# 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.

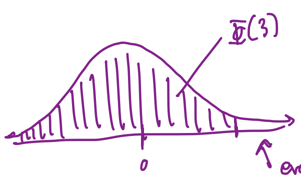
↳ more efficient

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, so may not get many values out here to occur by chance ⇒ not many values to contribute to our 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

$= \{x : f(x) > 0\}$ $X \sim f$

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 $\phi$ includes support of $f$.

$$\theta = E[g(X)] = \int_{\mathcal{X}} g(x) \frac{f(x)}{\phi(x)} \phi(x)\, dx$$
$$= E\left[g(Y) \frac{f(Y)}{\phi(Y)} \mathbb{I}(Y \in \mathcal{X})\right]$$

$Y \sim \phi$

$\phi$ is called the importance sampling function

$\phi$ MUST be a density (i.e. must integrate to 1 and always $\geq 0$).

An estimator of $\phi$ is given by the *importance sampling algorithm*:

1. Sample $X_1, \ldots, X_m \sim \phi$
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 $\phi$
2. easy to evaluate $f$ (even if 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

② Estimate of $P(X=1)$ could be proportion of ones in sample.

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

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

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

$$EY = \sum_{i=1}^{6} \mathbb{I}(Y=1) \cdot \frac{1}{6} = \frac{1}{6}$$

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

Expected # of 1's in $m$ rolls:

$$E\left[\sum Y_i\right] = \sum_{i=1}^{m} EY_i = \frac{m}{6}.$$

Properties of resulting estimator → Proportion of 1's in sample:

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

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

We can consider the "coefficient of variation" $CV[X] = \frac{\sqrt{\text{Var}[X]}}{E[X]}$. ← relative measure of variation used in chemistry and physics, etc.

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

If we want coefficient of variation = 5%, then

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

$$\frac{5}{36m} = \left[\frac{1}{6} \cdot (0.05)\right]^2$$

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

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

This increases prob. of rolling a 1 to 0.5 but now we aren't drawing from the target distribution (a fair die).

Now
$$P(X=1) = 1/2$$
$$P(X=2) = P(X=3) = 0$$
$$P(X=4) = P(X=5) = P(X=6) = 1/6$$

We can correct this by

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

Then the expectation of the sample mean of $Y$

$$E\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] = \frac{1}{m}\sum_{i=1}^{m} EY_i = EY_1 = \frac{1}{3}\cdot\frac{1}{2} + 0\left[0+0+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right] = \frac{1}{6} \checkmark$$

But the variance is

$$E(Y^2) = \frac{1}{3^2}\cdot\frac{1}{2} = \frac{1}{18}$$

$$\text{Var}\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] = \frac{1}{m^2}\sum \text{Var}\, Y_i = \frac{1}{m}\text{Var}\, Y_1 = \frac{1}{m}\left[\frac{1}{18} - \left(\frac{1}{6}\right)^2\right] = \frac{1}{36m}$$

So to achieve CV of 5% we would only need

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

$$\vdots$$

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

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

## 2.3 Choosing $\phi$

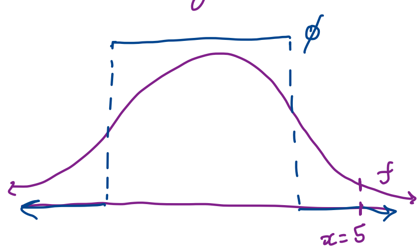
In order for the estimators to avoid excessive variability, it is important that $\boxed{f(x)/\phi(x)}$ is (1) bounded and that (2) $\phi$ has heavier tails than $f$.

importance weights

If these requirements aren't met, then some importance weights will be huge.

**Example 2.3**

If we ignore this require that $\phi(x) > 0$ when $f(x) > 0$.

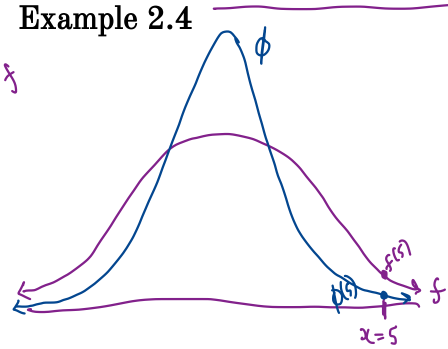



Then $\frac{f(5)}{\phi(5)} = \frac{f(5)}{0}$ unbounded!

AND we can never draw $x=5$ from $\phi$.

**Example 2.4**

select $\phi$ w/ lighter tails than $f$




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

Thus $x=5$ draw will have large weight associated with it

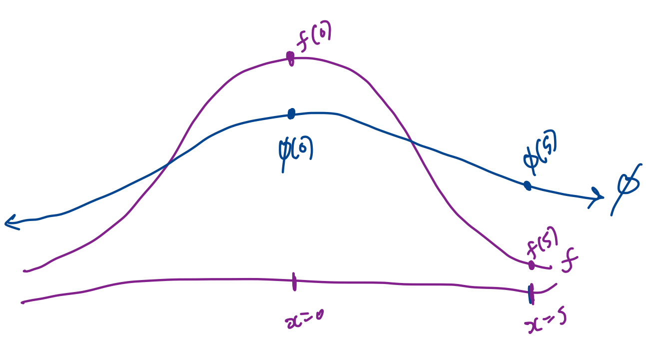
Integral approx will be poor (very high variability).

