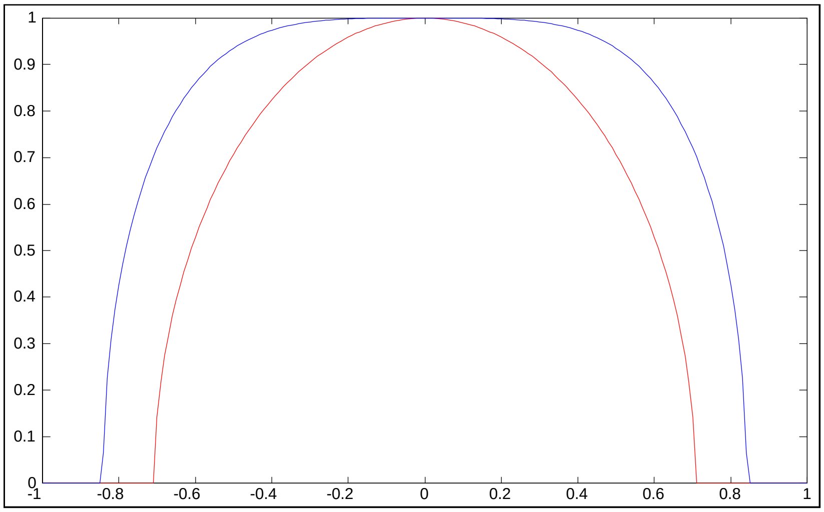


Figure 2.12:

16. Evaluate the following integral by simulation:

$$\int_0^1 \int_0^1 e^{(x+y)^2} dxdy.$$

(Hint: Note that if U_1, U_2 are independent Uniform[0,1] random variables, $E[g(U_1, U_2)] = \int_0^1 \int_0^1 g(x,y)dxdy$ for any function g).

17. Find the covariance $cov(U, e^U)$ by simulation where U is uniform[0,1] and compare the simulated value to the true value.

18. Find by simulation the area of the region $\{(x,y); -1 < x < 1,\ \ y > 0, \sqrt{1-2x^2} < y < \sqrt{1-2x^4}\}$. The boundaries of the region are graphed below.

19. For independent uniform random numbers $U_1, U_{2,\ldots.}$ define the random variable $N = \min imum\{n; \sum_{i=1}^n U_i > 1\}$.

    Estimate $E(N)$ by simulation. Repeat for larger and larger numbers of simulations. What do you think is the value of $E(N)$?

20. Give a precise algorithm for generating observations from a distribution with probability density function

$$f(x) = \frac{(x-1)^3}{4}$$

for $1 \leq x \leq 3$. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. .

21. Give a precise algorithm for generating observations from a distribution with probability density function $\frac{(x-20)}{200}$ for $20 \leq x \leq 40$. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. .

22. Give a precise algorithm for generating observations from a distribution with a density function of the form $f(x) = cx^3 e^{-x/2}$ for $x > 0$ and appropriate constant c. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. .

23. Give a precise algorithm for generating observations from a discrete distribution with $P[X = j] = (2/3)(1/3)^j$; $j = 0, 1, ...$Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. .

24. Give a precise algorithm for generating observations from a distribution with probability density function $f(x) = e^{-x}, 0 \leq x < \infty$. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. Compute as well the sample variance and compare withe the sample mean. How large would the simulation need to be if we wanted to estimate the mean within 0.01 with a 95% confidence interval?

25. Give a precise algorithm for generating observations from a distribution which has probability density function $f(x) = x^3, 0 < x < \sqrt{2}$. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. Determine the standard error of the sample mean. How large would the simulation need to be if we wanted to estimate the mean within 0.01 with a 95% confidence interval?

26. Give a precise algorithm for generating observations from a discrete distribution with probability function

x=	0	1	2	3	4	5
P[X=x]=	0.1	0.2	0.25	0.3	0.1	0.05

Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. Compare the sample mean and variance with their theoretical values. How large would the simulation need to be if we wanted to estimate the mean within 0.01 with a 95% confidence interval?

27. Give an algorithm for generating observations from a distribution which has cumulative distribution function function $F(x) = \frac{x+x^3+x^5}{3}$, $0 < x < 1$. Record the time necessary to generate the sample mean of 5,000 random variables with this distribution. (Hint: Suppose we generate X_1 with c.d.f. $F_1(x)$ and X_2 with c.d.f. $F_2(x)$, X_3 with c.d.f. $F_3(x)$ We

then generate $J = 1$,2, or 3 such that $\mathrm{P}[J = j] = p_j$ and output the value X_J. What is the c.d.f. of the random variable output?)

28. Consider independent random variables X_i with c.d.f.

$$\begin{aligned} F_i(x) &= x^2, \qquad i = 1 \\ &= \frac{e^x - 1}{e-1}, \quad i = 2 \\ &= xe^{x-1}, \qquad i = 3 \end{aligned}$$

for $0 < x < 1$. Explain how to obtain random variables with c.d.f.$G(x) = \Pi_{i=1}^3 F_i(x)$ and $G(X) = 1 - \Pi_{i=1}^3 (1 - F_i(x))$.

(Hint: consider the c.d.f. of the minimum and maximum).

29. Suppose we wish to estimate a random variable X having c.d.f. $F(x)$ using the inverse transform theorem, but the exact cumulative distribution function is not available. We do, however, have an unbiased estimator $\widehat{F}(x)$ of $F(x)$ so that $0 \leq \widehat{F}(x) \leq 1$ and $E\,\widehat{F}(x) = F(x)$ for all x. Show that provided the uniform variate U is independent of $\widehat{F}(x)$, the random variable $X = \widehat{F}^{-1}(U)$ has c.d.f. $F(x)$.

30. Give an algorithm for generating a random variable with probability density function

$$f(x) = 30(x^2 - 2x^3 + x^4), \quad 0 < x < 1$$

Discuss the efficiency of your approach.

31. The interarrival times between consecutive buses at a certain bus stop are independent uniform$[0, 1]$ random variables starting at clock time $t = 0$. You arrive at the bus stop at time $t = 1$. Determine by simulation the expected time that you will have to wait for the next bus. Is it more than $1/2$? Explain.

32. What is the probability density function of $X = a(1 - \sqrt{U})$ where $U \sim U[0, 1]$?

33. Develop an algorithm for generating variates from the density:

$$f(x) \;=\; 2/\sqrt{\pi}e^{2a - x^2 - a^2/x^2}, \quad x > 0$$

34. Develop an algorithm for generating variates from the density:

$$f(x) \;=\; \frac{2}{e^{\pi x} + e^{-\pi x}}, \qquad -\infty < x < \infty$$

35. Explain how the following algorithm works and what distribution is generated.

(a) . LET $I = 0$

(b) Generate $U \sim U[0,1]$ and set $T = U$.

(c) Generate U^* . IF $U \leq U^*$ return $X = I + T$.

(d) Generate U . If $U \leq U^*$ go to c.

(e) $I = I + 1$. Go to b

36. .Obtain generators for the following distributions:

(a) *Rayleigh*

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, x \geq 0 \tag{2.31}$$

(b) *Triangular*

$$f(x) = \frac{2}{a}(1 - \frac{x}{a}), 0 \leq x \leq a \tag{2.32}$$

37. Show that if (X,Y) *are independent standard normal variates, then* $\sqrt{X^2+Y^2}$ *has the distribution of the square root of a chi-squared(2) (i.e. exponential(2)) variable and* $arctan(Y/X))$ *is uniform on* $[0, 2\pi]$.

38. Generate the pair of random variables (X,Y)

$$(X,Y) = R(cos\Theta, \ sin\Theta) \tag{2.33}$$

where we use a random number generator with poor lattice properties such as the generator $x_{n+1} = (383x_n + 263)$ mod(10,000) to generate our uniform random numbers. Use this generator together with the Box-Mueller algorithm to generate 5000 pairs of independent random normal numbers. Plot the results. Do they appear independent?

39. Assume that a option has payoff at expiry one year from now ($T = 1$) given by the function $g(S_T) = 0, S_T < 20$, and $g(S_T) = \frac{S_T - 20}{S_T}, S_T > 20$. What is the approximate present value of the option assuming that the risk-neutral interest rate is 5 percent, the current price of the stock is 20, and the annual volatility is 20 percent. Determine this by simulating 1000 stock prices S_T and averaging the discounted return from a corresponding option. Repeat with 100000 simulations. What can you say about the precision?

40. (Log-normal generator)Describe an algorithm for generating log-normal random variables with probability density function given by

$$g(x|\eta,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(logx - log\eta - \sigma^2/2)^2/2\sigma^2\}. \tag{2.34}$$

41. (*hedging with futures*). I need to buy 1000 barrels of heating oil on November 1 1998. On June 1, I go long a December futures contract which allows me to purchase 1000 barrels of heating oil on December 1 for $20 per barrel. Suppose we have observed that the price of heating oil is lognormally distributed with monthly volatility 2 percent. The spot interest rate is presently 5 percent per annum

    (a) What is the value of the oil future on November 1 as a function of the current price of oil?

    (b) Determine by simulation what is the standard deviation of the value of my portfolio on November 1 assuming I sell the futures contract at that time.

    (c) How much difference would it have made if I had purchased the optimal number of futures rather than 1000?

42. (*Multivariate Normal generator*) Suppose we want to generate a multivariate normal random vector $(X_1, X_2, ..., X_N)$ having mean vector $(\mu_1, ..., \mu_N)$ and covariance matrix the $N \times N$ matrix Σ. The usual procedure involves a decomposition of Σ into factors such that $A'A = \Sigma$. For example, A could be determined from the Cholesky decomposition, in Matlab, A=*chol(sigma),* which provides such a matrix A which is also upper triangular, in the case that Σ is positive definite. Show that if $Z = (Z_1, ..., Z_N)$ is a vector of independent standard normal random variables then the vector $X = (\mu_1, ..., \mu_N) + ZA$ has the desired distribution.

43. (Ahrens-Dieter) Show that the rejection algorithm of Ahrens and Dieter ($b = 1$) has rejection constant c that is bounded for all $\alpha \epsilon (0, 1]$ and approaches 1 as $\alpha \to 0$.

44. What distribution is generated by the following algorithm where U is uniform[0, 1] and V is uniform $[-\sqrt{2/e}, \ \sqrt{2/e}]$?

    (a) GENERATE U, V

    (b) PUT $X = V/U$

    (c) IF $-ln(U) < X^2/4$, GO TO a.; ELSE RETURN X.

45. *(Euler vs. Milstein Approximation)* Use the Milstein approximation with step size .001 to simulate a geometric Brownian motion of the form

$$dS_t = .07S_t dt + .2S_t dW_t$$

Compare both the Euler and the Milstein approximations using different step sizes, say $\Delta t = 0.01, 0.02, 0.05, 0.1$ and use each approximation to price an at-the-money call option assuming $S_0 = 50$ and expiry at $T = 0.5$. How do the two methods compare both for accurately pricing the call option and for the amount of computing time required?

46. *(Cox, Ingersoll, Ross model for interest rates)* Use the Milstein approximation to simulate paths from a CIR model of the form

$$dr_t = k(b - r_t) + \sigma\sqrt{r_t}dW_t$$

and plot a histogram of the distribution of r_1 assuming that $r_0 = .05$ for $b = 0.10$. What is the effect of the parameters k and b?

Chapter 3

Variance Reduction Techniques.

3.1 Introduction.

Suppose you are assigned the task of simulating the behaviour of the following model used to describe an asset price.

Example.

A stock market opens at 10 a.m. and closes at 4:00 p.m. Over that period, it is assumed the market price of the asset follows a geometric Brownian motion with parameters (r, σ^2). However, during market hours a significant news event breaks (e.g. a change in interest rates, a statement from the Federal Reserve, a political event, a relevant major announcement, weather event etc.) occasionally according to a Poisson process with parameter λ and when it does there is an immediate adjustment or "shock" altering the price of the asset by a an amount which is lognormally distributed with mean 1, and parameter σ_2^2. When the market is closed, there is still trading on the global market, although at a lower level, say as a geometric Brownian motion with parameters r, σ_0^2. The "shocks" occur in this period at a lower rate, say λ_0. Explain how you would simulate this process, say for a 30 day period, and use the simulation to price a European Put option on this asset.

In such a problem, the quantity of interest, often called the *performance measure* in the simulation literature, is an expected value of complex function. The function is usually written as a computer program involving some simulated random variables, and whether our random variables are generated by inverse transform, or acceptance-rejection or some other method, it is ultimately a function of a number of uniform variates $U_1, U_2, \ldots$ which are input to the simulation. These uniform random variables determine such quantities as the normally distributed increments of the logarithm of the process and the time and the magnitude of the "shocks". The simulation is being used to estimate an integral

$$E(T) = \int\int..\int T(u_1, u_2, ... u_d)du_1 du_2 \ldots du_d \tag{3.1}$$

over the unit cube in say d dimensions where d is large, and where T is, for example, the expected return from the option under the risk-neutral measure. We now study techniques for evaluating such integrals, beginning with the much simpler case of an integral in one dimension.

Discrete Event Simulation.

Simulation of processes such as networks or queues are examples of *discrete event simulations*(DES), designed to describe systems that are assumed to change instantaneously in response to sudden or discrete events. These are models that can be categorized by a state, with changes only at discrete time points. In modeling an inventory system, for example, the arrival of a batch of raw materials can be considered as such an event which changes the state of the system. A system driven by a system of differential equations in continuous time is an example of a system that is not a DES because the changes occur continuously in time. Typically in a system we identify one or more *performance measures* by which the system is to be judged, and *parameters* which may be adjusted to improve the system performance. Examples are the delay for an air traffic control system, waiting times for bank teller scheduling system, delays or throughput for computer networks, response times for the location of fire stations or supply depots, etc.

One approach to DES is *future event simulation* which proceeds by scheduling one or more future events, choosing the future event in the future event set which has minimum time, updating the state of the system and the clock accordingly and then repeating this whole procedure. A stock price which moves by discrete amounts may be considered a DES. In fact this approach is often used in valuing american options by monte Carlo methods when we use a binomial or trinomial tree.

3.2 Variance reduction for one-dimensional Monte-Carlo Integration.

Consider evaluating an integral of the form $\theta = \int_0^1 f(u)du$ by Monte-Carlo methods. One simple approach, called *crude Monte Carlo* is to randomly sample $U_i \sim U[0,1]$ and then average to obtain $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(U_i)$. This average is obviously an unbiased estimator of the integral and the variance of the estimator is $var(f(U_1))/n$.

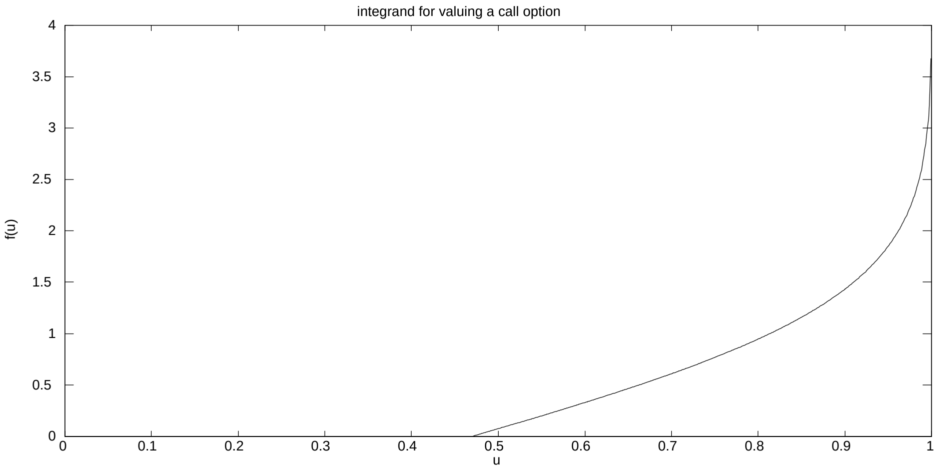


Figure 3.1:

Example: A crude simulation

For a simple example that we will use throughout, consider an integral which might be used to price a call option. Indeed we saw in section 2.8 that if a European option has payoff function given by $V_o(x)$ as a function of the future value of the stock, then the option can be valued using the discounted future payoff from the option under the risk neutral measure; $e^{-rT}E[V_0(S_0e^X)\,]$ where the random variable X has a normal distribution with mean $rT-\sigma^2T/2$ and variance σ^2T . We have also seen that any random variable can, in theory, be generated by inverse transform (although this is not recommended for the normal distribution). Let us for the moment ignore this recommendation and suppose that we have generated X from a single uniform random variable U using $X=F^{-1}(U)$ where F is the normal $(rT-\sigma^2T/2,\,\sigma^2T)$ cumulative distribution function. Then the value of the option can be written as an expectation with respect to the uniform random variable U,

$$e^{-rT}E[V_0(S_0\exp\{F^{-1}(U)\})]=\int_0^1 f(u)du \quad \text{with } f(u)=e^{-rT}V_0(S_0\exp\{F^{-1}(u)\})$$

This function is graphed in Figure **??**.

We have seen that a simple crude Monte Carlo estimator corresponds to evaluating this function and a large number of randomly selected values of $U_i\sim U[0,1]$ and then averaging the results. For example the following function in Matlab evaluates $f(u)$.

```
function v=callopt2(u,S0,ex,r,sigma,T)
% value of the integrand for a call option with exercise price ex, r=annual interest
% rate, sigma=annual vol, S0=current stock price. u=uniform (0,1) input to
```

%generate normal variate by inverse transform. T=maturity

%For Black-Scholes, integrate over (0,1).

x=S0*exp(norminv(u,r*T-sigma^2*T/2,sigma*sqrt(T))); % stock price at time T=$S_0 \exp\{\Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2 T)\}$

v=exp(-r*T)*max((x-ex),0); % This is the discounted to present value of the call option

In the case of initial stock price \$10, exercise price=\$10, annual vol=0.20, r= 5%, $T = .25$ (three months), this is run as

```
U=unifrnd(0,1,1,10000);
mean(callopt2(U,10,10,.05,.2,.25))
```

and this provides an approximate value of the option of 0.4743. We may confirm this with the black-scholes formula, again in *Matlab*, *[CALL,PUT] = BLSPRICE(S0,ex,r,T,sigma,0).* The last argument is the dividend yield which we assumed 0. This provides the result *CALL = 0.4615* indicating that our simulation was reasonably accurate- out by 2 percent or so. In fact one of the advantages of simulation is that it provides a simple estimator of accuracy. In general, when n simulations are conducted, the accuracy is measured by the standard error of the sample mean; $\sigma_f/\sqrt{n}$ where $\sigma_f^2 = var(f(U))$. In this case, this is easily estimated.

```
Sf=sqrt(var(callopt2(U,10,10,.05,.2,.25)));
Sf/sqrt(length(U))
```

giving the standard deviation or standard error of 0.0067. Since approximately normal variables are within 2 standard deviations of their mean (with probability around 95%) we can assert with confidence 95% that the true price of the option is within the interval $0.4743 \pm 2(0.0067)$. and this interval does, in this case, capture the true value of the option. We will look at the efficiency of various improvements in this method, and to that end, we record the value of the variance of the estimator based on a single uniform variate in this case;

$$\sigma^2_{crude} = \sigma_f^2 = var(f(U)) \approx 0.4467.$$

Then the crude Monte Carlo estimator using n function evaluations or n uniform variates has variance approximately $.0.4467/n$. If I were able to adjust the method so that the variance in the numerator were halved, then I could achieve the same accuracy from a simulation using half the number of function evaluations. For this reason, when we compare two different methods for conducting a simulation, the ratio of variances corresponding to a fixed number of function evaluations can also be interpreted roughly as the ratio of computer time required for a given predetermined accuracy. We will often compare various new methods of estimating the same function based on variance reduction schemes and quote the efficiency gain over crude Monte-Carlo sampling.

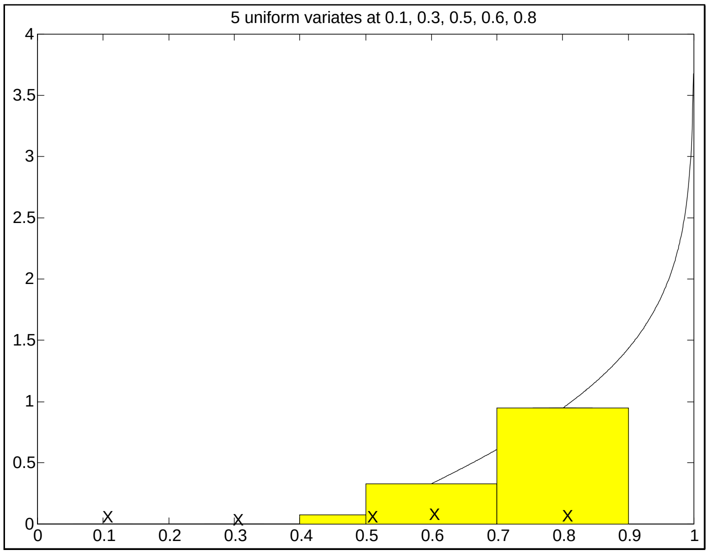


Figure 3.2:

$$\text{Efficiency} = \frac{\text{variance of Crude Monte Carlo Estimator}}{\text{Variance of new estimator}} \tag{3.2}$$

where both numerator a denominator correspond to estimators with the same number of function evaluations (since this is usually the more expensive part of the computation). An efficiency gain of 100 would indicate that the crude Monte Carlo estimator would require 100 times the number of function evaluations to achieve the same variance.

Consider a crude estimator obtained from 5 $U[0,1]$ variates, $0.1, 0.2, 0.5, 0.6, 0.8$. The crude Monte Carlo estimator in the case $n = 5$ is displayed in Figure 3.2, the estimator being the sum of the areas of the marked rectangles. For this particular choice of 5 uniform variates, note that there appears to be an underestimate of the integral because two of random numbers generated were smaller than 0.5 (and contributed 0) and the other three appear to be on average slightly too small. Of course another selection of 5 uniform random numbers may prove to be even more badly distributed.

There are various ways of improving the efficiency of this estimator, many of which partially emulate numerical integration techniques. First we should note that most numerical integrals, like $\hat{\theta}$, are weighted averages of the values of the function at certain points U_i but choice of these points is normally made to attempt reasonable balance in values, and the weights to provide accurate approximations for polynomials of certain degree. For example, the trapezoidal rule corresponds to the U_i equally spaced and the weights are $\frac{1}{n}$ so that the

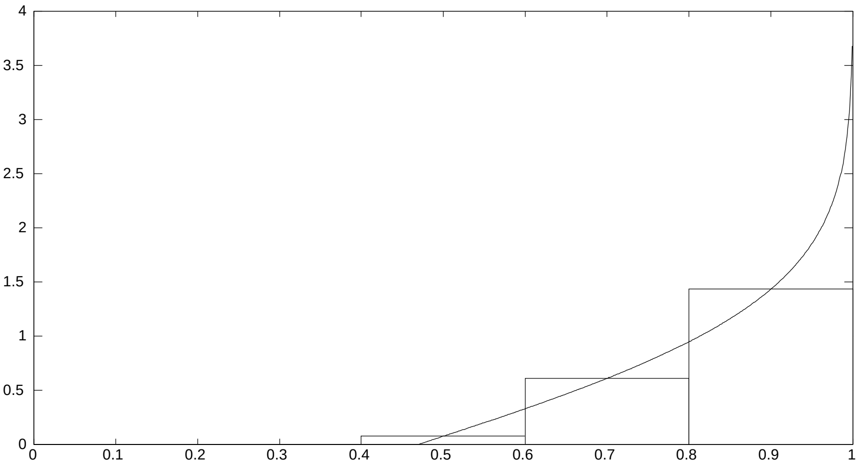


Figure 3.3:

"estimator" of the integral is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(i/n) \tag{3.3}$$

or more precisely, incorporating different weights at the boundary points, $\frac{1}{2n}\{f(0)+ 2f(1/n)+\ldots+2f(1-\frac{1}{n})+f(1)\}$. The balance in large and small values of the function is evident in such a rule, as shown in Figure 3.3. In this case the observations are equally spaced.

Simpson's rule is to generate equally spaced points and weights that(except for endpoints) alternate $2/3n,\ 4/3n,\ 2/3n\ \ldots$. In this case, when n is *even*, the integral is estimated with

$$\hat{\theta} = \frac{1}{3n}\{f(0)+4f(1/n)+2f(2/n)+\ldots+4f(\frac{n-1}{n})+f(1)\} \tag{3.4}$$

The trapezoidal rule is exact for linear functions and Simpson's rule, for quadratic functions.

The analogy between variance reduction techniques and numerical integration methods is a useful one, since it indicates in what direction we should move in order to provide increased accuracy over simple random sampling. We may either vary the weights attached to the individual points or vary the U_i themselves. Notice that as long as the U_i have marginal distribution that is $U[0,1]$, we can introduce any degree of dependence among them (in order to come closer to equal spacing) and $\hat{\theta}$ as defined above will continue to be an unbiased estimator.

Using Antithetic Variates.

Consider the case $n = 2$. Then $\hat{\theta} = \frac{1}{2}\{f(U_1) + f(U_2)\}$ has mean $\int_0^1 f(u)du$ and variance given by $\frac{1}{2}\{var f(U_1) + cov[f(U_1), f(U_2)]\}$ assuming both U_1, U_2 are uniform. In the independent case the covariance term disappears. Notice, however, that if we are able to introduce a *negative covariance*, the resulting variance of $\hat{\theta}$ will be smaller than that of a crude Monte Carlo estimator. When f is monotone, $f(1-U_1)$ decreases when $f(U_1)$ increases, substituting $U_2 = 1-U_1$ has the desired effect (in fact we will show later that we cannot do any better when the function is monotone). Such a value of U_2, balancing by a negative correlation the variability in U_1, is termed an *antithetic variate.* In our example, because the function to be integrated is monotone, there is a negative correlation between $f(U_1)$ and $f(1-U_1)$ and the variance is decreased over simple random sampling. To determine the extent of the variance reduction using antithetic random numbers, suppose we generate $100,000$ uniform variates U and use as well the values of $1-U$ as (for a total of $200,000$ function evaluations as before).

F=(callopt2(U,10,10,.05,.2,.25)+callopt2(1-U,10,10,.05,.2,.25))/2;

This results in *mean(F)=0.46186* and *var(F)=0.1121.* Since each of the 100000 components of F results from two function evaluations, the variance should be compared with $\sigma^2_{crude}/2 = .2234$. The efficiency gain due to the use of antithetic random numbers is .2234/.1121 or about two so roughly half as many function evaluations give the same precision as obtainable with crude Monte Carlo. The introduction of antithetic variates has had the same effect as increasing the sample size by a factor of 2 with the added benefit that only one half as many uniform variates are required.

Stratified Sample.

One of the reasons for the inaccuracy of both crude and antithetic Monte Carlo estimators in the above example is the large interval in which the function is zero, but we nevertheless sample there. We would prefer to concentrate our sample in the region where the function is positive- indeed where it varies more, use larger sample sizes. A method also designed to achieve a better balance, is the use of a *stratified sample.* Suppose, for example, we choose $U_1 \sim U[0,a]$ and $U_2 \sim U[a,1]$. Then the estimator $\hat{\theta}_{st} = af(U_1) + (1-a)f(U_2)$ is unbiased for θ. Moreover,

$$var(\hat{\theta}_{st}) = a^2 var[f(U_1)] + (1-a)^2 var[f(U_2)] + 2a(1-a)cov[f(U_1), f(U_2)] \; . \tag{3.5}$$

Even when U_1, U_2 are independent, there may be a dramatic improvement in variance if the variability of f in the intervals $[0,a]$ and $[a,1]$ is substantially less than in the whole interval $[0,1]$.

Let us return to the example above. Suppose for simplicity we choose independent values of U_1, U_2. In this case

$$var(\hat{\theta}_{st}) \; = \; a^2 var[f(U_1)] + (1-a)^2 var[f(U_2)]. \tag{3.6}$$

For example for $a = .7$, this results in a variance of around 0.0440 obtained from the following

```
var(a*callopt2(unifrnd(0,a,1,50000),10,10,.05,.2,.25)+(1-a)*callopt2(unifrnd(a,1,1,50000),10,10,.05,.2,.25))
```

which gives a variance of about 0.0440. Since each componento the vector above coresponds to two function evaluations we should compare with a crude Monte Carlo estimator with $n = 2$ having variance $\sigma_f^2/2 = 0.2234$.This corresponds to an efficiency gain of .2234/.0440 or around 5. We can afford to use one fifth the sample size by simply stratifying the sample into two strata. The improvement is limited by the fact that we are still sampling in a region in which the function is 0 (although now slightly less often).

