$ is a consistent point estimator for μ

Central limit theorem (CLT): Given $\sigma^2 = Var\big(h(\boldsymbol{X})\big) < \infty$, we have $\sqrt{n}(\bar{Y} - \mu)/\sigma \Rightarrow N(0,1)$. This means:

- $\bar{Y} = \mu + O_p\left(\frac{1}{\sqrt{n}}\right)$ has a canonical square-root convergence rate
- $\left[\bar{Y} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{Y} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$ is a $(1-\alpha)$-level confidence interval for μ, where S^2 is the sample variance of $h(\boldsymbol{X}_i)$ and $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of $N(0,1)$

Difference with Standard Statistics..



Difference with Standard Statistics..

Example – Discrete-event / queueing model:

- Input variates: $X = (X^1, X^2, \dots)$ = interarrival/service time sequences
- System logic: $h(\cdot)$ = a map that "transforms" the sequences into the average waiting time over some time horizon
- Target output quantity: $E[h(X)]$ = expected average waiting time

Their distributions can be distorted

Can be replaced by simpler "metamodel"

Can output other additional quantities too from the simulation

In a single-server system,

$h(X^1, \dots, X^T)$ = waiting time of the T-th customer can be expressed as a recursion that outputs W^T, with $W^1 = 0$, $W^t = (W^{t-1} + X^t)_+$ for $t = 1, \dots, T$ (Lindley's recursion)

Overview of Topics

- Variance reduction: Approaches to reduce MC error / speed up MC
  - Introduce most common techniques
  - Motivation
  - Connection to bias reduction
  - Connection to some applications in stochastic optimization and machine learning

- Rare-event simulation: Estimation of tail probabilities
  - Large-deviations-based importance sampling
  - Cross-entropy method
  - Multilevel splitting

Monte Carlo Error

- μ is a target quantity to be estimated, e.g., $\mu = E[h(X)]$

- $\hat{\mu}$ is a Monte Carlo estimator using, say, n simulation runs

- Mean squared error (MSE) = $E(\hat{\mu} - \mu)^2$ = $\text{bias}^2 + \text{variance}$

where

$\text{bias} = E[\hat{\mu}] - \mu$

$\text{variance} = Var(\hat{\mu})$

- For unbiased estimator (i.e., $bias$ = 0), the lower the variance, the better

Variance Reduction

Suppose $\hat{\mu}$ is the average of i.i.d. unbiased sample with (per-run) variance σ^2. Two views:

View 1: Half-width of confidence interval = $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

View 2: High-probability estimation discrepancy

$$P(|\hat{\mu} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

To reduce half-width or estimation discrepancy, we reduce the estimation variance σ^2/n by either:

- Increase sample size
- Reduce σ^2

Our focus next. When is reducing σ^2 worthwhile?

Variance Reduction

Methods to reduce σ^2:

- Importance sampling
- Control variates
- Multilevel MC
- Conditional MC
- Common random numbers
- Stratification
- Quasi MC…

Variance Reduction

Methods to reduce σ^2:

- Importance sampling
- Control variates
- Multilevel MC
- Conditional MC
- Common random numbers
- Stratification
- Quasi MC…

Need to use some information about system structure – no free lunch

Variance Reduction

Methods to reduce σ^2:

- Importance sampling
- Control variates
- Multilevel MC
- Conditional MC
- Common random numbers
- Stratification
- Quasi MC…

Need to use some information about system structure – no free lunch

Variance Reduction

Methods to reduce σ^2:

- Importance sampling
- Control variates
- Multilevel MC
- Conditional MC
- Common random numbers
- Stratification
- Quasi MC…

Need to use some information about system structure – no free lunch

Importance Sampling

Goal: Estimate $E[h(X)]$ under $X \sim P$

Naïve MC: Generate X_i from P and output $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$

IS idea:

- Use a different distribution $\tilde{P}$ to generate X, which gives a biased estimate
- Multiply by a likelihood ratio to de-bias the estimate

Importance Sampling

Mathematically,

- Use $\tilde{P}$ to generate X, and output $h(X)L(X)$ where L is known as the likelihood ratio

$$L(X) = \frac{dP}{d\tilde{P}}(X) = \text{Radon-Nikodym derivative between } P \text{ and } \tilde{P}$$

- With n simulation runs, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i)L(X_i)$$

where X_i is generated from the IS distribution $\tilde{P}$

Importance Sampling

Requirement: $h(x)dP(x)$ is absolutely continuous w.r.t. $d\tilde{P}(x)$

i.e., if $d\tilde{P}(x) = 0$, then $h(x)dP(x) = 0$, or
if $h(x)dP(x) \neq 0$, then $d\tilde{P}(x) \neq 0$

Example: To estimate $P(X > 10)$ for $X \sim N(0,1)$, can we use $\tilde{P} =$

- $N(20,20)$
- $Uniform(-20,20)$
- $Exp(1)$

Importance Sampling

Claim: The IS $h(X)L(X)$, where $X \sim \tilde{P}$, is unbiased

Importance Sampling

Claim: The IS $h(X)L(X)$, where $X \sim \tilde{P}$, is unbiased

Reasoning:

$$E[h(X)] = \int h(x)dP(x) = \int h(x)\frac{dP(x)}{d\tilde{P}(x)}d\tilde{P}(x) = \tilde{E}[h(X)L(X)]$$

Only makes sense if $h(x)\frac{dP(x)}{d\tilde{P}(x)}$ is well-defined, i.e., absolute continuity condition holds

A “change of measure” idea

Optimal Importance Sampling

IS is used in the hope that the variance is reduced compared to crude MC. How do we know variance can be reduced?

- Consider an IS P^* defined as

$$\frac{dP^*}{dP}(X) = \frac{|h(X)|}{E|h(X)|}$$

Then using P^* gives the least variance among all legitimate IS

Furthermore,

- If $h \geq 0$, then the variance of an IS sample from P^* has zero variance
- If $h(x) = I(x \in A)$, i.e., we estimate the probability of event A, then the best IS distribution

$$\frac{dP^*}{dP}(X) = \frac{|h(X)|}{E|h(X)|} = \frac{I(X \in A)}{P(A)}$$

is the conditional distribution given A (very intuitive!)

Optimal Importance Sampling

- While P^* is not implementable (in a meaningful way), the above guides us that a good IS $\tilde{P}$ approximates the conditional distribution given the considered event

- IS is a powerful technique in that it can enhance estimation efficiency substantially (exponentially) if carefully designed

- IS is a “double-edged sword”: It can also hurt efficiency substantially (very high variance) if poorly designed, despite its unbiasedness

Control Variate

In estimating $E[h(X)]$, suppose we can also generate an "auxiliary" quantity Y whenever we simulate $h(X)$

Suppose we know the information $E[Y] = \mu$

We can use Y as a control variate (CV). This means we output

$$h(X) + \beta(Y - \mu)$$

from our simulation run

Claim: The CV estimator is unbiased

Control Variate

- More generally, given we can simulate $(h(X), Y)$ together, and know the information $E[Y] = \mu$ for $Y \in R^d$, the CV output is
$$h(X) + \beta'(Y - \mu)$$

- If we have n simulation runs, we output
$$\frac{1}{n}\sum_{i=1}^{n}\left(h(X_i) + \beta'(Y_i - \mu)\right)$$

where $(h(X_i), Y_i)$ is the i-th simulation outcome

Control Variate

How can CV reduce variance?

- We choose β such that the variance is reduced

$$Var\big(h(X) + \beta'(Y - \mu)\big) = Var\big(h(X)\big) + 2\beta' Cov(h(X), Y) + \beta' Var(Y)\beta$$

Minimizing over β, we get $\beta = -Var(Y)^{-1}Cov(h(X), Y)$

- The variance is

$$Var\big(h(X)\big) - Cov(h(X), Y)'Var(Y)^{-1}Cov(h(X), Y)$$

- To achieve variance reduction, $h(X)$ and Y needs to have non-zero correlation, and the higher in magnitude the better, i.e., Y provides some information for $h(X)$ (very intuitive!)

Control Variate

- In practice, the optimal $\beta^* = -Var(Y)^{-1}Cov(h(X), Y)$ is unknown and estimated via its sample counterpart

$$\hat{\beta} = -\widehat{Var}(Y)^{-1}\widehat{Cov}(h(X), Y)$$

- Final output is

$$\frac{1}{n}\sum_{i=1}^{n}\left(h(X_i) + \hat{\beta}'(Y_i - \mu)\right)$$

This introduces a small bias (negligible relative to the standard deviation. How to show?)

Control Variate

Control variate estimator is equivalent to estimating the intercept of a linear regression that regresses $h(X)$ against $Y - \mu$

- $\widehat{Var}(Y)^{-1}\widehat{Cov}(h(X), Y)$ is the vector of estimated coefficients for $Y - \mu$
- The final control variate output $\overline{h(X)} - \widehat{Var}(Y)^{-1}\widehat{Cov}(h(X), Y)(\bar{Y} - \mu)$ is the estimated intercept
- Estimating the intercept of the regression has a lower variance than estimating the mean of the response variable

When Do We Need Variance Reduction?

