2 Importance Sampling

Can we do better than the simple Monte Carlo estimator of

$$\theta = E[g(X)] = \int g(x)f(x)dx \approx \frac{1}{m}\sum_{i=1}^{m} g(X_i)$$

where the variables $X_1, \ldots, X_m$ are randomly sampled from f?

Yes!!

Goal: estimate integrals with lower variance than the simplest Monte Carlo approach.

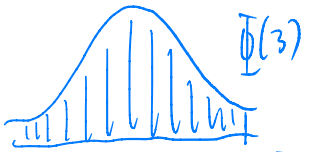
↳ more efficient estimaton

To accomplish this, we will use *importance sampling*.

2.1 The Problem

If we are sampling an event that doesn't occur frequently, then the naive Monte Carlo estimator will have high variance.

Example 2.1 Monte Carlo integration for the standard Normal cdf. Consider estimating $\Phi(-3)$ or $\Phi(3)$. (HW 6)



↖ events out here are rare => we may not get many samples in the MC estimator

We want to improve accuracy by causing rare events to occur more frequently than they would under the naive Monte Carlo sampling framework, thereby enabling more precise estimation.

For very rare events, extremely large reductions in the variance of the MC estimator are possible.

2.2 Algorithm

Consider a density function $f(x)$ with support $\mathcal{X}$. Consider the expectation of $g(X)$,

$$\theta = E[g(X)] = \int_{\mathcal{X}} g(x)f(x)dx.$$

Let $\phi(x)$ be a density where $\phi(x) > 0$ for all $x \in \mathcal{X}$. Then the above statement can be rewritten as

↑ support of ϕ includes support of f.

$$\theta = E[g(X)] = \int_{\mathcal{X}} g(x)\frac{f(x)}{\phi(x)}\phi(x)\,dx$$

ϕ is called the importance sampling function (similar to an envelope in accept-reject)

ϕ MUST be a density (i.e. integrate to 1 and always ≥ 0).

An estimator of ϕ is given by the *importance sampling algorithm*:

1. Sample $X_1, \ldots, X_m$ from ϕ
2. Compute

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)\frac{f(X_i)}{\phi(X_i)}$$

For this strategy to be convenient, it must be

(1) easy to sample from ϕ

(2) easy to evaluate f (even if it's not easy to sample from f

X = result of rolling 1 fair six-sided die.

Example 2.2 Suppose you have a fair six-sided die. We want to estimate the probability that a single die roll will yield a 1. Want to estimate $P(X=1)$.

We could:

① Roll a die m times (expect $\approx \frac{m}{6}$ ones)

② A point estimate of $P(X=1)$ would be proportion of ones in the sample.

The variance of this estimator is $\frac{5}{36m}$ if the die is fair.

$$X = \{1, \dots, 6\} \qquad f(x) = \begin{cases} \frac{1}{6} & x = 1, \dots, 6 \\ 0 & \text{o.w.} \end{cases}$$

Define $Y = \begin{cases} 1 & \text{if } X=1 \\ 0 & \text{o.w.} \end{cases} \Rightarrow Y \sim \text{Bernoulli}\left(\frac{1}{6}\right)$

$$EY = \sum_{i=1}^{6} \mathbb{I}[y=1] \cdot \frac{1}{6} = \frac{1}{6}$$

$$\text{Var}\, Y = p(1-p) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$$

Expected # of 1s in m rolls:

$$E(\Sigma Y) = \sum_{i=1}^{m} EY_i = \frac{m}{6}$$

Proportion of 1s in the sample:

$$E\left(\frac{\Sigma Y_i}{m}\right) = \frac{1}{m} E(\Sigma Y_i) = \frac{1}{6}$$

$$\text{Var}\left(\frac{\Sigma Y_i}{m}\right) = \frac{1}{m^2} \text{Var}\, \Sigma Y_i = \frac{1}{m^2} \sum_{i=1}^{m} \text{Var}\, Y_i = \frac{1}{m^2} \sum_{i=1}^{m} \frac{5}{36} = \frac{5}{36m}$$

relative measure of variation, used in Chemistry, physics, etc.

We can consider the "coefficient of variation" $CV[X] = \frac{\sqrt{\text{Var}\, X}}{E[X]}$

So $CV\left[\frac{\Sigma Y_i}{m}\right] = \frac{\sqrt{\text{Var}\, \Sigma Y_i/m}}{E\, \Sigma Y_i/m} = \frac{\sqrt{5/36m}}{1/6}$ ← estimator

If we want a CV of 5%, then

$$\frac{\sqrt{\frac{5}{36m}}}{\frac{1}{6}} = 0.05$$

$$\frac{5}{36m} = \left[\frac{1}{6}(.05)\right]^2$$

$$\frac{5}{36 \cdot [6(.05)]^2} = m \Rightarrow m = 2000 \text{ rolls!}$$

Example 2.2 Suppose you have a fair six-sided die. We want to estimate the probability that a single die roll will yield a 1. (cont'd)

To reduce the # of rolls, we could consider biasing the die by replacing the faces bearing 2 and 3 with additional 1s.

