Lecture Notes for

MA5233, Advanced Scientific Computing

By

Liu Jian Guo

matjgl@math.nus.edu.sg

Department of Mathematics, National University of Singapore

March 2003

Chapter Four - Monte Carlo Methods

1 Introduction

In this chapter we will describe an overview of Monte Carlo Methods and their components. Monte Carlo Methods are numerical methods that can be described as statistical simulation methods that utilize sequences of random numbers to perform the simulation. In Monte Carlo Methods a physical process is quantized and simulated, therefore firstly a brief review of probability theory and statistics is given. Afterwards we discuss the random number sequences and their generation. Then the two basic Monte Carlo methods for integration are discussed, followed by the variance reduction techniques for increasing the efficiency. In the end we introduce the application of these methods, namely Brownian Motion and stochastic differential equations.

2 Basic Probability Theory and Statistics

2.1 Basic Concepts

The standard textbook will call *experiment* a physical or mathematical process with certain *outcomes*. A very common *experiment* will be "tossing an unbiased coin", whose *outcomes* are "head" or "tail". We assume the tossing is done fairly and that is what makes the outcome *random*. The outcome of an experiment is not always a number; for example, it can be "heads" or "tails". However, usually numerical values are associated with the outcomes of the experiment as follows:

Let A denote some *event* associated with some possible outcome of the experiment. Then:

$$P(A) = \frac{N(A)}{N},$$

where $N(A)$ is the number of outcomes leading to A and N is the total number of outcomes. $P(A)$ is the probability value of the occurrence of the event A. Probability is expressed on a scale between 0 and 1; a rare event has a probability close to 0, a very common event has a probability close to 1.

The probability of an event has been defined as its long-run relative frequency. A whole series of independent trials under identical conditions are conducted and $n(A)$ is the number of experiments in which A occurs. *Relative frequency* of the event A is the number , which for very large values of n approaches $P(A)$. The following result is known as *the Law of Large Numbers*:

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n},$$

where $0 \leq P(A) \leq 1$.

2.2 Combination of Events

A *sample space* refers to the collection of all possible outcomes of a random experiment. Two events A_1 and A_2 will be called *mutually exclusive* if they cannot occur simultaneously.

Addition Law of Probabilities: If $A_1, A_2, \ldots, A_n$ are n mutually exclusive events, we have:

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k)$$

If the events are not mutually exclusive, then we have the following result: Let

$$P_1 = \sum_{i=1}^{n} P(A_i), \quad P_2 = \sum_{1\leq i<j\leq n} P(A_iA_j), \quad P_3 = \sum_{1\leq i<j<k\leq n} P(A_iA_jA_k), \ldots$$

Then,

$$P\left(\bigcup_{k=1}^{n} A_k\right) = P_1 - P_2 + P_3 - P_4 + \cdots \pm P_n$$

2.3 Dependent Events

In observing the outcomes of an experiment one is often interested in how the outcome of an event A is influenced by that of another event B. This situation is described by the *conditional probability* of A on the hypothesis B and defined as:

$$P(A/B) = \frac{P(AB)}{P(B)}$$

Often we observe experiments that are *independent*, that is, the outcome of one experiment has no influence on the outcome of the other. If A_1 and A_2 are independent events, then the probability that they occur at the same time is given by

$$P(A_1A_2) \sim \frac{n(A_1A_2)}{n} = \frac{n(A_1A_2)}{n(A_2)}\frac{n(A_2)}{n} \sim P(A_1/A_2) * P(A_2) = P(A_1) * P(A_2)$$

2.4 Random Variables

A *random variable* ξ is a numerical function $\xi = f(E_i)$, whose value depends on the event E_i. It is *random* because the event E_i is random and it is *variable* because the assignment of the value may vary over the real axis. Random variables are very useful because they allow the quantification of random processes and facilitate numerical manipulations. A random variable ξ can be:

1. **discrete** (meaning that it has a *discrete distribution*): if ξ takes only a finite or countably infinite number of discrete values and we have

$$\sum_x P(\xi = x) = 1$$

   i.e., a coin is tossed ten times. The random variable X is the number of tails that are noted. X can only take the values $0, 1, \ldots, 10$, so X is a discrete random variable.

2. **continuous** (meaning that it has a *continuous distribution*):

$$P(x_1 < \xi \leq x_2) = \int_{x_1}^{x_2} p_\xi(x)\, dx$$

   where $p(x)$ is a nonnegative integrable function and is called *probability density (or distribution) function (pdf)*. The probability of any range of values is the corresponding area under the *pdf* curve, and the area under the entire curve is always 1. i.e., a light bulb is burned until it burns out. The random variable Y is its lifetime in hours. Y can take any positive real value, so Y is a continuous random variable.

If $f(x)$ is a probability density (or distribution) function then it must obey two conditions:

1. The total probability for all possible values of the random variable X is 1;
2. The probability function can never be negative: $p(x) \geq 0$ for all x.

Cumulative Distribution Function (CDF) gives the probability that the random variable is less than or equal to x. Formally it is defined as:

1. For ξ - discrete: CDF $= \sum_{-\infty}^{x} p(x)$, and is a step function.
2. For ξ - continuous: CDF $= \int_{-\infty}^{x} p_\xi(x)\, dx$, and is a continuous monotone increasing function.

Uniform Distribution Uniform distributions model (some) continuous and (some) discrete random variables. The values of a uniform random variable are uniformly distributed over an interval. i.e., if buses arrive at a given bus stop every 15 minutes, and you arrive at the bus stop at a random time, the time you wait for the next bus to arrive could be described by a uniform distribution over the interval from 0 to 15.

A discrete uniform distribution has equal probability at each of its n values.

If ξ is a continuous random variable with a probability function

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

Then ξ is said to have a *uniform distribution*.

Joint Probability Distribution Consider now two random variables ξ_1, ξ_2. First suppose they are *discrete*, and then they have a *joint probability distribution*, where ξ_1 and ξ_2 range over all possible values. This is characterized by:

$$P\{(\xi_1, \xi_2) \in \Omega\} = \sum_{(\xi_1,\xi_2)\in\Omega} p(\xi_1, \xi_2).$$

Now, suppose that they are *continuous*. Then, we have:

$$P\{(\xi_1, \xi_2) \in \Omega\} = \int_\Omega p(\xi_1, \xi_2)\, d\xi_1\, d\xi_2.$$

Expected Value The expected value (or population mean) of a random variable indicates its average or central value. It is a useful summary value (a number) of the variable's distribution. Stating the expected value gives a general impression of the behavior of some random variable without giving full details. It is also called *"the first moment"*.

The mathematical *expectation* or *mean value* of a discrete random variable x is the following quantity:

$$E[X] = \sum_x xp(x)$$

i.e., when a die is thrown, each of the possible faces 1, 2, 3, 4, 5, 6 (the ξ's) has a probability of 1/6 (the $p(\xi)$'s) of showing. The expected value of the face showing is therefore:

$$\mu = E[X] = (1 * 1/6) + (2 * 1/6) + (3 * 1/6) + (4 * 1/6) + (5 * 1/6) + (6 * 1/6) = 3.5$$

Notice that, in this case, $E[X]$ is 3.5, which is not a possible value of X.

Given the *discrete* random variable X. Consider a new random variable $Y = g(X)$ defined as some function of X. The mathematical expectation of Y in terms of probability distribution of X is:

$$E(Y) = \sum_x g(x)p(x)$$

More generally, if ξ is a random variable defined as a function of two other random variables X and Y as $g(X, Y)$ with joint probability distribution $p(x, y)$. Then, the expected value of ξ is calculated by:

$$E[\xi] = E[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y).$$

If X and Y are independent random variables, then $E[X, Y] = E[X]E[Y]$.

