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 $X_1, \dots, X_m$ are randomly sampled from f?

Yes!! Probably.

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

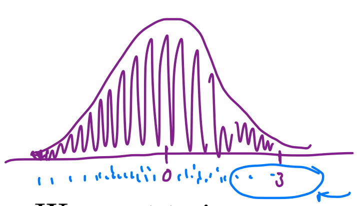
↳ more efficient estimation.

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 could be 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 reduction 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 the support of f

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

$$= \int_{\mathcal{X}_\phi} g(x) \frac{f(x)}{\phi(x)} \phi(x) dx = E\left[g(Y)\frac{f(Y)}{\phi(Y)}\right], Y \sim \phi$$

ϕ is called the importance sampling function

ϕ must be a density (integrate to 1 and ≥ 0).

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

1. Sample $X_1, \dots, X_m \sim \phi$
2. Compute $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} g(X_i) \underbrace{\frac{f(X_i)}{\phi(X_i)}}_{\text{importance weights}}$

For this strategy to be convenient, it must be

(1) easy to sample from ϕ

(2) easy to evaluate f (even if its 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 fair die m times

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

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

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

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

$$EY = \frac{1}{6}$$

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

Estimator: Proportion of 1's in the sample

$$E\left[\frac{\sum Y_i}{m}\right] = \frac{1}{m} E\left(\sum Y_i\right) = \frac{1}{6}$$

$$\text{Var}\left[\frac{\sum Y_i}{m}\right] = \frac{1}{m^2} \sum \text{Var}(Y_i) = \frac{5}{36} \cdot \frac{1}{m}$$

relative measure of variability (commonly used in chemistry or physics).

$$CV(X) = \frac{\sqrt{\text{Var}(X)}}{E(X)}$$

We can consider the "coefficient of variation"

estimator

$$\text{So} \quad CV\left[\frac{\sum Y_i}{m}\right] = \frac{\sqrt{\text{Var}\left(\frac{\sum Y_i}{m}\right)}}{E\left(\frac{\sum Y_i}{m}\right)} = \frac{\sqrt{\frac{5}{36m}}}{1/6}$$

If we want CV of 5% then:

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

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

$$\frac{5}{36\left[\frac{1}{6}(.05)\right]^2} = m \longrightarrow m = 2000 \text{ rolls!}$$

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

This increases probability of rolling a 1 to $\frac{1}{2}$ but now we are not sampling from the target dsn!
↳ 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 a 1 by $\frac{1}{3}$

let $Y_i = \begin{cases} 1/3 & \text{if } X_i = 1 \\ 0 & \text{o.w.} \end{cases}$

Then the expectation of the sample mean $\left(\frac{\sum Y_i}{m}\right)$:
$$E\left(\frac{\sum_{i=1}^m Y_i}{m}\right) = \frac{1}{m}\sum_{i=1}^m EY_i = EY = \frac{1}{3}\cdot\frac{1}{2} + 0\left[\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right] = \frac{1}{6}$$

$$EY^2 = \frac{1}{3^2}\cdot\frac{1}{2} + 0^2\left[\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right] = \frac{1}{18}$$

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

So to achieve a CV of 5% we would only need:
$$\frac{\sqrt{\frac{1}{36m}}}{\frac{1}{6}} = .05$$
$$\vdots$$
$$m = 400 \text{ rolls.}$$

This is successful because an importance sampling function (rolling a die w/ 3 ones) is used to over sample a portion of the state space that we cared about receives low prob under the target dsn. and correcting the bias.

2.3 Choosing ϕ

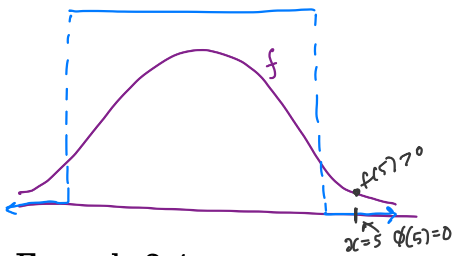
In order for the estimators to avoid excessive variability, it is important that $f(x)/\phi(x)$ is (1) bounded and that (2) ϕ has heavier tails than f.

importance weights.

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

Example 2.3

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

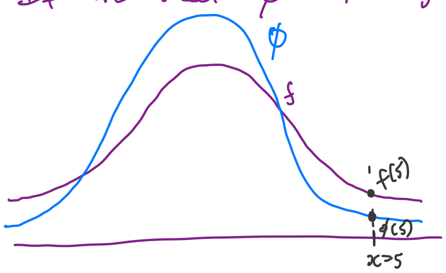


Then $\frac{f(5)}{\phi(5)} = \frac{f(5)}{0}$ unbounded!

