Some variance reduction methods

Felipe Uribe

Computational Engineering
School of Engineering Sciences
Lappeenranta-Lahti University of Technology (LUT)

Special Course on Inverse Problems
Lappeenranta, FI — January-February, 2024

Why variance reduction?

- We have seen that standard MC typically has an error variance of the form σ^2/n. We get a better answer with larger n, but the computing time grows with n.

- Sometimes we can find a way to reduce σ instead. We construct a new Monte Carlo problem with the same answer as our original one but with a lower $\sigma \Longrightarrow$ variance reduction techniques.

Why variance reduction?

- We have seen that standard MC typically has an error variance of the form σ^2/n. We get a better answer with larger n, but the computing time grows with n.

- Sometimes we can find a way to reduce σ instead. We construct a new Monte Carlo problem with the same answer as our original one but with a lower $\sigma \Longrightarrow$ variance reduction techniques.

- We can group the methods in the following categories:
  - Type-1 (**using clever samples**): antithetic sampling, stratification, and common random numbers.
  - Type-2 (**using things we know**): conditioning and control variates.
  - Type-3 (**using auxiliary densities**): importance sampling and its variants.

- These methods are also used in combination with MCMC.

This lecture...

- The lecture is based on multiple references. However, we mostly follow Chapters 8 and 9 of the book by **Art Owen**[1], which is freely available online.

[1] A. B. Owen. *Monte Carlo theory, methods and examples*. `artowen.su.domains/mc/`, 2018.

Variance reduction: type-1 methods (“using clever samples”)

Antithetic sampling: intro

- Random variables X, Y on the same probability space are antithetic, if they have the same distribution and their covariance is negative.
- When we are using Monte Carlo averages of quantities $f(\boldsymbol{x}_i)$ then the randomness in the algorithm leads to some error cancellation. In antithetic sampling, we try to get even more cancellation.
- An antithetic sample $\widetilde{\boldsymbol{x}}$ is one that gives the opposite value of $f(\boldsymbol{x})$, i.e., being low when $f(\boldsymbol{x})$ is high and vice versa. Ordinarily, we get an opposite f by sampling at a point $\widetilde{\boldsymbol{x}}$ that is *somehow* opposite to $\boldsymbol{x}$.
- Let $\mu = \mathbb{E}[\boldsymbol{X}]$ for $\boldsymbol{X} \sim \pi$, where π is a symmetric density on $\mathbb{R}^d$. Here, symmetry is with respect to reflection through the *center point* $\boldsymbol{c}$ of $\mathbb{R}^d$.

Antithetic sampling: estimator

- If we reflect $\boldsymbol{x}$ through $\boldsymbol{c}$, we have $\widetilde{\boldsymbol{x}} - \boldsymbol{c} = -(\boldsymbol{x} - \boldsymbol{c})$, and we get the point $\widetilde{\boldsymbol{x}} = 2\boldsymbol{c} - \boldsymbol{x}$. For basic examples, when $\pi = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ then $\widetilde{\boldsymbol{x}} = -\boldsymbol{x}$. When $\pi = \mathcal{U}(0,1)^d$, we have $\widetilde{\boldsymbol{x}} = 1 - \boldsymbol{x}$ (componentwise).

- The antithetic sampling estimate of μ is:

$$\mu \approx \hat{\mu}_{\text{anti}} = \frac{1}{n} \sum_{i=1}^{n/2} f(\boldsymbol{x}_i) + f(\widetilde{\boldsymbol{x}}_i), \tag{1}$$

  where $\boldsymbol{x}_i \overset{\text{idd}}{\sim} \pi$ and n is an even number. This estimator is also **unbiased**.

- The rationale for antithetic sampling is that each value of $\boldsymbol{x}$ is *balanced* by its opposite $\widetilde{\boldsymbol{x}}$, satisfying $(\boldsymbol{x} + \widetilde{\boldsymbol{x}})/2 = \boldsymbol{c}$.

Antithetic sampling

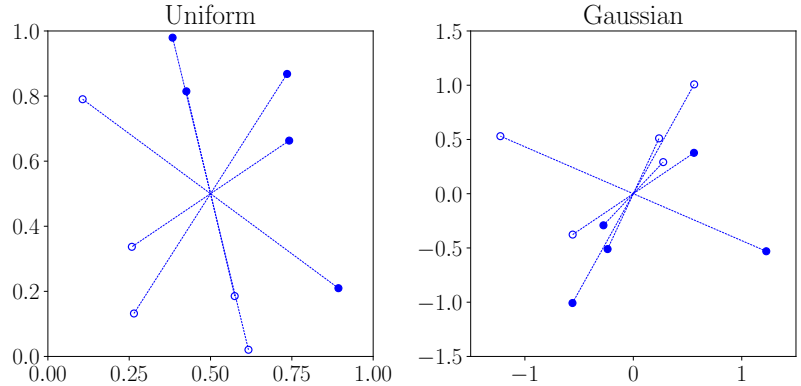

Figure: Five points and their antithetics. Left: from a standard uniform. Right: from a standard Gaussian.

Antithetic sampling: variance

- Whether the balance is helpful or not depends on f. If f is nearly linear, we could obtain a large improvement.
- The variance of antithetic sampling is:

$$\mathbb{V}[\hat{\mu}_{\text{anti}}] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n/2} f(\boldsymbol{x}_i) + f(\widetilde{\boldsymbol{x}}_i)\right] = \frac{n/2}{n^2}\mathbb{V}\left[f(\boldsymbol{X}) + f(\widetilde{\boldsymbol{X}})\right] \tag{2}$$

$$= \frac{1}{2n}\left(\mathbb{V}[f(\boldsymbol{X})] + \mathbb{V}\left[f(\widetilde{\boldsymbol{X}})\right] + 2\mathbb{C}\text{ov}\left[f(\boldsymbol{X}), f(\widetilde{\boldsymbol{X}})\right]\right) = \frac{\sigma^2}{n}(1+\rho) \tag{3}$$

- Since $-1 \leq \rho \leq 1$, we obtain $0 \leq \sigma^2(1+\rho) \leq 2\sigma^2$. In the best case, antithetic sampling gives the exact answer from just one pair of function evaluations. In the worst case, it doubles the variance.

Antithetic sampling: when it works?

- Hence, the variance of standard MC and antithetics can be written as:

$$\begin{bmatrix} \mathbb{V}[\hat{\mu}] \\ \mathbb{V}[\hat{\mu}_{\text{anti}}] \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \sigma_e^2 \\ \sigma_o^2 \end{bmatrix}; \tag{4}$$

  antithetic sampling eliminates the variance of f_o but doubles the contribution from f_e.

- **Tip**: antithetic sampling reduces the variance if $\rho < 0$ (e.g., monotone function), or equivalently if $\sigma_o^2 > \sigma_e^2$. This analysis is appropriate when the most of the computation is in evaluating f.

