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 estimator

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

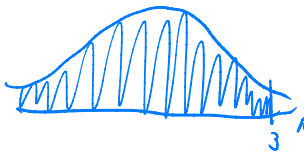
2.1 The Problem

rare event

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)

$\Phi(3) = P(X \leq 3)$



events out here are rare => we may not get a lot of 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.

place more "importance" on the rare events than we normally would by upweighting their chance of occurence and the correcting our estimator.

For very rare events, 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 $g(X)$,

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

where $X \sim f$.

Importance part.

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 covers the support of f.

downweighting back to what we need w/ f.

$$\theta = E[g(X)] = \int_{\mathcal{X}} g(x) \frac{f(x)}{\phi(x)} \phi(x)\,dx = E\left[g(X)\frac{f(X)}{\phi(X)}\right] \quad \text{where } X \sim \phi$$

upweighting rare events

importance weight

ϕ is called the importance sampling function

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

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).

Let 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 the die m times
② Use a point estimator of $P(X=1)$ as proportion of ones in the sample.

This is a MC approach.
- You are sampling $X_1, \dots, X_m$ from f.
- estimate $P(X=1)$
$\frac{1}{m}\sum_{i=1}^{m} I(X_i = 1)$.

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

Why?

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

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

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

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

Expected # of 1s in m rolls:

$E\left[\sum_{i=1}^{m} Y\right] = \sum_{i=1}^{m} EY = \frac{m}{6}$

$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ ← proportion of 1s in our sample.

$$E(\hat{\theta}) = E\left(\frac{1}{m}\sum_{i=1}^{m} Y_i\right) = \frac{1}{m} E\left(\sum_{i=1}^{m} Y_i\right) = \frac{1}{m} \cdot \frac{m}{6} = \frac{1}{6}.$$

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

relative measure of variability in a r.v. used a lot in chemistry and physics.

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

If we want a CV of 5%, then what value of m do we need?

$$CV\left(\frac{\sum Y_i}{m}\right) = \frac{\sqrt{\text{Var}\left(\frac{\sum Y_i}{m}\right)}}{E\left(\frac{\sum Y_i}{m}\right)} = \frac{\sqrt{5/36m}}{1/6} \overset{\text{want}}{=} .05$$

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

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

want to lower this.

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 probability of rolling a 1 to 0.5, but now we aren't sampling from the target dsn (a fair die roll).

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 $Y_i = \begin{cases} \frac{1}{3} & \text{if } X=1 \\ 0 & \text{otherwise.} \end{cases}$

Then the expectation of the sample mean of $\frac{\sum Y_i}{m}$ is

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

But the variance is

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

$$EY^2 = \left(\frac{1}{3}\right)^2 \cdot \frac{1}{2} = \frac{1}{18}$$

So to acheive a CV of 5%, we would need only

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

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

This die rolling example is successful because our importance sampling function (rolling a die w/ 3 ones) is used to over sample a portion of the state space that receives lower probability under the target dsn. and importance sampling corrects that bias.

2.3 Choosing ϕ

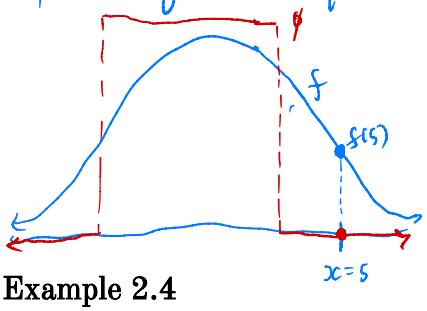
In order for the estimators to avoid excessive variability, it is important that $f(x)/\phi(x)$ is bounded and that ϕ has heavier tails than f.

more efficiency

importance weight

If this requirement is not met, then some importance weights will be huge, (we will increase variance of estimators)

Example 2.3

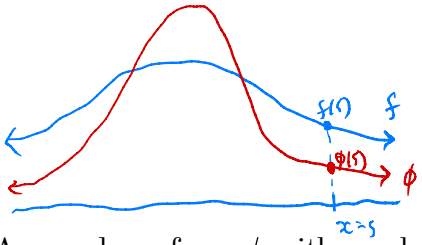


If we ignore requirement that $\phi(x) > 0$ when $f(x) > 0$,

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

Then $\frac{f(5)}{\phi(5)}$ will be large

Thus, $x=5$ draw will have a large weight
$\Rightarrow$ make integral approximation 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.

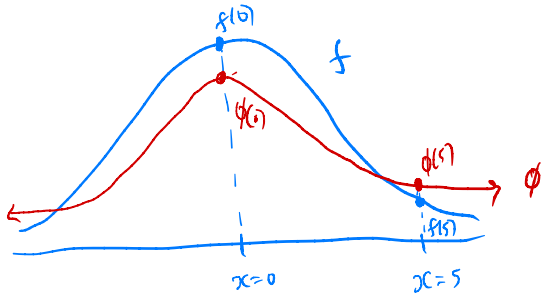
what we don't want!

Strategy – choose the function ϕ so that $f(x)/\phi(x)$ is large only when $g(x)$ is small. (want $g(x)\frac{f(x)}{\phi(x)}$ to be close to constant)

use visualization!