This increases the prob. of rolling a 1 to 0.5, but now we are sampling from not the target dsn (a fair die).

Now $P(X=1) = \frac{1}{2}$

$P(X=2) = P(X=3) = 0$

$P(X=4) = P(X=5) = P(X=6) = \frac{1}{6}$

Can correct this by

- weighting each roll of 1 by $1/3$
- Let $Y_i = \begin{cases} \frac{1}{3} & \text{if } X=1 \\ 0 & \text{otherwise.} \end{cases}$

Then the expectation of the sample mean

$$E\left(\sum_{i=1}^{m} \frac{Y_i}{m}\right) = \frac{1}{m}\sum_{i=1}^{m} EY_i = EY = \frac{1}{3}\cdot\frac{1}{2} + 0\left[0+0+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right] = \frac{1}{6} \checkmark$$

But the variance is

$$Var\left(\sum_{i=1}^{m} \frac{Y_i}{m}\right) = \frac{1}{m^2}\sum_{i=1}^{m} Var\, Y_i = \frac{1}{m} Var\, Y = \frac{1}{m}\left[\frac{1}{18} - \frac{1}{6^2}\right] = \frac{1}{36\cdot m}$$

$EY^2 = \frac{1}{3^2}\cdot\frac{1}{2} = \frac{1}{18}$

So to achieve a CV of 5% we would need only

$$\frac{\sqrt{\frac{1}{36m}}}{\frac{1}{6}} = .05$$

$$\frac{1}{36m} = \left(\frac{1}{6}\times .05\right)^2$$

$$m = \frac{1}{36\left(\frac{1}{6}\times .05\right)^2} = 400 \text{ rolls.}$$

The die rolling example is successful because an importance sampling function (rolling a die w/3 ons) is used to over-sample a portion of the state space that recieves lower prob. under the target dsn and importance weighting corrects the bias.

2.3 Choosing ϕ

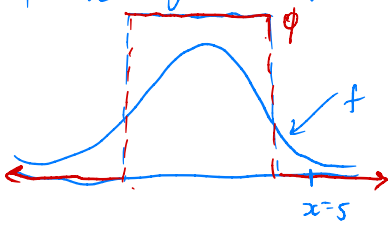
In order for the estimators to avoid excessive variability, it is important that $f(x)/\phi(x)$ is bounded and that ϕ has heavier tails than f.

importance weights

If this requirement is not met, then some importance weights will be huge.

Example 2.3

If we ignore requirement that $\phi(x) > 0$ when $f(x) > 0$,

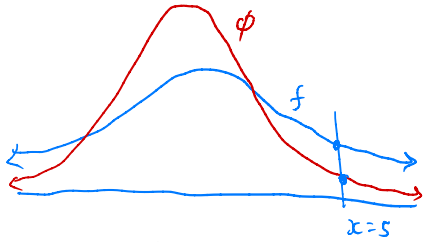


Then $\frac{f(5)}{\phi(5)} = \frac{f(5)}{0}$ unbounded!

AND, can't draw $X=5$ from ϕ!

Example 2.4

If we select ϕ with lighter tails than f



$\frac{f(5)}{\phi(5)}$ will be large if $\phi(5)$ is small.

Thus $x=5$ draw has large weight and integral approx. will be poor.

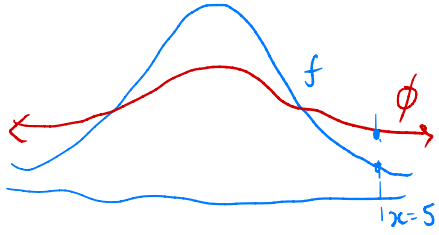
A rare draw from ϕ with much higher density under f than under ϕ will receive a huge weight and inflate the variance of the estimate.

Strategy – choose the function ϕ so that $f(x)/\phi(x)$ is large only when $g(x)$ is small.

Example 2.5

If we select an appropriate ϕ,

$f(0)/\phi(0)$ will be large and



$\frac{f(5)}{\phi(5)}$ will be small. Rare draw from f should have small weight.

The importance sampling estimator can be shown to converge to θ under the SLLN so long as the support of ϕ includes all of the support of f.

2.4 Compare to Previous Monte Carlo Approach

Common goal – estimate an integral $\int h(x)\,dx$.

Step 1 Do some derivations.

a. Find an appropriate f and g to rewrite your integral as an expected value.