- Because antithetic samples have dependent values within pairs. We can define $y_i = f_e(\boldsymbol{x}_i) = (f(\boldsymbol{x}_i) + f(\widetilde{\boldsymbol{x}}_i))/2$, for $i = 1, \ldots, m = n/2$, then

$$\hat{\mu}_{\text{anti}} = \frac{1}{m} \sum_{i=1}^{m} y_i, \qquad \sigma^2_{\text{anti}} = \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \hat{\mu}_{\text{anti}})^2. \tag{5}$$

Antithetic sampling: example (I)

Consider the expected logarithmic return of a portfolio:

- There are K stocks and the portfolio has proportion $\lambda_k \geq 0$ in stock k, with $\sum_{k=1}^{K} \lambda_k = 1$.
- The expected logarithmic return is defined as

$$\mu(\lambda) = \mathbb{E}\left[\log\left(\sum_{k=1}^{K} \lambda_k \, \exp(X_k)\right)\right], \tag{6}$$

  where $\boldsymbol{X} \in \mathbb{R}^K$ is the vector of returns.

- If one keeps reinvesting/rebalancing the portfolio at N regular time intervals then, by the LLN, our fortune grows as $\exp(N\mu + \mathcal{O}(N))$, assuming of course that the $\boldsymbol{X}$ for each time period are idd.

Antithetic sampling: example (II)

- The log-optimal choice λ is the allocation that maximizes μ. Finding a model for the distribution of $\boldsymbol{X}$ and then choosing λ are challenging problems. We focus on the problem of evaluating $\mu(\lambda)$ for a given λ.
- We take $\lambda_k = 1/K$ with $K = 500$. We also suppose that each marginal distribution is $X_k \sim \mathcal{N}(\delta, \sigma^2)$ but that $\boldsymbol{X}$ has the $t(0, \nu, \boldsymbol{\Sigma})$ copula. Here $\delta = 0.001$ and $\sigma = 0.03$ ($\approx$ one week time frame). And $\nu = 4$ with covariance is $\boldsymbol{\Sigma} = \rho \mathbf{1}_K \mathbf{1}_K^\mathsf{T} + (1-\rho)\mathbf{I}_k^\mathsf{T}$ for $\rho = 0.3$.
- Letting $f(\boldsymbol{X}) = \log\left(\sum_{k=1}^K \exp(X_k)/K\right)$, the MC estimate is $\hat{\mu} = 1/n \sum_{i=1}^n f(\boldsymbol{X}_i)$.
- The antithetic to $\boldsymbol{X}_i$ has components $\widetilde{X}_{ik} = 2\delta - X_{ik}$.
- Continue on code...

Stratified sampling: intro

- The idea in stratified sampling is to split up the domain D of $\boldsymbol{X}$ into separate regions, take a sample of points from each region, and combine the results.
- We might do better by *oversampling* within the important strata and *undersampling* those in which f is nearly constant.
- To use stratified sampling, we must know the sizes $\omega_j = \mathbb{P}[\boldsymbol{X} \in D_j]$ of the strata, and we must also know how to sample $\boldsymbol{X} \sim \pi_j$ for $j = 1, \ldots, J$.
- When we are defining strata, we naturally prefer ones we can sample from. If however, we know ω_j but are unable to sample from π_j, then we use *post-stratification*.

Stratified sampling

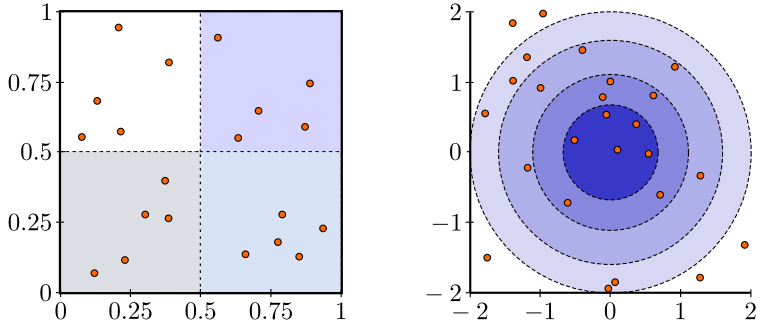

Figure: Left: 20 points in $[0, 1]^2$ of which 5 are sampled uniformly from within each quadrants ($J = 4$). Right: 25 points from a standard Gaussian. There are 4 concentric rings separating the distribution into $J = 5$ equally probable strata with 3 points sampled from each.

Stratified sampling: estimator

- Let $\boldsymbol{X}_{ij} \sim \pi_j$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ be sampled independently. The stratified sampling estimate is

$$\mu \approx \hat{\mu}_{\text{strat}} = \sum_{j=1}^{J} \frac{\omega_j}{n_j} \sum_{i=1}^{n_j} f(\boldsymbol{X}_{ij}); \tag{7}$$

  this estimator is also **unbiased**.

- As done previously for antithetic sampling, we now study the variance of the estimator $\hat{\mu}_{\text{strat}}$ to determine when stratification is advantageous, and to see how to design an effective stratification.

- We define $\mu_j = \mathbb{E}_{\pi_j}[f(\boldsymbol{x})]$ and $\sigma_j^2 = \mathbb{V}_{\pi_j}[f(\boldsymbol{x})]$ to be the jth stratum mean and variance, respectively.

Stratified sampling: variance

- The variance of the stratified sampling estimate is

$$\mathbb{V}[\hat{\mu}_{\text{strat}}] = \sum_{j=1}^{J} \omega_j^2 \frac{\sigma_j^2}{n_j}; \tag{8}$$

  an immediate consequence is that $\mathbb{V}[\hat{\mu}_{\text{strat}}] = 0$ for integrands f that are constant within strata D_j.

- The variance of $f(\boldsymbol{X})$ can be decomposed into within- and between-stratum components[2]

$$\sigma^2 = \mathbb{V}[f(\boldsymbol{X})] = \mathbb{E}[\mathbb{V}[f(\boldsymbol{X} \mid Z)]] + \mathbb{V}[\mathbb{E}[f(\boldsymbol{X} \mid Z)]], \quad Z = 1, \ldots, J \tag{9a}$$

$$= \sum_{j=1}^{J} \omega_j \sigma_j^2 + \sum_{j=1}^{J} \omega_j (\mu_j - \mu)^2 = \sigma_A^2 + \sigma_B^2; \tag{9b}$$

[2] See this Link to check this property of the variance.

