Unit 2: Monte Carlo Integration and Variance Reduction

Chapter 6 in “Statistical Computing with R”

Anna Ly

Department of Mathematical and Computational Sciences
University of Toronto Mississauga

January 19, 2026

Overview

1. Introductory Monte Carlo Methods
2. Empirical CDF
3. Standard Error and CLT
4. Bonus Monte Carlo Problems
5. Importance Sampling
6. Antithetic Variables

Introductory Monte Carlo Methods

Monte Carlo Methods

- Two major classes of numerical problems that arise in statistical inference are **optimization problems** and **integration problems.**
- Thus, we are often led to consider numerical solutions.
- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.

Simple Monte Carlo Integration

Let $g(x)$ be a function defined on some interval $[a, b] \subseteq \mathbb{R}$. Suppose we want to compute:

$$\int_a^b g(x)dx, \quad \text{(Assuming this integral exists).}$$

Recall that if X is a random variable with density $f_X(x)$, then the mathematical expectation of the random variable $Y = g(X)$ is:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Monte Carlo Methods (Probability Review)

Convergence in Probability

For a sequence of random variables $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we say the sequence converges in probability to another variable, X, on that space $X_n \xrightarrow{p} X$ if

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ such that } \forall n > N, \mathbb{P}\{|X_n - X| > \epsilon\} < \epsilon$$

Alternative version:

$$\forall \epsilon > 0, \lim_{n \to \infty} \mathbb{P}\{|X_n - X| > \epsilon\} = 0$$

Monte Carlo Methods (Probability Review)

Weak Law of Large Numbers

The weak law of large numbers states that given a collection of independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n$ that all arise from the same distribution X with finite mean ($\mathbb{E}[X] < \infty$), the sample mean **converges in probability** to the expected value. That is,

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{p} \mathbb{E}[X]$$

Technically, we could also use the strong law of large numbers, in which we'd need to discuss almost sure convergence. I'll explain it for completeness, but it is not necessary for examinations.

Monte Carlo Methods (Probability Review?)

Almost Sure Convergence

A sequence of random variables $(X_n, n \geq 1)$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ converges almost surely (a.s.) to a random variable X on that space, denoted $X_n \overset{a.s.}{\to} X$ if $\mathbb{P}\{\lim_{n\to\infty} X_n \text{ converges to } X\} = 1$. Equivalently,

$$\mathbb{P}\{\forall\epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, |X_n - X| < \epsilon\} = 1,$$

or,

$$\mathbb{P}\{\exists\epsilon > 0, \forall N \in \mathbb{N} \text{ s.t. } \exists n > N, |X_n - X| \geq \epsilon\} = 0$$

- Almost sure convergence implies convergence in probability.
- If you like theoretical probability theory, I recommend starting out by reading Knowing the Odds by Walsh (easy read... for a probability textbook).

Rant (Probability Review?)

- Convergence in probability and almost sure convergence look similar. In layman's terms, how do they differ, and how do we see that almost sure convergence is stronger?
- There are good answers in this stackexchange post if you are curious.
- To pass this class, you don't need to know the difference. But it is worth exploring if you will be taking probability theoretic courses in the future.

Monte Carlo Methods (Probability Review?)

Strong Law of Large Numbers

Given a collection of independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n$ that all arise from the same distribution X with finite mean ($\mathbb{E}[X] < \infty$), the sample mean **converges almost surely** to the expected value. That is,

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X]$$

Monte Carlo Methods

Simple Monte Carlo (MC) Estimator

Suppose we want to integrate $\theta = \int_0^1 g(x)dx$. If $X_1, X_2, \dots, X_n$ is a random sample of *Uniform*$(0,1)$ then by the **strong law of large numbers**:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \rightarrow \mathbb{E}[g(X_i)] \;\; a.s.$$

(Where $X_i \sim Uniform(0,1)$ for $i \in \{1, 2, \dots, n\}$).

Example

Compute a MC estimate of:

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

and compare the estimate with the exact value.

Solution. Solving the exact value:

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

Algorithm:

1. Generate $x_i \sim Uniform(0,1)$ for $i = 1, 2, \ldots, n$.
2. Calculate $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} e^{-x}$.

Adjustments for Finite Intervals: Method 1

- To compute $\int_a^b g(x)dx$, make a change of variables so that the limits of integration are 0 to 1.
- To do this, consider the transformation $y = (x - a)/(b - a)$.
- It follows that,
$$\int_a^b g(x)dx = \int_0^1 g(y(b-a)+a)(b-a)dy$$
- Hence consider the function $h(x) = g(y(b-a)+a)(b-a)$. Then you can follow the steps as before to compute $\int_0^1 h(x)dx$.

