Chapter 5

Generating Random Variables

The success of a Monte Carlo scheme depends on its ability to repeatedly sample from a given distribution. At the heart of such sampling procedures is a mechanism for generating iid sequences of random numbers that are uniformly distributed on $[0, 1]$. In reality, such "random" numbers are not truly random. They are part of a very long sequence of numbers that are generated by completely deterministic computer algorithms to mimic the behavior of genuine random numbers. For this reason, such numbers are said to be *pseudo-random*. For all practical purposes, one can treat this sequence as if it were truly random. In MATLAB®, the function "`rand`" is used for generating such uniform random numbers.

Even though MATLAB has a large built-in collection of algorithms to generate samples from commonly used probability distributions, occasionally one would like to sample from some very specialized distributions. In this chapter, we will discuss some commonly used techniques for generating random variables or random vectors: the inverse transform method, the acceptance-rejection method, and the Cholesky factorization method for multivariate normal distributions.

5.1 Inverse Transform Method

The inverse transform method is by far the simplest and most commonly used method for generating random variables. The key observation is that any random variable can be represented as a function of a uniformly distributed random variable.

Consider the problem of generating samples from a probability distribution whose cumulative distribution function F is known. Define the *in-*

verse of F to be

$$F^{-1}(u) = \min\{x : F(x) \geq u\}$$

for every $u \in [0,1]$. See Figure 5.1.

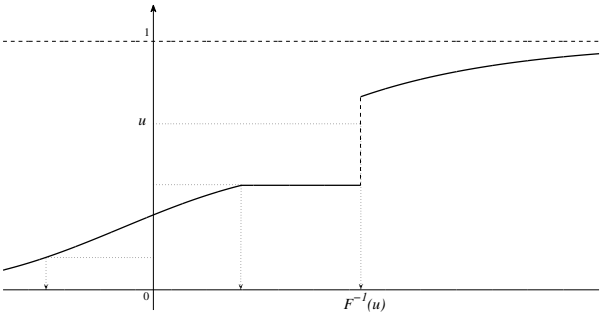


Figure 5.1: Inverse transform method.

Theorem 5.1. *If U is uniform on $[0,1]$, then the cumulative distribution function of $X = F^{-1}(U)$ is F.*

PROOF. Observe that for any $x \in \mathbb{R}$,

$$\{U < F(x)\} \subseteq \{X \leq x\} \subseteq \{U \leq F(x)\}.$$

It follows that

$$F(x) = \mathbb{P}\{U < F(x)\} \leq \mathbb{P}(X \leq x) \leq \mathbb{P}(U \leq F(x)) = F(x),$$

or

$$\mathbb{P}(X \leq x) = F(x).$$

This completes the proof. ■

The inverse transform method is based on the preceding theorem. Once one obtains the inverse of F, then $X = F^{-1}(U)$ will have the desired distribution if U is uniformly distributed on $[0,1]$.

Pseudocode for the inverse transform method:

generate a sample U from the uniform distribution on $[0,1]$
set $X = F^{-1}(U)$.

Example 5.1. A Pareto distribution with parameters (a, b) has the density

$$f(x) = \begin{cases} ab^a/x^{a+1} & \text{if } x \geq b, \\ 0 & \text{otherwise.} \end{cases}$$

Here a and b are both positive constants. Determine F^{-1}.

SOLUTION: It is straightforward to compute the cumulative distribution function for the Pareto distribution from its density:

$$F(x) = \begin{cases} 1 - (b/x)^a & \text{if } x \geq b, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$F^{-1}(u) = \frac{b}{(1-u)^{1/a}}$$

for all $u \in [0, 1]$. ■

Example 5.2. Consider a discrete random variable X with distribution

$$\mathbb{P}(X = x_i) = p_i,$$

where $x_i \in \mathbb{R}$ and $p_i > 0$ are given such that

$$x_1 < x_2 < \cdots < x_m, \quad \sum_{i=1}^{m} p_i = 1.$$

Determine the inverse of its cumulative distribution function.

SOLUTION: Define $q_0 = 0$ and $q_k = p_1 + \cdots + p_k$ for $k = 1, \ldots, m$. The cumulative distribution function of X is

$$F(x) = q_j, \quad j = \max\{1 \leq i \leq m : x \geq x_i\},$$

with the convention that $\max\{\emptyset\} = 0$. It is straightforward to verify that

$$F^{-1}(u) = x_i, \quad \text{if } q_{i-1} < u \leq q_i$$

for $u \in [0, 1]$. ■

5.2 Acceptance-Rejection Method

The acceptance-rejection method is another commonly used technique for generating random variables. Unlike the inverse transform method, it is not restricted to the univariate probability distributions.

For illustration, suppose that we are interested in generating samples from a target probability distribution with density f. Let g be an *alternative* density function, from which we know how to generate samples. Furthermore, assume that there exists a constant c such that

$$\frac{f(x)}{g(x)} \leq c \tag{5.1}$$

for all x. The algorithm is as follows. Generate a trial sample, say Y, from the density g. This sample will be accepted with probability

$$\frac{1}{c}\frac{f(Y)}{g(Y)}.$$

A trial sample that is not accepted will be discarded. Repeat this procedure until the desired sample size is reached.

Pseudocode for the acceptance-rejection method:

$(*)$ generate a trial sample Y from the density g
generate a sample U from the uniform distribution on $[0,1]$
accept Y if

$$U \leq \frac{1}{c}\frac{f(Y)}{g(Y)};$$

otherwise, discard Y and go to step $(*)$.

It is not difficult to explain why the acceptance-rejection method produces samples from the density f. Let X be a sample from the algorithm. Then

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A \mid \text{sample } Y \text{ is accepted}).$$

By the law of total probability in Theorem 1.11,

$$\mathbb{P}(Y \text{ is accepted}) = \int \frac{1}{c}\frac{f(y)}{g(y)} g(y) dy = \frac{1}{c}$$

$$\mathbb{P}(Y \in A,\ Y \text{ is accepted}) = \int_A \frac{1}{c}\frac{f(y)}{g(y)} g(y)\, dy = \frac{1}{c}\int_A f(y)\, dy.$$

It follows that

$$\mathbb{P}(X \in A) = \int_A f(y)\, dy,$$

and hence the density of X is f. This calculation also shows that the overall probability of accepting a trial sample is $1/c$, or on average c samples from g are needed to generate one sample from f. Therefore, it is preferable that the constant c be close to 1 so that only a small fraction of samples from g will be rejected or wasted.

Example 5.3. Consider a bounded univariate density function f that is zero outside some interval $[a, b]$. It is very easy to design an acceptance-rejection algorithm that uses the uniform distribution on $[a, b]$ as the alternative sampling distribution. That is, $g(x) = 1/(b-a)$ for all $a \le x \le b$. The smallest, hence the optimal, constant c that satisfies the requirement (5.1) is

$$c = \max_{x \in [a,b]} \frac{f(x)}{g(x)} = (b-a) \max_{x \in [a,b]} f(x).$$

Carry this idea over to higher dimensions and construct an acceptance-rejection algorithm to generate samples that are uniformly distributed on the unit disc

$$\left\{(x, y) : x^2 + y^2 \le 1\right\}.$$

SOLUTION: The density function of the uniform distribution on the unit disc is given by

$$f(x, y) = \begin{cases} 1/\pi & \text{if } x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let g be the probability density function of the uniform distribution on the rectangle

$$[-1, 1] \times [-1, 1].$$

That is,

$$g(x, y) = \begin{cases} 1/4 & \text{if } -1 \le x, y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The constant c is given by

$$c = \max_{-1 \le x, y \le 1} \frac{f(x, y)}{g(x, y)} = \frac{4}{\pi} \approx 1.273.$$

The acceptance probability of a trial sample (X, Y) from the density g is

$$\frac{1}{c}\frac{f(X,Y)}{g(X,Y)} = \begin{cases} 1 & \text{if } X^2 + Y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $c \approx 1.273$, on average 1.273 samples from the uniform distribution on $[-1,1] \times [-1,1]$ are needed to generate one sample from the uniform distribution on the unit disc. Note that to generate a sample (X, Y) from the uniform distribution on $[-1,1] \times [-1,1]$, it suffices to generate X and Y independently from the uniform distribution on $[-1,1]$.

Pseudocode:

$(*)$ generate two independent samples X and Y uniformly from $[-1,1]$
accept (X, Y) if $X^2 + Y^2 \le 1$, otherwise reject (X, Y) and go to $(*)$.

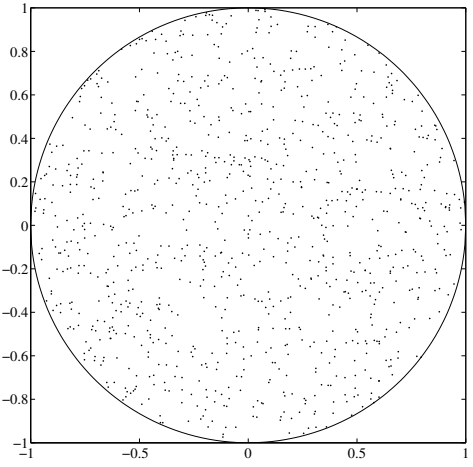


Figure 5.2: Uniform samples on a unit disc.

In the actual simulation, a total of 1255 trial samples are used to generate 1000 samples in the unit disc. ■

Example 5.4. Construct an acceptance-rejection scheme to generate samples from a Gamma distribution with density

$$f(x) = 2\sqrt{\frac{x}{\pi}}e^{-x}, \quad x \ge 0.$$

The alternative sampling distribution is assumed to be exponential. Which exponential distribution yields the most efficient sampling scheme?

SOLUTION: Suppose that the alternative sampling distribution is exponential with rate λ. That is, $g(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The smallest constant c that satisfies (5.1) is

$$c = \max_{x \geq 0} \frac{f(x)}{g(x)} = \max_{x \geq 0} \frac{2}{\sqrt{\pi}} \frac{\sqrt{x} e^{-x}}{\lambda e^{-\lambda x}}.$$

Note that the maximum is infinity if $\lambda \geq 1$. Therefore, we should only consider those $\lambda < 1$. In this case, the maximum is attained at $x^* = 0.5/(1-\lambda)$ and

$$c = \sqrt{\frac{1}{2e\pi\lambda^2(1-\lambda)}}.$$

Since on average c samples from g are needed to generate a sample from f, the optimal λ^* should minimize c, or equivalently, maximize $\lambda^2(1-\lambda)$. It follows that $\lambda^* = 2/3$, and the corresponding c is approximately 1.257.

Pseudocode:

$(*)$ generate a sample Y from the exponential distribution with rate $2/3$
generate a sample U from the uniform distribution on $[0, 1]$
accept Y if

$$U \leq \frac{1}{c} \frac{f(Y)}{g(Y)} = \sqrt{\frac{2eY}{3}} e^{-Y/3}$$

otherwise, discard Y and go to step $(*)$.

The simulation results are presented in the histogram in Figure 5.3 with $\lambda = 2/3$. On average, 1.257 samples from the exponential distribution are needed to generate one sample of this Gamma distribution. The sample size is 10000, and in total 12488 trial samples are drawn from the exponential distribution. ■

5.3 Sampling Multivariate Normal Distributions

Multivariate normal distributions are commonly used in financial engineering to model the joint distribution of multiple assets. Sampling from such distributions becomes less straightforward when the components are correlated; see the spread call option pricing problem considered in Example 4.3. In this section, we discuss a general scheme based on the Cholesky

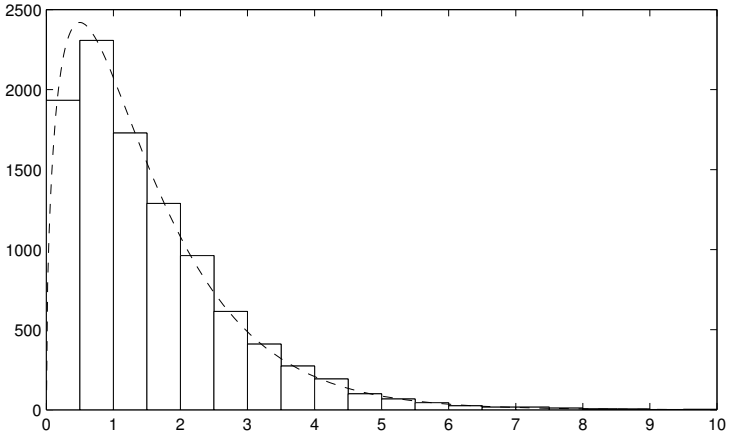


Figure 5.3: Samples of a Gamma distribution.

factorization. It is assumed that we are able to draw independent samples from the one-dimensional standard normal distribution; see Exercise 5.7.

Consider the d-dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Without loss of generality, we assume that Σ is nonsingular (see Remark 5.1). Recall the following property. Let Z be a d-dimensional standard normal random vector. Then for any $d \times d$ matrix A,

$$X = \mu + AZ \tag{5.2}$$

is jointly normal with mean μ and covariance matrix AA'. Therefore, if one can

(a) sample from the d-dimensional standard normal distribution,

(b) find a matrix A such that $AA' = \Sigma$,

then (5.2) leads to an algorithm for generating samples from $N(\mu, \Sigma)$.

Note that (a) can be easily achieved. Indeed, if $Z_1, \ldots, Z_d$ are independent standard normal variables, then $Z = (Z_1, \ldots, Z_d)'$ is a d-dimensional standard normal random vector. The answer to (b) is not unique. There are many matrices A that satisfy $AA' = \Sigma$. A particularly convenient choice is given by the *Cholesky factorization* of Σ, which assumes that A is a lower

triangular matrix:

$$A = \begin{bmatrix} A_{11} & 0 & 0 & \cdots & 0 \\ A_{21} & A_{22} & 0 & \cdots & 0 \\ A_{31} & A_{32} & A_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{d1} & A_{d2} & A_{d3} & \cdots & A_{dd} \end{bmatrix}.$$

The advantage of choosing a lower triangular matrix is that one can derive explicit formulas for A_{ij}, and these formulas can be evaluated numerically by a simple recursion.

Omitting the details of derivation, we give the formula for a solution to this Cholesky factorization:

$$A_{11} = \sqrt{\Sigma_{11}}, \quad A_{j1} = \Sigma_{j1}/A_{11}, \quad j > 1,$$

$$A_{ii} = \sqrt{\Sigma_{ii} - \sum_{k=1}^{i-1} A_{ik}^2}, \quad A_{ji} = \left(\Sigma_{ji} - \sum_{k=1}^{i-1} A_{jk}A_{ik}\right)/A_{ii}, \quad j > i \geq 2.$$

Pseudocode for Cholesky factorization:

set $A = 0$ [$d \times d$ zero matrix]
set $v = 0$ [$d \times 1$ zero vector]
for $i = 1, 2, \ldots, d$
    for $j = i, \ldots, d$
        set $v_j = \Sigma_{ji}$
        for $k = 1, \ldots, i-1$
            set $v_j = v_j - A_{jk}A_{ik}$
        set $A_{ji} = v_j/\sqrt{v_i}$
return A.

Pseudocode for generating one sample from $N(\mu, \Sigma)$:

find a matrix A such that $AA' = \Sigma$ through Cholesky factorization
generate independent samples $Z_1, \ldots, Z_d$ from $N(0, 1)$
set $Z = (Z_1, \ldots, Z_d)'$
set $X = \mu + AZ$.

Example 5.5. Some option payoffs depend on multiple assets. Assume that under the risk-neutral probability measure the prices of these underlying assets are all geometric Brownian motions

$$S_t^{(i)} = S_0^{(i)} \exp\left\{\left(r - \frac{1}{2}\sigma_i^2\right)t + \sigma_i W_t^{(i)}\right\}, \quad i = 1, \ldots, d,$$

where $W = (W^{(1)}, W^{(2)}, \cdots, W^{(d)})$ is a d-dimensional Brownian motion with covariance matrix $\Sigma = [\Sigma_{ij}]$ such that $\Sigma_{ii} = 1$ for all i. Consider an outperformance option with maturity T and payoff

$$\left(\max\left\{c_1 S_T^{(1)}, \ldots, c_d S_T^{(d)}\right\} - K\right)^+.$$

Estimate the price of this option.

SOLUTION: The price of the option is the expected value of the discounted payoff:

$$v = E\left[e^{-rT}\left(\max\left\{c_1 S_T^{(1)}, \ldots, c_d S_T^{(d)}\right\} - K\right)^+\right].$$

In order to estimate v, we need to generate samples of $(S_T^{(1)}, \cdots, S_T^{(d)})$, or equivalently, those of $(W_T^{(1)}, \cdots, W_T^{(d)})$, which is a jointly normal random vector with mean 0 and covariance matrix $T\Sigma$.

Pseudocode:

find a matrix A such that $AA' = \Sigma$ through Cholesky factorization
for $i = 1, \ldots, n$
  generate independent samples $Z_1, \ldots, Z_d$ from $N(0,1)$
  set $Z = (Z_1, \ldots, Z_d)'$
  set $Y = AZ$
  for $k = 1, \ldots, d$
    set $S_k = S_0^{(k)} \exp\left\{(r - \sigma_k^2/2)T + \sigma_k\sqrt{T}Y_k\right\}$
  set $H_i = e^{-rT}(\max\{c_1 S_1, \ldots, c_d S_d\} - K)^+$
compute the estimate $\hat{v} = \dfrac{1}{n}(H_1 + \cdots + H_n)$
compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}$.

The numerical results are reported in Table 5.1 for an outperformance option with $d = 4$ underlying assets. The parameters are given by

$$S_0^{(1)} = 45, \quad S_0^{(2)} = 50, \quad S_0^{(3)} = 45, \quad S_0^{(4)} = 55, \quad r = 0.02, \quad T = 0.5,$$

$$\sigma_1 = \sigma_2 = \sigma_3 = 0.1, \ \sigma_4 = 0.2, \ \Sigma = \begin{bmatrix} 1.0 & 0.3 & -0.2 & 0.4 \\ 0.3 & 1.0 & -0.3 & 0.1 \\ -0.2 & -0.3 & 1.0 & 0.5 \\ 0.4 & 0.1 & 0.5 & 1.0 \end{bmatrix}.$$

Table 5.1: Monte Carlo simulation for an outperformance option

	Sample size $n = 2500$			Sample size $n = 10000$		
Strike price K	50	55	60	50	55	60
M.C. Estimate	6.9439	3.5159	1.5406	7.0714	3.4197	1.4730
S.E.	0.1250	0.1009	0.0711	0.0624	0.0501	0.0346

The Cholesky factorization of Σ yields

$$A = \begin{bmatrix} 1.0000 & 0 & 0 & 0 \\ 0.3000 & 0.9539 & 0 & 0 \\ -0.2000 & -0.2516 & 0.9469 & 0 \\ 0.4000 & -0.0210 & 0.6069 & 0.6864 \end{bmatrix},$$

which satisfies $AA' = \Sigma$. ■

Remark 5.1. Let $X = (X_1, \ldots, X_d)'$ be a jointly normal random vector with mean μ and covariance matrix Σ. When Σ is singular, there exists a subset of the components of X whose covariance matrix is nonsingular, and such that every other component of X can be written as a linear combination of the components within this subset. Therefore, simulating X amounts to simulating this subset of components, which is a jointly normal random vector itself with a nonsingular covariance matrix.

Exercises

Pen-and-Paper Problems

5.1 Use the inverse transform method to generate samples from the following distributions. Write down the pseudocode.

(a) The uniform distribution on $[a, b]$.

(b) The exponential distribution with rate λ.

(c) The Weibull distribution with parameters (α, β) whose density is

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(d) The Cauchy distribution whose density is

$$f(x) = \frac{1}{\pi} \frac{1}{(1+x^2)}, \quad x \in \mathbb{R}.$$

5.2 A random variable X is said to be *geometric with parameter p* if it takes values in $\{1, 2, \ldots\}$ and

$$\mathbb{P}(X = i) = (1-p)^{i-1} p, \quad i = 1, 2, \ldots.$$

Show that if Y is an exponential random variable with rate $\lambda = -\log(1-p)$, then (denote by $[x]$ the integer part of x)

$$X = 1 + [Y]$$

is geometric with parameter p. Use this observation to generate samples of X from the uniform distribution on $[0, 1]$. Write down the pseudocode.

5.3 Write down the pseudocode for generating samples of a discrete random variable with infinitely many possible values $x_1 < \cdots < x_n < \cdots$ and

$$\mathbb{P}(X = x_n) = p_n.$$

You may want to use the inverse transform method and the loop command "while" in MATLAB.

5.4 Consider a mixture probability distribution whose cumulative distribution function is

$$F(x) = \sum_{i=1}^{m} p_i F_i(x).$$

Here F_i is a cumulative distribution function itself for each i and $p_1, \ldots, p_m$ are some positive numbers such that

$$\sum_{i=1}^{m} p_i = 1.$$

Assume that we know how to generate samples from F_i for each i. Write down an algorithm for generating samples from the mixture F. *Hint:* Suppose that $Y_1, \ldots, Y_m$ are independent and the cumulative distribution function of Y_i is F_i for each i. Let I be an independent random variable such that $\mathbb{P}(I = i) = p_i$. What is the cumulative distribution function of Y_I?

5.5 Suppose that we wish to generate samples from a probability distribution with density

$$f(x) = \begin{cases} \frac{1}{2}x^2 e^{-x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

by the acceptance-rejection method. The alternative density function is chosen to be

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

for some $\lambda \in (0, 1)$. Determine the best λ that minimizes the average number of trial samples needed to generate a sample from f.

5.6 Assume that X has density f. Design an acceptance-rejection scheme to draw samples of X conditional on $X > a$ for some given level a, with f as the alternative sampling density. Write down the pseudocode. On average, how many trial samples from f are needed to generate a desired sample? Do you think the assumption that X has a density is really necessary?

5.7 Let θ be uniform on $[0, 2\pi]$ and $R = \sqrt{2S}$ where S is an exponential random variable with rate one. Assume that θ and S are independent. Show that

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

are independent standard normal random variables. This is the *Box–Muller method* of generating standard normal random variables. *Hint:* Compute the joint density of X and Y using polar coordinates.

5.8 It happens quite often that one wishes to draw samples from a density function f that takes the form $f(x) = Ch(x)$, where h is a known nonnegative function and C is the *unknown* normalizing constant that satisfies

$$\int_{\mathbb{R}} f(x)\,dx = C\int_{\mathbb{R}} h(x)\,dx = 1.$$

Notationally, it is denoted by $f(x) \propto h(x)$. Let g be a density function from which one knows how to draw samples. Assume that for some positive constants k_1 and k_2,

$$k_1 \leq \frac{h(x)}{g(x)} \leq k_2$$

for every x. Show that

$$\max_x \frac{f(x)}{g(x)} \leq \frac{k_2}{k_1}.$$

Use this observation to design acceptance-rejection schemes to draw samples from the following distributions:

$$f(x) \propto e^{-x^2}(1 - e^{-\sqrt{1+x^2}}), \quad f(x) \propto (1 + e^{-x^2})(1 + x^2)^{-1}.$$

This method applies to higher dimensional distributions as well.

5.9 Argue that a Cholesky factorization for a 2×2 covariance matrix of the form

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

is given by

$$A = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

That is, A is lower triangular and $AA' = \Sigma$. This is exactly the matrix C obtained in Example 4.3.

MATLAB® Problems

5.A Write a function that uses the inverse transform method to draw one sample of a discrete random variable X with

$$\mathbb{P}(X = i) = p_i, \quad i = 1, \ldots, m,$$

where p_i are all positive constants and $p_1 + \cdots + p_m = 1$. The function should include the parameters m and $p = (p_1, \ldots, p_m)$ as the input. Test your algorithm by generating 10000 samples for

$$m = 4, \quad p = (0.1, 0.2, 0.3, 0.4).$$

Compare the empirical frequencies with p.