A general stratified sample estimator is constructed as follows. We subdivide the interval $[0,1]$ into convenient subintervals $0 = x_0 < x_1 < \ldots x_k = 1$, select n_i variates $V_{ij} \sim U[x_{i-1}, x_i]$. Then the unbiased estimator of θ is

$$\hat{\theta}_{st} \; = \; \sum_{i=1}^{k} (x_i - x_{i-1}) \frac{1}{n_i} \sum_{j=1}^{n_i} f(V_{ij}) \tag{3.7}$$

with variance, in the case of independent V_{ij}:

$$var(\hat{\theta}_{st}) \; = \; \sum_{i=1}^{k} (x_i - x_{i-1})^2 \frac{1}{n_i} var[f(V_{i1})]. \tag{3.8}$$

Once again, if we choose our intervals so that the variation within intervals is small, this provides a substantial improvement over crude Monte Carlo. The optimal choice of sample sizes within intervals are

$$n_i \propto (x_i - x_{i-1}) \sqrt{var[f(V_{i1})]}$$

and the intervals should be chosen so that the variances $var[f(V_{i1})]$ are small. $n_i \propto (x_i - x_{i-1}) \sqrt{var[f(V_{i1})]}$. In general, *optimal sample sizes are proportional to the length of interval times the standard deviation of function evaluated at a uniform random variable on the interval.* The following function was designed for a given selection of intervals to first estimate the variances, then determine appropriate sample sizes, and finally compute the stratified random sample estimator **??** and its variance **??**.

```
function [est,v,n]=stratified(x,nsample)
est=0;
n=[];
```

```
m=length(x);
for i=1:m-1
v= var(callopt2(unifrnd(x(i),x(i+1),1,1000),10,10,.05,.2,.25));
n=[n (x(i+1)-x(i))*sqrt(v)];
end
n=floor(nsample*n/sum(n));
v=0;
for i=1:m-1
F=callopt2(unifrnd(x(i),x(i+1),1,n(i)),10,10,.05,.2,.25);
est=est+(x(i+1)-x(i))*mean(F);
v=v+var(F)*(x(i+1)-x(i))^2/n(i);
end
```

A call to *[est,v,n]=stratified([0 .6 .85 1],100000)* for example generates a stratified sample with the strata the three intervals[0, 0.6],[0.6, 0.85],[0.8, 051] and outputs the estimate 0.4617, its variance 3.5×10^{-7} and the approximately optimal choice of sample sizes $n = 26855, 31358, 41785$. To compare this with a crude Monte Carlo estimator, note that a total of 99998 function evaluations are used so the efficiency gain is $\sigma^2_{crude}/(99998 \times 3.5 \times 10^{-7})$=12.8 so this stratified random sample can account for an improvement in efficiency of about a factor of 13.

Within a stratified random sample we may also introduce antithetic variates designed to provide negative covariance. For example we may use antithetic pairs within an interval if we believe that the function is monotone in the interval and also between intervals as well. For example we may set $V_{ij} = x_{i-1} + (x_i - x_{i-1})U$ and $V_{(i+1)j} = x_{i+1} - (x_{i+1} - x_i)U$ to obtain antithetic pairs within intervals. For a simple example of this applied to the above call option valuation, consider the estimator based on strata [0.47 0.84],[0.84 1]. Here we have not bothered to sample to the left of 0.47 since the function is 0 there.

$$\hat{\theta}_{str,ant} = \frac{0.37}{2}[f(.47 + .37U) + f(.84 - .37U)] + \frac{0.16}{2}[f(.84 + .16U) + f(1 - .16U)]$$

To assess this estimator, we shortened the call *callopt2* to a function of one argument *fn*,

```
function f=fn(u)
f=callopt2(u,10,10,.05,.2,.25);
```

and then evaluated, for U a vector of 100000 uniform,

```
mean(F)                    % gives 0.46156.
var(F)                     %  gives 0.00146.
```

to obtain the result 0.46156. and the variance of the same vector is 0.00146. This should be compared with the crude Monte-Carlo estimator having the same number $n = 4$ of function evaluations as each of the components of the vector $F : \sigma^2_{crude}/4 = 0.4467/4 = .1117$. The gain in efficiency is therefore .1117/.0014 or approximately 80. The above stratified-antithetic simulation with 100,000 input variates and 400,000 function evaluations is equivalent to a crude Monte Carlo simulation with sample size 32 million! Variance reduction makes the difference between a simulation that is feasible on a laptop and one that would require a very long time on a mainframe computer..

Control Variates.

There are two techniques that permit using knowledge about a function with shape similar to that of f . First, we consider the use of a *control variate*. Notice that for arbitrary function $g(u)$,

$$\int f(u)du \;=\; \int g(u)du \;+\; \int (f(u)-g(u))du. \tag{3.9}$$

If the integral of g is known, then we may substitute it for the first term above and calculate the second by crude Monte Carlo, resulting in estimator

$$\hat{\theta}_{cv} \;=\; \int g(u)du \;+\; \frac{1}{n}\sum_{i=1}^{n}[f(U_i)-g(U_i)] \tag{3.10}$$

and the variance is reduced over that of crude Monte Carlo by a factor

$$var[f(U)]/var[f(U)-g(U)], \quad U \sim U[0,1]. \tag{3.11}$$

Let us return to our example. By some experimentation, (which could involve a preliminary crude simulation) we note that the function

$$g(u) = 6[(u-.47)^{+}]^{2} + (u-.47)^{+}$$

provides a reasonable approximation to the function $f(u)$. Moreover, the integral of the function $g(.)$ is easy to obtain. The comparison is seen in Figure 3.4.

The improvement in variance is seen in the figure. By crude Monte Carlo, the variance of the estimator is determined by the variability in the function $f(u)$ over its full range. By using a control variate, the variance of the estimator is determined by the variance of the difference between the two functions, which in this case is quite small. We used the following matlab functions;

```
function g=GG(u)          % the function $g(u)$
% control variate for callopt2.
u=max(0,u-.47);
g=6*u.^2+u;
function [est,var1,var2]=control(f,g,intg,n)
%[est,var1,var2]=control(f,g,intg,n)
%runs a simulation on the function f using control variate g (both character strings) n times.
% intg is the integral of g          % intg=$\int_0^1 g(u)du$
```

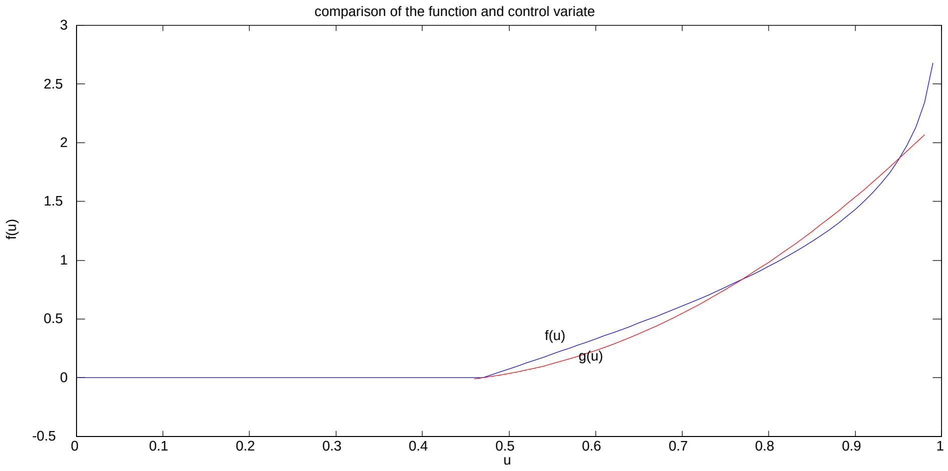


Figure 3.4:

%outputs estimator est and variances var1,var2, with and without control variate.

```
U=unifrnd(0,1,1,n);
FN=eval(strcat(f,'(U)'));          % f(u)
CN=eval(strcat(g,'(U)'));          % g(u)
est=intg+mean(FN-CN);
var1=var(FN);
var2=var(FN-CN);
```

and then the call *[est,var1,var2]=control('fn','GG',2*(.53)ˆ3+(.53)ˆ2/2,100000)* yielding the estimate 0.4602 and variances var1 $= 0.4371$, var2 $= 0.0138$ for an efficiency gain of around 32.

Importance Sampling.

A second technique that is similar is that of *importance sampling.* Again we depend on having a manageable function g that is similar to f but in this case, rather than minimize the difference between the two functions, we choose $g(u)$ such that $f(u)/g(u)$ has little variability over the unit interval. We also require that g is of sufficiently tractable form that we can generate variates from a density proportional to it, i.e. a density of the form $cg(u)$, $0 < u < 1$. This implies, of course, that the function g must be non-negative and have a finite integral. Note that

$$\int f(u)du = E\left[\frac{f(Z)}{cg(Z)}\right] \tag{3.12}$$

where Z has the density function $cg(z)$ and this can be estimated by

$$\hat{\theta}_{im} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(Z_i)}{cg(Z_i)} \tag{3.13}$$

for independent $Z_i \sim cg(z)$. The variance is

$$var\{\hat{\theta}_{im}\} = \frac{1}{n}var\{\frac{f(Z_i)}{cg(Z_i)}\} = \frac{1}{n}\left\{\int \frac{f^2(u)}{cg(u)}du - \theta^2\right\}. \tag{3.14}$$

Returning to our example, we might consider using the same function as before for $g(u)$. However, it is not easy to generate variates from a density proportional to this function g by inverse transform since this would require solving a cubic equation. Instead, let us consider something much simpler, the density function $cg(u) = c_1(u - .47)^+$ having cumulative distribution function $c_2\ [(u - .47)^+]^2$ and inverse c.d.f. $F^{-1}(u) = 0.47 + 0.53\sqrt{u}$. The following function simulates an importance sample estimator:

```
function [est,v]=importance(f,U)
%runs a simulation on the function f using importance density g (defined herein) n times.
% f should be 'fn' obtained from callopt2.
%outputs all the individual estimators (should be averaged) and variance.
%U=unifrnd(0,1,1,n);
%IM is the inverse cf of the importance distribution
IM=.47+.53*sqrt(U);
%IMdens is the density of the importance sampling distribution at IM
IMdens=2*(IM-.47)/(.53)^2;
FN=eval(strcat(f,'(IM)'));
est=FN./IMdens;
v=var(FN./IMdens);
```

The function was called with *[est,v]=importance('fn',unifrnd(0,1,1,100000));* giving estimate mean(*est*) = *0.4610* with variance v = *0.0128* for an efficiency gain of around 35 over crude Monte Carlo.

Combining Monte Carlo Estimators.

We have now seen a number of different variance reduction techniques (this is far from an exhaustive list). With each is a variance formula that would tell us what the gain in efficiency is over crude Monte Carlo if we were able to calculate the integrals appearing in the variance formula. Normally these, too, must be estimated from the sample. Thus it is often not clear *a priori* which sampling procedure and estimator is best. For example if a function f is monotone on $[0, 1]$ then an antithetic variate can be introduced with an estimator of the form

$$\hat{\theta}_{a1} = \frac{1}{2}[f(U) + f(1-U)]\ ,\quad U \sim U[0,1] \tag{3.15}$$

but if the function is increasing to a maximum somewhere around $\frac{1}{2}$ and then decreasing thereafter we might prefer

$$\hat{\theta}_{a2} = \frac{1}{4}[f(U/2) + f((1-U)/2) + f((1+U)/2) + f(1-U/2)] \ . \qquad (3.16)$$

Notice that any weighted average of these two unbiased estimators would also provide an unbiased estimator of θ . The large number of potential variance reduction techniques is an embarrassment of riches causing the usual dilemna; which tool do I use and how do I know it is better than the others? Fortunately, choosing a single method is rarely necessary or desirable. Instead it is preferable to use a weighted average of the available estimators with the optimal choice of the weights provided by regression. More generally suppose that we have n estimators or statistics Y_i , $i = 1, \ . \ . \ n$, all unbiased estimators of the same parameter θ so that that $E(Y_i) = \theta$ for all i . In vector notation, letting $Y' = (Y_1, ..., Y_n)$, we write $E(Y) = Z\theta$ where Z is the vector $Z' = (1, 1, ..., 1)$. Let us suppose for the moment that we know the variance-covariance matrix V of the vector $Y_1, . \ . \ . \ Y_n$.

Theorem 17 *(best linear combinations of estimators)*

The linear combination of the Y_i which is an unbiased estimator of θ and has minimum variance among all linear unbiased estimators is $\sum_i b_i Y_i$ where the vector b is given by

$$b^t = (Z^t V^{-1} Z)^{-1} Z^t V^{-1}.$$

The variance of the resulting estimator is $b^t V b = 1/(Z^t V^{-1} Z)$.

In practice, of course, we almost never know the variance-covariance matrix of a vector of estimators Y . However, with independent replicated values of these estimators, this covariance matrix can easily be estimated from data and the above weights leading to the optimal linear combination computed.

Let us return to the example and attempt to find the best combination of the many estimators we have considered so far. To this end, let

$$\begin{aligned}
Y_1 &= \frac{0.53}{2}[f(.47 + .53u) + f(1 - .53u)] \quad \text{an antithetic estimator,} \\
Y_2 &= \frac{0.37}{2}[f(.47 + .37u) + f(.84 - .37u)] + \frac{0.16}{2}[f(.84 + .16u) + f(1 - .16u)], \\
Y_3 &= 0.37[f(.47 + .37u)] + 0.16[f(1 - .16u)], \\
Y_4 &= \int g(x)dx \ + \ [f(u) - g(u)], \\
Y_5 &= \hat{\theta}_{im}, \quad \text{the importance sampling estimator (\textbf{??}).}
\end{aligned}$$

Y_2 is an antithetic and stratified estimator, Y_3 a simpler version of a stratified-antithetic estimator, Y_4 is a control variate estimator and Y_5 the importance sampling estimator all for a single input random variate. In order to determine the optimal linear combination we need to generate simulated values of all 5 estimators using the same uniform random numbers as inputs. We determine the best linear combination of these estimators using

```
function [o,v,b,V]=optimal(U)
% generates optimal linear combination of five estimators and outputs
% average estimator, variance  and weights
Y1=(.53/2)*(fn(.47+.53*U)+fn(1-.53*U));
Y2=.37*.5*(fn(.47+.37*U)+fn(.84-.37*U))+.16*.5*(fn(.84+.16*U)+fn(1-.16*U));
Y3=.37*fn(.47+.37*U)+.16*fn(1-.16*U);
intg=2*(.53)^3+.53^2/2;
Y4=intg+fn(U)-GG(U);
Y5=importance(’fn’,U);
X=[Y1’ Y2’ Y3’ Y4’ Y5’];
mean(X)
V=cov(X);
Z=ones(5,1);
V1=inv(V);
b=V1*Z/(Z’*V1*Z);
o=mean(X*b);
v=1/(Z’*V1*Z);
```

and one run of this estimator, called with *[o,v,b,V]= optimal(unifrnd(0,1,1,100000))* yields $o = 0.4615$, $v = 1.1228\times10^{-5}$, $b\prime = [-0.5505, 1.4490, 0.0998, 0.0491, -0.0475]$. The answer is accurate to at least four decimals which is not surprising since the variance per uniform random number is $v = 1.1228e - 005$. Consequently the variance of the mean of 100,000 estimators is 1.1228×10^{-10}, the standard error is around .00001 so we should expect accuracy to at least 4 decimal places. Note that some of the weights are negative and others are greater than one. This is common since the technique being used is regression. The effect of some estimators may be, on subtraction, to render the function more linear and accommodate it to another estimator, for example. The efficiency gain is an impressive 0.4467/0.000011228 or about 40,000. However since there are 10 function evaluations for each uniform variate, the efficiency when we adjust for the number of function evaluations is 4,000. This simulation using 100,000 uniform random numbers and taking a couple of minutes on a Pentium (233 Mhz) is equivalalent to *four billion simulations by crude Monte Carlo, a major task on the largest computers available!*

If I were intending to use this simulation method repeatedly, I might well wish to see whether some of the estimators can be omitted without too much loss of information. Since the variance of the optimal estimator is $1/(Z^tV^{-1}Z)$, we might choose an estimator for deletion (for example deleting the $i'th$ row and column of V) which has the least effect on this quantity or its reciprocal $Z^tV^{-1}Z$. In particular, if we let $V_{(i)}$ be the matrix V with the $i'th$ row and column deleted and $\sum V^{jk}$ as the sum of all elements of the matrix V^{-1},then we can identify $\sum V^{jk} - \sum V_{(i)}^{jk}$ as the loss of information when the $i'th$

estimator is deleted. Since not all estimators have the same number of function evaluations, we should adjust this information by $FE(i)$ =number of function evaluations required by the $i'th$ estimator. In other words, if an estimator i is to be deleted, it should be the one corresponding to

$$\min_i \{ \frac{\sum V^{jk} - \sum V^{jk}_{(i)}}{FE(i)} \}.$$

Since we know the variance of thecombined estimator per function evaluation, we should drop the $i'th$ estimator if this minimum is less than the information per function evaluation in the combined estimator. In the above example with all five estimators included, $\sum V^{jk} = 88757$ (with 10 function evaluations per uniform variate) so the information per function evaluation is $8,876$.

i	$\sum V^{jk} - \sum V^{jk}_{(i)}$	$FE(i)$	$\frac{\sum V^{jk} - \sum V^{jk}_{(i)}}{FE(i)}$
1	88,048	2	44024
2	87,989	4	21,997
3	28,017	2	14,008
4	55,725	1	55,725
5	32,323	1	32,323

In this case, if we were to elimate one of the estimators, our choice would likely be number 3 since it contributes the least information per function evaluation. However, since all contirube more than $8,876$ per function evaluation, we should likely retain all five.

Common Random Numbers.

We now discuss another variance reduction technique, closely related to control and antithetic variates called *common random numbers.* It is a common problem to need to estimate the difference in performance between two systems. For example, we know the variance of the sample mean and we wish to estimate by Monte Carlo the difference between the variance of a robust estimator of location and that of the mean. Alternatively we may be considering investing in a new piece of equipment that will speed up processing at one node of a network and we wish to estimate the expected improvement in performance. In general, suppose that we wish to estimate by Monte Carlo the difference between two expectations, say

$$Eh_1(X) - Eh_2(Y) \tag{3.17}$$

where X has cumulative distribution function F_X and Y has c.d.f. F_Y. Notice that

$$var[h_1(X) - h_2(Y)] = var[h_1(X)] + var[h_2(Y)] - 2cov\{h_1(X), h_2(Y)\} \tag{3.18}$$

and this is *small* if we can induce a high degree of *positive correlation* between the generated variates X and Y. This is precisely the opposite problem that led to antithetic random numbers, where we wished to induce a high degree of negative correlation. The following theorem supports the use of both common and antithetic random numbers.

Theorem 18 *(maximum/minimum covariance)*

Suppose h_1 and h_2 are both non-decreasing (or both non-increasing) functions. Subject to the constraint that X , Y have cumulative distribution functions F_X, F_Y respectively, the covariance

$$cov[h_1(X), h_2(Y)]$$

is maximized when $Y = F_Y^{-1}(U)$ and $X = F_X^{-1}(U)$ (i.e. for common uniform$[0,1]$ random numbers) and is minimized when $Y = F_Y^{-1}(U)$ and $X = F_X^{-1}(1-U)$ (i.e. for antithetic random numbers).

Proof. We will sketch a proof of the theorem. The following representation of covariance is useful: define

$$H(x,y) = P(X > x, Y > y) - P(X > x)P(Y > y). \tag{3.19}$$

Then the covariance between $h_1(X)$ and $h_2(Y)$, in the case of both h_1 and h_2 monotone differentiable functions, is given by the formula:

$$cov(h_1(X), h_2(Y)) \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)h_1'(x)h_2'(y)dxdy. \tag{4.20}$$

This formula can be verified by twice integrating by parts. The formula shows that our objective in the case of increasing functions h_i and maximizing the covariance is the maximization of $P(X > x, Y > y)$ subject to the constraint that they have the required marginal distributions. Suppose, for example, the distribution were discrete, at the points indicated in the following figure. We wish to maximize $P[X > x, Y > y]$ subject to the constraint that the probabilities $P[X > x]$ and $P[Y > y]$ are held fixed. Note that if there is any weight attached to the points in the lower right quadrant (labelled "LR"), this weight can be reassigned to the points in the upper right quadrant without affecting $P[X > x]$ and by so doing, increasing $P[X > x, Y > y]$. Similarly, any points in the upper left quadrant with positive probability can have this probability moved as well to the upper right quadrant. For the maximum, then, there should be no weight in the quadrants UL and LR for any choice of x. In other words, $X > x$ if and only if $Y > y$ or equivalently, X is a monotone increasing function of Y or they are both increasing functions of a common uniform variate. ■

We now consider a simple but powerful generalization of control variates. The general method is to achieve a reduction in variance by writing an estimator

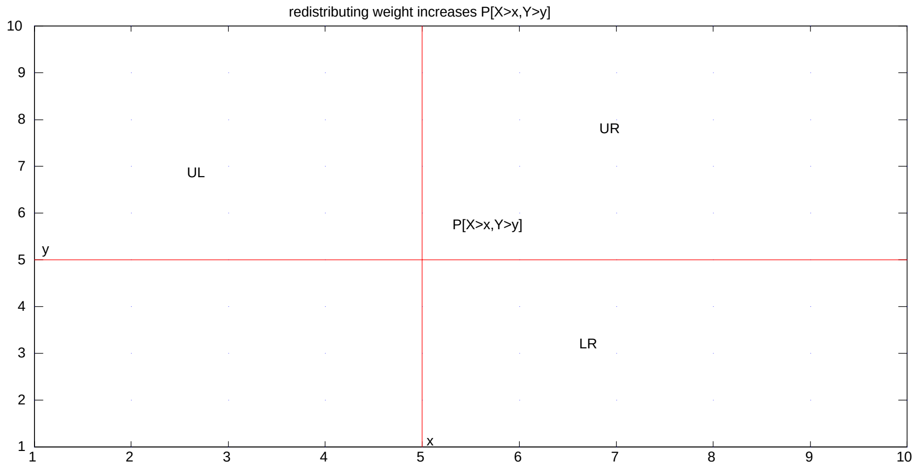


Figure 3.5:

T as the sum of two uncorrelated components, the second of which has mean 0, say.

$$T = T_1 + T_2 \tag{3.20}$$

where $E(T_2) = 0$ and $cov(T_1, T_2) = 0$. Then it is easy to see that T_1 has the same mean as T and variance that is smaller (unless $T_2 = 0$ with probability 1).

One special case is variance reduction by *conditioning*. In order to define conditional expectation, assume random variables X, Y, Z all have finite variances and define $E[X|Y]$ as the unique (with probability one) function of Y which minimizes $E\{X - g(Y)\}^2$. Define $cov(X, Y|Z) = E[XY|Z] - E[X|Z]E[Y|Z]$. The variance reduction is based on the following theorem:

Theorem 19 *(a)* $E(X) = E\{E[X|Y]\}$
(b) $cov(X, Y) = E\{cov(X, Y|Z)\} + cov\{E[X|Z], E[Y|Z]\}$

The theorem is used as follows. Suppose we are given $\hat{\theta}$, an unbiased estimator of θ and Z is some arbitrary conditioning variate. Then $T_1 = E[\hat{\theta}|Z]$, also an unbiased estimator of θ, and $var(T_1) = var(\hat{\theta}) - var\{\hat{\theta} - T_1\}$. In other words, any variable Z, when conditioned on, can only decrease the variance of the estimator, with the decrease most significant if Z is minimally correlated with $\hat{\theta}$. Reducing variance by conditioning involves searching for a variate whose conditional expectation with the original estimator is computable and which explains most of the variability in $\hat{\theta}$.

3.3 Simulations from the Stationary Distribution of a Markov Chain.

It is often the case that we wish to simulate from a finite ergodic Markov chain in its equilibrium or stationary state, but this stationary distribution does not take a simple form permitting one of the standard techniques. We do assume that we are able to generate transitions in the Markov Chain, however. In other words if the chain is presently in state i, we are able to generate from the distribution proportional to $P_{ij}, j = 1, ...K$. One possibility that is often used is to begin the Markov chain in some initial state and run it for a long time (called the initial transient) until we are quite sure that it is in equillibrium, and then use a subsequent portion of this chain, discarding the initial transient. Clearly this is not an efficient use of resources if the initial transient is long, but if it is shortened, we run the risk of introducing bias into our simulations. There is, however, a method which permits simulation directly from the stationary distribution of the Markov chain due to Propp and Wilson (1995). Let us suppose we are able to generate transitions in the Markov chain using a function of the form $\phi(i, U_t)$ where U_t is a uniform$[0, 1]$ distribution, so if $X_t = i$, then the next state of the chain at time $t + 1$ is generated as $X_{t+1} = \phi(i, U_t)$. Note that by composition we can generate the chain over an interval, for example $F_s^t(i) = \phi(...\phi(\phi(\phi(\phi(i, U_s), U_{s+1}), U_{s+2}), U_{s+3})..., U_{t-1})$ will generate the value of X_t given that $X_s = i$. Now imagine an infinite sequence U of independent uniform $U_t, t = ..., -3, -2, -1$ used to generate the state of a chain at time 0. Let us imagine for the moment that there is a value of M such that $F^0_{-M}(i)$ is a constant function of i. In this case we say that *coalescence* has occurred in the interval. This means that no matter where we start the chain at time $-M$ it ends up at the same point at time 0. In this case, it is quite unnecesary to simulate the chain over the whole infinite time interval $-\infty < t \leq 0$ since it had to be somewhere at time $t = -M$ and *no matter where it was, it ended up at the same point at time* $t = 0$. In this event, we can safely consider the common value of the chain at time 0 to be generated from the stationary distribution since it is exactly the same value as if we had run the chain from $t = -\infty$. Now there is an easy way to check whether coalescence has occurred in an interval if the state space of the Markov chain is ordered. For example suppose the states are numbered $1, 2, ..., K$. Then it is often possible to arrange that the function $\phi(i, U)$ is monotonic in its first argument for each value of U. This is the case, for example when we use inverse transform to generate the value, for example $\phi(i, U) = \inf\{j; \sum_{l=1}^{j} P_{il} > U\}$ provided that the partial sums $\sum_{l=1}^{j} P_{il}$ are monotonic functions of i. Notice then that the functions $F^0_{-M}(i)$ are all monotonic functions of i and so if $F^0_{-M}(1) = F^0_{-M}(K)$ then it must be a constant function. Notice also that if there is any time in an interval $[s, t]$ at which coalescence occurs so that F_s^t is a constant function, the same will be true of any interval containing it $[S, T] \supset [s, t]$.

It is easy to prove that coalescense occurs for sufficiently large M. For an ergodic finite Markov chain, there is some step size L such that every transition has positive probability $P[X_{t+L} = j|X_t = i] > \epsilon$ for all i, j. Then the probability of coalescence in an interval of length L is at least $\varepsilon^{K+1} > 0$ and since there are infinitely many intervals disjoint of length L in $[-\infty, 0]$ and the event that there is a coalescense in each are independent, the probability that coalescense occurs somewhere in $[-\infty, 0]$ is 1..

We now detail the Propp Wilson algorithm

1. Set $M = 1, X_U = K, X_L = 1$
2. Generate $U_{-M}....U_{-M/2+1}$
3. For $t = -M$ to -1 repeat
    (a) obtain $X_L = \phi(X_L, U)$ and $X_U = \phi(X_U, U)$
    (b) If $X_L = X_U$ stop and output $X(0) = X_L$
4. $M = 2M$
5. Go to 2.

It is important to notice in this algorithm that the random variable U_t once generated is NOT generated again on a subsequent pass when M is doubled. If it were the algorithm would be biased. It is reused at each pass until coalescense.

3.4 Coupling and Perfect Simulations.

A very elegant and simple method of generating two correlated simulations is that of coupling. This is similar to the method of control variates; it generates a positive correlation, but it has one remarkable advantage. In some cases, only approximate information about the parameter is necessary to generate the random variable X.

Let us begin with a simple example. Suppose we wish to generate random variables, say $\mathcal{N}(\mu)$ having a $N(\mu, 1)$ distribution. If we wish a maximum possible correlation with $\mathcal{N}(0)$ for example, then the simplest possibility, equivalent to the use of common random numbers, is to use $\mathcal{N}(\mu) = \mu + Z$ where Z is $N(0, 1)$. This provides a correlation of 1 with $\mathcal{N}(0)$. Coupling is related to acceptance-rejection. The idea is to generate $\mathcal{N}(\mu)$ and $\mathcal{N}(0)$ by acceptance-rejection, ensuring that provided that the point chosen lies under both densities, the same point is used. More precisely, we use the following algorithm:

(1) Choose a point (x, y) uniformly distributed under the $N(0, 1)$ probability density function.

(2) Define $\mathcal{N}(\mu) = x$ provided that the point also lies under the $N(\mu, 1)$ p.d.f. Otherwise define $\mathcal{N}(\mu) = x + iW$ where $W =$ the width of the normal p.d.f. at height $y = 2\sqrt{-2\ln(\sqrt{2\pi}y)}$ and i is the unique whole number chosen

so that the point $(x + iW, y)$ does lie under the $N(\mu, 1)$ probability density function.

For a particular value of y, note that $\mathcal{N}(\mu)$ is a piecewise constant, non-decreasing function of μ. For example the following graph.

FIGURE

We could, of course, generalize to an arbitrary variance generating $N(\mu, \sigma^2)$ using $\sigma\mathcal{N}(\mu/\sigma)$. let us denote such a random variable by $\mathcal{N}(\mu, \sigma^2|z, y)$.

Now the fact that this graph is piecewise constant means in fact we do not always need precise information concerning the value of μ in order to generate the random variable $\mathcal{N}(\mu)$. Suppose for example we were given an upper and lower bound on the mean $a \leq \mu \leq b$. Suppose furthermore that we generated $\mathcal{N}(a)$ and $\mathcal{N}(b)$ and note that they are identical. The monotonicity insures that then $\mathcal{N}(\mu)$ will take on this common value. Of course it is quite possible that $\mathcal{N}(a)$ and $\mathcal{N}(b)$ are not identical, and in this case it may be necessary to find a tighter bound on on the value of μ of the above form. But when there is some computational effort involved in tightening this bound, we have saved this effort with some (hopefully large) probability.

(FINISH........)

