Chapter 6

The Monte Carlo Method

6.1 The Monte Carlo method

6.1.1 Introduction

A basic problem in applied mathematics, is to be able to calculate an integral

$$I = \int f(x)dx,$$

that can be one-dimensional or multi-dimensional. In practice, the calculation can seldom be done analytically, and numerical methods and approximations have to be employed.

One simple way to calculate an integral numerically, is to replace it with an approximation, *Riemann sum*, leaning on the definition of the Riemann integral. For a one-dimensional integral, over the interval $[a, b]$, say, this means that the domain of integration is divided into several subintervals of length Δx, say,

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b \quad \text{where} \quad x_i = x_{i-1} + \Delta x \quad \text{for} \;\; i = 1, \ldots, n.$$

By Taylor expansion, the integral over an interval is given by

$$\int_{x_{i_1}}^{x_{i-1}+\Delta x} f(x)dx = \Delta x \frac{f(x_{i-1}) + f(x_{i-1} + \Delta x)}{2} - \frac{(\Delta x)^3}{12} f^{''}(\chi)$$

for some $\chi_i \in (x_{i-1}, x_{i-1} + \Delta x)$. It follows that the integral over the whole interval $[a, b]$ is given by

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_{i-1}+\Delta x} f(x)dx = \sum_{i=1}^{n} \Delta x w_i f(x_i) - \frac{(b-a)^3}{12n^2} f^{''},$$

where

$$f^{''} = \frac{1}{n} \sum_{i=1}^{n} f^{''}(\chi_i) \quad \text{and} \quad w_i = \begin{cases} 1/2 & \text{for} \quad i = 0, \\ 1 & \text{for} \quad i = 1, \ldots, n-1, \\ 1/2 & \text{for} \quad i = n. \end{cases}$$

Notice that the error is proportional to $1/n^2$, and that the function f has to calculated $n + 1$ times.

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 $n^{-2} \approx N^{-2/d}$.

Notice that a higher order methods, that use more terms of the Taylor expansion of f, give smaller approximation errors, at the cost of also having to calculate derivatives of f. When the dimension d is large, the above indicated extension of the one-dimensional approach will be very time consuming for the computer.

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 with Monte Carlo integration, is that the approximation error does not depend on the smoothness of the functions that is integrated, whereas for the above indicated method, the error increases with the size of f'', and the method breaks down if f is not smooth enough.

6.1.2 Monte Carlo in probability theory

We will see how to use the Monte Carlo method to calculate integrals. However, as probabilities and expectations can in fact be described as integrals, it is quite immediate how the Monte Carlo method for ordinary integrals extends to probability theory.

For example, to calculate the expected value $\mathbf{E}\{g(X)\}$ of a function g of a continuously distributed random variable X with probability density function f, using the Monte Carlo integration, we notice that

$$\mathbf{E}\{g(X)\} = \int g(x)f(x)dx.$$

This integral is then calculated with the Monte Carlo method.

To calculate the probability $\mathbf{P}\{X \in O\}$, for a set O, we make similar use of the fact that

$$\mathbf{P}\{X \in O\} = \int I_O(x)f(x)dx \quad \text{where} \quad I_O(x) = \begin{cases} 1 & \text{if} \quad x \in O, \\ 0 & \text{if} \quad x \notin O. \end{cases}$$

6.2 Monte Carlo integration

Consider the d-dimensional integral

$$I = \int f(x)dx = \int_{x_1=0}^{x_1=1} \cdots \int_{x_d=0}^{x_d=1} f(x_1, \ldots, x_d)dx_1 \ldots dx_d$$

of a function f over the unit hypercube $[0,1]^d = [0,1] \times \ldots \times [0,1]$ in $\mathbb{R}^d$. Notice that the integral can be interpreted as the expectation $\mathbf{E}\{f(X)\}$ of the random variable $f(X)$, where X is an $\mathbb{R}^d$-valued random variable with a uniform distribution over $[0,1]^d$, meaning that the components $X_1, \ldots, X_d$ are independent and identically uniformly distributed over $[0,1]$, i.e., $X_1, \ldots, X_d$ are random numbers.

The *Monte Carlo approximation* of the integral is given by

$$E = \frac{1}{n}\sum_{i=1}^{n} f(x_i),$$

where $\{x_i\}_{i=1}^n$ are independent observations of X, i.e., independent random observations of a $\mathbb{R}^d$-valued random variable, the components of which are random numbers.