Stratified sampling: post-stratification (proportional)

- Post-stratification: if we know ω_j but we cannot sample $X \sim \pi_j$. The idea is to sample $\boldsymbol{X}_i \sim \pi$ and assign it to their strata afterwards. The estimators remain the same.
- The main difference is that n_j are now random. A natural choice for stratum sample sizes is **proportional allocation**, $n_j = n\omega_j$. In this case, the estimators reduce to

$$\hat{\mu}_{\text{strat,p}} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} f(\boldsymbol{X}_{ij}) \qquad \sigma^2_{\text{strat,p}} = \frac{1}{n}\sum_{j=1}^{J}\omega_j\sigma_j^2. \tag{10}$$

- We can compare iid and proportional stratification in one equation

$$\begin{bmatrix} \mathbb{V}[\hat{\mu}] \\ \mathbb{V}[\hat{\mu}_{\text{strat,p}}] \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \sigma_A^2 \\ \sigma_B^2 \end{bmatrix}; \tag{11}$$

- **Tips**: a good stratification scheme is one that reduces the within-stratum variance σ_A^2, ideally $\sigma_B^2 \gg \sigma_A^2$. If sampling from π_j is slower than sampling from π, we loose any efficiency gain from stratification.

Stratified sampling: post-stratification (non-proportional)

- A proportional allocation is not necessarily the most efficient. Optimal sample allocation can be achieved using **Neyman allocation**, and the formulation allows for unequal sampling costs from the different strata.
- To minimize variance, we use

$$n_j \propto \frac{\omega_j \sigma_j}{\sqrt{c_j}}, \tag{12}$$

  where c_j is the (expected) cost to generate $\boldsymbol{X}$ from π_j and then compute $f(\boldsymbol{X})$.
- Non-proportional allocations carry some risk. The optimal allocation can be worse than the proportional allocation discussed before.
- There are also results on how to construct optional strata. In general, we want strata within which f is as flat as possible.

Stratified sampling: example (I)

Compound Poisson models (random process with jumps) are commonly used for rainfall:

- The number of rainfall events (storms) in the coming month is $S \sim \text{Poi}(\lambda)$ with $\lambda = 2.9$.
- The depth of rainfall in a storm s is $d_s \sim \text{Weib}(k, \sigma)$ with shape $k = 0.8$ and scale $\sigma = 3$ (cm) and the storms are independent.
- If the total rainfall is below 5 centimeters then an emergency water allocation will be imposed. The total rainfall is $X = \sum_{s=1}^{S} d_s$ taking the value 0 when $S = 0$.
- It is easy to get the mean and variance, but here we want $\mathbb{P}[\boldsymbol{X} < 5]$, that is $\mathbb{E}[f(\boldsymbol{X})]$ where $f(\boldsymbol{X}) = \mathbb{1}_{\boldsymbol{X}<5}$.
- Continue in code...

Stratified sampling: example (II)

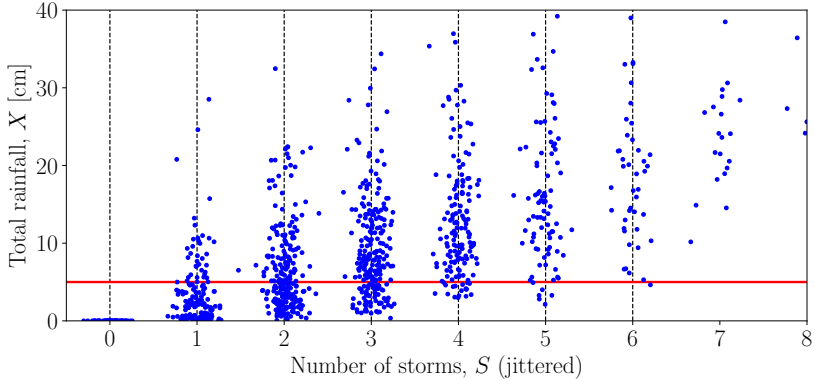

Figure: 1000 simulations of the compound Poisson model for rainfall. We define 7 strata.

Common random numbers: intro and estimator

- Suppose that f and g are closely related functions and that we want to find $\mathbb{E}[f(\boldsymbol{x}) - g(\boldsymbol{x})]$ for $\boldsymbol{x} \sim \pi$.
- Maybe $f(\boldsymbol{x}) = h(\boldsymbol{x}; \boldsymbol{\theta})$ with a parameter $\boldsymbol{\theta} \in \mathbb{R}^m$. To study its effect, we look at $g(\boldsymbol{x}) = h(\boldsymbol{x}; \widetilde{\boldsymbol{\theta}})$, for some $\widetilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}$.
- Because $\mathbb{E}[f(\boldsymbol{X}) - g(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X})] - \mathbb{E}[g(\boldsymbol{X})]$, we have two options:

$$\widehat{D}_{\text{com}} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i), \qquad \widehat{D}_{\text{ind}} = \frac{1}{n_1}\sum_{i=1}^{n_1} f(\boldsymbol{X}_{i1}) - \frac{1}{n_2}\sum_{i=1}^{n_2} g(\boldsymbol{X}_{i2}), \tag{13}$$

  where $\boldsymbol{X}_i \sim \pi$ (left: common random numbers (CRN)) and $\boldsymbol{X}_{ij} \sim \pi$ (right: independent random numbers).

Common random numbers: variance

- Taking $n = n_1 = n_2$, the sample variances are :

$$\mathbb{V}\left[\widehat{D}_{\text{com}}\right] = \frac{1}{n}\left(\sigma_f^2 + \sigma_g^2 - 2\rho\sigma_f\sigma_g\right), \qquad \mathbb{V}\left[\widehat{D}_{\text{ind}}\right] = \frac{1}{n}\left(\sigma_f^2 + \sigma_g^2\right). \tag{14}$$

- When $\rho > 0$, we are better off using common random numbers. Retaining some common random numbers requires considerable care in synchronization.

- The same problem arises if we are comparing $\mathbb{E}[f(\boldsymbol{X})]$ for $\boldsymbol{X} \sim \pi$ and $\mathbb{E}\left[f(\widetilde{\boldsymbol{X}})\right]$ for $\widetilde{\boldsymbol{X}} \sim \pi$.

- **Application:** CRN applies when we are comparing two or more alternative configurations (of a system) instead of investigating a single configuration.

Common random numbers: couplings

- **Example:** if a first simulation has $X_i \overset{\text{idd}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and a second has $\widetilde{X}_i \overset{\text{idd}}{\sim} \mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$, then we can sample $Z_i \overset{\text{idd}}{\sim} \mathcal{N}(0,1)$ and use

$$\widehat{D}_{\text{com}} = \frac{1}{n}\sum_{i=1}^{n} f(\mu + \sigma Z_i) - f(\widetilde{\mu} + \widetilde{\sigma} Z_i). \tag{15}$$

- More generally, when $\boldsymbol{X}_i$ is generated via a transformation $T(\boldsymbol{U}_i; \theta)$ of $\boldsymbol{U}_i \sim \mathcal{U}(0,1)^d$, then we can average $f(T(\boldsymbol{U}_i; \theta)) - f(T(\boldsymbol{U}_i; \widetilde{\theta}))$.