Example 2.5

If we select appropriate ϕ



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

$f(5)/\phi(5)$ 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.

$$\theta = \int h(x)dx = \int_{-\infty}^{\infty} g(x) f(x) dx = Eg(X) \quad \text{where } X \sim f.$$

b. For **importance sampling** only,

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

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

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

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

$\frac{f(X_i)}{\phi(X_i)}$ 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 ① and importance sampling with two different importance sampling distributions, ϕ. ② a) and b).

STEP 1 Derive things.

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

$$\theta = \int_0^1 \frac{e^{-x}}{1+x^2} dx = \int_0^1 \frac{1}{1+x^2} e^{-x} dx$$

looks like $f(x)$

$$= \int_0^\infty \frac{1}{1+x^2} 1(x \leq 1) e^{-x} dx = E\left[\frac{1}{1+X^2} \cdot 1(X \leq 1)\right]$$

where $X \sim Exp(1)$.

Need: $E[g(X)]$ where $X \sim Exp(1)$

$$\Rightarrow \int_0^\infty g(x) e^{-x} dx$$

Compare to $\int_0^1 \frac{e^{-x}}{1+x^2} dx$

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

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

Option ① MC Integration (no 1b step)

Option ② Importance sampling with

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

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

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

Check this

you can check this integrates to 1 and is always ≥ 0 ⇒ valid pdf.

$$\phi_b(x) \propto e^{-x} \quad \text{for} \quad 0 \le x \le 1.$$

Need
1) integrate to 1.
2) $\phi_B(x) \ge 0 \quad \forall x \in \mathbb{R}$. ✓

$$\int_0^1 c\, e^{-x} dx = 1 \quad \text{need.}$$

$$c \int_0^1 e^{-x} dx = 1$$

$$c = \frac{1}{\int_0^1 e^{-x} dx}$$

$$\int_0^1 e^{-x} dx = -e^{-x} \Big|_0^1 = 1 - e^{-1}$$

$$\Rightarrow \phi_b(x) = \frac{e^{-x}}{1 - e^{-1}} \qquad x \in [0,1].$$

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

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

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

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

This is always true

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

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

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

$$= E\left[\frac{1-e^{-1}}{1+X^2}\right]$$ This is always true.

STEP 2 Make a plan.

Option 1:

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

Option 2a:

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

Option 2b

① Sample $m = 1000$, $X_1, \dots, X_m \overset{iid}{\sim} \phi_b$ using the inverse transform method.

a) Sample $U_1, \dots, U_m \overset{iid}{\sim} \text{Uniform}(0,1)$.

b) Set $X_i = F^{-1}(U_i)$.

$$F_{\phi_b}(x) = \int_0^x \frac{e^{-y}}{1-e^{-1}}\,dy = -\frac{e^{-y}}{1-e^{-1}}\Big|_0^x = \frac{1}{1-e^{-1}}\left[-e^{-x} - (-e^{-0})\right]$$

$$= \begin{cases} 0 & x < 0 \\ \dfrac{1-e^{-x}}{1-e^{-1}} & x \in [0,1) \\ 1 & x \geq 1 \end{cases}$$

$$U = F_{\phi_b}(x) = \frac{1-e^{-x}}{1-e^{-1}}$$

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

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

$$-x = \log\left(1 - U(1-e^{-1})\right)$$

double check this.

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

② Compute $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \frac{1-e^{-1}}{1+X_i^2}$

Want these to be constant across x values. This will give us the lowest variance in the estimator.

Which will be the best? How can we compare them?

Can look at $\frac{h(x)}{f(x)} \propto \frac{f(x)g(x)}{f(x)}$, $\frac{h(x)}{\phi_a(x)} = \frac{f(x)g(x)}{\phi_a(x)}$, $\frac{h(x)}{\phi_b(x)} = \frac{f(x)g(x)}{\phi_b(x)}$

Recall: could use a general function $\gamma(x)$

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i) \qquad X_i \overset{iid}{\sim} f \qquad \text{where in importance sampling, } \gamma(x) = \frac{g(x)\cdot f(x)}{\phi(x)}.$$

$$\Rightarrow \operatorname{Var}\hat{\theta} = \frac{\operatorname{Var} g(X_i)}{m} \quad \text{where} \quad \operatorname{Var} g(X_i) = E\left[\left(g(X_i) - Eg(X_i)\right)^2\right]$$

$$\widehat{\operatorname{Var}}(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(g(X_i) - Eg(X_i)\right)^2, \quad X_i \overset{iid}{\sim} f.$$

$\frac{1}{m}$ can use instead $\frac{1}{m-1}$

$= \theta$

$\Rightarrow$ estimate using $\hat{\theta}$

$$\Rightarrow \widehat{\operatorname{Var}}(\hat{\theta}) = \frac{\frac{1}{m}\sum_{i=1}^{m}\left(g(X_i) - \hat{\theta}\right)^2}{m}$$

importance sampling

$$\widehat{\operatorname{Var}}(\hat{\theta}) = \frac{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{g(X_i)f(X_i)}{\phi(X_i)} - \hat{\theta}\right)^2}{m}$$

Want:

$\hat{\theta}$

$\widehat{\operatorname{Var}}(\hat{\theta})$

} each method 1, 2a, 2b.

Compare to θ using integrate( ) function.