- Not always needed, given nowadays' computational power to generate many simulation runs easily
- Useful when reducing σ^2 is much more effective than increasing n. Examples:
  - Rare-event simulation: one may need an "exponential" n to achieve a meaningful estimation error
  - When each output sample has a huge variance: gradient estimators that suffer from the "curse of horizon"
  - When many estimators are needed: E.g., running gradient descent to solve an optimization
  - Super-canonical convergence: distort the fundamental convergence speed in n from square-root scaling in CLT to faster

Rare-Event Simulation

- In rare-event probability estimation, i.e., the target quantity p = P(rare-event) is very small, we want a point estimate $\hat{p}$ to be close to p relative to the magnitude of p
- By Markov inequality,

$$P(|\hat{p} - p| > \epsilon p) \leq \frac{\sigma^2}{n\epsilon^2 p^2}$$

- The needed n to achieve a relative discrepancy of ϵ with confidence $1 - \alpha$ is $\geq \frac{\sigma^2}{\alpha\epsilon^2 p^2}$

Rare-Event Simulation

Consider simulating $P(X > 10)$ where $X \sim N(0,1)$

- Crude Monte Carlo: run n simulation runs and obtain

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i > 10)$$

- Suppose we want to estimate p within 5% of the truth, with confidence 95%, then we need a sample size to be

$$\frac{\sigma^2}{5\% \times\ 5\%^2 \times p^2}$$

where $\sigma^2 = p(1-p)$

- This number turns out to be around 10^{26}
- If we simulate 1 million normal variables in a millisecond, we need 3 million years to finish

Rare-Event Simulation

- Required sample size n to achieve a relative discrepancy of ϵ with confidence $1-\alpha$ is $\frac{\sigma^2}{\alpha\epsilon^2 p^2}$
- Define relative error = $\frac{\sqrt{Var(Z)}}{E[Z]} = \frac{\sigma}{p}$
- Crude MC has RE = $\frac{\sqrt{p(1-p)}}{p} \approx \frac{1}{\sqrt{p}}$
- When p is tiny, to control relative error (and needed n), we need variance reduction to reduce σ to $O(p)$ (as opposed to $O(\sqrt{p})$ in naïve MC)
- This requirement motivates rare-event simulation techniques, including IS, multi-level splitting / subset simulation (Au & Beck '01, Dean & Dupuis '09, Villen-Altamirano '94), cross-entropy methods (De Boer '05, Rubinstein & Kroese '13)…

Exponential Tilting

Can we efficiently estimate $P(X > 10)$ for $X \sim N(0,1)$ using IS?

Exponential tilting is a convenient framework to design IS: Assume P has light tail, i.e., the logarithmic moment generating function $\psi(\theta) = \log E\left[e^{\theta X}\right]$ exists for θ in a neighborhood of 0.

Consider an exponential family $d\tilde{P}_\theta(x) = e^{\theta x - \psi(\theta)} dP(x)$

Example:

- $P = N(\mu, \Sigma) \rightarrow \widetilde{P_\theta} = N(\tilde{\mu}, \Sigma)$
- $P = \text{Exp}(\lambda) \rightarrow \widetilde{P_\theta} = Exp(\tilde{\lambda})$

Find θ that gives a low variance via tail analysis. Can be generalized to other problems (more to come later)

Exponential Tilting

Known fact: $P(X > \gamma) \approx \frac{1}{\sqrt{2\pi}\gamma} e^{-\frac{\gamma^2}{2}}$ as $\gamma \to \infty$

Crude MC has $RE^2 = \frac{1}{p} \sim \gamma e^{\gamma^2/2}$, (more than) exponential in γ

Exponential Tilting

Consider an IS distribution $\tilde{P} = N(\gamma, 1)$

Likelihood ratio: $L = \frac{dP}{d\tilde{P}} = \frac{e^{-\frac{x^2}{2}}}{e^{-\frac{(x-\gamma)^2}{2}}} = e^{-x\gamma+\frac{\gamma^2}{2}}$

Relative error:

$$RE^2 = \frac{\widetilde{Var}(I(X > \gamma)L)}{p^2} = \frac{\left(\tilde{E}[I(X > \gamma)^2 L^2] - p^2\right)}{p^2} = \frac{\tilde{E}[I(X > \gamma)^2 L^2]}{p^2} - 1$$

where second moment of IS:

$$\tilde{E}[I(X > \gamma)^2 L^2] = \tilde{E}[L^2; X > \gamma] = \tilde{E}\left[e^{-2x\gamma+\gamma^2}; X > \gamma\right] = e^{-\gamma^2}\tilde{E}\left[e^{-2\gamma(x-\gamma)}; X > \gamma\right] \leq e^{-\gamma^2}$$

So

$RE^2 \leq O(\gamma^2) - 1$, polynomial in $\gamma \Rightarrow$ substantial improvement over CMC

Derivative Estimation

Goal: Estimate the derivative of $f(\theta)$, where $f(\theta)$ can only be observed with random noise

- $f(\theta) = E[h(\theta; X)]$, and we can generate unbiased estimate $h(\theta; X)$
- We denote $\hat{f}(\theta)$ an unbiased copy for $f(\theta)$

Motivation:

- Optimization: Stochastic gradient descent / stochastic approximation
- Sensitivity analysis: What if a (distributional or system) parameter perturbs?
- Uncertainty quantification: Constructing confidence intervals using the delta method

Variance reduction improves the efficiency of some challenging stochastic derivative estimators

Derivative Estimation: Example

Policy gradient in reinforcement learning

A Markov decision process with:

- State: s
- Action: a
- Transition kernel $P(s'|s,a)$ (simulatable from a model)
- Reward function: $r(\cdot)$
- Policy: Given state s, use action a with probability $p_\theta(a|s)$ ← can be a neural network, Gaussian mixture etc.

To maximize the cumulative reward $f(\theta) = E_\theta[\sum_{t=1}^{T} r(S_t, A_t)]$, we run gradient descent which requires estimating $f'(\theta)$

Zeroth vs First-Order Methods

- Efficiency of derivative estimation depends on our level of knowledge on $f(\theta)$:
  - Zeroth-order: Only noisy function evaluation is available
    - Finite-difference methods
  - First-order: Unbiased estimator for derivative is available
    - Infinitesimal perturbation analysis
    - Likelihood ratio / score function method
    - Measure-valued differentiation
    - Other variants, e.g., smoothed IPA…
- The bias in zeroth-order method pays a price on efficiency. If applicable, first-order methods are preferred
- If the gradient of a vector function is needed, the basic approach is to separately estimate the derivative for each direction (but things are more subtle when applying it in optimization)

Zeroth vs First-Order Methods

- Efficiency of derivative estimation depends on our level of knowledge on $f(\theta)$:
  - Zeroth-order: Only noisy function evaluation is available
    - Finite-difference methods
  - First-order: Unbiased estimator for derivative is available
    - Infinitesimal perturbation analysis
    - Likelihood ratio / score function method
    - Measure-valued differentiation
    - Other variants, e.g., smoothed IPA…
- The bias in zeroth-order method pays a price on efficiency. If applicable, first-order methods are preferred
- If the gradient of a vector function is needed, the basic approach is to separately estimate the derivative for each direction (but things are more subtle when applying it in optimization)

Finite Difference

Suppose $\hat{f}(\theta)$ is a "black-box", i.e., we have no access to what's inside f

Finite-difference is based on the first principle of differentiation:

$$\frac{\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)}{2\delta}$$

where $\hat{f}(\theta+\delta)$ and $\hat{f}(\theta-\delta)$ refer to two independent copies

When we have n simulation budget, we output

$$\sum_{i=1}^{n}\frac{\hat{f}_i(\theta+\delta)-\hat{f}_i(\theta-\delta)}{2\delta}$$

The above is called central finite difference. Can also define forward and backward finite difference (but generally less efficient)

Finite Difference

Bias:

$$E\left[\frac{\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)}{2\delta}\right]-f'(\theta)=\frac{f(\theta+\delta)-f(\theta-\delta)}{2\delta}-f'(\theta)=\frac{1}{3!}f'''(\theta)\delta^2+\cdots$$

Variance on n pairs:

$$\frac{1}{n}Var\left(\frac{\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)}{2\delta}\right)=\frac{\sigma^2(\theta+\delta)+\sigma^2(\theta-\delta)}{4n\delta^2}$$

$$\text{MSE} = Bias^2+Var=C_1\delta^4+\frac{C_2}{n\delta^2}$$

Optimal choice of perturbation size δ is tuned to balance squared bias and variance, to order $1/n^{\frac{1}{6}}$, giving root MSE of order $1/n^{\frac{1}{3}} \Rightarrow$ worse than canonical rate of $1/\sqrt{n}$

For forward / backward FD, the root MSE is even worse, of order $1/n^{\frac{1}{4}}$

Finite Difference with Common Random Numbers

- When estimating the mean difference of two systems, $X^1 - X^2$, we use the same stream of random numbers in the computer to drive the simulation of X^1 and X^2

- $Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2Cov(X_1, X_2)$, and the common random numbers usually make $Cov(X_1, X_2) > 0$, so that the output has less variance than crude MC