- The construction above is a **coupling** of the random vectors $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$. Any joint distribution on $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ with $\boldsymbol{X} \sim \pi$ and $\widetilde{\boldsymbol{X}} \sim \widetilde{\pi}$ is a coupling.

Common random numbers: implementation

- We want to estimate $\mu_j = \mathbb{E}[h(\boldsymbol{X};\theta_j))]$, for $j = 1,\ldots,m$ and using n random inputs $\{\boldsymbol{X}_i\}_{i=1}^n$. In the simplest case, $m = 2$ and we are interested in $\mu_1 - \mu_2$.
- We can run a nested loop over samples indexed by i and parameter values indexed by j. There are two main approaches that we can take, depending on which is the outer loop.
- CRN requires *synchronization* of the random number streams, which ensures that in addition to using the same random numbers to simulate all configurations, a specific random number used for a specific purpose in one configuration is used for exactly the same purpose in all other configurations.

Common random numbers: algorithms

Algorithm 1: Version 1: common random numbers

```
1 setseed(seed);
2 μ̂_j = 0, 1 ≤ j ≤ m;
3 for i = 1 to n do
4 |   X_i ~ π;
5 |   μ̂_j = μ̂_j + h(X_i; θ_j), 1 ≤ j ≤ m;
6 end
7 μ̂_j = μ̂_j/n, 1 ≤ j ≤ m;
```

Algorithm 2: Version 2: common random numbers

```
1 for j = 1 to m do
2 |   setseed(seed);
3 |   μ̂_j = 0;
4 |   for i = 1 to n do
5 |   |   X_i ~ π;
6 |   |   μ̂_j = μ̂_j + h(X_i; θ_j);
7 |   end
8 |   μ̂_j = μ̂_j/n;
9 end
```

Variance reduction: type-2 methods (“using things we know”)

Conditioning: intro and estimator

- Sometimes we can do part of the problem in closed form, and then do the rest of it by MC or some other numerical method.
- Assume that $\boldsymbol{X} \in \mathbb{R}^k$ and $\boldsymbol{Y} \in \mathbb{R}^{d-k}$ are random vectors and we want to estimate $\mathbb{E}[f(\boldsymbol{X}, \boldsymbol{Y})]$. The standard estimator is $\hat{\mu} = 1/n \sum_{i=1}^{n} f(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, where $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathbb{R}^d$ are independent samples from the joint distribution.
- Define $h(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}]$, then we can also estimate[3]:

$$\hat{\mu}_{\text{cond}} = \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{X}_i), \tag{16}$$

  where $\boldsymbol{X}_i$ are sampled independently from the distribution of $\boldsymbol{X}$. This method is called *conditioning* or **conditional Monte Carlo**.

[3] Note that $\mathbb{E}[f(\boldsymbol{X}, \boldsymbol{Y})] = \mathbb{E}[\mathbb{E}[f(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{X}]] = \mathbb{E}[h(\boldsymbol{X})]$.

Conditioning: variance

- The variance of the conditional MC estimator is:

$$\mathbb{V}[\hat{\mu}_{\text{cond}}] = \frac{1}{n}\mathbb{V}[h(\boldsymbol{X})] = \frac{1}{n}\mathbb{V}[\mathbb{E}[f(\boldsymbol{X},\boldsymbol{Y}) \mid \boldsymbol{X}]]\,. \tag{17}$$

- From the properties of the variance, we know:

$$\mathbb{V}[f(\boldsymbol{X},\boldsymbol{Y})] = \mathbb{E}[\mathbb{V}[f(\boldsymbol{X},\boldsymbol{Y}) \mid \boldsymbol{X}]] + \mathbb{V}[\mathbb{E}[f(\boldsymbol{X},\boldsymbol{Y}) \mid \boldsymbol{X}]]\,; \tag{18}$$

  hence, conditional Monte Carlo cannot have higher variance than crude MC sampling of f.

- Conditioning is a special case of de-randomization which is sometimes called **Rao–Blackwellization**.

- De-randomization by conditioning always reduces variance, it is not always worth doing. We could find our estimate is less efficient, if computing h costs much more than f.

Conditioning: example

- Let $C = \{(x, y) \mid a \leq x \leq b, 0 \leq y \leq f(x)\}$ and assume that $f(x) \leq c$ holds for $x \in [a, b]$. Then the MC estimate of the integral is

$$\widehat{\mathrm{vol}}(C) = \frac{c(b-a)}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \leq f(X_i)}, \quad (X_i, Y_i) \sim \mathcal{U}([a, b] \times [0, c]). \tag{19}$$

- The conditional expectation is:

$$h(x) = \mathbb{E}[f(X, Y) \mid X = x] = \frac{1}{\pi_X(x)} \int_{-\infty}^{\infty} \mathbb{1}_{Y \leq f(x)} \pi_{XY}(x, y) \, \mathrm{d}y = \frac{f(x)}{c}. \tag{20}$$

- Conditioning yields the estimate:

$$\widehat{\mathrm{vol}}(C) = \frac{c(b-a)}{n} \sum_{i=1}^{n} \frac{f(X_i)}{c} = \frac{(b-a)}{n} \sum_{i=1}^{n} f(X_i). \tag{21}$$

Conditioning: final comments

- Conditioning can be used in combination with other variance reduction methods. The most straightforward way is to apply those other methods to the problem of estimating $\mathbb{E}[h(\boldsymbol{X})]$.

- The combination of conditioning with stratified and/or antithetic sampling is simple, provided that the distribution of $\boldsymbol{X}$ is amenable to stratification or has some natural symmetry that we can exploit in antithetic sampling.

- Conditioning brings a dimension reduction in addition to the variance reduction, because the dimension k of $\boldsymbol{X}$ is smaller than the dimension d of $(\boldsymbol{X}, \boldsymbol{Y})$.

- In the Rao–Blackwell theorem, the quantity being conditioned on has to obey quite stringent conditions. Those conditions are usually not needed in MC applications.

Control variates: intro

- Control variates provide a way to exploit closed form results. With control variates we use some other problem, quite similar to our given one, but for which an exact answer is known.