5.B Assume that there are $(d + 1)$ underlying assets, whose prices under the risk-neutral probability measure are geometric Brownian motions

$$S_t^{(i)} = S_0^{(i)} \exp\left\{\left(r - \frac{1}{2}\sigma_i^2\right)t + \sigma_i W_t^{(i)}\right\}, \quad i = 1, \ldots, d+1,$$

where $W = (W^{(1)}, \ldots, W^{(d+1)})$ is a $(d + 1)$-dimensional Brownian motion with covariance matrix $\Sigma = [\Sigma_{ij}]$ such that $\Sigma_{ii} = 1$ for every i. Write a function to estimate the price of an exchange option with maturity T and payoff

$$X = \left[\sum_{i=1}^{d} c_i S_T^{(i)} - S_T^{(d+1)}\right]^+.$$

The function should have input parameters d, r, T, $\sigma_1, \ldots, \sigma_{d+1}$, $c_1, \ldots, c_d$, Σ, $S_0^{(1)}, \ldots, S_0^{(d+1)}$, and the sample size n. Report your estimate and standard error for the price of the option, given

$$d = 3, \;\; r = 0.05, \;\; T = 1, \;\; c_1 = c_2 = c_3 = 1/3, \;\; \sigma_1 = \sigma_2 = 0.2,$$

$$\sigma_3 = \sigma_4 = 0.3, \;\; \Sigma = \begin{bmatrix} 1 & 0.1 & 0.2 & 0.2 \\ 0.1 & 1 & 0.3 & -0.3 \\ 0.2 & 0.3 & 1 & 0.5 \\ 0.2 & -0.3 & 0.5 & 1 \end{bmatrix},$$

$$S_0^{(1)} = S_0^{(2)} = S_0^{(3)} = 50, \;\; S_0^{(4)} = 45, \;\; n = 10000.$$

5.C Assume that under the risk-neutral probability measure the prices of two underlying assets are geometric Brownian motions

$$S_t^{(i)} = S_0^{(i)} \exp\left\{\left(r - \frac{1}{2}\sigma_i^2\right)t + \sigma_i W_t^{(i)}\right\}, \quad i = 1, 2,$$

where $W = (W^{(1)}, W^{(2)})$ is a two-dimensional Brownian motion with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

for some $-1 < \rho < 1$. Consider a two-asset barrier option with maturity T and payoff

$$\left[S_T^{(1)} - K\right]^+ \cdot 1_{\left\{\min\left(S_{t_1}^{(2)}, \ldots, S_{t_m}^{(2)}\right) \geq b\right\}}$$

where $0 < t_1 < \cdots < t_m = T$ are given dates. Write a function to estimate the price of this option. The function should have input parameters $S_0^{(1)}$, $S_0^{(2)}$, r, σ_1, σ_2, ρ, K, b, T, m, $(t_1, \ldots, t_m)$, and the sample size n. Report your estimate and standard error for

$$S_0^{(1)} = 50, \;\; S_0^{(2)} = 40, \;\; r = 0.03, \;\; \sigma_1 = 0.2, \;\; \sigma_2 = 0.4, \;\; \rho = 0.2,$$

$$K = 50, \;\; b = 38, \;\; T = 1, \;\; m = 50, \;\; t_i = iT/m, \;\; n = 10000.$$

5.D Suppose that under the risk-neutral probability measure, the stock price is a geometric Brownian motion with jumps:

$$S_t = S_0 \exp\left\{\left(\bar{r} - \frac{1}{2}\sigma^2\right)t + \sigma W_t + \sum_{i=1}^{N_t} Y_i\right\},$$

where W is a standard Brownian motion, $N = \{N_t : t \geq 0\}$ is a Poisson process with rate λ, and Y_i's are iid double exponential random variables with density

$$f(x) = p \cdot \eta_1 e^{-\eta_1 x} 1_{\{x \geq 0\}} + (1 - p) \cdot \eta_2 e^{\eta_2 x} 1_{\{x < 0\}}$$

for some $p \in (0,1)$ and positive constants η_1, η_2. Assume that W, N, and $\{Y_i\}$ are independent. The parameter $\bar{r}$ no longer equals the risk-free interest rate r. Instead, it should be chosen so that $E[e^{-rT} S_T] = S_0$, or equivalently,

$$\bar{r} = r - \lambda \left(\frac{p\eta_1}{\eta_1 - 1} + \frac{(1-p)\eta_2}{\eta_2 + 1} - 1 \right);$$

see [17]. The price of an option with payoff X and maturity T still takes the form

$$v = E[e^{-rT} X],$$

where the expected value is taken under the risk-neutral probability measure. Write a function to price the call option with strike price K and maturity T. The function should have input parameters r, σ, λ, p, η_1, η_2, S_0, K, T, and the sample size n. Report your estimate and standard error for

$$r = 0.05,\ \sigma = 0.2,\ \lambda = 3,\ p = 0.4,\ \eta_1 = \eta_2 = 20,$$

$$S_0 = 50,\ K = 50,\ T = 0.5,\ n = 10000.$$

Chapter 6

Variance Reduction Techniques

The efficiency of a Monte Carlo estimate is often characterized by its variance. The smaller the variance, the more efficient the estimate. In this chapter, we will discuss some commonly used variance reduction techniques in Monte Carlo simulation, including antithetic sampling, the control variate method, and stratified sampling. These methods can also be combined to further improve efficiency. It should be noted that there is no "panacea" in this business of variance reduction. The method of choice is very much problem dependent.

6.1 Antithetic Sampling

In plain Monte Carlo simulation, samples are independent and identically distributed. The idea of antithetic sampling is to reduce the variance by introducing samples that are *negatively* correlated. To be more precise, consider the following two schemes for estimating $E[X]$. In both scenarios, the estimate is the sample average.

1. *Plain Monte Carlo:* $2n$ iid samples $X_1, \ldots, X_n, X_{n+1}, \ldots X_{2n}$. The estimate is
$$\frac{1}{2n}\sum_{i=1}^{2n} X_i.$$

2. *Antithetic Sampling:* $2n$ samples (or n pairs of samples)
$$\begin{array}{cccc} X_1 & X_2 & \cdots & X_n \\ Y_1 & Y_2 & \cdots & Y_n \end{array}.$$

Pairs of samples (X_i, Y_i) are iid; Y_i has the same distribution as X_i; X_i and Y_i are *dependent*. The estimate is

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i + Y_i}{2}.$$

Let $\sigma^2 = \mathrm{Var}[X_i]$. The plain Monte Carlo estimate is unbiased, and its variance is

$$\mathrm{Var}\left(\frac{1}{2n}\sum_{i=1}^{2n} X_i\right) = \frac{1}{2n}\sigma^2.$$

Since Y_i has the same distribution as X_i, the antithetic sampling estimate $\hat{v}$ is again unbiased and

$$\mathrm{Var}[\hat{v}] = \frac{1}{4n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i + Y_i).$$

Suppose that the correlation coefficient between X_i and Y_i is β. It follows that

$$\mathrm{Var}[X_i + Y_i] = \mathrm{Var}[X_i] + \mathrm{Var}[Y_i] + 2\mathrm{Cov}(X_i, Y_i) = 2\sigma^2 + 2\beta\sigma^2,$$

and

$$\mathrm{Var}[\hat{v}] = \frac{1}{2n}\sigma^2 + \frac{\beta}{2n}\sigma^2.$$

Therefore, antithetic sampling achieves variance reduction when X_i and Y_i are *negatively* correlated, i.e., when $\beta < 0$. The improvement is characterized by the magnitude of β — the stronger the negative correlation, the more significant the variance reduction. Note that if the samples X_i and Y_i are made to be positively correlated, then antithetic sampling will actually increase the variance and make the estimate less accurate!

Remark 6.1. The standard deviation associated with the antithetic sampling estimate $\hat{v}$ is

$$\sqrt{\frac{1}{n}\mathrm{Var}\left[\frac{X_i + Y_i}{2}\right]}.$$

Replacing the variance by sample variance, we obtain the standard error

$$\text{S.E.} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\left[\frac{X_i + Y_i}{2}\right]^2 - n\hat{v}^2\right)}.$$

6.1.1 Generating Antithetic Samples

Now the question is: given a sample X_i, how should the antithetic sample Y_i be defined and generated? To ease exposition, we will omit the index i and denote the samples by X and Y, respectively.

In theory, it is always possible to construct a negatively correlated antithetic sample. Indeed, by Theorem 5.1 one can write $X = F^{-1}(U)$, where U is a random variable uniformly distributed on $[0, 1]$ and F^{-1} is the inverse of the cumulative distribution function of X. An antithetic sample of X can be defined as

$$Y = F^{-1}(1 - U).$$

Clearly Y has the same distribution as X since $1 - U$ is also uniformly distributed on $[0, 1]$. Furthermore, X and Y are negatively correlated because X is an increasing function of U and Y is a decreasing function of U; see Exercise 6.4.

In practice, the inverse function F^{-1} is usually not available. However, the previous discussion does suggest that if $X = h(U)$ for some function h, then an antithetic sample can be defined as

$$Y = h(1 - U).$$

If h is monotone, then by the same token X and Y are negatively correlated. However, when h fails to be monotone, one has to exercise caution because the correlation between X and Y could be positive. We can also extend this discussion to the more general cases.

a. $\boldsymbol{X = h(Z)}$ **where** $\boldsymbol{Z}$ **is** $\boldsymbol{N(0,1)}$**.** One can write $Z = \Phi^{-1}(U)$, and thus the antithetic sample is

$$Y = h(\Phi^{-1}(1 - U)) = h(-\Phi^{-1}(U)) = h(-Z).$$

b. $\boldsymbol{X = h(U_1, \ldots, U_k)}$ **where** $\boldsymbol{\{U_1, \ldots, U_k\}}$ **are iid uniform on** $[0, 1]$**.** The antithetic sample is

$$Y = h(1 - U_1, \ldots, 1 - U_k).$$

c. $\boldsymbol{X = h(Z_1, \ldots, Z_k)}$ **where** $\boldsymbol{\{Z_1, \ldots, Z_k\}}$ **are iid** $\boldsymbol{N(0,1)}$**.** The antithetic sample is

$$Y = h(-Z_1, \ldots, -Z_k).$$

Remark 6.2. Antithetic sampling does not exploit much knowledge of the underlying stochastic models, and thus its effectiveness is limited. It is often used as part of a larger scheme to achieve greater variance reduction.

6.1.2 Examples of Antithetic Sampling

For all the examples in this section, we assume that the underlying stock price S is a geometric Brownian motion under the risk-neutral probability measure. That is,

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\},$$

where r is the risk-free interest rate.

Example 6.1. Use antithetic sampling to estimate the price of a call option with maturity T and strike price K. Compare with the plain Monte Carlo estimate.

SOLUTION: The call option payoff is an increasing function of S_T, and hence an increasing function of W_T. We expect that antithetic sampling will reduce the variance. The pseudocode for the plain Monte Carlo scheme is given in Example 4.1.

Pseudocode for antithetic sampling:

for $i = 1, 2, \ldots, n$

generate a sample Z from $N(0,1)$

set $S_i = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right\}$

set $X_i = e^{-rT}(S_i - K)^+$

set $S_i = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T - \sigma\sqrt{T}Z\right\}$

set $Y_i = e^{-rT}(S_i - K)^+$

compute the estimate $\hat{v} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{X_i + Y_i}{2}$

compute the standard error

$$\text{S.E.} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\left[\frac{X_i + Y_i}{2}\right]^2 - n\hat{v}^2\right)}.$$

The simulation results are given in Table 6.1 for call options with different strike prices. The parameters are

$$r = 0.05,\ \ \sigma = 0.2,\ \ T = 1,\ \ S_0 = 50,\ \ n = 10000.$$

The true values are obtained from the Black–Scholes formula. As expected, the antithetic sampling method does reduce the variance, but only to some extent.

Table 6.1: Call option: antithetic sampling versus plain Monte Carlo

Strike price	$K = 40$		$K = 50$		$K = 60$	
	antithetic	plain	antithetic	plain	antithetic	plain
True value	12.2944		5.2253		1.6237	
Estimate	12.2638	12.3376	5.2679	5.2142	1.6527	1.6251
S.E.	0.0231	0.0680	0.0372	0.0521	0.0287	0.0304

Note that for a give n, we use n pairs of samples (X_i, Y_i) for the antithetic sampling scheme and use $2n$ samples for the plain Monte Carlo scheme. ■

Example 6.2. Consider a discretely monitored down-and-out barrier option with maturity T and payoff

$$(S_T - K)^+ \cdot 1_{\{\min(S_{t_1}, \cdots, S_{t_m}) \geq b\}}.$$

The monitoring dates $0 < t_1 < t_2 < \cdots < t_m = T$ are prefixed. Compare the antithetic sampling estimate with the plain Monte Carlo estimate.

SOLUTION: This is a path-dependent option and its payoff is monotonically increasing with respect to the stock price. Therefore, we expect that antithetic sampling will reduce the variance. Note that the sample path $(S_{t_1}, \ldots, S_{t_m})$ is generated sequentially by

$$S_{t_{i+1}} = S_{t_i} \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)(t_{i+1} - t_i) + \sigma\sqrt{t_{i+1} - t_i}Z_{i+1}\right\},$$

where Z_i's are iid standard normal random variables. The antithetic sample path $(\bar{S}_{t_1}, \ldots, \bar{S}_{t_m})$ is given by

$$\bar{S}_{t_{i+1}} = \bar{S}_{t_i} \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)(t_{i+1} - t_i) - \sigma\sqrt{t_{i+1} - t_i}Z_{i+1}\right\}$$

with the same initial price $\bar{S}_0 = S_0$. The pseudocode for the plain Monte Carlo scheme is very straightforward and thus omitted.

Pseudocode for antithetic sampling:

for $i = 1, 2, \ldots, n$

  for $j = 1, 2, \ldots, m$

    generate a sample Z from $N(0,1)$

    set $S_j = S_{j-1} \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)(t_j - t_{j-1}) + \sigma\sqrt{t_j - t_{j-1}}Z\right\}$

    set $\bar{S}_j = \bar{S}_{j-1} \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)(t_j - t_{j-1}) - \sigma\sqrt{t_j - t_{j-1}}Z\right\}$

  set $X_i = e^{-rT}(S_m - K)^+ \cdot 1_{\{\min(S_1,\cdots,S_m)\geq b\}}$

  set $Y_i = e^{-rT}(\bar{S}_m - K)^+ \cdot 1_{\{\min(\bar{S}_1,\cdots,\bar{S}_m)\geq b\}}$

compute the estimate $\hat{v} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{X_i + Y_i}{2}$

compute the standard error

$$\text{S.E.} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\left[\frac{X_i + Y_i}{2}\right]^2 - n\hat{v}^2\right)}.$$

The simulation results are reported in Table 6.2. The parameters are given by

$$r = 0.05,\ \ \sigma = 0.2,\ \ T = 1,\ \ S_0 = 50,\ \ b = 45,$$
$$m = 12,\ \ t_i = iT/m,\ \ n = 10000.$$

Table 6.2: Barrier option: antithetic sampling versus plain Monte Carlo

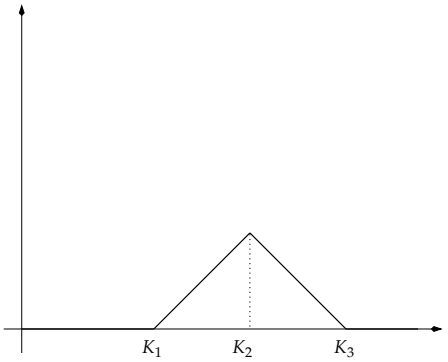
Strike price	$K = 40$		$K = 50$		$K = 60$	
	antithetic	plain	antithetic	plain	antithetic	plain
Estimate	9.8839	9.9773	4.7669	4.8529	1.5651	1.5672
S.E.	0.0423	0.0767	0.0399	0.0528	0.0283	0.0308

As expected, antithetic sampling reduces the variance to some extent. ■

Example 6.3. Use antithetic sampling to estimate the price of a butterfly spread option with maturity T and payoff

$$(S_T - K_1)^+ + (S_T - K_3)^+ - 2(S_T - K_2)^+,$$

where $0 < K_1 < K_3$ and $K_2 = (K_1 + K_3)/2$. Compare with the plain Monte Carlo estimate.



Butterfly spread option

SOLUTION: We omit the pseudocode since it is exactly the same as that of Example 6.1, except that the call option payoff should be replaced by the butterfly spread option payoff. The numerical comparison is presented in Table 6.3, where the parameters are

$$r = 0.05,\ \sigma = 0.2,\ S_0 = 50,\ K_1 = 45,\ K_2 = 50,\ K_3 = 55,\ n = 10000.$$

The true values are calculated from the Black–Scholes formula.

Table 6.3: Butterfly option: antithetic sampling versus plain Monte Carlo

Maturity	$T = 1$		$T = 0.5$		$T = 0.25$	
	antithetic	plain	antithetic	plain	antithetic	plain
True value	0.9192		1.3183		1.8156	
Estimate	0.9204	0.9180	1.3057	1.3373	1.8411	1.8311
S.E.	0.0123	0.0102	0.0152	0.0115	0.0166	0.0120

This example shows that when the payoff function is not monotone, the antithetic sampling method may increase the variance. ■

6.2 Control Variates

Suppose that we are interested in estimating the expected value $E[X]$. The difference between the plain Monte Carlo scheme and the control variates method is as follows.

1. *Plain Monte Carlo:* n iid samples $X_1, \ldots, X_n$. The estimate is

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. *Control Variates:* n samples $\{X_i\}$ and n control variate samples $\{Y_i\}$

$$\begin{array}{cccc} X_1 & X_2 & \cdots & X_n \\ Y_1 & Y_2 & \cdots & Y_n \end{array}.$$

Pairs of samples (X_i, Y_i) are iid; Y_i has *known* expected value $\bar{\mu}$. The estimate is

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n} X_i - b\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{\mu}\right),$$

where b is a fixed constant.

Clearly, both estimates are unbiased. In general, the scheme of control variates is set up in such a way that sampling the pair (X_i, Y_i) requires little extra computational effort compared with sampling X_i alone.

Denote by σ_X^2 the variance of X_i, σ_Y^2 the variance of Y_i, and β the correlation coefficient between X_i and Y_i. The variance of the plain Monte Carlo estimate is

$$\frac{1}{n}\sigma_X^2.$$

As for the control variate estimate $\hat{v}$, write

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n} H_i, \quad H_i = X_i - b(Y_i - \bar{\mu}).$$

Since $H_1, \ldots, H_n$ are iid random variables, it follows that the variance of the control variate estimate equals

$$\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[H_i] = \frac{1}{n}\left[\sigma_X^2 - 2b\beta\sigma_X\sigma_Y + b^2\sigma_Y^2\right].$$

The size of variance reduction depends on the coefficient b. The *optimal* choice of b is the one that minimizes the variance of the control variate estimate:

$$b^* = \beta\frac{\sigma_X}{\sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}[Y]}. \tag{6.1}$$

If one uses b^*, the variance of the control variate estimate becomes

$$\frac{1}{n}(1-\beta^2)\sigma_X^2.$$

In other words, it reduces the variance of the plain Monte Carlo estimate by a factor of β^2.

There are usually many possible ways to select the control variate Y. Ideally, it should have a strong correlation with X, be it positive or negative. For this reason, the control variate is often chosen to have a structure similar to X.

Remark 6.3. The standard deviation associated with the control variate estimate $\hat{v}$ is

$$\sqrt{\frac{1}{n}\mathrm{Var}[H_i]}.$$

Replacing the variance by the sample variance, we obtain the standard error of $\hat{v}$:

$$\text{S.E.} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}.$$

6.2.1 The Optimal Coefficient b^*

The optimal coefficient b^*, given by (6.1), is an unknown quantity in general. A common approach is to estimate b^* by the sample variance and covariance, using the same samples (X_i, Y_i). That is,

$$\hat{b}^* = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{6.2}$$

Since the sample size n is large, the estimate $\hat{b}^*$ will be close to the true b^*. However, it should be pointed out that by using $\hat{b}^*$, we introduce bias. The reason is that in general

$$E\left[\hat{b}^*\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{\mu}\right)\right] \neq 0,$$

since $\hat{b}^*$ is no longer independent of $\{Y_i\}$. In practice, this bias is usually very small compared to the standard error of the estimate, and can be safely ignored.

If one wishes to eliminate this bias, one can estimate b^* from samples other than (X_i, Y_i) [such samples are sometimes called the *pilot samples*]. By using pilot samples, the estimate of b^* will be independent of $\{Y_i\}$, and hence the unbiasedness of the control variate estimate will be preserved. Typically, the size of pilot samples is much smaller than n so that it does not incur too much extra computational effort.

Yet another way to eliminate the bias is to forgo optimality and prefix b in a sensible way. For example, if X and Y are positively correlated and have *very* similar structures, one can let $b = 1$ or choose a b from prior experience; see Exercise 6.7. This naive approach can be quite effective sometimes.

6.2.2 Examples of Control Variates

For all the examples in this section, the underlying stock price S is assumed to be a geometric Brownian motion under the risk-neutral probability measure. That is,

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\},$$

where r is the risk-free interest rate.

Example 6.4. Underlying stock price as control variate. Use the method of control variates to estimate the price of a call option with maturity T and strike price K. Since

$$E[e^{-rT}S_T] = S_0,$$

the discounted stock price $e^{-rT}S_T$ can serve as a control variate.

SOLUTION: The control variate estimate is the sample average of iid copies of

$$H = e^{-rT}(S_T - K)^+ - b(e^{-rT}S_T - S_0).$$

We will either use $b = 1$ or the estimate $\hat{b}^*$ from formula (6.2).

Pseudocode for control variate method:

for $i = 1, 2, \ldots, n$

generate Z from $N(0,1)$

set $S = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}Z\right\}$

set $X_i = e^{-rT}(S - K)^+$

set $Y_i = e^{-rT} S - S_0$

set $b = 1$ or $b = \hat{b}^*$

for $i = 1, 2, \ldots, n$

set $H_i = X_i - bY_i$

compute the estimate $\hat{v} = \dfrac{1}{n} \sum_{i=1}^{n} H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)} \left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2 \right)}$.

The numerical results are shown in Table 6.4, where the parameters are given by

$$S_0 = 50, \ \ r = 0.05, \ \ \sigma = 0.2, \ \ T = 1, \ \ n = 10000.$$

Table 6.4: Call option: control variates versus plain Monte Carlo

Strike price	$K = 45$			$K = 55$		
	$b = 1$	$b = \hat{b}^*$	Plain MC	$b = 1$	$b = \hat{b}^*$	Plain MC
True value	8.3497			3.0200		
Estimate	8.3036	8.3104	8.3459	3.0572	3.0279	3.0001
S.E.	0.0262	0.0206	0.0860	0.0604	0.0308	0.0581

As we can see, the estimate with $b = \hat{b}^*$ is the most accurate one. Furthermore, the performance of the control variate estimates is better for the smaller strike price. The reason is that as K becomes smaller, the correlation between the payoff $(S_T - K)^+$ and the stock price S_T becomes stronger. ■

Example 6.5. Analytically tractable derivatives as control variate. Estimate the price of a discretely monitored average price call option with maturity T and payoff $(\bar{S} - K)^+$. Here $\bar{S}$ is the arithmetic mean of stock prices:

$$\bar{S} = \frac{1}{m} \sum_{k=1}^{m} S_{t_k},$$

where $0 < t_1 < \cdots < t_m = T$ are given dates. Use the average price call option with geometric mean as the control variate.

SOLUTION: The price of an average price call option with geometric mean can be explicitly evaluated; see Example 2.5. Denote the price by p. Then the control variate estimate for the price of an average price call option with arithmetic mean is the sample average of iid copies of

$$e^{-rT}\left(\bar{S}-K\right)^{+}-b\left[e^{-rT}(\bar{S}_G-K)^{+}-p\right],$$

where

$$\bar{S}_G=\left(\prod_{k=1}^{m}S_{t_k}\right)^{1/m}.$$

Again, we let $b=1$ or $b=\hat{b}^*$ from formula (6.2).