Adjustments for Finite Intervals: Method 2

- If $X \sim \mathit{Uniform}(a, b)$ then

$$f_X(x) = \frac{1}{b-a}, \quad a < x < b, \quad \mathbb{E}[g(X)] = \int_a^b \frac{1}{b-a} g(X) dx$$

- Notice that:

$$\int_a^b g(x)dx = (b-a)\int_a^b \frac{1}{(b-a)} g(x)dx = (b-a)\mathbb{E}[g(X)]$$

- Thus, if $X_1, X_2, \ldots, X_n$ is a random sample of $\mathit{Uniform}(a, b)$, we can use the strong law of large numbers to estimate $\mathbb{E}[g(X)]$.

Example

Use both methods to compute an MC estimate of:

$$\theta = \int_2^4 e^{-x} dx$$

and compare the estimate with the exact value.

Solution. The true value is:

$$\int_2^4 e^{-x} dx = e^{-2} - e^{-4}$$

For method 1, we need to consider:

$$h(x) = (4-2)\exp\{-y(4-2)+2\} = 2e^{-2y+2}$$

Monte Carlo Methods

The algorithm for the first method:

1. Generate $x_i \sim \mathit{Uniform}(0, 1)$ for $i = 1, 2, \ldots, n$.
2. Calculate $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} 2e^{-2x_i+2}$.

The algorithm for the second method (personal preference):

1. Generate $x_i \sim \mathit{Uniform}(2, 4)$ for $i = 1, 2, \ldots, n$.
2. Calculate $\hat{\theta} = 2\frac{1}{n} \sum_{i=1}^{n} e^{-x_i}$.

Unbounded Interval Case

Use the Monte Carlo approach to estimate the standard Gaussian cdf:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Where $x > 0$. Try both methods and compare the results to `pnorm()`.

In general, when dealing with unbounded intervals, it's best to try to eliminate them. In this case, since we are given $x > 0$ we have:

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= 0.5 + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\end{aligned}$$

Remaining Solution. For the first method, we could use the substitution $y = t/x$ and $dt = xdy$ (we have $a = 0, b = x$). Then,

$$\theta = \frac{1}{\sqrt{2\pi}} \int_0^1 x \exp\{(xy)^2/2\} dy = \frac{1}{\sqrt{2\pi}} \mathbb{E}[xe^{-(xY)^2/2}]$$

Algorithm for method I:

- Generate $u_1, \dots, u_n \sim \mathit{Uniform}(0, 1)$.
- Compute $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} xe^{-(xu_i)^2/2}$.
- If $x \geq 0$, estimate $\Phi(x)$ by $0.5 + \hat{\theta}$. Otherwise, compute $\Phi(x) = 1 - \Phi(-x)$ (using the symmetry property of the Gaussian distribution.)

Monte Carlo Methods

Algorithm for method II:

- Generate $u_1, \ldots, u_n \sim \mathit{Uniform}(0, x)$.
- Compute $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x \frac{1}{\sqrt{2\pi}} e^{-(u_i)^2/2}$.
- If $x \geq 0$, estimate $\Phi(x)$ by $0.5 + \hat{\theta}$. Otherwise, compute $\Phi(x) = 1 - \Phi(-x)$ (using the symmetry property of the Gaussian distribution.)

Probability Review - Conditional Expectation

Let X be a random variable with pdf $f_X(x)$. We know that its expected value is $\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx$. However, what if we knew that the values of X are between some interval $[a, b]$?

Conditional Expectation

$$\mathbb{E}[X | a \leq X \leq b] = \frac{\mathbb{E}[X \mathbf{1}_{\{a \leq X \leq b\}}]}{\mathbb{P}(a \leq X \leq b)} = \frac{\int_a^b x f_X(x) dx}{\int_a^b f_X(x) dx}$$

Monte Carlo Methods

Example

Let Z be a standard Gaussian distribution. Compute a simple Monte Carlo estimate of $\mathbb{E}[Z| -2 \leq Z \leq 1]$.

Solution. Recall:

$$\mathbb{E}[Z| -2 \leq Z \leq 1] = \frac{\int_{-2}^{1} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz}{\int_{-2}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz}$$

Our plan of attack:

1. Do a Monte Carlo estimate for the numerator, $\int_{-2}^{1} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$.
2. Do a Monte Carlo estimate for the denominator, $\int_{-2}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$.
3. Divide part (1) by part (2).