For an integral

$$I = \int_{[a,b]} f(x)dx = \int_{x_1=a_1}^{x_1=b_1} \cdots \int_{x_d=a_d}^{x_d=b_d} f(x_1, \ldots, x_d)dx_1 \ldots dx_d$$

over a hyperrectangle $[a,b]^d = [a_1,b_1] \times \ldots \times [a_d,b_d]$ in $\mathbb{R}^d$, the sample $\{x_i\}_{i=1}^n$ should be independent observations of a $\mathbb{R}^d$-valued random variable X that is uniformly distributed over $[a,b]$ instead, i.e., the components $X_1, \ldots, X_d$ of X should have uniform distributions over $[a_1,b_1], \ldots, [a_d,b_d]$, respectively.

This approximation converges, by the law of large numbers, as $n \to \infty$, to the real value I of the integral. The convergence is in the probabilistic sense, that there is never a guarantee that the approximation is so and so close I, but that it becomes increasingly unlikely that it is not, as $n \to \infty$.

To study the error we use the *Central Limit Theorem* (CLT), telling us that the sample mean of a random variable with expected value μ and variance σ^2, is approximately normal $\mathrm{N}(\mu, \sigma^2/n)$-distributed.

For the Monte Carlo approximation E of the integral I, the CLT gives

$$\mathbf{P}\left(a\frac{\sigma(f)}{\sqrt{n}} < E - I < b\frac{\sigma(f)}{\sqrt{n}}\right) = \mathbf{P}\left(a\frac{\sigma(f)}{\sqrt{n}} < \frac{1}{n}\sum_{i=1}^n f(x_i) - I < b\frac{\sigma(f)}{\sqrt{n}}\right) \approx \Phi(b) - \Phi(a).$$

Here, making use of the Monte Carlo method again,

$$\sigma^2(f) = \int (f(x) - I)^2 dx \approx \frac{1}{n}\sum_{i=1}^n (f(x_i) - E)^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2 - E^2 = \widehat{\sigma^2(f)}.$$

In particular, the above analysis shows that the error of the Monte Carlo method is of the order $n^{-1/2}$, regardless of the dimension d of the integral.

Example 6.1. Monte Carlo integration is used to calculate the integral

$$\int_0^1 \frac{4}{1+x^2}dx,$$

which thus is approximated with

$$E = \frac{1}{n}\sum_{i=1}^n \frac{4}{1+x_i^2},$$

where x_i are random numbers. A computer program for this could look as follows:

```
E=0, Errorterm=0
For 1 to n
    Generate a uniform distributed random variable x_i.
    Calculate y=4/(1+x_i^2)
    E=E+y and Errorterm=y^2+Errorterm
End
E=E/n
Error=sqrt(Errorterm/n-E^2)/sqrt(n)
```

6.3 More on Monte Carlo integration

6.3.1 Variance reduction

One way to improve on the accuracy of Monte Carlo approxiamtions, is to use *variance reduction* techniques, to reduce the variance of the integrand. There are a couple of standard techniques of this kind.

It should be noted that a badly performed attempt to variance reduction, at worst leads to a larger variance, but usually nothing worse. Therefore, there is not too much to lose on using such techniques. And it is enough to feel reasonbly confident that the technique employed really reduces the variance: There is no need for a formal proof of that belief!

It should also be noted that variance reduction techniques often carry very fancy names, but that the ideas behind always are very simple.

6.3.2 Stratified sampling

Often the variation of the function f that is to be integrated varies over different parts of the domain of integration. In that case, it can be fruitful to use *stratified sampling*, where the domain of integration is divided into smaller parts, and use Monte Carlo integration on each of the parts, using different sample sizes for different parts.