Pseudocode for control variate method:

for $i=1,2,\ldots,n$
  for $k=1,2,\ldots,m$
    generate Z from $N(0,1)$
    set $S_k=S_{k-1}\exp\left\{\left(r-\frac{1}{2}\sigma^2\right)(t_k-t_{k-1})+\sigma\sqrt{t_k-t_{k-1}}Z\right\}$
  compute the arithmetic mean $\bar{S}$ of $S_1,\ldots,S_m$
  compute the geometric mean $\bar{S}_G$ of $S_1,\ldots,S_m$
  set $X_i=e^{-rT}(\bar{S}-K)^{+}$
  set $Y_i=e^{-rT}(\bar{S}_G-K)^{+}-p$
set $b=1$ or $b=\hat{b}^*$
for $i=1,2,\ldots,n$
  set $H_i=X_i-bY_i$
compute the estimate $\hat{v}=\dfrac{1}{n}\sum_{i=1}^{n}H_i$
compute the standard error S.E. $=\sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^{n}H_i^2-n\hat{v}^2\right)}$.

We report the simulation results in Table 6.5. The parameters are given by

$$r=0.05,\ \ T=1,\ \ \sigma=0.2,\ \ m=12,\ \ t_i=iT/m,\ \ S_0=50,$$

and the sample size is $n=10000$.

Table 6.5: Asian option: control variates versus plain Monte Carlo

Strike price	$K = 45$			$K = 55$		
	$b = 1$	$b = \hat{b}^*$	plain	$b = 1$	$b = \hat{b}^*$	plain
Estimate	6.4590	6.4607	6.3755	1.1477	1.1456	1.1863
S.E.	0.0017	0.0013	0.0554	0.0018	0.0011	0.0288

The variance reduction achieved by the control variate method is significant even with the naive choice of $b = 1$. It reflects the strong correlation between the arithmetic mean $\bar{S}$ and the geometric mean $\bar{S}_G$. ■

6.3 Stratified Sampling

Consider the problem of estimating $\mu = E[X]$. Suppose that there exists a random variable Y such that

1. Y takes finitely many possible values $\{y_1, \ldots, y_k\}$ and the probability mass function
$$p_i = \mathbb{P}(Y = y_i), \quad i = 1, \ldots, k$$
is known.

2. For each $i = 1, \ldots, k$, one is able to sample X *conditional* on $Y = y_i$.

The random variable Y partitions the sample space into k subsets according to its value. Each subset is called a *stratum*, and Y is said to be a *stratification variable*.

It follows from the tower property of conditional expectations (Theorem 1.12) that

$$\mu = E[X] = E[E[X|Y]] = \sum_{i=1}^{k} p_i E[X|Y = y_i]. \tag{6.3}$$

Since p_i is known, we only need to estimate $E[X|Y = y_i]$ for each i. This is made possible by the assumption that one can sample X conditional on Y. More precisely, a stratified sampling scheme divides the total sample size n into

$$n = n_1 + \cdots + n_k$$

and draws n_i samples $\{X_{ij} : j = 1, \ldots, n_i\}$ from the stratum determined by $Y = y_i$, that is, n_i samples of X conditional on $Y = y_i$. The stratified sampling estimate is defined to be

$$\hat{\mu} = \sum_{i=1}^{k} p_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}. \tag{6.4}$$

Thanks to (6.3) and that $E[X_{ij}] = E[X|Y = y_i]$, the estimate $\hat{\mu}$ is clearly unbiased. Since X_{ij}'s are all independent, it follows that

$$\mathrm{Var}[\hat{\mu}] = \sum_{i=1}^{k} p_i^2 \cdot \frac{1}{n_i^2} \sum_{j=1}^{n_i} \mathrm{Var}[X_{ij}] = \sum_{i=1}^{k} p_i^2 \cdot \frac{1}{n_i} \sigma_i^2,$$

where

$$\sigma_i^2 = \mathrm{Var}[X|Y = y_i].$$

Comments on stratified sampling.

- There are no fixed rules for choosing the stratification variable Y. In general, Y is selected in such a way that it is strongly correlated with X and the sampling of X conditional on $Y = y_i$ does not require too much additional computational effort.
- The sample size allocation $\{n_1, \ldots, n_k\}$ is assigned in advance. There are two commonly used approaches to select n_i, both of which lead to variance reduction. See Section 6.3.1.
- In general, the value of σ_i^2 is unknown. However, it can be estimated by the sample variance of $\{X_{ij} : j = 1, \ldots, n_i\}$:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2, \quad \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

  Consequently, the standard error of $\hat{\mu}$ is just

$$\text{S.E.} = \sqrt{\sum_{i=1}^{k} p_i^2 \cdot \frac{1}{n_i} s_i^2}\,.$$

6.3.1 Allocating Samples in Each Stratum

The sample size in each stratum is often selected as a fixed fraction of the total sample size n. That is, $n_i = nq_i$, where

$$q_i > 0, \quad q_1 + \cdots + q_k = 1. \tag{6.5}$$

A very simple strategy is to let $q_i = p_i$, which is said to be the *proportional allocation.*

Another frequently used strategy aims to minimize the variance of $\hat{\mu}$. That is, one chooses the allocation $\{q_1, \ldots, q_k\}$ that minimizes

$$\mathrm{Var}[\hat{\mu}] = \sum_{i=1}^{k} p_i^2 \cdot \frac{1}{n_i}\sigma_i^2 = \frac{1}{n}\sum_{i=1}^{k} p_i^2 \cdot \frac{1}{q_i}\sigma_i^2$$

under the constraints of (6.5). This optimization problem can be easily solved (see Remark 6.4), and the solution is the *optimal allocation*

$$q_i^* = \frac{p_i\sigma_i}{\sum_{m=1}^{k} p_m\sigma_m}. \tag{6.6}$$

In practice, since σ_i's are unknown, pilot samples are often used to obtain estimates for $\{\sigma_i\}$ and $\{q_i^*\}$.

Remark 6.4. The optimal allocation can be solved by the Cauchy–Schwartz inequality [14]: for any real numbers a_i and b_i

$$\sum_{i=1}^{k} a_i^2 \cdot \sum_{i=1}^{k} b_i^2 \geq \left(\sum_{i=1}^{k} a_i b_i\right)^2$$

with equality if and only if (assuming $b_i \neq 0$)

$$\frac{a_1}{b_1} = \cdots = \frac{a_k}{b_k}.$$

It follows that

$$\sum_{i=1}^{k} p_i^2 \frac{1}{q_i}\sigma_i^2 = \sum_{i=1}^{k} q_i \cdot \sum_{i=1}^{k} p_i^2 \frac{1}{q_i}\sigma_i^2 \geq \left(\sum_{i=1}^{k} p_i\sigma_i\right)^2$$

with equality if and only if

$$\frac{q_1}{p_1\sigma_1} = \cdots = \frac{q_k}{p_k\sigma_k}.$$

Solving the equations in the last display along with the constraints in (6.5) leads to the optimal allocation (6.6).

6.3.2 Variance Decomposition

Compared with the plain Monte Carlo scheme, stratified sampling always achieves variance reduction with either the proportional allocation or the optimal allocation. We only need to verify this claim for the proportional allocation since by definition, the variance associated with the optimal allocation cannot exceed that of the proportional allocation.

Note that for the proportional allocation, $q_i = p_i$ and the variance of the corresponding estimate is

$$\sum_{i=1}^{k} p_i^2 \cdot \frac{1}{n_i}\sigma_i^2 = \frac{1}{n}\sum_{i=1}^{k} p_i \sigma_i^2.$$

Since the plain Monte Carlo estimate from n samples has variance $\mathrm{Var}[X]/n$, it suffices to show that

$$\mathrm{Var}[X] \geq \sum_{i=1}^{k} p_i \sigma_i^2. \tag{6.7}$$

To this end, let $\mu_i = E[X|Y = y_i]$ and observe that

$$E[X^2] = E[E[X^2|Y]] = \sum_{i=1}^{k} p_i E[X^2|Y = y_i] = \sum_{i=1}^{k} p_i(\sigma_i^2 + \mu_i^2),$$

$$E[X] = E[E[X|Y]] = \sum_{i=1}^{k} p_i E[X|Y = y_i] = \sum_{i=1}^{k} p_i \mu_i,$$

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i \mu_i^2 - \left(\sum_{i=1}^{k} p_i \mu_i\right)^2.$$

The extra term

$$\sum_{i=1}^{k} p_i \mu_i^2 - \left(\sum_{i=1}^{k} p_i \mu_i\right)^2 \tag{6.8}$$

is nonnegative since it equals the variance of a discrete random variable that takes value μ_i with probability p_i. The desired inequality (6.7) follows readily.

Remark 6.5. The variance reduction achieved by the proportional allocation can also be understood from the variance decomposition formula

$$\mathrm{Var}[X] = E[\mathrm{Var}[X|Y]] + \mathrm{Var}(E[X|Y]),$$

which holds for arbitrary random variables X and Y (see Exercise 6.12). The variance from proportional allocation is indeed

$$\sum_{i=1}^{k} p_i \sigma_i^2 = E[\mathrm{Var}[X|Y]],$$

while the extra term (6.8) equals $\mathrm{Var}(E[X|Y])$. From this point of view, one can say that the proportional allocation eliminates the variance *between* strata, but not *within* strata.

6.3.3 Basic Strategies in Stratified Sampling

We should illustrate how one chooses the stratification variable Y and generates samples from the conditional distributions. The specific values that Y takes (that is, the y_i's) are not important. They are merely the indicators of strata. The essence of a stratification variable is a partition of the sample space.

The most basic algorithm in stratified sampling is the stratification of the uniform distribution on $[0, 1]$. To be more concrete, suppose that one is interested in estimating $\mu = E[h(U)]$, where U is uniformly distributed on $[0, 1]$ and h is an arbitrary function.

a. Stratification of Uniform Distribution on $[0, 1]$. A common strategy is to partition the unit interval into k strata of equal length:

$$I_1 = \left[0, \frac{1}{k}\right), \ \cdots, \ I_k = \left[\frac{k-1}{k}, 1\right].$$

Define Y so that $Y = i$ if and only if $U \in I_i$. It follows that for each i

$$p_i = \mathbb{P}(Y = i) = \mathbb{P}(U \in I_i) = \frac{1}{k}.$$

b. Stratified Sampling Estimate. Let n_i be the sample size for the stratum $\{Y = i\} = \{U \in I_i\}$. Denote by $\{U_{ij} : j = 1, \ldots, n_i\}$ the iid samples from the conditional distribution of U given $U \in I_i$. The estimate is

$$\hat{\mu} = \sum_{i=1}^{k} \frac{1}{k} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} h(U_{ij}).$$

c. Sample from Conditional Distributions. It is trivial that conditional on $U \in I_i$, U is uniformly distributed on the interval I_i. Sampling from this uniform conditional distribution is straightforward. Indeed, let V_{ij} be iid samples from the uniform distribution on $[0,1]$. Then

$$U_{ij} = \frac{i-1}{k} + \frac{1}{k} V_{ij}$$

are iid samples from the uniform distribution on I_i.

The stratified sampling scheme for estimating μ is now ready. Below is the pseudocode.

Pseudocode for stratified sampling:

specify the sample size n_i in stratum I_i for each i
for $i = 1, 2, \ldots, k$
    generate iid samples $\{V_{ij} : j = 1, \ldots, n_i\}$ uniformly from $[0,1]$
    set $U_{ij} = (i-1)/k + V_{ij}/k$ for $j = 1, \ldots, n_i$
    compute the sample mean and standard deviation in stratum I_i

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} h(U_{ij}), \quad s_i = \sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} [h(U_{ij}) - \hat{\mu}_i]^2}$$

compute the stratified sampling estimate $\hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \hat{\mu}_i$

compute the standard error S.E. $= \frac{1}{k} \sqrt{\sum_{i=1}^{k} \frac{1}{n_i} s_i^2}$.

Now consider the more general problem of estimating $E[h(X)]$. Denote by F the cumulative distribution function of X. By the inverse transform method, one can write $E[h(X)] = E[\bar{h}(U)]$, where $\bar{h} = h \circ F^{-1}$ and U is a random variable uniformly distributed on $[0,1]$. Thus we have reverted to the previous setting. For this strategy to work, it is necessary that F^{-1} can be evaluated in a very efficient manner. For many financial engineering problems, this is not difficult because X can be chosen as a standard normal random variable and the inverse of its cumulative distribution function can be evaluated by the MATLAB® built-in function "norminv".

Example 6.6. Assume that under the risk-neutral probability measure the stock price is a geometric Brownian motion

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\}.$$

Design a stratified sampling scheme to estimate the price of a call option with maturity T and strike price K.

SOLUTION: The payoff is a function of S_T and thus a function of W_T. Write

$$W_T = \sqrt{T}\Phi^{-1}(U)$$

for some random variable U that is uniformly distributed on $[0,1]$. Then the price of the call option is $v = E[h(U)]$, where

$$h(u) = \left(S_0 \exp\left\{-\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}\Phi^{-1}(u)\right\} - e^{-rT}K\right)^+.$$

The pseudocode for stratified sampling is exactly the same as the one given in the preceding discussion. In the numerical simulation, we use proportional allocation and let

$$S_0 = 50,\ \ r = 0.05,\ \ \sigma = 0.2,\ \ T = 1,\ \ n = 10000.$$

The results are reported in Table 6.6.

Table 6.6: Stratified sampling for call option

Strike price	$K = 40$		$K = 50$		$K = 60$	
# of strata	$k = 25$	$k = 100$	$k = 25$	$k = 100$	$k = 25$	$k = 100$
True value	12.2944		5.2253		1.6237	
Estimate	12.2972	12.2979	5.2340	5.2358	1.6334	1.6279
S.E.	0.0127	0.0061	0.0124	0.0072	0.0136	0.0057

Compared with the results from the plain Monte Carlo scheme in Example 4.1, stratified sampling reduces the variance significantly. ■

Example 6.7. Use stratified sampling to estimate the price of a spread call option with maturity T and payoff

$$(X_T - Y_T - K)^+,$$

where X and Y are the prices of two underlying assets. Assume that under the risk-neutral probability measure,

$$\begin{aligned} X_t &= X_0 \exp\left\{\left(r - \frac{1}{2}\sigma_1^2\right)t + \sigma_1 W_t\right\}, \\ Y_t &= Y_0 \exp\left\{\left(r - \frac{1}{2}\sigma_2^2\right)t + \sigma_2 B_t\right\}, \end{aligned}$$

where (W, B) is a two-dimensional Brownian motion with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

SOLUTION: The payoff is a function of (X_T, Y_T), which is in turn a function of (θ, η), where

$$\theta = \frac{W_T}{\sqrt{T}}, \quad \eta = \frac{B_T}{\sqrt{T}}.$$

By assumption, (θ, η) is a jointly normal random vector with mean 0 and covariance matrix Σ. There are many ways to define a stratification variable. We will discuss two of them: (1) stratification of the random variable θ alone; (2) simultaneous stratification of the random vector (θ, η). In both cases we use proportional allocation.

Stratification of θ alone: In this approach, we first perform stratified sampling on θ. Since θ is a standard normal random variable, this can be done as in Example 6.6. The second step is to sample η conditional on the value of θ. This is not difficult because given $\theta = x$, η is normally distributed with mean ρx and variance $1 - \rho^2$; see Appendix A.

Pseudocode for stratifying θ alone:

specify the sample size n_i in stratum I_i for each i
for $i = 1, 2, \ldots, k$
    for $j = 1, 2, \ldots, n_i$
        generate a sample V_{ij} from the uniform distribution on $[0, 1]$
        set $U_{ij} = (i-1)/k + V_{ij}/k$
        set $\theta_{ij} = \Phi^{-1}(U_{ij})$
        generate a sample η_{ij} from $N(\rho\theta_{ij}, 1 - \rho^2)$
        set H_{ij} as the discounted option payoff given (θ_{ij}, η_{ij})

compute the sample mean and standard deviation in stratum I_i

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} H_{ij}, \quad s_i = \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(H_{ij} - \hat{\mu}_i)^2}$$

compute the stratified sampling estimate $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} \hat{\mu}_i$

compute the standard error S.E. $= \frac{1}{k}\sqrt{\sum_{i=1}^{k} \frac{1}{n_i} s_i^2}$.

The simulation results are reported in Table 6.7 with sample size $n = 10000$ for

$$X_0 = 50, \ Y_0 = 45, \ r = 0.05, \ \sigma_1 = 0.2, \ \sigma_2 = 0.3, \ \rho = 0.5, \ T = 1.$$

Table 6.7: Stratified sampling for spread call option

Strike price	$K = 0$		$K = 5$		$K = 10$	
# of strata	$k = 25$	$k = 100$	$k = 25$	$k = 100$	$k = 25$	$k = 100$
Estimate	7.9277	7.8958	4.8622	4.9584	2.7980	2.7988
S.E.	0.0782	0.0781	0.0630	0.0623	0.0484	0.0476

Compared with Example 4.3, the improvement of stratified sampling over plain Monte Carlo is almost negligible. The reason is that the stratification will only eliminate the variance *between strata*. In this case, since θ and η are positively correlated, so are X_T and Y_T. Therefore, the difference $X_T - Y_T$, or more precisely, the average of the difference $X_T - Y_T$ in each stratum, has little variation across the strata. Thus the variance reduction through stratification is inconsequential. One can also stratify η instead θ. The simulation result will only be slightly better.

Simultaneous stratification of (θ, η): Now let us consider another approach where we stratify θ and η simultaneously. To do this, recall from Example 4.3 that

$$CC' = \Sigma, \quad C = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$$

is a Cholesky factorization of Σ. Suppose that U and V are independent and uniformly distributed on $[0, 1]$. Then

$$Z = \begin{bmatrix} \Phi^{-1}(U) \\ \Phi^{-1}(V) \end{bmatrix}$$

is a two-dimensional standard normal random vector and CZ is jointly normal with mean 0 and covariance matrix $CC' = \Sigma$. The scheme stratifies both U and V simultaneously by partitioning the interval $[0,1]$ into k subintervals of equal length for U and for V. That is, there will be $k \times k = k^2$ strata $\{I_{ij}\}$ with

$$I_{ij} = \left[\frac{i-1}{k}, \frac{i}{k}\right) \times \left[\frac{j-1}{k}, \frac{j}{k}\right), \quad i, j = 1, \ldots, k.$$

To generate a sample from stratum I_{ij}, we generate u and v uniformly from $[0,1]$ and let

$$U = (i-1)/k + u/k, \;\; V = (j-1)/k + v/k.$$

Then we let

$$\begin{bmatrix} \theta \\ \eta \end{bmatrix} = C \begin{bmatrix} \Phi^{-1}(U) \\ \Phi^{-1}(V) \end{bmatrix}.$$

The pseudocode is straightforward and thus omitted. In order to match the number of strata in the previous simulation, we let $k = 5$ and $k = 10$ so that the total number of strata is $5^2 = 25$ and $10^2 = 100$, respectively. The simulation results are given in Table 6.8. We can see that this stratification strategy outperforms the previous one and the plain Monte Carlo scheme.

Table 6.8: Stratified sampling for spread call option

Strike price	$K = 0$		$K = 5$		$K = 10$	
# of strata k^2	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Estimate	7.8979	7.8805	4.9766	4.9542	2.7966	2.8041
S.E.	0.0241	0.0139	0.0228	0.0137	0.0208	0.0129

It is worth noting that determining a good stratification variable is not always straightforward. Only in very specialized settings can one explicitly solve for the optimal stratification strategy. For more details, see [12]. ■

Exercises

Pen-and-Paper Problems

6.1 Determine the antithetic variable and express it in terms of the original random variable X.

(a) X is normally distributed with mean μ and variance σ^2.

(b) X is lognormally distributed with parameters (μ, σ^2).

(c) X is uniformly distributed on the interval $[a, b]$.

(d) X is exponentially distributed with rate λ.

(e) X has a symmetric density function. That is, the density function f satisfies $f(x) = f(-x)$ for all x.

6.2 Suppose that X is a Bernoulli random variable with parameter p. Show that if $2p > 1$, then the antithetic variable of X is

$$Y = 1 - XZ,$$

where Z is a Bernoulli random variable with parameter $(1-p)/p$, independent of X. What is the antithetic variable when $2p < 1$ or $2p = 1$?

6.3 Suppose that X and Y have the same distribution (X and Y could be dependent). Let $0 \leq \alpha \leq 1$ and define

$$Z = \alpha X + (1-\alpha)Y.$$

Show that the variance of Z is minimized when $\alpha = 1/2$. This explains why in antithetic sampling $(X+Y)/2$ is used as the estimate instead of the more general $\alpha X + (1-\alpha)Y$.

6.4 Let X be an arbitrary random variable. Let f be a monotonically increasing function and g a monotonically decreasing function. Show that $f(X)$ and $g(X)$ are negatively correlated. That is,

$$\mathrm{Cov}[f(X), g(X)] \leq 0.$$

Hint: Let Y be an independent random variable that has the same distribution as X. Argue that $[f(X) - f(Y)] \cdot [g(X) - g(Y)] \leq 0$ and then take expected value on both sides.

6.5 Use antithetic sampling to estimate $P(Z \geq b)$, where Z is a standard normal random variable and b is a positive constant.

(a) Write down the antithetic sampling estimate.

(b) Compute the magnitude of variance reduction.

6.6 Construct an antithetic sampling algorithm to estimate $E[\exp\{U\}]$, where U is uniformly distributed on $[0,1]$. Compute the magnitude of variance reduction.

6.7 Show that the control variate method with $b = 1$ reduces the variance if and only if

$$2\beta\sigma_X > \sigma_Y.$$

6.8 When a random variable X has a known expected value μ, it can be used as the control variate in the estimation of $\mathbb{P}(X \le a)$ for some given constant a. The estimate is the sample average of iid copies of

$$1_{\{X \le a\}} - b(X - \mu),$$

for some constant b. Determine the optimal b and the magnitude of variance reduction when

(a) X is uniformly distributed on $[0,1]$ and $a \in (0,1)$;

(b) X is a standard normal random variable and $a \in \mathbb{R}$.

6.9 Suppose that one would like to estimate the price of a discretely monitored lookback put option with maturity T and payoff

$$X = \left(K - \min_{i=1,\dots,m} S_{t_i}\right)^+,$$

where $0 < t_1 < \cdots < t_m = T$ are given dates. The stock price is assumed to be a geometric Brownian motion under the risk-neutral probability measure. Which one of the following will you choose as the control variate? Explain your reasoning.

(a) $Y = e^{-rT} S_T$ with $\bar{\mu} = E[Y] = S_0$.

(b) $Y = e^{-rT}(K - S_T)^+$ with $\bar{\mu} = E[Y] =$ BLS_Put.

(c) $Y = e^{-rT}\left(K - \min_{0 \le t \le T} S_t\right)^+$, where $\bar{\mu} = E[Y]$ can be explicitly calculated; see Appendix C.

6.10 Let U be a random variable uniformly distributed on $[0,1]$. Consider the problem of estimating $E[U^2]$ by stratified sampling. Stratify U by partitioning the interval $[0,1]$ into $k = 2$ strata of equal length.

(a) Write down the estimate with the proportional allocation.

(b) Find the optimal allocation and write down the corresponding estimate.

Evaluate the magnitude of variance reduction from these two estimates.

6.11 Let X be exponentially distributed with rate λ. Design a stratified sampling scheme to estimate $\mu = E[h(X)]$. Write down the pseudocode.

6.12 Let X and Y be two arbitrary random variables. Define the *conditional variance* of X given Y by

$$\mathrm{Var}[X|Y] = E[X^2|Y] - (E[X|Y])^2.$$

Show that $\mathrm{Var}[X|Y] \geq 0$ and the variance decomposition formula

$$\mathrm{Var}[X] = E[\mathrm{Var}[X|Y]] + \mathrm{Var}[E[X|Y]]$$

holds.