- Often used in the comparisons among multiple simulation-based alternatives, i.e., ranking and selection

Finite Difference with Common Random Numbers

When applying to finite difference, we simulate $\hat{f}(\theta+\delta)$ and $\hat{f}(\theta-\delta)$ using the same stream of random numbers

For typical discrete-event systems, $Var\left(\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)\right)=O(\delta)$

Bias:

$$E\left[\frac{\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)}{2\delta}\right]-f'(\theta)=\frac{f(\theta+\delta)-f(\theta-\delta)}{2\delta}-f'(\theta)=\frac{1}{3!}f'''(\theta)\delta^2+\cdots$$

Variance on n pairs:

$$\frac{1}{n}Var\left(\frac{\hat{f}(\theta+\delta)-\hat{f}(\theta-\delta)}{2\delta}\right)=\frac{O(\delta)}{4n\delta^2}=O\left(\frac{1}{4n\delta}\right)$$

So

$$\text{MSE} = Bias^2+Var=C_1\delta^4+\frac{C_2}{n\delta}$$

Optimal choice of perturbation size δ is of order $1/n^{\frac{1}{5}}$, giving root MSE of order $1/n^{\frac{2}{5}}$ ⇒ an improvement

Infinitesimal Perturbation Analysis / Pathwise differentiation

- Exchange the derivative with expectation operators

$$\frac{d}{d\theta} E[h(\theta; X)] = E\left[\frac{d}{d\theta} h(\theta; X)\right]$$

- If we can simulate $\frac{d}{d\theta} h(\theta; X)$ directly, then we can output

$$\frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\theta} h(\theta; X_i)$$

as an unbiased estimator for $f'(\theta) = \frac{d}{d\theta} E[h(\theta; X)]$

Infinitesimal Perturbation Analysis / Pathwise differentiation

If $h(\theta; X)$ is almost surely differentiable at θ and Lipschitz continuous

$$|h(\theta_1; X) - h(\theta_2; X)| \leq M|\theta_1 - \theta_2|$$

for θ_1, θ_2 in a neighborhood of θ, where $E[M] < \infty$, then IPA is valid

Example: $h(\theta; \boldsymbol{X}) = \max\{X_1, \theta X_2\}$

Example: $h(\theta; \boldsymbol{X}) = I(\theta X_2 > X_1)$

Likelihood Ratio / Score Function Method

Suppose the parameter θ is in the probability distribution P_θ that generates X

We write $f(\theta) = E_\theta[h(X)]$

Then

$$\frac{d}{d\theta} E_\theta[h(X)] = E_\theta[h(X) S_\theta(X)]$$

where $S_\theta(X)$ is the score function

$$S_\theta(x) = \frac{d}{d\theta} \log f_\theta(x) = \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}$$

Likelihood Ratio / Score Function Method

Reasoning:

$$\frac{d}{d\theta} E_\theta[h(X)] = \frac{d}{d\theta} \int h(x) f_\theta(x) dx = \int h(x) \frac{d}{d\theta} f_\theta(x) dx$$

$$= \int h(x) \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int h(x) S_\theta(x) f_\theta(x) dx = E_\theta[h(X) S_\theta(X)]$$

Alternately via IS:

$$\frac{d}{d\theta} E_\theta[h(X)] = \frac{d}{d\tilde{\theta}} \int h(x) f_{\tilde{\theta}}(x) dx \Big|_{\tilde{\theta}=\theta} = \frac{d}{d\tilde{\theta}} \int h(x) \frac{f_{\tilde{\theta}}(x)}{f_\theta(x)} f_\theta(x) dx \Big|_{\tilde{\theta}=\theta}$$

$$= \int h(x) \frac{\frac{d}{d\tilde{\theta}} f_{\tilde{\theta}}(x)|_{\tilde{\theta}=\theta}}{f_\theta(x)} f_\theta(x) dx = E_\theta[h(X) S_\theta(X)]$$

Likelihood Ratio / Score Function Method

Suppose $f_\theta(x)$ is continuously differentiable in θ almost everywhere in x, $h(x) \in L_q$, $\left|\frac{d}{d\tilde{\theta}} f_{\tilde{\theta}}(x)\right| \leq M(x)$ where $M(x) \in L_p$, for $\tilde{\theta}$ in a neighborhood of θ and almost everywhere in x, with $\frac{1}{p} + \frac{1}{q} = 1$. Then LR/SF is valid

Comparisons

IPA:

- More structural assumptions on h
- Less variance

LR/SF:

- Minimal structural assumptions on h, but some conditions on f_θ
- High variance especially if the time horizon is long ("curse of horizon")

LR/SF is more flexible but has (potentially much) higher variance. IPA is usually viewed as superior if it can be implemented

Curse of Horizon

The score function of multiple (independent or Markovian) variables is the sum of individual score functions

⇒ LR/SF method involves multiplying with a long summation if the time horizon is long ⇒ high variance

Example:

$f(\theta) = E_\theta[\sum_{t=1}^{T} c(S_t, A_t)]$, with transition probability $P(s'|s,a)$ and policy parametrized by $P_\theta(a|s)$

An LR/SF estimate of $f'(\theta)$ (using one simulated copy) is $\sum_{t=1}^{T} c(S_t, A_t) \sum_{t=1}^{T} S_\theta(A_t|S_t)$, where $S_\theta(a|s) = \frac{d}{d\theta} \log P_\theta(a|s)$

Reminiscent of a “blow-up” issue in IS when applied poorly (more later)

Reparametrization Trick

Consider taking the gradient w.r.t. θ

$$\frac{d}{d\theta} E_\theta[h(X)]$$

where LR/SF could have a large variance. Can we possibly bypass this?

Reparametrization trick (Rezende & Mohamed '15), equivalent to the "push-out" method (Rubinstein '92):

Suppose we can generate $X \sim P_\theta$ by a "flow" f_θ that maps a simple variable, e.g., a normal variable ϵ, to the complicated variable X

Then $E_\theta[h(X)] = E_\epsilon\left[h\left(f_\theta(\epsilon)\right)\right]$

and we can use IPA as long as $h \circ f_\theta$ satisfies the needed smoothness conditions

Reparametrization Trick

Consider taking the gradient w.r.t. θ

$$\frac{d}{d\theta} E_\theta[h(X)]$$

where LR/SF could have a large variance. Can we possibly bypass this?

Reparametrization trick (Rezende & Mohamed '15), equivalent to the "push-out" method (Rubinstein '92):

Suppose we can generate $X \sim P_\theta$ by a "flow" f_θ that maps a simple variable, e.g., a normal variable ϵ, to the complicated variable X

Then $E_\theta[h(X)] = E_\epsilon\left[h\left(f_\theta(\epsilon)\right)\right]$

and we can use IPA as long as $h \circ f_\theta$ satisfies the needed smoothness conditions

lower variance

may not be differentiable, esp. in black-box setups

A Control Variate Remedy

Train a differentiable approximating model $\hat{h}$ for h.
The estimator (Grathwohl et al. '18)

$$h(X)S_\theta(X) - \hat{h}(X)S_\theta(X) + \frac{d}{d\theta}\hat{h}\left(f_\theta(\epsilon)\right)$$

is unbiased for

$$\frac{d}{d\theta}E_\theta[h(X)]$$

and has potentially much lower variance than $h(X)S_\theta(X)$ when $\hat{h}$ is a good approximation

$\hat{h}(X)S_\theta(X)$ acts as a CV for $h(X)S_\theta(X)$ with an accurate mean estimate $\frac{d}{d\theta}\hat{h}\left(f_\theta(\epsilon)\right)$

Stochastic Variance Reduced Gradient Descent

To solve an empirical minimization problem $\min_{\theta} \hat{f}(\theta)$, where $\hat{f}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(\cdot)$, $i = 1, \ldots, n$, and $\hat{f}_i(\cdot)$ is a noisily observed function via one sample that is differentiable in θ (e.g., loss minimization for a statistical model)

Gradient descent: $\theta_{t+1} = \theta_t - \eta_t \nabla \hat{f}(\theta_t)$ where $\nabla \hat{f}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \nabla \widehat{f}_i(\cdot)$, η_t = step size

- Each gradient in the iteration is computed directly from differentiating $\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_i(\cdot)$

Stochastic gradient descent: $\theta_{t+1} = \theta_t - \eta_t \nabla \widehat{f}_{i_t}(\theta_t)$ where i_t = randomly selected index

- Each gradient in the iteration is computed using the gradient of one sampled $\widehat{f}_i(\cdot)$

Stochastic Variance Reduced Gradient Descent

SVRG (Johnson & Tong '13):

$$\theta_{t+1} = \theta_t - \eta_t \left(\nabla \hat{f}_{i_t}(\theta_t) - \nabla \hat{f}_{i_t}(\tilde{\theta}) + \nabla \hat{f}(\tilde{\theta}) \right)$$

where $\tilde{\theta}$ = sufficiently optimal solution

$\nabla \hat{f}_{i_t}(\tilde{\theta})$ acts as a CV for $\nabla \hat{f}_{i_t}(\theta_t)$ to decrease variance in gradient estimation
⇒ allow use of constant step size without sacrificing high variability

Super-Canonical Convergence

- Canonical MC rate (measured by root MSE) = $O(1/\sqrt{n})$
- Can we achieve faster rate $o(1/\sqrt{n})$ by variance reduction?

