Lecture 3

Crude Monte Carlo methods and importance sampling

The aim of this method is to evaluate a multidimensional integral, such as

$$\int_a^b dx_1 \int_a^b dx_2 \dots \int_a^b dx_N f(x_1, \dots, x_N)$$

Performance estimation of a deterministic method
We can make an estimation of the performance of a deterministic method in the following way:
Suppose we have an integral between (a, b), and we want to approximate it with

$$\int_a^b f(x)dx \simeq \sum_i f(x_i)\Delta x$$

Of course, the greater the number of points x_i that we divide segment $[a, b]$ in, the better the approximation will be.

Now assume we want to estimate an N-dimensional integral with a grid: $\{x_1 \times x_2 \times x_3 \times \dots \times x_N\}$, where each x_i corresponds to 100 grid points. The total number of grid points will be $\mathcal{N} = 100^N = 10^{2N}$ grid points. The typical memory size of a computer is of the order of 10 Gb, or 10^10 bytes. In our example, the **maximal dimension** in which we can estimate the integral using the whole memory of our computer will be $N \simeq 5$. Which is a relatively small number compared to common problems where multidimensional integrals have to be calculated, and this number only becomes smaller when taking into account problems with multiple particles.

Simplest "crude" Monte Carlo method
With this motivation, we introduce the Monte Carlo method for approximating one dimensional integrals, called the **hit-or-miss** method. Consider the integral

$$\int_a^b f(x)dx = (b - a) \int_a^b f(x)p(x)dx \tag{3.1}$$

where $p(x) = 1/(b - a)$ is a uniform PDF with which we can apply the Central Limit Theorem. Namely,

- we can calculate the expectation value of $f(x)$ with respect to $p(x)$ as:

$$\langle f \rangle = \int_a^b f(x)dx = \lim_{N\to\infty} (b-a) \frac{\sum_{i=1}^N f(x_i)}{N}$$

  where random values x_i are taken from the uniform PDF $p(x)$.

- For a finite sample size of N random values, the integral is approximated as

$$\int_a^b f(x)dx \approx \frac{b-a}{N} \sum_{i=1}^N f(x_i) \tag{3.2}$$

  and the error in the estimation can be calculated.

- The statistical error of such an approximation corresponds to $\pm\sigma/\sqrt{N}$, where the variance is

$$\sigma^2 = \lim_{N\to\infty} (b-a) \left[\frac{\sum_i f^2(x_i)}{N} - \left(\frac{\sum_i f(x_i)}{N} \right)^2 \right] \equiv \langle f^2 \rangle - \langle f \rangle^2$$

The nice thing about this simple method is that it can be generalized to D-dimensions.

Monte Carlo method with importance sampling
Suppose we are interested in calculating the area of a particular region inside our domain. Then, we could start sampling random points inside our domain and count how many of them are inside our region of interest. However, this leads to a lot of useless sampled points. Then, in order to sample the important part of our domain, we consider the integral

$$\int f(x)p(x)dx$$

where now $p(x)$ is a non-uniform PDF and takes large values in the region of interest. If we use CLT theorem, we can see that any integral can be written as

$$I = \int f(x)dx = \int \left[\frac{f(x)}{p(x)} \right] p(x)dx$$

Notice that if we compare this with the previous crude Monte Carlo method (equation 3.2), the result is a bit different.

	Crude method	Importance sampling
Integral	$\frac{1}{N}\sum f(x_i)$	$\frac{1}{N}\sum \frac{f(x_i)}{p(x_i)}$
Variance	$\int f^2(x)dx - (\int f(x)dx)^2$	$\int [\frac{f(x)}{p(x)}]p(x)dx - (\int f(x)dx)^2$

Table 3.1: Comparison between the two Monte Carlo methods

In importance sampling, we have an additional degree of freedom with which we can reduce the statistical error in the approximation, and that is adjusting the PDF $p(x)$.

Generating a non-uniform PDF

The procedure is simple:

1. Take random values u_i from a uniform PDF.
2. Use them to generate random values x_i for the non-uniform PDF $p(x)$.
3. This can be done with the relation $x_i = P^{-1}(u_i)$, where $P(x)$ is the cumulative function.

Example: Generate random values from a Gaussian

Let our objective $p(x)$ be

$$p(x) = 2xe^{-x^2}, x > 0$$

normalized to unity by the condition $\int_0^\infty p(x)dx = 1$. Its cumulative function is

$$\begin{aligned} P(x) &= \int_0^x p(x')dx' = \int_0^x 2x'e^{-(x')^2}dx' \\ &= e^{-(x')^2}\Big|_0^x = 1 - e^{-x^2} \end{aligned}$$

If we invert the relation $x(u)$, we find

$$\begin{aligned} u &= 1 - e^{-x^2} \\ e^{-x^2} &= 1 - u \\ -x^2 &= \ln(1-u) \\ x &= \sqrt{-\ln(1-u)} \end{aligned}$$

where u is sampled from a uniform distribution. But actually $1 - u$ is also a uniform distribution from 0 to 1. So we can actually use $x = \sqrt{-\ln(u)}$ to sample a Gaussian distribution.

Normal Random Generator

The previous example did not quite generate samples from a Gaussian distribution, since it had the $2x$ term in the front. In order to sample from a normal distribution, we can use two different methods:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Box-Muller method

The first step is to generate 2 random values from a uniform PDF u_1, u_2, and use the polar transformation:

$$\begin{aligned} r &= \sqrt{-2\ln u_1} \\ \varphi &= 2\pi u_2 \end{aligned}$$

and then define

$$\begin{aligned} z_0 &= r\cos\varphi \\ z_1 &= r\sin\varphi \end{aligned}$$

in order to produce two independent random variables that are distributed according to $p(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Finally, we have to adjust the mean value and the variance according to

$$\begin{aligned} x_0 &= \mu + z_0\sigma \\ x_1 &= \mu + z_1\sigma \end{aligned}$$

3.1 Exercise 3: Volume inside a sphere in D dimensions

Area of a circle
The first part of the third exercise will consist in calculating the area of a circle. That is

$$A_{\text{exact}} = \pi R^2$$

The way to do this is to generate two random variables $x, y \in (0,1)$ with R being the desired radius of the circle. If we repeat this process N times and count the number of times where the pair of values (x_i, y_i) is inside the circle, we will have an approximation of the area as

$$A \approx 4R^2 \frac{N_{hit}}{N_{iter}}$$

where N_{hit} is the number of pairs (x_i, y_i) such that $\sqrt{x_i^2 + y_i^2} < 1$. The statistical error is estimated as $\varepsilon_{stat} = \sigma/\sqrt{N}$, and we can compare it with the actual error $\varepsilon_{actual} = |A - A_{\text{exact}}|$ in a log-log plot for different values of N_{iter}.

Volume inside a sphere
To calculate the volume inside a sphere, we generate three random variables according to a uniform distribution and compare the results with

$$V = \frac{4}{3}\pi R^3$$

Volume inside a D-dimensional sphere
Finally, we generalize the case to the D-dimensional case, where the volume of this sphere is estimated as

$$V \approx (2R)^D \frac{N_{hit}}{N_{iter}}$$