6.13 The variance decomposition formula in Exercise 6.12 suggests the following variance reduction technique. Consider the problem of estimating $\mu = E[h(X)]$. Let Y be a random variable for which

$$f(Y) = E[h(X)|Y]$$

is explicitly known. Let $\{Y_1, \ldots, Y_n\}$ be iid copies of Y, and define the estimate to be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i).$$

Show that $\hat{\mu}$ is unbiased and its variance is always no greater than the variance of the plain Monte Carlo estimate with the same sample size. This variance reduction technique is called the **method of conditioning**. Discuss the difference between the method of conditioning and stratified sampling.

6.14 Let (X, Y) be a jointly normal random vector with mean 0 and covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Write down explicitly the estimate by the method of conditioning (conditional on Y) for

(a) $\mu = \mathbb{P}(XY \geq a)$ for some $a \in \mathbb{R}$;

(b) $\mu = \mathbb{P}(\min\{X, Y\} \geq a)$ for some $a \in \mathbb{R}$;

(c) $\mu = E[(e^X - e^Y)^+]$.

MATLAB® Problems

In Exercises 6.A – 6.D, assume that the underlying stock price is a geometric Brownian motion under the risk-neutral probability measure:

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\},$$

where W is a standard Brownian motion and r is the risk-free interest rate.

6.A Design simulation schemes to estimate the price of the butterfly option considered in Example 6.3.

(a) Control variate method: use the underlying stock price as the control variate and let $b = \hat{b}^*$, the sample estimate of b^*.

(b) Stratified sampling: stratify W_T into $k = 100$ strata and use the proportional allocation.

Report your simulation results with sample size $2n$. The parameters are the same as those given in Example 6.3.

6.B Consider a straddle with strike price K and maturity T. The option payoff is

$$X = (S_T - K)^+ + (K - S_T)^+.$$

Write a function to compare the following schemes:

(a) Plain Monte Carlo with sample size $2n$.

(b) Antithetic sampling with n pairs of samples.

(c) Control variate method with sample size $2n$. Use the underlying stock price as the control variate. Let $b = \hat{b}^*$.

(d) Stratified sampling with sample size $2n$. Stratify W_T into k strata and use the proportional allocation.

The function should have input parameters S_0, r, σ, K, T, k, and n. Report your estimates and standard errors for

$$S_0 = 50,\ r = 0.02,\ \sigma = 0.2,\ K = 50,\ T = 1,\ k = 100,\ n = 10000.$$

Why is the performance from the antithetic sampling and the control variate method poor?

6.C Consider a discretely monitored lookback call option with fixed strike price K and maturity T. The option payoff is

$$X = \left(\max_{i=1,\ldots,m} S_{t_i} - K \right)^+,$$

where $0 < t_1 < \cdots < t_m = T$ are given dates. Write a function to compare the performance of the following simulation schemes.

(a) Plain Monte Carlo with sample size $2n$.

(b) Antithetic sampling with n pairs of samples.

(c) Control variate method with sample size $2n$. Use $b = \hat{b}^*$, the sample estimate of b^*. The control variate is

i. the underlying stock price;

ii. the call option with strike price K and maturity T;

iii. the discretely monitored average price call option with geometric mean and maturity T, whose payoff is

$$Y = \left(\bar{S}_G - K\right)^+, \quad \bar{S}_G = \left(\prod_{i=1}^{m} S_{t_i}\right)^{1/m};$$

iv. the continuous time lookback call option with strike price K and maturity T, whose payoff is

$$Y = \left(\max_{0\le t\le T} S_t - K\right)^+.$$

Hint: You will need to sample from $(S_{t_1}, \ldots, S_{t_m}, \max_{0\le t\le T} S_t)$. Use Exercise 4.7 (d).

The function should have input parameters $S_0, r, \sigma, K, T, m, (t_1, \ldots, t_m)$, and n. Report your estimates and standard errors for

$$S_0 = 50,\ r = 0.05,\ \sigma = 0.2,\ K = 55,\ T = 1,$$

$$m = 50,\ t_i = iT/m,\ n = 10000.$$

6.D Antithetic sampling and stratified sampling can be combined. For example, suppose that one is interested in estimating $\mu = E[h(Z)]$, where Z is a standard normal random variable. Observe that antithetic sampling is equivalent to estimating $\mu = E[\bar{h}(Z)]$ by the plain Monte Carlo method, where

$$\bar{h}(Z) = \frac{1}{2}[h(Z) + h(-Z)].$$

Now one can use stratified sampling to further improve the performance. Write a function to estimate the price of call options, combining antithetic sampling and stratified sampling. The function should have input parameters r, σ, T, K, S_0, the number of strata k, and the sample size n. Report your simulation results for

$$r = 0.05,\ \sigma = 0.2,\ T = 1,\ K = 50,\ S_0 = 50,\ n = 10000,$$

and $k = 25, 100$, respectively. Use the proportional allocation.

In Exercises 6.E – 6.G, assume that the prices of the two underlying stocks are both geometric Brownian motions under the risk-neutral probability measure:

$$\begin{aligned} X_t &= X_0 \exp\left\{\left(r - \frac{1}{2}\sigma_1^2\right)T + \sigma_1 W_t\right\}, \\ Y_t &= Y_0 \exp\left\{\left(r - \frac{1}{2}\sigma_2^2\right)T + \sigma_2 B_t\right\}. \end{aligned}$$

Again, r is the risk-free interest rate, whereas (W, B) is a two-dimensional Brownian motion with covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

6.E Suppose that we wish to estimate the price of a spread call option with maturity T and payoff

$$H = (X_T - Y_T - K)^+.$$

Let $Z_1 = W_T/\sqrt{T}$ and $Z_2 = B_T/\sqrt{T}$.

(a) What is the joint distribution of (Z_1, Z_2)?

(b) Given $Z_2 = z$, what is the distribution of Z_1?

(c) What is the conditional expected value $E[H|Z_2 = z]$?

(d) Write a function to estimate the option price by the method of conditioning; see Exercise 6.13. The function should have input parameters r, σ_1, σ_2, ρ, X_0, Y_0, T, K, and the sample size n. Report your estimate and standard error for

$$X_0 = 50,\ Y_0 = 45,\ r = 0.05,\ \sigma_1 = 0.2,\ \sigma_2 = 0.3,\ \rho = 0.5,\ T = 1,$$

with $n = 10000$ and $K = 0, 5, 10$, respectively. Compare your results with those of Example 6.7.

6.F Suppose that we wish to estimate the price of a basket call option with maturity T and payoff

$$H = (c_1 X_T + c_2 Y_T - K)^+.$$

Let $Z_1 = W_T/\sqrt{T}$ and $Z_2 = B_T/\sqrt{T}$. Write a function to compare the following simulation schemes.

(a) Method of conditioning (conditional on Z_2). Be careful with the sign of $c_2 Y_T - K$ when calculating $E[H|Z_2 = z]$.

(b) Stratified sampling. Always use the proportional allocation.

i. Stratify Z_1 alone into k^2 strata.

ii. Stratify Z_2 alone into k^2 strata.

iii. Stratify Z_1 and Z_2 simultaneously into $k \times k = k^2$ strata.

The function should have input parameters r, σ_1, σ_2, ρ, X_0, Y_0, T, K, c_1, c_2, k, and the sample size n. Report your estimates and standard errors for

$$r = 0.1,\ \sigma_1 = 0.2,\ \sigma_2 = 0.3,\ X_0 = 50,\ Y_0 = 50,\ T = 1,\ K = 55,$$

$$\rho = 0.7,\ c_1 = 0.5,\ c_2 = 0.5$$

with $k = 10$ and sample size $n = 10000$.

6.G Consider the problem of estimating the price of a two-asset barrier option with maturity T and payoff

$$(X_T - K)^+ \cdot 1_{\{\min(Y_{t_1},\ldots,Y_{t_m})\geq h\}},$$

where $0 < t_1 < \cdots < t_m = T$ are prefixed dates.

(a) Write a function to estimate the option price by the control variate method. Let $(X_T - K)^+$ be the control variate and use $b = \hat{b}^*$.

(b) Write a function to estimate the option price by the method of conditioning, conditional on $(Y_{t_1}, \ldots, Y_{t_m})$ or equivalently $(B_{t_1}, \ldots, B_{t_m})$. *Hint:* Use Exercise 2.16 to argue that one can write $W = \rho B + \sqrt{1-\rho^2}Q$, where Q and B are independent standard Brownian motions. Use this to compute

$$E\left[(X_T - K)^+ \cdot 1_{\{\min(Y_{t_1},\ldots,Y_{t_m})\geq h\}} \,\middle|\, Y_{t_1}, \ldots, Y_{t_m}\right].$$

Compare your estimates and standard errors with those of Exercise 5.C for

$$X_0 = 50,\ Y_0 = 40,\ r = 0.03,\ \sigma_1 = 0.2,\ \sigma_2 = 0.4,\ \rho = 0.2,\ K = 50,$$

$$h = 38,\ T = 1,\ m = 50,\ t_i = iT/m,\ n = 10000.$$

This page intentionally left blank

Chapter 7

Importance Sampling

Importance sampling is a frequently used variance reduction technique in Monte Carlo simulation. It is particularly powerful for estimating small probabilities or expected values that are largely determined by events of small probabilities. In a nutshell, importance sampling draws samples from an alternative sampling distribution and compensates by multiplying the outcome with appropriate likelihood ratios.

The systematic construction of efficient importance sampling schemes requires advanced mathematical knowledge that is beyond the scope of this book. For this reason, we will limit our discussion to the basic ideas and rudimentary strategies of importance sampling, as well as the cross-entropy method, which is a simulation-based technique for selecting alternative sampling distributions. We will also address applications of importance sampling to option pricing and risk analysis.

7.1 Basic Ideas of Importance Sampling

Importance sampling is often referred to as a *change of measure* technique. To be more concrete, consider the problem of estimating the expected value

$$\mu = E[h(X)].$$

Suppose that X has a density $f(x)$. The plain Monte Carlo scheme will simulate iid samples $\{X_i\}$ from the density f and take the sample average of $\{h(X_i)\}$ as the estimate.

The basic idea of importance sampling comes from the observation that

for an *arbitrary* density function $g(x)$ [see Remark 7.1], one can write

$$\mu = \int_{\mathbb{R}} h(x) f(x)\, dx = \int_{\mathbb{R}} h(x) \frac{f(x)}{g(x)} g(x)\, dx = E\left[h(Y)\frac{f(Y)}{g(Y)}\right],$$

where Y is a random variable with density g. Therefore, one can draw iid samples $\{Y_i\}$ from the alternative density g and use the sample average of

$$h(Y_i)\frac{f(Y_i)}{g(Y_i)}$$

to estimate μ. Importance sampling is different from plain Monte Carlo in that samples are now generated from a different probability distribution (hence the name "change of measure") and in order to preserve unbiasedness, each outcome $h(Y_i)$ is weighed by the *likelihood ratio*

$$\frac{f(Y_i)}{g(Y_i)}.$$

In the preceding discussion, we have assumed that the random variable X has a density. The situation is similar when X is a discrete random variable. Suppose that the distribution of X is given by

$$\mathbb{P}(X = x_k) = p(x_k), \quad k = 1, 2, \ldots.$$

Consider a random variable Y with an alternative discrete probability distribution

$$\mathbb{P}(Y = x_k) = \bar{p}(x_k), \quad k = 1, 2, \ldots.$$

The importance sampling scheme generates iid copies of Y, say $\{Y_i\}$, and the estimate is the sample average of

$$h(Y_i)\frac{p(Y_i)}{\bar{p}(Y_i)}.$$

This estimate is unbiased as well.

Remark 7.1. To be more accurate, the alternative density function g should have the property that $f(x) = 0$ whenever $g(x) = 0$. In other words, the probability distribution associated with the density f is *absolutely continuous* with respect to the probability distribution associated with the density g. Analogous absolute continuity condition applies to the discrete probability distributions as well. Throughout the section, this requirement is implicitly imposed.

Remark 7.2. To ease exposition, we have assumed that X is a random variable. But the discussion obviously applies to the case where X is a general random vector, provided that one replaces the density function or the probability mass function by the joint density function or the joint probability mass function, respectively.

7.1.1 Guideline for Selecting Alternative Distributions

The key question in importance sampling is the choice of the alternative sampling distribution. With a poor choice, the variance of the importance sampling estimate might even exceed that of the plain Monte Carlo estimate. In order to achieve the greatest variance reduction, we should determine the alternative sampling distribution that minimizes the variance of the importance sampling estimate.

To this end, let us assume that the random variable X has a density f and the importance sampling scheme generates iid samples $\{Y_i\}$ from an alternative density g. The variance of the importance sampling estimate is

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)}\right] = \frac{1}{n}\text{Var}\left[h(Y)\frac{f(Y)}{g(Y)}\right],$$

where Y is a representative sample with density g. Consider the special case where h is nonnegative. If one picks the alternative sampling density g to be proportional to $h(x)f(x)$, that is,

$$g(x) = ch(x)f(x)$$

for some constant c, then the importance sampling estimate has zero variance! To implement such an alternative density requires the knowledge of c. However,

$$1/c = \int_{\mathbb{R}} h(x)f(x)\,dx = E[h(X)] = \mu,$$

which is the very quantity we wish to estimate. Thus the zero variance estimate has no practical value. Nevertheless, this discussion is not pointless. It leads to the general guideline for selecting an alternative sampling distribution, which states that the density g should be chosen to "mimic" $h(x)f(x)$.

7.1.2 Importance Sampling for Normal Distributions

Importance sampling for normal distributions is of particular interest in financial engineering. For a normal distribution, the alternative sampling

distribution is usually chosen from the class of normal distributions with the same variance.

To be more concrete, consider the problem of estimating $E[h(X)]$ where X is a standard normal random variable. Let f be the density of $N(0,1)$. The alternative sampling distribution is chosen as $N(\theta,1)$ for some $\theta \in \mathbb{R}$. Denote its density by $g(x)$. The corresponding likelihood ratio is

$$\frac{f(x)}{g(x)} = \exp\left\{-\theta x + \frac{1}{2}\theta^2\right\}.$$

In order to choose the parameter θ, recall that the principle is to pick a density $g(x)$ that mimics $h(x)f(x)$. This leads to the heuristic method of *mode matching*. That is, θ is chosen so as to match the mode of $h(x)f(x)$ with the mode of $g(x)$, which is precisely θ itself. In other words, we let $\theta = x^*$, where x^* maximizes $h(x)f(x)$, or equivalently,

$$\theta = \operatorname{argmax}_x h(x)f(x).$$

The mode x^* can often be computed numerically. The above discussion easily extends to the multivariate normal distributions; see for example, Exercise 7.H.

Example 7.1. Assume that under the risk-neutral probability measure, the price of the underlying stock is a geometric Brownian motion

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\}.$$

Use importance sampling to estimate the price of a binary call option with maturity T and payoff

$$1_{\{S_T \geq K\}}.$$

SOLUTION: The plain Monte Carlo scheme is straightforward. We include the simulation results in Table 7.1. As a benchmark, the theoretical values are calculated from Example 2.3. The parameters are given by

$$S_0 = 50, \quad r = 0.01, \quad \sigma = 0.1, \quad T = 1, \quad n = 10000.$$

It is clear from the simulation results that as the strike price K becomes larger, the performance of the plain Monte Carlo estimate deteriorates. The reason is similar to that of Example 4.5. That is, with a moderate sample size, only a few or even none of the samples will reach the strike price K when K is large, which leads to poor estimates.

Table 7.1: Plain Monte Carlo for binary option

	$K = 50$	$K = 60$	$K = 70$	$K = 80$
Theoretical value	0.5148	0.0377	4.541×10^{-4}	1.643×10^{-6}
Estimate	0.5097	0.0392	6.000×10^{-4}	0
S.E.	0.0050	0.0019	2.449×10^{-4}	0
R.E.	0.98%	4.95%	40.81%	NaN

As for the importance sampling scheme, observe that the price of the option is

$$E\left[e^{-rT}1_{\{S_T \geq K\}}\right] = E\left[e^{-rT}1_{\{X \geq b\}}\right],$$

where $X = W_T/\sqrt{T}$ is a standard normal random variable and

$$b = \frac{\log(K/S_0) - (r - \sigma^2/2)T}{\sigma\sqrt{T}}.$$

In other words, the price of the option is $v = E[h(X)]$ where

$$h(x) = e^{-rT}1_{\{x \geq b\}}.$$

Let f be the density of the standard normal distribution. Then

$$h(x)f(x) = \begin{cases} 0 & \text{if } x < b, \\ e^{-rT}f(x) & \text{if } x \geq b, \end{cases}$$

and the mode of $h(x)f(x)$ is $x^* = \max\{b, 0\}$. The method of mode matching suggests $N(x^*, 1)$ as the alternative sampling distribution. Therefore, the importance sampling estimate is the sample average of iid copies of

$$h(Y)\frac{f(Y)}{g(Y)} = \begin{cases} 0 & \text{if } Y < b, \\ \exp\{-rT - x^*Y + (x^*)^2/2\} & \text{if } Y \geq b, \end{cases}$$

where Y is distributed as $N(x^*, 1)$.

Pseudocode for importance sampling:

set $x^* = \max\{b, 0\}$
for $i = 1, 2, \ldots, n$
    generate a sample Y from $N(x^*, 1)$
    set $H_i = 0$ if $Y < b$
    set $H_i = \exp\left\{-rT - x^*Y + (x^*)^2/2\right\}$ if $Y \geq b$

$$\text{compute the estimate } \hat{v} = \frac{1}{n}\sum_{i=1}^{n} H_i$$

$$\text{compute the standard error S.E.} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}.$$

The simulation results are reported in Table 7.2. We use the same parameters and the sample size is again $n = 10000$.

Table 7.2: Importance sampling for binary call option

	$K = 50$	$K = 60$	$K = 70$	$K = 80$
Theoretical value	0.5148	0.0377	4.541×10^{-4}	1.643×10^{-6}
Estimate	0.5129	0.0379	4.717×10^{-4}	1.658×10^{-6}
S.E.	0.0050	0.0005	9.086×10^{-6}	3.796×10^{-8}
R.E.	0.97%	1.46%	1.93%	2.29%

The importance sampling scheme produces very accurate estimates even for large strike prices. ■

Example 7.2. The setup is the same as Example 7.1. Use importance sampling to estimate the price of a call option with maturity T and strike price K. Compare with the plain Monte Carlo scheme.

SOLUTION: The price of the call option is $v = E[h(X)]$, where X is a standard normal random variable and

$$h(x) = \left(S_0 \exp\left\{-\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}x\right\} - e^{-rT}K\right)^+.$$

We first run the plain Monte Carlo simulation and estimate the option price for a variety of strike prices. The parameters are

$$S_0 = 50,\ \ r = 0.05,\ \ \sigma = 0.2,\ \ T = 1,\ \ n = 10000.$$

The simulation results are reported in Table 7.3. The behavior of the plain Monte Carlo scheme is qualitatively similar to that of Example 7.1, that is, it performs poorly when the option is deep out of the money. The reason is the same—only for very few samples will the stock price at maturity exceed the strike price K when K is large. The theoretical values are obtained from the Black–Scholes formula.

Table 7.3: Plain Monte Carlo for call option

	$K = 50$	$K = 60$	$K = 80$	$K = 100$	$K = 120$
Theoretical value	5.2253	1.6237	0.0795	0.0024	6.066×10^{-5}
Estimate	5.2224	1.5943	0.0945	0.0035	0
S.E.	0.0740	0.0432	0.0104	0.0025	0
R.E.	1.42%	2.71%	11.00%	70.11%	NaN

Denote by f the density of $N(0,1)$ and by g the alternative sampling density, which is assumed to be $N(\theta,1)$. The mode matching method suggests that $\theta = x^*$, where x^* is the maximizer of $h(x)f(x)$. A typical picture of $h(x)f(x)$ is given in Figure 7.1.

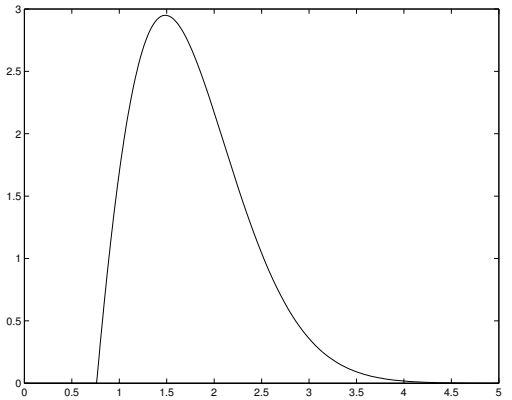


Figure 7.1: A representative picture of $h(x)f(x)$.

By taking the derivative of $h(x)f(x)$ and setting it to zero, it is easy to see that the maximizing x^* satisfies the equation

$$S_0 \exp\left\{-\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}x^*\right\}(\sigma\sqrt{T} - x^*) + e^{-rT}Kx^* = 0, \qquad (7.1)$$

which can be solved numerically by the bisection method (see Remark 7.3).

Pseudocode for importance sampling:

```
solve for x* from equation (7.1) by the bisection method
for i = 1, 2, ..., n
    generate a sample Y from N(x*, 1)
    set H_i = h(Y) exp{−x*Y + (x*)^2/2}
```

compute the estimate $\hat{v} = \dfrac{1}{n}\sum_{i=1}^{n} H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}$.

The results from importance sampling are reported in Table 7.4. Compared with the plain Monte Carlo scheme, the importance sampling scheme yields much better estimates, especially when the strike price K is large.

Table 7.4: Importance sampling for call option

	$K = 50$	$K = 60$	$K = 80$	$K = 100$	$K = 120$
Theoretical value	5.2253	1.6237	0.0795	0.0024	6.066×10^{-5}
Estimate	5.2349	1.6310	0.0802	0.0024	5.957×10^{-5}
S.E.	0.0247	0.0112	0.0008	3.0×10^{-5}	8.631×10^{-7}
R.E.	0.47%	0.68%	1.00%	1.26%	1.45%
x^*	0.9849	1.4849	2.6002	3.6014	4.4569

Remark 7.3. The Bisection Method. Let h be a continuous function. The bisection method is a very simple algorithm for finding a root of the equation

$$h(x) = 0.$$

It is based on the intermediate value theorem, which asserts that if $h(x_1)$ and $h(x_2)$ have different signs, then there must exist a root on the interval $[x_1, x_2]$. To be more precise, suppose that $h(x_1) < 0 < h(x_2)$. Consider the midpoint

$$x_m = (x_1 + x_2)/2.$$

If $h(x_m) = 0$, then we have found a root. If $h(x_m) < 0$, then there must be a root on the interval $[x_m, x_2]$. Replace x_1 by x_m and repeat. Similarly, if $h(x_m) > 0$, then there must exist a root on $[x_1, x_m]$. Replace x_2 by x_m and repeat. Each iteration will halve the width of the interval, and the algorithm always converges to a root. The pseudocode for the bisection method is included. The algorithm will not end until a prescribed precision is achieved.

Pseudocode for solving $h(x) = 0$ by the bisection method:

initialize x_1 and x_2 — $h(x_1)$ and $h(x_2)$ must have opposite signs
prescribe a precision level ε
while $|x_1 - x_2| \geq \varepsilon$

set $x_m = (x_1 + x_2)/2$
if $h(x_m) = 0$, set $x_1 = x_m$ and $x_2 = x_m$
if $h(x_m) \neq 0$
    set $x_1 = x_m$ if $h(x_m)$ has the same sign as $h(x_1)$
    set $x_2 = x_m$ otherwise
return x_1. ■

7.1.3 Importance Sampling for General Distributions

For a general random variable X, the alternative sampling distribution is often chosen from the so-called *exponential tilt* family. This is suggested by the asymptotic analysis of rare events under the right scaling [31, 33].

a. X is continuous: Let f denote the density of X. The exponential tilt family consists of probability distributions with density

$$f_\theta(x) = \frac{1}{E[e^{\theta X}]} e^{\theta x} f(x)$$

for some $\theta \in \mathbb{R}$.

b. X is discrete: Assume that the probability mass function of X is p, that is, $\mathbb{P}(X = x) = p(x)$. The exponential tilt family consists of probability distributions with probability mass function

$$p_\theta(x) = \frac{1}{E[e^{\theta X}]} e^{\theta x} p(x)$$

for some $\theta \in \mathbb{R}$.