For the case of *continuous* random variables the expectation will have the following definitions:

$$E[X] = \int_{-\infty}^{\infty} xp(x)\, dx,$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)\,dx,$$
$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)p(x,y)\,dx\,dy,$$

The expected value of the *uniformly distributed* random variable in the interval $[a, b]$ is:

$$E[X] = \int_a^b x\frac{1}{b-a}\,dx = \frac{a+b}{2}$$

Variance The variance of a random variable is a non-negative number which gives an idea of how widely spread the values of the random variable are likely to be; the larger the variance, the more scattered the observations on average. Stating the variance gives an impression of how closely concentrated round the expected value the distribution is; it is a measure of the 'spread' of a distribution about its average value. It is also called *"the second moment"*.

Mathematically, Var (x), is the mean squared value $E[(X-\mu)^2]$ of the difference $(X-\mu)$, where $\mu = E[X]$ is the mean value of X. It follows that:

$$\text{Var } (X) = E(X^2) - \mu^2$$

For two independent random variables X and Y, we have the following property:

$$\text{Var } (X+Y) = \text{Var } (X) + \text{Var } (Y)$$

The value Var $(x) = \sigma^2(x)$ and σ is called the *standard deviation* of the variable x. The variance and standard deviation of a random variable are always non-negative.

Covariance The expected value of the product of the deviations of two random variables from their means: $E[(X-\mu_x)(Y-\mu_y)]$. Uncorrelated variables have zero covariance. The covariance of a variable with itself is its variance. Correlation is quantified by the *correlation coefficient* of the random variables x and y and is expressed as:

$$\text{corr } (X,Y) = \frac{E\left[(X-E[X])(Y-E[Y])\right]}{\sigma(X)\sigma(Y)}$$

If X and Y are independent, then corr $(X, Y) = 0$.

2.5 Important Probability Distributions

2.5.1 Binomial Distribution

Binomial distributions model (some) discrete random variables. Typically, a binomial random variable is the number of successes in a series of trials, for example, the number of 'heads' occurring when a coin is tossed 50 times.

Closely related to this topic is the concept of *Bernoulli trials*. They are the identical independent experiments in each of which an event A may occur with probability $p = P(A)$, or fail to occur with probability $q = 1 - p$. In the case of n consecutive Bernoulli trials,

each event can be described as a sequence of $1011\ldots0001$ consisting of n digits where success in the i^{th} trial is denoted by 1 and failure by 0 in the i^{th} place. We calculate: $P(\#\text{success } = k) = p^k q^{n-k}$.

Now consider the random variable ξ equal to the total number of successes in n Bernoulli trials.

$$P(\xi = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}$$

The probability distribution of ξ is known as **binomial distribution** and it is specified by two parameters p, probability of a single success, and n, the number of trials. For binomial distribution, we have the following:

$$E[\xi] = \mu = np \quad \text{and} \quad \text{Var } (\xi) = np(1-p)$$

For this distribution, the total number of trials is always fixed in advance and for each single trial there can be only two outcomes, namely: "success" and "failure". Probability of success is the same for all trials and they are independent from each other.

2.5.2 Poisson Distribution

Poisson distributions also are used to model (some) discrete random variables. Typically, a Poisson random variable is a count of the number of events that occur in a certain time interval or spatial area. For example, the number of cars passing a fixed point in a 5 minute interval, or the number of calls received by a switchboard during a given period of time.

A random variable ξ is said to have a *Poisson distribution* if

$$p_\xi(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ for } k = 0, 1, 2,$$

where $e = 2.718\ldots$ is the base of natural logarithm. The Poisson distribution is specified by the parameter λ , the mean value of ξ , that is $\lambda = E[\xi]$.

2.5.3 Normal Distribution

Normal distribution is used to model the continuous random variables although this requirement does not usually hold in practice. For example, height of people of the same gender at a given age for a given group is adequately described by a Normal random variable even though height values can be only positive. (They do not extend over the whole real line.)

A random variable has a *normal distribution (Gaussian)* if

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

The density is a symmetrical, bell-shaped curve, centered at its expected value μ. The variance is σ^2. In the equation on the left above, we have the simplest case of normal distribution known as *Standard Normal Distribution*, written as $N(0, 1)$, has expected value zero and variance one. More generally, the random variable with normal probability distribution is expressed by the equation on the right given above, in which case it has mean μ and variance σ^2.

Closely related is the concept of *confidence intervals*. *Confidence interval* is a random interval that has a known probability (the "confidence coefficient" or "confidence level") of including the true value of a parameter. Defines an interval within which the true population parameter is likely to lie. It can be thought of as a measure of the precision of a sample statistic.
Confidence coefficient (also known as the "confidence level".) is the level of confidence (e.g. 95%) associated with a confidence interval.

In a normally distributed sample 66% of the useful information is expected to lie within a standard deviation interval around the mean, and 99% of the data is within $[\mu \pm 3\sigma]$.

2.5.4 Central Limit Theorem

The *Central Limit Theorem* states that whenever a random sample of size n is taken from *any* distribution with mean μ and variance σ^2, then the sample mean will be *approximately* normally distributed with mean μ and variance σ^2/n. The larger the value of the sample size n, the better the approximation to the normal.

This is very useful when it comes to inference. For example, it allows us (if the sample size is fairly large) to use hypothesis tests, which assume normality even if our data appear non-normal. This is because the tests use the sample mean, which the Central Limit Theorem tells us will be approximately normally distributed. We illustrate this fact with the help of the following example:

Example Let X be the random variable $X = \max(T1, T2, T3)$, where the T_j are three independent uniformly distributed random variables in the interval $[0, 1]$. We will perform the following tasks:

First, we will show that the cumulative distribution function for X is $F(x) = x^3$ and the probability density function for X is $f(x) = 3x^2$. We need to find cumulative density function for $X = \max(T_1, T_2, T_3)$ where T_i are random variables in the interval $[0, 1]$. Since T_i are independent of each other we can write the following equality.

$$\text{CDF } (x) = P(X \leq x) = P(T_1 \leq x)P(T_2 \leq x)P(T_3 \leq x)$$

We know that T_i are uniformly distributed in the interval $[0, 1]$ then their CDF will be the identity function

$$\text{CDF }_T(x) = P(T_i \leq x) = x$$

Replacing this in the first equation we get

$$F(x) = x^3$$

Then probability density function would be simply the derivative of this function

$$f(x) = F'(x) = 3x^2$$

In order to support our result, we construct 10^4 samples of X using three uniformly distributed random numbers (T_i's in this case) and make a histogram with a bin size 0.01. Then, we compare the density function $f(x)$ with our histogram Figure 1 shows our histogram and the plot of our density function, $f(x)$.

As a second part of our example, we will consider the *ensemble average* of the sum of n independent random variables. Let S_n such a sum:

$$S_n = \sum_{k=1}^{n} X_k$$

Let the brackets $\langle \cdot \rangle$ denote an *ensemble average*. Then, the mean value μ_n is given by $\mu_n = \langle S_n \rangle = \frac{3}{4}n$, and the variance, $V_n = \langle (S_n - m_n)^2 \rangle = \frac{3}{80}n$. To verify these results, we have to find the mean for $S_n = \sum_{k=1}^{n} X_k$ which is equivalent to the expected value. Then using linearity of expectations we can get

$$\mu_n = \langle S_n \rangle = E[S_n] = \sum_{k=1}^{n} E[X_k]$$

Expected value of X_k can be computed using the pdf function $f(x)$ from the first part

$$E[X_k] = \int_0^1 x f(x)\, dx = \int_0^1 3x^3\, dx = \frac{3}{4}$$

Replacing this, we finally get

$$\mu_n = \frac{3}{4}n$$

The variance is

$$V_n = \sum_k \text{Var } (X_k) = \frac{3}{80}n$$

Finally, we will construct a histogram of the S_{100} values as defined above. The bin widths that we are going to consider will be chosen such that 100 equally spaced bins will capture approximately 99% of the data (this property follows from the central limit theorem, according to which three standard deviations shall capture approximately 99% of the whole sample). Figure 2 shows our histogram. In this figure, we compare it with a Gaussian density function with the appropriate mean and variance.