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

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

**Example 2.5**

If we select an appropriate $\phi$,




$\frac{f(0)}{\phi(0)}$ will be large but

not concerned because we still have a lot of draws here from $\phi$.

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

Rare draw from $f$ will have small weights.

The importance sampling estimator can be shown to converge to $\theta$ under the SLLN so long as the support of $\phi$ includes all of the support of $f$.

## 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.

$$\theta = \int h(x)dx = \int g(x)f(x)\,dx = E[g(X)], \quad X \sim f$$

b. For **importance sampling** only,

smart

Find an appropriate $\phi$ to rewrite $\theta$ as an expectation with respect to $\phi$.

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

make some plots? think about importance weights...

$$\theta = \int g(x)f(x)dx = \int g(x)\frac{f(x)}{\phi(x)}\cdot\phi(x)\,dx = E\left[g(Y)\frac{f(Y)}{\phi(Y)}\right], \quad Y \sim \phi.$$

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

- For **Monte Carlo integration**

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

- For **importance sampling**

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

     importance weights.

**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 ① MC integration and ② importance sampling with two different importance sampling distributions, $\phi$. a) and b)

$\boxed{\text{STEP 1}}$ derive things.

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

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

Option ① MC integration (no step 1b).

② Importance sampling with

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

b) $\phi \sim \text{Exp}(1)$ truncated/rescaled to have support $0 \leq x \leq 1$.

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

← you can check this is a valid pdf because it is always $\geq 0$ & integrates to 1.

Want $\phi_b(x) \propto e^{-x}$ for $x \in [0,1]$.

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

$\int_0^1 \phi_b(x) dx = 1.$

$\Rightarrow$ solve for c.

$\boxed{\text{STEP 1b}}$ Option 2a) $\phi_a(x) = 1 \quad 0 \leq x \leq 1$

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

$$= \int_0^\infty \frac{1}{1+x^2} \cdot \mathbb{I}(x \leq 1) \cdot \frac{e^{-x}}{1} \cdot 1 \cdot \mathbb{I}[0 \leq x \leq 1] dx.$$

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

$$= E\left[\frac{1}{1+Y^2} e^{-Y}\right] \quad Y \sim \text{Unif}(0,1).$$

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

$$\theta = E_f[g(X)] = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{\phi(x)} \phi(x) dx = \int_0^\infty \frac{1}{1+x^2} \mathbb{I}(x \leq 1) \cdot \frac{\cancel{e^{-x}}}{\cancel{e^{-x}}/1-e^{-1}} \cdot \frac{e^{-x}}{1-e^{-1}} \mathbb{I}[0 \leq x \leq 1] dx$$

Want $\phi_b(x) = ce^{-x}$ $x \in [0,1]$.

$$\int_0^1 \phi_b(x) = \int_0^1 ce^{-x}\,dx = c\int_0^1 e^{-x}\,dx$$

$$= c\left[-e^{-x}\right]_0^1 = c\left[-e^{-1} - (-e^{-0})\right] = c(1-e^{-1}).$$

$$\Rightarrow \quad c = \frac{1}{1-e^{-1}}$$

$$= E\left[\frac{1+e^{-1}}{1+Z^2}\mathbb{I}(Z\le 1)\right]$$

$Z \sim$ truncated Exp(1).

$\boxed{\text{STEP 2}}$ Make a plan

Option 1

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

Option 2a

1. Sample $Y_1, \dots, Y_m \sim \text{Unif}(0,1)$
2. Compute $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{1+Y_i^2}\underbrace{\mathbb{I}[Y_i \le 1]}_{\text{always be true in this case.}} e^{-Y_i}\right]$

Option 2b:

1. Sample $Z_1, \dots, Z_m \sim \text{truncated}_{[0,1]}\, \text{Exp}(1)$

How to do this step? Let's try inverse cdf method!

2. Compute $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{1+Z_i^2}\mathbb{I}(Z_i \le 1)\cdot(1-e^{-1})\right]$

First need cdf

$$F_{\phi_b}(x) = \int_0^x \frac{e^{-y}}{1-e^{-1}}\,dy = -\frac{e^{-y}}{1-e^{-1}}\Bigg|_0^x = \begin{cases} 0 & x<0 \\ \frac{1-e^{-x}}{1-e^{-1}} & \text{for } x\in[0,1]. \\ 1 & x>1 \end{cases}$$

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

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

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

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

Option 2b:

1. Sample $Z_1, \dots, Z_m \sim$ truncated$_{[0,1]}$ Exp(1)

a) Sample $U_1, \dots, U_m \sim Unif(0,1)$
b) set $Z_i = -\log(1 - U_i(1-e^{-1}))$
for $i = 1, \dots, m$.

2. Compute $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \left[ \frac{1}{1+Z_i^2} \mathbb{I}(Z_i \le 1) \cdot (1 - e^{-1}) \right]$

$\boxed{\text{STEP 3}}$ Do it in R.

Which one will be best? (lowest variance)

Can compare

$\frac{f(x)}{\phi(x)}$ compared to where $g(x)$ is large/small.

$\downarrow$

look $f(x) \cdot g(x)$. Compared to $f$, $\phi_a$, $\phi_b$

$\frac{\cancel{f}(x) \cdot g(x)}{\cancel{f}(x)}$, $\frac{f(x)\, g(x)}{\phi_a(x)}$, $\frac{f(x) g(x)}{\phi_b(x)}$

want these to be as constant as possible.
this will give us lowest variance.