The parameter θ is called the *tilting parameter*. The distribution determined by f_θ or p_θ is said to be the *exponential tilt distribution of X with parameter θ*. Note that when X is normally distributed, the exponential tilt family is the collection of all normal distributions with the same variance.

It is beyond the scope of this book to state a general rule for selecting the best member from the exponential tilt family. But some of the ideas can be illustrated through the following example, where we estimate the probability of a large loss in a credit risk model. Importance sampling can be much more efficient than the plain Monte Carlo in this context since such probabilities are usually very small.

Example 7.3. Consider a much simplified credit risk model with m independent obligors. Denote by p_k the probability that the k-th obligor defaults and c_k the loss resulting from its default. Assuming that $c_k = 1$ for every k, the total loss is

$$L = \sum_{k=1}^{m} c_k X_k = \sum_{k=1}^{m} X_k,$$

where X_k is the default indicator for the k-th obligor, that is, $\{X_k\}$ are independent Bernoulli random variables such that

$$\mathbb{P}(X_k = 1) = p_k, \quad \mathbb{P}(X_k = 0) = 1 - p_k.$$

Assuming that x is a large threshold, use importance sampling to estimate the tail probability $\mathbb{P}(L > x)$.

SOLUTION: To exclude the trivial case, throughout the discussion we assume that

$$x > E[L] = \sum_{k=1}^{m} p_k.$$

Otherwise the event $\{L > x\}$ is not rare, and the plain Monte Carlo scheme suffices to produce an accurate estimate.

The alternative sampling distribution will be chosen from the exponential tilt family; see Exercise 7.7. Given $\theta \in \mathbb{R}$, let Y_k's be independent Bernoulli random variables such that

$$\bar{p}_k = \mathbb{P}(Y_k = 1) = \frac{1}{E[e^{\theta X_k}]} e^{\theta} \cdot \mathbb{P}(X_k = 1) = \frac{p_k e^{\theta}}{1 + p_k(e^{\theta} - 1)},$$

$$\mathbb{P}(Y_k = 0) = 1 - \bar{p}_k$$

for every k. Letting

$$\bar{L} = \sum_{k=1}^{m} Y_k,$$

the corresponding importance sampling estimate is the sample average of iid copies of

$$H = 1_{\{\bar{L} > x\}} \prod_{k=1}^{m} \left(\frac{p_k}{\bar{p}_k}\right)^{Y_k} \left(\frac{1 - p_k}{1 - \bar{p}_k}\right)^{1 - Y_k}.$$

Since we would like to have $\bar{p}_k > p_k$ for every k so that it is more likely for the total loss to reach the threshold x under the alternative sampling distribution, θ is restricted to be positive.

It remains to choose an appropriate θ in order to achieve as much variance reduction as possible. Since H is an unbiased estimate, it suffices to make $E[H^2]$ as small as possible. Observe that

$$\begin{aligned} E[H^2] &= E\left[1_{\{\bar{L}>x\}} \prod_{k=1}^{m} \left(\frac{p_k}{\bar{p}_k}\right)^{2Y_k} \left(\frac{1-p_k}{1-\bar{p}_k}\right)^{2(1-Y_k)}\right] \\ &= E\left[1_{\{\bar{L}>x\}} e^{-2\theta\bar{L}+2\phi(\theta)}\right], \end{aligned}$$

where

$$\phi(\theta) = \sum_{k=1}^{m} \log[1 + p_k(e^{\theta} - 1)].$$

It follows that, for $\theta > 0$

$$E[H^2] \leq e^{-2\theta x + 2\phi(\theta)}. \tag{7.2}$$

The idea is to find a θ that minimizes the upper bound (see Exercise 7.8), or equivalently,

$$-2\theta x + 2\phi(\theta).$$

It can be easily argued that the minimizing θ^* is the unique positive solution to the equation

$$\phi'(\theta) = x \tag{7.3}$$

when $x > E[L]$. The bisection method can be used to numerically solve for θ^*.

Pseudocode:

solve for the unique positive root θ^* from (7.3) by the bisection method
compute the corresponding $\bar{p}_k$ for $k = 1, \ldots, m$
for $i = 1, 2, \ldots, n$
  generate Y_k from Bernoulli with parameter $\bar{p}_k$ for $k = 1, \ldots, m$
  set $L = Y_1 + \cdots + Y_m$
  set $H_i = \prod_{k=1}^{m} \left(\frac{p_k}{\bar{p}_k}\right)^{Y_k} \left(\frac{1-p_k}{1-\bar{p}_k}\right)^{1-Y_k}$ if $L > x$; otherwise set $H_i = 0$
compute the estimate $\hat{v} = \frac{1}{n}\sum_{i=1}^{n} H_i$
compute the standard error S.E. $= \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}$.

The simulation results are presented in Table 7.5. The parameters are given by

$$m = 1000, \quad p_k = 0.01 \cdot [1 + e^{-k/m}], \quad k = 1, 2, \ldots, m.$$

The sample size is $n = 10000$.

Table 7.5: Credit risk model: Importance sampling versus plain Monte Carlo

	$x = 20$		$x = 30$		$x = 40$	
	IS	plain	IS	plain	IS	plain
Estimate	0.1482	0.1533	6.936×10^{-4}	0.0011	1.534×10^{-7}	0
S.E.	0.0019	0.0036	1.375×10^{-5}	0.0003	3.847×10^{-9}	0
R.E.	1.30%	2.35%	1.98%	30.14%	2.51%	NaN

Importance sampling proves to be much more efficient than the plain Monte Carlo scheme. ■

7.2 The Cross-Entropy Method

The cross-entropy method is a relatively new simulation technique originated by Reuven Rubinstein [28]. It is a very versatile methodology that can be employed to improve the efficiency of Monte Carlo simulation or solve difficult combinatorial optimization problems. We should mainly focus on the application of the cross-entropy method to importance sampling. A comprehensive treatment can be found in [29].

The cross-entropy method is essentially a simulation-based technique for selecting alternative sampling distributions. In order to illustrate the main idea, consider the generic problem of estimating the expected value

$$\mu = E[h(X)]$$

by importance sampling. To ease exposition, assume that X is a random variable with density $f(x)$ and h is a nonnegative function. The alternative sampling distribution is usually restricted to a parameterized family of density functions $\{f_\theta(x)\}$ that contains the original density f. A particularly popular choice is the exponential tilt family described in Section 7.1.3. For this reason, the reference parameter θ is often said to be the *tilting parameter* as well.

As we have discussed previously in Section 7.1.1, the importance sampling estimate based on the alternative sampling density

$$g^*(x) = \frac{1}{\mu} h(x) f(x) \tag{7.4}$$

has zero variance. Even though such a sampling distribution is impractical as it requires the knowledge of μ, it leads to the heuristic principle that an alternative sampling density "close" to $g^*(x)$ should be a good choice for importance sampling. The cross-entropy method aims to solve for the density $f_\theta(x)$ that is *closest* to $g^*(x)$ in the sense of *Kullback–Leibler cross entropy* or *relative entropy*, which is defined by

$$R(g^* \| f_\theta) = \int_{\mathbb{R}} \log \frac{g^*(x)}{f_\theta(x)} \cdot g^*(x)\, dx.$$

That is, the cross-entropy method chooses the minimizing density of the minimization problem

$$\min_\theta R(g^* \| f_\theta) \tag{7.5}$$

as the alternative sampling density for importance sampling. Plugging in formula (7.4), it follows that

$$\begin{aligned} R(g^* \| f_\theta) &= \int_{\mathbb{R}} g^*(x) \log g^*(x)\, dx - \int_{\mathbb{R}} g^*(x) \log f_\theta(x)\, dx \\ &= \int_{\mathbb{R}} g^*(x) \log g^*(x)\, dx - \frac{1}{\mu} \int_{\mathbb{R}} h(x) f(x) \log f_\theta(x)\, dx. \end{aligned}$$

Since neither the first term nor μ depends on θ, the minimization problem (7.5) is equivalent to the maximization problem

$$\max_\theta \int_{\mathbb{R}} h(x) f(x) \log f_\theta(x)\, dx. \tag{7.6}$$

This maximization problem does not admit explicit solutions in general. The cross-entropy method produces a simple, simulation-based, iterative algorithm to solve for the maximizing θ. At each step of the iteration, the maximizing θ is approximated by an *explicitly* computable quantity from a relatively small number of pilot samples. The extra computational cost incurred by these pilot samples and iterations is often significantly outweighed by the resulting variance reduction.

Remark 7.4. The Kullback–Leibler cross entropy is a measure of how close two probability distributions are. It is always nonnegative and takes value zero if and only if the two probability distributions coincide; see Exercise 7.10. Even though there are other definitions of distance between two probability distributions, the cross entropy is convenient as it is analytically simpler to work with.

Remark 7.5. Even though we have assumed that X is a continuous random variable for notational clarity, the cross-entropy method easily extends to the general cases where X is discrete or a random vector. Similarly, θ can also be a vector itself. All the ensuing discussions and formulas are still valid as long as one replaces the density functions with probability mass functions or their multivariate versions and replaces derivatives with gradients.

7.2.1 The Basic Cross-Entropy Algorithm

Consider the maximization problem (7.6). Since X has density f, one can rewrite

$$\int_{\mathbb{R}} h(x) f(x) \log f_\theta(x)\, dx = E\left[h(X) \log f_\theta(X)\right].$$

Under mild conditions such as $f_\theta(x)$ is differentiable with respect to θ and so on, the maximizer is the solution to the equation

$$0 = \frac{\partial}{\partial \theta} E[h(X) \log f_\theta(X)] = E\left[h(X) \frac{\partial}{\partial \theta} \log f_\theta(X)\right].$$

This equation is not explicitly solvable in general. However, if one replaces the expected value by sample average and considers the corresponding stochastic version

$$0 = \frac{1}{N} \sum_{k=1}^{N} h(X_k) \frac{\partial}{\partial \theta} \log f_\theta(X_k), \tag{7.7}$$

where $\{X_k\}$ are iid copies of X, then it is often possible to obtain a solution in *closed form*; see Exercise 7.11. Note that when θ is a vector, the partial derivative $\partial/\partial\theta$ denotes the gradient with respect to θ.

The solution to (7.7) takes a particularly simple form in the context of normal distributions. More precisely, we have the following lemma, whose proof is straightforward and thus omitted.

Lemma 7.1. *Suppose that $f(x)$ is the density of $N(0, I_m)$ and $f_\theta(x)$ denotes the density of $N(\theta, I_m)$ where $\theta = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$. Then the solution to (7.7) is simply*

$$\hat{\theta} = \frac{\sum_{k=1}^N h(X_k) X_k}{\sum_{k=1}^N h(X_k)},$$

where $\{X_1, \ldots, X_N\}$ are iid samples from $N(0, I_m)$.

To summarize, the cross-entropy method generates N iid pilot samples $X_1, \ldots, X_N$ from the original density $f(x)$ and compute $\hat{\theta}$ from equation (7.7). Once $\hat{\theta}$ is obtained, estimate $\mu = E[h(X)]$ by importance sampling with the alternative sampling density $f_{\hat{\theta}}(x)$. The pilot sample size N is usually chosen to be much smaller than the sample size for the estimation of μ. Below is the pseudocode.

Pseudocode for the basic cross-entropy algorithm:

generate N iid pilot samples $X_1, \ldots, X_N$ from density $f(x)$
obtain $\hat{\theta}$ by solving (7.7)
for $i = 1, 2, \ldots, n$
    generate Y_i from the alternative sampling density $f_{\hat{\theta}}(x)$
    set $H_i = h(Y_i) f(Y_i) / f_{\hat{\theta}}(Y_i)$.
compute the estimate $\hat{v} = \dfrac{1}{n} \sum_{i=1}^n H_i$
compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)} \left(\sum_{i=1}^n H_i^2 - n\hat{v}^2 \right)}$.

Example 7.4. Assume that under the risk-neutral probability measure, the underlying stock price is a geometric Brownian motion

$$S_t = S_0 \exp\left\{ \left(r - \frac{1}{2}\sigma^2 \right) t + \sigma W_t \right\},$$

where r is the risk-free interest rate. Estimate the price of a call option with strike price K and maturity T. Compare with Example 7.2.

SOLUTION: The price of the call option is $v = E[h(X)]$, where X is a standard normal random variable and

$$h(x) = \left(S_0 \exp\left\{ -\frac{1}{2}\sigma^2 T + \sigma\sqrt{T} x \right\} - e^{-rT} K \right)^+.$$

The family of alternative sampling densities is $\{f_\theta : \theta \in \mathbb{R}\}$, where f_θ is the density of $N(\theta, 1)$. The solution $\hat{\theta}$ to equation (7.7) is given by Lemma 7.1. Below is the pseudocode.

Pseudocode for call option by the cross-entropy method:

generate N iid pilot samples $X_1, \ldots, X_N$ from $N(0,1)$
set $\hat{\theta} = \sum_{k=1}^{N} h(X_k) X_k / \sum_{k=1}^{N} h(X_k)$
for $i = 1, 2, \ldots, n$
    generate a sample Y from $N(\hat{\theta}, 1)$
    set $H_i = h(Y) \exp\{-\hat{\theta} Y + \hat{\theta}^2/2\}$
compute the estimate $\hat{v} = \dfrac{1}{n} \sum_{i=1}^{n} H_i$
compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)} \left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2 \right)}$.

The numerical results are reported in Table 7.6. The pilot sample size is $N = 2000$, and the sample size for importance sampling is $n = 10000$. The parameters are again given by

$$S_0 = 50, \;\; r = 0.05, \;\; \sigma = 0.2, \;\; T = 1.$$

Table 7.6: Basic cross-entropy method for call option

	$K = 50$	$K = 60$	$K = 80$	$K = 100$		$K = 120$
True value	5.2253	1.6237	0.0795	0.0024		6.066×10^{-5}
Estimate	5.2166	1.6273	0.0805	0.0024	NaN	NaN
S.E.	0.0243	0.0109	0.0008	3.2×10^{-5}	NaN	NaN
R.E.	0.47%	0.67%	0.99%	1.32%	NaN	NaN
$\hat{\theta}$	1.2057	1.7525	2.8247	3.3781	NaN	NaN

Compared with the simulation results in Example 7.2, the performance of the basic cross-entropy scheme is indistinguishable from that of the importance sampling scheme using the mode matching method, when K is not exceedingly large. However, around $K = 100$, the basic cross-entropy scheme starts to behave erratically: sometimes it yields a great estimate, and sometimes it produces an "NaN", which stands for "Not a Number"

in MATLAB®. When $K = 120$, almost all the results from the cross-entropy method are "NaN". It is because in the estimation of $\hat{\theta}$, the denominator

$$\sum_{k=1}^{N} h(X_k)$$

is more likely to become zero as K increases. This example suggests that a modification of the basic cross-entropy algorithm is necessary when the simulation is dominated by events with very small probabilities. Such development will be discussed later in the book. ■

Example 7.5. Consider a discretely monitored average price call option with payoff $(\bar{S} - K)^+$ and maturity T. Here $\bar{S}$ is the arithmetic mean

$$\bar{S} = \frac{1}{m} \sum_{k=1}^{m} S_{t_k}$$

for a given set of dates $0 < t_1 < \cdots < t_m = T$. Assume that under the risk-neutral probability measure the price of the underlying asset is a geometric Brownian motion

$$S_t = S_0 \exp\left\{ \left(r - \frac{1}{2}\sigma^2 \right) t + \sigma W_t \right\},$$

where r is the risk-free interest rate. Estimate the option price.

SOLUTION: For $1 \le k \le m$, let $Z_k = (W_{t_k} - W_{t_{k-1}})/\sqrt{t_k - t_{k-1}}$. Then $Z_1, \ldots, Z_m$ are iid standard normal random variables and $X = (Z_1, \ldots, Z_m)$ is an m-dimensional standard normal random vector. The stock prices at time t_k can be written as a function of X:

$$S_{t_k} = S_0 \exp\left\{ \left(r - \frac{1}{2}\sigma^2 \right) t_k + \sigma \sum_{j=1}^{k} \sqrt{t_j - t_{j-1}}\, Z_j \right\}. \tag{7.8}$$

Consequently, the discounted option payoff can also be written as a function of X, say

$$h(X) = e^{-rT} (\bar{S} - K)^+. \tag{7.9}$$

Denote the joint density function of X by f, that is, for $x = (x_1, \ldots, x_m)$,

$$f(x) = \left(\frac{1}{\sqrt{2\pi}} \right)^m \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} x_i^2 \right\}.$$

Suppose that the family of alternative sampling densities is $\{N(\theta, I_m) : \theta \in \mathbb{R}^m\}$. Let f_θ be the density of $N(\theta, I_m)$, that is,

$$f_\theta(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^m \exp\left\{-\frac{1}{2}\sum_{i=1}^m (x_i - \theta_i)^2\right\}$$

for $\theta = (\theta_1, \ldots, \theta_m)$. The importance sampling estimate is the sample average of iid copies of

$$h(Y)\frac{f(Y)}{f_\theta(Y)} = h(Y)\exp\left\{-\sum_{i=1}^m \theta_i Y_i + \frac{1}{2}\sum_{i=1}^m \theta_i^2\right\},$$

where $Y = (Y_1, \ldots, Y_m)$ has distribution $N(\theta, I_m)$, that is, $\{Y_1, \ldots, Y_m\}$ are independent and Y_i is distributed as $N(\theta_i, 1)$ for each i.

A good tilting parameter θ can be determined by the basic cross-entropy scheme. Indeed, let $X_1, \ldots, X_N$ be iid pilot samples from the original distribution $N(0, I_m)$. Then the solution to equation (7.7) is given by Lemma 7.1, i.e.,

$$\hat{\theta} = \frac{\sum_{k=1}^N h(X_k) X_k}{\sum_{k=1}^N h(X_k)}.$$

Below is the pseudocode.

Pseudocode for average price call by the cross-entropy method:

generate iid pilot samples $X_1, \ldots, X_N$ from $N(0, I_m)$
set $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m) = \sum_{k=1}^N h(X_k)X_k / \sum_{k=1}^N h(X_k)$
(▷) for $i = 1, 2, \ldots, n$
  for $j = 1, 2, \ldots, m$
    generate Y_j from $N(\hat{\theta}_j, 1)$
    set $S_{t_j} = S_{t_{j-1}} \exp\left\{(r - \sigma^2/2)(t_j - t_{j-1}) + \sigma\sqrt{t_j - t_{j-1}}Y_j\right\}$
(□) compute the discounted payoff multiplied by the likelihood ratio

$$H_i = e^{-rT}(\bar{S} - K)^+ \cdot \exp\left\{-\sum_{j=1}^m \hat{\theta}_j Y_j + \frac{1}{2}\sum_{j=1}^m \hat{\theta}_j^2\right\}$$

compute the estimate $\hat{v} = \dfrac{1}{n}\sum_{i=1}^n H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^n H_i^2 - n\hat{v}^2\right)}$.

We should mention that the first two lines of the pseudocode calculate the tilting parameter using the cross-entropy method. The details of evaluating $h(X_k)$ are left out for two reasons: (i) it is straightforward from (7.9), the definition of h; (ii) it is indeed very much similar to the lines from ($\triangleright$) to ($\square$). We will fill in the missing details below for the purpose of illustration. However, such details will not be given for the rest of the book in order to avoid repetition.

More details for the first two lines in the previous pseudocode:

$$
\begin{aligned}
&\text{for } k = 1, 2, \ldots, N \\
&\quad \text{for } j = 1, 2, \ldots, m \\
&\quad\quad \text{generate } Z_j \text{ from } N(0,1) \\
&\quad\quad \text{set } S_{t_j} = S_{t_{j-1}} \exp\left\{(r - \sigma^2/2)(t_j - t_{j-1}) + \sigma\sqrt{t_j - t_{j-1}}Z_j\right\} \\
&\quad \text{set } X_k = (Z_1, \ldots, Z_m) \\
&\quad \text{set } R_k = h(X_k) = e^{-rT}(\bar{S} - K)^+ \\
&\text{set } \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m) = \textstyle\sum_{k=1}^N R_k X_k / \sum_{k=1}^N R_k
\end{aligned}
$$

The numerical results are reported in Table 7.7. We compare the cross-entropy method with the plain Monte Carlo for

$$S_0 = 50, \;\; r = 0.05, \;\; \sigma = 0.2, \;\; T = 1, \;\; m = 12, \;\; t_i = \frac{i}{m}.$$

The sample size for importance sampling is $n = 10000$ and the pilot sample size for the cross-entropy method is $N = 2000$. The cross-entropy method reduces the variance significantly, especially when K is large.

Table 7.7: Cross-entropy method versus plain Monte Carlo

Strike price	$K = 50$		$K = 60$		$K = 70$	
	CE	plain MC	CE	plain MC	CE	plain MC
Estimate	3.0637	3.0998	0.3404	0.3178	0.0168	0.0237
S.E.	0.0155	0.0429	0.0033	0.0144	0.0004	0.0042
R.E.	0.51%	1.38%	0.96%	4.54%	2.44%	17.8%

It should be noted that the mode matching method can also be applied to this problem. Even though there are m unknowns, it can be reduced to a single equation, which can then be solved by the bisection method [11]. ■

7.2.2 The General Iterative Cross Entropy Algorithm

The basic cross-entropy algorithm is the simple version of a more general iterative procedure for solving the maximization problem (7.6). The latter is particularly useful in the context of simulating events with very small probabilities.

Every iteration in this general scheme involves two phases. In the i-th iteration, one first generates iid samples from the density function $f_\theta(x)$ with $\theta = \hat{\theta}^{i-1}$ being the current candidate of the tilting parameter. The tilting parameter $\hat{\theta}^{i-1}$ is then updated to $\hat{\theta}^i$ based on these samples. Here we have used a superscript instead of a subscript because it is very common that θ is a vector and we would like to reserve the subscript for the individual components of θ. As in the basic algorithm, $\hat{\theta}^i$ often admits *analytical* formulas. If $\hat{\theta}^*$ is the tilting parameter from the final iteration, then $f_{\hat{\theta}^*}$ is used as the alternative sampling density in the importance sampling scheme to estimate $\mu = E[h(X)]$, the quantity of interest.

The iterative cross-entropy algorithm is based on the following observation. Define the likelihood ratio

$$\ell_\theta(x) = \frac{f(x)}{f_\theta(x)}. \tag{7.10}$$

Consider the maximization problem (7.6). Fixing an *arbitrary* tilting parameter, say ν, we can rewrite the integral in (7.6) as

$$\int_{\mathbb{R}} h(x)\ell_\nu(x) \log f_\theta(x) \cdot f_\nu(x)\, dx = E\left[h(Y)\ell_\nu(Y) \log f_\theta(Y)\right],$$

where Y is a random variable with density $f_\nu(x)$. Consequently, the maximizer to (7.6) satisfies the equation

$$0 = E\left[h(Y)\ell_\nu(Y)\frac{\partial}{\partial\theta}\log f_\theta(Y)\right]. \tag{7.11}$$

As before, we replace the expected value by sample average and solve the equation

$$0 = \frac{1}{N}\sum_{k=1}^{N} h(Y_k)\ell_\nu(Y_k)\frac{\partial}{\partial\theta}\log f_\theta(Y_k),$$

where $Y_1, \ldots, Y_N$ are iid pilot samples from the density f_ν. This leads to the following updating rule for $\hat{\theta}$:

The updating rule of $\hat{\theta}$. Suppose that $\hat{\theta}^j$ is the value of the tilting parameter at the end of the j-th iteration. Let $\hat{\theta}^{j+1}$ be the solution to the equation

$$0 = \frac{1}{N}\sum_{k=1}^{N} h(Y_k)\ell_{\hat{\theta}^j}(Y_k)\frac{\partial}{\partial\theta}\log f_\theta(Y_k), \tag{7.12}$$

where $Y_1, \ldots, Y_N$ are iid pilot samples from the density $f_{\hat{\theta}^j}$.