As it can be seen in this figure, the histogram resembles the behavior of the Gaussian density function (i.e., the normal distribution), which is in completely in accordance with the *central limit theorem*.

3 Random Number Generation

Random number generation is the procedure of creating numbers, which seem as samples randomly selected from a known distribution. Generating random numbers on a computer, which executes completely deterministic algorithms, cannot in any way yield true random numbers. Therefore, the process is called as *pseudo-random* or *quasi-random* number generation. There are various algorithms designed for this purpose and a random number generation method is considered to be *good* if it outputs a sequence of numbers that are uniformly distributed, statistically independent and reproducible. It is also desirable that the method will run fast and require minimum memory capacity.

3.1 Uniform random number generation

Linear congruential generators are the most commonly used method for generating uniform random numbers. These methods produce a sequence where each single number determines its successor with the following simple recursive formula

$$X_{i+1} = (aX_i + c) \bmod m$$

where the *multiplier* a, the *increment* c and the *modulus* m are nonnegative integers. The starting value in the recursion, X_0, is called the *seed*. Often c is taken to be 0, and in this case, the generator is called as a *multiplicative congruential generator* and can be as good as any linear congruential method if the multiplier and modulus are chosen exquisitely carefully. We generally want to generate random numbers uniformly distributed over the interval $[0, 1]$ - $U(0, 1)$- and this is accomplished by dividing the output of the congruential generator by the modulus m.

The maximum period or cycle length of a congruential method is m since X_i is determined by X_{i-1} and there are m possible different values for X's. In other words, such a sequence will recur as soon as a number is repeated and this happens after at most m iterations. If the constants m, a, and c are not chosen carefully the period will be much less than this. One simple rule says that a should be relatively prime to m for a sequence to have a period equal to m.

Linear congruential methods are widely used since they are extremely fast, requiring very few operations per call, but they have the disadvantage that it is not free of sequential correlation on successive calls. If k numbers at a time are used to plot points in a k dimensional space (with each coordinate between 0 and 1), then the points will not tend to *fill up* the k-dimensional space, but rather will lie on $(k-1)$-dimensional *planes*.

3.2 Non-uniform random number generation

Sampling of random variates from a non-uniform distribution is usually done by applying a transformation to uniform variates. The algorithms for these transformations differ in speed, in accuracy, in storage requirements, and in complexity of coding. Some methods apply to almost any distribution and hence are "universal". Some other methods are devised to generate particular distributions and only used to generate those specific distributions.

3.2.1 Inverse CDF Method

Let ξ be a random variable uniformly distributed over the interval $[0, 1]$; then $\eta = F^{-1}(\xi)$ will be a random variable with the cumulative density function F. We can generate any random distribution by changing the CDF function F. The proof is straightforward: If ξ is uniform random variable in the interval $[0, 1]$ then

$$P(\xi \leq x) = \begin{cases} 0, & \text{if } x \leq 0 \\ x, & \text{if } 0 < x \leq 1 \\ 1, & \text{if } 1 < x \end{cases}$$

We want $P(\eta \leq x) = F(x)$ so using the previous equation we can show

$$P(\xi \leq x) = x$$

$$P(\xi \leq F(x)) = F(x)$$
$$P(F^{-1}(\xi) \leq x) = F(x)$$
$$P(\eta \leq x) = F(x)$$

This method is illustrated in Figure 3.

Inverse CDF method has the advantage that basic relationships among a set of uniform deviates (such as order relationships) may result in similar relationships among the set of deviates from the other distribution. However, this method is not widely used as its simplicity suggest because it is relatively difficult inverse of some distribution functions of interest or it is very slow to evaluate the inverse function.

The inverse CDF method also applies to discrete distributions, but of course, we cannot take the inverse of the distribution function. Suppose we want to generate random variables for the discrete random variable η, which has mass points $m_1 < m_2 < m_3 < \ldots$ with probabilities $p_1, p_2, p_3, \ldots$ and with the following distribution function

$$F(x) = \sum_{i, m_i \leq x} p_i$$

To use the inverse CDF method for this distribution, we first generate a realization then ξ of a uniform random variable. We then deliver the realization of the target distribution as η, where η satisfies the relationship

$$F(\eta_-) < \eta \leq F(\eta)$$

This is illustrated in Figure 4. Without loss of generality, we often assume that the mass points of a discrete uniform distribution are the integers $1, 2, 3, \ldots$. The special case in which there are k mass points and they all have equal probability is called the *discrete uniform distribution*, and use of the inverse CDF method is particularly simple: the value is $[\xi_k]$.

3.2.2 Box-Muller Method

Box-Muller method is a way to obtain exact normal random variables by means of a one-to-one transformation of two $U(0,1)$ random variables. Let ξ_1 and ξ_2 be two independent $U(0,1)$ variables and define

$$\theta = 2\pi\xi_1, \qquad R = \sqrt{-2\pi \ln \xi_2}$$

then, Box-Muller method showed that

$$X = R\cos\theta, \quad Y = R\sin\theta$$

are two independent random variables with mean 0 and standard deviation 1 - $N(0,1)$. While this method is valid in principle, there is a serious difficulty if 1 and 2 are actually adjacent random numbers produced by a linear congruential generator. Due to the fact that ξ_1 would depend on ξ_2, it can be shown that the generated variates x and y are not truly independently normally distributed

As an example, we want to study whether the resulting random variables (X, Y) for Box-Muller algorithm are members of the following joint probability density function:

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

Now, the first task is to capture random variable pairs (X, Y) from the Box-Muller algorithm and determine how many *falls* within a pre-defined region on the xy-plane. We will consider only the central *bin*, which is centered over the origin of the xy-plane.

Let the xy-plane be subdivided into bins:

$$B_{jk} = \{(x, y) : |x - x_j| \leq h/2, \text{ and } |y - y_k| \leq h/2\}$$

centered at $x_j = jh$ and $y_k = kh$.

If we take n pairs, (X_t, Y_t), $t = 1, \ldots, n$, then the bin counts are

$$N_{jk} = \#\{(X_t, Y_t) \in B_{jk} : 1 \leq t \leq n\}$$

The expected counts are

$$\langle N_{jk} \rangle = n \int_{B_{jk}} f(x, y)\, dx\, dy$$

where f is given by (1).

The midpoint rule for the integral in (2) is second order accurate. In terms of local truncation error, this means that

$$\int_{B_{jk}} f(x, y)\, dx\, dy = h^2 f(x_j, y_k) + O(h^4).$$

In two dimensions, we do not want h so small because that would create too many bins. Therefore, we want a more accurate integration formula -rather than $O(h^4)$. To do so, we can expand $f(x, y)$ in a *Taylor series* at the bin center (x_j, y_k) so that

$$\int_{B_{jk}} f(x, y)\, dx\, dy = h^2 f(x_j, y_k) + \frac{h^4}{24}(f_{xx} + f_{yy})(x_j, y_k) + O(h^6).$$

Now, if we want to compute the h value for the central bin where $x_j = 0$ and $y_k = 0$ such that there are 1000 points in the bin out of total $n = 10^6$ experiments so putting the values in the following equation and solving the polynomial we find the h value.

$$\langle N_{00} \rangle / n = \int_{B_{00}} f(x, y)\, dx\, dy = h^2 f(0, 0) + \frac{h^4}{24}(f_{xx} + f_{yy})(0, 0)$$

$$1000/10^6 = h^2 \frac{1}{2\pi} - \frac{h^4}{24} \frac{1}{\pi}$$

$$h = .079287409$$

Actually there are four different roots to the polynomial but two of them are negative and cannot be used.