"Super"-effective control functionals (Oates et al., 2017): To estimate $E[h(X)]$, find $s(X)$ such that the CV estimator

$$h(X) - s(X) + E[s(X)]$$

has extremely low variance, by approximating h via a dense class of functional approximation s whose mean $E[s(X)]$ is known

Super-Canonical Convergence

How can such s be constructed?

Synthesis of two ideas:

Reproducing Kernel Hilbert Space (RKHS): An RKHS H is a Hilbert space of real-valued functions h: $\Omega \to R$ with an inner product $<\cdot,\cdot>$ and there exists a positive definite symmetric function k: $\Omega \times \Omega \to R$ such that

- $k(\cdot, x) \in H, \forall x \in \Omega$;
- $h(x) = < h(\cdot), k(\cdot, x) >, \forall x \in \Omega, \forall h \in H$.

k: the reproducing kernel of this Hilbert space.

Stein's identity: Assume the distribution P of X is analytically known, then for any $g(x)$ in the Stein class of P,

$$E_P[\mathcal{A}_P g(X)] = 0$$

where $\mathcal{A}_P g(x) = \nabla \log p\,(x) g(x) + \nabla g(x)$ is called Stein's operator.

- The class $\mathcal{A}_P k(x,\cdot)$ forms another RKHS H_0 where $E_P[\mathcal{A}_P k(x, X)] = 0$ $\Rightarrow$ A rich class of CV

Super-Canonical Convergence

- Using simulation data $\big(x_i, h(x_i)\big)$, run kernel ridge regression on the space $H_+ \coloneqq C + H_0$:

$$s_m(x) \coloneqq \ argmin_{g \in H_+} \left\{ \frac{1}{m} \sum_{i=1}^{m} \big(h(x_i) - g(x_i)\big)^2 \ + \lambda \|g\|_{H_+}^2 \right\}$$

- Apply CV $s_m(\cdot)$ in a "test set" to get the estimator

$$\hat{\theta}_{CF} = \frac{1}{n-m} \sum_{j=m+1}^{n} \left\{ h(x_j) - s_m\big(x_j\big) + E_P\left[s_m\big(x_j\big) \right] \right\}$$

- Super-canonical convergence: Suitably choosing regularization parameter λ and allocating "training" and "testing" set size,

$$\text{root MSE} = O\big(n^{-\frac{3}{4}}\big)$$

- Can use "Stein-kernelized" IS as well (Liu & Lee '17)
- Can use other functional approximations as well (Portier & Segers '19, Henderson & Glynn '02, Maire '03)

Bias Reduction

Rather than variance reduction, we sometimes face the need of bias reduction. Examples:

- When generating X from the original distribution is prohibited. E.g.,
  - Off-policy evaluation: An alternate policy employed in experiments to inform the performance of a target policy
  - Transfer learning / covariate shift: The test data set is generated from a different distribution from the training set, and only the training set has "label"
  - Bayesian statistics: Intractable posterior distribution (and known only up to normalizing constant)
- When the target quantity is not a simple "expectation". E.g., zeroth-order derivative estimation, and more

IS for Bias Reduction

IS is also a bias reduction tool:

- Suppose can only generate $X \sim Q$, different from P in the target $E_P[h(X)]$
- We can use $\sum_i h(X_i) \frac{dP}{dQ}(X_i)$ as an unbiased estimator
- More generally, even if we only know P up to a normalizing constant (common for Bayesian posterior), call this f, we can use the self-normalized estimator

$$\frac{\sum_i h(X_i) \frac{f}{q}(X_i)}{\sum_i \frac{f}{q}(X_i)}$$

as a nearly unbiased estimator

Simultaneous Bias and Variance Reduction

- If Q is very different from P in the target $E_P[h(X)]$, the likelihood ratio $\frac{dP}{dQ}$ could be huge (i.e., sample from P concentrates in region untouched by Q)

- The IS estimator $\frac{1}{n}\sum_i h(X_i)\frac{dP}{dQ}(X_i)$, although unbiased, has high variance (recall the "double-edged sword" comment earlier – though not exactly meant there)

- If we can train a model $\hat{h}$ to approximate h such that $E_P[\hat{h}(X)]$ is (approximately) known, we can combine IS and CV to obtain the doubly robust estimator:

$$\frac{1}{n}\sum_i \left(h(X_i) - \hat{h}(X_i)\right)\frac{dP}{dQ}(X_i) + E_P[\hat{h}(X)]$$

For some of the mentioned problems, it is possible to get such an $\hat{h}$

Bias Reduction

When the target quantity is not a simple “expectation”. E.g.,

- Zeroth-order derivative estimation: Bias caused by using finite difference in approximating derivative (we’ve seen)
- Steady-state estimation of a stochastic process: If we stop simulating the process at a finite time, there would be an “initial transient” bias
- Function-of-expectation estimation: To estimate $f(E[h(X)])$, if we use $f\left(\hat{E}[h(X)]\right)$, there would be bias. This includes settings in stochastic optimization and nested simulation etc.
- Discretizing continuous-time processes: Simulating a continuous-time process $X_\cdot$ at only the time-discretized values $X_{t_1}, \ldots, X_{t_m}$ would cause bias

Multilevel Monte Carlo

Biased simulation estimator $\hat{\mu}_k$ to estimate μ typically has a tuning parameter k that controls the tradeoffs among *bias*, *variance* or *computation load*, e.g.,

- Steady-state estimation: process run-length
- Function-of-expectation estimation: sample size for estimating the expectation
- Continuous-time processes: discretization scale

W.l.o.g., when $k = \infty$, $\hat{\mu}_\infty$ is unbiased

Multilevel Monte Carlo

MLMC is an approach to allocate different simulation run budget to different levels of k to maximize efficiency

Consider the telescoping sum:

$$E[\hat{\mu}_L] = E[\hat{\mu}_0] + \sum_{k=1}^{L} (E[\hat{\mu}_k] - E[\hat{\mu}_{k-1}])$$

so that

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \hat{\mu}_0^{(i)} + \sum_{k=1}^{L} \frac{1}{N_k} \sum_{i=1}^{N_k} (\hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)})$$

has the same bias as $\hat{\mu}_L$.

- The budget $N_k, k = 1, \dots, L$ is carefully allocated to simulate $\hat{\mu}_0^{(i)}$ and $\hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)}, k = 1, \dots, L$ (to achieve variance reduction)
- Efficiency gain is only possible if $\hat{\mu}_{k-1}^{(i)}$ and $\hat{\mu}_k^{(i)}$ can be simulated in a coupled manner

Multilevel Monte Carlo

Let us first look at the performance of a simple biased estimator $\hat{\mu}_k$

Suppose

- Bias = $b(k)$
- Variance = $\frac{\sigma_k^2}{N}$
- Computation cost per run = $c(k)$

To achieve a root MSE = $\sqrt{Var + Bias^2}$ within ϵ, we need (in terms of order):

- Bias = $b(k) \leq \epsilon \Rightarrow k$ needs to be $b^{-1}(\epsilon)$
- Var = $\frac{\sigma_k^2}{N} \approx \frac{\sigma^2}{N} \leq \epsilon^2 \Rightarrow N$ needs to be $\frac{\sigma^2}{\epsilon^2}$
- Total computation cost = (cost per run) × (sample size) = $c(k)N = c\left(b^{-1}(\epsilon)\right)\frac{\sigma^2}{\epsilon^2}$

Multilevel Monte Carlo

MLMC:

$$\sum_{k=0}^{L} \frac{1}{N_k} \sum_{i=1}^{N_k} (\hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)})$$

where for convenience, $\hat{\mu}_{-1}^{(i)} = 0$. How to allocate N_k?

Solve

$$\min_{N_k, 0 \le k \le L} \quad Total\ cost$$
$$\text{subject to} \quad \sqrt{Var + Bias^2} \le \epsilon$$

- Bias = $b(L) \le \epsilon \Rightarrow L$ needs to be $b^{-1}(\epsilon)$
- Var = $\sum_{k=0}^{L} \frac{\tilde{\sigma}_k^2}{N_k}$ where $\tilde{\sigma}_k^2$ is the variance of $\hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)}$ under coupling
- Total computation cost = $\sum_{k=0}^{L} c(k) N_k$

Multilevel Monte Carlo

To obtain the optimal allocation, solve the Lagrangian

$$\sum_{k=0}^{L} c(k)N_k + \lambda\left(\sum_{k=0}^{L}\frac{\tilde{\sigma}_k^2}{N_k} - O(\epsilon)\right)$$

which gives $c(k) - \frac{\lambda\tilde{\sigma}_k{}^2}{N_k^2} = 0$ or $N_k = \frac{\sqrt{\lambda}\tilde{\sigma}_k}{\sqrt{c(k)}}$. Plugging back, we get

- Var = $\frac{1}{\sqrt{\lambda}}\sum_{k=0}^{L}\tilde{\sigma}_k\sqrt{c(k)}$ which needs to be = ϵ^2 $\Rightarrow \sqrt{\lambda} = \frac{\sum_{k=0}^{L}\tilde{\sigma}_k\sqrt{c(k)}}{\epsilon^2}$
- Total computation cost $= \sqrt{\lambda}\sum_{k=0}^{L}\tilde{\sigma}_k\sqrt{c(k)} = \frac{\left(\sum_{k=0}^{L}\tilde{\sigma}_k\sqrt{c(k)}\right)^2}{\epsilon^2} = \frac{\left(\sum_{k=0}^{b^{-1}(\epsilon)}\tilde{\sigma}_k\sqrt{c(k)}\right)^2}{\epsilon^2}$

Multilevel Monte Carlo

Discretizing stochastic differential equation: To simulate $E[f(X_T)]$ where X_T is the solution of an SDE, we use $\frac{1}{N}\sum_{i=1}^{N} f\left(\hat{X}_{T/k}\right)$ where $\hat{X}_{T/k}$ is the solution obtained from an Euler discretization with time increment T/k.

