Welcome to Advanced Data Analysis (PHYS 605)

Prof. Claudia Gomes da Rocha

claudia.gomesdarocha@ucalgary.ca

Department of Physics and Astronomy
Faculty of Science, University of Calgary

UNIVERSITY OF CALGARY

Monte Carlo Methods

Monte Carlo Methods use “**chance**” or random numbers to sample all possible states of a system. Its underlying concept is to use randomness to solve problems that may be deterministic in principle.

Monte Carlo Methods are mainly used in optimization problems, numerical integration (multidimensional integrals with complicated boundary conditions), generating draws from a probability distribution, determining uncertainties in parameters model, and generating *synthetic data* (generating something that looks like the data).

Monte Carlo Methods are also useful for simulating systems with many coupled degrees of freedom (e.g., fluids, disordered materials, cellular structures, kinetic models of gases, …).

Other examples include modelling phenomena with significant uncertainty in inputs such as calculation of risks in business, prediction of failure simulators, sampling, etc.

The name *Monte Carlo* is inspired from **games of chance** played in the Monte Carlo casinos in Monaco.

Monte Carlo Methods

First, we introduce Monte Carlo method as a technique to carry out integrals of single and multidimensional functions.

Evaluation of the integral of a function $g(x)$ for which an analytic solution is either unavailable or too complicated to obtain exactly.

$$I = \int_A g(x)dx$$

The goal is to derive a method to approximate this integral by randomly drawing *N* samples from the domain A. For simplicity, let's assume that this domain is an interval $[a, b]$.

The method starts with the drawing of *N* samples from a uniform distribution

$$f(x) = \begin{cases} \dfrac{1}{b-a}\,, & if\ a \leq x \leq b \\ 0\,, & otherwise \end{cases}$$

Monte Carlo Methods

For the purpose of this example, the expectation value of $g(x)$ of the random variable X is:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx \simeq \frac{1}{N}\sum_{i=1}^{N} g(x_i)$$

$$E[g(x)] = \frac{1}{b-a}\underbrace{\int_{a}^{b} g(x)dx}_{I} \simeq \frac{1}{N}\sum_{i=1}^{N} g(x_i)$$

$$I = (b-a)E[g(x)] \simeq \frac{(b-a)}{N}\sum_{i=1}^{N} g(x_i)$$

which can be calculated by drawing *N* random uniform samples x_i from the domain range, then calculating $g(x_i)$ and evaluating the sum. ➔ *Basic Monte Carlo Integration Method*

Monte Carlo Methods

The value of the integral and its associated error can be estimated as

$$I \pm \frac{s_g}{\sqrt{N}}$$

as the precision of the integral depends on the number of samples drawn.

$$s_g^2 = \frac{1}{N-1}\sum_{i=1}^{N}(g(x_i) - \bar{g})^2 = \frac{1}{(b-a)}\int_a^b (g(x) - \bar{g})^2 dx$$

Monte Carlo Methods

But for sufficiently large N ➔ $\hat{I} \simeq \mathcal{N}(I, \sigma_I^2)$

One can define a confidence interval for I can be estimated as

$$I \pm z_\alpha \, \sigma_I$$

$$\sigma_I^2 = Var[I] = Var\left[\frac{(b-a)}{N}\sum_{i=1}^{N} g(x_i)\right]$$

where z_α is the standard score (z-score) at a significance level α representing how many standard deviations a data point is from the mean of a set of data.