Phrased mathematically, we patition the integration domain $M = [0,1]^d$ into k regions $M_1, \ldots, M_k$. For the region M_j we use a sample of size n_j of observation $\{x_{ij}\}_{i=1}^{n_j}$ of a random variable X_j with a uniform distribution over M_j. The resulting Monte Carlo approximation E of the integral I becomes

$$E = \sum_{j=1}^{k} \frac{\mathrm{vol}(M_j)}{n_j} \sum_{i=1}^{n_j} f(x_{ij}),$$

with the corresponding error

$$\Delta_{SS} = \sqrt{\sum_{j=1}^{k} \frac{\mathrm{vol}(M_j)^2}{n_j} \sigma_{M_j}^2(f)},$$

where

$$\sigma_{M_j}^2(f) = \left(\frac{1}{\mathrm{vol}(M_j)} \int_{M_j} f(x)^2 dx - \left(\frac{1}{\mathrm{vol}(M_j)} \int_{M_j} f(x) dx \right)^2 \right).$$

The variances $\sigma_{M_j}^2(f)$ of the differents parts of the partition, in turn, are again estimated by means of Monte Carlo integration.

In order for startified sampling to perform optimal, on should try to select

$$n_j \sim \mathrm{vol}(M_j) \sigma_{M_j}(f).$$

6.3.3 Importance sampling

An alternative to stratified sampling, is *importance sampling*, where the redistribution of the number of sampling points is carried out by means of replacing the uniform distribution with another distribution of sampling points.

First notice that

$$I = \int f(x)dx = \int \frac{f(x)}{p(x)} p(x)dx,$$

If we select p to be a probability density function, we may, as an alternative to ordinary Monte Carlo integration, generate random observations $x_1, \dots, x_n$ with this probability density function, and approximate the integral I with

$$E = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{p(x_i)},$$

The error of this Monte Carlo approximation is $\sigma(f/p)/\sqrt(n)$, where $\sigma^2(f/p)$ is estimated as before, with

$$\widehat{\sigma^2(f/p)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{f(x_i)}{p(x_i)} \right)^2 - E^2,$$

In analogy with the selection of the diffrent sample sizes for stratified sampling, it is optimal to try select $p(x)$ as close in shape to $f(x)$ as possible. (What happens if the fucntion f to be integrated is itself a probability density function?)

6.3.4 Control variates

One simple approach to reduce variance, is try to employ a *control variate* g, which is a function that is close to f, and with a known value $I(g)$ of the integral. Writing

$$I = \int f(x)dx = \int (f(x) - g(x))dx + \int g(x)dx = \int (f(x) - g(x))dx + I(g),$$

with g close to f, the variance of $f - g$ should be smaller than that of f, and the integral $I = I(f)$ is approximated by the sum E of the Monte Carlo approxiamtion of that integral and $I(g)$:

$$E = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i)) + I(g).$$

6.3.5 Antithetic variates

Whereas ordinary Monte Carlo integration uses random samples built of independent observations, it can be advantageous to use samples with pairs of observations that are negatively correlated with each other. This is based on the fact

$$\mathbf{Var}\{f_1 + f_2\} = \mathbf{Var}\{f_1\} + \mathbf{Var}\{f_2\} + 2\mathbf{Cov}\{f_1, f_2\}.$$

Example 6.2. Let f be a monotone function of one variable (i.e., f is either increasing or decreasing). In order to calculate the integral

$$I = \int_0^1 f(x)dx.$$

using observed random numbers $\{x_i\}_{i=1}^k$, can use the Monte Carlo approximation

$$I \approx E = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{f(1-x_i)}{2}.$$

This motivation of this approximation is that $1-x_i$ is an observation of a random number when x_i is. As x_i and $1-x_i$ obviously are negatively correlated, so should be $f(x_i)$ and $f(1-x_i)$. Thus the error of this Monte Carlo approximation should be small.

In the above example, the random variable $f(Y) = f(1-X)$ has the same distribution as the random variable $f(X)$ that is sampled for ordinary Monte Carlo integration. In addition $f(Y)$ and $f(X)$ are negatively correlated. We summarize these two properties, that can be very useful to calculate the integral of f, by saying that $f(Y)$ is an *antithetic variate* to $f(X)$.