Equation (7.12) is of exactly the same form as the basic cross-entropy equation (7.7), except that X_k is replaced by Y_k and $h(X_k)$ by $h(Y_k)\ell_{\hat{\theta}^j}(Y_k)$. Therefore, just like (7.7) it is often explicitly solvable; see Exercise 7.11. In particular, we have the following result, which is an immediate corollary of Lemma 7.1.

Lemma 7.2. *Suppose that $f(x)$ is the density of $N(0, I_m)$ and $f_\theta(x)$ is the density of $N(\theta, I_m)$ where $\theta = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$. Then the solution to (7.12) is simply*

$$\hat{\theta}^{j+1} = \frac{\sum_{k=1}^N h(Y_k)\ell_{\hat{\theta}^j}(Y_k)Y_k}{\sum_{k=1}^N h(Y_k)\ell_{\hat{\theta}^j}(Y_k)} = \frac{\sum_{k=1}^N h(Y_k)e^{-\langle\hat{\theta}^j, Y_k\rangle}Y_k}{\sum_{k=1}^N h(Y_k)e^{-\langle\hat{\theta}^j, Y_k\rangle}},$$

where $\{Y_1, \ldots, Y_N\}$ are iid samples from $N(\hat{\theta}^j, I_m)$ and $\langle\cdot,\cdot\rangle$ is the inner product of two vectors defined by

$$\langle\theta, y\rangle = \sum_{i=1}^m \theta_i y_i, \quad \text{if } \theta = (\theta_1, \ldots, \theta_m),\ y = (y_1, \ldots, y_m).$$

Note that in the one-dimensional case (i.e., $m = 1$), the inner product is just the regular product of numbers.

If the initial tilting parameter is reasonably chosen, the number of iterations necessary to reach a good final tilting parameter is very small in practice: four or five iterations are in general sufficient, and for many problems one or two iterations are all that is needed. Since $\hat{\theta}^j$'s are estimates based on samples, we cannot expect the convergence in the classical sense that $|\hat{\theta}^{j+1} - \hat{\theta}^j| \to 0$ as j tends to infinity. The "convergence" is reached when $\hat{\theta}^j$ starts to oscillate with small variations.

The choice of the initial tilting parameter $\hat{\theta}^0$ is quite flexible in general. For example, in many situations it suffices to let $\hat{\theta}^0 = \theta^0$, where θ^0 corresponds to the original distribution $f(x)$, that is, $f = f_{\theta^0}$. With this initial

choice, $\ell_{\hat{\theta}^0}(x) = 1$ and equation (7.12) reduces to (7.7). In other words, the basic cross-entropy algorithm is equivalent to the *first* iteration of the general iterative scheme with $\hat{\theta}^0 = \theta^0$. However, when the simulation is related to events with very small probabilities, the choice of $\hat{\theta}^0$ is less straightforward. The reason is as follows. As we have observed in Example 7.4, when the simulation is largely determined by rare events, the basic cross-entropy algorithm, and hence the first iteration of the general scheme, will frequently produce "NaN" if one lets $\hat{\theta}^0 = \theta^0$.

In financial engineering, many quantities of interest are the expected values of random variables that can only be nonzero on sets of the form $\{R \geq a\}$. The random variable R can be, for example, the underlying asset price or the total loss in a risk model, and a can be some given level related to the strike price, barrier, or loss threshold, and so on. The difficulty arises when a is large. The rule of thumb for a good initial tilting parameter $\hat{\theta}^0$ is that it should be chosen in an "economical" way, or equivalently, the original distribution should be tilted just enough, to ensure that $\{R \geq a\}$ is no longer rare. This in general means that we should let

$$E_{\hat{\theta}^0}[R] = a,$$

where $E_{\hat{\theta}^0}[\cdot]$ denotes the expected value taken under the alternative sampling density $f_{\hat{\theta}^0}$. This expected value may or may not admit an analytical expression, but it can often be approximated.

We should demonstrate how to judiciously choose an initial tilting parameter through examples in this section and postpone the discussion of a general initialization technique to the next.

Example 7.6. Let us revisit the problem of estimating the call option price in Example 7.4. The basic cross-entropy method fails when the strike price K is exceedingly large. Design an iterative cross-entropy scheme to resolve this issue.

SOLUTION: Recall that the call option price is $v = E[h(X)]$, where X is a standard normal random variable and

$$h(X) = \left(S_0 \exp\left\{ -\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}X \right\} - e^{-rT}K \right)^+ .$$

The family of alternative sampling densities is $\{f_\theta : \theta \in \mathbb{R}\}$, where f_θ is the density of $N(\theta, 1)$. The updating rule of $\hat{\theta}^j$ is given by Lemma 7.2 and is straightforward. The key issue is the initialization of the tilting parameter, especially when K is large.

A good initial tilting parameter $\hat{\theta}^0$ should alter the original distribution just enough so that $\{h(Y) > 0\}$ is no longer a rare event if Y is distributed as $N(\hat{\theta}^0, 1)$, or equivalently, the stock price exceeds K with nontrivial probability under the alternative sampling density $f_{\hat{\theta}^0}$. Under these considerations, a natural choice of $\hat{\theta}^0$ is such that

$$E\left(S_0 \exp\left\{-\frac{1}{2}\sigma^2 T + \sigma\sqrt{T}Y\right\} - e^{-rT}K\right) = 0,$$

where Y is normally distributed as $N(\hat{\theta}^0, 1)$. It is straightforward to calculate the expected value and solve the equation to obtain

$$\hat{\theta}^0 = \frac{1}{\sigma\sqrt{T}} \log\left(\frac{K}{S_0}\right) - \frac{r}{\sigma}\sqrt{T}. \tag{7.13}$$

Below is the pseudocode. The total number of iterations is denoted by IT_NUM.

Pseudocode for call option by the iterative cross-entropy method:

initialize $\hat{\theta}^0$ by (7.13) and set the iteration counter $j = 0$

(*) generate N iid samples $Y_1, \ldots, Y_N$ from $N(\hat{\theta}^j, 1)$

set $\hat{\theta}^{j+1} = \sum_{k=1}^N h(Y_k) e^{-\hat{\theta}^j Y_k} Y_k / \sum_{k=1}^N h(Y_k) e^{-\hat{\theta}^j Y_k}$

set the iteration counter $j = j + 1$

if $j =$ IT_NUM set $\hat{\theta} = \hat{\theta}^j$ and continue, otherwise go to step (*)

for $i = 1, 2, \ldots, n$

generate a sample Y from $N(\hat{\theta}, 1)$

set $H_i = h(Y) \exp\{-\hat{\theta}Y + \hat{\theta}^2/2\}$

compute the estimate $\hat{v} = \dfrac{1}{n}\sum_{i=1}^n H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^n H_i^2 - n\hat{v}^2\right)}$.

The numerical results are reported in Table 7.8. The pilot sample size is $N = 2000$ and the sample size for importance sampling is $n = 10000$. The number of iteration is IT_NUM $= 5$. The parameters are again given by

$$S_0 = 50, \;\; r = 0.05, \;\; \sigma = 0.2, \;\; T = 1.$$

Compared with the simulation results of the basic cross-entropy algorithm in Example 7.2, the improvement from the general iterative cross-entropy method is negligible when K is moderate. The reason is that if one regards the basic cross-entropy algorithm as a special case of the general iterative scheme with $\hat{\theta}^0 = \theta^0 = 0$ and IT_NUM $= 1$, then the convergence is already reached with one iteration. On the other hand, the general scheme consistently yields very accurate estimates when K is large.

Table 7.8: Iterative cross-entropy method for call option

	$K = 50$	$K = 60$	$K = 80$	$K = 100$	$K = 120$
True value	5.2253	1.6237	0.0795	0.0024	6.066×10^{-5}
Estimate	5.2603	1.6338	0.0791	0.0024	6.042×10^{-5}
S.E.	0.0240	0.0109	0.0008	2.987×10^{-5}	8.667×10^{-7}
R.E.	0.46%	0.67%	1.01%	1.25%	1.43%
$\hat{\theta}$	1.2330	1.7632	2.8631	3.8209	4.6657

Table 7.9: Iteration of the tilting parameter $\hat{\theta}^j$

	$\hat{\theta}^0$	$\hat{\theta}^1$	$\hat{\theta}^2$	$\hat{\theta}^3$	$\hat{\theta}^4$	$\hat{\theta}^5 = \hat{\theta}$
$K = 50$	-0.2500	1.1770	1.2127	1.2381	1.2092	1.2330
$K = 60$	0.6616	1.7681	1.7468	1.7802	1.7814	1.7632
$K = 80$	2.1000	2.8451	2.8605	2.8731	2.8676	2.8631
$K = 100$	3.2157	3.8165	3.8223	3.8196	3.8162	3.8209
$K = 120$	4.1273	4.6518	4.6508	4.6479	4.6472	4.6657

The iteration of $\hat{\theta}^j$ is given in Table 7.9. We can see that the convergence is achieved within one or two iterations for all strike prices. ■

Example 7.7. Revisit Example 7.5 for the pricing of average price call options. When the strike price K gets larger, the plain Monte Carlo scheme or the basic cross-entropy method will start to fail. Design an iterative cross-entropy scheme to estimate the option price.

SOLUTION: Recall that the discounted option payoff is denoted by $h(X)$, where $X = (Z_1, \ldots, Z_m)$ and

$$Z_k = (W_{t_k} - W_{t_{k-1}})/\sqrt{t_k - t_{k-1}}, \quad 1 \le k \le m.$$

The joint density function of X and the family of alternative sampling densities are given by $f(x)$ and $\{f_\theta(x)\}$, respectively. Furthermore, for

$x = (x_1, \ldots, x_m)$ and $\theta = (\theta_1, \ldots, \theta_m)$, the likelihood ratio $\ell_\theta(x)$ is given by

$$\ell_\theta(x) = \frac{f(x)}{f_\theta(x)} = \exp\left\{ -\sum_{i=1}^m x_i\theta_i + \frac{1}{2}\sum_{i=1}^m \theta_i^2 \right\}.$$

In the general iterative cross-entropy scheme, the updating rule for the tilting parameter is given by Lemma 7.2:

$$\hat{\theta}^{j+1} = \frac{\sum_{k=1}^N h(Y_k)\ell_{\hat{\theta}^j}(Y_k)Y_k}{\sum_{k=1}^N h(Y_k)\ell_{\hat{\theta}^j}(Y_k)} = \frac{\sum_{k=1}^N h(Y_k)e^{-\langle\hat{\theta}^j, Y_k\rangle}Y_k}{\sum_{k=1}^N h(Y_k)e^{-\langle\hat{\theta}^j, Y_k\rangle}},$$

where Y_k's are iid samples from $f_{\hat{\theta}^j}$ or $N(\hat{\theta}^j, I_m)$.

A good choice of $\hat{\theta}^0$ should tilt the original distribution just enough so that $\{h(Y) > 0\}$ happens with moderate probability if Y is distributed as $N(\hat{\theta}^0, I_m)$, or equivalently, $\bar{S}$ exceeds K with nontrivial probability if $(Z_1, \ldots, Z_m)$ is treated as a $N(\hat{\theta}^0, I_m)$ random vector, in which case it follows from (7.8) that

$$E[\bar{S}] = \frac{1}{m}\sum_{k=1}^m E[S_{t_k}] = \frac{1}{m}\sum_{k=1}^m S_0 \exp\left\{ rt_k + \sigma\sum_{i=1}^k \sqrt{t_i - t_{i-1}}\hat{\theta}_i^0 \right\}.$$

A convenient choice is to set $\hat{\theta}^0 = (a, \ldots, a)$ and let $E[\bar{S}] = K$, that is, a satisfies

$$\frac{1}{m}\sum_{k=1}^m S_0 \exp\left\{ rt_k + a\sigma\sum_{i=1}^k \sqrt{t_i - t_{i-1}} \right\} = K. \tag{7.14}$$

This equation can be easily solved by the bisection method. Below is the pseudocode for the iterative cross-entropy scheme. The number of iterations is denoted by IT_NUM.

Pseudocode for average price call by the iterative cross-entropy method:

```
Solve for a from equation (7.14) by the bisection method
initialize θ̂^0 = (a, ..., a) and set the iteration counter j = 0
(*) generate N iid samples Y_1, ..., Y_N from N(θ̂^j, I_m)
set θ̂^{j+1} = Σ_{k=1}^N h(Y_k) e^{-⟨θ̂^j, Y_k⟩} Y_k / Σ_{k=1}^N h(Y_k) e^{-⟨θ̂^j, Y_k⟩}
set the iteration counter j = j + 1
if j = IT_NUM set θ̂ = θ̂^j and continue, otherwise go to step (*)
for i = 1, 2, ..., n
```

for $k = 1, 2, \ldots, m$

generate Y_k from $N(\hat{\theta}_k, 1)$

set $S_{t_k} = S_{t_{k-1}} \exp\left\{(r - \sigma^2/2)(t_k - t_{k-1}) + \sigma\sqrt{t_k - t_{k-1}}Y_k\right\}$

compute the discounted payoff multiplied by the likelihood ratio

$$H_i = e^{-rT}(\bar{S} - K)^+ \cdot \exp\left\{-\sum_{k=1}^{m} \hat{\theta}_k Y_k + \frac{1}{2}\sum_{k=1}^{m} \hat{\theta}_k^2\right\}$$

compute the estimate $\hat{v} = \dfrac{1}{n}\sum_{i=1}^{n} H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)}\left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2\right)}$.

The numerical results are reported in Table 7.10. The parameters are again given by

$$S_0 = 50, \;\; r = 0.05, \;\; \sigma = 0.2, \;\; T = 1, \;\; m = 12, \;\; t_i = \frac{i}{m}.$$

The sample size is always $n = 10000$ for importance sampling. The pilot sample size for the cross-entropy method is $N = 2000$ and the number of iterations is IT_NUM $= 5$.

Table 7.10: Iterative cross-entropy scheme for Asian option

	$K = 50$	$K = 60$	$K = 70$	$K = 80$	$K = 100$
Estimate	3.0749	0.3413	0.0166	4.983×10^{-4}	2.083×10^{-7}
S.E.	0.0151	0.0028	0.0002	6.538×10^{-6}	3.389×10^{-9}
R.E.	0.49%	0.81%	1.11%	1.31%	1.63%

The performance of the iterative cross-entropy scheme is indistinguishable from that of the basic cross-entropy scheme when K is not very large. When K reaches 80 and above, the basic cross-entropy scheme frequently gives an "NaN", while the iterative scheme consistently yields accurate estimates. ■

Example 7.8. Consider a multi-asset basket call option with maturity T and payoff

$$\left(c_1 S_T^{(1)} + \cdots + c_d S_T^{(d)} - K\right)^+.$$

Under the risk-neutral probability measure, the underlying stock prices are assumed to be geometric Brownian motions:

$$S_t^{(i)} = S_0^{(i)} \left\{ \left(r - \frac{1}{2}\sigma_i^2 \right) t + \sigma_i W_t^{(i)} \right\}, \quad i = 1, \ldots, d,$$

where $W = (W^{(1)}, \ldots, W^{(d)})$ is a d-dimensional Brownian motion with covariance matrix $\Sigma = [\Sigma_{ij}]$ such that $\Sigma_{ii} = 1$ for all i. Use the cross-entropy method to estimate the option price.

SOLUTION: Let A be a Cholesky factorization of Σ, that is, A is a lower triangular matrix with $AA' = \Sigma$. It follows that

$$\frac{1}{\sqrt{T}} W_T = \frac{1}{\sqrt{T}} (W_T^{(1)}, \ldots, W_T^{(d)}) = AX \tag{7.15}$$

for some d-dimensional standard normal random vector $X = (Z_1, \ldots, Z_d)'$. Clearly, the discounted option payoff can be written as $h(AX)$, where

$$h(y) = \left(\sum_{i=1}^{d} c_i S_0^{(i)} \left\{ -\frac{1}{2}\sigma_i^2 T + \sigma_i \sqrt{T} y_i \right\} - e^{-rT} K \right)^+$$

for $y = (y_1, \ldots, y_d)$. Except for the payoff function, the setup is very similar to Example 7.7. Let $f(x)$ be the joint density function of X or $N(0, I_d)$. The family of alternative sampling densities is given by $\{f_\theta(x) : \theta \in \mathbb{R}^d\}$, where f_θ is the joint density function for $N(\theta, I_d)$. The updating rule in the iterative cross-entropy scheme is again

$$\hat{\theta}^{j+1} = \frac{\sum_{k=1}^{N} h(Y_k) e^{-\langle \hat{\theta}^j, Y_k \rangle} Y_k}{\sum_{k=1}^{N} h(Y_k) e^{-\langle \hat{\theta}^j, Y_k \rangle}},$$

where $\hat{\theta}^j$ denotes the value of the tilting parameter at the end of the j-th iteration and Y_k's are iid samples from $N(\hat{\theta}^j, I_d)$.

To initialize the tilting parameter, observe that if X were $N(\theta, I_d)$, then it follows from (7.15) that $W_T/\sqrt{T}$ would have distribution $N(A\theta, \Sigma)$. In particular, $W_T^{(i)}/\sqrt{T}$ would be distributed as $N(\eta_i, 1)$, where $\eta = (\eta_1, \ldots, \eta_d)' = A\theta$, which implies that

$$E\left[\sum_{i=1}^{d} c_i S_T^{(i)} \right] = \sum_{i=1}^{d} c_i S_0^{(i)} \exp\left\{ rT + \sigma_i \sqrt{T} \eta_i \right\}.$$

Therefore, a convenient choice will be to choose $\eta = (a, \ldots, a)'$ so that the above expected value equals K, or

$$\sum_{i=1}^{d} c_i S_0^{(i)} \exp\left\{rT + \sigma_i \sqrt{T} a\right\} = K, \tag{7.16}$$

and then set $\hat{\theta}^0 = A^{-1}\eta$. The solution to equation (7.16) can be easily obtained by the bisection method. Below is the pseudocode.

Pseudocode for basket call by the iterative cross-entropy method:

solve for a from equation (7.16) by the bisection method
set $\eta = (a, \ldots, a)'$
initialize $\hat{\theta}^0 = A^{-1}\eta$ and set the iteration counter $j = 0$
(*) generate N iid samples $Y_1, \ldots, Y_N$ from $N(\hat{\theta}^j, I_d)$
set $\hat{\theta}^{j+1} = \sum_{k=1}^{N} h(AY_k) e^{-\langle \hat{\theta}^j, Y_k \rangle} Y_k / \sum_{k=1}^{N} h(AY_k) e^{-\langle \hat{\theta}^j, Y_k \rangle}$
set the iteration counter $j = j + 1$
if $j =$ IT_NUM set $\hat{\theta} = \hat{\theta}^j$ and continue, otherwise go to step (*)
for $i = 1, 2, \ldots, n$
generate $Y = (Y_1, \ldots, Y_d)$ from $N(\hat{\theta}, I_d)$
compute the discounted payoff multiplied by the likelihood ratio

$$H_i = h(AY) \cdot \exp\left\{ -\sum_{k=1}^{m} \hat{\theta}_k Y_k + \frac{1}{2} \sum_{k=1}^{m} \hat{\theta}_k^2 \right\}$$

compute the estimate $\hat{v} = \dfrac{1}{n} \sum_{i=1}^{n} H_i$

compute the standard error S.E. $= \sqrt{\dfrac{1}{n(n-1)} \left(\sum_{i=1}^{n} H_i^2 - n\hat{v}^2 \right)}$.

We consider a numerical example with $d = 4$ underlying assets. The simulation results are shown in Table 7.11 for

$$S_0^{(1)} = 45,\ S_0^{(2)} = S_0^{(3)} = 50,\ S_0^{(4)} = 55,\ r = 0.03,\ T = 0.5,\ c_1 = 0.4,$$
$$c_2 = c_3 = c_4 = 0.2,\ \sigma_1 = \sigma_2 = 0.1,\ \sigma_3 = \sigma_4 = 0.2,$$
$$\Sigma = \begin{bmatrix} 1.0 & 0.5 & -0.3 & 0.4 \\ 0.5 & 1.0 & 0.3 & 0.5 \\ -0.3 & 0.3 & 1.0 & 0.7 \\ 0.4 & 0.5 & 0.7 & 1.0 \end{bmatrix},\quad n = 10000.$$

We use IT_NUM $= 5$ iterations and the pilot sample size is $N = 2000$. ■

Table 7.11: Cross-entropy method for basket call option

Strike price	$K = 50$		$K = 60$		$K = 70$	
	CE	plain	CE	plain	CE	plain
Estimate	1.2945	1.3240	0.0081	0.0090	3.8384×10^{-6}	0
S.E.	0.0071	0.0213	0.0001	0.0015	5.9022×10^{-8}	0
R.E.	0.55%	1.61%	1.11%	16.28%	1.54%	NaN

7.2.3 Initialization in Rare Event Simulation

We have seen from the previous discussion that the initialization in the iterative cross-entropy schemes becomes less straightforward when the simulation involves events with very small probabilities. The usual choice of $\hat{\theta}^0 = \theta^0$, where θ^0 corresponds to the original distribution, can be problematic in this context. For example, in the estimation of a very small probability $\mathbb{P}(X \in A)$, most likely the indicator $h(X_i) = 1_{\{X_i \in A\}}$ will be zero for all pilot samples X_i, rendering the basic cross-entropy scheme (7.7) or the first iteration of the updating rule (7.12) meaningless. However, it has been demonstrated that one can often resolve this issue by judiciously choosing a good initial tilting parameter $\hat{\theta}^0$ so that the general iterative scheme will converge to a good final tilting parameter within a few iterations.

In this section, we discuss a different, systematic initialization technique that takes the guesswork out of $\hat{\theta}^0$. It is indeed a *separate* iterative scheme to obtain a good initial tilting parameter $\hat{\theta}^0$. To fix ideas, consider the problem of estimating

$$\mu = E\left[H(X;\alpha) \cdot 1_{\{F(X)\geq\alpha\}}\right]$$

for some functions H and F and some constant α. We assume that H is nonnegative, and is strictly positive with high probability given $F(X) \geq \alpha$. The parameter α is called the *rarity parameter* — the bigger α is, the smaller the probability of $\{F(X) \geq \alpha\}$ becomes. Most of the estimation problems in financial engineering are of this type. For example, to estimate the price of an option, usually one can let H be the discounted payoff of the option, F some function of the underlying asset price, and α a parameter that represents the source of rarity (e.g., a large strike price or a low barrier price). Another example is that in the problem of estimating loss probabilities, often one can let H be one, F the size of the loss, and α a given large loss threshold. For notational simplicity, in all the subsequent discussions we assume that X is a random variable or a random vector with density f. The extension to discrete distributions is straightforward. The family of alterna-

tive sampling densities is denoted by $\{f_\theta(x)\}$ and $f(x) = f_{\theta^0}(x)$ for some θ^0.

The main idea is as follows. Recall that equation (7.11) with $h(x)$ replaced by $H(x;\alpha)1_{\{F(x)\geq\alpha\}}$ yields an optimal tilting parameter for estimating μ, regardless of the value of the tilting parameter ν. Now consider a different, iterative approach. Let

$$\alpha_1 \leq \alpha_2 \cdots \leq \alpha_m = \alpha$$

be an increasing sequence. Then solving equation (7.11), with $h(x)$ replaced by $H(x;\alpha_1)1_{\{F(x)\geq\alpha_1\}}$ and $\nu = \theta^0$, yields an optimal tilting parameter, say θ^1, for estimating

$$E\left[H(X;\alpha_1)\cdot 1_{\{F(X)\geq\alpha_1\}}\right].$$

With this new tilting parameter θ^1, solving equation (7.11) once more, with $h(x)$ replaced by $H(x;\alpha_2)1_{\{F(x)\geq\alpha_2\}}$ and $\nu = \theta^1$, yields an optimal tilting parameter for estimating

$$E\left[H(X;\alpha_2)\cdot 1_{\{F(X)\geq\alpha_2\}}\right].$$

Repeating this process, we will obtain an optimal tilting parameter for estimating μ in the end.