3.5 Some Multivariate Applications in Finance.

3.5.1 Asian Options.

Consider as an example a discretely sampled Asian call option on an asset with price process $S(t)$. An Asian option is like a European option but with the value function dependent not on the closing price of the underlying but on an average. An Asian call options pays an amount equal to $max(0, \bar{S}_k - K)$ where $\bar{S}_k = \frac{1}{k}\sum_{i=1}^{k} S(iT/k)$. Here k depends on the frequency of sampling (e.g. if $T = .25$ and sampling is weekly, then $k = 13$). If $S(t)$ follows a geometric Brownian motion, this is the sum of lognormally distributed random variables (rather than normally distributed ones) and as a result the distribution of the partial sums is very difficult to obtain. However, the distribution of the *geometric mean* is relatively simpler where the geometric mean of n values $x_1, ..., x_n$ is $(x_1x_2...x_n)^{1/n}$. Our objective is to determine the value of the option $E(V_1) = E\{e^{-rT}max(0, \bar{S}_k - K)\}$. Since we expect geometric means to be close to arithmetic means, we may use as a control variate the random variable $V_2 = e^{-rT}max(0, \tilde{S}_k - K)$ where $\tilde{S}_k = \{\prod_{i=1}^{k} S(iT/k)\}^{1/k}$. Assume that V_1 and V_2 obtain from the same simulation and are therefore possibly correlated. Of course V_2 is only useful as a control variate if its expected value can be determined analytically or numerically more easily than V_1. Fortunately, in this case, we may use the relation between a geometric Brownian motion and the normal random walk to determine the distribution of the geometric mean. Since $S(t) = S_0e^{Y(t)}$ where $Y(t)$ is a Brownian motion with drift $r - \sigma^2/2$

and diffusion σ, it follows that $\tilde{S}_k$, has the same distribution as does

$$S_0 \exp\{\frac{1}{k}\sum_{i=1}^{k} Y(iT/k)\}. \tag{3.21}$$

This is a weighted average of the independent normal increments of the process and therefore normally distributed. In particular if

$$\begin{aligned}\bar{Y} &= \frac{1}{k}\sum_{i=1}^{k} Y(iT/k)\\ &= \frac{1}{k}[k(Y(T/k)) + (k-1)\{Y(2T/k) - Y(T/k)\} + (k-2)\{Y(3T/k) - Y(2T/k)\}\\ &\quad +... + \{Y(T) - Y((k-1)T/k)\}],\end{aligned}$$

then

$$\mu_Y = E(\bar{Y}) = \frac{r - \sigma^2/2}{k}\sum_{i=1}^{k} iT/k = (r - \frac{\sigma^2}{2})\frac{k+1}{2k}T$$

and

$$\begin{aligned}\sigma_Y^2 &= var(\bar{Y}) = \frac{1}{k^2}\{k^2 var(Y(T/k)) + (k-1)^2 var\{Y(2T/k) - Y(T/k)\} + ...\}\\ &= \frac{T\sigma^2}{k^3}\sum_{i=1}^{k} i^2 = \frac{T\sigma^2(k+1)(2k+1)}{6k^2}.\end{aligned}$$

The closed form solution for the price $E(V_2)$ in this case is therefore easily obtained because it reduces to the same integral over the lognormal density that leads to the Black-Scholes formula. Recall that the Black-Scholes formula gives

$$E(e^{-rT}(S_0\exp\{N((r - \frac{\sigma^2}{2})T, \sigma^2 T)\} - K)^+ = E(S_0 \exp\{N(-\frac{\sigma^2 T}{2}, \sigma^2 T)\} - Ke^{-rT})^+.$$

In fact

$$\begin{aligned}E(V_2) &= E\{e^{-rT}(S_0 e^{\overline{Y}} - K)^{\ +}\}, \overline{Y} \sim N((r - \frac{\sigma^2}{2})\frac{k+1}{2k}T, \frac{T\sigma^2(k+1)(2k+1)}{6k^2})\\ &= E\{S_0 \exp\{rT(\frac{1-k}{2k}) + \frac{\sigma^2 T(1-k^2)}{12k^2}\}\exp\{N(-\frac{\widetilde{\sigma}^2 T}{2}, \widetilde{\sigma}^2 T)\} - Ke^{-rT})^{\ +}\}\end{aligned}$$

with

$$\widetilde{\sigma}^2 = \frac{\sigma^2(k+1)(2k+1)}{6k^2}.$$

Thus $E(V_2)$ is given by the Black-Scholes formula with S_0 replaced by $S_0\exp\{rT(\frac{1-k}{2k})+\frac{\sigma^2T(1-k^2)}{12k^2}\}$ and σ^2 by $\frac{\sigma^2(k+1)(2k+1)}{6k^2}$. Of course when $k=1$, this gives exactly the same result as the basic Black-Scholes because in this case, the asian option corresponds to the average of a single observation.

Now the most elementary form of control variate suggests using the estimator

$$E(V_2)+V_1-V_2 \tag{3.22}$$

where the random variables V_1, V_2 result from the same simulation. This expression may be regarded as a simple approximation to V_1 when observations on V_2 are available. A better approximation is obtained by regression. Since elementary regression yields

$$V_1-E(V_1)=\beta(V_2-E(V_2))+\epsilon \tag{3.23}$$

where

$$\beta=\frac{cov(V_1,V_2)}{var(V_2)} \tag{3.24}$$

and the errors ϵ have expectation 0, it follows that $E(V_1)+\epsilon=V_1-\beta(V_2-E(V_2))$ an unbiased estimator of $E(V_1)$ having smallest variance among all linear combinations of V_1 and V_2. Now when $\beta\ =1$ this reduces to the simpler form of the control variate technique. However, this form is generally better in terms of maximizing efficiency. Of course it is necessary to estimate the covariance and the variances in the definition of β from the simulations themselves.

In practice, of course, there is not a single simulation but many and the random variables V_1, V_2 above are replaced by their averages over many simulations. The following table is similar to that in Boyle, Broadie and Glasserman(1995), compares the variance of the crude Monte Carlo estimator with that of an estimator using a simple control variate. In this case, $K=100, k=50, r=0.10, T=0.2$ and standard errors are estimated from 10,000 simulations. Since the efficiency is the ratio of the number of simulations required for a given degree of accuracy, or alternatively the ratio of the variances, this table indicates efficiency gains due to the use of a control variate of several hundred. Further gains can be achieved using the modified control variate described above.

Table 4.1. Standard Errors for Arithmetic Average Asian Options.

SIGMA	K/S	STANDARD ERROR	STANDARD ERROR
		OF CRUDE	OF CONTROL
0.2	0.9	0.0558	0.0007
	1.0	0.0334	0.00064
	1.1	0.00636	0.00046
0.4	0.9	0.105	0.00281
	1.0	0.0659	0.00258
	1.1	0.0323	0.00227

The following function implements the control variate for an asian option and was used to produce the above table. We avoid looping in the function in order to speed up computations.

```
function [v1,v2,sc]=asian(r,S0,sig,T,K,k,n)
%computes the value of an asian option V1 and control variate V2
%S0=initial price, K=strike price
%sig = sigma, k=number of time increments in interval [0.T]
%sc is value of the score function for the normal inputs with respect to
% r the interest rate parameter.
%Repeats for a total of n simulations.
v1=[]; v2=[]; sc=[];
mn=(r-sig^2/2)*T/k;
sd=sig*sqrt(T/k);
Y=normrnd(mn,sd,k,n);
sc= (T/k)*sum(Y-mn)/(sd^2);
Y=cumsum([zeros(1,n); Y]);
S = S0*exp(Y);
v1= exp(-r*T)*max(mean(S)-K,0);
v2=exp(-r*T)*max(S0*exp(mean(Y))-K,0);
disp(['standard errors ' num2str(sqrt(var(v1)/n)) ' num2str(sqrt(var(v1-v2)/n))])
```

For example we might confirm the last row of the above table using the command
asian(.1,100/1.1,.4,.2,100,50,10000);.

3.5.2 Girsanov's Lemma.

In the above, we implemented just one variance reduction scheme. There are many other possibilities. We expect the option to have a payoff closely related to the closing value of the stock $S(T)$. It might be reasonable to stratify the sample; i.e. sample more often when $S(T)$ is large, and there are several ways to implement this. One is to use importance sampling and generate $S(T)$ from a geometric Brownian motion with drift larger than rS_t so that it is more likely that $S(T) > K$. But if we do this we need to then multiply by the ratio of the two probability density functions or the density of one process with respect to the other. This density is given by a result called Girsanov's lemma and a very simple form of this lemma appears below.

Theorem 20 *(Girsanov) Consider an Ito process generated by the equation*

$$dS_t = \mu(S_t)dt + \sigma(S_t)dW_t. \tag{3.25}$$

Let the distribution of this process be P. Suppose we generate a similar process with the same diffusion term but different drift term

$$dS_t = \mu_0(S_t)dt + \sigma(S_t)dW_t. \tag{3.26}$$

Assume that in both cases, the process starts at the same initial value S_0 and let the distribution of this process be P_0. Then the "likelihood ratio" or the density $\frac{dP}{dP_0}$ of P with respect to P_0 is

$$\frac{dP}{dP_0} = \exp\{\int_0^T \frac{\mu(S_t) - \mu_0(S_t)}{\sigma^2(S_t)}dS_t - \int_0^T \frac{\mu^2(S_t) - \mu_0^2(S_t)}{2\sigma^2(S_t)}dt\}$$

Proof. Despite the claim to the left, this is not technically a proof, but an argument in favour of the above formula. Let us consider the conditional distribution of a small increment in the process S_t under the model (**??**). Since this distribution is conditionally normal distributed it has conditional probability density function given the past

$$\frac{1}{\sqrt{2\pi dt}} \exp\{-(dS_t - \mu(S_t)dt)^2/(2\sigma^2(S_t)dt) \tag{3.27}$$

and under the model (**??**), it has the conditional probability density

$$\frac{1}{\sqrt{2\pi dt}} \exp\{-(dS_t - \mu_0(S_t)dt)^2/(2\sigma^2(S_t)dt) \tag{3.28}$$

The ratio of these two probability density functions is

$$\exp\{\frac{\mu(S_t) - \mu_0(S_t)}{\sigma^2(S_t)}dS_t - \frac{\mu^2(S_t) - \mu_0^2(S_t)}{2\sigma^2(S_t)}dt\}$$

But the joint probability density function over a number of disjoint intervals is obtained by multiplying these conditional densities together and this results in

$$\begin{aligned} &\Pi_t \exp\{\frac{\mu(S_t) - \mu_0(S_t)}{\sigma^2(S_t)}dS_t - \frac{\mu^2(S_t) - \mu_0^2(S_t)}{2\sigma^2(S_t)}dt\} \\ = \quad &\exp\{\int_0^T \frac{\mu(S_t) - \mu_0(S_t)}{\sigma^2(S_t)}dS_t - \int_0^T \frac{\mu^2(S_t) - \mu_0^2(S_t)}{2\sigma^2(S_t)}dt\} \end{aligned}$$

where the product of exponentials results in the sum of the exponents, or, in the limit as the increment dt approaches 0, the corresponding integrals. ■

Girsanov's result is very useful in conducting simulations because it permits us to change the distribution under which the simulation is conducted. In general, if we wish to determine an expected value under the measure P, we may conduct a simulation under P_0 and then multiply by $\frac{dP}{dP_0}$ or if we use a subscript on E to denote the measure under which the expectation is taken,

$$E_P V(S_T) = E_{P_0}[V(S_T)\frac{dP}{dP_0}].$$

Suppose for example I wish to determine by simulation the expected value of $V(r_T)$ for an interest rate model

$$dr_t = \mu(r_t)dt + \sigma dW_t \tag{3.29}$$

for some choice of function $\mu(r_t)$. Then according to Girsanov's theorem, we may simulate under the much simpler Brownian motion model $dr_t = \mu_0 dt + \sigma dW_t$ and then average the values of

$$V(r_T)\frac{dP}{dP_0} = V(r_T)\exp\{\int_0^T \frac{\mu(r_t) - \mu_0}{\sigma^2}dr_t - \int_0^T \frac{\mu^2(r_t) - \mu_0^2}{2\sigma^2}dt\}$$

We can then choose the constant μ_0 to (approximately) produce minimum variance of the above average.

Call option with stochastic interest rates.

Consider for an example the pricing of an option, say a call option under an assumption of stochastic interest rates. We will use the method of conditioning, although there are other potential variance reduction tools here. Suppose the asset price, (under the risk-neutral probability measure, say) follows a model of the form

$$dS_t = r_t S_t dt + \sigma S_t dW_t^{(1)} \tag{3.30}$$

where the spot interest rate model is the Brennan-Schwartz model,

$$dr_t = a(b - r_t)dt + \sigma_0 r_t dW_t^{(2)} \tag{3.31}$$

where $W_t^{(1)}$, $W_t^{(2)}$ are independent Brownian motion processes. Here b is the long run average of the interest rates and the parameter $a > 0$ governs how quickly reversion to b occurs.

We wish to use simulation to price a derivative, say a call option.

Control Variates. The first method might be to use crude Monte Carlo; i.e. to simulate both the process S_t and the process r_t, evaluate the option at expiry, say $V(S_T, T)$ and then discount to its present value by multiplying by $exp\{-\int_0^T r_t dt\}$. However, in this case we can exploit the knowledge that the interest rates are independent of the Brownian motion process $W_t^{(1)}$ which drives the asset price process. For example, suppose that the interest rate function r_t *were known* (equivalently: condition on the value of the interest rate process). While it may be difficult to obtain the value of an option under the model (**??**),(**??**) it is easier under the model which assumes constant interest rate c. Let us call this constant interest rate model for asset prices with the same initial price S_0 and driven by the equation

$$dZ_t = cZ_t dt + \sigma Z_t dW_t^{(1)} \tag{3.32}$$

model "0" and denote expectations under this distribution by E^0. The value of the constant c will be determined later. Assume that we simulated the asset prices under this model and then valued a call option, say. Then since

$$\ln(Z_T/S_0) \text{ has a } N((c-\frac{\sigma^2}{2})T, \sigma^2 T) \text{ distribution}$$

we could use the Black-Scholes formula to determine the conditional expected value

$$\begin{aligned} E^0[exp\{-\int_0^T r_t dt\}(Z_T - K)^+ | r_s, 0 &< s < T] = E[(S_0 e^{(c-\overline{r})T} e^W - e^{-\overline{r}T} K)^+ | \overline{r}] \\ &= BS(S_0 e^{(c-\overline{r})T}, K, \overline{r}, T, \sigma) \end{aligned}$$

where W has a $N(-\sigma^2 T/2, \sigma^2 T)$ distribution and $\overline{r} = \frac{1}{T}\int_0^T r_t dt$ is the average interest rate over the period. The function BS is the Black-Scholes formula with arguments in the same order as the Matlab function *blsprice*. In other words by replacing the the interest rate by its average over the period and the initial value of the stock by $S_0 e^{(c-\overline{r})T}$, the Black-Scholes formula provides the value for an option on an assset driven by (**??**) conditional on the value of $\overline{r}$. This is a useful control variate for the problem. Its unconditional expected value can be determined by generating the interest rate processes and averaging values of $BS(S_0 e^{(c-\overline{r})T}, K, \overline{r}, T, \sigma)$. Finally we may estimate the required option price using an average of values of

$$exp\{-\int_0^T r_t dt\}[(S_T - K)^+ - (Z_T - K)^+]\} + E\{BS(S_0 e^{(c-\overline{r})T}, K, \overline{r}, T, \sigma)\}$$

for S_T and Z_T generated from the same simulation (i.e. using common random numbers).

The choice of the costant c can be made either for convenience or to minimize the variance of the estimators $exp\{-\int_0^T r_t dt\}[(S_T-K)^+ - (Z_T - K)^+]\}$. One simple and effective choice is $c = \overline{r}$ since this means that the second term is $E\{BS(S_0, K, \overline{r}, T, \sigma)\}$.

Importance Sampling The expectation under the correct model could also be determined by multiplying this random variable by the ratio of the two likelihood functions and then taking the expectation under E^0. In other words, by Girsanov's Theorem, $EV(S_T, T) = E^0\{V(S_T, T)\frac{dP}{dP_0}\}$ where P and P_0 are the measures corresponding to the P and P_0 processes respectively. The required Radon-Nykodym derivative is

$$\frac{dP}{dP_0} = exp\{\int_0^T \frac{r_t - c}{\sigma^2 r_t} dr_t - \int_0^T \frac{r_t^2 - c^2}{2\sigma^2} dt\} \tag{3.33}$$

The resulting estimator of the value of the option is therefore an average over all simulations of the value of

$$V(r_T, T)exp\{-\int_0^T r_t dt + \int_0^T \frac{r_t - c}{\sigma^2 r_t} dS_t - \int_0^T \frac{r_t^2 - c^2}{2\sigma^2} dt\} \tag{3.34}$$

where the trajectories r_t are simulated under the constant interest rate model (??). In other words, $r_t = \exp\{(c - \sigma^2/2)t + \sigma W_t^{(1)}\}$ for standard Brownian motion $W_t^{(1)}$.

The drift parameter in this model c can be chosen to minimize the variance of the estimator.

3.6 Simulating Barrier and lookback options

Suppose we observe a stochastic process X_t over the interval $0 \leq t \leq T$. As is often done with financial time series we record the initial value or *open* of the time series $O = X_0$ the terminal value or *close* $C = X_T$, the maximum over the period or the *high* $H = max\{X_t;\ 0 \leq t \leq T\}$ and the minimum or the *low* $L = min\{X_t; 0 \leq t \leq T\}$. The recording of all four variables is common in practice but the use of all is rare. For example, the variance or volatility parameter is commonly estimated using only the open and close O, C although the information available for this parameter in the four random variables O, C, H, L is about seven times as great for Brownian motion. More commonly, in fact, volatility is determined as the "implied volatility" from the price of a derivative sold on the open market which has X_t as the underlying asset price. The implied volatility is the value of the volatility parameter which produces the market price of a given option (usually a heavily-traded or benchmark option). For example suppose a particular option with strike price K, maturity T, and initial value of the stock S_0 is traded on the market at a price given by V. Then we may solve the equation $BS(S_0, K, r, T, \sigma) = V$ for the implied volatility parameter σ. This estimate of volatiliy may differ substantially from the historical volatility obtained in the Black-Scholes model by computing the sample variance of the returns $\log(S_{t+1}/S_t)$. Nevertheless since it agrees with the market price of the option it expected to more closely reflect the risk-neutral distribution Q and is therefore used. The disadvantage of this method of calibrating the parameter is that its value will depend on the strike price of the option, as well as the time and maturity parameters.

For many of the properties of the process, both for calibrating volatility parameters and for valuing products that depend on the tail behaviour of the distributions, the vector of values (H, L, O, C) is substantially more informative than is (O, C) and should generally be used, particularly if the product may be a function of the maximum or the minimum. Clearly the extreme observations of a process is not only highly informative for the volatility but also for important measures related to the risk associated with a given investment. In general, measures of risk such as VaR (Value at Risk) should also be adapted to observations of the high and low of a process.

Properties of the joint distribution of these random variables conditional on O are well-known in certain special cases. For example when X_t is a Brownian motion with zero drift the joint distribution is given in Billingsley(1968). Such results permit us to calculate the joint distribution for the single most important model for security prices, the geometric Brownian motion model. This joint distribution is important for the valuation of derivatives that involve the maximum of the process; options such as barrier options, look-back options, caps, floors, etc. The assumption that the underlying asset follows a geometric Brownian motion is standard in the valuation of such derivative products. However, it has also been well-known for some time that the distribution of asset prices *do not* follow a geometric Brownian motion, and at best this is a fairly crude approximation applicable only on a large scale. Many alternatives have been suggested which attempt to accommodate the larger-than-Gaussian tails experienced in the market, including mixtures of normal distributions, processes with jumps, geometric Brownian motion subordinated to a random clock, and the stable processes.

The application of the joint distribution to option pricing is well developed, and there are many path-dependent options whose valuation requires both the close and the extrema of a process. For example a barrier option has payoff function a value of the close C conditional on the extrema (H, L) lying in some region, usually an interval. The option may be knocked-out (i.e. the option has value 0 if the process leaves the interval) or knocked in (the option only has value if the process enters the interval at some point). Look-back options have payoff that is a function of both the high and the close or the low and the close. For example a look-back put option has payoff given by $(H - C)$ equivalent to the return obtained by selling the stock at its high and covering the short position at the close. A look-back call option is similar, with payoff of the form $(C - L)$. Hindsight options, sometimes called fixed-strike look-back options, have payoff which depends only on the distribution of the high or low, for example $(H - K)^+$ in the case of a hindsight call option. There is a large number of papers devoted to the valuation of such options. For details, see the references in Broadie, Glasserman and Kou (1996). We begin with the result on the distribution of highs and lows.

Theorem 21 *Suppose* $\sigma(x)$ *and* $\lambda(t)$ *are positive real-valued functions such that* $g(x) = \int^x \frac{1}{\sigma(y)}dy$ *and* $\tau(t) = \int_0^t \lambda^2(s)ds$ *are well defined on* $\Re^+$ *and let* $\tau^{-1}(t)$ *denote the inverse function of* τ. *Suppose a process* X_t *having real parameters* ν *and diffusion coefficient* $\sigma(x) > 0$ *satisfies the stochastic differential equation:*

$$dX_t = [\nu + \frac{1}{2}\sigma'(X_t)]\sigma(X_t)\lambda^2(t)dt + \sigma(X_t)\lambda(t)dW_t. \tag{3.35}$$

Define $H = max\{X_t; 0 \leq t \leq T\}$ *to be the* high *over the period* $[0, T]$, $O = X_0, X_T = C$.
(a) Then with f_0 *representing the probability density function of* $X_T = C - O$ *in the case* $\nu = 0$, *we have*

$$U_H = \frac{f_0(2g(H) - g(O) - g(C))}{f_0(g(C) - g(O))} \sim U[0,1]$$

and U_H *is independent of* C. *(b) For each value of* T**,** $Z_H = (g(H) - g(O))(g(H) - g(C))$ *is independent of* O, C, *and has an exponential distribution with mean* $\frac{1}{2}\int_0^T \lambda^2(s)ds$. *Similarly for the low,*

$$U_L = \frac{f_0(2g(L) - g(O) - g(C))}{f_0(g(C) - g(O))} \sim U[0,1]$$

and $Z_H = (g(L) - g(O))(g(L) - g(C))$ *is independent of* O, C, *and has an exponential distribution with mean* $\frac{1}{2}\int_0^T \lambda^2(s)ds$.

Proof. First note that under the monotonically increasing transformation $Z_t = g(X_t)$ and using Ito's lemma, Z_t satisfies a stochastic differential equation of the form;

$$dZ_t = \nu\lambda^2(t)dt + \lambda(t)dW_t, \quad 0 \leq t \leq T \tag{3.36}$$

If we now apply a time change and consider the process $Z_{\tau^{-1}(t)}$ it is easy to see that this process is a Brownian motion with drift, i.e. it satisfies the equation

$$dZ_t = \nu dt + dW_t, \quad 0 \leq t \leq \tau^{-1}(T). \tag{3.37}$$

Therefore it is sufficient to prove the result for a Brownian motion process with T replaced by $\tau^{-1}(T)$. Assume without loss of generality that $Z_0 = 0$. Now let P, E denote probabilities and expectations in model (4.35) and P_0, E_0 be probabilities and expectations in the same model with zero drift i.e. $\nu = 0$. Assume without loss of generality that $Z_0 = 0$. Now a Brownian motion process can be considered as a limit of a sequence of simple random walks so the first step is to verify a result for simple random walks, one in which the process jumps up or down with equal probability 1/2. Suppose that Figure 3.6 is a rescaled sample path from such a process. Consider two values z and s both possible values for the process. Notice that for each sample path ending at s which passes above a barrier at the point z there is a corresponding path ending at $2z-s$ obtained by reflecting the original path from the first time it crosses the barrier at z. In fact there is a one-one correspondence between such paths. It follows that there is the same number of paths ending at $s<z$ and with maximum $\geq z$ as paths ending at the reflection of s,namely $2z-s$. Since for a simple random walk all paths have the same probability, it follows that for a simple random walk

$$P[\max_{t<T} Z_t \geq z, Z_T = s] = P[Z_T = 2z-s].$$

Now since the Brownian motion process with drift $\nu = 0$ is a limit of such simple random walks, the same result holds provided we interpret objects like $P[Z_T = 2z-s]$ as a probability density function. Note also that the conditional distribution of a Brownian motion process satisfying $dZ_t = \nu dt + dW_t$ given Z_T does not depend on the value of the drift term ν. Therefore

$$\begin{aligned} P[\max_{t<T} Z_t &\geq z|Z_T = s] = P_0[\max_{t<T} Z_t \geq z|Z_T = s] \\ &= \frac{P_0[\max_{t<T} Z_t \geq z, Z_T = s]}{P_0[Z_T = s]} = \frac{P_0[Z_T = 2z-s]}{P_0[Z_T = s]}, \quad \text{for } s<z. \end{aligned}$$

Now recall that by the inverse transform property, if a random variable X has a continuous cumulative distribution function $F(x)$, then $F(X)$ has a uniform$[0,1]$ distribution. The same is true if we replace $F(x)$ by the survivor function $P[X \geq x]$. It follows if we denote $H = \max_{t<T} Z_t$, that conditional on $Z_T = s$, the random variable $f_0(2H-s)/f_0(s)$ has a uniform $[0,1]$ distribution where f_0 is the normal $(0,T)$ probability density function. Since this distribution does not depend on the value of s, the uniform random variable $f_0(2H-Z_T)/f_0(Z_T)$ is independent of the the value Z_T.(b) Now assume that X_t satisfies equation (**??**) with $X_0 = 0$. We have seen that the random variable $U_H = f_0(2H-C)/f_0(C)$ has a uniform$[0,1]$ distribution independent of C where f_0 is the normal$(0,\tau^{-1}(T))$ probability density function. On taking logarithms and simplifying

$$-\ln(U_H) = \frac{2}{\tau^{-1}(T)} H(H-C)$$

has an exponential distribution with mean 1 and therefore $H(H-C)$ has an exponential distribution with mean $\frac{1}{2}\tau^{-1}(T)$. More generally if we remove the assumption that $O=0$,

$$(H-O)(H-C) \sim \exp(\frac{1}{2}\tau^{-1}(T)) \text{ independently of } O, C.$$

The results for the low follow by symmetry. ■

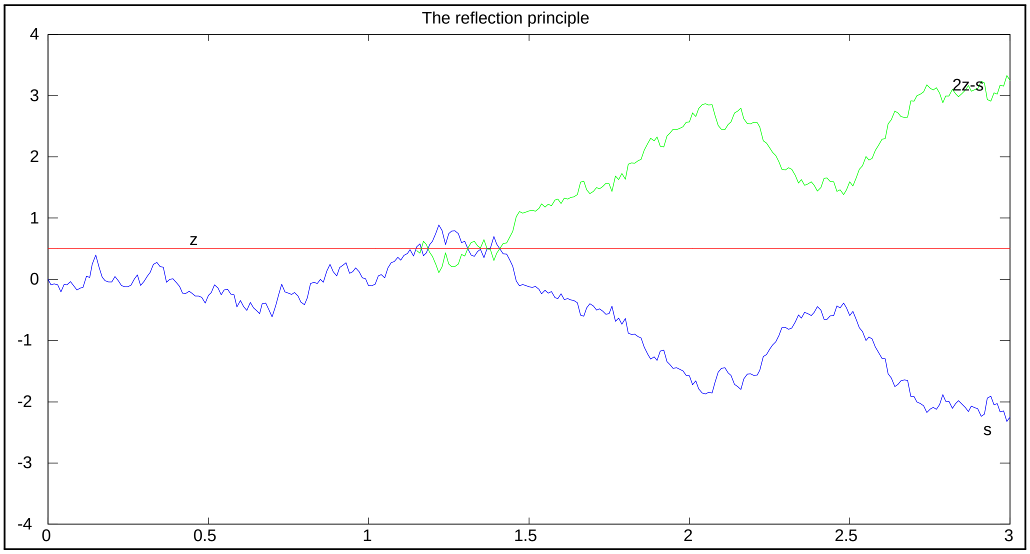


Figure 3.6:

Corollary 22 *For a Brownian motion process,*

$$\begin{aligned}(H-O)(H-C) &\sim \exp(\frac{1}{2}\sigma^2 T) \text{ independently of } O, C \text{ and}\\ (L-O)(L-C) &\sim \exp(\frac{1}{2}\sigma^2 T) \text{ independently of } O, C.\end{aligned}$$

Corollary 23 *For a Geometric Brownian motion process,*

$$\begin{aligned}\ln(H/O)\ln(H/C) &\sim \exp(\frac{1}{2}\sigma^2 T) \text{ independently of } O, C \text{ and}\\ \ln(L/O)\ln(L/C) &\sim \exp(\frac{1}{2}\sigma^2 T) \text{ independently of } O, C.\end{aligned}$$

These two corollaries may be used to directly simulate a value for the high given the value of the close. For example, for a Brownian motion process, we need only generate a random exponential variate $E \sim \exp(\frac{1}{2}\sigma^2 T)$ and then solve the equation $(H-O)(H-C) = E$ for $H > \max(C, O)$.

The joint probability density function of the high and the close of a Brownian motion is easily obtained from the above theorem. In particular if $X = (C -$

$O)/\sqrt{T}$, $n = \nu\sqrt{T}$ and $Y = (H - O)/\sqrt{T}$, then the joint probability density function of (Y, X) takes the form;

$$f(y,x) = \frac{\sqrt{2}\,(2\,y - x)\,e^{-\frac{(2\,y-x)^2}{2}+nx-\frac{n^2}{2}}}{\sqrt{\pi}}, \quad for \quad -\infty < x < y, \quad y > 0. \quad (3.38)$$

Note that X is a complete sufficient statistic for the parameter n . Moreover, the probability density function of $Y|X = x$ is

$$f_{Y|X}(y|x) = 2(2y - x)e^{-2y(y-x)}, \quad y \geq x \quad (3.39)$$

With $Z = Y(Y - X)$, the conditional density of Z is

$$f_{Z|X}(z|x) = 2e^{-2z}, \quad z \geq 0 \quad (3.40)$$

and so is exponential with mean $1/2$, providing an another simple derivation of the exponentially distributed random variable.