Monte Carlo Methods

For part 1, note that:

$$\int_{-2}^{1} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = (1-(-2)) \int_{-5}^{5} \frac{z}{(1-(-2))} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 3\mathbb{E}\left[\frac{z}{\sqrt{2\pi}} e^{-z^2/2}\right].$$

Part 2 is similar:

$$\int_{-2}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = (1-(-2)) \int_{-5}^{5} \frac{1}{(1-(-2))} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 3\mathbb{E}\left[\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\right].$$

Monte Carlo Methods

The algorithm is as follows:

1. Generate $U_i \sim Uniform(-2, 1)$ for $i = 1, 2, \ldots, n$.
2. Compute $\widehat{\mathbb{E}}[Z\mathbf{1}_{\{-2 \leq X \leq 1\}}] = 3\frac{1}{n}\sum_{i=1}^{n} \frac{u_i}{\sqrt{2\pi}} e^{-u_i^2/2}$.
3. Compute $\widehat{\mathbb{P}}(-2 \leq Z \leq 1) = 3\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-u_i^2/2}$.
4. Deliver $\widehat{\mathbb{E}}[Z| -2 \leq Z \leq 1] = \frac{\widehat{\mathbb{E}}[Z\mathbf{1}_{\{-2 \leq X \leq 1\}}]}{\widehat{\mathbb{P}}(-2 \leq Z \leq 1)}$.

Monte Carlo Methods

For many distributions that we care about, we cannot find the CDF in closed form:

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

For example, there are the Gaussian, Gamma, Beta, Logistic, etc. Note that,

$$\mathbb{E}[I(X \leq x)] = \int_{-\infty}^{\infty} I(X \leq x) f_Z(z)dz = \int_{-\infty}^{x} f_X(x)dx = \mathbb{P}(X \leq x) = F_X(x).$$

We can estimate $\mathbb{E}[I(X \leq x)]$ using the strong law of large numbers, i.e.,

$$\mathbb{E}[I\widehat{(X \leq} x)] = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

Monte Carlo Methods

Consider the following approach:

Hit or Miss (I guess they never miss, huh?)

- Generate a random sample $X_1, \ldots, X_n$ from the distribution X.
- For each observation X_i, compute:

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x, \\ 0 & X_i > x \end{cases}$$

- Compute $\widehat{F(x)} = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$

Example

Let $X \sim N(0,1)$. Use the hit or miss method to estimate $\mathbb{P}(X > 2)$ and $\mathbb{P}(0 < X < 1)$. Compare this value with the Rfunction `pnorm()`.

Solution. Consult the R codes.

Empirical CDF

Empirical

- What does “empirical” mean? According to Merriam-Webster, it means “originating in or based on observation or experience”.
- As its name implies, the empirical CDF is dependent on the observations themselves.
- You’ve also seen this function before in the “hit or miss” approach.

Order Statistics Review

- Suppose you had a random sample $X_1, \ldots, X_n$.
- Let $X_{1:n} = \min\{X_1, \ldots, X_n\}$ represent the smallest value, $X_{2:n}$ represent the second smallest value, ...
- Let $X_{n:n} = \max\{X_1, \ldots, X_n\}$ represent the largest value, $X_{n-1:n}$ represent the second largest value, ...
- You might have seen the notation $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ instead, which is commonly used in undergraduate textbooks.
- However, order statistics researchers tend to use the latter notation instead, and it will be the notation I will use.

Empirical CDF

Suppose we have a random sample $X_1, \dots, X_n$, and its order statistics are $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ where $X_{1:n}$ represents the lowest value and $X_{n:n}$ represents the highest. Then,

Empricial CDF (or Empirical Distribution Function)

$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$

Another way to write this is,

$$F_n(x) = \begin{cases} 0 & x < x_{1:n}, \\ \frac{i}{n} & x_{i:n} \leq x < x_{i+1:n}, \quad i \in \{1, 2, \dots, n-1\}, \\ 1 & x \geq x_{n:n} \end{cases}$$

Empirical CDF

Example

Using R, generate a standard Gaussian random sample. Create the following lines and put them on the same plot:

- The empirical cumulative distribution function (ECDF).
- The standard Gaussian cumulative distribution function.

Solution. Consult the R codes.

Empirical CDF

More About F_n (Part 1)

For a fixed t, what distribution is $I(X_i < t)$?

Solution. Given that indicator functions denote binary outcomes such that 1 tends to correspond to a success condition and 0 corresponds to a failure condition, $I(X_i < t) \sim Bernoulli(p)$. Again, p corresponds to the probability that the success condition occurs, which is just $\mathbb{P}(X_i < t) = F(t)$.

Empirical CDF

More About F_n (Part 2)

What is the distribution of $V = nF_n(t)$? What is the mean and variance?

Solution.

$$V = \sum_{i=1}^{n} I(X_i \leq x) \sim Binomial(n, F(t))$$

Since Binomial represents the sum of binary outcomes (Bernoulli). Thus, the mean and variance is:

$$\mathbb{E}[V] = nF(t), \quad \mathbb{V}ar[V] = nF(t)[1 - F(t)]$$

Empirical CDF

More About F_n (Part 3)

Show that $F_n(t)$ is an unbiased estimator for $F(t)$.

