2 Importance Sampling

↗ more efficient

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 estimation.

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

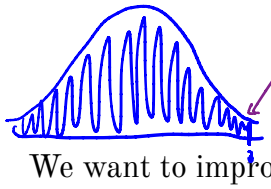
2.1 The Problem

(why would our simple MC estimators have high variance to begin with?)

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 $\Phi(3)$. (HW6)

$\Phi(3) = P(X \leq 3)$, $X \sim N(0,1)$.



events out here are rare ⇒ depending on what proposal distribution he choose in our MC estimator, we may not get many samples out (to plug in to our sample mean step ②)

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

$= \{x : g(x) > 0\}$

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

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

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

$\{x : \phi(x) > 0\} \supseteq \{x : f(x) > 0\} = \mathcal{X}$

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

$$= E\left[g(Y) \cdot \frac{f(Y)}{\phi(Y)}\right] \text{ for } Y \sim \phi$$

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

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

ϕ MUST be a density! (integrate to 1 & always ≥ 0).

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 (and ϕ) even if it is not easy to sample from f.

X = result of rolling one 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. $P(X=1)$.

We could

① Roll a 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.

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

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

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

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

Expected # of 1's in m rolls

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

Proportion of 1's (estimator) in the sample:

$$E\left[\frac{\sum Y_i}{m}\right] = \frac{1}{m}\cdot\frac{m}{6} = \frac{1}{6}$$ unbiased estimator

$$\text{Var}\left[\frac{\sum Y_i}{m}\right] \overset{\text{indep.}}{=} \frac{1}{m^2}\sum \text{Var}\,Y_i \overset{\text{iid}}{=} \frac{1}{m}\text{Var}\,Y_1 = \frac{1}{m}\cdot\frac{5}{36}$$ ← variance of our estimator

We could consider the "coefficient of variation" $CV[X] = \frac{\sqrt{\text{Var}(X)}}{E(X)}$ ← relative measure of variation (chemistry, physics).

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

Say we want a CV of 5%. Want $\frac{\sqrt{5/(36m)}}{1/6} = .05$ ... solve for m ... $m = 2000$ rolls!

goal: Can we do better? i.e. use less rolls to get the same CV = .05 with a different estimator.

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

This will increase the probability of rolling a 1 to 0.5, but now we are not sampling from target distribution. (a fair die).

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

Now:
$P(X=1) = \frac{1}{2}$
$P(X=2) = P(X=3) = 0$
$P(X=4) = P(X=5) = P(X=6) = \frac{1}{6}$

estimator: $\frac{1}{m}\sum_{i=1}^{m} Y_i$

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

But the variance:

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

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

So to achieve CV of 5% we would only need

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

$\therefore$ solve for m

$\Rightarrow m = 400$ rolls.

⇒ This die rolling example is successful is because an importance function (rolling die w/ 3 sides = 1) is used to over-sample a portion of the statespace that recieves lower probability under the target and then importance weighting to correct the bias.

2.3 Choosing ϕ

(0) support of ϕ must include support of f.

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.

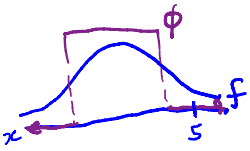
target / importance

importance weights

If this requirement is not met, some importance weight $\frac{f(x)}{\phi(x)}$ will be huge.

(0) **Example 2.3**

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

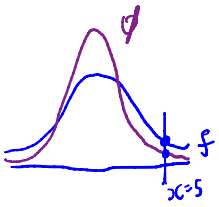


Then $\frac{f(5)}{\phi(5)} = \frac{f(5)}{0}$ unbounded! → breaks requirement (1).

And, we can't draw $x = 5$ from ϕ.

(2) **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 $x = 5$ (few) draw has large weight ⇒ approx will be poor.

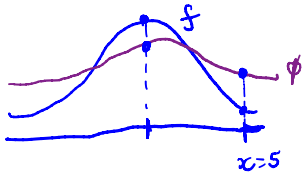
A rare draw from ϕ with much higher density under f than under ϕ will receive a huge weight and inflate the variance of the estimate.

opposite of our goal.

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,



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

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

We care about something out in the tails ⇒ this makes sense to try.

parameter of interest

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 (Naive) 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.

$$\theta = \int h(x)dx = \int g(x)f(x)dx = E[g(x)] \text{ w.r.t. } X \sim f.$$

if we want to improve on step a)

b. For **importance sampling** only,

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

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

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** (naive)
  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 weights)

Step 3 Program it.

may also choose to approximate $\text{Var}(\hat{\theta})$.

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, ϕ. a) & b)

STEP 1 Derive things.

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

$\Rightarrow$ find $g(x)$ s.t. $\theta = E\,g(X) = \int_0^\infty g(x) f(x)\,dx$. $\Rightarrow g(x) = \frac{1}{1+x^2} \mathbb{I}[x \leq 1]$

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

Option ① MC integration w/ $X \sim Exp(1)$ (no 1b step).

Option ② Importance sampling w/

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

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 by making sure it integrates to 1 ($\phi_b(x) \geq 0$ ✓)

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

$\theta = E_f[g(X)] = \int_0^\infty g(x) \frac{f(x)}{\phi(x)} \phi_a(x)\,dx$

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

$= E\left[\frac{1}{1+X^2}\mathbb{I}(X \leq 1)\,e^{-X}\right]$, $X \sim \phi_a = \text{Uniform}(0,1)$.

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

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

$= E\left[\frac{1+e^{-1}}{1+X^2}\mathbb{I}(X \leq 1)\right]$, $X \sim \phi_b$

STEP 2 Make a plan.

Option 1:

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

Option 2a:

1. Sample $X_1, \dots, X_m$ from $\text{Unif}[0,1]$.
2. $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1+X_i^2} e^{-X_i} \mathbb{I}[X_i \leq 1]$

This will always hold.

Option 2b:

1. Sample $X_1, \dots, X_m$ from ϕ_6

Note this ϕ_6 is not a named distribution. => We need to sample from it.
Inverse transform!