Note that by symmetry,

$$U_L = \frac{f_0(2g(L) - g(X_T) - g(X_0))}{f_0(g(X_T) - g(X_0))} \sim U[0,1] \quad (3.41)$$

and $Z_L =(g(L_T) - g(X_0))(g(L_T) - g(X_T))$ has an exponential distribution where L denotes the low.

There is a uniform statistic related to U_H used by Redekop (1995) to test the local Brownian nature of various financial time series. For a Brownian motion process, the statistic

$$\frac{H - O}{2H - O - C} \quad (3.42)$$

is uniformly $[0,1]$ distributed. Redekop observes that the observations on this statistic are far too often close to or equal the extreme values 0 or 1. Equivalently, we may use the statistic

$$\frac{C - O}{2H - O - C} \sim U[-1,1]. \quad (3.43)$$

The uniformity of the statistics U_H and U_L is useful for simulating the values of the high or low and the close of the Brownian motion process without replicating its path. This may be used, for example, to price a European option with knock-out barrier at the point m. Suppose the process is geometric Brownian motion, possibly under a time change.

$$dX_t = [\nu + 1/2]\sigma\lambda^2(t)X_t dt + \sigma\lambda(t)X_t dW_t. \quad (3.44)$$

In this case $\sigma(X_t) = \sigma X_t$ for constant σ and $g(x) = \ln(x)$. Then $\ln(X_T/X_0)$ has a normal distribution under P_0 with mean 0 and variance $\sigma^2 \int_0^T \lambda^2(s)ds$.

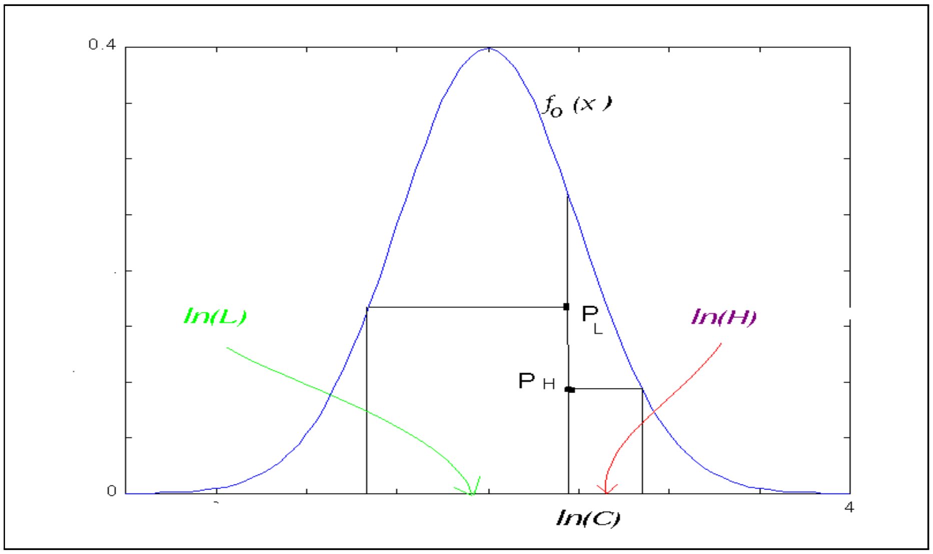


Figure 3.7:

3.6.1 One Barrier

Without generating a path for the process, we may simulate the high and the close geometrically as follows. We begin by generating both high and close under P_0 in the case of zero drift. Consider a graph of the P_0 probability density function $f_0(x)$ of $\ln(C)$ as shown in Figure 3.7.

If we chose to simulated the close using acceptance-rejection, we would choose a point P_H at random uniformly distributed in the region below the graph of this density. Then the x-coordinate of this point is a variate generated from the probability density function $f_0(x)$. Remarkably, the y-coordinate of such a randomly selected point can be used to generate the value of the high. Assume that x-coordinate is the simulated value of $\ln(C)$. Suppose that we extend a line horizontally to the right from this point until it strikes the graph of the probability density and then consider the abscissa of this point- this value is the simulated value of $2\ln(H) - \ln(C)$. It is clear that the corresponding value of $\ln(H)$ does not exceed a boundary at the point m if and only if the point P is below the graph of the probability density function but **not** in the shaded region obtained by reflecting the right hand tail of the density about the vertical line $x = m - \ln(O)$ in Figure 3.8. Thus a knock-out option with payoff function given by $\psi(\ln(C))I(H < e^m)$ can be considered a vanilla European option with payoff function $\psi^*(x) = \psi(x),\ x \leq m,\ \ \psi^*(x) = -\psi(2m - x),\ \ x \geq m.$

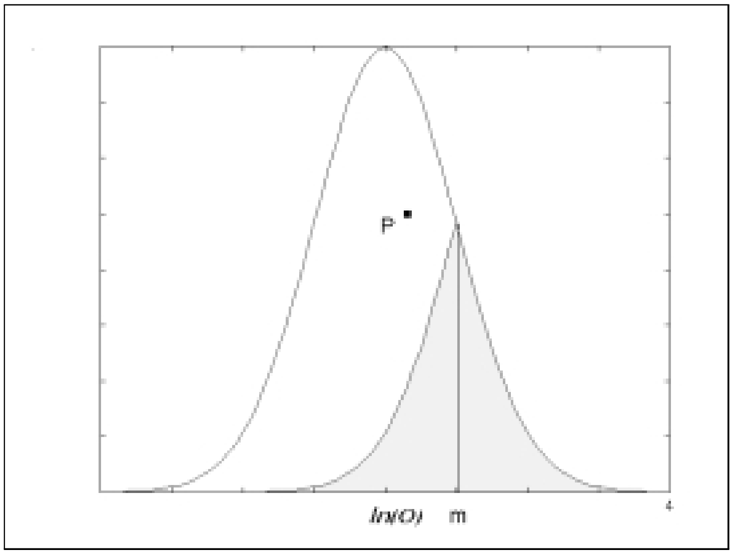


Figure 3.8:

Indeed any option whose value depends on the high and the close of the process can be similarly valued. Either of the points P_H, P_L may be sampled by generating its x coordinate $\ln(C)$ from the density $f_0(x)$ and then subsequently the y coordinate as $Uf_0(\ln(C)),\quad U \sim U[0,1]$.

We now consider briefly the case of non-zero drift. If the original process is a Brownian motion, then the martingale measure will necessarily have zero drift and this consideration is unnecessary. However, for valuing options on a geometric Brownian motion, the drift in the process $\log(X_t)$, though typically small, is non-zero. Fortunately, all that needs to be changed in the above is the marginal distribution of $\ln(C)$ since all conditional distributions given the value of C are the same as in the zero-drift case. For example, in Figure 3.9, a point P has been selected uniformly distributed in the shaded region under $f(x)$, the graph of the probability density function of $\ln(C)$. If this point had been also under the graph of $f_0(x)$ as well, we would have used it as P_H exactly as before to generate the value of the high H. However, in this case, the point was not below the graph of $f_0(x)$ and so we replaced the y-coordinate of the point by another $Uf_0(C)$ where U is $U[0,1]$.

If we wish to price an option with a more general payoff function $\psi(H,C)$ increasing in C, it may be preferable to use importance sampling, for example generate C from a density with more weight in the right tail. In fact since option

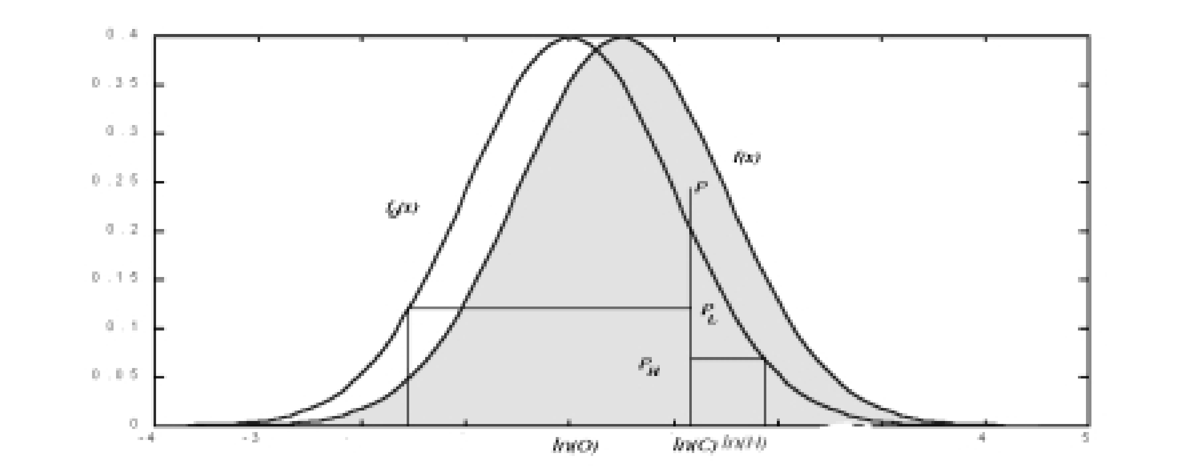


Figure 3.9:

payoff functions are generally simpler functions than many probability densities, it is often desirable to use them as importance sampling distributions. As an example below, we consider a European knock-out call option with exercise price E and knock-out upper barrier at e^m. *Assume for simplicity in the remainder of this example zero drift and that we have already transformed the problem to Brownian motion (i.e. E, C, m, etc are logarithms of the prices).* The payoff function is the triangular region below in Figure 3.10. Suppose we generate a point C at random with probability density proportional to this function. If we repeat this, averaging the length of the line segments $f_0(C) - f_0(2m - C)$, we obtain an estimator of the value of the option.

We similarly show how to value a down and in put option using the figure. Consider a knock-in boundary at $m < E$ where E is the option exercise price. Then the payoff function is $\max(0, E-C)$ when $L < m$, and otherwise the payoff is 0. In order to breach the boundary, a point must be selected from the shaded region in Figure 3.11. The piecewise linear function is the payoff function. Note that the integral of the payoff function over points chosen from the shaded region is equivalent to the integral for points chosen under the normal curve with mean $2m - O$. In other words, in the case of geometric Brownian motion, we can establish the value of this option using the Black-Scholes formula for the price of a call option with the same parameters but with current price of the stock (on a log scale) $2m - O$.

There is a similar geometric view of the conditional distribution of the close C given the high H. Suppose we wish to generate a point from the conditional

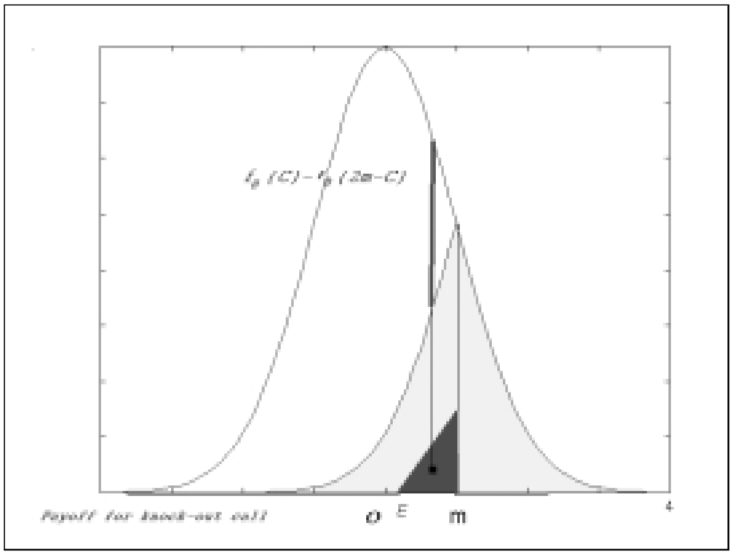


Figure 3.10:

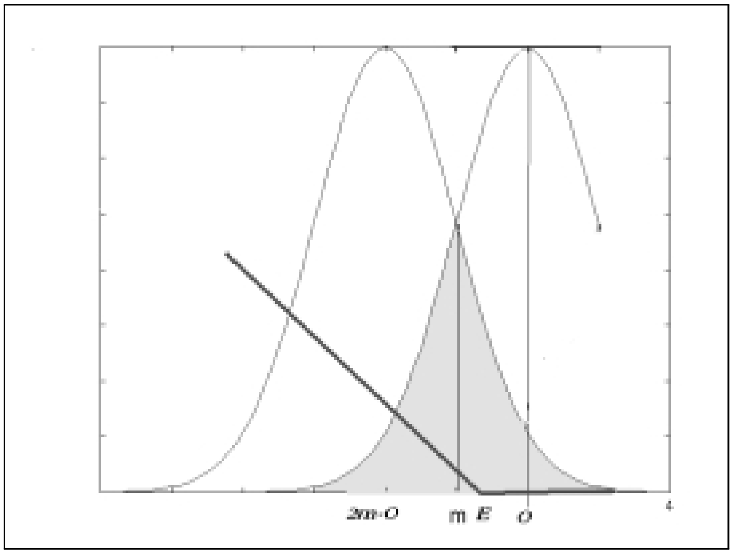


Figure 3.11:

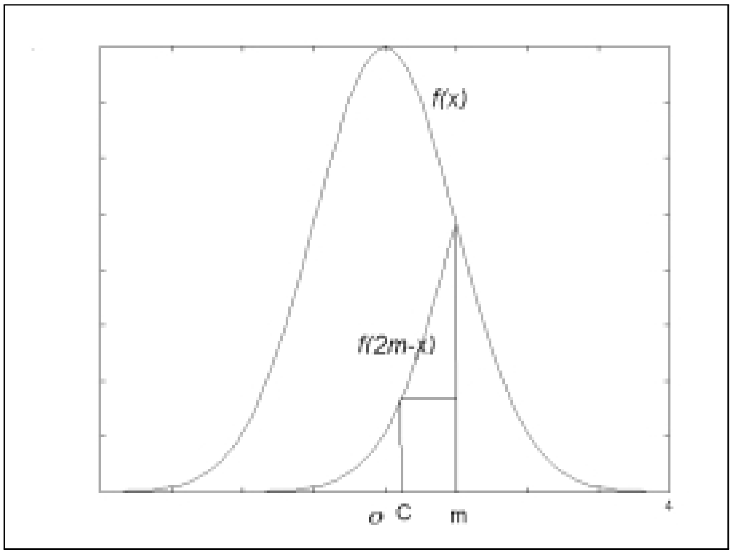


Figure 3.12:

distribution of C given $H = m$. Then a point is chosen uniformly on the interval $[0, f(m)]$ and then projected horizontally to the left until it strikes the graph of $f(2m-x)$. The x-coordinate of the point of intersection is the generated value of the close. This is illustrated in Figure 3.12.

A similar figure illustrates the distribution of two simple statistics that figure prominently in the test of fit of Redekop (1995).

In figure 3.13, consider the distribution of the close given the value of $2H-C$. Clearly the close is distributed uniformly along the horizontal stripe, i.e. $U[O-(2H-C),(2H-C)]$. Since the value of H is half way between the point C and $2H-C$, it follows that the conditional distribution of H is uniform $U[O, 2H-C]$. One advantage of this uniformity is that it is independent of the distribution f_0. For example it holds whatever the scale parameter of the normal distribution is, or even if the distribution is a variance mixture of normal random variables. Thus it applies to a Brownian motion subordinated to an independent random clock.

This example simply indicates that simulating look-back and barrier options can often be reduced a problem of finding a certain integral or area in a figure. Thus, the array of variance reduction tools that are discussed in this chapter may be applied to problems of this type.

We close this section with a brief discussion of a similar figure which applies to the discrete case. Suppose that the stock price can onl move up or down by a fixed increment Δ as for a simple random walk. Consider the probability histogram of the increment $C-O$ supported by a lattice

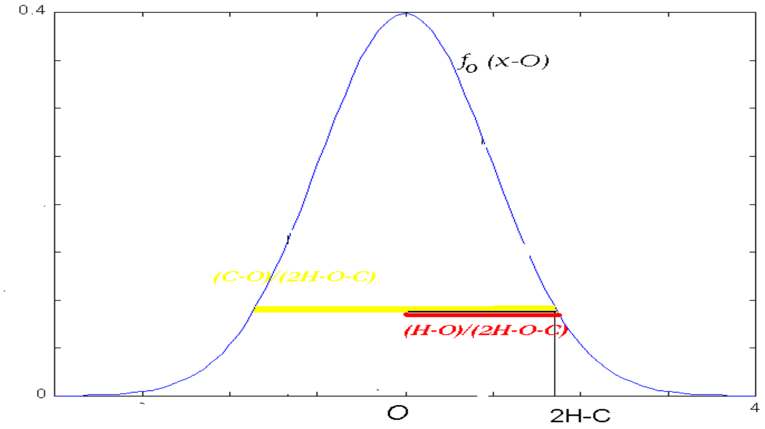
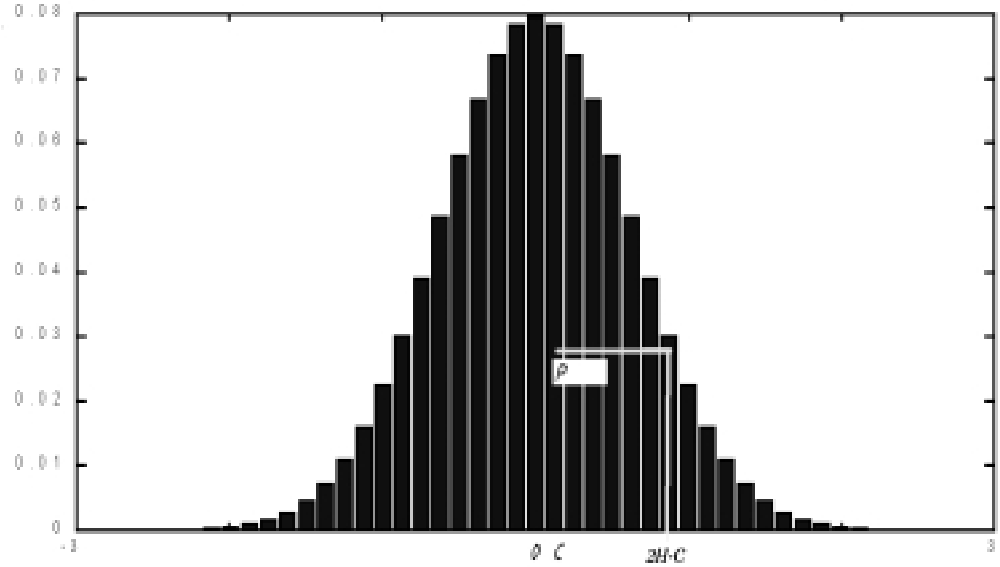


Figure 3.13:



Once again consider a point chosen uniformly and at random under this histogram. The closest point on the lattice to the x-coordinate of this point can be considered a generated value of the close. Moreover, if we run a horizontal line to the right, then the last bar, say y, passed through before it leaves the histogram corresponds, with one small adjustment, to the generated value of $2H - C$ as in the continuous case. The adjustment stems from the fact that the difference between this point and the close is $2(H-C)$ which is an even multiple of the size of a jump. Thus, if we define $H = \Delta[\frac{y-C}{2\Delta}]$, where the square brackets indicate "the integer part of ", then H is a simulated value of the high.

3.6.2 One Factor, Two barriers

We have discussed a simple device above for generating jointly the high and the close or the low and the close of a process given the value of the open. The joint distribution of H, L, C given the value of O or the distribution of C in the case of upper and lower barriers is more problematic. Consider a single factor model and two barriers- an upper and a lower barrier. Note that the high and the low in any given interval is dependent, but if we simulate a path in relatively short segments, by first generating n increments and then generating the highs and lows within each increment, then there is an extremely low probability that the high and low of the process will both lie in the same short increment. For example for a Brownian motion with the time interval partitioned into 5 equal subintervals, the probability that the high and low both occur in the same increment is less than around 0.011 whatever the drift. If we increase the number of subintervals to 10, this is around 0.0008. This indicates that provided we are willing to simulate highs, lows and close in ten subintervals, pretending that within subintervals the highs and lows are conditionally independent, the error in our approximation is very small.

An alternative, more computationally intensive, is to differentiate the infinite series expression for the probability $P(H \le b, L \ge -a, C = u|O = 0]$ (see for example Billingsley, (1968), p. 79) to obtain the joint probability density and attempt to generate from this density. C can be generated by acceptance rejection in the presence of the barriers since the density is dominated by the density in the absence of any barriers. This requires evaluation of the infinite series.

An alternative allows generating simulated values of the close with absorbing barriers at $-a$, b without using an infinite series. It is well-known that the reflection principle applies to the two-boundary case as well. Assume that the process is standard Brownian motion with drift zero.The above results are primarily useful for an option that depends only o the closing price and either the high or the low over a period. More generally we would like to simulate a closing under the condition that the process remains within a certain interval, e.g. that $H < b$ and $L > -a$. To this end, define a function $frac^+(x) = x - \lfloor x \rfloor$ if $x > 0$ and otherwise $frac^+(x) = 0$.

Theorem 24 *For a Brownian motion process,* $P[-a < L < H < b|C = u] = 1 - P[frac^+(\frac{H}{a+b}) > \frac{b}{a+b}|C = u] - P[frac^+(\frac{-L}{a+b}) > \frac{a}{a+b}|C = u]$

Proof. The following formula is useful for a case in which all three of H, L, C are required . Assume for simplicity that $X(t)$ is a Brownian motion with $\sigma = 1$, $X(0) = 0$,and $H = \max\{X(t); 0 < t < 1\}, L = \min\{X(t); 0 < t < 1\}$ and $C = X(1)$.Then (see for example He, Keirsted, Rebholz, Theorem 2.1) for

$-a < u < b$,

$$P[-a < L < H < b|C = u] = \sum_{n=-\infty}^{\infty}\{\frac{\phi(u-2n(a+b))}{\phi(u)} - \frac{\phi(u-2n(a+b)+2a)}{\phi(u)}\}$$

Note that for $n > 0$,

$$\begin{aligned} P[H &> n(a+b)|C = u] = \frac{\phi(u-2n(a+b))}{\phi(u)}, \\ P[H &> n(a+b) - a|C = u] = \frac{\phi(u-2n(a+b)+2a)}{\phi(u)} \end{aligned}$$

and for $n = -m < 0$,

$$\begin{aligned} P[L &< -m(a+b)|C = u] = \frac{\phi(u+2m(a+b))}{\phi(u)} \\ P[L &< -m(a+b) - a|C = u] = \frac{\phi(u+2m(a+b)+2a)}{\phi(u)} \end{aligned}$$

With these substitutions,

$$\begin{aligned} P[-a &< L < H < b|C = u] = 1 - \frac{\phi(u+2a)}{\phi(u)} \\ -\sum_{n=1}^{\infty} P[n(a+b) &> H > n(a+b) - a|C = u] \\ +\sum_{m=1}^{\infty} P[-m(a+b) - a &< L < -m(a+b)|C = u] \\ &= 1 - P[L < -a|C = u] \\ -P[frac^{+}(\frac{H}{a+b}) &> \frac{b}{a+b}] \\ +\sum_{m=1}^{\infty} P[m + \frac{a}{a+b} &> \frac{-L}{a+b} > m|C = u] \\ &= 1 - P[frac^{+}(\frac{H}{a+b}) > \frac{b}{a+b}] - P[frac^{+}(\frac{-L}{a+b}) > \frac{a}{a+b}] \end{aligned}$$

since

$$\begin{aligned} \sum_{m=1}^{\infty} P[m + \frac{a}{a+b} &> \frac{-L}{a+b} > m|C = u] = \sum_{m=1}^{\infty} P[m - \frac{b}{a+b} > \frac{-L-a-b}{a+b} > m-1|C = u] \\ &= P[frac^{+}(\frac{-L}{a+b}) < \frac{a}{a+b}, \quad \text{and} \quad -L > (a+b)|C = u] \\ &= P[-L > (a+b)|C = u] - P[frac^{+}(\frac{-L}{a+b}) > \frac{a}{a+b}, \quad \text{and} \quad -L > (a+b)|C = \end{aligned}$$

■

The most important feature of this result is that the two probabilities on the right side depend only on the joint distribution of two random variables H and C or L and C.

Let us now assume that we are interested in a barrier or path dependent option whose underlying follows a Geometric Brownian motion. This result can be used to simulate an option that depends on the closing price of a stock in one way if the stock price remains in a given interval $Oe^{-a} < L < H < Oe^b$, but if it breaches a lower barrier at Oe^{-a} or upper barrier at Oe^b it pays a different amount possibly also a function of the closing price. These payoffs could be positive or negative. If it breaches an upper barrier at Oe^b it pays another amount V_b. If both barriers are breached then the payoff is a third amount V_{ab}. For example, suppose the payoff is $V(C)$ if $Oe^{-a} < L < H < Oe^b$. Assume that given the value of C, the payoff of the option is given by

$$V(C,H,L,O)=\begin{cases} V(C) & \text{if} \quad Oe^{-a} < L < H < Oe^b \\ V_a(C) & \text{if} \quad Oe^{-a} > L \text{ and } H < Oe^b \\ V_b(C) & \text{if} \quad Oe^{-a} < L \text{ and } H > Oe^b \\ V_{ab}(C) & \text{if} \quad Oe^{-a} > L \text{ and } H > Oe^b \end{cases}$$

Define random weights as follows: suppose for given C, we may generate H and L either independently of one another or with an arbitrary degree of dependence. These generate random weights

$$\begin{array}{lcc} W_a(C)=1 & \text{if} & L < Oe^{-a} \\ W_a^+(C)=1 & \text{if} & frac^+(\frac{-\ln(L)}{a+b}) > \frac{a}{a+b} \\ W_b(C)=1 & \text{if} & H > Oe^b \\ W_b^+(C)=1 & \text{if} & frac^+(\frac{\ln(H)}{a+b}) > \frac{b}{a+b} \end{array}$$

and otherwise, each of these weights is 0. Then the above theorem shows that $1 - W_a^+(C) - W_b^+(C)$ is an unbiased estimator conditional on C of the probability $P[Oe^{-a} < L < H < Oe^b|C]$ and therefore $W_a^+(C) + W_b^+(C)$ is an unbiased estimator of $P[Oe^{-a} > L \text{ or } H > Oe^b|C]$. Similarly, the conditional expected value of $W_a^+(C) + W_b^+(C) - W_b(C)$ is $P[Oe^{-a} > L \text{ and } H < Oe^b|C]$ and of $W_a^+(C) + W_b^+(C) - W_a(C)$ is $P[Oe^{-a} < L \text{ and } H > Oe^b|C]$. Therefore the weighted average

$$\begin{aligned} &V(C)(1 - W_a^+(C) - W_b^+(C)) + V_a(C)(W_a^+(C) + W_b^+(C) - W_b(C)) \\ &+V_b(C)(W_a^+(C) + W_b^+(C) - W_a(C)) + V_{ab}(C)(W_a(C) - W_a^+(C) + W_b(C) - W_b^+(C)) \end{aligned}$$

provides an unbiased estimator of $E[V(C,H,L,O)|C]$ for each C. These random weights may be replaced by an average of a number of such randomly generated weights for each value of C. The weights can be negative- for example $(1 - W_a^+(C) - W_b^+(C))$ can equal -1. These values are adjustment for a small degree of overcounting that occurs when both barriers are crossed. We have found that this scheme is particularly efficient if H and L are generated using antithetic uniform random numbers; for example solving for $H > \max(C,O)$ the equation

$$\ln(H/O)\ln(H/C) = -\frac{\sigma^2 T}{2}\ln(U)$$

gives an adjustment to the geometric mean of the open and close:

$$H = \sqrt{OC}exp\{\frac{1}{2}\sqrt{\ln(C/O)^2 - 2\sigma^2 T\ln(U)}\}$$

and similarly

$$L = \sqrt{OC}exp\{-\frac{1}{2}\sqrt{\ln(C/O)^2 - 2\sigma^2 T\ln(1-U)}\}$$

where U is Uniform$[0,1]$. This choice leads to a very small probability that the weights $(1 - W_a^+(C) - W_b^+(C))$ are equal to -1.

3.7 Problems

1. Use both crude and antithetic random numbers to integrate the function

   $$\int_0^1 \frac{e^u - 1}{e-1}du.$$

   What is the efficiency gain attributed to the use of antithetic random numbers?

2. How large a sample size would I need, using antithetic and crude Monte Carlo, in order to estimate the above integral, correct to four decimal places, with probability at least 95%?

3. Under what conditions on f does the use of antithetic random numbers completely correct for the variability of the Monte-Carlo estimator? i.e. When is $var(f(U) + f(1-U)) = 0$?

4. Show that if we use antithetic random numbers to generate two normal random variables X_1, X_2, having mean $rT - \sigma^2 T/2$ and variance $\sigma^2 T$, this is equivalent to setting $X_2 = 2(rT - \sigma^2 T/2) - X_1$. In other words, it is not necessary to use the inverse transform method to generate normal random variables in order to permit the use of antithetic random numbers.

5. Use a stratified random sample to integrate the function

$$\int_0^1 \frac{e^u - 1}{e-1}du.$$

What do you recommend for intervals (two or three) and sample sizes? What is the efficiency gain?

1. Use a combination of stratified random sampling and an antithetic random number in the form
$$\frac{1}{2}[f(U/2)+f(1-U/2)]$$
to integrate the function
$$\int_0^1 \frac{e^u-1}{e-1}du.$$
What is the efficiency gain?

