DIRECT MONTE CARLO AND IMPORTANCE SAMPLING

Biswaranjan Meher
P346 DIY REPORT
Roll No- 2011050
School of Physical Sciences
National Institute of Science Education and Research, Bhubaneswar

December 1, 2022

ABSTRACT

In solving modern-day physical/ mathematical problems, definite integration is the most crucial calculation that is commonly employed. Estimation of those integrals is approximately calculated by various techniques or algorithms known as the Midpoint rule, Trapezoidal rule, Simpson method, Monte Carlo method, etc. However, the Monte Carlo method gave a better approximation of the integral up to the given precision among those. But, in the case of probability estimation (the issue concerned in this report), the Monte Carlo method gave a comparatively high error bar. This report is mainly focused on describing an approximation method known as **Importance Sampling** that helps to reduce the error bar, i.e. the standard deviation.

1 INTRODUCTION

In the probability estimation, the following integration, where p(x) is the probability density function of x, defines the expectation value of a continuous random variable x:

$$\int xp(x)dx \tag{1}$$

However, using a computer to calculate this value is a difficult task. Numerical approximation techniques are employed to precisely estimate this integration, one of which is known as Monte Carlo Integration.
However, estimating the integral using the Monte Carlo integration technique gives a comparatively high standard deviation. Now, to improve the estimator, a new variant or algorithm is proposed, known as Importance Sampling to address the problem by reducing the standard deviation (or variance).

2 BACKGROUND

2.1 Monte Carlo method

Monte Carlo estimation is an algorithm that involves repeated random sampling to estimate an integration numerically. It is a widely used method to estimate multi-variable or higher-dimensional integrals.
To estimate an integral of the form:

$$\int_a^b f(x)dx \tag{2}$$

where x is a random variable drawn from a PDF p(x).
The estimator is defined as:

$$I = \frac{b-a}{N}\sum_{i=1}^{N} f(x_i) \tag{3}$$

The variance and standard deviation are defined as:

$$\sigma_f^2 = \frac{1}{N}\sum_{i=1}^{N} f(x_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right)^2 \tag{4}$$

$$Std.\,deviation = \sigma_f \tag{5}$$

In order to calculate a d-dimensional integral, it is natural to try to extend the one-dimensional approach. When doing so, the number of times the function f has to be calculated increases to N = $(n+1)^d \approx n^d$ times, and the approximation error will be proportional to $n^{-2} \approx N^{-2/d}$. One key advantage of the Monte Carlo method to calculate integrals numerically is that it has an error that is proportional to $n^{-1/2}$, regardless of the dimension of the integral. A second important advantage of Monte Carlo integration is that the approximation error does not depend on the smoothness of the functions that are integrated.

2.2 Monte Carlo in Probability theory

We recently saw how to use the Monte Carlo method to calculate integrals in the previous subsection. Now, probabilities and expectation values can also be described as integrals and can be estimated by the Monte Carlo method. Thus, it extends to the probability theory.
For example, to calculate the expectation value E[f(X)] of a function f of a continuously distributed random variable X with probability density function p, using the Monte Carlo integration, we notice that:

$$E[f(X)] = \int f(x)p(x)dx \tag{6}$$

This integral is then calculated with the Monte Carlo method.
To calculate the probability $P(X \in O)$, for a set O, we make use of the fact that:

$$P(X \in O) = \int I(x)p(x)dx \tag{7}$$

where

$$I(x) = \begin{cases} 1 & x \in O \\ 0 & X \notin O \end{cases} \tag{8}$$

NOTE: This whole algorithm is based on the assumption that x was distributed according to the PDF p similar to the case of X, i.e. for x, we can efficiently draw samples from the PDF p of the random variable X.
PROBLEM: What if we can't do that????
SOLUTION: Introducing an alternative distribution and observing the results...gave birth to...