Solution.

$$\mathbb{E}[nF_n(t)] = n\mathbb{E}[F_n(t)], \;\; \mathbb{E}[nF_n(t)] = nF(t)$$
$$\Rightarrow \;\; n\mathbb{E}[F_n(t)] = nF(t) \;\; \Rightarrow \;\; \mathbb{E}[F_n(t)] = F(t).$$

More About F_n (Part 4)

What is the variance of $F_n(t)$?

Solution.

$$\mathbb{V}ar[nF_n(t)] = nF(t)[1 - F(t)], \;\; \mathbb{V}ar[nF_n(t)] = n^2\mathbb{V}ar[F_n(t)],$$
$$\Rightarrow \;\; nF(t)[1 - F(t)] = n^2\mathbb{V}ar[F_n(t)] \;\; \Rightarrow \;\; \mathbb{V}ar[F_n(t)] = \frac{F(t)[1 - F(t)]}{n}$$

Empirical CDF

More About F_n (Part 5)

For what value of $F(t)$ does the maximum variance occur for $F_n(t)$?

Solution. Let $h(p) = \frac{p(1-p)}{n} = \frac{p}{n} - \frac{p^2}{n}$ (where $p := F(t)$). Then,

$$h'(p) = \frac{1}{n} - \frac{2p}{n} \overset{\text{set}}{=} 0 \quad \Rightarrow \quad p = \frac{1}{2}$$

$h''(p) = \dfrac{-2}{n} < 0$ (unless I mention otherwise, you don't need to take the second derivative.)

Hence the maximum variance occurs when $F(t) = \frac{1}{2}$, which means the maximum variance is $\mathbb{V}ar[F_n(t)] = \frac{1/2(1-1/2)}{n} = \frac{1}{4n}$.

Standard Error and CLT

Standard Error of $\hat{\theta}$

Suppose we want to compute the standard error for $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$. Earlier, we claimed that if we had a random sample $X_1, \ldots, X_n \sim \mathit{Uniform}(0, 1)$ then by the SLLN we have:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \to \mathbb{E}[g(X_i)] \ \ a.s.$$

Variance for $\hat{\theta}$

Find the variance of $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$. Assume the variance of $\mathbb{V}ar[g(X_i)] = \sigma^2$.

Solution.

$$\mathbb{V}ar[\hat{\theta}] \overset{\text{i.i.d}}{=} \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar[g(X_i)] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Standard Error of $\hat{\theta}$

Estimating Variance and Standard Error for $\hat{\theta}$

How can we estimate $\mathbb{V}ar[g(X_i)] = \sigma^2$? Use this to estimate $\mathbb{V}ar[\hat{\theta}]$ and the corresponding standard error.

Solution. Assume random sample $X_1, \ldots, X_n \sim Uniform(0,1)$, then by the SLLN:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[g(x_i) - \overline{g(x)}\right]^2 \to \mathbb{E}\left[\left\{g(X_i) - \overline{g(X)}\right\}^2\right] = \mathbb{V}ar[g(X_i)] \ \ a.s.$$

Thus,

$$\widehat{\mathbb{V}ar[\hat{\theta}]} = \frac{\hat{\sigma}^2}{n} = \frac{\sum_{i=1}^{n}\left[g(x_i) - \overline{g(x)}\right]^2}{n^2}$$

$$\widehat{SE[\hat{\theta}]} = \sqrt{\widehat{\mathbb{V}ar[\hat{\theta}]}} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left[g(x_i) - \overline{g(x)}\right]^2}$$

Probability Review?

Weak Convergence

If $(S, \mathcal{T})$ is a nice topological space (a Polish space, or a complete separable measure space) and $\mu_1, \mu_2, \mu_3, \ldots$ are probability measures on this space then we say that μ_n converges weakly to μ, $\mu_n \stackrel{w}{\to} \mu$ if:

$$\int f d\mu_n \to \int f d\mu$$

For all bounded, continuous functions $f : S \to \mathbb{R}$.

Probability Review?

Convergence in Distribution

Given random variables $X_1, X_2, \ldots$ (not necessarily on the same probability space), we say X_n converges in distribution to X, denoted $X_n \xrightarrow{D} X$ if the probability distribution of X_n converges weakly to that of X.

Equivalently, for all bounded continuous functions f,

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(x)]$$

In layman's terms, convergence in distribution just means that the CDFs converge.

Central Limit Theorem of $\hat{\theta}$

We can use the CLT to help create confidence intervals for our Monte Carlo estimates.