2. In the case $f(x) = \frac{e^x-1}{e-1}$, use $g(x) = x$ as a control variate to integrate over [0,1]. Show that the variance is reduced by a factor of approximately 60. Is there much additional improvement if we use a more general quadratic function of x?

3. In the case $f(x) = \frac{e^x-1}{e-1}$, consider using $g(x) = x$ as a control variate to integrate over [0,1]. Note that regression of $f(U)$ on $g(U)$ yields $f(U)-E(f(U)) = \beta[g(U)-Eg(U)]+\varepsilon$ where the error term ε has mean 0 and is uncorrelated with $g(U)$ and $\beta = cov(f(U), g(U))/var(g(U)$. Therefore, taking expectations on both sides and reorganising the terms, $E(f(U)) = f(U) - \beta[g(U) - E(g(U))]$. The Monte-Carlo estimator
$$\frac{1}{n}\sum_{i=1}^{n}\{f(U_i) - \beta[g(U_i) - E(g(U_i))]\}$$
is an improved control variate estimator, equivalent to the one discussed above in the case $\beta = 1$. Determine how much better this estimator is than the basic contol variate case $\beta = 1$ by performing simulations. Show that the variance is reduced by a factor of approximately 60. Is there much additional improvement if we use a more general quadratic function of x?

4. A call option pays an amount $V(S) = 1/(1+\exp(S(T)-k))$ at time T for some predetermined price k. Discuss what you would use for a control variate and conduct a simulation to detemine how it performs, assuming geometric Brownian motion for the stock price, interest rate 5%, annual volatility 20% and various initial stock prices, values of k and T.

5. It has been suggested that stocks are not log-normally distributed but the distribution can be well approximated by replacing the normal distribution by a student t distribution. Suppose that the daily returns X_i are independent with probability density function $f(x) = c(1+(x/b)^2)^{-2}$ (the re-scaled student distribution with 3 degrees of freedom). We wish to estimate a weekly $Var_{.95}$, a value $-p$ such that $P[\sum_{i=1}^{5} X_i < p] = .05$. If we wish to do this by simulation, suggest an appropriate method involving importance sampling. Implement and estimate the variance reduction.

6. Suppose, for example, I have three different simulation estimators Y_1, Y_2, Y_3 whose means depend on two unknown parameters $\theta_1,\ \theta_2$. In particular, suppose Y_1, Y_2, Y_3, are unbiased estimators of $\theta_1, \theta_1+\theta_2, \theta_2$ respectively. Let us assume for the moment that $var(Y_i) = 1,\ \ cov(Y_i, Y_j) = -1/2$. I want to estimate the parameter θ_1. Should I use only the estimator Y_1 which is the unbiased estimator of θ_1, or some linear combination of Y_1, Y_2, Y_3? Compare the number of simulations necessary for a certain degree of accuracy.

7. Consider the *systematic sample* estimator based on the trapezoidal rule:

$$\hat{\theta} = \frac{1}{n}\sum_{i=0}^{n-1} f(V + i/n), \qquad V \sim U[0, \frac{1}{n}]$$

   Discuss the bias and variance of this estimator. In the case $f(x) = x^2$, how does it compare with other estimators such as crude Monte Carlo and antithetic random numbers requiring n function evaluations. Are there any disadvantages to its use?

8. In the case $f(x) = \frac{e^x-1}{e-1}$, use $g(x) = x$ as a control variate to integrate over [0,1]. Find the optimal linear combination using estimators (4.15) and (4.16), an importance sampling estimator and the control variate estimator above. What is the efficiency gain over crude Monte-Carlo?

Chapter 4

Quasi- Monte Carlo Multiple Integration

4.1 Introduction

When integrating in one dimension, a numerical method with N equally spaced points will generally have bias that approaches 0 at the rate $1/N$ when the function has one derivative. This is because

$$\int_{j/N}^{(j+1)/N} f(x)dx - \frac{1}{N}f(y) \;\leq\; \frac{1}{N^2} sup_z \; |f'(z)| \tag{4.1}$$

for y in the interval $j/N \leq y \leq (j+1)/N$. If the function is known to have more bounded derivatives, then numerical integrals can be found which use N points but which have smaller error. Indeed quadrature formulae permit approximating an integral of a polynomial of degree $2N-1$ *exactly* using only N points together with (non-constant) weights attached to those points. By contrast, a Monte Carlo integral with N points has zero bias but standard deviation that is a constant multiple of $1/\sqrt{N}$. Thus the numerical integral has a faster rate of decrease of bias then the rate at which the Monte Carlo integral decreases its standard deviation, and this is a large part of the reason we may prefer numerical integration to Monte Carlo methods in one dimension.

The situation changes substantially in 2 dimensions. Now, if N points are to be distributed over a uniform lattice in some region, the distance between adjacent points will be of order $1/\sqrt{N}$ and this is the order of the bias in a numerical integral. This is the same order as the standard deviation of a Monte Carlo integral. Furthermore, the situation results in a preference for the Monte Carlo integral over such numerical methods for an $s-$dimensional integral when $s \geq 3$. However, there are methods of improving on the placement of the points in a numerical integral to decrease the bias. Quasi-random samples, analogous to equally spaced points in one dimension, are discussed by Niederreiter (1978).

Niederreiter shows that for sufficiently smooth functions and intelligent choice of points, one can achieve the much better rate of convergence.

We have seen a number of methods designed to reduce the dimensionality of the problem. Perhaps the most important of these is conditioning, which can reduce an s dimensional integral to a one-dimensional one. In the multidimensional case, variance reduction has an increased importance because of the high variability induced by the dimensionality of crude methods. The other variance reduction techniques such as regression and stratification carry over to the multivariable problem with little change, except for the increased complexity of determining a reasonable stratification in such problems.

4.2 Errors in numerical Integration

We consider the problem of numerical integration in s dimensions. For $s = 1$ there are classical integration methods, like the trapezoidal rule:

$$\int_0^1 f(u)du \approx \sum_{n=0}^{m} w_n f(\frac{n}{m}), \tag{4.2}$$

where $w_o = w_m = 1/(2m)$, and $w_n = 1/m$ for $1 \leq n \leq m-1$. The trapezoidal rule is exact for any function that is linear and so we can assess the error of integration by using a linear approximation through the points $(\frac{j}{m}, f(\frac{j}{m}))$ and $(\frac{j+1}{m}, f(\frac{j+1}{m}))$.For

$$\frac{j}{m} < x < \frac{j+1}{m}$$

analogous to the Taylor series expansion, if the function has a continuous second derivative,

$$f(x) = f(\frac{j}{m}) + (x - \frac{j}{m})m[f(\frac{j+1}{m}) - f(\frac{j}{m})] + O(x - \frac{j}{m})^2.$$

Integrating both sides between $\frac{j}{m}$ and $\frac{j+1}{m}$, notice that

$$\int_{j/m}^{(j+1)/m} \{f(\frac{j}{m}) + (x - \frac{j}{m})m[f(\frac{j+1}{m}) - f(\frac{j}{m})]\}dx = \frac{f(\frac{j+1}{m}) + f(\frac{j}{m})}{2m}$$

is the area of the trapezoid and the error in the approximation is

$$O(\int_{j/m}^{(j+1)/m} (x - \frac{j}{m})^2) = O(m^{-3}).$$

Adding these errors of approximation over the m trapezoids gives $O(m^{-2})$. Consequently, the error in the trapezoidal rule approximation is $O(m^{-2})$, provided that f has a continuous second derivative on $[0, 1]$.

We now consider the multidimensional case, $s \geq 2$. Suppose we evaluate the function at all of the $(m+1)^s$ points of the form $(\frac{n_1}{m}, \ldots, \frac{n_s}{m})$ and use this to approximate the integral. The classical numerical integration methods use Cartesian product of one-dimensional integration rules. For example, the s-fold Cartesian product of the trapezoidal rule is

$$\int_{I^s} f(\mathbf{u})d\mathbf{u} \approx \sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} w_{n_1} \cdots w_{n_s} f(\frac{n_1}{m}, \ldots, \frac{n_s}{m}), \tag{4.3}$$

where $\bar{I}^s = [0,1]^s$ is the closed s-dimensional unit cube and the w_n are as before. The total number of nodes is $N = (m+1)^s$. From the previous error bound it follows that the error now is $O(m^{-2})$, provided that the second partial derivatives of f are continuous on $\bar{I}^s$. We know that the error cannot be smaller as the above formula can be applied to the case where the function depends on only one variable. In terms of the number N of nodes or function evaluations, since $m = O(N^{1/s})$, the error is $O(N^{-2/s})$, which with increasing dimension s changes drastically. For example if we required $N = 100$ nodes to achieve a required precision in the case $s = 1$, to achieve the same precision for a $s = 5$ dimensional integral using this approach we would need to evaluate the function at a total of $100^s = 10^{10}$ or ten billion nodes. This phenomena is often called the "curse of dimensionality".

A decisive step in overcoming the problem of dimensionality was the development of the Monte Carlo method which is based on random sampling. By the central limit theorem, even a crude Monte Carlo estimate for numerical integration yields a probabilistic error bound of the form $O_P(N^{-1/2})$ in terms of the number N of nodes (or function evaluations) and this holds under a very weak regularity condition on the function f. The remarkable feature here is that this order of magnitude does not depend on the dimension s. This is true even if the integration domain is complicated. *Note however that the definition of "O" has changed from one that essentially considers the worst case scenario to one that measures the average or probabilistic behaviour of the error.*

However, the Monte Carlo method has several deficiencies which may limit its usefulness:

1. There are only probabilistic error bounds (there is no guarantee that the expected accuracy is achieved in a particular case -an alternative approach would optimize the "worst-case" behaviour);

2. The regularity of the integrand is not reflected. The probabilistic error bound $O_P(N^{-1/2})$ holds under a very weak regularity condition but no extra benefit is derived from any additional regularity of the integrand. For example the estimator is no more precise if we know that the function f has several continuous derivatives. Of course in many cases we do not know whether the integrand is smooth and so this property is sometimes an advantage.

3. Generating truly independent random numbers is virtually impossible - in practice we use pseudorandom numbers to approximate independence.

4.2.1 Low discrepancy sequences

The quasi-Monte Carlo method places attention on the objective, approximating an integral, rather than attempting to imitate the behaviour of independent uniform random variates. Our objective is to approximate the integral using a average of the function at N points, and we may attempt to choose the points so that the approximation is more accurate.

$$\int_{I^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}).$$

Quasi Monte-Carlo yields a much better result, giving us the deterministic error bound $O(N^{-1}(logN)^{s-1})$ for suitably chosen sets of nodes and for integrands with a relatively low degree of regularity. Even smaller error bounds can be achieved for sufficiently regular integrands. The sets of nodes producing this high accuracy are obtained from various well-known sequences.

Suppose, as with a crude Monte Carlo estimate, we approximate

$$\int_{I^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}).$$

with $\mathbf{x_1}, \ldots, \mathbf{x_N} \in \bar{I}^s$. The difference is that now the nodes are deterministic, chosen so as to guarantee a small error. The criterion for the choice of deterministic points depends on the numerical problem at hand. For the problem of numerical integration, the selection criterion is easy to find and leads to the concepts of $uniformly\ distributed\ sequence$ and *discrepancy*, which can be viewed as a quantitative measure for the deviation from uniform distribution.

A basic requirement for a low discrepancy sequence is that we obtain a convergent method:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}) = \int_{I^s} f(\mathbf{u})\mathbf{du},$$

and this should hold for a reasonable class of integrands. This suggests that the desirable nodes $\mathbf{x_1}, \ldots, \mathbf{x_N}$ are those which are "evenly distributed" over $\bar{I}^s$. Various notions of discrepancy have been considered as quantitative measures for the deviation from the uniform distribution and we will discuss only two.

4.2.2 Definition: Measures of Discrepancy.

If $\mathcal{B}$ is a nonempty family of Lebesgue-measurable subsets of $\bar{I}^s$, then a general notion of discrepancy of the set $P = \{\mathbf{x_1}, \ldots , \mathbf{x_N}\}$ is given by

$$D_N(\mathcal{B}, P) = \sup_{B\in\mathcal{B}} |\frac{\#\ of\ points\ in\ B}{N} \;-\; \lambda_s(B)|, \tag{4.4}$$

where $\lambda_s(B)$ denotes the Lebesgue measure of B in $\mathcal{R}^s$. By suitable specialization of the family $\mathcal{B}$ we obtain the most important concepts of discrepancy:

The *star discrepancy*:

$$D_N^*(P) = D_N(J^*, P), \tag{4.5}$$

where J^* is the family of all subintervals of $\bar{I}^s$ of the form $\prod_{i=1}^s [0, u_i)$.
The *extreme discrepancy:*

$$D_N(P) = D_N(J, P), \tag{4.6}$$

where J is the family of all subintervals of $\bar{I}^s$ of the form $\prod_{i=1}^s [u_i, v_i)$.

Note that the star discrepancy is a natural one in statistics, since it measures the maximum difference between the empirical cumulative distribution function of the points $\{\mathbf{x_1}, \ldots , \mathbf{x_N}\}$ and the uniform distribution of measure on the unit cube. In order to provide error bounds for the quasi-Monte Carlo approximation we need a notion of total variation.

4.2.3 Definition: Total Variation

If f is sufficiently differentiable then the variation of f on $\bar{I}^s$ in the sense of Hardy and Krause is

$$V(f) = \sum_{k=1}^{s} \sum_{1\leq i_1<\cdots<i_k\leq s} V^{(k)}(f; i_1, \ldots , i_k), \tag{4.7}$$

where

$$V^{(k)}(f; i_1, \ldots , i_k) = \int_0^1 \cdots \int_0^1 |\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}|_{x_j=1, j\neq i_1,\ldots,i_k} dx_{i_1} \cdots dx_{i_k}. \tag{4.8}$$

We have the following inequality:

4.2.4 Theorem: Koksma - Hlawka inequality

If f has bounded variation $V(f)$ on $\bar{I}^s$ in the sense of Hardy and Krause, then, for any $\mathbf{x_1}, \ldots, \mathbf{x_N} \in I^s$, we have

$$|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}) - \int_{I^s} f(\mathbf{u})\mathbf{du}| \;\leq\; V(f)D_N^*(\mathbf{x_1}, \ldots, \mathbf{x_N}). \tag{4.9}$$

Since $D_N^*(P) \leq D_N(P)$, we also have

$$|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}) - \int_{I^s} f(\mathbf{u})\mathbf{du}| \;\leq\; V(f)D_N(\mathbf{x_1}, \ldots, \mathbf{x_N}). \tag{4.10}$$

This result allows us to separate the effects of the integrand from those of the sequence and explains why the discrepancy plays a central role in the theory of quasi-Monte Carlo methods.

The error analysis based on this result demonstrates that small errors are guaranteed if point sets with small star or extreme discrepancy are used. Such sequences are called low-discrepancy sequences

In the one dimensional case the best rate of convergence is $O(N^{-1}\log N)$, $N \geq 2$. It is achieved, for example, by the **van der Corput sequence**.

4.2.5 Examples of low discrepancy sequences

In higher dimensions there exist several constructions:

4.2.6 Halton Sequences.

A Halton sequence is obtained by reversing the digits in the representation of some sequence of integers in a given base. To begin with, consider one-dimensional case $s = 1$ (this is the so-called **van der Corput sequence)** and base $b = 2$. Take the base b representation of the sequence of natural numbers; $1, 10, 11, 100, 101, 110, 111, 1000, 1001, 1010, 1011, 1100, 1101, ...$ and then map these into the unit interval $[0, 1]$. In general, the integer $\sum_{k=0}^{t} a_k b^k$ is mapped into the point $\sum_{k=0}^{t} a_k b^{k-t-1}$. These binary digits are mapped into (0,1) in the following three steps;

1. Write n using its binary expansion. e.g. $13 = 1(8) + 1(4) + 0(2) + 1(1)$ becomes 1101.

2. Reverse the order of the digits. e.g. 1101 becomes 1011.

3. Determine the number that this is the binary decimal expansion for. e.g. $1101 = 1(\frac{1}{2}) + 1(\frac{1}{4}) + 0(\frac{1}{8}) + 1(\frac{1}{16}) = \frac{13}{16}$.

Thus 1 generates 1/2, 10 generates $0(\frac{1}{2}) = 1(\frac{1}{4})$, 11 generates $1(\frac{1}{2})+1(\frac{1}{4})$ and the sequence of positive integers generates the points $1/2, 1/4, 3/4, 1/8, 5/8, 3/8, 7/8, ...$ which are fairly evenly spaced, and perfectly spaced if the number of nodes N is of the form $2^k - 1$. In higher dimensions, say in s dimensions, we choose s distinct primes, $b_1, b_2, ...b_s$ (usually the smallest) and generate, from the same integer m , the s components of the vector according the above method. For example, we consider the case $s = 3$ and use bases $b_1 = 2$, $b_2 = 3, b_3 = 5$. The first few vectors , $(\frac{1}{2}, \frac{1}{3}, \frac{1}{5}), (\frac{1}{4}, \frac{2}{3}, \frac{2}{5}), (\frac{3}{4}, \frac{1}{9}, \frac{3}{5}), ...$are generated in the table below.

m	repres base 2	first component	repres. base 3	second comp	repres base 5	third comp
1	1	1/2	1	1/3	1	1/5
2	10	1/4	2	2/3	2	2/5
3	11	3/4	10	1/9	3	3/5
4	100	1/8	11	4/9	4	4/5
5	101	5/8	12	7/9	10	1/25
6	110	3/8	20	2/9	11	6/25
7	111	7/8	21	5/9	12	11/25
9	1000	1/16	22	8/9	13	16/25
10	1001	9/16	100	1/27	14	21/25

Figure 4.1 provides a plot of the first 500 points in the above Halton sequence of dimension 3.

There appears to be greater uniformity than a sequence of random points would have. Some patterns are discernible on the two dimensional plot of the first 100 points, for example see Figures 4.2, 4.3.

.However, notice that the plot of 100 pairs of independent uniform random numbers in Figure 4.4 shows more clustering and more holes in the point cloud.

These points were generated with the following function for producing the Halton sequence.

```
function x=halton(n,s)
%x has dimension n by s and is the first n terms of the halton sequence of
%dimension s.
p=primes(s*6); p=p(1:s); x=[];
for i=1:s
 x=[x (corput(n,p(i)))'];
end
function x=corput(n,b)
% converts integers 1:n to from van den corput number with base b
m=floor(log(n)/log(b));
n=1:n;                A=[];
for i=0:m
 a=rem(n,b);        n=(n-a)/b;
A=[A ;a];
```

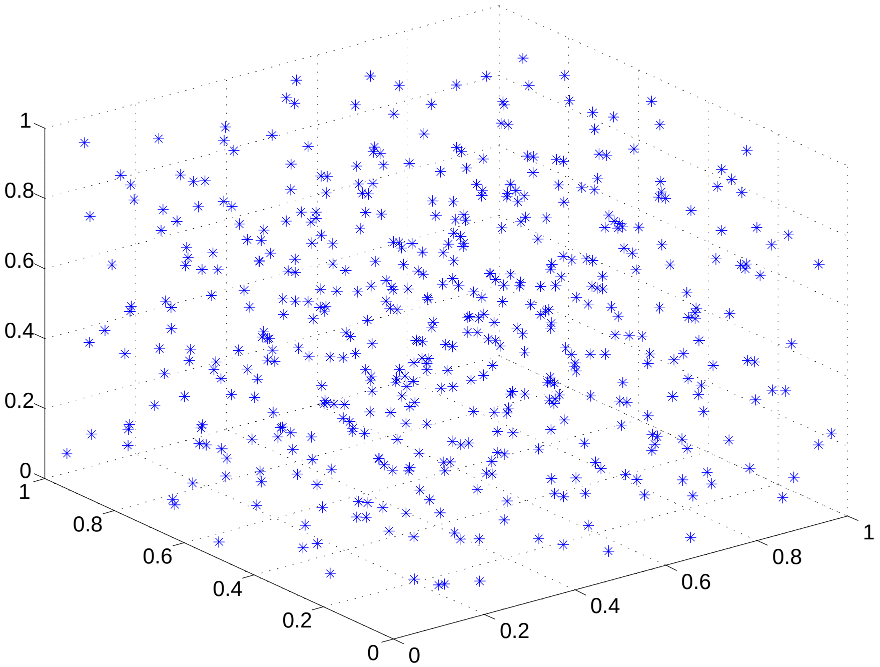


Figure 4.1:

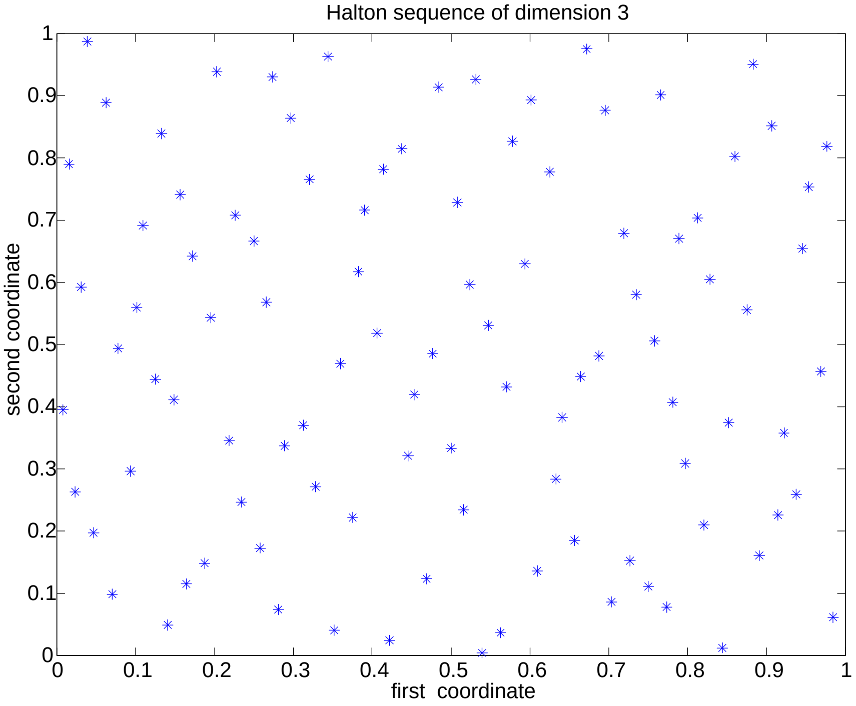

Halton sequence of dimension 3
second coordinate
first coordinate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4.2:

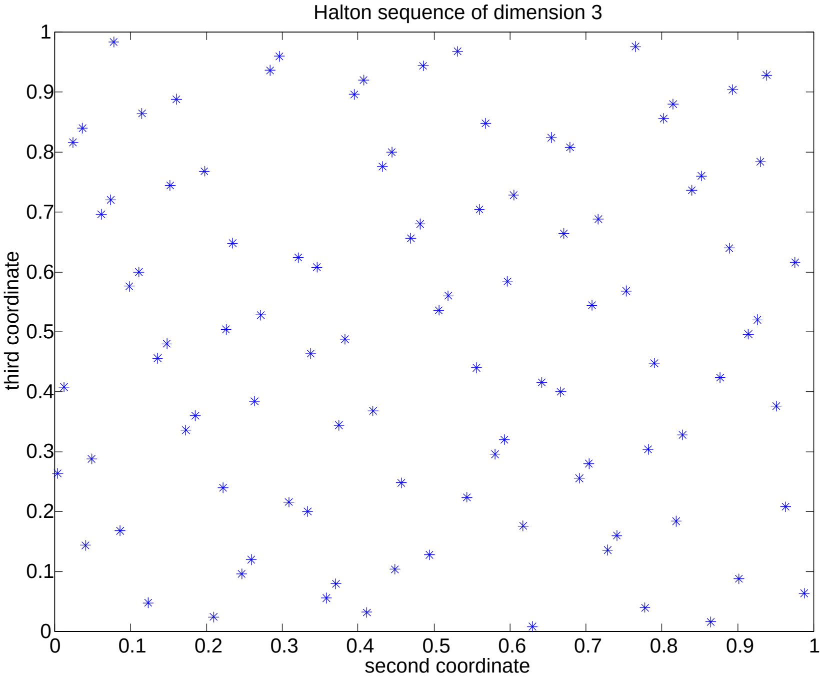

Halton sequence of dimension 3
third coordinate
second coordinate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4.3:

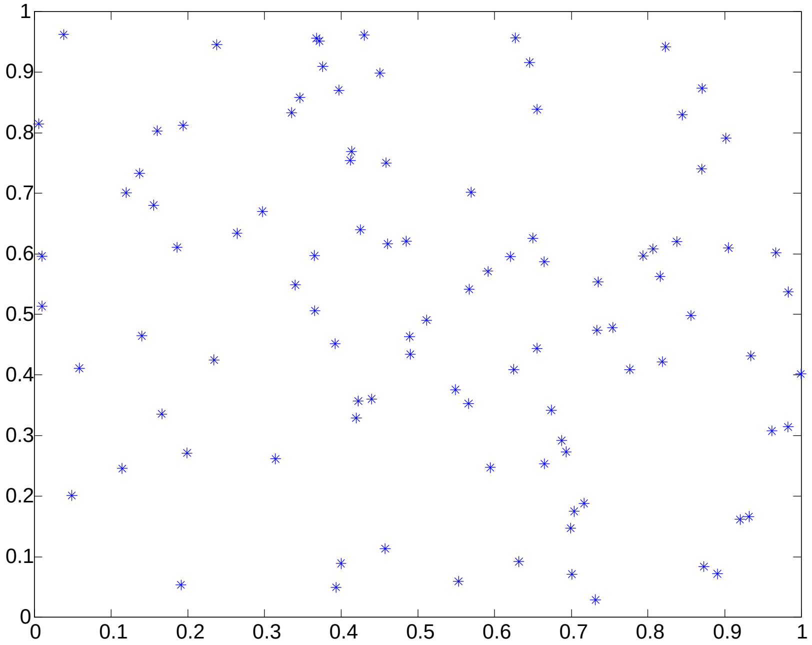


Figure 4.4: 100 independent $U[0,1]$ pairs

```
end
x=((1./b').^(1:(m+1)))*A;
```

The Halton sequence is reasonably uniform for small dimensions, but it is easy to see that if s is large, the uniformity degrades rapidly. The performance is enhanced by permuting the coefficients a_k prior to mapping into the unit interval. The **Faure** sequence is obtained in this way. It is similar to the Sobol's sequence below in that each dimension is a permutation of a van der Corput sequence; however, the prime used for the base is chosen as the smallest prime greater than or equal to the dimension (B.L. Fox, 1996, ACM Trans. Math. Software). Other suggestions for permuting the digits in a Halton sequence include using only every $l'th$ term in the sequence so as to destroy the cycle.

In practice, in order to determine the effect of using one of these low discrepancy sequences we need only substitute such a sequence for the vector of independent uniform random numbers used by a simulation. For example if we wished to simulate a process for 10 time periods, then value a call option and average the results, we could replace the 10 independent uniform random numbers that we used to generate one path by an element of the Halton sequence with $s = 10$.

Suppose we return briefly to the call option example treated in Chapter 3. The true value of this call option was 0.4615 according to the Black-Scholes formula. If however we substitute the van den Corput sequence for the sequence of uniform random numbers,

```
mean(fn(corput(100000,2)))
```

we obtain an estimate of 0.4614 very close to the correct value.

4.2.7 Sobol Sequence

The Sobol sequence is generated such that the first 2^m terms of each dimension for $m = 0, 1, \ldots$ are a permutation of the corresponding terms of the van der Corput sequence (P. Bratley and B.L. Fox, 1998, ACM Trans. Math. Software). We begin with a set of direction numbers $v_i = \frac{m_i}{2^i}, i = 1, 2,$ where the m_i are odd positive integers less than 2^i. The values of m_i are chosen to satisfy a recurrence relation using the coefficients of a primitive polynomial in the Galois Field of order 2. For example corresponding to a primitive polynomial

$$z^p + c_1 z^{p-1} + ...c_{p-1} z + c_p$$

is the recursion

$$m_i = 2c_1 m_{i-1} + 2^2 c_2 m_{i-2} + ... + 2^p c_p m_{i-p}$$

where the addition is carried out using binary arithmetic. For the Sobol sequence, we then replace the binary digit a_k by $a_k v_k$.

The Sobel and Faure sequences are particular cases of $(t, s) - nets$. In order to define then we need the concept of an elementary interval.

4.2.8 Definition: elementary interval

An elementary interval in base b is n interval E in I^s of the form

$$E = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \frac{(a_j+1)}{b^{d_j}} \right), \tag{4.11}$$

with $d_j \geq 0$, $0 \leq a_j \leq b^{d_j}$ and a_j, d_j are integers.

4.2.9 Definition: (t, m, s) - net

Let $0 \leq t \leq m$ be integers. A $(t.m.s)$ - net in base b is a finite sequence with b^m points from I^s such that every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence.

4.2.10 Definition: (t, s) - sequence

An infinite sequence of points $\{\mathbf{x_i}\} \in I^s$ is a (t,s)-sequence in base b if for all $k \geq 0$ and $m > t$, the finite sequence $\mathbf{x_{kb^m}}, \ldots, \mathbf{x_{(k+1)b^{m-1}}}$ forms a (t,m,s) - net in base b.

It is known that for a (t, s)-sequence in base b the low discrepancy is ensured:

$$D_N^* \leq C\frac{(\log N)^s}{N} + O(\frac{(\log N)^{s-1}}{N}). \tag{4.12}$$

Special constructions of such sequences for $s \geq 2$ have the smallest discrepancy that is currently known (H. Niederreiter, 1992, *Random Number Generation and Quasi-Monte Carlo Methods*).