The motivation is that the cross-entropy method replaces equations like (7.11) by their stochastic versions such as (7.12). If α_{j+1} is chosen so that $F(Y) \geq \alpha_{j+1}$ happens with some moderate probability, given that Y has density f_{θ^j}, then the updating rule (7.12) will lead to a good estimate for θ^{j+1}. In other words, to reach the level α, a collection of intermediate levels are introduced, and the tilting parameter is adjusted gradually to make the transition from one level to the next no longer a rare event.

From these considerations, a natural choice of the sequence $\{\alpha_1,\ldots,\alpha_m\}$ is to recursively define α_{j+1} as the $(1-\rho)$-quantile of the distribution of $F(Y)$, where Y has density f_{θ^j} and $0<\rho<1$ is some fraction. That is,

$$\alpha_{j+1} = \min\{x : \mathbb{P}(F(Y)\leq x)\geq 1-\rho\}.$$

Even though these quantiles are nearly impossible to quantify analytically, they can be easily estimated from the samples of $F(Y)$. Moreover, since samples of $F(Y)$ will be drawn anyways to update the tilting parameter, these *same samples* can be used to estimate the quantile α_{j+1}.

To be more precise, denote by $\bar{\theta}^j$ the tilting parameter at the end of the j-th iteration. Let $\bar{\theta}^0 = \theta^0$. In the $(j+1)$-th iteration, one generates iid samples $Y_1, \ldots, Y_N$ from $f_{\bar{\theta}^j}$. Let $V_k = F(Y_k)$ and consider the order statistics

$$V_{(1)} \leq \cdots \leq V_{(N)}.$$

An estimate for the $(1-\rho)$-quantile is simply $\bar{\alpha}_{j+1} = V_{(N_0)}$, where $N_0 = [N(1-\rho)]$, the integer part of $N(1-\rho)$. Using these same samples, one can solve for $\bar{\theta}^{j+1}$ from equation (7.12) with $\nu = \bar{\theta}^j$ and $h(x)$ replaced by $H(x;\bar{\alpha}_{j+1})1_{\{F(x)\geq\bar{\alpha}_{j+1}\}}$. That is, $\bar{\theta}^{j+1}$ is the solution to the equation

$$0 = \frac{1}{N}\sum_{k=1}^{N} H(Y_k;\bar{\alpha}_{j+1})1_{\{F(Y_k)\geq\bar{\alpha}_{j+1}\}}\ell_{\bar{\theta}^j}(Y_k)\frac{\partial}{\partial\theta}\log f_\theta(Y_k). \tag{7.17}$$

The iteration will end if $\bar{\alpha}_{j+1} \geq \alpha$. Note that in equation (7.17), only those samples Y_k that satisfy $F(Y_k) \geq \bar{\alpha}_{j+1}$ contribute to the updating of the tilting parameter. For this reason, these samples are called *elite samples*. By construction, the fraction of elite samples is approximately ρ. As before, equation (7.17) admits explicit solutions. In particular, in the case of normal distributions we have, analogous to Lemma 7.2,

$$\bar{\theta}^{j+1} = \frac{\sum_{k=1}^{N} H(Y_k;\bar{\alpha}_{j+1})1_{\{F(Y_k)\geq\bar{\alpha}_{j+1}\}}e^{-\langle\bar{\theta}^j,Y_k\rangle}Y_k}{\sum_{k=1}^{N} H(Y_k;\bar{\alpha}_{j+1})1_{\{F(Y_k)\geq\bar{\alpha}_{j+1}\}}e^{-\langle\bar{\theta}^j,Y_k\rangle}}. \tag{7.18}$$

Below is the pseudocode. We would like to repeat that the final tilting parameter from this iterative scheme will *not* be used for importance sampling, but rather as the *initial tilting parameter* $\hat{\theta}^0$ for the general iterative cross-entropy algorithm. The parameter ρ should not be too small because it determines the fraction of samples that will be used for updating the tilting parameter. Nor should it be too large because otherwise the growth of $\bar{\alpha}_j$ will be too slow. In practice, ρ is usually chosen between 5% and 10%.

Pseudocode for the initialization of the iterative cross-entropy scheme:

choose ρ between 5% and 10% and set $N_0 = [N(1-\rho)]$
set $\bar{\theta}^0 = \theta^0$ and the iteration counter $j = 0$
(*) generate N iid samples $Y_1, \ldots, Y_N$ from density $f_{\bar{\theta}^j}(x)$
set $V_k = F(Y_k)$ and the order statistics $V_{(1)} \leq \cdots \leq V_{(N)}$
set $\bar{\alpha}_{j+1} = V_{(N_0)}$ and $\bar{\theta}^{j+1}$ as the solution to (7.17)

set the iteration counter $j = j + 1$
if $\bar{\alpha}_j \geq \alpha$ set $\hat{\theta}^0 = \bar{\theta}^j$ and stop, otherwise go to step (*).

Example 7.9. Let us revisit Example 7.6. The call option price can be written as $v = E[H(X;\alpha)1_{\{F(X)\geq\alpha\}}]$, where X is a standard normal random variable and

$$\alpha = K, \quad F(X) = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}X\right\},$$

$$H(X;\alpha) = e^{-rT}\left(F(X) - \alpha\right)^+.$$

In the numerical experiment, we set $\rho = 10\%$ and $N = 2000$. Consequently, $N_0 = [2000 \times (1 - 0.10)] = 1800$. The other parameters are given by

$$S_0 = 50, \;\; r = 0.05, \;\; \sigma = 0.2, \;\; T = 1.$$

The solution to equation (7.17) is given by (7.18), which is used to update $\bar{\theta}^j$. The initial tilting parameter is simply $\bar{\theta}^0 = \theta^0 = 0$. The number of iterations, the final rarity parameter $\bar{\alpha}$, and the final tilting parameter $\bar{\theta}$ are reported in Table 7.12.

Table 7.12: Initialization of the cross-entropy method for call option

	$K = 50$	$K = 60$	$K = 80$	$K = 100$	$K = 120$
# of iterations	1	1	2	2	3
Final $\bar{\alpha}$	66.3546	66.9077	100.8898	101.8036	150.7960
Final $\bar{\theta}$	2.1133	2.1284	3.8432	3.8940	5.7056

Note that the final tilting parameter $\bar{\theta}$ is close to be optimal for estimating the price of a call option with strike price $\bar{\alpha}$. From this point of view, the results are quite consistent with Table 7.9. One can use the general iterative cross-entropy scheme, with the initial tilting parameter $\hat{\theta}^0$ set as the final tilting parameter $\bar{\theta}$, to produce a nearly optimal tilting parameter for importance sampling within a couple of iterations. ■

7.3 Applications to Risk Analysis

Popular risk measures such as value-at-risk or expected tail loss involve tail probabilities that are usually very small. Importance sampling is particularly powerful in the estimation of such quantities. To fix ideas, let L

represent the loss and x be a large threshold. Consider the tail probability $\mathbb{P}(L > x)$ and the expected tail loss $E[L|L > x]$. A generic importance sampling estimate for the tail probability takes the form

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{\bar{L}_i > x\}} w_i,$$

where $\{\bar{L}_1, \ldots, \bar{L}_n\}$ are iid loss samples from some alternative distribution and $\{w_1, \ldots, w_n\}$ are the corresponding likelihood ratios. As for the expected tail loss, observe that it is closely related to the tail probability as one can easily show (see Example 1.10) that

$$E[L|L > x] = \frac{E[L1_{\{L>x\}}]}{\mathbb{P}(L > x)}. \tag{7.19}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} \bar{L}_i 1_{\{\bar{L}_i > x\}} w_i$$

is an unbiased estimate for the numerator $E[L1_{\{L>x\}}]$, it is natural to estimate the expected tail loss by

$$\frac{\sum_{i=1}^{n} \bar{L}_i 1_{\{\bar{L}_i > x\}} w_i}{\sum_{i=1}^{n} 1_{\{L_i > x\}} w_i}.$$

The following lemma is concerned with the ratio of two sample means, which will be useful for constructing confidence intervals for the expected tail loss. For the sake of completeness, a proof is provided. If the reader is only interested in using it for confidence intervals, the proof can be safely skipped.

Lemma 7.3. *Assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid random vectors such that*

$$E[X_i] = \mu, \;\; E[Y_i] = \nu \neq 0.$$

Define

$$r = \frac{\mu}{\nu}, \quad R_n = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}, \quad \sigma_n^2 = \frac{\sum_{i=1}^{n} (X_i - R_n Y_i)^2}{(\sum_{i=1}^{n} Y_i)^2}.$$

Then as $n \to \infty$, R_n converges to r with probability one and the distribution of

$$\frac{R_n - r}{\sigma_n}$$

converges to the standard normal distribution.

PROOF. Without loss of generality, we assume $\nu > 0$; otherwise one can simply replace Y_i by $-Y_i$ and X_i by $-X_i$. Define $\bar{X}_n = \sum_{i=1}^n X_i/n$ and $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. By the strong law of large numbers, with probability one

$$\lim_{n\to\infty} R_n = \lim_{n\to\infty} \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\mu}{\nu} = r.$$

Moreover, it follows from straightforward computation, the strong law of large numbers, and the convergence of R_n to r, that with probability one

$$\begin{aligned}
\lim_{n\to\infty} (\sqrt{n}\sigma_n \bar{Y}_n)^2 &= \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - 2R_n \sum_{i=1}^n X_i Y_i + R_n^2 \sum_{i=1}^n Y_i^2 \right) \\
&= E[X^2] - 2rE[XY] + r^2 E[Y^2] \\
&= E[(X - rY)^2] \\
&= \mathrm{Var}(X - rY),
\end{aligned}$$

where (X, Y) has the same distribution as (X_i, Y_i). Observe that

$$\frac{R_n - r}{\sigma_n} = \frac{\bar{X}_n - r\bar{Y}_n}{\sigma_n \bar{Y}_n} = \frac{\sqrt{n}(\bar{X}_n - r\bar{Y}_n)}{\sqrt{\mathrm{Var}(X - rY)}} \cdot \frac{\sqrt{\mathrm{Var}(X - rY)}}{\sqrt{n}\sigma_n \bar{Y}_n}.$$

The distribution of the first term converges to the standard normal distribution by the central limit theorem, whereas the second term converges to one with probability one (here we have used the assumption $\nu > 0$). The lemma follows readily. ■

In order to illustrate the main idea, we focus on a popular class of portfolio credit risk models that uses *normal copula* to model the dependence structure among defaults. The special case with independent defaults has been investigated in Example 7.3 and will prove useful to the analysis of general credit risk models.

To be more concrete, assume that the model under consideration has m obligors. Let X_k be the default indicator for the k-th obligor and c_k the loss resulting from the default of the k-th obligor. Assuming for clarity that $c_k = 1$ for every k, the total loss L is

$$L = \sum_{k=1}^m c_k X_k = \sum_{k=1}^m X_k.$$

The default probability of the k-th obligor is assumed to be p_k. In other words, X_k is a Bernoulli random variable with parameter p_k, i.e.,

$$\mathbb{P}(X_k = 1) = p_k, \quad \mathbb{P}(X_k = 0) = 1 - p_k.$$

To model the dependence structure of the defaults, auxiliary random variables $\{Y_1, \ldots, Y_m\}$ are introduced so that

$$\begin{aligned} X_k &= 1_{\{Y_k > y_k\}}, \\ Y_k &= \rho_k Z + \sqrt{1-\rho_k^2}\varepsilon_k, \end{aligned}$$

where $y_k \in \mathbb{R}$ and $-1 < \rho_k < 1$ for each k, and $\{Z, \varepsilon_1, \ldots, \varepsilon_m\}$ are iid standard normal random variables. Since Y_k is a standard normal random variable itself and $\mathbb{P}(X_k = 1) = p_k$, it follows that

$$y_k = \Phi^{-1}(1-p_k) = -\Phi^{-1}(p_k).$$

Observe that X_k's are *dependent* because of the single common factor Z. In general, it is possible to introduce multiple common factors. However, as far as importance sampling is concerned, the difference is only notational.

We should first discuss the estimation of the tail probability $\mathbb{P}(L > x)$. It follows from the tower property that $\mathbb{P}(L > x) = E[h(Z)]$, where

$$h(z) = \mathbb{P}(L > x | Z = z).$$

Therefore, we can divide the estimation into two subproblems: (1) estimate $h(z)$ for $z \in \mathbb{R}$; (2) estimate $E[h(Z)]$ where Z is a standard normal random variable.

The key observation in estimating $h(z)$ is that $\{Y_1, \ldots, Y_m\}$ are *independent* conditional on Z. The conditional independence implies that $h(z)$ equals the loss probability of a credit risk model with independent defaults, where the default probability of the k-the obligor is (abusing notation)

$$p_k(z) = \mathbb{P}(Y_k > y_k | Z = z) = \Phi\left(-\frac{y_k - \rho_k z}{\sqrt{1-\rho_k^2}}\right). \tag{7.20}$$

Thus it follows from Example 7.3 that an efficient importance sampling scheme for estimating $h(z)$ is to use the alternative sampling distribution under which the default probability of the k-th obligor is given by

$$\bar{p}_k(z) = \frac{p_k(z)e^{\theta^*(z)}}{1 + p_k(z)(e^{\theta^*(z)} - 1)}, \tag{7.21}$$

where the tilting parameter $\theta^*(z) \in \mathbb{R}$ is defined in the following fashion. If

$$x > \sum_{k=1}^m p_k(z) = E\left[\sum_{k=1}^m X_k \middle| Z = z\right] = E[L | Z = z], \tag{7.22}$$

then $\theta^*(z)$ is defined to be the unique positive solution to an equation analogous to (7.3), namely,

$$x = \frac{d}{d\theta}\phi(\theta; z) = \sum_{k=1}^{m} \frac{p_k(z)}{p_k(z) + [1 - p_k(z)]e^{-\theta}}, \tag{7.23}$$

where

$$\phi(\theta; z) = \sum_{k=1}^{m} \log[1 + p_k(z)(e^{\theta} - 1)].$$

When (7.22) fails, $h(z)$ is not a small quantity and it suffices to use the plain Monte Carlo scheme, or equivalently, define $\theta^*(z) = 0$.

Importance sampling can also be applied to the estimation of $E[h(Z)]$. It is particularly important to do so when the correlations among $\{Y_k\}$ are strong, in which case a large loss is more likely due to a large value of Z, and thus altering the sampling distribution of Z becomes more imperative. Assume as usual that the family of alternative sampling densities is $\{f_\mu(x) : \mu \in \mathbb{R}\}$, where f_μ is the density of $N(\mu, 1)$. The original distribution of Z corresponds to $\mu = \mu^0 = 0$. A good tilting parameter μ can be determined by the general iterative cross-entropy method. The only trouble here is that h is not explicitly known and has be to be replaced by an approximation or an estimate, say $\hat{h}$, which leads to the following updating rule suggested by Lemma 7.2:

$$\hat{\mu}^{j+1} = \frac{\sum_{k=1}^{N} \hat{h}(\bar{Z}_k) e^{-\hat{\mu}^j \bar{Z}_k} \bar{Z}_k}{\sum_{k=1}^{N} \hat{h}(\bar{Z}_k) e^{-\hat{\mu}^j \bar{Z}_k}},$$

where $\bar{Z}_1, \ldots, \bar{Z}_N$ are iid pilot samples from $N(\hat{\mu}^j, 1)$. The initial tilting parameter is simply set to be $\hat{\mu}^0 = \mu^0 = 0$.

It remains to determine $\hat{h}$. For a given z, it is possible to take $\hat{h}(z)$ to be the importance sampling estimate of $h(z)$ as described previously. The drawback is that it will require nontrivial computational resource unless the sample size for simulating $h(z)$ is very small. We will adopt another approach which does not involve simulation. Following exactly the same steps that led to the upper bound (7.2), one can establish an upper bound for $h(z)$, that is, for any $\theta \geq 0$

$$h(z) = P(L > x | Z = z) \leq e^{-\theta x + \phi(\theta; z)}.$$

In particular,

$$\hat{h}(z) = \exp\{-\theta^*(z)x + \phi(\theta^*(z); z)\}$$

is an upper bound, as well as a good approximation, of h; see Exercise 7.8. Below is the pseudocode.

Pseudocode for the iterative cross-entropy method for $\hat{\mu}$:

initialize $\hat{\mu}^0 = 0$ and set the iteration counter $j = 0$

(*) for $i = 1, 2, \ldots, N$

generate sample $\bar{Z}_i$ from $N(\hat{\mu}^j, 1)$

calculate $P_k = p_k(\bar{Z}_i)$ for $k = 1, \ldots, m$ from (7.20)

set $l = P_1 + \cdots + P_m$

if $l < x$ solve $\theta = \theta^*(\bar{Z}_i)$ by the bisection method from (7.23);

otherwise set $\theta = \theta^*(\bar{Z}_i) = 0$

set $\hat{h}_i = \hat{h}(\bar{Z}_i) = \exp\{-\theta x + \phi(\theta; \bar{Z}_i)\}$

set $\hat{\mu}^{j+1} = \sum_{k=1}^N \hat{h}_k e^{-\hat{\mu}^j \bar{Z}_k} \bar{Z}_k / \sum_{k=1}^N \hat{h}_k e^{-\hat{\mu}^j \bar{Z}_k}$

set the iteration counter $j = j + 1$

if $j =$ IT_NUM set $\hat{\mu} = \hat{\mu}^j$, otherwise go to step (*)

Once the tilting parameter $\hat{\mu}$ is obtained, it is easy to construct an importance sampling estimate for the tail probability $\mathbb{P}(L > x)$ using the sample average of iid copies of

$$H = 1_{\{\bar{L} > x\}} \cdot \exp\left\{-\hat{\mu}\bar{Z} + \frac{1}{2}\hat{\mu}^2\right\} \cdot \prod_{k=1}^m \left(\frac{p_k(\bar{Z})}{\bar{p}_k(\bar{Z})}\right)^{\bar{X}_k} \left(\frac{1 - p_k(\bar{Z})}{1 - \bar{p}_k(\bar{Z})}\right)^{1-\bar{X}_k},$$

where $\bar{Z}$ has distribution $N(\hat{\mu}, 1)$, $p_k(\bar{Z})$ is given by (7.20), $\bar{p}_k(\bar{Z})$ is given by (7.21), $\bar{X}_k$'s are independent Bernoulli random variables conditional on $\bar{Z}$ with

$$\mathbb{P}(\bar{X}_k = 1 | \bar{Z}) = \bar{p}_k(\bar{Z}), \quad k = 1, \ldots, m,$$

and

$$\bar{L} = \sum_{k=1}^m \bar{X}_k.$$

In this algorithm, importance sampling is used for estimating both $h(z)$ and $E[h(Z)]$.

The above scheme can be modified slightly to yield an estimate for the expected tail loss. As we have mentioned previously, the sample average of iid copies of

$$\bar{L} \cdot H$$

is an unbiased estimate for $E[L1_{\{L>x\}}]$. Therefore, a natural estimate for $E[L|L > x]$ is simply

$$R_n = \frac{\sum_{i=1}^n \bar{L}_i H_i}{\sum_{i=1}^n H_i},$$

where H_i's are iid copies of H and $\bar{L}_i$'s are the corresponding losses. The standard error associated with this estimate, by Lemma 7.3, is

$$\text{S.E.} = \sqrt{\frac{\sum_{i=1}^n (\bar{L}_i H_i - R_n H_i)^2}{(\sum_{i=1}^n H_i)^2}}.$$

In other words, a $(1-\alpha)$ confidence interval of the expected tail loss is approximately

$$R_n \pm z_{\alpha/2} \cdot \text{S.E.}$$

where $z_{\alpha/2}$ is determined by $\Phi(-z_{\alpha/2}) = \alpha/2$. Below is the pseudocode for estimating the tail probability and the expected tail loss.

Pseudocode for the tail probability and expected tail loss:

use the iterative cross-entropy method to obtain a tilting parameter $\hat{\mu}$

for $i = 1, 2, \ldots, n$

generate $\bar{Z}$ from $N(\hat{\mu}, 1)$

calculate $P_k = p_k(\bar{Z})$ for $k = 1, \ldots, m$ from (7.20)

set $l = P_1 + \cdots + P_m$

if $l < x$ solve $\theta = \theta^*(\bar{Z})$ by the bisection method from (7.23);
otherwise set $\theta = \theta^*(\bar{Z}) = 0$

calculate $\bar{P}_k = \bar{p}_k(\bar{Z})$ for $k = 1, \ldots, m$ from (7.21)

generate $\bar{X}_k$ from Bernoulli with parameter $\bar{P}_k$ for $k = 1, \ldots, m$

set $\bar{L}_i = \bar{X}_1 + \cdots + \bar{X}_m$

if $\bar{L}_i > x$ set $H_i = e^{-\mu\bar{Z}+\mu^2/2} \cdot \prod_{k=1}^m \left(\frac{P_k}{\bar{P}_k}\right)^{\bar{X}_k} \left(\frac{1-P_k}{1-\bar{P}_k}\right)^{1-\bar{X}_k}$;
otherwise set $H_i = 0$

compute the estimate for tail probability $\hat{v} = \frac{1}{n}\sum_{i=1}^n H_i$

compute the standard error of $\hat{v} = \sqrt{\frac{1}{n(n-1)}\left(\sum_{i=1}^n H_i^2 - n\hat{v}^2\right)}$

compute the estimate for expected tail loss $\hat{r} = \frac{1}{n\hat{v}}\sum_{i=1}^n \bar{L}_i H_i$

compute the standard error of $\hat{r} = \frac{1}{n\hat{v}}\sqrt{\sum_{i=1}^n (\bar{L}_i H_i - \hat{r} H_i)^2}$.

If we fix the tilting parameter $\hat{\mu} = 0$ in the above importance sampling scheme, then it is equivalent to using importance sampling for the estimation of $h(z)$ but using plain Monte Carlo for the estimation of $E[h(Z)]$. We call such a scheme "partial importance sampling." We should include it in our numerical experiment to investigate the effect of the correlations among defaults on its performance. We expect that as the correlations grow stronger, the partial importance sampling will become less effective because a large loss is more likely due to a large value of the common factor Z.

Numerical Experiment. Consider a credit risk model where the parameters are given by

$$m = 1000, \;\; p_k = 0.01 \cdot [1 + e^{-k/m}], \;\; \rho_k = \rho, \;\; n = 10000.$$

For the iterative cross-entropy method, we use IT_NUM $= 5$ iterations and $N = 2000$ pilot samples for each iteration (in the actual simulation, the convergence is attained within one iteration, which means that the basic cross-entropy scheme would have sufficed). The parameter ρ captures the strength of the correlations among defaults. A larger value of $|\rho|$ means stronger correlations. When $\rho = 0$, the defaults are independent.

Table 7.13: Credit risk model: IS versus Plain MC versus Partial IS

$x = 26, \; \rho = 0.05$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	0.0180	1.59%	2.6180	1.06%	0.9785
Plain MC	0.0196	7.07%	2.7296	6.04%	0
Partial IS	0.0181	3.00%	2.5541	1.99%	0
$x = 30, \; \rho = 0.1$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	0.0159	1.53%	3.7131	1.30%	1.5803
Plain MC	0.0170	7.60%	3.3412	5.78%	0
Partial IS	0.0168	5.17%	3.5487	4.43%	0
$x = 60, \; \rho = 0.3$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	0.0156	1.49%	16.2838	1.25%	2.2538
Plain MC	0.0158	7.89%	15.9114	6.98%	0
Partial IS	0.0140	7.86%	16.2384	7.79%	0
$x = 145, \; \rho = 0.6$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	0.0157	1.52%	76.4701	1.31%	2.4254
Plain MC	0.0161	7.82%	69.7888	7.79%	0
Partial IS	0.0162	7.68%	85.2236	8.14%	0

Table 7.14: Credit risk model: IS versus Plain MC versus Partial IS

$x = 40,\ \rho = 0.1$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	5.8800×10^{-4}	1.76%	3.3147	1.30%	2.5216
Plain MC	6.0000×10^{-4}	40.81%	2.8333	17.48%	0
Partial IS	7.8932×10^{-4}	21.09%	4.3622	19.49%	0
$x = 50,\ \rho = 0.1$	TProb	R.E.	ETL $- x$	R.E.	$\hat{\mu}$
IS	1.4478×10^{-5}	1.96%	3.0765	1.33%	3.3562
Plain MC	0	NaN	NaN	NaN	0
Partial IS	8.3848×10^{-6}	52.04%	2.7119	37.39%	0

The numerical results are reported in Tables 7.13 and 7.14. The entry "TProb" means the estimate for the tail probability $\mathbb{P}(L > x)$, while "ETL $-$ x" means the estimate for $E[L|L > x] - x$.