- Suppose first that we want to find $\mu = \mathbb{E}[f(\boldsymbol{X})]$ and that we know the value $\theta = \mathbb{E}[h(\boldsymbol{X})]$, where $h(\boldsymbol{X}) \approx f(\boldsymbol{X})$. Using the MC estimators for each of these quantities:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) \qquad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i) \tag{22}$$

  we can estimate μ, using the (unbiased) **difference estimator**:

$$\hat{\mu}_{\text{diff}} = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)) + \theta = \hat{\mu} - \hat{\theta} + \theta. \tag{23}$$

Control variates: estimators

- The variance of the difference estimator is

$$\mathbb{V}[\hat{\mu}_{\text{diff}}] = \frac{1}{n}\mathbb{V}[f(\boldsymbol{X}) - h(\boldsymbol{X})]\,. \tag{24}$$

- If h is similar to f in the sense that the difference $f(\boldsymbol{X}) - h(\boldsymbol{X})$ has smaller variance than $f(\boldsymbol{X})$, we will reduce the variance. In this setting, $h(\boldsymbol{X})$ is called the control variate.

- The difference estimator is not the only way to use a control variate. The ratio and product estimators are also used:

$$\hat{\mu}_{\text{ratio}} = \frac{\hat{\mu}}{\hat{\theta}}\theta \qquad \hat{\mu}_{\text{prod}} = \frac{\hat{\mu}\hat{\theta}}{\theta}; \tag{25}$$

  however, the ratio and product estimators are usually biased.

Control variates: regression estimator (I)

- By far the most common way of using a control variate is through the regression. For a value $\beta \in \mathbb{R}$, the (unbiased) regression estimator of μ is:

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{X}_i) - \beta h(\boldsymbol{X}_i)) + \beta\theta = \hat{\mu} - \beta(\hat{\theta} - \theta); \tag{26}$$

  note that $\beta = 0$ gives standard MC and $\beta = 1$ yields the difference estimator.

- The variance of this estimator is:

$$\mathbb{V}[\hat{\mu}_\beta] = \frac{1}{n}\left(\mathbb{V}[f(\boldsymbol{X})] - 2\beta\mathbb{C}\mathrm{ov}\left[f(\boldsymbol{X}), h(\boldsymbol{X})\right] + \beta^2\mathbb{V}[h(\boldsymbol{X})]\right). \tag{27}$$

- **Intuition**: control variates create a new random vector $\boldsymbol{Z} = f(\boldsymbol{X}) + \beta(h(\boldsymbol{X}) - \theta)$, that allows us to leverage θ in order to compute $\mathbb{E}[f(\boldsymbol{X})]$ in an easier way.

Control variates: regression estimator (II)

- We can find the optimal value of β as:

$$\beta_{\text{opt}} = \arg\min_{\beta} \mathbb{V}[\hat{\mu}_\beta] = \frac{\mathbb{C}\text{ov}\left[f(\boldsymbol{X}), h(\boldsymbol{X})\right]}{\mathbb{V}[h(\boldsymbol{X})]} \quad \text{and} \quad \mathbb{V}\left[\hat{\mu}_{\beta_{\text{opt}}}\right] = \frac{\sigma^2}{n}(1-\rho^2); \tag{28}$$

  note that in the regression estimator, any control variate that correlates with f is helpful, even one that correlates negatively.

- Since we do not know β_{opt} in practice, it can be estimated as

$$\beta_{\text{opt}} \approx \hat{\beta} = \frac{\sum_{i=1}^{n}(f(\boldsymbol{X}_i) - \hat{\mu})(h(\boldsymbol{X}_i) - \hat{\theta})}{\sum_{i=1}^{n}(h(\boldsymbol{X}_i) - \hat{\theta})^2}; \tag{29}$$

  note that the estimator $\hat{\mu}_{\hat{\beta}}$ is no longer unbiased. But the bias is very small !

Control variates: regression estimator (III)

- The estimated variance of $\hat{\mu}_{\hat{\beta}}$ is

$$\hat{\sigma}^2_{\hat{\beta}} = \mathbb{V}\left[\hat{\mu}_{\hat{\beta}}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\left(f(\boldsymbol{X}_i) - \hat{\mu}_{\hat{\beta}} - \hat{\beta}(h(\boldsymbol{X}_i) - \hat{\theta})\right)^2. \tag{30}$$

  and a 99% confidence interval is $\hat{\mu}_{\hat{\beta}} \pm 2.58\,\hat{\sigma}_{\hat{\beta}}$.

- The variance with a control variate is never worse than the MC one. Whether the control variate is helpful ultimately depends on **how much it costs to use it**.

- A significant advantage of the regression estimator is that it generalizes easily to handle multiple control variates. The potential value is greatest when f is expensive but is approximately equal to a linear combination of inexpensive control variates.

Control variates: example I

Let's compute the integral:

$$I = \int_0^{\pi/4} \int_0^{\pi/4} f(x, y) \, \mathrm{d}x \, \mathrm{d}y, \tag{31}$$

where $f(\boldsymbol{X}) = f(x, y) = x^2 y^2 \sin(x + y) \log(x + y)$.

We are going to use the control variate $h(\boldsymbol{X}) = h(x, y) = x^2 y^2$, for which we know the integral is equal to $\theta = ((\pi/4)^6)/9$.

Continue in code...

Control variates: example II

- Consider one of the target applications, where $f(\boldsymbol{U})$ is the forward model solution at location $L/2$ and $\boldsymbol{U} \in \mathbb{R}^3$ is standard Gaussian.

- We can define a control variate to estimate the mean $\mathbb{E}[f(\boldsymbol{U})]$. For instance, a linearization of the map $\boldsymbol{U} \mapsto f(\boldsymbol{U})$. This coarse model $Y = h(\boldsymbol{U})$, is given by a multivariate linear regression:

$$Y_i = c_0 + \sum_{j=1}^{d} c_j U_{i,j} + \eta_i, \qquad i = 1, \ldots, n \tag{32}$$

  here, Y_i is the response for the i-th observation, c_0 is the regression intercept, c_j is the j-th predictor regression, $X_{i,j}$ is the j-th predictor for the i-th observation, and η_j is a Gaussian error term. Here, n is the number of observations used to train the regression.

- Using the ordinary least squares, the coefficients are $\boldsymbol{c} = (\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}\boldsymbol{Y}$. The mean of the control variate is $\theta = \mathbb{E}[\mathbf{U}\boldsymbol{c}] = c_0$.
  Continue in code...

Variance reduction: type-3 methods (“using auxiliary densities”)

Importance sampling: intro

- In many applications, we want to compute $\mu = \mathbb{E}[f(\boldsymbol{X})]$ where $f(\boldsymbol{X})$ is nearly zero outside a region A. The set A may have small volume, or it may be in the tail of the $\boldsymbol{X}$ distribution. A plain MC sample from π could fail to have even one point inside A.

- We must get some samples from the region A. We do this by sampling from a distribution that over-weights the important region, hence the name **importance sampling** (IS) [1].

- IS is more than just a variance reduction method. It can be used to study one distribution while sampling from another. As a result, we can use IS as an alternative to acceptance-rejection.

- IS is also an important prerequisite for *sequential Monte Carlo*, one of the state-of-the-art Bayesian inference techniques.