$\theta = \int_{-\infty}^{\infty} g(x) f(x)\,dx = E[g(X)]$ wrt. f.

b. For **importance sampling** only,

Find an appropriate ϕ to rewrite θ as an expectation with respect to ϕ.

$$\theta = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{\phi(x)} \phi(x)\,dx = E\left[g(X)\frac{f(X)}{\phi(X)}\right] \text{ wrt } \phi$$

NOTE: $\phi(x) > 0$ when $f(x) > 0$ required.

Step 2 Write pseudo-code (a plan) to define estimator and set-up the algorithm.

- For **Monte Carlo integration**
  1. Sample $X_1, \ldots, X_m \sim f$
  2. $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)$

- For **importance sampling**
  1. Sample $X_1, \ldots, X_m \sim \phi$
  2. $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)\frac{f(X_i)}{\phi(X_i)}$

importance weight.

Step 3 Program it.

2.5 Extended Example

In this example, we will estimate $\theta = \int_0^1 \frac{e^{-x}}{1+x^2} dx$ using MC integration and importance sampling with two different importance sampling distributions, ϕ.

STEP 1 derive things.

a) select $X \sim Exp(1)$ so $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$

$$\Rightarrow \theta = \int_0^1 \frac{e^{-x}}{1+x^2} dx = \int_0^\infty \underbrace{\frac{1}{1+x^2} 1(x \leq 1)}_{g(x)} \underbrace{e^{-x}}_{f(x)} dx = E\left[\frac{1}{1+X^2} 1(X \leq 1)\right] \text{ where } X \sim Exp(1)$$

Option ① MC integration (no 1b step)

Option ② Importance sampling with

a) $\phi \sim Uniform(0,1)$ $\phi_a(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$

b) $\phi \sim Exp(1)$ rescaled to have support $0 \leq x \leq 1$.

$$\Rightarrow \phi_b(x) = \begin{cases} \frac{e^{-x}}{1+e^{-1}} & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

← you can check this is a valid pdf because it integrates to 1.

STEP 1b Option 2a) $\phi_a(x) = 1 \quad 0 \leq x \leq 1.$

$$\theta = E_f[g(X)] = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{\phi(x)} \phi(x) dx$$

$$= \int_0^\infty \frac{1}{1+x^2} 1(x \leq 1) \frac{e^{-x}}{1} \cdot 1 \, dx$$

$$= E\left[\frac{1}{1+X^2} 1[X \leq 1] e^{-x}\right] \text{ wrt } \phi_a.$$

Option 2b) $\phi_b = e^{-x} / 1+e^{-1} \quad 0 \leq x \leq 1.$

$$\theta = E_f[g(X)] = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{\phi(x)} \phi(x) dx = \int_0^\infty \frac{1}{1+x^2} 1(x \leq 1) \frac{e^{-x}}{\frac{e^{-x}}{1+e^{-1}}} \cdot \frac{e^{-x}}{1+e^{-1}} dx$$

$$= E\left[\frac{1+e^{-1}}{1+x^2} 1(X \leq 1)\right] \text{ wrt } \phi_b.$$

STEP 2 Make a plan.

Option 1:

1. Sample $X_1, \dots, X_m$ from $Exp(1)$
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{1+X_i^2} 1(X_i \le 1) \right]$

Option 2a:

1. Sample $X_1, \dots, X_m \sim Unif(0,1)$
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1+X_i^2} \underbrace{1(X_i \le 1)}_{\text{This will always hold.}} e^{-X_i}$

Option 2b:

Want to sample from ϕ_3, but not a standard named dsn.

$\Rightarrow$ Inverse transform!

$$F_{\phi_3}(x) = \int_0^x \frac{e^{-y}}{1-e^{-1}} dy = -\frac{e^{-y}}{1-e^{-1}} \Big|_0^x = \frac{1-e^{-x}}{1-e^{-1}} \quad \text{for } x \in [0,1].$$

Let $u = F_{\phi_3}(x) = \frac{1-e^{-x}}{1-e^{-1}}$

$$u(1-e^{-1}) = 1-e^{-x}$$

$$e^{-x} = 1 - u(1-e^{-1})$$

$$F^{-1}(u) = x = -\log\left[1 - u(1-e^{-1})\right]$$

↓

Algorithm for 2b

1. Sample $X_1, \dots, X_m \sim \phi_3$ using the inverse transform method.
   a) Sample $U_1, \dots, U_m$ from $\text{Unif}(0,1)$
   b) Set $X_i = -\log\left[1 - U_i(1 - e^{-1})\right]$ for $i = 1, \dots, m$
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \frac{1 - e^{-1}}{1 + X_i^2} 1(X_i \le 1)$

this will always hold by def'n.

Which will be the best? We can compare

$f(x)g(x)$ to f, ϕ_a, and ϕ_b

can look at $\frac{f(x)g(x)}{f(x)}$, $\frac{f(x)g(x)}{\phi_a(x)}$, $\frac{f(x)g(x)}{\phi_b(x)}$ ← want these to be constant.

This will give us the lowest variance.