As the second task of our example, we will show the following moment equation holds.

$$\langle X^{2p} \rangle = \frac{1}{2\pi} \int_{-\infty}^{\infty} x^{2p} e^{-x^2/2}\, dx = 1 * 3 * 5 * \cdots (2p-1)$$

If we integrate by parts and repeat this process until the exponent of x in the integral becomes 0 we find the following

$$\langle X^{2p} \rangle = \frac{1}{2\pi} \int_{-\infty}^{\infty} x^{2p} e^{-x^2/2}\, dx = \frac{1}{2\pi} \Pi_{k=1}^{p} (2k-1) \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \Pi_{k=1}^{p} (2k-1)$$

We can find the variance using this result

$$\text{Var } (X^{2p}) = E[X^{4p}] - E[X^{2p}]^2$$

Furthermore, we can find the required number of samples to achieve an accuracy of 1% by using the following equation

$$\frac{\sigma}{\mu\sqrt{N}} = 0.01$$

where σ is the standard deviation and μ is the mean of samples.

4 Monte Carlo Methods

The name *Monte Carlo* was applied to class of mathematical methods first by scientists working on the development of nuclear weapons in Los Alamos in the 1940s. A definition of a Monte Carlo method would be one that involves deliberate use of random numbers in a calculation that has the structure of a stochastic process. By a stochastic process, we mean a sequence of states whose evolution is determined by random events.

Perhaps the earliest documented use of random sampling to find the solution to an integral is that of Compte de Buffon. In 1777, he described the following experiment. A needle of length L is thrown at random onto a horizontal plane ruled with straight lines a distance d apart ($d > L$). What is the probability p that the needle will intersect one of the lines? He performed the experiment of throwing the needle many times. He has also derived the value of p by mathematical analysis and showed that

$$p = \frac{2L}{\pi d}$$

Some years later, Laplace suggested that this idea could be used to estimate π from throws of a needle. This is indeed a Monte Carlo determination of π, however the rate of convergence is very slow.

Every Monte Carlo simulation that leads to quantitative results may be regarded as estimating the value of a multiple integral. There are many deterministic numerical integration methods for computation of ordinary integrals with well-behaved integrands. However if the function is not well behaved (does not have continuous derivatives of moderate order) or the integration is over multiple dimensions then integration methods like trapezoidal or Simpson rule are less convenient than Monte Carlo method. In such cases even though the result may be less accurate, Monte Carlo method is computationally easier.

In this section we will explain two simple techniques for computing one dimensional integrals by a Monte Carlo method. The extension to higher dimensions is sometimes obvious and sometimes rather difficult, depending on the subtlety of the technique under discussion. In contrast, for conventional numerical integration, this extension is nearly always difficult to compute.

4.1 Hit or Miss Monte Carlo

Consider the problem of calculating the following one dimensional integral where for simplicity we assume integrand $g(x)$ is bounded $0 < g(x) < c$;

$$I = \int_a^b g(x)\,dx$$

Let (X, Y) be a random vector uniformly distributed over the rectangle Ω which is defined as $\Omega = \{(x, y) : a \leq x \leq b, 0 \leq y \leq c\}$ (see Figure 5).

According to Figure 5 (X, Y) is in the area S if $Y \leq g(X)$. If p is the probability that (X, Y) falls into S then p can be calculated by;

$$p = \frac{\text{area } S}{\text{area } \Omega} = \frac{1}{c(b-a)} \int_a^b g(x)\,dx$$

In order to estimate p we can use N random vectors (X_i, Y_i) and count the number of hits or misses. Let N_H be the number of hits meaning N_H vectors fall into S then $\hat{p}$ is an estimator of p where

$$\hat{p} = \frac{N_H}{N}$$

It can be easily seen from (5) and (6) that an estimator for I is θ, where

$$\theta = c(b-a)\frac{N_H}{N}$$

Notice that each trial is a Bernoulli trial with hit probability p and expected value of N Bernoulli trials with successes probability p is np. Therefore,

$$E[\theta] = c(b-a)E\left[\frac{N_H}{N}\right] = pc(b-a) = I$$

and this shows that θ is an unbiased estimator of I. We also need to calculate the variance of θ in order to state how accurately θ estimates I. Remember that the variance for N Bernoulli trials with successes probability p is $np(1-p)$. Using this and (5), we get

$$\text{Var }(\theta) = c^2(b-a)^2\text{Var }(\hat{p}) = c^2(b-a)^2\frac{p(1-p)}{N} = \frac{I}{N}\left(c(b-a) - I\right)$$

4.2 Crude Monte Carlo

A second way to compute the integral in (4) is to represent it as an expected value of some random variable. Because of this approach, this method is also known as Sample-Mean Monte Carlo method. An equivalent of (4) is

$$I = \int_a^b \frac{g(x)}{f(x)} f(x)\, dx$$

where $f(x)$ is any p.d.f such that $f(x) > 0$ when $g(x) \neq 0$. Then,

$$I = E\left[\frac{g(X)}{f(X)}\right]$$

Let's say $f(x)$ is a uniform probability distribution function between a and b, then

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

We can estimate $E[g(X)]$ by its sample mean of size N and using (12) we get the estimator θ for integral I;

$$\theta = \frac{b-a}{N} \sum_{i=1}^{N} g(X_i)$$

θ is an unbiased estimator for I since we used sample mean for $g(x)$. Using the definition of variance, we can compute the variance of θ;

$$\text{Var } (\theta) = \frac{1}{N}\left[(b-a)\int_a^b g^2(x)\, dx - I^2\right]$$

4.3 Efficiency of Monte Carlo Methods

We now define the relative efficiency of two Monte Carlo methods. Let the methods call for χ_1 and χ_2 units of computing time, respectively and let the resulting estimates θ have variances of σ_1^2 and σ_2^2. Then the efficiency of method two relative to method one is defined as

$$\epsilon = \frac{\chi_1 \sigma_1^2}{\chi_2 \sigma_2^2}$$

If ϵ is less than 1 then we can conclude that the first method is more efficient then second method. Using the definition above let's compare the two methods we have seen so far. We will assume computing time for both *hit or miss* and *crude Monte Carlo* methods are approximately same and we will concentrate on the variances of these two methods to determine the efficiency. As we have shown in the previous sections the variances for these two methods were

$$\sigma_1^2 = \frac{I}{N}\left(c(b-a) - I\right) \quad \text{and } \sigma_2^2 = \frac{1}{N}\left[(b-a)\int_a^b g^2(x)\, dx - I^2\right]$$

where σ_1^2 is the variance of *hit or miss* Monte Carlo Method and σ_2^2 is the variance of *crude* Monte Carlo Method.

$$\sigma_1^2 - \sigma_2^2 = \frac{b-a}{N}\left[cI - \int_a^b g^2(x)\,dx\right] = \frac{b-a}{N}\int_a^b (c - g(x))g(x)\,dx$$

Remembering the assumption that $g(x) < c$; we can conclude

$$\sigma_1^2 - \sigma_2^2 \geq 0.$$

Thus ϵ is greater than or equal to 1 and we say that crude Monte Carlo method is more efficient than hit-or-miss Monte Carlo method.