IMPORTANCE SAMPLING

3 IMPORTANCE SAMPLING

The Monte Carlo approximation of the integral is given by:

$$E[f(X)] = \frac{1}{N}\sum_{i=1}^{N} f(x) \tag{9}$$

To address the problem discussed above, we introduce a proposal function q(x) into Eqn. (9), as:

$$E[f(X)] = \int f(X)\left(\frac{p(X)}{q(X)}\right)q(X)dx \tag{10}$$

For the above eqn., the computed integral would be:

$$E[f(X)] = \frac{1}{N}\sum_{i=1}^{N} f(x)\frac{p(x)}{q(x)} \tag{11}$$

where q(x) is the new PDF such that

$$q(x) = 0 \implies p(x) = 0$$

This condition is known as absolute continuity (not so important in this report).
Let's define the importance weight function:

$$w(x) = \frac{p(x)}{q(x)} \tag{12}$$

So, the final equation becomes:

$$E[f(X)] = \frac{1}{N}\sum_{i=1}^{N} w(x)f(x) \tag{13}$$

Now, even if we got the theoretical knowledge of the solution to the problem addressed before, it is a difficult task to implement it, especially in determining the proposal function q(x).

3.1 Determination of q(x)

The proposal function q(x) should be determined with a logic:
Since the integral value of $\int f(X)p(X)$ has the major contribution from the region where |f(X)|p(X) is large, we need to find a q(X) such that it is large at the same region. In other words, we need to find a q(X) such that it has a major contribution to the estimated value of the integral.

4 QUESTION TO BE SOLVED

Given the current equation:

$$I(\Delta L, \Delta V_{TH}) = \frac{50}{0.1 + \Delta L}(0.6 - \Delta V_{TH})^2 \tag{14}$$

where $\Delta L \approx N(0, 0.01^2)$ and $\Delta V_{TH} \approx N(0, 0.03^2)$.
Calculate the probability of I> 275 units and the standard deviation corresponding to the methods, i.e. Direct Monte Carlo estimation and Importance Sampling.

4.1 Analytical advancement for the solution

4.1.1 Direct Monte Carlo estimation

In this method, we approximate an integral by the sample mean of a function of simulated random variables. Mathematically,

$$E[f(X)] \approx \int p(X)f(X) = \frac{1}{N}\sum_{i=1}^{N} f(x) \tag{15}$$

In our problem of probability estimation, we have:

$$I_A(x) = \begin{cases} 1 & I > 275 \\ 0 & I < 275 \end{cases} \tag{16}$$

Replacing in our equation,

$$E[f(X)] \approx \int p(X)f(X) = \frac{1}{N}\sum_{i=1}^{N} I_A(x) \tag{17}$$

4.1.2 Importance Sampling method

With Importance Sampling, we try to reduce the standard deviation of the Monte Carlo estimation by choosing a proposal distribution from which we simulate the random variables. It involves multiplying the integrand by 1 (in a tricky fashion) to yield the expectation value of a quantity that varies less than the original integrand over the integration region. Mathematically,

$$E[f(X).\frac{p(X)}{q(X)}] \approx \frac{1}{N}\sum_{i=1}^{N} w(x)f(x) \tag{18}$$

where $w(x) = \frac{p(x)}{q(x)}$.
In our problem above, we can alter the mean and standard deviation of ΔL and ΔV_{TH} in the hopes that more of our sampling points fall in the failure region. So, let's define 2 new distributions with altered σ^2: $\Delta\hat{L} \approx N(0, 0.02^2)$ and $\Delta\hat{V_{TH}} \approx N(0, 0.06^2)$.

5 CODES AND OUTPUTS

The code for the solution to the problem given above will be attached along with the report separately in GitHub (link provided in the Google Classroom) and Google Classroom. Please refer to that for codes and Outputs.

6 CONCLUSIONS AND DISCUSSIONS

The mean probability and the standard deviation as asked in the question have been estimated using both methods.

- The mean probabilities estimated by both methods are nearly equal up to the precision 10^{-3} and sometimes up to 10^{-4}.
- The standard deviation estimated by Importance Sampling is much less (approximately by a factor of 3-5) as compared to the estimation by Direct Monte Carlo, as expected to be.

About the code:

- The code uses the Python inbuilt functions NumPy, SciPy, and Matplotlib:
  - To construct a Gaussian distribution of random variables, which was not possible in the RNG function in my library as it gives only a list containing random numbers.
  - To find the standard deviation in a given list of variables.
  - To plot the necessary graphs to show the convergence speed of the estimation in the two methods.

7 REFERENCES

(1) NISER P346 Lecture Notes: Numerical Integration Slides:
`https://classroom.google.com/u/0/c/NTM3OTY3MDcwMjQz`
(2) `https://astrostatistics.psu.edu/su14/lectures/cisewski_is.pdf`
(3) `https://towardsdatascience.com/importance-sampling-introduction-e76b2c32e744`