Monte Carlo Methods

Generalization to multivariate integrals:

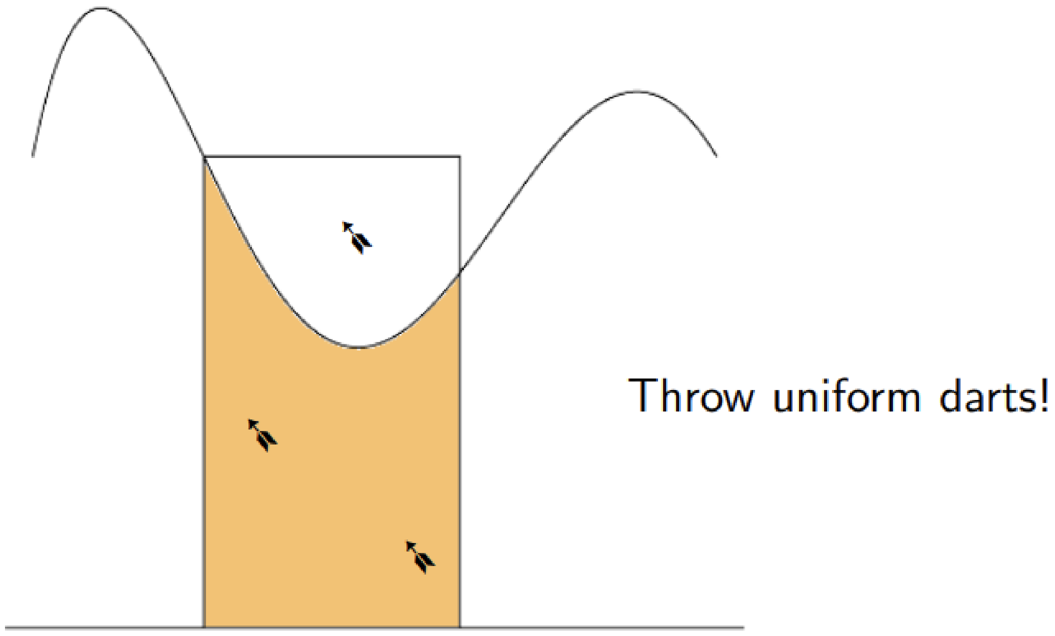
$$\int_{a_1}^{b_1}\int_{a_2}^{b_2}\int_{a_3}^{b_3} g(x,y,z)dz\,dy\,dx \approx (b_1-a_1)\,(b_2-a_2)\,(b_3-a_3)\frac{1}{N}\sum_{i=1}^{N} g(x_i,y_i,z_i)$$

for example, to estimate $\int_{-2}^{2}\int_{0.5}^{1}\int_{0}^{1}\frac{e^{-x^2/2y}}{(x^2z+1)}dz\,dy\,dx$.

If we have a multivariate space with n variables with a n-dimensional volume domain V, then the Monte Carlo basic integration can be generalized to

$$I \simeq \frac{V}{N}\sum_{i=1}^{N} g(x_i,y_i,z_i)$$

Hit-or-Miss Monte Carlo (integrate function)



$$\int_a^b g(x)dx = I = pc(b-a) \approx \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right] c(b-a)$$

- Sample uniformly from a rectangular region $[a, b] \times [0, c]$
- The probability a dart will be below the curve is
$$p = \frac{I}{c(b-a)}$$
- So, if we can estimate p, we can estimate I!
- How to estimate p? Throw N “uniform darts” at the rectangle.
- Let X be the number of times the dart is under the curve $y = g(x)$.
- Then $\hat{p} = \frac{X}{N}$ and with an approximate confidence interval that can be obtained as
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Hit-or-Miss Monte Carlo (integrate function)

Example:

$$I = \int_0^3 e^x dx = e^3 - 1 \approx 19.08554$$

- An upper bound of the rectangle is e^3
- 100,000 uniform draws over $[0,3] \times [0, e^3]$

Try it!

$\hat{p}$	I	95% CI
0.31365	19.06217885	(18.88849413, 19.23586357)
0.31552	19.01216586	(18.83860388, 19.18572785)
0.31729	19.11882006	(18.94499712, 19.29264301)
0.31642	19.06639681	(18.89270180, 19.24009185)
0.31672	19.08447380	(18.91073460, 19.25821302)

We can use this method to estimate the area of a circle and estimate π!