The thesis of K.S. Tang (1998) provides a thorough investigation into various improvements in Quasi-Monte Carlo sampling, as well as the evidence of the high efficiency of these methods when valuing Rainbow Options in high dimensions. Papageorgiou and Traub (1996) tested what Tezuka called generalized Faure points. They concluded that these points were superior to Sobol points for the model problem. Particularly important for financial computation, a reasonably small error could be achieved with few evaluations. For example, just 170 generalized Faure points were sufficient to achieve an error of less than one part in a hundred for a 360 dimensional problem. See also Traub and Wozniakowski, (1994) and Paskov and Traub,(1995).

Chapter 5

Sensitivity Analysis, Estimating Derivatives and the Greeks.

Estimating the sensitivity of a simulation with respect to changes in the parameter values is an important part of establishing the validity of the conclusions. If a simulation estimates an expected value at certain value of the parameters with 0.32 ± 0.05 but the derivative with respect to one parameter, say the volatility parameter σ, is 5, this indicates that a change of the volatility of only 0.02 or 2 percent would result in a change in the average of the order of 0.1. Since volatility typically changes rapidly by far more than one percent, then the apparent precision of the estimate $0.32 \pm .005$ is very misleading.

Of particular importance in finance are certain derivatives of an option price or portfolio value with respect to the parameters underlying the Black Scholes model. These are called the "Greeks", because many of them (not to mention many parameters and constants used in Statistics, Physics, Mathematics, and the rest of Science) are denoted by greek letter. Suppose V denotes the value of a portfolio based on an asset $S(t)$ whose volatility parameter is σ when the current spot interest rate is r. Then if $V = V(S(t), t, \sigma, r)$, the most important derivatives are;

Name	Symbol		Value in BS model
Delta	Δ	$\frac{\partial V}{\partial S}$	$\Phi(d_1)$
Gamma	Γ	$\frac{\partial^2 V}{\partial S^2}$	$\frac{\phi(d_1)}{s\sigma\sqrt{T-t}}$
rho	ρ	$\frac{\partial V}{\partial r}$	$K(T-t)e^{-r(T-t)}\Phi(d_2)$
Theta	Θ	$\frac{\partial V}{\partial t}$	$\frac{s\sigma\phi(d_1)}{2\sqrt{T-t}} - rKe^{-r(T-t)}\Phi(d_2)$
Vega	$\mathcal{V}$	$\frac{\partial V}{\partial \sigma}$	$s\phi(d_1)\sqrt{T-t}$

In some cases there are analytic formulae for these quantities and for a

European call option in the Black-Scholes model, these formulae are given above where

$$\begin{aligned} d_1 &= \frac{1}{\sigma\sqrt{T-t}}\{\ln(\frac{s}{K})+(r+\frac{\sigma^2}{2})(T-t)\} \\ d_2 &= d_1-\sigma\sqrt{T-t} \end{aligned}$$

and ϕ,Φ are the standard normal probability density function and cumulative distribution function respectively. These derivatives are calculated typically not only because they are relevant to a hedging strategy (especially Δ and Γ) but also because they give an idea as to how rapidly the value of our portfolio is effected when there is an adverse change in one of the parameters.

As an example of the use of these derivatives, it is common to *immunize* a portfolio against changes in one or more parameters. For example suppose I own a portfolio whose value $P(S,t)$ depends on the price of a stock or index S. I wish to immunize this portfolio against changes in S by investing directly in the stock S and in an option on this stock whose value is given by $V(S,t)$ at time t. Suppose I add to my portfolio x_S units of the stock and x_O units of the option so that the value of the new portfolio is

$$P(S,t)+x_SS+x_oV(S,t).$$

In order to ensure that this value changes as little as possible when S changes, set the value of its delta and gamma (first and second derivative with respect to S) equal to zero. This gives two equations in the two unknown values of x_s,x_o.

$$\begin{aligned} \Delta_P+x_s+x_0\Delta_o &= 0 \\ \Gamma_P+x_o\Gamma_o &= 0 \end{aligned}$$

where Δ_P,Δ_0 are the deltas for the original portfolio and option respectively and Γ_P,Γ_o are the gammas. The solution gives

$$\begin{aligned} x_o &= -\frac{\Gamma_P}{\Gamma_o} \\ x_s &= \Delta_o\frac{\Gamma_P}{\Gamma_o}-\Delta_P \end{aligned}$$

and the hedged portfolio has value

$$P(S,t)+(\Delta_o\frac{\Gamma_P}{\Gamma_o}-\Delta_P)S-(\frac{\Gamma_P}{\Gamma_o})V(S,t).$$

The availability of two instruments, the stock and a single option on the underlying S allow us to adjust a portfolio so that the first two derivatives of its value function with respect to S are both zero. The portfolio is therefore protected against reasonably small changes in S. Similarly, with more options on the same stock, one could arrange that the portfolio is immunized or protected against adverse movements in the other parameters as well, including the interest rate and the volatility parameter. This hedged portfolio clearly requires derivatives of the value function, and for more complicated models than the Black-Scholes, we require simulation methods not only for valuing options and portfolios, but also for determining these derivatives with respect to underlying parameters.

Consider now an important question in stress or sensitivity testing, the problem of estimating an expected value at many different values of an underlying parameter. One very surprising feature of importance sampling is that simulations conducted at one value of a parameter θ can also be used to estimate and expected value corresponding to *all other values of the parameter.* The estimation of an expectation under one value of a parameter using simulations conducted at another is sometimes called the "what if" problem. Denote the probability density function of a random variable or vector X under θ by $f_\theta(x)$ and assume these densities all have common support. An expectation calculated under this value of the parameter will be denoted $E_\theta(.)$. If we want to estimate the expected value of a statistic $V(X)$, under different values ψ of the parameter note that

$$m(\psi) \;=\; E_\psi V(X) \;=\; E_\theta[V(X)f_\psi(X)/f_\theta(X)]. \tag{5.1}$$

There may be many reasons for our interest in the function $m(\psi)$. A derivative is priced using current values for the asset price, interest rate, volatility parameter etc. and we may wish to graph the price over a range of (possible future) values of the parameters. The necessity for estimating derivatives in order to immunize of hedge a portfolio is discussed above. The *likelihood ratio estimator* $V(X)f_\psi(X)/f_\theta(X)$ where $X \sim f_\theta$ is an unbiased (importance sample) estimator of $m(\psi)$. This means that *a simulation at θ permits unbiased estimation of the whole function $m(\psi) \;=\; E_\psi T(X)$, and thereby also its derivatives.*

However, this simple result must be tempered by a study of the precision of this estimator. For θ the true value of the parameter (so X is generated under θ) and for $\psi \neq \theta$, the likelihood ratios $f_\psi(X)/f_\theta(X) \to 0$ with probability one as the sample size (usually the dimension of the observation vector X) $n \to \infty$. This means that the likelihood is very much smaller for some value of the parameter ψ far from the true value than it is at the true value. This would seem to imply that for large sample sizes, the function $f_\psi(X)/f_\theta(X)$ is very close to zero and so the expectation $E_\theta[V(X)f_\psi(X)/f_\theta(X)]$. should be close to zero. However, if we substitute $V(X) = 1$, then the expectation $E_\theta[f_\psi(X)/f_\theta(X)] \;=\; 1$ for all ψ and all n. So for large n we have found a random variable $f_\psi(X)/f_\theta(X)$ which is very close to 0 and indeed converges to 0 as $n \to \infty$, and yet its expected value remains 1 for all n. This apparent contradiction is resolved upon recalling that the likelihood ratios are not uniformly integrable, and do not behave in the limit as $n \to \infty$ in the same way in probability as they do in expectation. This likelihood ratios for large n are close to 0 with high probability but take larger and larger values with decreasingly small probabilities.

This means that for fairly large sample sizes, the likelihood ratio $f_\psi(X)/f_\theta(X)$ is rather unstable. It takes very large values over a small range of values of

X and is close to zero for others and consequently it has very large variance. In fact as the sample size $\to\infty$, the variance of the likelihood ratio $var_\theta(f_\psi(X)/f_\theta(X))\to\infty$ very quickly. The process of "averaging" in such a situation takes a long time to approximate an expected value. This argues against the use of the likelihood ratios as suggested above, at least for large sample sizes, since moment-type estimators based on these likelihood ratios will tend to be very unstable in mean and variance, particularly when ψ is far from θ. This problem may be partially alleviated if variance reduction or alternative techniques are employed.

5.1 Estimating Derivatives.

Let us begin by examining the estimation of the derivative $m'(\theta)=\frac{\partial}{\partial\theta}E_\theta V(X)$ in general when we are only able to evaluate the function $V(X)$ by simulation, so there is error in its valuation. We could conduct independent simulations at two different values of the parameters, say at $\theta+h$, $\theta-h$, average the values of $V(X)$ under each, resulting say in the estimators $\hat{m}(\theta+h)$ and $\hat{m}(\theta-h)$ and then take as our estimator the difference

$$\frac{\hat{m}(\theta+h)-\hat{m}(\theta-h)}{2h} \tag{5.2}$$

but this crude estimator suffers from a number of disadvantages;

- It requires twice as many simulations as we conduct at a single point.
- It is heavily biased if h is large unless the function $m(\theta)$ is close to linear.
- It has very large variance it h is small.

Now we have seen some methods for ameliorating the last of these problems. Since we are estimating a difference, use of common random numbers in the simulations at the two parameter values $\theta+h$ and $\theta-h$ should reduce the variability somewhat, but this still leaves open the problem of estimating the derivative, essentially the limit of such a slope estimate.

5.1.1 The Score Function Estimator.

There are two alternatives that are popularly used, *Perturbation Analysis*, which depends on pathwise differentiation, and the *score function* or *Likelihood ratio* method. Both have the advantage that a simulation at a single parameter value allows estimation of the function and its derivative both. We begin by introducing the *score function method.* The idea behind the score function method is very simple, and it involves interchanging derivative and integral. We wish to estimate $m'(\theta)=\frac{\partial}{\partial\theta}\int V(x)f_\theta(x)dx$ and under some regularity conditions called the *Cramér conditions* we may interchange the integral and derivative (for example as required by the Cramer-Rao inequality)

$$\begin{aligned} m'(\theta) &= \frac{\partial}{\partial\theta}\int V(x)f_\theta(x)dx \\ &= \int V(x)\frac{\partial f_\theta(x)}{\partial\theta}dx \;=\; E_\theta[V(X)S(\theta)] \end{aligned} \tag{5.3}$$

where $S(\theta)$ denotes the score function or

$$S(\theta) = S(\theta, x) = \frac{\partial ln[f_\theta(x)]}{\partial\theta}. \tag{5.4}$$

Since the score function has expected value 0, i.e. $E_\theta S(\theta) = 0$, the quantity $E_\theta[V(X)S(\theta)]$ is just the covariance between $V(X)$ and $S(\theta)$ and this can be estimated using the sample covariance. In particular if we have a total of n independent simulations at parameter value θ,

$$cov(V(\widehat{X), S}(\theta, X)) = \frac{1}{n}\sum_{i=1}^{n} V(X_i)S(\theta, X_i) - \frac{1}{n}\sum_{i=1}^{n} V(X_i)\frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i)$$

provides an estimator of the sensitivity $\frac{\partial}{\partial\theta}E_\theta V(X)$.

Example. A Monte-Carlo Estimator of rho.

Suppose are interested in the ρ for an option with payoff function at maturity given by $V(S(T), T)$. Assume the Black-Scholes model so that the distribution of $S(T)$ under the Q measure is lognormal with mean $\eta = S_0 \exp\{rT\}$ and volatility $\sigma\sqrt{T}$. For brevity we denote $S(T)$ by S. Then if S has the log-normal distribution with mean η, $S = e^Y$ where $Y \sim N(log(\eta) - \sigma^2 T/2,\ \sigma^2 T)$. Note that if g is the corresponding probability density function,

$$\begin{aligned} \frac{\partial log(g)}{\partial\eta} &= \frac{Y - log(\eta) + \sigma^2 T/2}{\eta\sigma^2 T} \\ \frac{\partial log(g)}{\partial r} &= \frac{Y - log(\eta) + \sigma^2 T/2}{\eta\sigma^2 T}\frac{\partial\eta}{\partial r} \\ &= \frac{\log(S/S_0) - rT + \sigma^2 T/2}{\sigma^2} \end{aligned} \tag{5.5}$$

Thus an estimator of ρ can be obtained from the sample covariance, over a large number of simulations, of the values of $V(S,T)$ and $\frac{\partial log(g)}{\partial r}$ or equivalently the sample covariance between $V(S,T)$ and $\sigma^{-2}\log(S/S_0)$.

This score function estimator can be expressed as a limit of likelihood ratio estimators However, the score function is more stable than is the likelihood ratio for large sample size because its moment behaviour is, unlike that of the likelihood ratio, similar to its behaviour in probability. Under the standard regularity conditions referred to above, the score function $S(\theta) = S_n(\theta)$ for an independent sample of size n satisfies a law of large numbers

$$\frac{1}{n} S_n(\theta) \rightarrow E[S_1(\theta)] = 0 \tag{5.6}$$

and a central limit theorem;

$$\frac{1}{\sqrt{n}} S_n(\theta) \rightarrow N(0, J_1(\theta)) \tag{5.7}$$

in distribution where the limiting variance $J_1(\theta) = var[S_1(\theta)]$. When the dimension of X is high, however, the score function estimator still suffers from too much variability.

Among all random functions $G(X;\theta)$ which satisfy $\frac{\partial}{\partial\theta} E_\theta V(X) = E_\theta[(V(X)G(X;\theta)]$ for all V, the score function cannot be improved on in the sense that it has the smallest possible variance.

Conditioning the Score Function Estimator.

Note that

$$m\prime(\theta) = E_\theta[V(X)S(\theta)] = E_\theta\{E_\theta[V(X)|S(\theta)]S(\theta)\} \tag{5.8}$$

The conditional expectation $E_\theta[V(X)|S(\theta)]$ in the above product is to be estimated by Monte-Carlo provided that we are able to generate the variates conditional on the value of the score function. The outside integral $E_\theta\{.\}$over the distribution of $S(\theta)$ may be conducted either analytically or numerically, using our knowledge of the asymptotic distribution of the score function.

For brevity, denote $S(\theta)$ by S and its marginal probability density function by $f_S(s)$.

Let $X_{si}, i = 1, ...n$ be variates all generated with the **conditional** distribution of X given $S = s$ for the fixed parameter θ. Then based on a sample of size n, the suggested estimator is:

$$\int (\frac{1}{n} \sum_{i=1}^{n} V(X_{si})) s f_S(s) ds \tag{5.9}$$

There are some powerful advantages to (**??**), particularly when the data is generated from one of the distributions in an exponential family. The exponential family of distributions is a broad class which includes most well-known continuous and discrete distribution families such as the normal, lognormal, exponential, gamma, binomial, negative binomial, geometric, and Poisson distributions.

X_1 is said to have an *exponential family distribution* if its density with respect to some dominating measure (usually a counting measure or Lebesgue measure) takes the form:

$$f_\theta(x_1) \ = \ e^{\eta(\theta)Y(x_1)}h(x_1)c(\theta)$$

for some functions $\eta(\theta)$, $c(\theta)$, $Y(x_1)$ and $h(x_1)$.

When the input consists of a random sample of size n from such an exponential family distribution, the statistic $Y_n \ = \ \sum_{i=1}^n Y(X_i)$ has a distribution also of the exponential family form and is *sufficient* for the family of distributions. By this we mean that the *conditional distribution of* $(X_1, \ . \ . \ . \ X_n)$ *given the statistic* Y_n *is independent of the parameter* θ. Furthermore, provided $\eta'(\theta) \neq 0$, conditioning on the score function is equivalent to conditioning on Y_n . The score function is always a function of the sufficient statistic. Suppose we denote it by $S(Y_n, \theta)$. Thus, denoting a conditional variate X given $Y_n = y$ by X_y , we may estimate $m'(\theta)$ using

$$\widehat{m\prime(\theta)} = \int E[V(X)|Y_n = y]S(y,\theta)G_n(dy) = \int V(X_y)S(y,\theta)G_n(dy) \qquad (5.10)$$

where G_n is the distribution of the sufficient statistic Y_n . For general sample size, $V(X_y)$ in the integrand is replaced by an average of the terms of the form $V(X_y)$. Similarly, we estimate $m(\psi)$ using simulations at the parameter value θ by

$$\widehat{m(\psi)} \ = \ \int V(X_y)e^{y(\eta(\psi)-\eta(\theta))}\frac{c(\psi)}{c(\theta)}G_n(dy). \qquad (5.11)$$

When we are attempting to estimate derivatives $m'(\theta)$ simultaneously at a number of different values of θ , perhaps in order to fit the function with splines or represent it graphically, there are some very considerable advantages to the estimator underlying (**??**). Because the conditional expectation does not depend on the value of θ , we may conduct the simulation (usually at two or more values of t) at a single convenient θ . The estimated conditional expectation will be then used in an integral of the form (**??**) for all underlying values of θ. Similarly, a single simulation can be used to estimate $m(\psi)$ for many different values of ψ .

There are a number of simple special cases of exponential family where the conditional distributions are easily established and simulated. Note that the variables can be generated sequentially beginning with X_1 so the following distributional results are adequate.

5.1.2 Example.

1. **(Exponential Distribution).** Suppose X_i are exponentially distributed with probability density function $f_\theta(x) \ = \ \frac{1}{\theta}e^{-x/\theta}$. Then given $\sum_{i=1}^n X_i = y$ the values $X_1, \ X_1 + X_2, \ . \ . \ . \ \sum_{i=1}^{n-1} X_i$ are distributed as $n-1$ Uniform $[0, y]$ order statistics.

2. **(Gamma distribution).** Suppose X_i are distributed as independent gamma (α, θ) variates with probability density function

$$f_\theta(x) = \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}. \tag{5.12}$$

Then the distribution of X_1/y given $\sum_{i=1}^n X_i = y$ has the Beta $(\alpha, n\alpha)$ distribution.

3. **(Normal distribution).** Suppose X_i have a $N(\theta, \sigma^2)$ distribution. Then the distribution of X_1 given $\sum_i X_i = y$ is $N(y/n,\ (1-\frac{1}{n})\sigma^2)$.

4. **(Binomial distribution).** Suppose X_i are distributed as binomial (n, θ) variates. Then given $\sum_{i=1}^m X_i = y$, X_1 has a hypergeometric distribution with parameters (mn, n, y).

5. **(Poisson distribution).** Suppose X_i have the Poisson (θ) distribution. Then given $\sum_{i=1}^n X_i = y$, the distribution of X_1 is binomial $(y, 1/n)$.

6. **(Geometric distribution).** Suppose X_i have the geometric distribution. Then given $\sum_{i=1}^n X_i = y$, the distribution of X_1 is a negative hypergeometric with probability function

$$f(x) = \frac{\binom{y-x-1}{n-2}}{\binom{y-1}{n-1}}.$$

7. **(log-Normal Distribution)** Suppose X_i have the log-normal distribution with mean η. Recall that $X_i = e^{Y_i}$ where $Y_i \sim N(log(\eta) - \sigma^2/2,\ \sigma^2)$. Note that

$$\frac{\partial log(g)}{\partial \eta} = \frac{Y - log(\eta) + \sigma^2/2}{\eta\sigma^2} \tag{5.13}$$

and therefore the conditional distribution of X_1 given the sufficient statistic $\prod_{i=1}^n X_i = x$ is of the form e^{Z_i} where $Z_i \sim N(log(x)/n,\ (1-\frac{1}{n})\sigma^2)$ and so Y_i has conditional mean $x^{1/n}exp\{(1-\frac{1}{n})\frac{\sigma^2}{2}\}$ and volatility parameter $(1-\frac{1}{n})\sigma^2$.

5.1.3 Example. Estimating Vega.

Suppose we wish to estimate $\frac{\partial V}{\partial \sigma}$ where V is the value of an option on an asset, and σ is the volatility parameter in the asset price equation. Consider for example a European option whose present value can be written as an expected value under the risk neutral distribution

$$V = E\{e^{-rT}v(S_T)\}$$

where S_T, the terminal value of the asset, is assumed to have a lognormal distribution with mean $S_0 e^{rT}$ and variance parameter $\sigma^2 T$. Denote the corresponding lognormal probability density function of $S(t)$ by

$$g(s) = \frac{1}{s\sigma\sqrt{2\pi T}} exp\{-(log(s) - log(S_0) - rT + \sigma^2 T/2)^2/2\sigma^2 T\}.$$

For brevity, write

$$g(s) = \frac{1}{s\sigma\sqrt{2\pi T}} exp\{-R^2/2\sigma^2 T\}$$

where we have denoted $R = R(s) = (log(s) - rT - log(S_0) + \sigma^2 T/2)$. Then the score function with respect to the parameter σ^2 is

$$\frac{\partial \log(g)}{\partial \sigma} = \frac{R^2 - \sigma^2 T}{\sigma^3 T} - \frac{R}{\sigma}$$

Therefore, by (**??**), an unbiased estimator of *vega* is the sample covariance, over all simulations,

$$\widehat{cov}(e^{-rT} v(S_T), \frac{R^2(S_T) - \sigma^2 T}{\sigma^3 T} - \frac{R(S_T)}{\sigma}).$$

Notice that $\frac{R(S_T)}{\sigma\sqrt{T}} = Z$ has the standard normal distribution, and in terms of Z ,

$$S_T = S_0 exp\{rT - \sigma^2 T/2 + \sigma\sqrt{T}Z\}. \tag{5.14}$$

Then the covariance in (**??**) is

$$E\{e^{-rT} v(S_0 exp\{rT - \sigma^2 T/2 + \sigma\sqrt{T}Z\}) \left[\frac{Z^2 - 1}{\sigma} - \sqrt{T}Z\right]\}$$

since S_T is generated from (**??**). This reduces to a simple one-dimensional integral with respect to a normal probability density function and we can either simulate this quantity or use a numerical integration. Because of the high variability of the score function, it is desirable to use variance reduction in evaluating this estimator.One of the simplest numerical integration techniques when expectation can be written with respect to a normal probability density function is *Gaussian Quadrature* mentioned below.

5.1.4 Gaussian Quadrature.

We consider general integrals of the form

$$\int_{-\infty}^{\infty} h(t)\phi(t)dt \tag{5.15}$$

where ϕ is the standard normal probability density function. Suppose such an integral is approximated by a weighted sum,

$$\sum_{i=1}^{k} w_i h(t_i). \tag{5.16}$$

The weights, w_i, are chosen so that the approximation is exact for the first k moments of the standard normal distribution. In other words, it is required that

$$\sum_{i=1}^{k} w_i t_i^r = \mu_r, \; r = 0, 1, \ldots, k-1 \tag{5.17}$$

where $\mu_r = 0$ for r odd and $\mu_r = r!/(r/2)!2^{r/2}$ for r even. For arbitrary t_i these weights give an approximation that is exact for polynomials of degree at most $k-1$. However, if we are free not only to choose the weights, but also the points t_i , we can achieve a better approximation, one that is exact for polynomials of degree $2k-1$. In this case, we must choose the abscissae to be the k roots of the *Hermite polynomials* ,

$$p_k(x) = (-1)^k [\phi(x)]^{-1} \frac{d^k \phi(x)}{dx^k}, \tag{5.18}$$

to give an approximation which is exact for polynomials of degree $2k-1$. The Hermite polynomials with degree $k \leq 5$ are:

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - 1, \quad p_3(x) = x^3 - 3x,$$

$$p_4(x) = x^4 - 6x^2 - 3, \quad p_5(x) = x^5 - 10x^3 - 15x.$$

Finding the roots of these polynomials, and solving for the appropriate weights, the corresponding approximations to the integral $\int_{-\infty}^{\infty} h(t)\phi(t)dt$ are:

$$\begin{aligned} \int_{-\infty}^{\infty} h(t)\phi(t)dt \approx (1/2)h(-1) + (1/2)h(1), \quad k &= 2 \\ \int_{-\infty}^{\infty} h(t)\phi(t)dt \approx (2/3)h(0) + (1/6)h(\pm\sqrt{3}), \quad k &= 3 \end{aligned}$$

In general, if we wish to evaluate the expected value of a function of $X \sim N(\mu, \sigma^2)$, the approximations are

$$\begin{aligned} E[h(X)] \approx (1/2)h(\mu - \sigma) + (1/2)h(\mu + \sigma), \quad k &= 2 \\ E[h(X)] \approx (2/3)h(\mu) + (1/6)h(\mu + \sqrt{3}\sigma) + (1/6)h(\mu - \sqrt{3}\sigma), \quad k &= 3 \end{aligned}$$

This last formula is exact for h a polynomial of degree 5.

5.2 Infinitesimal Perturbation Analysis: Pathwise differentiation.

There is an alternate method for sensitivity analysis which often competes favourably with the score function method above and which exploits information on the derivative of the performance measure. As a preliminary example, let us return to the problem of estimating a greek (e.g. rho, vega, delta or theta) for a European option. In this case, we wish to estimate the derivative of the option price $V = E\{e^{-r(T-t)}v(S_T)\}$ with respect to some parameter (e.g. r, σ, S_0, t) where S_T has a lognormal distribution with mean $S_t e^{r(T-t)}$ and variance parameter $\sigma^2(T-t)$, and $v(S_T)$ is the value of the option on expiry when the stock price is S_T. Call the parameter θ for the present. Suppose we generate $S_T = S_0 \exp(r(T-t) - \sigma^2(T-t)/2 + \sigma Z\sqrt{T-t})$ for a standard normal random variable Z. Then differentiating directly with respect to the parameter provided such a derivative exists and can be moved under the expectation sign, yields

$$\frac{\partial}{\partial\theta}E\{e^{-rT}v(S_T)\} = e^{-rT}v'(S_T)\frac{\partial}{\partial\theta}S_T$$

Thus, to estimate the derivative, an average of simulated values of the form

$$\frac{1}{n}\sum_{i=1}^{n}[e^{-rT}v'(S_{Ti})\frac{\partial}{\partial\theta}S_{Ti}] \tag{5.19}$$

where $S_{Ti} = S_0\exp(rT - \sigma^2 T/2 + \sigma\sqrt{T}Z_i)$ is the i'th simulated closing value. In fact if the function $v(.)$ is close to being constant, then this estimator will have variance close to zero and will be quite accurate, likely more accurate than the score function estimator described in the last section. Consider the case of a European call option with strike K, $v(S_T) = (S_T - K)^+$ and $v'(S_T) = 1_{[S_T > K]}$. Note that the derivative exists everywhere except at the point K, but the derivative at the point $S_T = K$ does not exist. To see if we can circumvent this problem, can we find a sequence of everywhere differentiable functions $v_n(x)$ such that $v_n(x) \to v(x)$ and $v_n'(x) \to v'(x)$ for all $x \neq k$? If so then we can show that with v_n replacing v in (**??**), we obtain a consistent estimator of $\frac{\partial}{\partial\sigma}E\{e^{-rT}v_n(S_T)\}$ and then using the Lebesgue dominated convergence theorem, we may carry this consistency over to $v(x)$. In this case, we might choose

$$v_n(x) = \begin{cases} n(x - K + \frac{1}{4n})^2, & \text{for } K - \frac{1}{4n} < x < K + \frac{1}{4n} \\ (x-K), & \text{for } x > K + \frac{1}{4n} \end{cases}$$

and $v_n(x) = 0$ for $x < K - \frac{1}{4n}$, a continuously differentiable function which agrees with $v(x)$ both in its value and its derivative everywhere except in the diminishing interval $(K - \frac{1}{4n}, K + \frac{1}{4n})$. More generally when $v(x)$ increases at most linearly in x as $x \to \infty$, it is possible to find a dominating function, but if the payoff function $v(x)$ increased at a faster rate, this may not be possible.

So generally speaking if there are a finite number of points where the derivative does not exist, and the payoff function is bounded above by linear functions of the stock price, an estimator of the form (**??**) can be used.

In general the suggested method corresponds to the foillowing simple steps;

1. write the expected value we wish to determine in terms of the parameters (as explicit arguments) and random variables whose distribution does not depend on these parameters (e.g. $U[0,1]$ or $N(0,1)$.) The simplest way to do this may be to use the inverse transform.
2. Differentiate this expected value with respect to the parameter of interest, passing the derivative under the expected value sign.
3. Simulate or numerically determine this expected value.