Simple biased estimator: $b(k) = 1/k$, $c(k) = k$ under regularity conditions

Total computation cost = $c\left(b^{-1}(\epsilon)\right)\frac{\sigma^2}{\epsilon^2} = O\left(\frac{1}{\epsilon^3}\right)$

MLMC: We use a coupling of $f\left(\hat{X}_{T/M^k}\right)$ and $f\left(\hat{X}_{T/M^{k-1}}\right)$ where k denotes the grid scale. Then $b(k) = M^{-k}$, $c(k) = M^k$, $\tilde{\sigma}_k^2 = M^{-k}$

Total computation cost $= \frac{\left(\sum_{k=0}^{b^{-1}(\epsilon)} \tilde{\sigma}_k\sqrt{c(k)}\right)^2}{\epsilon^2} = O\left(\frac{1}{\epsilon^2}(\log\epsilon)^2\right)$

Multilevel Monte Carlo

Steady-state estimation of Markov chain: To estimate $E[f(X_\infty)]$ for a Markov chain $X_t, t = 1,2, \ldots$, we simulate up to X_k and output $\frac{1}{N}\sum_{i=1}^{k} f\left(X_k^{(i)}\right)$.

Simple biased estimator: $b(k) = \beta^k$ for $\beta < 1$, $c(k) = k$ under regularity conditions

Total computation cost = $c\left(b^{-1}(\epsilon)\right)\frac{\sigma^2}{\epsilon^2} = O\left(\frac{1}{\epsilon^2}\log\left(\frac{1}{\epsilon}\right)\right)$

MLMC:

No improvement beyond logarithmic: ready to check because $\frac{1}{\epsilon^2}$ is an intrinsic cost even with unbiased estimator. However, MLMC can still provide improvement in the following sense

Multilevel Monte Carlo

MLMC can turn a biased estimator into an unbiased estimator, by using a randomized level

This facilitates statistical inference of the simulation output, e.g., if we can estimate $Ef(X_\infty)$ by using unbiased $Y_1, \dots, Y_n$, then we can use standard normality tool for constructing confidence interval for $Ef(X_\infty)$ without worrying about the bias caused by the run-length per run

Consider

$$E[\hat{\mu}_\infty] = \sum_{k=0}^{\infty} (E[\hat{\mu}_k] - E[\hat{\mu}_{k-1}]) = E\left[\sum_{k=0}^{\infty} \frac{(\hat{\mu}_k - \hat{\mu}_{k-1})I(N \geq k)}{P(N \geq k)}\right] = E\left[\sum_{k=0}^{N} \frac{\hat{\mu}_k - \hat{\mu}_{k-1}}{P(N \geq k)}\right]$$

where N is a random level independent of $\hat{\mu}_k$'s

$\sum_{k=0}^{N} \frac{\hat{\mu}_k - \hat{\mu}_{k-1}}{P(N \geq k)}$

- is an unbiased estimator of $\mu = E[\hat{\mu}_\infty]$ (under technical condition that $E[\sum_{k=0}^{\infty} |\hat{\mu}_k - \hat{\mu}_{k-1}|] < \infty$)
- has finite variance under suitable choice of N (e.g., if N is power-law decay in the Markov chain example)

Multilevel Monte Carlo

MLMC is actively studied in applying to stochastic gradient descent, MCMC..

The unbiased estimation made possible by MLMC is an example of exact estimation, which belongs to a wider umbrella of study on unbiased estimators for stochastic processes that include:

Perfect sampling / Coupling from the past

Regenerative simulation

Rare-Event Simulation

Goals:

- Estimate $p = P(h(\boldsymbol{X}) \in A)$ for some rare-event set A, and $h(\boldsymbol{X})$ is the output of a stochastic model
- Understand how the rare event arises (i.e., the "most likely path to catastrophe")

Applications: risk analysis in industrial processes, operations, insurance, finance…

- $h(\boldsymbol{X})$ can be a portfolio return, workload of an operations system..
- A can denote the extreme set $\{y: y \geq \gamma\}$ for an exceedance threshold γ

Compared to extreme value theory (using data directly), rare-event simulation can be viewed as a model-based approach to quantify extreme risk

Rare-Event Simulation

Recall:

- We want an estimate $\hat{p}$ to be close to p relative to the magnitude of p
- By Markov inequality, the needed sample size n to achieve a relative discrepancy of ϵ with confidence $1-\alpha$ is $\geq \frac{\sigma^2}{\alpha\epsilon^2 p^2}$
- Crude MC has relative error (RE) = $O\left(\frac{1}{\sqrt{p}}\right)$
- If RE = $\frac{\sigma}{p}$ grows slowly in p, the algorithm is efficient

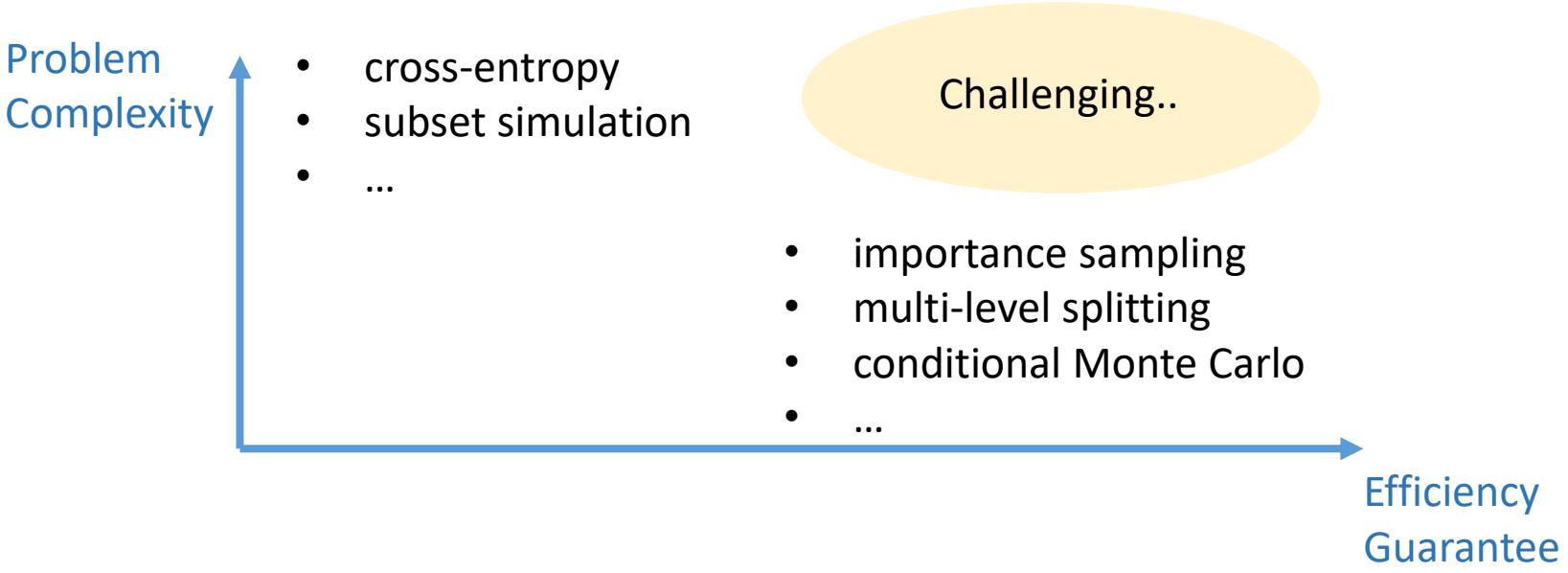
Rare-Event Simulation

Recall:

- We want an estimate $\hat{p}$ to be close to p relative to the magnitude of p
- By Markov inequality, the needed sample size n to achieve a relative discrepancy of ϵ with confidence $1-\alpha$ is $\geq \frac{\sigma^2}{\alpha\epsilon^2 p^2}$
- Crude MC has relative error (RE) = $O\left(\frac{1}{\sqrt{p}}\right)$
- If RE = $\frac{\sigma}{p}$ grows slowly in p, the algorithm is efficient
  - This view is taken assuming we use i.i.d. copies and take average to estimate p (e.g., IS). We can have a more general estimation procedure (e.g., cross-entropy, multi-level splitting)
  - How is "slowly" quantified?