Importance sampling: intro

- Consider again the problem of finding $\mathbb{E}_\pi[f(\boldsymbol{X})]$:

$$\mu = \int_{\mathbb{R}^d} f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \frac{f(\boldsymbol{x})\pi(\boldsymbol{x})}{\pi_{\text{bias}}(\boldsymbol{x})}\pi_{\text{bias}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_{\pi_{\text{bias}}}\left[f(\boldsymbol{x})\frac{\pi(\boldsymbol{x})}{\pi_{\text{bias}}(\boldsymbol{x})}\right], \tag{33}$$

  where π_{bias} is the so-called importance or *biasing density* ($\text{supp}(f(\boldsymbol{x})\pi(\boldsymbol{x})) \subseteq \text{supp}(f(\boldsymbol{x})\pi_{\text{bias}}(\boldsymbol{x}))$). Moreover, the adjustment factor $\pi(\boldsymbol{x})/\pi_{\text{bias}}(\boldsymbol{x})$ is called the *likelihood ratio*.

- The variance $\sigma^2 = \mathbb{V}_\pi[f(\boldsymbol{X})]$ can be written analogously as:

$$\sigma^2_{\text{IS}} = \int_{\mathbb{R}^d} \frac{(f(\boldsymbol{x})\pi(\boldsymbol{x}))^2}{\pi_{\text{bias}}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} - \mu^2 = \mathbb{E}_{\pi_{\text{bias}}}\left[\frac{(f(\boldsymbol{x})\pi(\boldsymbol{x}) - \mu\pi_{\text{bias}}(\boldsymbol{x}))^2}{\pi^2_{\text{bias}}(\boldsymbol{x})}\right]. \tag{34}$$

Importance sampling: estimators

- The IS estimate of μ is

$$\mu \approx \hat{\mu}_{\mathsf{IS}} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) w(\boldsymbol{X}_i) \qquad \text{with} \qquad w(\boldsymbol{X}_i) = \frac{\pi(\boldsymbol{X}_i)}{\pi_{\mathsf{bias}}(\boldsymbol{X}_i)}, \tag{35}$$

  where $\{\boldsymbol{X}_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} \pi_{\mathsf{bias}}$, and each value $w(\boldsymbol{X}_i)$ represents a *weight* that corrects for the use of the biasing density and ensures that the IS estimator remains unbiased, i.e., $\mathbb{E}_{\pi_{\mathsf{bias}}}[\hat{\mu}_{\mathsf{IS}}] = \mu$.

- Moreover, the variance of the IS estimator is

$$\begin{aligned}
\mathbb{V}_{\pi_{\mathsf{bias}}}[\hat{\mu}_{\mathsf{IS}}] &= \frac{1}{n}\left(\mathbb{E}_{\pi_{\mathsf{bias}}}\left[(f(\boldsymbol{x})w(\boldsymbol{x}) - \mu)^2\right]\right) \\
&= \frac{1}{n}\Big(\int_{\mathbb{R}^d} \frac{(f(\boldsymbol{x})\pi(\boldsymbol{x}))^2}{\pi_{\mathsf{bias}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \mu^2\Big) = \frac{1}{n}\left(\mathbb{E}_{\pi_{\mathsf{bias}}}\left[(f(\boldsymbol{x})w(\boldsymbol{x}))^2\right] - \mu^2\right).
\end{aligned}$$

Importance sampling: optimal biasing density (I)

- We can also approximate the 99% confidence interval for μ similar to the MC case, i.e.,

$$\hat{\mu}_{\mathsf{IS}} \pm 2.58 \frac{\hat{\sigma}_{\mathsf{IS}}}{\sqrt{n}} \qquad \text{where} \quad \hat{\sigma}^2_{\mathsf{IS}} = \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{X}_i) w(\boldsymbol{X}_i) - \hat{\mu}_{\mathsf{IS}})^2. \tag{36}$$

Importance sampling: optimal biasing density (I)

- We can also approximate the 99% confidence interval for μ similar to the MC case, i.e.,

$$\hat{\mu}_{\mathsf{IS}} \pm 2.58 \frac{\hat{\sigma}_{\mathsf{IS}}}{\sqrt{n}} \qquad \text{where} \quad \hat{\sigma}^2_{\mathsf{IS}} = \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{X}_i) w(\boldsymbol{X}_i) - \hat{\mu}_{\mathsf{IS}})^2. \tag{39}$$

- To choose a good biasing distribution requires some educated guessing and possibly numerical search. **Rule**: π_{bias} should have tails at least as heavy as π (domination !).

- We can also try to find the optimal biasing density as follows. Aiming to reduce the variance, we require a π_{bias} such that,

$$\pi^{\star}_{\mathsf{bias}}(\boldsymbol{x}) = \arg\min_{\pi_{\mathsf{bias}}} \mathbb{E}_{\pi_{\mathsf{bias}}} \left[(f(\boldsymbol{x}) w(\boldsymbol{x}))^2 \right], \tag{37}$$

  where $\pi^{\star}_{\mathsf{bias}}$ is the so-called *optimal biasing density*.

Importance sampling: optimal biasing density (II)

- The minimizer of eq. (37) can be found by applying Jensen's inequality

$$\mathbb{E}_{\pi_{\text{bias}}}\left[(f(\boldsymbol{x})w(\boldsymbol{x}))^2\right] \geq \left(\mathbb{E}_{\pi_{\text{bias}}}[|f(\boldsymbol{x})|\, w(\boldsymbol{x})]\right)^2.$$

  The relation is strict if $|f(\boldsymbol{x})|\, w(\boldsymbol{x})$ is constant. Hence, the optimal biasing density, generating a zero-variance estimate, is given by

$$\pi^{\star}_{\text{bias}}(\boldsymbol{x}) \propto |f(\boldsymbol{x})|\, \pi(\boldsymbol{x}) = \frac{1}{\mu}\, |f(\boldsymbol{x})|\, \pi(\boldsymbol{x}).$$

- Although zero-variance biasing densities are not usable, they provide insight into the design of a good IS scheme, e.g., the cross-entropy method.

- The likelihood ratio also reveals a dimension effect for IS. Some weights can become significantly larger than others.

Importance sampling: example (I)

- We want to estimate the integral:

$$\int_0^{10} f(x)\,\mathrm{d}x \qquad \text{with} \quad f(x) = \exp(-2\,|x-5|). \tag{38}$$

- Problem with standard MC: this function is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution ($\pi = \mathcal{U}(0, 10)$), many of the points are contributing very little to this expectation.

- Something more like a Gaussian function with mean at 5 and small variance, say, 1, would provide greater precision: $\pi_{\text{bias}} = \mathcal{N}(5, 1)$. Hence:

$$\mathbb{E}_{\pi_{\text{bias}}}[f(x)w(x)] = \int_0^{10} 10\exp(-2\,|x-5|)\frac{\frac{1}{10}}{\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-5)^2}{2})}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)\mathrm{d}x. \tag{39}$$

- Continue in code...