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.

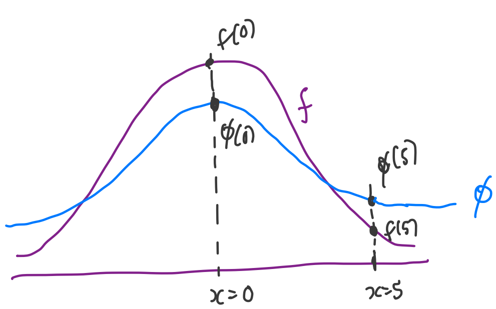
Thus $x = 5$ draw has large weight and integral approximation will be poor.

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 ϕ so that $f(x)/\phi(x)$ large only when $g(x)$ is small.

Example 2.5

If we select an appropriate ϕ,



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

$\frac{f(0)}{\phi(0)}$ will be large (depending on g, could be fine).

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.

$* \ \mathcal{X}_\phi \supseteq \mathcal{X}_f$

2.4 Compare to Previous Monte Carlo Approach

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

Step 1 Do some derivations.

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

Want $\theta = \int h(x)\,dx = \int_{\mathcal{X}} g(x) f(x)\,dx = E[g(X)]$, $X \sim f$ (with support $\mathcal{X}$).

b. For **importance sampling** only,

$\phi(x) > 0$ when $f(x) > 0$ required!

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

Want $\theta = \int_{\mathcal{X}_\phi} g(x) \frac{f(x)}{\phi(x)} \phi(x)\,dx = E\left[g(Y)\frac{f(Y)}{\phi(Y)}\right]$, $Y \sim \phi$ (with support $\mathcal{X}_\phi$).

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

(a) • For **Monte Carlo integration**

1. Sample $X_1, \dots, X_m \sim f$
2. Compute $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)$.

(b). • For **importance sampling**

1. Sample $Y_1, \dots, Y_m \sim \phi$
2. Compute $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(Y_i) \underbrace{\frac{f(Y_i)}{\phi(Y_i)}}_{\text{importance.}}$

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 (1) and importance sampling (2) with two different importance sampling distributions, ϕ. a) and b).

$\boxed{\text{STEP 1: DERIVE THINGS}}$

(1) Select f to be the pdf of Exp(1) so, $f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases} = e^{-x}\mathbb{I}(x \geq 0)$.

MC

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

limits of int. matches support of Exp(1) ✓

Importance Sampling with

a) Normal(0,1).

b) t_1 (t dsn w/ 1 degree of freedom.).

(2) a)

ϕ is pdf of $N(0,1)$.

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

$$= \int_{-\infty}^\infty \frac{1}{1+x^2}\mathbb{I}(x \leq 1)\frac{e^{-x}\mathbb{I}(x \geq 0)}{\phi(x)} \phi(x) dx.$$

$$= E\left[\frac{1}{1+Y^2}\mathbb{I}(Y \leq 1)\frac{e^{-Y}\mathbb{I}(Y \geq 0)}{\phi(Y)}\right], \quad Y \sim N(0,1).$$

b) let τ denote the pdf of t_1 r.v.

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

$$= \int_{-\infty}^\infty \frac{1}{1+x^2}\mathbb{I}(x \leq 1)\frac{e^{-x}\mathbb{I}(x \geq 0)}{\tau(x)} \tau(x) dx$$

$$= E\left[\frac{1}{1+Z^2}\mathbb{I}(Z \leq 1)\frac{e^{-Z}\mathbb{I}(Z \geq 0)}{\tau(Z)}\right], \quad Z \sim t_1$$

STEP 2: MAKE A PLAN

Option 1:

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

Option 2a:

1. Sample $Y_1, \ldots, Y_m$ from $N(0,1)$
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{1+Y_i^2} \mathbb{I}(Y_i \leq 1) \cdot \frac{e^{-Y_i} \mathbb{I}(Y_i \geq 0)}{\phi(Y_i)} \right]$

Option 2b:

1. Sample $Z_1, \ldots, Z_m$ from t_1
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{1+Z_i^2} \mathbb{I}(Z_i \leq 1) \cdot \frac{e^{-Z_i} \mathbb{I}(Z_i \geq 0)}{\tau(Z_i)} \right]$ where τ is the pdf of a t_1 r.v.

Which will be the best?

We can compare $\underline{f(x)g(x)} = h(x)$. to f, ϕ, τ

can look at $\cancel{\dfrac{f(x)g(x)}{f(x)}}, \dfrac{f(x)g(x)}{\phi(x)}, \dfrac{f(x)g(x)}{\tau(x)}$.

☆☆☆ We want these ratios to be constant.

This will give the lowest variance ☆☆☆