Rare-Event Simulation

To obtain efficiency guarantees from variance reduction, we need to leverage structural knowledge/analysis (“No free lunch”)



Efficiency Notions

Introduce a rarity parameter, say γ, that models the level of rarity of p_γ such that as $\gamma \to \infty$, $p_\gamma \to 0$

E.g., γ can be the exceedance threshold in $P(h(\boldsymbol{X}) \geq \gamma)$

Recall IS:

- Call P the original distribution generating $\boldsymbol{X}$, and $\tilde{P}$ the IS distribution and $\tilde{E}[\cdot]$ the corresponding expectation
- Call $Z_\gamma = h(\boldsymbol{X})L(\boldsymbol{X})$ the output of IS

A popular efficiency notion is asymptotic optimality / logarithmic efficiency:

$$\lim_{\gamma \to \infty} \frac{\log \tilde{E}\left[Z_\gamma^2\right]}{\log \tilde{E}\left[Z_\gamma\right]} = 2$$

Efficiency Notions

$$\lim_{\gamma\to\infty}\frac{\log\tilde{E}\left[Z_\gamma^2\right]}{\log\tilde{E}\left[Z_\gamma\right]}=2$$

- This criterion means the exponential decay rates of $\tilde{E}\left[Z_\gamma^2\right]$ and $\tilde{E}\left[Z_\gamma\right]^2$ in terms of γ are the same
- Suppose we have a large deviations asymptotic $\tilde{E}\left[Z_\gamma\right]=p_\gamma\sim poly(\gamma)e^{-\gamma I}$ for a decay rate I. Then a polynomial-growth RE is a sufficient condition
- Note: $\tilde{E}\left[Z_\gamma^2\right]\geq\tilde{E}\left[Z_\gamma\right]^2\Rightarrow\frac{\log\tilde{E}\left[Z_\gamma^2\right]}{\log\tilde{E}\left[Z_\gamma\right]}\geq 2\Rightarrow$ The above holds if

$$\limsup_{\gamma\to\infty}\frac{\log\tilde{E}\left[Z_\gamma^2\right]}{\log\tilde{E}\left[Z_\gamma\right]}\leq 2$$

which is usually the focus of analysis

- Asymptotic optimality is easier to satisfy for many problems than bounded RE: $\limsup_{\gamma\to\infty}\frac{\tilde{E}\left[Z_\gamma^2\right]}{\tilde{E}\left[Z_\gamma\right]^2}<\infty$

Efficiency Notions

How do we:

- come up with a good IS and
- show it is efficient?

In the light-tailed case,

- Large deviations (LD) asymptotic for p_γ means it decays exponentially in the rarity parameter γ
- We trace the LD rate function and use the associated exponential tilting to construct IS
- IS can be viewed as a strengthened estimate (more accurate) of LD (more crude), if we know the LD (i.e., some level of analytical tractability)
- There are complications however

Let's consider some simple examples to illustrate the principle, then discuss more generally

Importance Sampling on Processes

We can keep the dependence property of a stochastic process in the IS

e.g., if we estimate $E[h(X_1, \dots, X_n)]$ where $X_i \sim P$ i.i.d., we can construct IS that keeps the i.i.d. structure but each tilting to $\tilde{P}$, with likelihood ratio

$$L(X_1, \dots, X_n) = \prod_{i=1}^{n} \frac{dP}{d\tilde{P}}(X_i)$$

Similarly, if we estimate $E[h(X_1, \dots, X_n)]$ where $X_i, i = 1,2\dots$ is a Markov chain with initial distribution $P(x)$ and transition $P(x'|x)$, we can construct IS that keeps the Markov structure but tilting to $\tilde{P}$, with likelihood ratio

$$L(X_1, \dots, X_n) = \frac{P(X_1)}{\tilde{P}(X_2)} \prod_{i=2}^{n} \frac{P(X_i|X_{i-1})}{\tilde{P}(X_i|X_{i-1})}$$

Importance Sampling and Large Deviations

Consider efficiently estimating $p_n = P(X_1 + \cdots + X_n > na)$ where X_i are i.i.d.

If $a > E[X_i]$, then by LLN $p_n \to 0$ as $n \to \infty$. How fast?

If X_i is light-tailed, then it satisfies a large deviations asymptotic

$$\lim_{n\to\infty} \frac{1}{n} \log p_n = -I(a)$$

where $I(a) = \sup\{\theta a - \psi(\theta)\}$ is the rate function given by the Legendre transform of $\psi(\theta) = \log E\left[e^{\theta X}\right]$, the logarithmic moment generating function of X_i

In other words, p_n decays exponentially in n with rate $I(a)$

Importance Sampling and Large Deviations

Upper bound:

- Chernoff's inequality implies

$$P(X_1 + \cdots + X_n > na) \leq e^{-na\theta + n\psi(\theta)}$$

for any $\theta \geq 0$

- Choose the best θ

$$P(X_1 + \cdots + X_n > na) \leq e^{-n \sup_{\theta \geq 0}\{a\theta - \psi(\theta)\}} = e^{-nI(a)}$$

Importance Sampling and Large Deviations

Lower bound:

- Let θ^* be the solution of the optimization that gives $I(a) \Rightarrow$ root of $\psi'(\theta^*) = a$
- Consider a change of measure that exponentially tilts dP to $d\tilde{P} = e^{\theta^* x - \psi(\theta^*)} dP$
- Letting $S_n = X_1 + \cdots + X_n$,

$$P(S_n > na) = \tilde{E}\left[e^{-\theta^* S_n + n\psi(\theta^*)}; S_n > na\right] = e^{-nI(a)}\tilde{E}\left[e^{-\theta^*(S_n - na)}; S_n > na\right]$$
$$\geq e^{-nI(a)}\tilde{E}\left[e^{-\theta^* n\bar{Y}}; 0 < \bar{Y} < \epsilon\right] \geq e^{-n(I(a) + \theta^*\epsilon)}$$

for any $\epsilon > 0$, where $\bar{Y} = (S_n - na)/n$

Importance Sampling and Large Deviations

The lower bound proof hints an IS that tilts dP to $d\tilde{P} = e^{\theta^* x - \psi(\theta^*)} dP$ on each X_i, keeping the i.i.d. structure intact

This IS is indeed asymptotically optimal, since

$$\tilde{E}[L^2; S_n > na] = \tilde{E}\left[e^{-2\theta^* S_n + 2n\psi(\theta^*)}; S_n > na\right]$$
$$= e^{-2nI(a)} \tilde{E}\left[e^{-2\theta^*(S_n - na)}; S_n > na\right] \leq e^{-2nI(a)}$$

But why exactly does it work?

A Reasoning

The stepwise exponential tilting with parameter θ^*, leading to mean a, is approximately the conditional probability of each step given the rare event occurs (recall our guideline previously)

- The Markov chain generated by

$$P^*(X_{m+1} \in dy | S_m = s_m) = P(X_{m+1} \in dy) \frac{P(S_n > na | S_{m+1} = s_m + y)}{P(S_n > na | S_m = s_m)}$$

is a zero-variance measure that, when used in IS, unbiasedly estimates the rare-event problem

- For fixed m, as $n \to \infty$,

$$P^*(X_{m+1} \in dy | S_m = s_m) = P(X_{m+1} \in dy) \frac{P(S_{n-m-1} > na - s_m - y)}{P(S_{n-m} > na - s_m)} \to P(X_{m+1} \in dy) e^{\theta^* y - \psi(\theta^*)}$$

The zero-variance measure representation holds more generally for Markov chains (leading to the adaptive IS method)

Most Likely Path

The most likely sample path to achieve the rare event is the path when every step increments by a. All other paths have exponentially smaller likelihood

Consider a more general large deviations principle

$$\lim_{n\to\infty}\frac{1}{n}\log P\big((S_m)_{1\le m\le n}\in A\big)=\lim_{n\to\infty}\frac{1}{n}\log P\left(\left(\tilde{S}_t\right)_{0\le t\le 1}\in\tilde{A}\right)=-I\big(\tilde{A}\big)$$

where $\tilde{S}_t = S_{\lceil nt\rceil}$ is the continuous interpolation of S_m,

$$I\big(\tilde{A}\big)=\inf\left\{\int_{t=0}^{1}J\big(x'(t)\big)dt : x(0)=0, x(t)\in\tilde{A}\right\}$$

where is the sample-path rate function obtained from a variational problem, and $J(y)=\sup_{\theta}\{\theta y-\psi(\theta)\}$ is the instantaneous rate function

This problem has a linear solution $x^*(t)$.

By picking smaller $\tilde{A}$ and applying the same principle, we see other paths occurs with exponentially smaller likelihood conditional on the rare event

Simulating Random Walk First Passage

Consider estimating $p_b = P(\tau < \infty)$ where $\tau = \min\{n: S_n > b\}$, $S_n = X_1 + \cdots + X_n$, X_i i.i.d. with mean $E[X_i] < 0$, and $b > 0$

- A random walk with negative drift, starting from zero, may never hit a positive level $\Rightarrow p_b < 1$
- As $b \to \infty$, $p_b \to 0$
- It's a rare-event simulation problem and, when running crude MC, we may never know when to stop..