6.4 Simulation of random variables

6.4.1 General theory for simulation of random variables

The following technical lemma is a key step to simulate random variables in a computer:

Lemma 6.1. *For a distribution function F, define the* generalized right-invers *$F^{\leftarrow}$ by*

$$F^{\leftarrow}(y) \equiv \min\{x \in (0,1) : F(x) \geq y\} \quad \textit{for } y \in (0,1).$$

We have

$$F^{\leftarrow}(y) \leq x \Leftrightarrow y \leq F(x).$$

Proof. [3]For $F(x) < y$ there exists an $\epsilon > 0$ such that $F(x) < y$ for $z \in (-\infty, x+\epsilon]$, as F is non-decreasing and continuous from the right. This gives

$$F^{\leftarrow}(y) = \min\{z \in (0,1) : F(z) \geq y\} > x.$$

On the other hand, for $x < F^{\leftarrow}(y)$ we have $F(x) < y$, since

$$F(x) \geq y \Rightarrow F^{\leftarrow}(y) = \min\{z \in (0,1) : F(z) \geq y\} \leq x.$$

Since we have shown that $F(x) < y \Leftrightarrow x < F^{\leftarrow}(y)$, it follows that $F^{\leftarrow}(y) \leq x \Leftrightarrow y \leq F(x)$. □

[3] This proof is not important for the understanding of the rest of the material.

From a *random number*, i.e. a random variable that is *uniformly distributed over the interval* $[0,1]$, a random variable with any other desired distribution can be simulated, at least in theory:

Theorem 6.1. *If F is a distribution function and ξ a random number, then $F^{\leftarrow}(\xi)$ is a random variable with distribution function F.*

Proof. Since the uniformly distributed random variable ξ has distribution function $F_\xi(x) = x$ for $x \in [0,1]$, Lemma 6.1 shows that

$$F_{F^{\leftarrow}(\xi)}(x) = \mathbf{P}\{F^{-1}(\xi) \leq x\} = \mathbf{P}\{\xi \leq F(x)\} = F_\xi(F(x)) = F(x). \quad \square$$

When using Theorem 6.1 in practice, it is not necessary to know an analytical expression for $F^{\leftarrow}$: It is enough to know how to calculate $F^{\leftarrow}$ numerically.

If the distribution function F has a well-defined ordinary invers F^{-1}, then that inverse coincides with the generalized right-inverse $F^{\leftarrow} = F^{-1}$.

Corollary 6.1. *Let F be a continuous distribution function. Assume that there exists numbers $-\infty \leq a < b \leq \infty$ such that*

- $0 < F(x) < 1$ *for* $x \in (a,b)$*;*
- $F : (a,b) \to (0,1)$ *is strictly increasing and onto.*

Then the function $F : (a,b) \to (0,1)$ is invertible with invers $F^{-1} : (0,1) \to (a,b)$. Further, if ξ is a random number, then the random variable $F^{-1}(\xi)$ has distribution function F.

Corollary 6.1 might appear to be complicated, at first sight, but in practice it is seldom more difficult to make use of it than is illustrated in the following example, where F is invertible on $(0,\infty)$ only:

Example 6.3. The distribution function of an $\exp(\lambda)$-distribution with mean $1/\lambda$ $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ has the invers

$$F^{-1}(y) = -\lambda^{-1}\ln(1-y) \quad \text{for } y \in (0,1).$$

Hence, if ξ is a random number, then Corollary 6.1 shows that

$$\eta = F^{-1}(\xi) = -\lambda^{-1}\ln(1-\xi) \quad \text{is} \quad \exp(\lambda)\text{-distributed.}$$

This give us a recepy for simulating $\exp(\lambda)$-distributed random variables in a computer.

It is easy to simulate random variables with a discrete distribution:

Theorem 6.2 (Table Method). *Let f be the probability density function for a discrete random variable with the possible value $\{y_1, y_2, y_3, \ldots\}$. If ξ is a random number, then the random variable*

$$\eta = \begin{cases} y_1 & \text{if} \quad 0 > \xi \le f(y_1) \\ y_2 & \text{if} \quad f(y_1) < \xi \le f(y_1) + f(y_2) \\ y_3 & \text{if} \quad f(y_1) + f(y_2) < \xi \le f(y_1) + f(y_2) + f(y_2) + f(y_3) \\ & \vdots \end{cases}$$

is a discrete random variable with the possible value $\{y_1, y_2, y_3, \ldots\}$ and probability density function $f_\eta = f$.