Central Limit Theorem (CLT)

As $n \to \infty$,

$$\frac{\hat{\theta} - \mathbb{E}[\hat{\theta}]}{\sqrt{\mathbb{V}ar(\hat{\theta})}} \xrightarrow{d} N(0, 1)$$

Where $\xrightarrow{d}$ represents "convergence in distribution".

This implies:

- Hence, if n is sufficiently large, $\hat{\theta}$ is approximately Gaussian with mean $\hat{\theta}$.
- We can use the Gaussian distribution to apply error bounds on the Monte Carlo estimation.

Central Limit Theorem of $\hat{\theta}$

Example

Estimate the hit-or-miss estimators for $\mathbb{P}(Z < 2)$ and $\mathbb{P}(Z < 2.5)$ and construct approximate 95% confidence intervals for these estimates.

Solution. Consult the R codes.

Bonus Monte Carlo Questions

Bonus Monte Carlo Questions

- From Wikipedia: "**Monte Carlo methods**, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle."
- Hence, a lot of theoretical statisticians and probability theorist will use Monte Carlo methods; it's simply just using pseudo-randomness to simulate outcomes.
- There are many different ways to use Monte Carlo methods; we'll give two extra problems. But most people use these to validate their results.
- Pretty much everyone in my statistics cohort used this for their thesis, except the ones doing pure measure theoretic probability theory.

Bonus Monte Carlo Questions

Example

Derive a Monte Carlo estimator of S using the Geometric distribution as defined below, and then approximate a value for S:

$$S = \sum_{i=0}^{\infty} (i^2 + 2)^{-5} 5^{-i}, \quad i \in \{1, 2, \dots\}$$

Solution. Recall, for a geometric distribution, $\mathbb{P}(X = i) = (1 - p)^{i-1} p$ for $i = 1, 2, \dots$ So then, let $g(x)$ be a valid function, so:

$$\mathbb{E}[g(X)] = \sum_{i=1}^{\infty} g(i) \mathbb{P}(X = i) = \sum_{i=1}^{\infty} g(i)(1 - p)^{i-1} p = p \sum_{i=0}^{\infty} g(i + 1)(1 - p)^{i}$$

Bonus Monte Carlo Questions

Now just let $g(i+1) = (i^2+2)^{-5}$. Below now looks very similar to S; except we need to find a value for p:

$$\mathbb{E}[g(X)] = p\sum_{i=0}^{\infty}(i^2+2)^{-5}(1-p)^i$$

Let $(1-p)^i = \left(\frac{1}{5}\right)^i$. Then, $p = 4/5$. Thus, we have that:

$$\mathbb{E}[g(X)] = \frac{4}{5}\sum_{i=0}^{\infty}(i^2+2)^{-5}(5)^{-i} = \frac{4}{5}S \quad \Rightarrow \quad S = \frac{5}{4}\mathbb{E}[g(X)]$$

Let $X_1, \dots X_n \sim \mathit{Geometric}(p = 4/5)$. Then, using SLLN we can estimate:

$$\widehat{\mathbb{E}[g(X)]} = \frac{1}{n}\sum_{i=0}^{n}(i^2+2)^{-5}$$

Bonus Monte Carlo Questions

Thus, consider the following algorithm:

- Randomly sample $X_1, \ldots X_n \sim \mathit{Geometric}(p = 4/5)$.
- Compute $\widehat{\mathbb{E}[g(X)]} = \frac{1}{n} \sum_{i=0}^{n} (i^2 + 2)^{-5}$.
- Deliver $\hat{S} = \frac{5}{4} \frac{1}{n} \sum_{i=0}^{n} (i^2 + 2)^{-5}$.

Bonus Monte Carlo Questions

Example

- Consider the unit circle: it is centered at (0, 0) with radius 1.
- Suppose the random vector $\mathbf{U} = (U_1, U_2)$ is uniformly distributed in this square, where $U_i \sim \mathit{Uniform}(-1, 1)$ for $i = 1, 2$.
- Let us compute the probability that the random point (U_1, U_2) in the square is also within the inscribed circle. Technically:

$$\mathbb{P}(\mathbf{U} \in \mathit{Circle}) = \mathbb{P}(U_1^2 + U_2^2 \leq 1) = \frac{\pi}{4}$$

Now, write a Monte Carlo algorithm to estimate π.