Simulating Random Walk First Passage

Suppose X_i is light-tailed. Use an exponentially tilted change of measure on each i.i.d. X_i with $\psi(\theta^*) = 0$

$$P(\tau < \infty) = \tilde{E}\left[e^{-\theta^* S_\tau + \tau\,\psi(\theta^*)}; \tau < \infty\right] = e^{-\theta^* b}\tilde{E}\left[e^{-\theta^*(S_\tau - b)}; \tau < \infty\right] = e^{-\theta^* b}\tilde{E}\left[e^{-\theta^* Y}\right] \sim e^{-\theta^* b}$$

where Y is the overshoot variable above threshold b. In the above, the individual likelihood ratios are multiplied till the stopping time τ (using its martingale property)

Under the change of measure, the random walk has a positive drift, so $P(\tau < \infty) = 1$

IS uses the same exponential tilting. This is asymptotically optimal and the simulation is guaranteed to stop

Simulating Random Walk First Passage

Suppose X_i is light-tailed. Use an exponentially tilted change of measure on each i.i.d. X_i with $\psi(\theta^*) = 0$

$$P(\tau < \infty) = \tilde{E}\left[e^{-\theta^* S_\tau + \tau\,\psi(\theta^*)}; \tau < \infty\right] = e^{-\theta^* b}\tilde{E}\left[e^{-\theta^*(S_\tau - b)}; \tau < \infty\right] = e^{-\theta^* b}\tilde{E}\left[e^{-\theta^* Y}\right] \sim e^{-\theta^* b}$$

where Y is the overshoot variable above threshold b. In the above, the individual likelihood ratios are multiplied till the stopping time τ (using its martingale property)

Under the change of measure, the random walk has a positive drift, so $P(\tau < \infty) = 1$

IS uses the same exponential tilting. This is asymptotically optimal and the simulation is guaranteed to stop

Simulating Random Walk First Passage

Why does $\psi(\theta^*) = 0$ arise?

The sample-path large deviations rate function is now

$$I = \inf\left\{\int_{t=0}^{\infty} J\big(x'(t)\big)dt : x(0) = 0, x(t) \geq 1\right\}$$

which gives linear solution $x^*(t)$ and θ^* is the solution to obtain $J(\cdot)$

Similar reasoning on IS and most likely path as before via conditional probability given rare event

Summary and Pitfall of Roadmap

So far we suggest:

- Analyze LD which hints an IS
- Show asymptotic optimality of IS
- Reason why it works by arguing its closeness to zero-variance conditional measure given rare event

However, an IS distribution close to the zero-variance measure does not guarantee it’s asymptotically optimal

Sometimes, the above roadmap fails

A General Framework

Estimate $P(X \in A)$ for light-tailed (multivariate) X with log. MGF $\psi(\theta)$

LD principle gives

$$P(X \in A) \approx e^{-I(A)}$$

where $I(A) = \inf_{a \in A} I(a)$, and $I(a) = \sup_{\theta} \{\theta' a - \psi(\theta)\}$

Suppose we solve the rate function and obtain a^*, with corresponding exponential tilting parameter θ^*

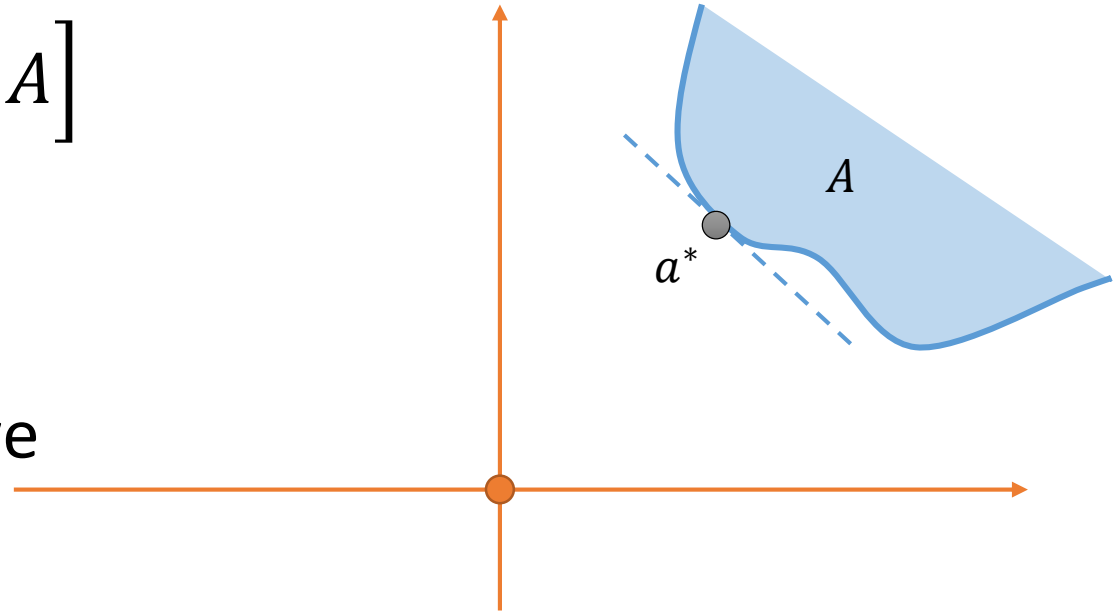
Consider IS that exponentially tilts P, giving

$$L = e^{-\theta^{*\prime} x + \psi(\theta^*)}$$

A General Framework

To analyze RE, consider

$$\tilde{E}[L^2; x \in A] = \tilde{E}\left[e^{-2\theta^{*\prime}x+2\psi(\theta^*)}; x \in A\right]$$
$$= e^{-2I(a^*)}\tilde{E}\left[e^{-2\theta^{*\prime}(x-a^*)}; x \in A\right]$$



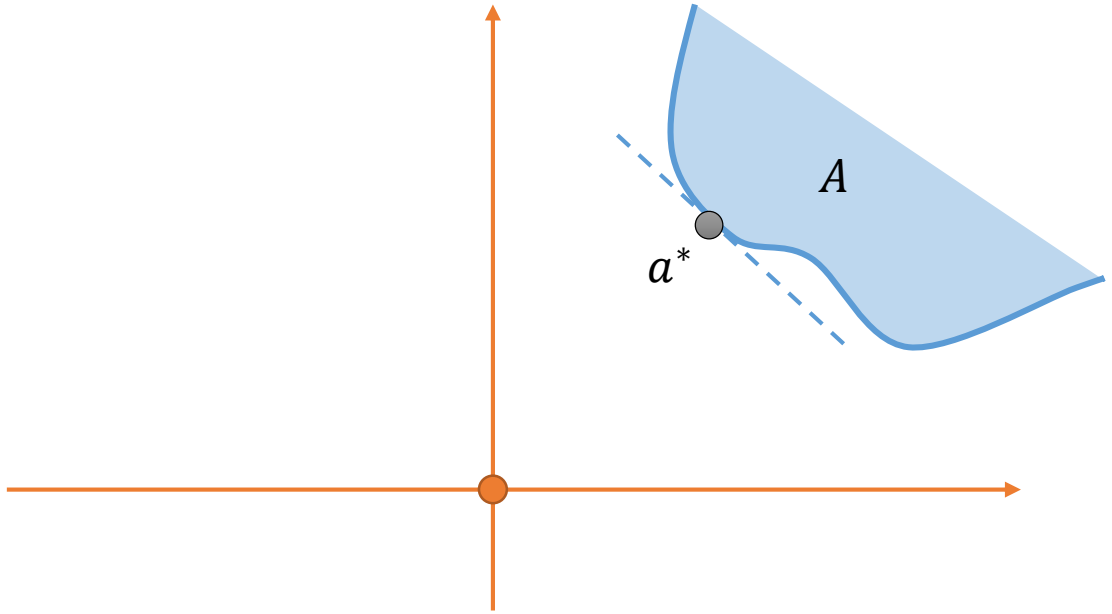
If $2\theta^{*\prime}(x - a^*) \geq 0$ for all $x \in A$, then we have asymptotic optimality

This is equivalent to saying the rare-event set is contained in the half-space cut by the tangent line touching a^* ($\theta^* = \nabla I(a^*)$)