5.2.1 Example. IPA estimate of Vega.

Again consider an estimate of $\frac{\partial V}{\partial \sigma}$ where $t = 0$,

$$V = E\{e^{-rT}v(S_T)\},$$

and S_T, the terminal value of the asset, has a lognormal distribution with mean S_0e^{rT} and volatility parameter $\sigma^2 T$. We wish to write S_T in terms of random variables with distributions that do not depend on the parameters. Recall that

$$S_T = S_0 exp\{rT - \sigma^2 T/2 + \sigma\sqrt{T}Z\}$$

with Z a standard normal random variable. Then provided that we can pass the derivative through the expected value,

$$\begin{aligned}\frac{\partial V}{\partial \sigma} &= E\{e^{-rT}v'(S_T)\frac{\partial S_T}{\partial \sigma}\}\\ &= E\{e^{-rT}v'(S_T)S_T(\sqrt{T}Z - \sigma T)\}.\end{aligned}$$

This can be simulated by generating values of Z and then $S_T = S_0exp\{rT - \sigma^2T/2 + \sigma\sqrt{T}Z\}$ and averaging the values of $e^{-rT}v'(S_T)S_T(\sqrt{T}Z - \sigma T)$. Alternatively, since this is a one dimensional integral, we can integrate the function against the standard normal p.d.f.ϕ i.e.

$$e^{-rT}\int_{-\infty}^{\infty} v'(S_0e^{rT-\sigma^2T/2+\sigma\sqrt{T}z})S_0e^{rT-\sigma^2T/2+\sigma\sqrt{T}z}(\sqrt{T}z - \sigma T)\phi(z)dz.$$

Note the similarity between this estimator and the score function estimator in the same problem. The primary difference is that v' is multiplied by a linear function of Z in this case, but v by a quadratic function of Z in the case of the score function. In part because of the higher variability of the score function, the perturbation analysis estimator is substantially better at least for a standard call option. The following function was used to compare the estimators and their standard errors.

```
function [price,vega,SE]=estvega(Z,S0,sigma,r,T,K)
% two estimators of vega , vega(1)=score function estimator, v(2)=IPA estimator
SE(1),SE(2) their standard errors.
% v=payoff function, vprime is its derivative.
%Z=randn(1,n)  is a vector of standard normal
ST=S0*exp(r*T+sigma*sqrt(T)*Z-.5*sigma^2*T);
v=max(0,ST-K);
v1=exp(-r*T)*(v.*((Z.^2-1)/sigma-sqrt(T)*Z));
vprime=ones(1,length(Z)).*(ST>K);
v2=exp(-r*T)*(vprime.*ST.*(sqrt(T)*Z-sigma*T));
vega=[mean(v1) mean(v2)];
SE=sqrt([var(v1) var(v2)]/length(Z));
price=exp(-r*T)*mean(v);
```

For example the call [price,vega,SE]=estvega(randn(1,500000),10,.2,.1,.25,9) results in the price of a call option on a stock worth \$10 and with 3 months or one quarter of a year to maturity, interest rate $r = .05$, annual volatility 0.20. The estimated price is \$1.1653 and the two estimates of vega are 0.8835 and 0.9297 with standard errors 0.0238 and 0.0059 respectively. Since the ratio of variances is approximately 4, the IPA estimator is evidently about 16 times as efficient as is the score function estimator in this case, although even the score function estimator provides reasonable accuracy. Not all derivatives can be estimated as successfully using IPA however. For example if we are interested in the Gamma or second derivative of a European call option with respect to S_t, $v(S_T) = (S_T - K)^+$ and $v''(x) = 0$ for all $x \neq K$. Thus, if we are permitted to differentiate twice under the expected value in

$$V = E\{e^{-rT}v(S_T)\}$$

we obtain

$$\Gamma = e^{-rT}E[v''(S_T)\frac{\partial^2 S_T}{\partial S_0^2}] = 0$$

which is clearly incorrect. The problem in this case is that the regularity required for the second interchange of derivative and espectation fails.

The Multivariate Case.

We wish to generate $X = (X_1, . . . X_n)$ with independent components and let the cumulative distribution function and the probability density function of X_i be denoted $F_{i\theta}(x)$ and $f_{i\theta}$ respectively. One again we wish to estimate the sensitivity or derivative of the expected value

$$m(\theta) = E_\theta V(X_1, \ldots X_n, \theta)$$

with respect to the parameter θ for some function V. Notice that to allow for the most general situation, we permit θ to not only affect the distribution of the variables X_i but also in some cases be an argument of the function V. Suppose we generate the random variables X_i by inverse transform from a vector of n independent uniform variates U_i according to $X_i = F_{i\theta}^{-1}(U_i)$. Then note that $\frac{\partial X_i}{\partial \theta} = -\frac{1}{f_{i\theta}(X_i)}\frac{\partial F_{i\theta}(X_i)}{\partial \theta}$. Thus, with $V^{(i)} = \frac{\partial V(X,\theta)}{\partial X_i}$, and $V^{(\theta)} = \frac{\partial V(X,\theta)}{\partial \theta}$ we have, under conditions permitting the interchange of derivative and integral,

$$m'(\theta) = E\{\sum_i \frac{\partial V(X,\theta)}{\partial X_i}\frac{\partial X_i}{\partial \theta} + \frac{\partial V(X,\theta)}{\partial \theta}\}$$

$$= E\left[\frac{\partial V(X,\theta)}{\partial \theta} - \sum_i V^{(i)}(X,\theta)\frac{1}{f_{i\theta}(X_i)}\frac{\partial F_{i\theta}(X_i)}{\partial \theta}\right] \tag{5.20}$$

This suggests a Monte Carlo estimator, an average over all (independent) simulations of terms of the form

$$\text{average}\{V^{(\theta)}(X,\theta) - \sum_i \frac{V^{(i)}(X,\theta)}{f_{i\theta}(X_i)}\frac{\partial F_{i\theta}(X_i)}{\partial \theta}\} \tag{5.21}$$

The unbiased estimator (**??**) is called the *Infinitesimal perturbation analysis estimator (IPA)*. Unfortunately, the conditions permitting the required interchange of derivative and integral are not always met and so the estimator may in some cases be biased. See Cao (1987a) for some conditions. When the conditions are met, note the relationship between terms in the perturbation analysis estimator and the score function estimator, obtained by integration by parts:

$$E_\theta[V(X,\theta)\frac{\partial \log(f_{\theta i}(X_i))}{\partial \theta}] = E_\theta \int V(X_1,...x_i,...X_n,\theta)\frac{\partial}{\partial \theta} f_{i\theta}(x_i)dx_i$$

$$\begin{aligned} &= E_\theta V(X_1,...x_i,...X_n,\theta)\frac{\partial}{\partial \theta}F_{i\theta}(x_i)dx_i|_{-\infty}^{\infty} - E_\theta \int V^{(i)}(X_1,...x_i,...X_n,\theta)\frac{\partial}{\partial \theta}F_{i\theta}(x_i)dx_i \\ &= -E_\theta\{V^{(i)}(X,\theta)\frac{\partial F_{i\theta}(X_i)/\partial \theta}{f_{i\theta}(X_i)}\}. \end{aligned}$$

Notice that for nearly constant functions V, the gradient $V^{(i)}$ is close to zero and the perturbation analysis estimator has small variance. In general, when it is unbiased, it seems to provide greater efficiency than the crude score function estimator. On the other hand, the comparison is usually carried out in specific cases, and there seems to be no general reason why perturbation analysis should be preferred. The lack of differentiability of payoff functions V can be a problem, introducing potential bias into perturbation analysis estimators. The infinitesimal perturbation analysis estimator is an infinitesimal or limiting version of the use of common random numbers as the following argument shows. Generating $X_{i\theta}$ as above, it is reasonable to estimate

$$\frac{m(\theta+\delta) - m(\theta-\delta)}{2\delta} \approx \frac{V(X_{\theta+\delta},\theta+\delta) - V(X_{\theta-\delta},\theta-\delta)}{2\delta}.$$

Taking limits as $\delta \to 0$ and assuming the gradient exists in a neighborhood of θ we arrive at the perturbation analysis estimator.

Infinitesimal perturbation analysis (IPA) assumes that the order of events in the perturbed path is the same as the order in the nominal path for a small enough δ, allowing a calculation of $V(x, \theta)$, the sensitivity of the sample performance for a particular simulation. It will generally give satisfactory results for European options, but may fail in simulations of lookback or barrier options if common random numbers fail to give payoffs that are close when parameter values are close.

In the more common circumstance that the function V does not directly depend on the parameter, the crude Monte Carlo IPA estimator (**??**) is an average over all (independent) simulations

$$-\sum_i \frac{\partial}{\partial X_i} V(X) \frac{\partial F_{i\theta}(X_i)/\partial\theta}{f_{i\theta}(X_i)} \tag{5.22}$$

where the derivatives of $\frac{\partial}{\partial X_i} V(X)$ may be derived through analysis of the system or through the implicit function theorem if the problem is tractable. In examples where IPA has been found to be unbiased, it has also been found to be consistent. When compared to the crude score function method for these examples, it has generally been found to be the more efficient of the two, although exceptions to this rule are easy to find.

IPA is based on the differentiation of the output process. Because of this, the conditions required for the exchange of the differentiation and expectation operators must be verified for each application. This makes IPA unsuitable as a "black-box" algorithm. By contrast, the score function method, together with its variance reduced variations, only impose regularity on the input variables and require no knowledge of the process being simulated. On the other hand, the score function method requires that the parameter whose sensitivity is investigated be a statistical parameter; i.e. index a family of densities, whereas perturbation analysis allows more general types of parameters.

5.2.2 Sensitivity of the value of a spread option to the correlation.

Consider two stocks or asset prices with closing values $S_1(T)$ and $S_2(T)$ jointly lognormally distributed with volatility parameters σ_1, σ_2, and correlation ρ. Of course all of the parameters governing this distribution are subject to change in the market, including the correlation ρ. We are interested in the price of a European call option on the spread in price between the two stocks, and in particular, the sensitivity of this price to changes in the correlation. Let the payoff function be

$$\begin{aligned} v(S_1(T), S_2(T)) &= \max(0, (S_1(T) - S_2(T) - K)) \qquad (5.23)\\ &= \max(0, [\exp\{rT - \sigma_1^2 T/2 + \sigma_1 Z_1\} - \exp\{rT - \sigma_2^2 T/2 + \sigma_2 Z_2\} - K]) \end{aligned}$$

for strike price K and correlated standard normal random variables Z_1, Z_2. Perhaps the easiest way to generate such random variables is to generate Z_1, Z_3 independent standard normal and then set

$$Z_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_3. \tag{5.24}$$

Then the sensitivity of the option price with respect to ρ is the derivative the discounted expected return

$$\frac{\partial}{\partial \rho} E[e^{-rT} v(S_1(T), S_2(T))] = E[-\sigma_2 \exp\{-\sigma_2^2 T/2 + \sigma_2 Z_2\} \frac{\partial}{\partial \rho} Z_2 I_A]$$

$$= E[-\sigma_2 \exp\{-\sigma_2^2 T/2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_3)\}(Z_1 - \frac{\rho}{\sqrt{1-\rho^2}} Z_3) 1_A] \tag{5.25}$$

where 1_A is the indicator function of the set

$$A = A(Z_1, Z_2) = [\exp\{rT - \sigma_1^2 T/2 + \sigma_1 Z_1\} - \exp\{rT - \sigma_2^2 T/2 + \sigma_2 Z_2\} > K]$$

and where Z_1, Z_3 are independent standard normal random variables and Z_2 satisfies (**??**). Thus an IPA estimator of the sensitivity is given by an average of terms of the form

$$-\sigma_2 \exp\{-\sigma_2^2 T/2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_3)\}(Z_1 - \frac{\rho}{\sqrt{1-\rho^2}} Z_3) 1_{A(Z_1, Z_2)}. \tag{5.26}$$

Of course variance reduction can be easily applied to this estimator, especially since there is a substantial set on which (**??**) is equal to 0.

5.3 Problems.

1. Assume that X has a normal $(\theta, 1)$ distribution and $V(X) = X + bX^2 + cX^3$. Show can estimate $\frac{\partial}{\partial \theta} E_\theta V(X) = 1 + 2b\theta + 3c(1+\theta^2)$ by randomly sampling n independent values of $X_i, i = 1, 2, ...n$ and using the estimator $\frac{1}{n}\sum_{i=1}^n V(X_i)(X_i - \theta)$. How would the variance of this compare with the variance of an alternative estimator $\frac{1}{n}\sum_{i=1}^n V'(X_i)$. How do they compare if V is close to being a linear function, i.e. if b, c are small?

Chapter 6

Estimation and Calibration.

6.1 Using Historical Data for Diffusion Models.

Typically a diffusion model for financial data includes parameters with unknown values which require estimation. For example the *CIR model* for interest rates, written in the form

$$dr_t = (\alpha + \beta r_t)dt + \sigma\sqrt{r_t}dW_t$$

has three unknown parameters that require estimation in order to use the model in valuing derivatives. According to the simplest discrete time approximation to the process, the *Euler Scheme*, the increments in the process over small intervals of time are approximately conditionally independent and normally distributed. Thus, approximately,

$$\Delta r - (\alpha + \beta r_t)\Delta t \sim N(0, \sigma^2 r_t \Delta t)$$

Thus, the parameters in the drift term can be obtained by weighted least squares; i.e. by minimizing the sum of the squared standardized normal variates.

$$min_{\alpha,\beta} w(\Delta r - (\alpha + \beta r_t)\Delta t)^2$$

where the weights w are proportional to the *reciprocal of the variances* $w = 1/(r_t\Delta t)$. The solution to this is standard in regression textbooks:

$$\hat{\beta} = \frac{\sum w\Delta r(r_t - \bar{r})}{\sum w(r_t - \bar{r})^2}, \quad \hat{\alpha} = \overline{\Delta r} - \hat{\beta}\bar{r} \tag{6.1}$$

where $\bar{r}$, $\overline{\Delta r}$ denote weighted averages; e.g. $\bar{r} = \sum wr_t / \sum w$.

Girsanov's Theorem allows us to use maximum likelihood estimation for any parameters that reside in the drift term of a diffusion. For example consider a model of the form

$$dX_t = a(X_t, \theta)dt + \sigma(X_t)dW_t$$

We suppose for the moment that the diffusion term $\sigma(X_t)$ is known and that the only unknown parameter(s) is θ. Then the Radon-Nykodym derivative of the measure induced by this process with respect to the corresponding martingale measure defined by

$$dX_t = \sigma(X_t)dW_t$$

is given (under the usual conditions) by Girsanov's Theorem

$$exp\{\int \frac{a(X_t,\theta)}{\sigma^2(X_t)}dX_t - \frac{1}{2}\int \frac{a^2(X_t,\theta)}{\sigma^2(X_t)}dt\}. \tag{6.2}$$

The maximum likelihood estimate of θ is obtained by maximizing this function. Setting the derivative of its logarithm equal to 0 results in the likelihood equation

$$\begin{aligned} \int \frac{\partial a}{\partial \theta}\sigma^{-2}(X_t)dX_t - \int \frac{a\partial a}{\partial \theta}\sigma^{-2}(X_t)dt &= 0 \\ or \int \sigma^{-2}(X_t)\frac{\partial a}{\partial \theta}(dX_t - a(X_t,\theta)dt) &= 0 \end{aligned}$$

Usually, of course, we have available only observations taken at discete time points $t_1 < t_2 < \ldots$ and the above integral will then be replaced by a sum

$$\sum_i \sigma^{-2}(X_{ti})\frac{\partial a(X_{ti},\theta)}{\partial \theta}(\Delta X_{t_i} - a(X_{t_i},\theta)\Delta t) = 0. \tag{6.3}$$

6.2 Estimating Volatility

While the Euler method permits estimation of parameters in the diffusion coefficient as well as those in the drift, there is no likelihood argument based on continuous time observations which allows estimation of diffusion coefficient parameters. This is because the infinite variation of a diffusion process in arbitrarily small time intervals for continuous time observations theoretically permit *exact* estimation of parameters in the diffusion coefficient with an arbitrarily short observed trajectory of the process. In other words information for estimating the diffusion coefficient obtains much more rapidly than for the drift, and in this respect the continuous time processes are quite different than their discrete analogues. Two diffusions (or even Brownian motion processes) with different diffusion coefficients are mutually singular so that we can theoretically determine from a single sample path the exact diffusion term. Practice is considerably different for several reasons. First, we never observe a process in continuous time. Second, processes like security prices, interest rates and exchange rates are only similar to certain diffusion processes when viewed over a longer time interval than a single day or week. Their local behaviour is very different; for example they evolve through a series of jumps of varying magnitudes. Third, there is information on any process for which derivatives are sold in the derivative market. The usual estimate of volatility is the "implied volatility" or the variance parameter which would, in the Black-Scholes formula, make the theoretical derivative prices equal to their observed values. While this is not exactly the historical volatility, it (provided that the model holds) is identical to the instantaneous value of the volatility since the risk neutral measure has the same diffusion coefficient as does the Ito process we assume derives the asset price.

Consider as an example the stock price of New Zealand Telecom listed on the New York Stock Exchange (Ticker NYSE:NZT). We downloaded three months of daily stock price data (Feb 22, 2000 to May 22, 2000) from the website http://finance.yahoo.com and on the basis of this, wish to estimate the volatility. The stock price over this period is graphed in Figure 6.1. It was downloaded to an excel file and loaded into *Matlab* from this file. Since the logarithm of daily stock prices is assumed to be a Brownian motion we may estimate the daily volatility using the sample variance of the first differences in these logarithms. To obtain the variance over a year, multiply by the number of trading days (around 252) in a year. Thus the annual volatility is estimated by sqrt(252*var(diff(log(telecomprice)))) which gives a value around 0.31. How does this historical estimate of volatility compare with the volatility as determined by option prices?

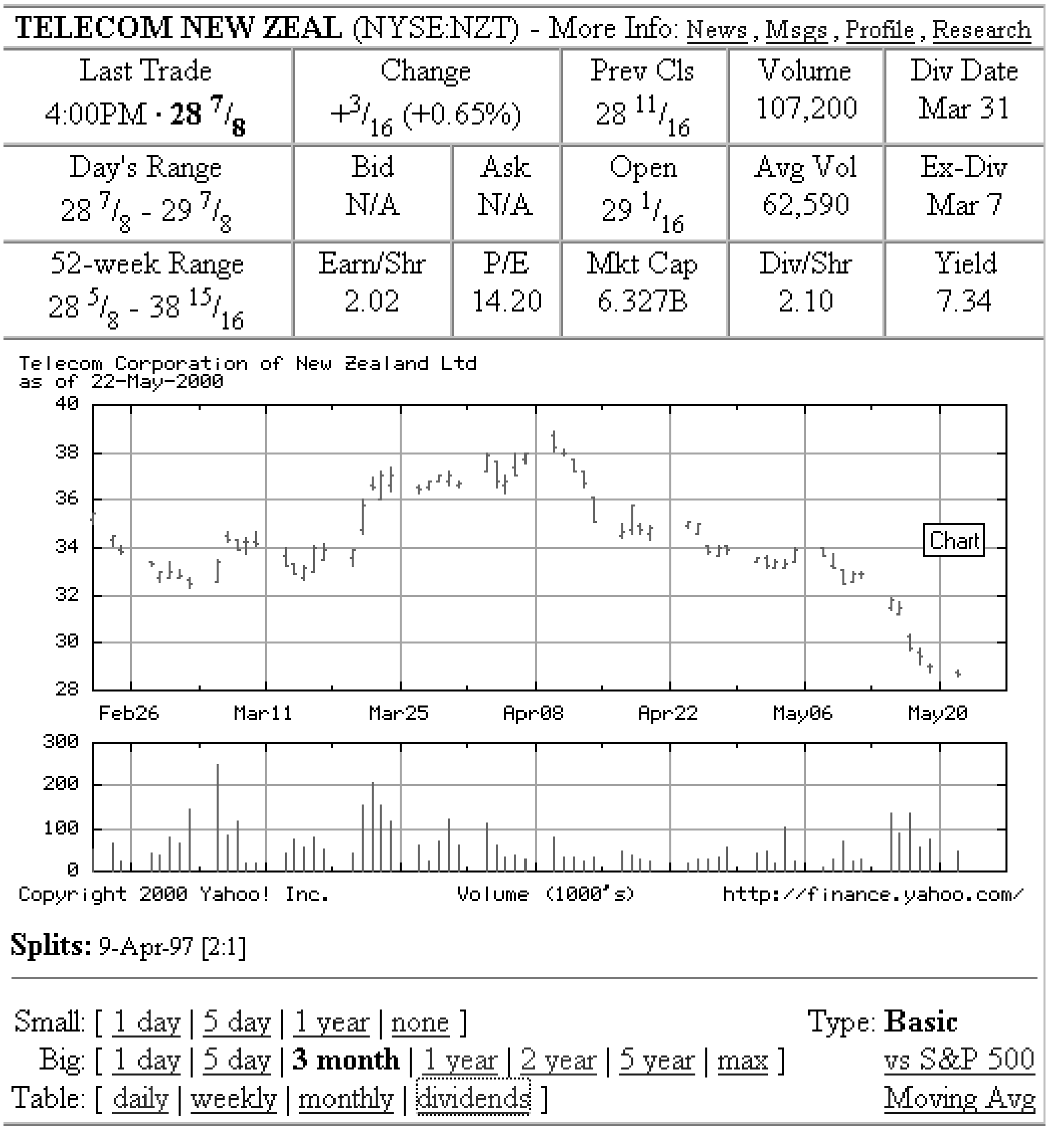
Figure 6.2 obtains from the Chicago Board of Options Exchange and provides the current price of calls and puts on NZT.

For example suppose there was a July put option, strike price \$30 that sold for $\$2\frac{1}{3}$. Suppose the current interest rate (in US dollars since these are US dollar prices) is 6%. This is roughly the interest rate on a short term risk free deposit like a treasury bill. Then the implied volatility is determined by finding the value of the parameter σ so that the Black-Scholes formula gives exactly this value for an option price, i.e. finding a value of σ so that PUT=2.33 where

[CALL,PUT] = BLSPRICE(28.875,30,.06,42/252,σ,0).

In this case, we obtain $\sigma = .397$ larger than the historical volatility over the past three months. Which value is "correct"? Apparently in this case the Q measure assigns a greater volatility than the stock has. Is there any obvious explanation? Certainly the market conditions for this company may well have changed enough to effect a recent change in the volatility. Moreover, remember that the distribution that matters in pricing an option is the Q measure, a distribution assigned by *the market for the option.* If you use any other distribution, you offer others an arbitrage at your expense. So in practice, volatility is backed out, where possible, from the price of derivatives on the given asset. Exceptions are made when the market for a given option or the underlying asset is very thin or when it is a very short time until maturity (in this case, other considerations affect the option price). In this case there are no traded options on telecom in the day and we have little choice but to use hisstorical volatility, perhaps comparing with implied volatility from previous days.

When we decide to uses historical data to estimate volatility there are more efficient estimators than the sample variance of the returns. Those which use the



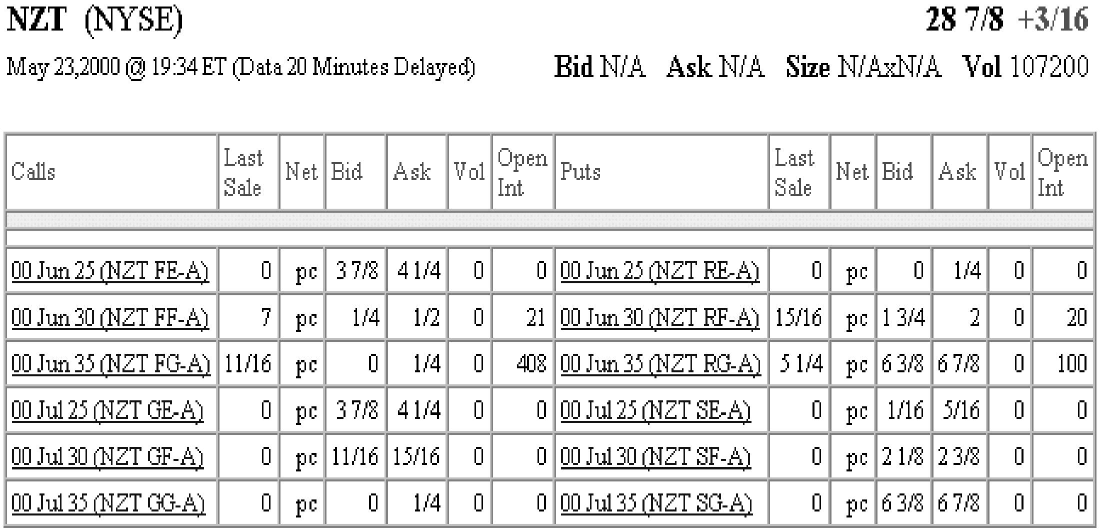
TELECOM NEW ZEAL (NYSE:NZT) - More Info: News , Msgs , Profile , Research

Last Trade 4:00PM · **$28\,^{7}/_{8}$**	Change $+^{3}/_{16}$ (+0.65%)		Prev Cls $28\,^{11}/_{16}$	Volume 107,200	Div Date Mar 31
Day's Range $28\,^{7}/_{8}$ - $29\,^{7}/_{8}$	Bid N/A	Ask N/A	Open $29\,^{1}/_{16}$	Avg Vol 62,590	Ex-Div Mar 7
52-week Range $28\,^{5}/_{8}$ - $38\,^{15}/_{16}$	Earn/Shr 2.02	P/E 14.20	Mkt Cap 6.327B	Div/Shr 2.10	Yield 7.34

Splits: 9-Apr-97 [2:1]

Small: [1 day | 5 day | 1 year | none] Type: **Basic**
Big: [1 day | 5 day | **3 month** | 1 year | 2 year | 5 year | max] vs S&P 500
Table: [daily | weekly | monthly | dividends] Moving Avg

Figure 6.1:



NZT (NYSE) **28 7/8 +3/16**

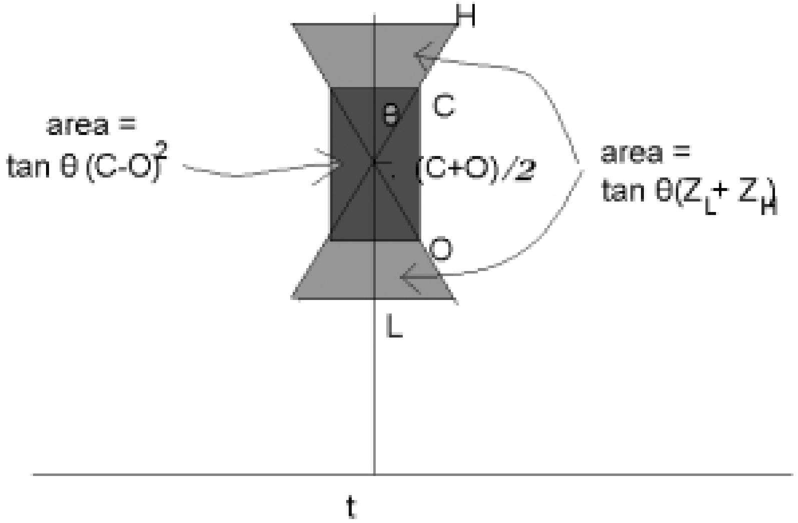
May 23,2000 @ 19:34 ET (Data 20 Minutes Delayed) **Bid** N/A **Ask** N/A **Size** N/AxN/A **Vol** 107200

Calls	Last Sale	Net	Bid	Ask	Vol	Open Int	Puts	Last Sale	Net	Bid	Ask	Vol	Open Int
00 Jun 25 (NZT FE-A)	0	pc	3 7/8	4 1/4	0	0	00 Jun 25 (NZT RE-A)	0	pc	0	1/4	0	0
00 Jun 30 (NZT FF-A)	7	pc	1/4	1/2	0	21	00 Jun 30 (NZT RF-A)	15/16	pc	1 3/4	2	0	20
00 Jun 35 (NZT FG-A)	11/16	pc	0	1/4	0	408	00 Jun 35 (NZT RG-A)	5 1/4	pc	6 3/8	6 7/8	0	100
00 Jul 25 (NZT GE-A)	0	pc	3 7/8	4 1/4	0	0	00 Jul 25 (NZT SE-A)	0	pc	1/16	5/16	0	0
00 Jul 30 (NZT GF-A)	0	pc	11/16	15/16	0	0	00 Jul 30 (NZT SF-A)	0	pc	2 1/8	2 3/8	0	0
00 Jul 35 (NZT GG-A)	0	pc	0	1/4	0	0	00 Jul 35 (NZT SG-A)	0	pc	6 3/8	6 7/8	0	0

Figure 6.2:

highs and lows in a period offer considerable gains in efficiency using commonly published data.

Particularly useful statistics in this regard are the exponentially distributed random variables $Z_H = \log(H/O)\log(H/C)$ and $Z_L = \log(L/O)\log(L/C)$ (in the case of geometric Brownian motion) introduced in Theorem 4.3.4, where (O, C, H, L) denote the open, close, high, low price over a period Δt. In this case both Z_H and Z_L have an exponential distribution with mean $\sigma^2\Delta t/2$ conditionally on the values of O, C. Therefore, it is independent of C. A similar argument leads to $z_L = \log(L/O)\log(L/C) \sim exp(\sigma^2\Delta t/2)$. Thus both of these statistics leads to an unbiased estimator of the parameter σ^2 . An alternative estimator of the scale parameter σ^2 is obtained from the increment C/O alone. Indeed the maximum likelihood estimator based on the distribution of this increment is $\{\log(C/O)\}^2/\Delta t$. Thus we have three estimators of the volatility parameter, $\{\log(C/O)\}^2/\Delta t, 2\log(L/O)\log(L/C)/\Delta t$, and $2\log(L/O)\log(L/C)/\Delta t$. While the first is independent of the other two given O, unfortunately the second and third are themselves not uncorrelated. In order to weight them optimally we need some information about their joint distribution. It follows that both $\{log(C/O)\}^2/\Delta t$ and $(Z_H + Z_L)/\Delta t$ provide unbiased estimators of the volatility parameter σ^2 and indeed the latter is independent of the



former.

These estimators are areas illustrated in Figure xx. Consider the plot corresponding to time t. The vertical scale is logarithmic so that logs are plotted. This plot is constructed using an arbitrarily chosen angle θ from the four values (O, C, H, L) using two lines ℓ_1, ℓ_2 through the point $(t, \frac{1}{2}(\log(O) + \log(C)))$ with slopes $\pm\, tan(\theta)$. Horizontal lines are drawn at the ordinate values $\log(H), \log(L), \log(O), \log(C)$ and using the points where $\log(O)$and $\log(C)$ strike the two lines as corners, a rectangle is constructed. The area of this rectangle $tan(\theta)(\log(C/O))^2$ is an unbiased estimator of $tan(\theta)\sigma^2\Delta t$ provided the Brownian motion has no drift. The second region consists of "wings" generated by the four points at which the horizontal line at $\log(H), \log(L)$ strike the lines ℓ_1, ℓ_2. The total area of this region (both wings) is $tan(\theta)(Z_L + Z_H)$ which is another unbiased estimator of $tan(\theta)\sigma^2 T$ independent of the first, and also independent of whether or not the underlying Brownian motion has drift. By comparing these areas, we can detect abnormal changes in the volatility, or changes in the drift of the process that will increase the observed value of $(\log(C/O))^2$ while leaving the second estimator unchanged. Because each estimator is based only on a single period, it is useful to provide as well a plot indicating whether there is a persistent change in either or both of the two estimators of volatility.