Historically hit-or-miss methods were the most popular methods since they were easy to interpret particularly if explained in a graphical set up. Another reason is efficiency was not the main concern in those days. The comparison between hit-or-miss and crude Monte Carlo methods also illustrates the general principle of Monte Carlo work; *if at any point of a Monte Carlo calculation we can replace an estimate by an exact value, we shall* reduce the sampling error in the final result.

It is interesting to note that when estimating the integral we do not need to know the function $g(x)$ explicitly. We need only the value of $g(x)$ for a given point.

Another point is when comparing the efficiencies of two methods, if the variances are not known, we can replace them with the sample variance S^2 and use the same efficiency formula.

5 Variance Reduction Techniques

As we described earlier, Monte-Carlo methods can be attractive to use especially in estimating multi-dimensional integrals. It is shown that each integral can be represented as an *expected value* and the problem of estimating an integral by the Monte-Carlo method is equivalent to the problem of estimating an unknown expected value.

Given an integration problem, the *variance* in the problem denotes the fact that how much the current estimation is closer to the actual integration value. Most of the time, we can reduce the variance - i.e., we can make a better estimation - by using the information that is currently known about the problem. This technique is called *Variance Reduction*. There are two extreme points that we need to consider:

- Variance reduction cannot be achieved if no information is available about the problem, and
- If we have complete information about the problem, then the variance is equal to zero. Of course, there is no need for variance reduction is no need simulation.

One important aspect of variance reduction is that how to gain the information that is used in the process. The most common way - also the most relevant one to our current subject of interest - is to make a series of direct (crude) Monte-Carlo simulations of the process. Information gathering is achieved through the successive simulations; information

gathered in one simulation is used to define the variance reduction methods that will refine and improve the efficiency of the second one and so on.

In the subsequent subsections, we will describe various variance reduction techniques that can be used to refine Monte-Carlo simulations and make them more efficient. During our discussions, we will assume that we are given a problem of estimating the multi-dimensional integral

$$I = \int g(x)\,dx, \qquad x \in D \subset R^n$$

Furthermore, we assume that $\int g(x)^2\,dx$ exists.

5.1 Importance Sampling

Given a sample, the basic idea behind this technique is to concentrate the distribution of the sample points in the parts of the interval that are of the *most importance* instead of spreading them our evenly. Therefore, the heart of the technique is to define the *importance function*.

We can represent the integral I as follows.

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

In the equation above, X is a random vector with p.d.f. $f(x)$ such that $f(x) \geq 0$ for each $x \in D \subset R^n$. The function $f(x)$ is called the **importance sampling distribution**.

Now, let $\xi = \frac{g(X)}{f(X)}$ be a random variable which is an estimator of I, with the variance

$$\text{Var }(\xi) = \int \frac{g(x)^2}{f(x)}\,dx - I^2$$

In order to estimate the integral I, we take a sample $X_1, \ldots, X_N$ from p.d.f. $f(x)$ and substitute its values in the sample-mean formula

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \frac{g(X_i)}{f(X_i)}$$

One important issue is how to choose the distribution of the random vector X such that the variance of ξ is minimized, which is the same as to minimize the variance of σ^2. It has been shown that the minimum of Var (ξ) is equal to

$$\text{Var }(\xi) = \left(\int |g(x)|\,dx\right)^2 - I^2$$

when the following p.d.f. is used

$$f(x) = \frac{|g(x)|}{\int |g(x)|\,dx}$$

The proof is sketched at (Rubinstein, 1981). Unfortunately, although it is correct, this method cannot be used for computing Var (ξ) because it involves the computation of I. The

problem is that the reason that we are doing importance sampling and using Monte-Carlo simulations is the computation of I is too hard. Now, if we already know the result of I, then we do not need to all these techniques at all.

Therefore, we need a pseudo-optimal, but useful method for finding the distribution $f(x)$. For example, the variance can be reduced if $f(x)$ is chosen in order to have a similar behavior as $|g(x)|$. Now, of course, we can have trouble in finding such a p.d.f. if $g(x)$ is not a well-behaved function. To overcome these difficulties, we may choose the sample distribution function such that the sample points $X_1, \ldots, X_N$ are chosen from the region D' where $D' \subset D$ and $D' = \{x | g(x) \geq 0\}$. This is the same as defining

$$f_x(x) = \frac{1}{\text{area D'}} \begin{cases} 1, & \text{if } g(x) \geq 0 \\ 0, & \text{if } g(x) < 0 \end{cases}$$

5.2 Correlated Sampling

Correlated Sampling is one of the most powerful variance reduction techniques. The primary goal of making a simulation study is to determine the changes in successive simulations so that their effects can be used to reduce the variance between our successive estimations - i.e., the ultimate goal, of course, is to find the actual solution to the problem. However, in the successive simulations by using crude Monte-Carlo method, the difference (changes) in those simulations may be too small to provide useful information, while the variance between the successive simulations is significant. Mostly, this is because of the fact that each simulation assumes independent random variables. Instead of using independent random variable, we can use the same random values in each simulation, in which case the results will be highly correlated and the variance reduction will be achieved largely. Thus, the aim of correlated sampling is to produce highly positive correlation between two similar processes so that the variance of the difference is considerably smaller than it would be if the two processes were statistically independent.

There is no general procedure that can be implemented in correlated sampling. However, in the following two situations, correlated sampling can be successfully employed.

- The value of a small change in a system is to be calculated.
- The difference in a parameter in two or more similar cases is of more interest than is absolute value.

To give an example, suppose that we want to estimate

$$\Delta I = I_1 - I_2$$

where

$$I_1 = \int g_1(x)\, dx, \quad x \in D_1 \subset R^n$$

$$I_2 = \int g_2(x)\, dx, \quad x \in D_2 \subset R^n$$

Then, the procedure for correlated sampling is as follows:

- Generate $X_1, \ldots, X_n$ from $f_1(x)$ and $Y1, \ldots, Y_n$ from $f_2(x)$, where $f_i(x)$ are p.d.f.'s.

- Estimate ΔI using

$$\Delta\mu = \mu_1 - \mu_2$$

where μ_i's are calculated as in (4) by using the same techniques we have used for *importance sampling*.

Now, the variance of $\Delta\mu$ is

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\text{cov}\ (\mu_1, \mu_2)$$

where

$$\sigma_1^2 = E[\mu_1 - I_1]$$
$$\sigma_2^2 = E[\mu_2 - I_2]$$
$$\text{cov}\ (\mu_1, \mu_2) = E[(\mu_1 - I_1)(\mu_2 - I_2)]$$

Now, if μ_1 and μ_2 are statistically independent, then

$$\text{cov}\ (\mu_1, \mu_2) = 0 \ \text{ and } \sigma^2 = \sigma_1^2 + \sigma_2^2$$

However, if the random variables X and Y are positively correlated and if $\frac{g_1(x)}{f_1(x)}$ is similar to $\frac{g_1(x)}{f_2(x)}$ in shape, then the random variables μ_1 and μ_2 will be positively correlated, that is $\text{cov}\ (\mu_1, \mu_2) > 0$, and the variance of $\Delta\mu$ may be greatly reduced.

5.3 Control Variates

In this technique, instead of estimating an expected value directly, the difference between the problem of interest and some analytical model is considered. Just like control variables in scientific method, control variates provide a base model for comparison and for calculating the difference between the current problem and the controlled one.

A variance reduction technique based on this approach can employ either only one or more than one control variates. In this note, we will focus only on the former case. The latter can be formulated as a general case of the former and is left as an exercise to the reader.