Inefficiency of IS

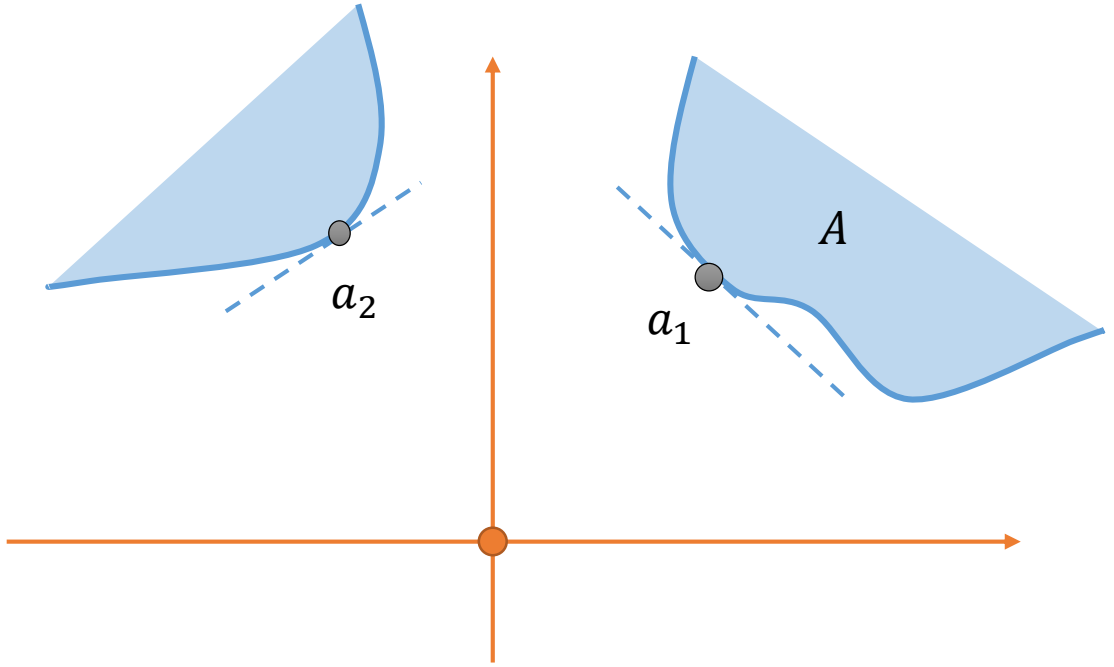
This is true if A is convex, because in this case $2\theta^{*\prime}(x - a^*) \geq 0$ is the first order optimality condition for $\min_{a \in A} I(a)$

However, the IS is not efficient if the condition is violated



Dominating Points

In general, we can divide $A = \bigcup_i A_i$ into individual "local" regions A_i where reach region has its own "dominating point" a_i, i.e., a point which $2\theta^{*\prime}(x - a_i) \geq 0$ for all $x \in A_i$



Dominating Points

For each region A_i, the exponential tilting to a_i, with likelihood ratio denoted L_i, is asymptotically optimal

We can use IS that consists of a mixture of individual exponential tilting, i.e.,

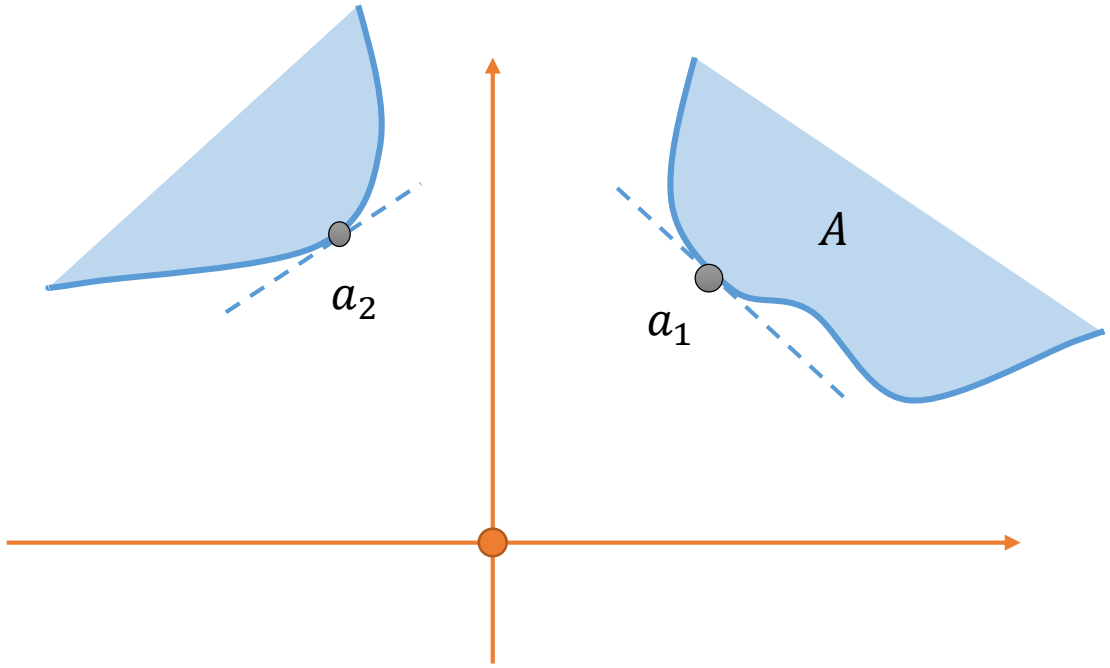
$$d\tilde{P} = \sum_i q_i dP_{a_i}$$

where q_i = mixing probability, dP_{a_i} = exponential tilting to point a_i

If $q_i > 0$ for all i, this scheme is asymptotically optimal because

$$\tilde{E}[L^2; X \in A] = \tilde{E}\left[\left(\frac{dP}{\sum_i q_i dP_{a_i}}\right)^2 ; X \in A\right]$$

$$\leq \sum_i \tilde{E}\left[\left(\frac{dP}{\sum_i q_i dP_{a_i}}\right)^2 ; X \in A_i\right] \leq \sum_i \frac{1}{q_i^2} \tilde{E}\left[L_i^2; X \in A_i\right]$$



Connecting Back..

How do dominating points and mixtures relate to the IS construction roadmap before?

In general, knowing the closeness to zero-variance measure and the most likely path (in the sense we discussed before) does not guarantee the stepwise exponentially tilted IS is efficient

Those IS are state-independent. To obtain efficient IS more generally, we use state-dependent schemes:

- Write the sample-path LD problem as a Hamilton-Jacobi-Bellman equation and formulate a state-dependent IS based on dynamic programming
- Sometimes this problem is difficult to solve, and subsolution suffices which can give rise to mixture-based IS

Conditional Monte Carlo

- To estimate $E[h(X)]$, we know $E[h(X)|Y]$ and can simulate Y
- We output $g(Y) = E[h(X)|Y]$ as one simulation run output
- With n simulation runs, we output

$$\frac{1}{n}\sum_{i=1}^{n} g(Y_i)$$

- Conditional Monte Carlo is unbiased and has variance no more than crude MC

$$Var(E[h(X)|Y]) \leq Var(h(X))$$

because of the law of total variance

$$Var\big(h(X)\big) = Var(E[h(X)|Y]) + E[Var(h(X)|Y)]$$

Conditional Monte Carlo

- Conditional Monte Carlo has most variance reduction when $h(X)$ and Y are least dependent; zero variance when Y is independent of $h(X)$ (contrast with control variate)

- Conditional Monte Carlo is known to provide dramatic variance reduction for heavy-tailed rare-event estimation problems

- Stratification uses the other term in the law of total variance

Cross-Entropy Method

- Automatic approach to search for optimal parameter θ over a class of IS P_θ (not necessarily exponential tilting)
- Idea: Minimize the "distance" between P_θ and P^*, the theoretically optimal IS
- We minimize the Kullback-Leibler divergence or the relative entropy

$$KL(P_\theta||P^*) = \int \log \frac{dP^*}{dP_\theta} dP^* = \int \log f^*(x)\ f^*(x)dx - \int \log f_\theta(x)\ f^*(x)dx$$

- It suffices to focus on maximizing

$$\int \log f_\theta(x)\ f^*(x)dx$$

- Plugging in $f^*(x) \propto I(x \in A)f(x)$, we get

$$\int \log f_\theta(x)\ f^*(x)dx = E[\log f_\theta(X); X \in A]$$

- That is, we maximize the expected log-likelihood within the set A

Cross-Entropy Method

- Implementing this idea needs a “warm-start”
- Consider estimating $P(X \in A_\gamma)$, where γ = rarity parameter

Iteratively run the following:

Given the current rarity level γ_t and IS $\tilde{f}_{\theta_t}$, simulate $X_1, \ldots, X_{n_t}$, and solve the sample average approximation

$$\max_\theta \frac{1}{n_t} \sum_{i=1}^{n_t} \log f_\theta(X_i)\, I\left(X_i \in A_{\gamma_t}\right) \frac{f(X_i)}{\tilde{f}_{\theta_t}(X_i)}$$

to obtain θ_{t+1}, and increase the rarity to γ_{t+1}

Cross-Entropy Method

- Other variants of this approach replace KL-divergence with the variance itself

- Other “automatic” methods for rare-event estimation include multi-level splitting

References

- Bucklew, J. (2004), *Introduction to Rare Event Simulation*, Springer

- Rubinstein, R. Y. & Kroese, D. P. (2004), *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, Springer

- Juneja, S. & Shahabuddin, P. (2006), Rare-event simulation techniques: An introduction and recent advances, *Handbooks in Operations Research and Management Science* 13:291-350

- Blanchet, J. & Lam, H. (2012), State-dependent importance sampling for rare-event simulation: An overview and recent advances", *Surveys in Operations Research and Management Science* 17(1):38-59