Solution. Consider the following function (note that $\mathbf{u} = (u_1, u_2)$):

$$g(u_1, u_2) = \begin{cases} 1 & u_1^2 + u_2^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases} = \mathbf{1}_{\{u_1^2 + u_2^2 \leq 1\}}(\mathbf{u})$$

Bonus Monte Carlo Questions

Then,

$$\mathbb{E}[g(\mathbf{U})] = \int_{-1}^{1}\int_{-1}^{1} \mathbf{1}_{\{u_1^2+u_2^2\leq 1\}}(\mathbf{u})\frac{1}{4}du_1 d_2 = \mathbb{P}(U_1^2 + U_2^2 \leq 1) = \frac{\pi}{4}$$

So,

$$\hat{\pi} = 4 \times \widehat{\mathbb{E}[g(\mathbf{U})]}$$

The algorithm is:

- Generate U_1 and U_2 i.i.d from $Uniform(0,1)$ for n amount of times.
- Compute $\widehat{\mathbb{E}[g(\mathbf{U})]} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{U_1^2 + U_2^2 \leq 1\}$.
- Deliver $\hat{\pi} = 4 \times \widehat{\mathbb{E}[g(\mathbf{U})]}$.

Extra Proof about ECDF

Example

If $(X_n, n \geq 1)$ is an i.i.d. sequence from a distribution F, then the ecdf $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ converges almost surely to $F_X(x)$.

Solution. Fix $x \in \mathbb{R}$. We established earlier that $I(X_i \leq x) \sim Binomial(p)$ where $p := \mathbb{P}(X_i \leq x)$ and thus $\mathbb{E}[I(X_i \leq x)] = \mathbb{P}(X_i \leq x) = F_X(x)$. By SLLN, we know that:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \overset{a.s.}{\to} \mathbb{E}[I(X_i \leq x)] = F_X(x),$$

which is what we wanted to show.

Importance Sampling

Motivation for Importance Sampling

First, we'll start with some motivation. Remember, when we wanted to estimate the integral over an interval $[a, b]$:

$$\theta = \int_a^b g(x)dx = (b-a)\mathbb{E}[g(x)], \text{ where } X \sim \textit{Uniform}(a, b).$$

We let the estimator be:

$$\hat{\theta} = (b-a)\widehat{\mathbb{E}[g(x)]} = \frac{(b-a)}{n}\sum_{i=1}^{n} g(x_i).$$

The problem with this method is that this does not work for unbounded intervals and if $g(\cdot)$ is not uniform then the choice of $X \sim \textit{Uniform}(a, b)$ would be poor. Hence, we may consider another weight function.

Importance Sampling

Let X be a random variable with density $f(x)$ where $f(x) > 0$ on the set $\{x : g(x) > 0\}$. Let $Y = g(X)/f(X)$ also be a random variable. Then,

$$\int g(x)dx = \int \frac{g(x)}{f(x)} f(x)dx = \mathbb{E}\left[\frac{g(X)}{f(X)}\right] = \mathbb{E}[Y]$$

And we can estimate $\mathbb{E}[Y]$ using simple Monte Carlo integration. That is,

1. Let $X_1, \ldots, X_n$ be a random variable generated from the distribution of X with probability density $f(x)$.
2. $f(x)$ is called the **importance function**.
3. We estimate $\mathbb{E}[Y]$ via:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$$

Importance Sampling

Some notes regarding importance sampling:

- Ideally, we want $f(\cdot)$ to be "close" to $g(\cdot)$ (the variance is small if it's nearly a constant).
- $f(\cdot)$ should be easy to simulate.
- We want f/g to be bounded; see Question 1 of Practice Problems 6 (posted on Quercus).
- **Picking an importance function is difficult! You are not expected to propose a good one on the spot on a test or a quiz.**

Importance Sampling

- Typically in statistics we are interested in estimating expected values.
- Let X be a random variable with density $f(X)$. Suppose I wanted to estimate $\mathbb{E}[g(X)] = \int g(X)f(X)dx$.
- If the expected value is hard to compute, we can introduce a new function $\phi(x)$ and employ the same trick:

$$\mathbb{E}[g(X)] = \int g(x)f(x)dx = \int g(x)f(x)\frac{\phi(x)}{\phi(x)}dx = \mathbb{E}_\phi\left[\frac{g(X)f(X)}{\phi(X)}\right]$$

- The estimator of $\widehat{\mathbb{E}[g(X)]}$ is:

$$\widehat{\mathbb{E}[g(X)]} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{f(X_i)}{\phi(X_i)}$$

Importance Sampling

Variance of Importance Sampling

If $\theta = \int g(x)dx = \int \frac{g(x)}{f(x)} f(x)dx$ and we use the estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$ then compute the variance of $\hat{\theta}$.