Now, we will describe how variance reduction is done by using only one *control variate*. Let y be a random sequence. There are two cases. In the first case, a *control variate* C for Y is defined as a random variate, which is correlated with Y and its expectation, μ_c, is known. The control variate C is used to build an estimate for that has a smaller variance than the estimator Y. Then, for any β,

$$Y(\beta) = Y - \beta(C - \mu_c),$$

is an unbiased estimator for μ. Then,

$$\text{Var}\ [Y(\beta)] = \text{Var}\ [Y] - 2\beta\text{cov}\ (Y, C) + \beta^2\text{Var}\ (C)$$

Hence if

$$2\beta\text{cov}\ (Y, C) > \beta^2\text{Var}\ (C)$$

then we achieve the required reduction in the variance of μ. The value of β that minimizes Var $[Y(\beta)]$ can be found as

$$\beta^* = \frac{\text{cov } (Y, C)}{\text{Var } (C)}$$

and the minimum variance is equal to

$$\text{Var } [Y(\beta^*)] = 1 - \rho_{YC}^2$$

Where ρ_{YC} is the **correlation coefficient** between Y and C. Thus, the more Y and C are correlated, the greater reduction is obtained in the variance.

In the second case, we have a control variate whose expected value is unknown, but is equal to μ, that is, $E[C] = E[Y] = \mu$. In this case, any linear combination

$$Y(\beta) = \beta Y + (1 - \beta)C,$$

will be an unbiased estimator of μ, and similarly, if Y and C are correlated, variance reduction will be achieved.

5.4 Stratified Sampling

In stratified techniques, the domain D of the independent variable x is broken into subintervals such that a random variable is sampled from every subinterval. For example, consider the one-dimensional integral

$$I = \int_0^1 g(x)\, dx, \qquad x \in D \subset R$$

Now, the simplest stratification is to divide $(0, 1)$ into M equal subintervals. An x is then chosen in each interval in succession,

$$x_m = \frac{1 - \xi_m}{M}$$

where $\ell = (m-1)(\text{mod } M)+1$, $m = 1, 2, \ldots, NM$, N is the size of the random sample. The ℓ's will cycle through the integer values 1 through M, and the x_m will be drawn at random in the ℓ^{th} interval. An estimator for I is given by

$$\xi = \frac{1}{N} \sum_n \left(\frac{1}{M} \sum_m g(x_m) \right)$$

Here, $\frac{1}{M} \sum_m g(x_m)(1/M)$ are the terms that are statistically independent. The variance is straightforward to calculate from these.

We can also choose the subinterval of different sizes. An estimator g_k for the integral I can be defined that allows varying subinterval size,

$$g_k = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\alpha_j - \alpha_{j-1}) \frac{1}{n_j} g(\alpha_{j-1} + (\alpha_j - \alpha_{j-1})\xi_{ij})$$

The notation ξ_{ij} indicates that a new random number is chosen for every combination of (i, j); the random number is then mapped onto the interval $(\alpha_j - \alpha_{j-1})$ in which n_j samples are to be used. A particular subinterval is sampled n_j times, and the mean of the g's evaluated within $(\alpha_j - \alpha_{j-1})$ is multiplied by the size of the interval. As it can be seen, this is a generalization of the trapezoidal rule for integration since the size of the interval changes.

The variance on a subinterval is proportional to the number of samples taken within the subinterval

$$n_j^2 \sim (\alpha_j - \alpha_{j-1}) \int_{\alpha_{j-1}}^{\alpha_j} g^2(x)\, dx - \left(\int_{\alpha_{j-1}}^{\alpha_j} g^2(x)\, dx \right)^2$$

The variance can be reduced with respect to n_j to determine the optimal value for n_j.

5.5 Antithetic Variates

In this technique, we seek two unbiased estimator Y' and Y'' for some unknown expected value I (in our case I is the unknown integral), having strong negative correlation. Note that $\frac{1}{2}(Y' + Y'')$ will be an unbiased estimator of I with variance

$$\text{Var } \left(\frac{Y' + Y''}{2} \right) = \frac{1}{4}\text{Var } (Y') + \frac{1}{4}\text{Var } (Y'') + \frac{1}{2}\text{cov } (Y', Y''),$$

and it follows that if cov (Y', Y'') is strongly negative, then the method of antithetic variates can be effective reducing the variance.

As an example, consider the integral,

$$I = \int_0^1 g(x)\, dx$$

which is equal to

$$I = \frac{1}{2} \int_0^1 (g(x) + g(1 - x))\, dx$$

The estimator of I is then,

$$Y = \frac{Y' + Y''}{2} = \frac{g(U) + g(1 - U)}{2}$$

Y is an unbiased estimator of I, because both $Y' = g(U)$ and $Y'' = g(1 - U)$ are unbiased estimators of I. If we take a sample size of N from the uniform distribution for example, we find

$$\mu = \frac{1}{2N} \sum_{i=1}^{N} (g(U_i) + g(1 - U_i))$$

Now, this estimator for I is more efficient than the estimator, μ_c, with crude Monte-Carlo Method with $a = 0$ and $b = 1$ only if

$$\text{Var } (\mu) \leq \frac{1}{2}\text{Var } (\mu_c)$$

6 Applications of Monte Carlo Methods

6.1 Brownian Motion

Brownian motion is a sophisticated random number generator, based on a process in plants discovered by Robert Brown in 1827. He found that small particles suspended in a fluid were in continuous movement and thus, described it as Brownian motion. His discovery did not receive much attention for a long time, until before the turn of the 20th century when Guoy's conviction and research (that Brownian motion constituted a clear demonstration of the existence of molecules in continuous motion) brought it to the attention of the Physics world. However, all nineteenth-century research remained at the qualitative level.

It was only in 1905 when a quantitative analysis was brought about, where Einstein succeeded in stating the mathematical laws governing the movements of particles on the basis of the principles of the kinetic-molecular theory of heat. According to this theory, bodies of microscopically visible size suspended in a liquid will perform irregular thermal movements called Brownian molecular motion, which can be easily observed in a microscope. Brownian motion was then more generally accepted because it could now be treated as a practical mathematical model. As a result, many scientific theories and applications related to it have been developed and they subsequently play major roles in the world of Physics.

A random variable $B(t)$ (see Figure 6) is called a Brownian Motion if it satisfies the following properties:

- $B(0) = 0$,
- $B(t)$ is a continuous function of t;
- B has independent, normally distributed increments: if

  $$0 = t_0 < t_1 < t_2 < \cdots < t_n$$

  and

  $$Y_1 = B(t_1) - B(t_0), Y_1 = B(t_2) - B(t_1), \cdots, Y_n = B(t_n) - B(t_{n-1})$$

  then
  - $Y_1, Y_2, \cdots, Y_n$ are independent
  - $E(Y_j) = 0, \forall j$,
  - Var $(Y_j) = t_j - t_{j-1}, \forall j$.

A standard Brownian motion can be simulated by taking

$$X_0 = 0, \qquad X_k = X_{k-1} + \sqrt{\Delta t} Z_k$$

Z_k are identically independently distributed standard normal and $X_k \sim X(k\Delta t)$.