Importance sampling: example (II)

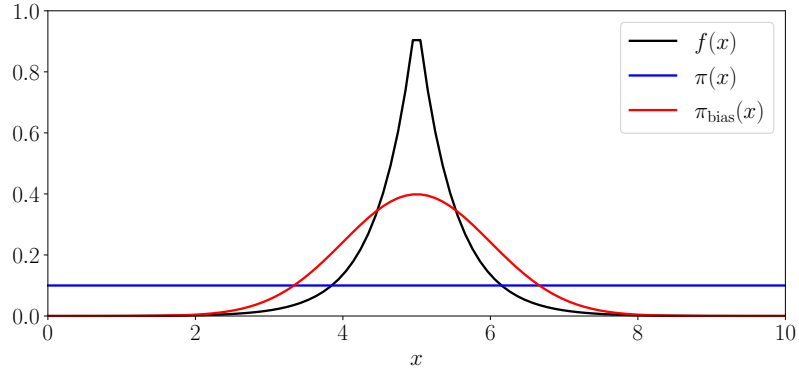

Figure: For the integration of $f(x)$, IS provided a substantial increase in precision. MC std ≈ 2.02, IS std ≈ 0.59.

Importance sampling: self-normalized

- Sometimes we can only compute an unnormalized version of π, $\overline{\pi}(\boldsymbol{x}) = c_1\pi(\boldsymbol{x})$ where $c_1 > 0$ is unknown. The same may be true for the biasing density, i.e., $\overline{\pi}_{\text{bias}}(\boldsymbol{x}) = c_2\pi_{\text{bias}}(\boldsymbol{x})$, where $c_2 > 0$ is unknown.

- In this case, we can compute the likelihood ratio $\overline{w} = \overline{\pi}(\boldsymbol{x})/\overline{\pi}_{\text{bias}}(\boldsymbol{x}) = (c_1/c_2)(\pi(\boldsymbol{x})/\pi_{\text{bias}}(\boldsymbol{x}))$, and use the **self-normalized IS** estimator:

$$\mu \approx \hat{\mu}_{\text{sIS}} = \frac{\sum_{i=1}^{n} f(\boldsymbol{X}_i)\,\overline{w}(\boldsymbol{X}_i)}{\sum_{i=1}^{n} \overline{w}(\boldsymbol{X}_i)}, \tag{40}$$

  where $\{\boldsymbol{X}_i\}_{i=1}^{n} \sim \pi_{\text{bias}}$.

- The self-normalized IS estimator requires a stronger condition on π_{bias}. We now need $\pi_{\text{bias}}(\boldsymbol{x}) > 0$ whenever $\pi(\boldsymbol{x}) > 0$, even if $f(\boldsymbol{x})$ is zero with high probability.

Importance sampling: diagnostics I

- IS uses unequally weighted samples. The weights are $w_i = w(\boldsymbol{X}_i) = \pi(\boldsymbol{X}_i)/\pi_{\text{bias}}(\boldsymbol{X}_i) > 0$ for $i = 1, \ldots, n$. We want to have a diagnostic to tell when the weights are problematic.

- A common metric is the effective sample size[4]:

$$n_{\text{ESS}} = \frac{\left(\sum_{i=1}^{n} w(\boldsymbol{x}_i)\right)^2}{\sum_{i=1}^{n} (w(\boldsymbol{x}_i))^2} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{N} w_i^2} = \frac{\left(\widetilde{W}\right)^2}{\widetilde{W^2}}, \tag{41}$$

  where $\widetilde{W}$ denotes the sum of the weights, $\widetilde{W^2}$ the sum of the squared weights, and $1 \leq n_{\text{ESS}} \leq n$.

- The weights are all the same when $n_{\text{ESS}} = n$. Conversely, if the weights are very unequal, the IS estimator is averaging only with $n_{\text{ESS}} \ll n$ samples and thus it is less accurate.

[4] A. B. Owen. *Monte Carlo theory, methods and examples*. artowen.su.domains/mc/, 2018, Ch.9 p.11.

Importance sampling: diagnostics II

- Another way to express the n_{ESS} is via the coefficient of variation of the weights $\text{cv}(\boldsymbol{w})$:

$$n_{\text{ESS}} = \frac{n}{1 + (\text{cv}(\boldsymbol{w}))^2}, \tag{42}$$

  where $\boldsymbol{w} = \{w_i\}_{i=1}^n$ is the vector of weights. Again, if n_{ESS} is too small, we know π_{bias} produces imbalanced weights.

- Effective sample sizes are imperfect diagnostics: When they are too small, we have a sign that they are problematic. When they are large, we still cannot conclude that IS has worked.

- Moreover, badly skewed weights could give a badly estimated mean along with a bad variance estimate that masks the problem.

- We can also use the variance as a diagnostic. When it is quite large, we would conclude that IS has not worked well.

Importance sampling: some comments

- IS and acceptance-rejection sampling are quite similar ideas.
- Some techniques used to find biasing densities are:
  - **Exponential tilting**: IS by changing the parameter $\boldsymbol{\theta}$ of a $\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})$ chosen from an exponential family.
  - **Modes and Hessians**: matching the Hessian of $\pi_{\text{bias}}(\boldsymbol{x})$ to that of $\pi(\boldsymbol{x})$ at the mode.
  - **Mixture IS**: $\pi_{\text{bias}}(\boldsymbol{x})$ comes from a mixture distribution. Mixtures of unimodal densities provide a flexible approximation to multimodal targets.
  - **Defensive IS**: we take a $\pi_{\text{bias}}(\boldsymbol{x})$ thought to be a good one and mix it with $\pi(\boldsymbol{x})$, i.e., $\pi_{\text{bias}}(\boldsymbol{x};\alpha) = \alpha_1\pi(\boldsymbol{x}) + \alpha_2\pi_{\text{bias}}(\boldsymbol{x})$.
  - **Cross-entropy method**: finds an optimal approximation in the Kullback–Leibler divergence sense.

Importance sampling: cross-entropy method

- The standard *cross-entropy* (CE) method [3] considers the problem of approximating $\pi^{\star}_{\text{bias}}(\boldsymbol{x})$ by a *parametric biasing* density $\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})$, with reference parameters $\boldsymbol{\theta}$.

- The approximation is selected from a family of densities $\Pi = \{\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ designed to be of simpler form than $\pi^{\star}_{\text{bias}}$.