Solution.

$$\begin{aligned}
\mathbb{V}ar(\hat{\theta}) &= \mathbb{V}ar\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}\right\} \\
&= \mathbb{V}ar\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left\{\frac{g(X_i)}{f(X_i)}\right\}
\end{aligned}$$

Importance Sampling

$$
\begin{aligned}
\mathbb{V}ar(\hat{\theta}) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(\left\{\frac{g(X_i)}{f(X_i)}\right\}^2\right) - \mathbb{E}\left(\left\{\frac{g(X_i)}{f(X_i)}\right\}\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\int \frac{g(x_i)^2}{f(x)^2} f(x_i)dx_i\right) - \left(\int \frac{g(x_i)}{f(x_i)} f(x_i)dx_i\right)^2\right] \\
&\overset{i.i.d}{=} \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\int \frac{g(x)^2}{f(x)} dx\right) - \left(\int g(x)dx\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\int \frac{g(x)^2}{f(x)} dx\right) - \theta^2\right] \\
&= \frac{1}{n}\left[\left(\int \frac{g(x)^2}{f(x)} dx\right) - \theta^2\right]
\end{aligned}
$$

Importance Sampling - Extra Slide

Example

Let $\hat{\theta}$ be an importance sampling estimator of $\theta = \int g(x)dx$, where the importance function f is a valid probability density function. Prove that if $\theta < \infty$ and $g(x)/f(x)$ is bounded, then the variance of the importance sampling estimator $\hat{\theta}$ is finite.

Since $f(x)/g(x)$ is bounded, suppose there exists $M \in \mathbb{R}$ where $f(x)/g(x) < M$. Then, the following steps hold:

$$
\begin{aligned}
\mathbb{V}ar(\hat{\theta}) &= \mathbb{V}ar\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{g(x)}{f(x)}\right\} \\
&\overset{i.i.d.}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left\{\frac{g(x)}{f(x)}\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}\left\{\frac{g(x)}{f(x)}\right\}^2 - \left\{\mathbb{E}\left[\frac{g(x)}{f(x)}\right]\right\}^2\right)
\end{aligned}
$$

Importance Sampling - Extra Slide

Note that,

$$\mathbb{E}\left\{\frac{g(x)}{f(x)}\right\}^2 = \int \left\{\frac{g(x)}{f(x)}\right\}^2 f(x)dx \le \int M^2 f(x)dx = M^2$$

Thus,

$$\begin{aligned}\mathbb{V}ar(\hat{\theta}) &\le \frac{1}{n^2}\sum_{i=1}^{n} M^2 - \theta^2 \\ &= \frac{1}{n}(M^2 - \theta^2) \\ &< \infty.\end{aligned}$$

Example

Let $X \sim N(0,1)$. Estimate $\mathbb{P}(X > 2.5)$ using importance sampling in R, where the importance function is an exponential density with mean 1.

Solution. Note that,

$$\frac{g(x)}{f(x)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{e^{-x}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} x}$$

Consider the algorithm:

1. Let $X_1, \ldots, X_n$ be a random variable generated from *Exponential*(1).
2. Compute:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2} + x_i\right\} \mathbf{1}_{\{x_i > 2.5\}}$$

Consult the R codes.

Importance Sampling - Extra Example

Example

Use the importance sampling approach to estimate

$$\theta = \int_0^1 \tan(x) e^{-x^2/2} dx$$

Solution. On a test I will actually tell you what to use, but in this case we'll let $f(x)$ be the density function of the standard Gaussian distribution. Thus,

$$\int_0^1 \tan(x) e^{-x^2/2} dx = \sqrt{2\pi} \int_0^1 \frac{1}{\sqrt{2\pi}} \tan(x) e^{-x^2/2} dx = \sqrt{2\pi}\, \mathbb{E}[\tan(X) \mathbf{1}_{\{0<X<1\}}]$$

Consider the algorithm:

1. Let $X_1, \ldots, X_n$ be a random variable generated from $N(0,1)$.
2. Compute:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\pi} \tan(X) \mathbf{1}_{\{0<X<1\}}$$

Antithetic Variables

Motivation: Antithetic Variables

Consider the variance of random variables X_1 and X_2:

$$\mathbb{V}ar\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\left[\mathbb{V}ar[X_1] + \mathbb{V}ar[X_2] + 2\mathbb{C}ov(X_1, X_2)\right]$$

Although $\mathbb{V}ar(\cdot) \geq 0$, it's possible for covariance to be negative.

Hence, variance is reduced if $\mathbb{C}ov(X_1, X_2) < 0$.

Rant: Why Reduce Variance?

Bias, Mean Square Error

$$Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta, \quad MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V}ar(\hat{\theta}) + Bias(\hat{\theta})^2$$

- Clearly if you have an unbiased estimator than your MSE is just the variance.
- A heuristic for a "better" estimator is one that has the smallest MSE.
- The above is not prescriptive; there are many different alternatives to the MSE. However, most statistics education and researchers use MSE since it's widely used, and honestly, convenient to compute.
- Hence, in early statistics research the primary goal was to compute MVUE (minimum variance unbiased estimators). However, estimators like the MLE works very well despite the fact that the MLE is not always unbiased.

Rant: Why Reduce Variance?