To give an example, let's consider a Brownian motion simulation - represented as (1) and (2) performed for 10 steps and each time step is $\Delta t = 0.1$. Our aim is to calculate the probability P of a trajectory extending to $X_k \geq 3$ for some k:

$$P = \Pr\ (X_k \geq 3 \text{ for some } t_k \leq 1)$$

Now, first we will calculate the precision that can be achieved using $N = 10^6$ trajectories. Then, P can be approximated by the following formula:

$$P = \frac{k}{N}$$

where k is the number of trajectories in which $X_k > 3$ for some $t_k < 1$. The accuracy can be then calculated by the formula :

$$\text{accuracy } = \frac{\sqrt{p(1-p)}}{p\sqrt{N}}$$

The following is the estimated value of P and the accuracy we can get using 10^6 trajectories.

$$\text{accuracy } = 0.02445733375403$$

$$P = 0.00166900000000 \pm 0.02445733375403$$

As it can be seen for its calculation, the P value above is a very rough approximation of the real P value. The P value can be expressed as an integral and can be computed using *importance sampling*. Let a trajectory be represented by the vector $\bar{Z}$:

$$Z = \{z_1, z_2, \ldots, z_{10}\}$$

The corresponding differential elements are $dZ = dz_1 dz_2 \ldots dz_{10}$.The local of the individual steps within the trajectory are represented by x:

$$X = \{x_1, x_2, \ldots, x_{10}\}$$

Each element of x is defined as follows:

$$x_0 = 0, \qquad x_k = x_{k-1} + \sqrt{\Delta t} Z_k$$

The probability was defined as

$$P = \Pr\ (X_k \geq 3 \text{ for some } t_k \leq 1)$$

Let the decision function be

$$\phi(Z) = \begin{cases} 1, & \text{if } x_k \geq 1 \text{ for some } t_k \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

We also know that the probability density function for z is the joint probability function for ten independent Gaussians which is

$$f(Z) = \frac{1}{(2\pi)^5} e^{-(z_1^2 + \cdots + z_{10}^2)/2}$$

Combining all together we can show the following

$$P = \Pr\ (X_k \geq 3 \text{ for some } t_k \leq 1) = E[\phi(Z)] = \int \cdots \int \phi(z_1, \cdots, z_{10}) f(z_0, \cdots, z_{10}) dz_1, \cdots, dz_{10}$$

For importance sampling we define two functions g and h as following

$$h(Y) = e^{5a^2 - a(y_1 + \cdots + y_{10})}, \qquad g(y) = \frac{1}{(2\pi)^5} e^{-((z_1-a)^2 + \cdots + (z_{10}-a)^2)/2}$$

Now we can show the following

$$g(Y) = \frac{f(Y)}{h(Y)}$$

We can equivalently say $h(y) = f(y)/g(y)$. Then we can implement importance sampling by the following equation

$$P = \int \phi(Z) f(Z)\, dZ = \int \phi(Y) h(y) g(Y)\, dY$$

Since $g(y)$ is the joint pdf for $Y_1, \ldots, Y_{10}$ values the previous integral can be interpreted as the moment for random variable $\phi(y)g(y)$. So we can find the expected value by generating random y values and finding the mean by

$$P = \frac{1}{N} \sum_{j=1}^{N} \phi(Y^j) g(Y^j)$$

where Y^j are each ten dimensional vectors.

6.2 Stochastic Differential Equations

Stochastic methods have become increasingly important in the analysis of a broad range of phenomena in natural sciences and economics. Many processes are described by differential equations where some of the parameters and/or the initial data are not known with complete certainty due to lack of information, uncertainty in the measurements, or incomplete knowledge of the mechanisms themselves. To compensate for this lack of information one introduces stochastic noise in the equations, either in the parameters or in the initial data, which results in stochastic differential equations.

Now let's see an example of Brownian motion simulation to stochastic differential equations in finance. Specifically, we will consider using simulations to estimate the price of a European call option. A call option of European type is a contract where the holder has the option of buying an asset A at time T with a pre-determined price X. It is clear then that the holder will decide to purchase the asset A at time T if the market price S of A at time T is larger than X.

The price of a call option is equal to the current expected value of the pay-off function ϕ which is defined as

$$\phi(S) = \max(0, S_T - X)$$

or

$$c_t = e^{-r(T-t)} E[\phi(S)]$$

One way to estimate the value of the call is to simulate a large number of sample values of S_T according to the assumed underlying process, and find the estimated call price as the average of the simulated values. By a Law of Large Numbers, this average will converge to

the actual call value, where the rate of convergence will depend on the way we simulate the sample path, and how many simulations we perform.

The assumed underlying price process is the geometric Brownian Motion

$$\frac{dS}{S} = rdt + \sigma dZ$$

or

$$dS = rSdt + \sigma SdZ$$

To simulate this, we need to discretise the process, by splitting the time between 0 and T into a number of periods of length Δt,

$$dt \sim \Delta, \quad dS \sim DeltaS = S_t - S_{t-1}$$

To simulate the term due to the Brownian motion (dZ), remember that one of the defining characteristics of Brownian motion is that the increment for any finite period of length t is distributed normally with mean zero and variance t. Since the variance is t, the standard deviation is $\sqrt{t}$. Thus, if Z_t is normally distributed with mean zero and unit variance, we can simulate the discrete process by

$$S_t = S_{t-1} + rS_{t-1}\Delta t + \sigma S_{t-1}\sqrt{\Delta t}Z_t$$

This expression can be used to generate a series of sample paths, which will approximate the sample path of a Brownian motion process as the sampling interval Δt decreases. If we generate standard normal random variables Z_t's and use them to find a value for S_T and repeat this process many times, the mean of the sample S_T's will approximate the actual value of S_T.

References

1. Gentle, J.E., *Random Number Generation and Monte Carlo Methods*, Springer Pub., 1998.

2. Hammersley, J. M., Handscomb, D.C., *Monte Carlo Methods*, John-Wiley and Sons, 1986.

3. Kalos, M.H., Whitlock, P.A., *Monte Carlo Methods, Volume I:Basics*, Wiley-Interscience Publication, John-Wiley & Sons, 1986.

4. Morgan, B.J.T., *Elements of Simulations*, Chapman and Hall Pub., 1984.

5. Rozanov, Y.A., *Probability Theory, A Concise Course*, Dover Publications, Inc. New York, 1969.

6. Rubinstein, R.Y., *Simulation and the Monte Carlo Method*, Wiley Series in Probability and Mathematical Statistics, 1981.

7. Teukolsky, S.A., Vetterling, W.T., Flannery B.P., *Numerical Recipes in C, Second Edition*, Cambridge University Press, 1999.

8. BUBL Information Service *http://bubl.ac.uk/link/p/probability.htm*

9. Introduction to Monte Carlo Methods *http://csep1.phy.ornl.gov/mc/mc.html*

Figure caption

Figure 1. Comparison of the density function with the actual data distribution.

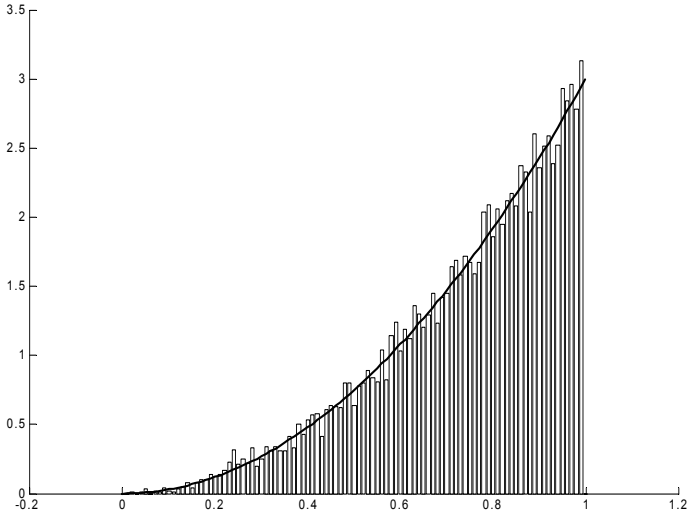
Figure 2. Comparison between the empirical histogram of S100 values and a Gaussian density function.