If the Brownian motion does indeed have zero drift we could combine the two estimators above, and the optimal linear combination is, with weights very slightly rounded,

$$\sigma^2_{BLU} \approx \frac{1}{7\Delta t}(\{\log(C/O)\}^2 + 6(Z_H + Z_L)) \tag{6.4}$$

where the weights have been determined using the fact that Z_H and Z_L are correlated with correlation coefficient $\approx -.338$.

How much better is this than the usual estimator of volatility $(\log(C/O))^2$?

Computation in the case $\sigma^2 T = 1$ yields

$$var(\hat{\sigma}^2_{BLU}) \approx 0.284$$

while $var(\log(C/O)^2) = 2$. The ratio is approximately 7. In other words, observations on the high, low, open, close permit about seven times the efficiency for estimating the volatility parameter. Related estimators have been suggested in the literature. For example, Parkinson (1980) in effect suggests the estimator

$$\frac{1}{4T^2 log\ 2}(\log(H/L))^2 \tag{6.5}$$

which is about five times as efficient as $(\log(C/O))^2/T$ and Rogers and Satchell (1991) suggest the estimator

$$(Z_H + Z_L) \tag{6.6}$$

which is nearly the same as σ^2_{BLU}. Perhaps the simplest high efficiency estimator, is suggested by Garman and Klass (1980) and takes the form

$$\hat{\sigma}^2_{GK} = \frac{1}{2}(\log(H/L))^2 - (2\ln(2) - 1)(\log(C/O))^2 \tag{6.7}$$

We also show empirically the effectiveness of incorporating the high low close information in a measure of volatility. For example, the plot below gives the eggtimer plot for the Dow Jones Industrial Index for the months of February and March 2000. The vertical scale is logarithmic since the Black Scholes model is such that the logarithm of the index is Brownian motion. A prepondrance of red rectangles shows periods when the drift dominates, whereas where the green tails are larger, the volatility is evidenced more by large values of the high or small values of the low, compared to the daily change. The cumulative sum of the areas of the regions below, either red or green, provide a measure of volatility. In the absence of substantial drift, both measure the same quantity. We can either plot this cumulative sum or a moving average of the above measures as in the graph below. The curve labelled “intra-day” measures the volatility as determinined by the high, low, open close for a given day and that labelled inter-day, the volatility as estimated from only the daily close/open (or close/close-the second almost identical curve) prices. Apparently for this period, from January 1999 to March 2000, the intra-day volatility was greater than the inter-day volatility. This is equally evident from the plot of the cumulative variance for the same period of time.

A consistent difference between the intra-day and the inter-day volatility would be easy to explain if the situation were reversed because one could argue that the inter-day measure contains a component due to the drift of the process and over this period there was a significant drift. A difference in this direction is more difficult to explain unless it is a failure of the Black-Scholes model.

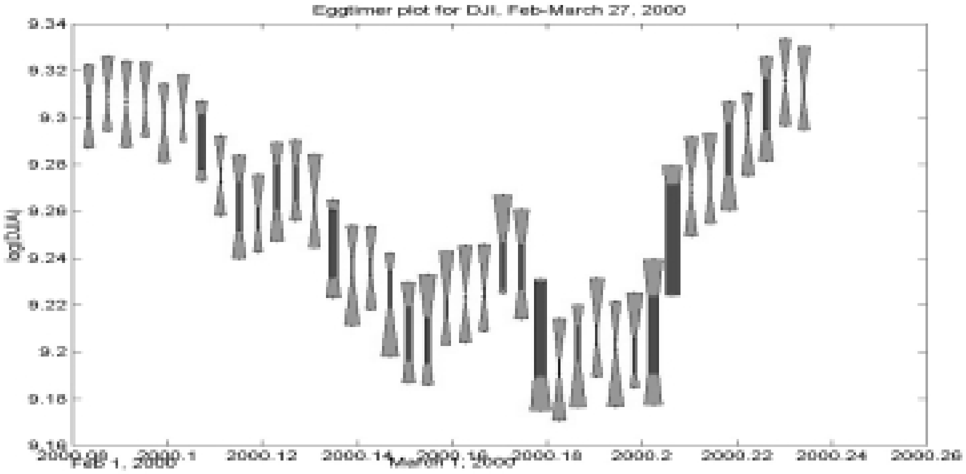


Figure 6.3:

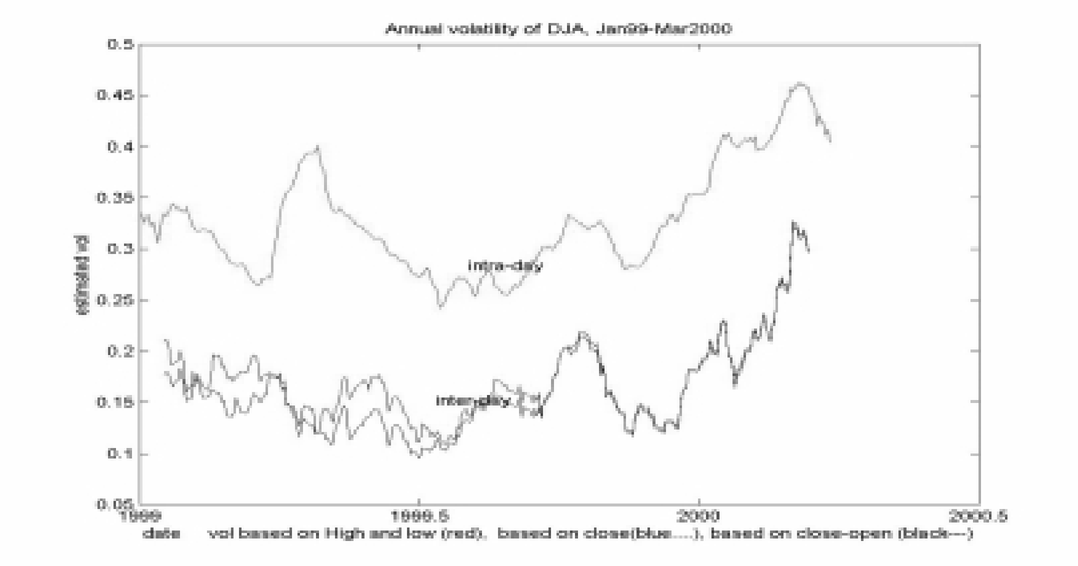


Figure 6.4:

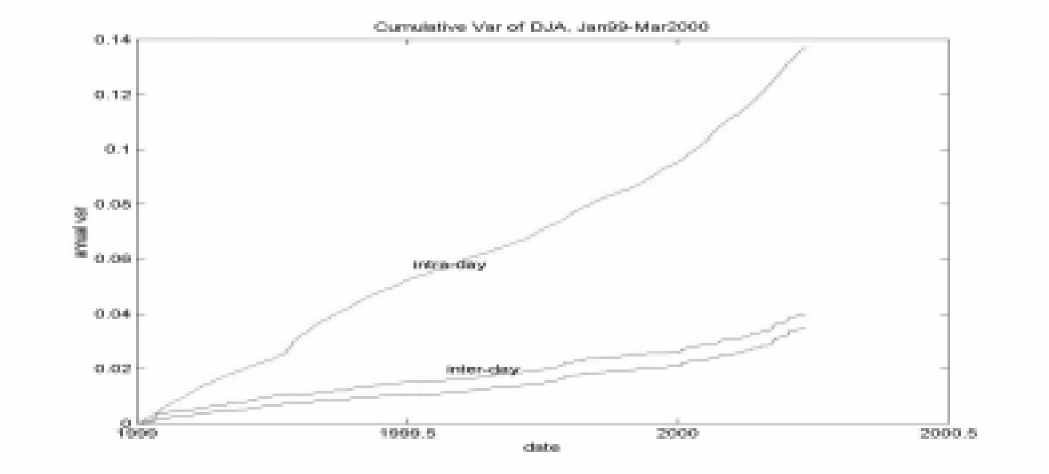


Figure 6.5:

If there is a consistent such failure, one might expect a similar behaviour in another market. If we generate a similar plot over the identical period of time for the NASDAQ index we find that the comparison is reversed. This, of course, could be explained by the greater drift of the technology dependent NASDAQ (relative to its volatility) compared to the relatively traditional market of the Dow Jones.

There is no doubt that this difference is real. In fact if we plot the cumulative value of the range of the index divided by the close $(H-L)/C$ as in Figure X below it confirms that the daily range as measured by this ratio is consistently smaller for the NASDAQ than for the Dow Jones for this period.

Although high, low, open, close data is commonly available for many financial time series, the quality of the recording is often doubtful. When we used older data from the Toronto Stock Exchange, there were a number of days in which the high or low were so far from open and close to be explicable only as a recording error (often the difference was almost exactly $10). When the data on highs and lows is accurate, there is substantial improvement in efficiency and additional information available by using it. But there is no guarantee that published data is correct. A similar observation on NYSE data is made by Wiggins (1991); "*In terms of the CUPV data base itself, there appear to be a number of cases where the recorded high or low prices are significantly out of line relative to adjacent closing prices*".

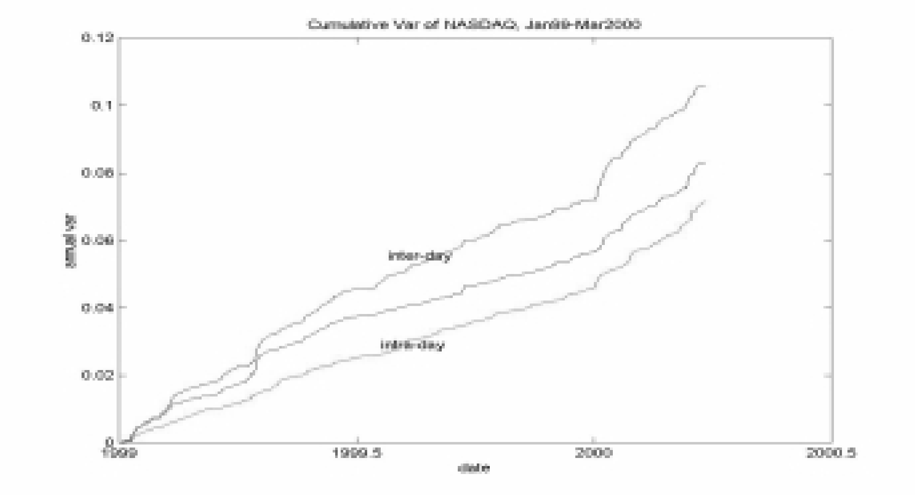


Figure 6.6:

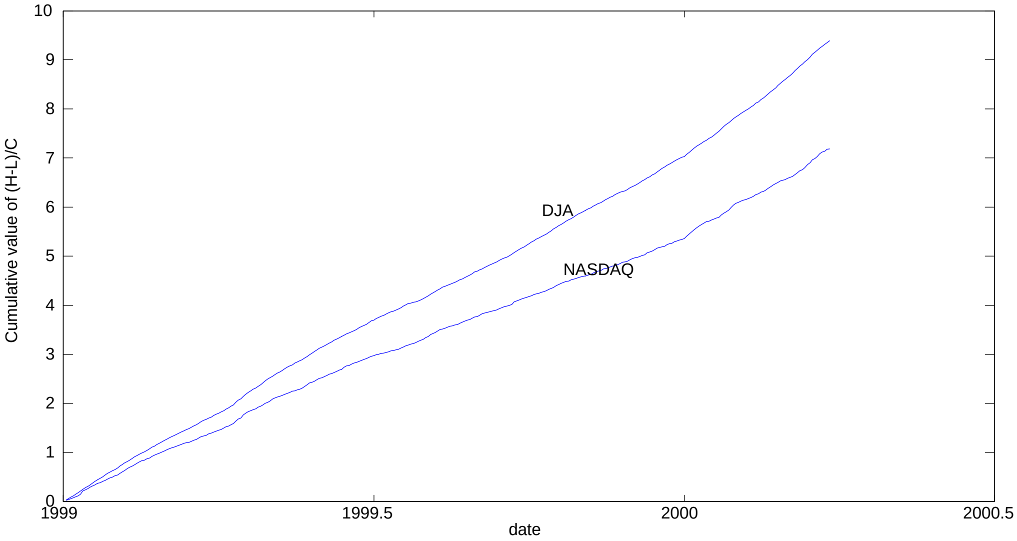


Figure 6.7:

6.3 Estimating Hedge ratios and Correlation Coefficients.

"The only perfect hedge is in a Japanese garden." In practice we are usually faced with requiring to hedge one asset using one or more other assets. These may include derivatives on the original asset, which are highly correlated over short periods, or simply similar investments (such as financial institutions) which react similarly to market conditions.

Suppose we wish to hedge one investment, say in stock C_2 using another, say C_1. Let us suppose that $\log(C_1/O_1), \log(S_2/O_2)$ have variances $\sigma_1^2\Delta t, \sigma_2^2\Delta t$ and correlation coefficient ρ. Assume at the open we are long 1 dollar worth of stock 2 and short h dollars worth of stock 1. Our total investment is $1-h_.$ Our total at the end of the period is $C_2/O_2 - hC_1/O_1$. The optimal hedge ratio, the value of h minimizing the variance of $C_2/O_2 - hC_1/O_1$ is given by

$$h = \frac{cov(C_2/O_2, C_1/O_1)}{var(C_1/O_1)} = \frac{e^{\rho\sigma_1\sigma_2\Delta t}-1}{e^{\sigma_1^2\Delta t}-1} \approx \rho\sigma_2/\sigma_1 \quad \text{when } \Delta t \text{ is small.}$$

While volatilities σ_1 ,σ_2 may be implied by derivatives on each of these assets, the correlation parameter ρ is typically not known and usually estimated from historical data.(If spread options on the difference between these two assets were marketed, then these might allow us to back out correlations from market prices as well). We consider using full observations on high, low, open, close data towards estimating ρ.

Since in the two or multi-factor case the joint distributions of highs, lows and closing values is unknown, we need to revert to a simpler alternative than likelihood methodology. One possibility is a semiparametric approach. We have seen that in the Black-Scholes model, tthe statistics

$$\begin{aligned}
Z_{H1} &= \log(H_1/O_1)\log(H_1/C_1) \\
Z_{H2} &= \log(H_2/O_2)\log(H_2/C_2) \\
Z_{L1} &= \log(L_1/O_1)\log(L_1/C_1) \\
Z_{L2} &= \log(L_2/O_2)\log(L_2/C_2)
\end{aligned}$$

all have marginal exponential distributions and each is independent of the close.

We consider here a semi-parametric approach. As an introduction, let us consider a simple transformation of the above exponential random variables g and assume that we can determine the correlation $cor(g(Z_{H1}), g(Z_{H2})) = a(\rho)$ as a function of ρ. For simplicity assume a location and scale change so that $E\{g(Z_{H1})\} = 0, var\{g(Z_{H1})\} = 1$. Then a simple estimating function for ρ can be constructed as

$$g(Z_{H1})g(Z_{H2}) + g(Z_{L1})g(Z_{L2}) - 2a(\rho) = 0. \tag{6.8}$$

TheGodambe information (or reciprocal of asymptotic variance) in this estimating function for the parameter ρ is

$$\frac{\{2a'(\rho)\}^2}{var\{g(Z_{H1})g(Z_{H2})+g(Z_{L1})g(Z_{L2})\}} \tag{6.9}$$

and this should be a guideline to the choice of transformation g. Larger values of this ratio correspond to smaller values of the asymptotic variance of the estimator. Another estimating equation for ρ that is commonly used is written in terms of the open and closing prices:

$$(C_2-O_2-\mu_2T)[\ C_1-O_1-\mu_1T-\frac{\rho\sigma_1}{\sigma_2}(C_2-O_2-\mu_2T)]=0$$

from which, if we use sample variances to estimate the variance parameters, we obtain the estimator

$$\hat{\rho}_C=\widehat{cor}(C_2-O_2,C_1-O_1) \tag{6.10}$$

where $\widehat{cor}$ denotes the sample correlation coefficient. By a similar argument this has Godambe information given by

$$\frac{1}{1-\rho^2}.$$

The estimating function in (**??**) is not only unbiased, it is conditionally unbiased given C_1, C_2, O_1, O_2. To see this note that the conditional expectation

$$E[g(Z_{H1})g(Z_{H2})+g(Z_{L1})g(Z_{L2})-2a(\rho)|C_1-O_1,C_2-O_2] \tag{6.11}$$

is a function of the complete sufficient statistic (C_1-O_1, C_2-O_2) for the drift terms (μ_1,μ_2) whose expectation is 0 for any value of these drift terms. It follows that the estimating functions in (**??**) and (**??**) are orthogonal. (Note: it is not difficult to show that the conditional expectation in (**??**) can be generated from two correlated Brownian bridges without knowledge of the drift terms (μ_1,μ_2) and is therefore a *bona fide* statistic, independent of these parameters.) Because they are orthogonal, the best linear combination of the two functions is easily obtained. The weights are proportional to

$$\frac{2a'(\rho)}{var\{g(Z_{H1})g(Z_{H2})+g(Z_{L1})g(Z_{L2})\}} \quad \text{and} \quad \frac{1}{var(C_1-O_1)(1-\rho^2)} \quad \text{respectively,}$$

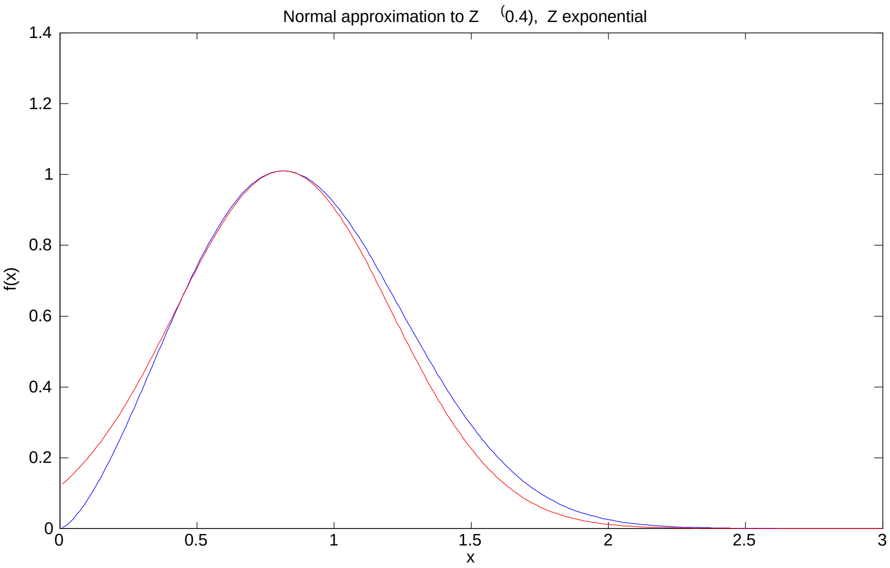
and the information in the optimal linear combination is given by the sum of the two informations

$$\frac{\{2a'(\rho)\}^2}{var\{g(Z_{H1})g(Z_{H2})+g(Z_{L1})g(Z_{L2})\}}+\frac{1}{1-\rho^2}. \tag{6.12}$$

The expression (**??**) is useful both to determine what transformation g contributes most to the estimation of ρ as well as to determine the extent of the contribution of the high-low information relative to an optimal combination with the open-close data.

There are various possibilities for the transformation g, the simplest being an ordinary standardization: We consider the class of the form $g(Z_{H1}^p) = (Z_{H1}^p - E(Z_{H1}^p))/\sqrt{var(Z_{H1}^p)}$ for some suitable value of $p > 0$. A transformation of the gamma distributions that make them very nearly normal is the cube root transformation($p = 1/3$). However, we will see later that a slightly different power $p = 0.4$ results in a distribution that is still close to normal but produces somewhat greater efficiency for estimating the correlation across a range of underlying values of ρ. See Figure **??** for the comparison of the density of $Z_H^{0.4}$ and an approximating normal density obtained by equating the two densities at the mode.

Normal approximation to $Z^{0.4}$

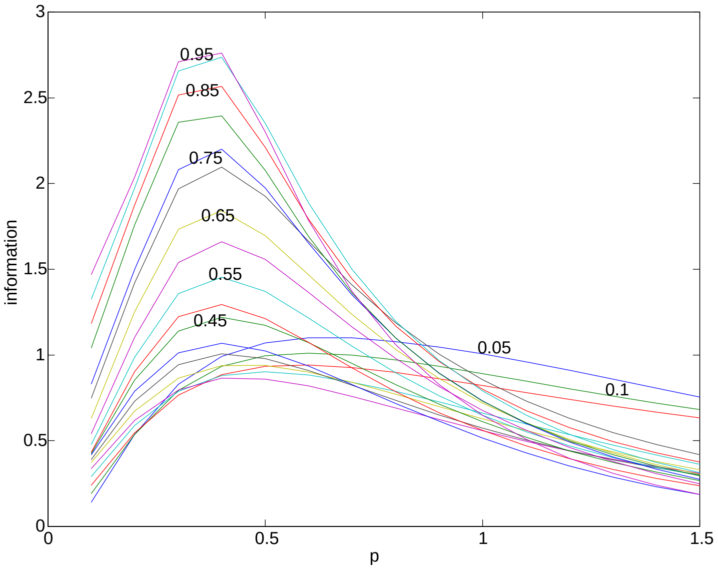


Standardization is achieved using the mean and variance

$$\beta_i = E(Z_{Hi}^{0.4}) = E(Z_{Li}^{0.4}) = \Gamma(1.4)\left[\frac{\sigma_i^2 T}{2}\right]^{0.4} = .6724(\sigma_i^2 T)^{0.4},$$

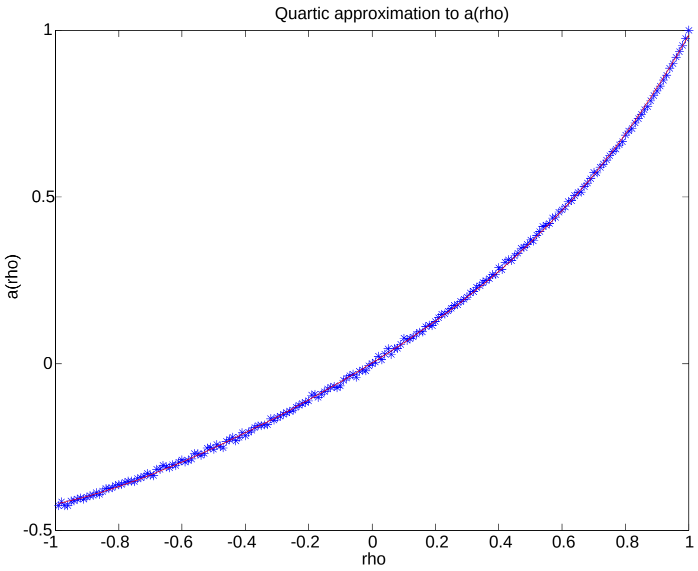
$$\gamma_i^2 = var(Z_{Hi}^{0.4}) = \{\Gamma(1.8) - \Gamma^2(1.4)\}\left[\frac{\sigma_i^2 T}{2}\right]^{0.8} = .0828(\sigma_i^2 T)^{0.8}. \qquad (6.13)$$

We begin by choosing a value of p which results in approximate maximization of the information in (**??**) . Since $a(\rho)$ is unknown, evaluating its derivative in the numerator of (**??**) must be done using either simulation or numerical methods. Our approximation leads to the following plot of the information for versus p for various values of ρ.



The optimal choice of p evidently depends to some degree on the underlying value of ρ (see the following figure) but it appears that the choice $p \simeq 0.4$ is reasonably efficient for most values of ρ. The level of these curves also indicates what increase of efficiency to expect by using the estimating function (**??**) in addition to (7.9). In fact the relative efficiency of this estimator relative to use of (**??**) is example in the case $\rho = 0.85$, the two terms in (**??**) are about 2.8 and 3.6 respectively indicating roughly an increase of 80 percent due to the additional information. The gains appear to be smaller for larger values of ρ. When $p = .4$ and $\rho = 0.15$, the values are around 0.75 and 0.98 respectively indicating around a 75 percent gain in efficiency. So although the information changes with ρ, the relative efficiency appears more stable. We emphasize that these efficiency figures are rough at this stage, since the derivative in the numerator or (3.9) has been estimated by a crude first difference.

In the case $p = .4$, we may obtain the function $a(\rho)$ by numerical means using the known but complicated form of the joint distribution (see He, Kierstead and xxx) or alternatively estimate it by simulation. In this case we used a number of approximations including a smooth regression of the form $a(\rho) \approx c(b^\rho - d)$ to estimate the function from simulations. However a fourth degree polynomial fit seemed adequate. The polynomial fit was $a(\rho) \approx 0.0903\rho^4 + 0.1163\rho^3 + 0.1898\rho^2 + 0.5867\rho + 0.0023$. The fit to the function is graphed below. The individual points are estimates of the correlation each based on 25,000 bivariate Brownian motion processes.



Inverting the simple approximation to $a(\rho)$ provides a highly simple and tractible estimator of the correlation between stocks based only on the correlation between the marginally exponential statistics Z_{Hi}, Z_{Li}. This estimator is

$$\hat{\rho}_1 = a^{-1}(\frac{1}{2}(\widehat{cor}(Z_{H1}^{0.4}, Z_{H2}^{0.4}) + \widehat{cor}(Z_{L1}^{0.4}, Z_{L2}^{0.4}))) \qquad (7.14)$$

A similar estimator obtains as well from the cross terms since $cor_\rho(Z_{H1}^{0.4}, Z_{H2}^{0.4}) = a(-\rho)$.

$$\hat{\rho}_2 = -a^{-1}(\frac{1}{2}(\widehat{cor}(Z_{H1}^{0.4}, Z_{L2}^{0.4}) + \widehat{cor}(Z_{L1}^{0.4}, Z_{H2}^{0.4}))) \qquad (7.15)$$

where again $\widehat{cor}$ denotes the sample correlation . There are two possible ways of using these estimators, either as estimators of the correlation between the two processes that that, unlike (3.5), is independent of any drift within the periods, or in combination with the the estimator (**??**). As anticipated, the estimator (**??**) adds considerably to the efficiency of (**??**) . The relative efficiencies are graphed below. We also approximate the gain in efficiency in (**??**) by assuming that the distribution of $g(Z_{Hi})$, g(Z$_{Li}$) is multivariate normal so that we can estimate the denominator of (**??**). In this case, (see Anderson, section xxx), if we put

$$X = [X_{H1}, X_{L1}, X_{H2}, X_{L2}]$$

where $X_{Hj} = g(Z_{Hj}), X_{Lj} = g(Z_{Lj})$, then X has covariance matrix

$$\begin{bmatrix} A & B(\rho) \\ B(\rho) & A \end{bmatrix}$$

where

$$\begin{aligned} B(\rho) &= \left[\begin{array}{ll} a(\rho) & a(-\rho) \\ a(-\rho) & a(\rho) \end{array}\right], \\ \text{and } A &= \left[\begin{array}{ll} 1 & a(-1) \\ a(-1) & 1 \end{array}\right] = B(1). \end{aligned}$$

Then the denominator in the first term of (3.7) is approximately (see Anderson( ),),

$$var\{g(Z_{H1})g(Z_{H2}) + g(Z_{L1})g(Z_{L2})\} = 2[1 + a^2(-1) + a^2(\rho) + a^2(-\rho)]$$

and (**??**) becomes

$$\frac{2\{a'(\rho)\}^2}{[1 + a^2(-1) + a^2(\rho) + a^2(-\rho)]} + \frac{1}{1-\rho^2}.$$

The ratio of the two terms, $(1-\rho^2)\ \frac{2\{a'(\rho)\}^2}{[1+a^2(-1)+a^2(\rho)+a^2(-\rho)]}$ provides the relative efficiency of the estimator $\hat{\rho}_1$ with respect to (**??**). Our approximation to the function $a(.)$, provides at best a crude approximation to the derivative in the numerator, but this would seem to indicate efficiencies in excess of 60 percent. An estimator which combines high-low and open-close information can increase substantially the information and this is confirmed by simulations. Figure XX shows the relative efficiencies of the estimator $\hat{\rho}_1$ and the best linear combination of $\hat{\rho}_1$ and $\hat{\rho}_C$ both with respect to $\hat{\rho}_C$. These efficiencies are obtained by simulation and a smoothed curve (loess in Splus) is shown through the points.The points are determined as an average of 500 simulations, each corresponding to sample size $n = 200$ having drift 0. The upper set of points and curve are the efficiencies (as measured by the ratio of sample variances) of the optimal linear combination of the three estimators, ρ_1, ρ_2, ρ_C with respect to the estimator ρ_C . Efficiency gains of one hundred percent and more are observed, especially when the true correlation is around 0.75. If we use the optimal linear combination of the two estimators ρ_C, ρ_1 only, there is very little loss of information over using all three. These points are labelled "." and the smooth nearly coincides with the upper curve except at the extreme ends. Finally the points labelled "*" and the lower curve is the efficiency of the estimator ρ_C alone. Evidently, it is virtually as efficient as ρ_C for values of ρ around 0.75. Any bias in the estimators is too small to be detected in a simulation of this magnitude.

FIGURE XX