Proof. One sees directly that the result is true. Alternatively, the theorem can be shown by application of Theorem 6.1. □

6.4.2 Simulation of normal distributed random variables

Normal distributed random variables can be simulated with Theorem 6.1, as the invers for the normal distribution function can be calculated numerically. However, sometimes it is desirable to have an alternative, more analytical algorithm, for simulation of normal random variates:

Theorem 6.3 (Box-Müller). *If ξ and η are independent random numbers, then we have*

$$Z \equiv \mu + \sigma\sqrt{-2\ln(\xi)}\cos(2\pi\eta) \quad N(\mu, \sigma^2) - \textit{distributed}$$

Proof. [4]For N_1 and N_2, independent $N(0,1)$-distributed, the two-dimensional vector (N_1, N_2) has radius $\sqrt{N_1^2 + N_2^2}$ that is distributed as the square-root of a $\chi(2)$-distribution. Moreover, it is a basic fact, that is easy to check, that a $\chi(2)$-distribution is the same thing as an $\exp(1/2)$-distribution.

By symmetry, the vector (N_1, N_2) has argument $\arg(N_1, N_2)$ that is uniformly distributed over $[0, 2\pi]$.

Adding things up, and using Example 6.3, it follows that, for ξ and η random numbers,

$$(N_1, N2) =_{\text{distribution}} \sqrt{-2\ln(\xi)}(\cos(2\pi\eta), \sin(2\pi\eta)). \quad \square$$

6.5 Software

The computer assignment is to be done in C and in Matlab. More precisely, you have to write the code in C and then incorporate it into Matlab with MEX. For a short example of how it can be done, see homepage → programming → C interface in Matlab.

[4] This proof is not important for the understanding of the rest of the material.

6.6 Laboration

1. It is well known that the number π can be calculated numerically as the integral

$$\pi = \int_0^1 \frac{4}{1+x^2} dx.$$

   - Use Monte Carlo integration to approximate the integral numerically. Do this for several "sample sizes" n, for example $n = 100, 1000, 10000, ...$ Perform an error estimate pretending that the real value of π is unknown and compare it with the actual error calculated using the real value of π. Begin with plotting the function $f(x) = 4/(1+x^2)$ to get a feeling for how it behaves (do this in Matlab).
   - Pick two variance reduction techniques (whichever you want) and re-calculate the integral by applying those. Do this for the same n values as before. Do you get more accurate estimates? Why (why not)?

   Do the above by writing a function in C that takes in one "sample size" n and returns the integral and the error estimates. Call then this function from Matlab with different n values. You can write a separate function for the variance reduction or change the original one so that it returns both the simple estimate and the ones obtained through variance reduction.

2. (2p)

   - Do a Monte Carlo calculation of the integral

$$\int_0^1 B(t)dt,$$

   where B is the quite irregular function given by

$$B(t) = \sum_{k=0}^{n} \frac{\sqrt{8}}{\pi} \frac{\sin(\frac{1}{2}(2k+1)\pi t)}{2k+1} n_k \quad \text{for } t \in [0,1],$$

   for a large n, and $\{n_k\}_{k=1}^{n}$ independent normal N$(0,1)$-distributed. Actually, if one lets $n \to \infty$, $B(t)$ becomes Brownian motion (see Chapter 5). This function, or stochastic process rather, is known to be continuous, but not differentaible in a single point in the interval $[0,1]$!

   - Improve your program in from Task 1 (the one where you calculated π) by adding one more variance reduction technique (again, pick whichever you like). Do you get better estimates? Why (why not)?

3. (2p)

   - In many applications, it is of interest to study worst case scenarios, and the *expected shortfall* $\mathbf{E}\{S_X(u)\}$ is a measure that is commonly used, for that purpose. The definition of expected shortfall is the expectation, of a suitable *loss random variable* X, given that the loss is greater than a certain threshold u:

$$\mathbf{E}\{S_X(u)\} = \mathbf{E}\{X|X > u\}.$$

Expected shortfalls can be difficult to calculate analyiclly, but with Monte Carlo simulations things simplify.

Assume that an insurance company has found that the probability to have a flood is p, and that if a flood occurs, then the loss is exponential distributed with parameter λ. In other words, we have the loss $X = YZ$, where Y is a Bernoulli(p)-distributed random varaible, and Z is an exp(λ)-distributed random variable with mean $1/\lambda$, independent of Y.

Select $p = 0.1$, $1/\lambda = 3.4$ and $u = 10$, and use Monte Carlo simulation to estimate the expected shortfall $\mathbf{E}\{S_X(u)\}$. Give bounds on the error of the estimation.

- Add one more variance reduction technique to the program from Task 1 (it should now have four in total). Compare those three and the original estimate. Which one is the best? Why?