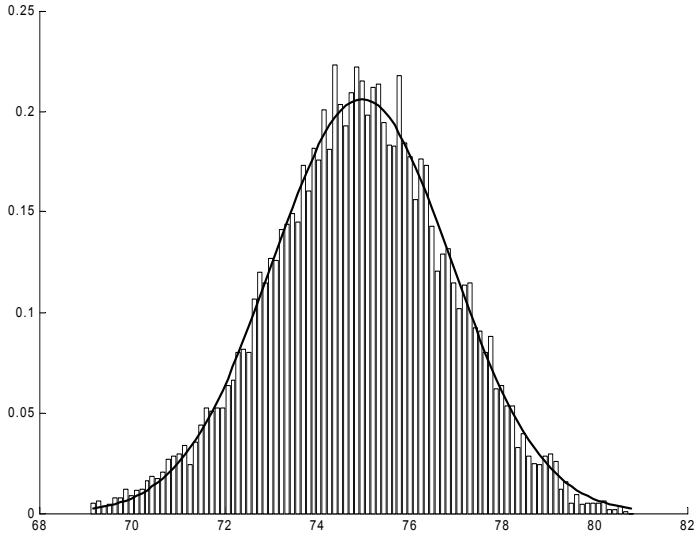
Figure 3. Inverse CDF method to convert a uniform random number to a number from a continuous distribution.

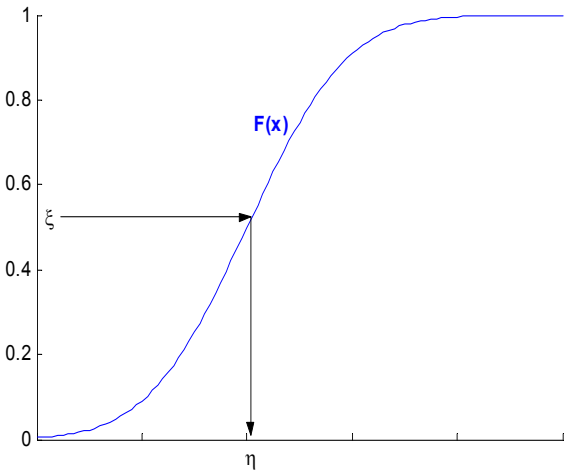
Figure 4. Inverse CDF method to convert a uniform random number to a number from a discrete distribution.

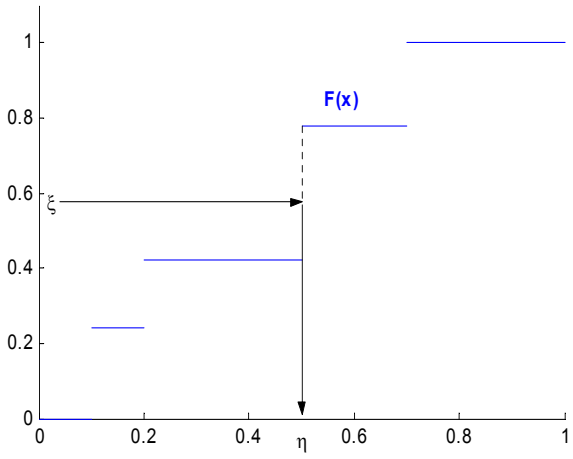
Figure 5. Hit or miss.

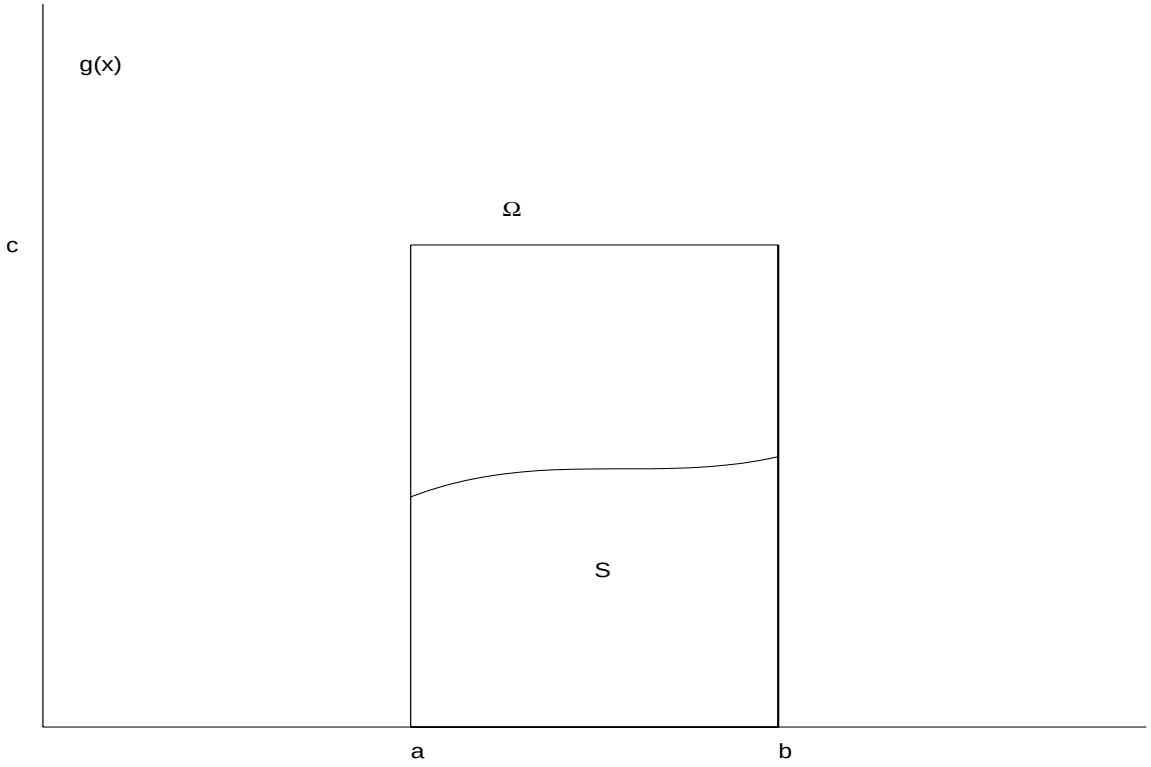
Figure 6. Continuous time Brownian motion.

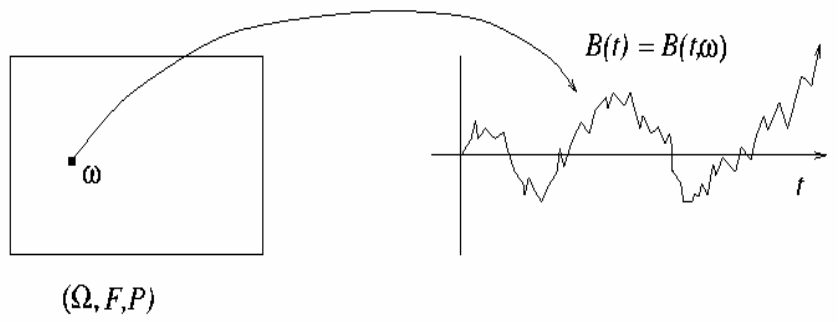

3.5
3
2.5
2
1.5
1
0.5
0
-0.2
0
0.2
0.4
0.6
0.8
1
1.2


0.25
0.2
0.15
0.1
0.05
0
68
70
72
74
76
78
80
82


1
0.8
0.6
0.4
0.2
0
F(x)
ξ
η


F(x)
ξ
η
0
0.2
0.4
0.6
0.8
1


g(x)
Ω
c
S
a
b


B(t) = B(t,ω)
ω
t
(Ω, F, P)