- Thereafter, the objective is to find $\boldsymbol{\theta}^{\star} \in \Theta$ such that the distance between the optimal and approximated biasing densities is minimal. The dissimilarity between these distributions is measured by the cross-entropy or Kullback–Leibler divergence (KLD)

$$
\begin{aligned}
D_{\text{KL}}\left(\pi^{\star}_{\text{bias}} || \pi_{\text{bias}}\right) &= \int_{\mathbb{R}^d} \ln\left(\frac{\pi^{\star}_{\text{bias}}(\boldsymbol{x})}{\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})}\right) \pi^{\star}_{\text{bias}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} \ln \pi^{\star}_{\text{bias}}(\boldsymbol{x})\ \pi^{\star}_{\text{bias}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_{\mathbb{R}^d} \ln \pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})\ \pi^{\star}_{\text{bias}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \qquad (43)
\end{aligned}
$$

Importance sampling: cross-entropy method

- The first term in eq. (43) is invariant with respect to any choice of π_{bias} and the problem reduces to the optimization task:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\, \mathbb{E}_{\pi^{\star}_{\text{bias}}}[\ln \pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})]\,, \tag{44}$$

  where $\boldsymbol{\theta}^{\star}$ denotes the optimal reference parameters. We can substitute the optimal IS biasing density into eq. (44) to express the optimization program as

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\, \mathbb{E}_{\pi}[\ln \pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})\ f(\boldsymbol{x})]\,. \tag{45}$$

- To efficiently evaluate eq. (45), we apply IS with biasing distribution $\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta}') \in \Pi$:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\, \mathbb{E}_{\pi_{\text{bias}}(\cdot;\boldsymbol{\theta}')}[\ln \pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{x})\, w(\boldsymbol{x};\boldsymbol{\theta}')] \quad \text{with} \quad w(\boldsymbol{x};\boldsymbol{\theta}') = \frac{\pi(\boldsymbol{x})}{\pi_{\text{bias}}(\boldsymbol{x};\boldsymbol{\theta}')}. \tag{46}$$

Importance sampling: cross-entropy method

- We can further employ the IS estimator of the expectation in eq. (46) to define the stochastic optimization problem:

$$\boldsymbol{\theta}^\star \approx \widehat{\boldsymbol{\theta}}^\star = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\, \mathcal{J}(\boldsymbol{\theta}) \quad \text{with} \quad \mathcal{J}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln \pi_{\text{bias}}(\boldsymbol{X}_i;\boldsymbol{\theta}) f(\boldsymbol{X}_i) w(\boldsymbol{X}_i;\boldsymbol{\theta}'), \quad (47)$$

  where $\{\boldsymbol{X}_i\}_{i=1}^n \overset{\text{iid}}{\sim} \pi_{\text{bias}}(\cdot;\boldsymbol{\theta}')$.

- If the biasing distribution belongs to the natural exponential family, the solution of the stochastic optimization problem can be computed analytically. For instance, if Π is a collection of Gaussian densities, the parameter $\boldsymbol{\theta}$ is selected from the space Θ containing mean vectors and covariance matrices.

- In this case, the reference parameter estimator $\widehat{\boldsymbol{\theta}}^\star$ has an explicit updating rule.

Importance sampling: cross-entropy method

- One still requires a good initial choice of $\boldsymbol{\theta}'$, such that a substantial number of samples from $\pi_{\text{bias}}(\boldsymbol{x}; \boldsymbol{\theta}')$ lie in the failure domain. This is addressed in the CE method by gradually approaching the optimal biasing density. The idea is to construct a sequence of intermediate sets $\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \leq \gamma_j\}$, with intermediate thresholds $\gamma_j \geq 0$.

- Starting from an initial reference parameter estimate $\widehat{\boldsymbol{\theta}}_0$, the sequential CE program reads

$$\widehat{\boldsymbol{\theta}}_{j+1} = \arg\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln \pi_{\text{bias}}(\boldsymbol{X}_i; \boldsymbol{\theta}) \widetilde{w}_i^{(j)} \quad \text{with} \quad \widetilde{w}_i^{(j)} = f(\boldsymbol{X}_i) \frac{\pi(\boldsymbol{X}_i)}{\pi_{\text{bias}}(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_j)}, \tag{48}$$

  where $\{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{iid}}{\sim} \pi_{\text{bias}}(\cdot; \widehat{\boldsymbol{\theta}}_j)$.

- The CE optimization eq. (47) is now solved at each level with respect to an intermediate optimal biasing density $\pi^{\star}_{\text{bias},j}(\boldsymbol{x})$ associated to a threshold γ_j.

Importance sampling: cross-entropy method

- Note that if π and π_{bias} belong to the same parametric family, the initial estimate of the reference parameters is typically selected as the parameters defining π (e.g., if Gaussian, $\widehat{\boldsymbol{\theta}}_0 = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$).

- In the CE method, $f(\boldsymbol{x})$ can be either an indicator function (if a rare event problem) or a likelihood (if a Bayesian problem).

- If $f(\boldsymbol{x})$ accounts for a rare event problem, each threshold γ_j is defined as the ρ-quantile of the sequence of values $\{f_i = f(\boldsymbol{X}_i)\}_{i=1}^n$. The value ρ is chosen to ensure that a good portion of the samples from $\pi_{\text{bias}}(\cdot; \boldsymbol{\theta}_j)$ fall in the next set set, usually $\rho \in [0.01, 0.1]$.

Importance sampling: cross-entropy method example (see Project)

- Consider one of the target applications, where $f(\boldsymbol{U})$ is the forward model solution at location $L/2$ and $\boldsymbol{U} \in \mathbb{R}^3$ is standard Gaussian.
- We can define a rare event problem of estimating the probability that the model response exceeds a maximum allowed threshold, i.e., $\mathbb{P}[\tau \leq f(\boldsymbol{x})]$. The thresholds are $\tau = \{50, 10, 1.5, 2\}$ for Poisson, Heat, Abel and Deconvolution problems, respectively.
- Assuming the biasing density belongs to a family of Gaussian distributions. Employ the CE method to find $\mathbb{P}[\tau \leq f(\boldsymbol{x})]$.
- Continue on code...

Variance reduction: final comments

- Variance reduction is an ongoing field of research in UQ, for both forward and inverse problems.
- Many of the methods exposed here can be extended to the case of inverse problems within the Bayesian framework.
- We will see that practical UQ for inverse problems requires a solid foundation on stochastic simulation (i.e., the methods discussed in the past lectures).

References

[1] H. Kahn et al. "Methods of reducing sample size in Monte Carlo computations". In: *Journal of the Operations Research Society of America* 1.5 (1953), pp. 263–278.

[2] A. B. Owen. *Monte Carlo theory, methods and examples*. `artowen.su.domains/mc/`, 2018.

[3] R. Y. Rubinstein. "Optimization of computer simulation models with rare events". In: *European Journal of Operational Research* 99.1 (1997), pp. 89–112.