In Table 7.13, we push up the threshold x as ρ increases in order for the tail probability to remain roughly the same magnitude. The efficiency of partial importance sampling clearly deteriorates as the correlations grow stronger. Note that the performance of partial importance sampling is indistinguishable from that of the plain Monte Carlo scheme when ρ is large, which suggests that using importance sampling for estimating $h(z)$ alone does not achieve much variance reduction overall. This justifies our intuition that a large loss is most likely due to a large value of the common factor Z. In Table 7.14, we fix ρ and let x vary. It is clear that only the importance sampling scheme consistently yields accurate estimates even when the tail probability becomes really small. ■

Exercises

Pen-and-Paper Problems

7.1 Write down the importance sampling estimates for the following quantities:

(a) $\mathbb{P}(X \geq m\alpha)$, where X is a binomial random variable with parameters (m, p) and $\alpha \in (0, 1)$ is a given constant. The alternative sampling distribution is binomial with parameters $(m, \bar{p})$.

(b) $E[h(X_1, \ldots, X_m)]$, where $(X_1, \ldots, X_m)$ is an m-dimensional standard normal random vector. The alternative sampling distribution is assumed to be $N(\theta, I_m)$ for some $\theta = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$.

(c) $E[h(X_1, \ldots, X_m)]$, where $(X_1, \ldots, X_m)$ is a jointly normal random vector with mean μ and nonsingular covariance matrix Σ. *Hint:* Use (b) and Cholesky's factorization.

(d) $P(X_1 + \cdots + X_m \geq a)$, where $(X_1, \ldots, X_m)$ is a random vector with joint density $f(x_1, \ldots, x_m)$. The alternative sampling density is assumed to be $g(x_1, \ldots, x_m)$.

7.2 Let X be a standard normal random variable. Assume that we wish to use importance sampling to estimate, for a given positive constant a,

$$E\left(e^{a\sqrt{X}} 1_{\{X \geq 0\}}\right).$$

The alternative sampling distribution is $N(\theta, 1)$ for some $\theta \in \mathbb{R}$. Use the mode matching method to determine the best θ.

7.3 The purpose of this exercise is to justify the method of mode matching for normal distributions. Suppose that one is interested in estimating the tail probability

$$\mathbb{P}(Z \geq b)$$

for a standard normal random variable Z and a large threshold b.

(a) Write down the importance sampling estimate with the alternative sampling distribution $N(\theta, 1)$.

(b) Compute the variance of this estimate.

(c) Assume that the variance is minimized at $\theta = \theta^*$. Write down the equation that θ^* satisfies.

(d) Solve the equation to obtain θ^*, using the extremely accurate approximation [4] that for large x

$$\Phi(-x) \approx \frac{\phi(x)}{x}, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

7.4 A Cautionary Example. The purpose of this exercise is to show that the method of mode matching will not work universally. Indeed, it can be much worse than the plain Monte Carlo. The setup is the same as Exercise 7.3 except that the quantity of interest is

$$\mathbb{P}(Z \geq b \text{ or } Z \leq -2b)$$

The alternative sampling distribution is assumed to be $N(\theta, 1)$.

(a) Use the method of mode matching to determine θ.

(b) Write down the importance sampling estimate H.

(c) Compute $E[H^2]$ explicitly and use the same approximation in Exercise 7.3 to argue that

$$\lim_{b\to\infty} E[H^2] = \infty.$$

The method of mode matching should be used with caution. The same can be said to the cross-entropy method. Indeed, when the problem exhibits certain kind of nonconvexity, looking for an alternative sampling distribution within the exponential tilt family may not be sufficient [9]. This counterexample was first constructed in [13]. See also Exercise 7.L.

7.5 Determine the exponential tilt family for the random variable X.

(a) X is normal with mean 0 and variance σ^2.

(b) X is Bernoulli with parameter p.

(c) X is Poisson with parameter λ.

(d) X is exponentially distributed with rate λ.

7.6 The exponential tilt family for a general random vector can be defined in the same fashion. Let $X = (X_1, \ldots, X_m)$ be an m-dimensional random vector.

(a) *X is continuous:* Let f denote the density of X. The exponential tilt family consists of density functions of the form

$$f_\theta(x) = \frac{1}{E[e^{\langle \theta, X\rangle}]} e^{\langle \theta, x\rangle} f(x)$$

for some $\theta \in \mathbb{R}^m$.

(b) *X is discrete:* Let p denote the probability mass function of X, that is, $\mathbb{P}(X = x) = p(x)$. The exponential tilt family consists of probability mass functions of the form

$$p_\theta(x) = \frac{1}{E[e^{\langle \theta, X\rangle}]} e^{\langle \theta, x\rangle} p(x)$$

for some $\theta \in \mathbb{R}^m$.

Determine the exponential tilt family for X when (i) X is $N(0, I_m)$; (ii) the components $X_1, \ldots, X_m$ are independent random variables.

7.7 Suppose that $X = Y_1 + \cdots + Y_m$ where Y_k's are independent discrete random variables. Fix an arbitrary θ. Let $\bar{Y}_1, \ldots, \bar{Y}_m$ be independent with distribution

$$\mathbb{P}(\bar{Y}_k = y) = \frac{1}{E[e^{\theta Y_k}]} e^{\theta y} \mathbb{P}(Y_k = y)$$

for each k. That is, $\bar{Y}_k$ has the exponential tilt distribution of Y_k with parameter θ. Show that $\bar{X} = \bar{Y}_1 + \cdots + \bar{Y}_m$ has the exponential tilt distribution of X with parameter θ. That is,

$$\mathbb{P}(\bar{X} = x) = \frac{1}{E[e^{\theta X}]} e^{\theta x} \mathbb{P}(X = x).$$

In other words, if one wishes to use the exponential tilt distribution of X with parameter θ for importance sampling, it suffices to use the exponential tilt distribution on each component Y_k with the *same* θ. This result justifies the alternative sampling distribution used in Example 7.3. Not surprisingly, a similar result holds when Y_k's are continuous random variables.

7.8 Consider the credit risk model in Example 7.3. Show that equation (7.3) is equivalent to

$$x = E[\bar{L}] = \sum_{k=1}^{m} \bar{p}_k.$$

Show that in this case, the standard deviation of $\bar{L}$ is bounded from above by $\sqrt{x}$, which is much smaller compared to x when x is large. Use this to explain why (7.2) gives a good upper bound for $E[H^2]$ when θ is taken to be the solution to equation (7.3).

7.9 Consider a slight generalization of the credit risk model in Example 7.3, where c_k's are not always one. Show that $\bar{L}$ has the exponential tilt distribution of L with parameter θ, where

$$\bar{L} = \sum_{k=1}^{m} c_k Y_k$$

and Y_k's are independent Bernoulli random variables with parameter

$$\bar{p}_k = \mathbb{P}(Y_k = 1) = \frac{1}{E[e^{\theta c_k X_k}]} e^{\theta c_k} p_k = \frac{p_k e^{\theta c_k}}{1 + p_k(e^{\theta c_k} - 1)}.$$

Hint: Apply Exercise 7.7 with $Y_k = c_k X_k$.

7.10 Let $f(x)$ and $g(x)$ be two density functions. Define the Kullback–Leibler cross entropy or relative entropy by

$$R(g\|f) = \int_{\mathbb{R}} \log \frac{g(x)}{f(x)} \cdot g(x)\, dx.$$

Show that $R(g\|f) \geq 0$ and $R(g\|f) = 0$ if and only if $g(x) = f(x)$. Similarly, if $p(x_i)$ and $q(x_i)$ are two probability mass functions, then the relative entropy is defined as

$$R(q\|p) = \sum_i \log \frac{q(x_i)}{p(x_i)} \cdot q(x_i)$$

Again, show that $R(q\|p) \geq 0$ and $R(q\|p) = 0$ if and only if $q = p$. *Hint:* Let X be a random variable with density f. Define $Y = g(X)/f(X)$. Argue that the function $h(x) = x \log x$ is convex and apply the Jensen's inequality (Exercise 1.23) on $R(g\|f) = E[h(Y)]$. The discrete case can be proved similarly.

7.11 Assume that X is a random variable with density f, and $\{f_\theta\}$ is the exponential tilt family defined in Section 7.1.3. That is,

$$f_\theta(x) = \frac{1}{E[e^{\theta X}]} e^{\theta x} f(x).$$

Let $H(\theta) = \log E[e^{\theta X}]$. Show that the solution $\hat{\theta}$ to equation (7.7) is determined by

$$H'(\hat{\theta}) = \frac{\sum_{k=1}^N h(X_k) X_k}{\sum_{k=1}^N h(X_k)}.$$

Similarly, the solution $\hat{\theta}^{j+1}$ to equation (7.12) is determined by

$$H'(\hat{\theta}^{j+1}) = \frac{\sum_{k=1}^N h(Y_k) \ell_{\hat{\theta}^j}(Y_k) Y_k}{\sum_{k=1}^N h(Y_k) \ell_{\hat{\theta}^j}(Y_k)}.$$

Argue that the exactly same formulas hold when X is a discrete random variable.

7.12 Consider the problem of estimating $E[h(X)]$, where X is a binomial random variable with parameters (m, p) and h is nonnegative. Assuming that the alternative sampling distribution is binomial with parameters (m, θ), write down the solution to the basic cross-entropy scheme (7.7) and to the general iterative scheme (7.12).

7.13 Consider the credit risk model in Example 7.3. Assuming that the iterative cross-entropy scheme is to be used in order to obtain a good tilting parameter θ, show that the updating rule, or the solution to equation (7.12), is determined by

$$\phi'(\hat{\theta}^{j+1}) = \frac{\sum_{i=1}^N \bar{L}_i e^{-\hat{\theta}^j \bar{L}_i} 1_{\{\bar{L}_i > x\}}}{\sum_{i=1}^N e^{-\hat{\theta}^j \bar{L}_i} 1_{\{\bar{L}_i > x\}}},$$

where $\bar{L}_1, \ldots, \bar{L}_N$ are iid copies of $\bar{L} = Y_1 + \cdots + Y_m$ and Y_k's are independent Bernoulli random variables with parameter

$$\bar{p}_k = \mathbb{P}(Y_k = 1) = \frac{p_k e^{\hat{\theta}^j}}{1 + p_k(e^{\hat{\theta}^j} - 1)}.$$

7.14 Let $Z = (Z_1, \ldots, Z_m)$ be an m-dimensional standard normal random vector. Denote the exponential tilt family of Z by $\{f_\theta(x) : \theta \in \mathbb{R}^m\}$, where f_θ is the density function of $N(\theta, I_m)$ for $\theta = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$. In some applications, one can restrict the alternative sampling distribution to a subset of the exponential tilt family where θ takes the form $\theta = (\mu, \ldots, \mu)$ for some $\mu \in \mathbb{R}$. Under this assumption, write down the solution to the basic cross-entropy scheme (7.7) and to the general iterative scheme (7.12).

MATLAB® Problems

7.A Compare importance sampling with the plain Monte Carlo scheme in the estimation of $\mathbb{P}(X \geq b)$.

(a) X is binomial with parameters (m, p) and $b = m\alpha$ with $p < \alpha < 1$. The alternative sampling distribution is binomial with parameters (m, α). Report your results for $p = 0.5, \alpha = 0.9$, and $m = 10, 100, 1000$, respectively. Sample size is $n = 10000$.

(b) X is exponential with rate λ and $b = a/\lambda$ with $a > 1$. The alternative sampling distribution is exponential with rate $1/b$. Report your results for $\lambda = 1$ and $a = 2, 5, 10$, respectively. Sample size is $n = 10000$.

7.B Let X, Y, Z be iid standard normal random variables. Use the basic cross-entropy scheme to estimate

$$\mathbb{P}(\min\{X + Y, Y + 2Z + 1\} \geq b)$$

The sample size is $n = 10000$. The pilot sample size is $N = 2000$.

(a) Compare with plain Monte Carlo for $b = 1, 2, 3$, respectively.

(b) Let $b = 5$. Does the basic cross-entropy method still work? Why?

In Exercises 7.C — 7.G, the underlying asset price is assumed to be a geometric Brownian motion under the risk-neutral probability measure:

$$S_t = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\},$$

where W is a standard Brownian motion and r is the risk-free interest rate.

7.C Consider the binary option in Example 7.1. Write a function to estimate the option price by the iterative cross-entropy method. The function should have input parameters S_0, r, σ, T, K, sample size n, pilot sample size N, and the number of iterations IT_NUM. Explain your choice of the initial tilting parameter. Report your estimate, relative error, and the final tilting parameter for

$$S_0 = 50,\ r = 0.01,\ \sigma = 0.1,\ T = 1,\ n = 10000,\ N = 2000,\ \text{IT\_NUM} = 5,$$

and $K = 50, 60, 70, 80$, respectively. How many iterations are actually needed for the tilting parameters to converge?

7.D Write a function to estimate the price of a put option with maturity T and strike price K, via importance sampling. Use the mode matching method to find a good tilting parameter. The function should have input parameters S_0, r, σ, K, T, n. Report your estimate, standard error, and the tilting parameter for

$$r = 0.1,\ \sigma = 0.2,\ T = 1,\ K = 30,\ n = 10000,$$

and $S_0 = 30, 50, 70$, respectively.

7.E Repeat Exercise 7.D, but use the iterative cross-entropy method to find a good tilting parameter. The function should have two additional input parameters, namely, the pilot sample size N and the number of iterations IT_NUM. Explain your choice of the initial tilting parameter. Set $N = 2000$ and IT_NUM $= 5$ in your numerical simulation.

7.F Consider a lookback call option with fixed strike price K and maturity T, whose payoff is

$$\left(\max_{i=1,\dots,m} S_{t_i} - K \right)^+$$

for some given dates $0 < t_1 < \cdots < t_m = T$. The plain Monte Carlo scheme or the basic cross-entropy method fails when the strike price K is large. Write a function to estimate the option price by the iterative cross-entropy method. The function should have input parameters $S_0, r, \sigma, K, T, m, (t_1, \dots, t_m)$, sample size n, pilot sample size N, and the number of iterations IT_NUM. Explain your choice of the initial tilting parameter. Report your estimate, standard error, and the final tilting parameter for

$$S_0 = 50,\ r = 0.02,\ \sigma = 0.2,\ T = 0.5,\ m = 12,\ t_i = iT/m,$$

$$n = 10000,\ N = 2000,\ \text{IT\_NUM} = 5,$$

and $K = 60, 80, 100$, respectively.

7.G Consider a binary down-and-in barrier option with maturity T and barrier b, whose payoff is

$$1_{\{\min(S_{t_1},\dots,S_{t_m}) \le b\}}$$

for a given set of dates $0 < t_1 < \cdots < t_m = T$. The plain Monte Carlo or the basic cross-entropy scheme fails when the barrier b is much lower than S_0. Note that the payoff is a function of $Z = (Z_1, \ldots, Z_m)$, where

$$Z_k = (W_{t_k} - W_{t_{k-1}})/\sqrt{t_k - t_{k-1}}, \quad 1 \le k \le m,$$

are iid standard normal random variables. Design an iterative cross-entropy scheme, with the alternative sampling distribution restricted to the subset of the exponential tilt family of Z given by Exercise 7.14. Explain your choice of the initial tilting parameter. Report your estimate, standard error, and the final tilting parameter for

$$S_0 = 50,\ r = 0.02,\ \sigma = 0.2,\ T = 1,\ m = 100,\ t_i = iT/m,$$

$$n = 10000,\ N = 2000,\ \text{IT\_NUM} = 5,$$

and $b = 40, 30, 20$, respectively.

7.H The setup is similar to Exercise 6.F. Let X and Y be the prices of two underlying stocks. Assume that they are both geometric Brownian motions under the risk-neutral probability measure and

$$\begin{aligned} X_T &= X_0 \exp\left\{\left(r - \frac{1}{2}\sigma_1^2\right) T + \sigma_1\sqrt{T}Z_1\right\}, \\ Y_T &= Y_0 \exp\left\{\left(r - \frac{1}{2}\sigma_2^2\right) T + \sigma_2\sqrt{T}Z_2\right\}, \end{aligned}$$

where Z_1 and Z_2 are two independent standard normal random variables. We wish to estimate the price of a basket call option with maturity T and payoff

$$h(Z_1, Z_2) = (c_1 X_T + c_2 Y_T - K)^+$$

by importance sampling. The alternative sampling distribution is assumed to be $N(\theta, I_2)$ for some $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$.

(a) The mode matching method sets the tilting parameter (θ_1, θ_2) as the maximizer of $h(z_1, z_2) f(z_1, z_2)$, where f is the joint density function of $Z = (Z_1, Z_2)$. It is equivalent to maximizing

$$h(z_1, z_2) \exp\left\{-\frac{1}{2}(z_1^2 + z_2^2)\right\}.$$

Find the maximizer (z_1^*, z_2^*) by the bisection method. *Hint:* Taking partial derivatives over z_1 and z_2 leads to two equations with two unknowns. Reduce the system of equations to a single equation with one unknown.

(b) Write a MATLAB function to estimate the call option price by importance sampling. The function should have input parameters r, σ_1, σ_2, X_0, Y_0, T, K, c_1, c_2, and sample size n. Report your estimate, standard error, and the tilting parameter (θ_1, θ_2) for

$$r = 0.1,\ \sigma_1 = 0.2,\ \sigma_2 = 0.3,\ X_0 = Y_0 = 50,\ T = 1,\ c_1 = c_2 = 0.5,$$

and strike $K = 80, 100, 120$, respectively. Sample size is $n = 10000$.

7.I Consider a multi-asset basket put option with maturity T and strike price K. The option payoff is

$$\left(K - \sum_{i=1}^{d} c_i S_T^{(i)} \right)^+ .$$

Assume that under the risk-neutral probability measure, the underlying stock prices are modeled by geometric Brownian motions:

$$S_t^{(i)} = S_0^{(i)} \left\{ \left(r - \frac{1}{2}\sigma_i^2 \right) t + \sigma_i W_t^{(i)} \right\}, \quad i = 1, \ldots, d,$$

where $W = (W^{(1)}, \ldots, W^{(d)})$ is a d-dimensional Brownian motion with covariance matrix $\Sigma = [\Sigma_{ij}]$ such that $\Sigma_{ii} = 1$ for all i. Write a function to estimate the option price by the iterative cross-entropy method. The function should have input parameters $d, r, K, T, \Sigma, S_0^{(i)}, \sigma_i, c_i$, sample size n, pilot sample size N, and the number of iterations IT_NUM. Explain your choice of the initial tilting parameter. Report your estimate, standard error, and the final tilting parameter for

$$d = 3,\ S_0^{(1)} = 50,\ S_0^{(2)} = 45,\ S_0^{(3)} = 55,\ r = 0.03,\ T = 1,$$

$$\sigma_1 = 0.2,\ \sigma_2 = \sigma_3 = 0.3,\ \Sigma = \begin{bmatrix} 1 & 0.2 & -0.2 \\ 0.2 & 1 & 0.5 \\ -0.2 & 0.5 & 1 \end{bmatrix},$$

$$c_1 = 0.4,\ c_2 = c_3 = 0.3,\ n = 10000,\ N = 2000,\ \text{IT\_NUM} = 5,$$

and $K = 50, 40, 30$, respectively.

7.J Consider a slight generalization of the simple credit risk model in Example 7.3, where the losses c_k are not always one. Write a function to estimate the tail probability $\mathbb{P}(L > x)$ and the expected tail loss $E[L|L > x]$ by importance sampling. The alternative sampling distribution is given by Exercise 7.9. Use the mode matching method to determine a good tilting parameter θ. The function should have input parameters $m, (p_1, \ldots, p_m), (c_1, \ldots, c_m), x$, and sample size n. Report your estimate, standard error, and the tilting parameter for

$$m = 1000,\ p_k = 0.01 \cdot [1 + e^{-k/m}],\ c_k = 1 + e^{k/m},\ n = 10000,$$

and $x = 50, 70, 90$, respectively.

7.K Consider the credit risk model in Example 7.3. Use the initialization technique described in Section 7.2.3 to find an initial tilting parameter for the cross-entropy scheme. Set $\rho = 10\%$, $\bar{\theta}^0 = 0$, and pilot sample size $N = 2000$. Report the final tilting parameter $\bar{\theta}$ and the number of iterations performed. Let $\hat{\theta}^0 = \bar{\theta}$ be the initial tilting parameter for the iterative cross-entropy scheme, and report your final tilting parameter $\hat{\theta}$ for importance sampling, with pilot sample size $N = 2000$ and the number of iterations IT_NUM $= 5$. *Hint:* Apply the results in Exercise 7.13.

7.L This exercise is to reinforce numerically what Exercise 7.4 has demonstrated. Consider a very simple problem of estimating

$$\mathbb{P}(Z \in A),$$

where Z is a standard normal random variable and $A = (-\infty, a] \cup [b, \infty)$ is a nonconvex set. Compare the following algorithms:

i. The plain Monte Carlo.

ii. Importance sampling using the mode matching method.

iii. Importance sampling using the iterative cross-entropy method with the initial tilting parameter $\hat{\theta}^0 = a$.

iv. Importance sampling using the iterative cross-entropy method with the initial tilting parameter $\hat{\theta}^0 = b$.

Perform ten sets of simulations for each scheme given $a = -2.2$ and $b = 2$, with sample size $n = 10000$, pilot sample size $N = 2000$, and the number of iterations IT_NUM $= 5$. Report your estimates, standard errors, and 95% confidence intervals. Describe your findings. From the simulation, you will see clearly that the standard error of an importance sampling estimate is not always trustworthy when the problem involves nonconvex sets, or more generally, nonconvex functions. Therefore, when one wishes to use importance sampling to estimate the price of an option with nonconvex payoff (e.g., an outperformance option), exercise caution because selecting an alternative sampling density from the exponential family may lead to erroneous estimates. For a general approach to building efficient importance sampling schemes, see [9, 10].

This page intentionally left blank

Chapter 8

Stochastic Calculus

Stochastic calculus is an essential mathematical tool in modern continuous time finance [1, 7, 16, 19, 23]. For example, stochastic integrals naturally arise in the analysis of self-financing portfolios and Itô formula connects option prices with solutions to partial differential equations. Recent years have also seen the development of many financial models where the underlying stock prices are described by diffusion processes or solutions to stochastic differential equations. They are generalizations of the classical Black–Scholes model, which uses geometric Brownian motions to model stock prices.

Stochastic calculus differs fundamentally from classical calculus. While classical calculus often involves integrals against functions that are nice and smooth, stochastic calculus has to deal with integrals with respect to random processes whose sample paths are rugged and nowhere differentiable. As a consequence, Itô formula, instead of the fundamental theorem of classical calculus, becomes the corner stone of stochastic calculus.

This chapter offers a quick introduction to stochastic calculus, including stochastic integrals, Itô formula, and stochastic differential equations. Since a mathematically rigorous treatment is beyond the scope of this book, we aim for an informal understanding of stochastic calculus. This means that, for example, some technical conditions such as measurability and integrability will be omitted in the statement of theorems (they are satisfied in nearly all practical applications), the modes of convergence will not be clearly specified, solutions to stochastic differential equations will not be distinguished as in the strong or weak sense, and so on. For the more mathematically inclined reader, rigorous treatment of stochastic calculus can be found in many advanced graduate level textbooks such as [18, 25, 26, 27].