- Suppose hypothetically you are choosing between living two apartments in Toronto. Supposing the price and square footage of the apartment are equal. You only care about the proximity to three things: the Saint George campus, a grocery store, and a cafe.
- So what are you going to do? Are you going to compute the **absolute distance** from your apartment, or are you going to compute the expected **squared distance**?
- Quote from one of my professors: "Math is God-given, statistics is man-made."

Antithetic Variables

Returning back to problems solved by a simple Monte Carlo estimator, we are interested in solving:

$$\theta = \int_0^1 g(x)dx.$$

We did this by generating from the uniform distribution. Note that if $U_i \sim Uniform(0, 1)$ then $1 - U_i \sim Uniform(0, 1)$.

Antithetic Variables

Covariance between Uniform

Compute $\mathbb{C}ov(U, 1 - U)$.

Solution.

$$
\begin{aligned}
\mathbb{C}ov(U, 1 - U) &= \mathbb{E}[U(1 - U)] - \mathbb{E}[U]\mathbb{E}[1 - U] \\
&= \mathbb{E}[U - U^2] - \frac{1}{2}\frac{1}{2} \\
&= \mathbb{E}[U] - \mathbb{E}[U^2] - \frac{1}{4} \\
&= \frac{1}{2} - \int_0^1 u^2 du - \frac{1}{4} \\
&= \frac{1}{2} - \frac{1}{3} - \frac{1}{4} \\
&= -\frac{1}{12}
\end{aligned}
$$

Antithetic Variables

The proof is omitted but a handwavy version is within the textbook (page 156).

Corollary

Consider $U_i \sim Uniform(0,1)$ for $i = 1, 2, \ldots, n$. If $g(X_1, \ldots, X_n)$ is a monotone function then

$$Y = g(F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n))$$

and

$$V = g(F_X^{-1}(1 - U_1), \ldots, F_X^{-1}(1 - U_n))$$

are negatively correlated.

The algorithm for the antithetic approach is as follows:

1. Generate $U_i \overset{i.i.d}{\sim} Uniform(0,1)$ for $i = 1, 2, \ldots, n/2$.
2. Deliver

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n/2}[g(u_i) + g(1-u_i)] = \frac{2}{n}\sum_{i=1}^{n/2}\frac{[g(u_i) + g(1-u_i)]}{2}$$

Antithetic Variables

Example

Using antithetic variables, estimate the following integral:

$$\int_0^1 \frac{1}{1+x}dx$$

Compare the result to the simple Monte Carlo estimator and the true value.

Solution. Note that between [0, 1] that $\frac{1}{1+x}$ is monotone decreasing. Hence, using the previous corollary we know that if we sample from $U_i \sim Uniform(0,1)$ for $i = 1, 2, \ldots, n$, then $\frac{1}{1+U_i}$ and $\frac{1}{1+(1-U_i)}$ are negatively correlated.
The true value is:

$$\int_0^1 \frac{1}{1+x}dx = \ln(x) + c\Big|_1^2 = \ln(2)$$

Antithetic Variables

The algorithm:

1. Generate $U_i \overset{i.i.d}{\sim} Uniform(0, 1)$ for $i = 1, 2, \ldots, n/2$.
2. Deliver

$$\hat{\theta} = \frac{2}{n} \sum_{i=1}^{n/2} \frac{[g(u_i) + g(1 - u_i)]}{2}$$

See the corresponding R codes.

Antithetic Variables

Example

Using antithetic variables, estimate $\Phi(2.9)$ and $\Phi(-3.2)$. Recall:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\} dt$$

Compare the result to the simple Monte Carlo estimator that we computed earlier.

Solution. Recall from before:

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= 0.5 + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.\end{aligned}$$

Antithetic Variables

Use the substitution $y = t/x$ and $dt = xdy$ Then,

$$\theta = \int_0^1 \frac{1}{\sqrt{2\pi}} x \exp\{-(xy)^2/2\} dy$$

Note that the function $g(y) = \frac{1}{\sqrt{2\pi}} x \exp\{-(xy)^2/2\}$ is monotone increasing between $[0, 1]$. Thus, we can use the antithetic approach. The algorithm is:

1. Generate $U_i \overset{i.i.d}{\sim} Uniform(0, 1)$ for $i = 1, 2, \ldots, n/2$.
2. Compute

$$\hat{\theta} = 0.5 + \frac{2}{n} \sum_{i=1}^{n/2} \frac{\left[\frac{1}{\sqrt{2\pi}} x \exp\{-(xu_i)^2/2\} + \frac{1}{\sqrt{2\pi}} x \exp\{-(x(1-u_i))^2/2\}\right]}{2}.$$

3. If $x > 0$, deliver $0.5 + \hat{\theta}$. Otherwise, deliver $1 - (0.5 + \hat{\theta})$.

See the corresponding R codes.