MONTE CARLO SIMULATION AND FINANCE

Don L. McLeish

September, 2004

Contents

Dedication: to be added

Acknowledgement 1 *I am grateful to all of the past students of Statistics 906 and the Master's of Finance program at the University of Waterloo for their patient reading and suggestions to improve this material, especially Keldon Drudge and Hristo Sendov. I am also indebted to my colleagues, Adam Kolkiewicz and Phelim Boyle for their contributions to my understanding of this material.*

Chapter 1

Introduction

Experience, how much and of what, is a valuable commodity. It is a major difference between an airline pilot and a New York Cab driver, a surgeon and a butcher, a succesful financeer and a cashier at your local grocers. Experience with data, with its analysis, experience constructing portfolios, trading, and even experience losing money (one experience we all think we could do without) are all part of the education of the financially literate. Of course, few of us have the courage to approach the manager of our local bank and ask for a few million so we can acquire this experience, and fewer still managers have the courage to acceed to our request. The "joy of simulation" is that you do not need to have a Boeing 767 to fly one, and that you don't need millions of dollars to acquire a considerable experience valuing financial products, constructing portfolios and testing trading rules. Of course if your trading rule is to buy condos in Florida because you expect boomers to all wish to retire there, a computer simulation will do little to help you since the ingredients to your decision are largely psychological (yours and theirs), but if it is that you should hedge your current investment in condos using financial derivatives real estate companies, then the methods of computer simulation become relevant.

This book concerns the simulation and analysis of models for financial markets, particularly traded assets like stocks, bonds. We pay particular attention to financial derivatives such as options and futures. These are financial instruments which derive their value from some associated asset. For example a call option is written on a particular stock, and its value depends on the price of the stock at expiry. But there are many other types of financial derivatives, traded on assets such as bonds, currency markets or foreign exchange markets, and commodities. Indeed there is a growing interest in so-called "real options", those written on some real-world physical process such as the temperature or the amount of rainfall.

In general, an option gives the holder a right, not an obligation, to sell or buy a prescribed asset (the underlying asset) at a price determined by the contract (the exercise or strike price). For example if you own a call option on shares of IBM with expiry date Oct. 20, 2000 and exercise price $120, then on October 20, 2000 you have the right to purchase a fixed number , say 100 shares of IBM at the price $120. If IBM is selling for $130 on that date, then your option is worth $10 per share on expiry. If IBM is selling for $120 or less, then your option is worthless. We need to know what a fair value would be for this option when it is sold, say on February 1, 2000. Determining this fair value relies on sophisticated models both for the movements in the underlying asset and the relationship of this asset with the derivative, and is the subject of a large part of this book. You may have bought an IBM option for two possible reasons, either because you are speculating on an increase in the stock price, or to hedge a promise that you have made to deliver IBM stocks to someone in the future against possible increases in the stock price. The second use of derivatives is similar to the use of an insurance policy against movements in an asset price that could damage or bankrupt the holder of a portfolio. It is this second use of derivatives that has fueled most of the phenomenal growth in their trading. With the globalization of economies, industries are subject to

more and more economic forces that they are unable to control but nevertheless wish some form of insurance against. This requires hedges against a whole litany of disadvantageous moves of the market such as increases in the cost of borrowing, decreases in the value of assets held, changes in a foreign currency exchange rates, etc.

The advanced theory of finance, like many areas where advanced mathematics plays an important part, is undergoing a revolution aided and abetted by the computer and the proliferation of powerful simulation and symbolic mathematical tools. This is the mathematical equivalent of the invention of the printing press. The numerical and computational power once reserved for the most highly trained mathematicians, scientists or engineers is now available to any competent programmer.

One of the first hurdles faced before adopting stochastic or random models in finance is the recognition that for all practical purposes, the prices of equities in an efficient market are *random variables*, that is while they may show some dependence on fiscal and economic processes and policies, they have a component of randomness that makes them unpredictable. This appears on the surface to be contrary to the training we all receive that every effect has a cause, and every change in the price of a stock must be driven by some factor in the company or the economy. But we should remember that random models are often applied to systems that are essentially causal when measuring and analyzing the various factors influencing the process and their effects is too monumental a task. Even in the simple toss of a fair coin, the result is predetermined by the forces applied to the coin during and after it is tossed. In spite of this, we model it as a random variable because we have insufficient information on these forces to make a more accurate prediction of the outcome. Most financial processes in an advanced economy are of a similar nature. Exchange rates, interest rates and equity prices are subject to the pressures of a large number of traders, government agencies, speculators, as well as the forces applied by international

trade and the flow of information. In the aggregate there is an extraordinary number of forces and information that influence the process. While we might hope to predict some features of the process such as the average change in price or the volatility, a precise estimate of the price of an asset one year from today is clearly impossible. This is the basic argument necessitating stochastic models in finance. Adoption of a stochastic model does neither implies that the process is pure noise nor that we are unable to forecast. Such a model is adopted whenever we acknowledge that a process is not *perfectly* predictable and the *non-predictable* component of the process is of sufficient importance to warrant modeling.

Now if we accept that the price of a stock is a random variable, what are the constants in our model? Is a dollar of constant value, and if so, the dollar of which nation? Or should we accept one unit of a index what in some sense represents a share of the global economy as the constant? This question concerns our choice of what is called the "numeraire" in deference to the French influence on the theory of probability, or the process against which the value of our assets will be measured. We will see that there is not a unique answer to this question, nor does that matter for most purposes. We can use a bond denominated in Canadian dollars as the numeraire or one in US dollars. Provided we account for the variability in the exchange rate, the price of an asset will be the same. So to some extent our choice of numeraire is arbitrary- we may pick whatever is most convenient for the problem at hand.

One of the most important modern tools for analyzing a stochastic system is simulation. Simulation is the imitation of a real-world process or system. It is essentially a model, often a mathematical model of a process. In finance, a basic model for the evolution of stock prices, interest rates, exchange rates etc. would be necessary to determine a fair price of a derivative security. Simulations, like purely mathematical models, usually make assumptions about the behaviour of the system being modelled. This model requires inputs, often

called the parameters of the model and outputs a result which might measure the performance of a system, the price of a given financial instrument, or the weights on a portfolio chosen to have some desirable property. We usually construct the model in such a way that inputs are easily changed over a given set of values, as this allows for a more complete picture of the possible outcomes.

Why use simulation? The simple answer is that is that it transfers work to the computer. Models can be handled which have greater complexity, and fewer assumptions, and a more faithful representation of the real-world than those that can be handled tractable by pure mathematical analysis are possible. By changing parameters we can examine interactions, and sensitivities of the system to various factors. Experimenters may either use a simulation to provide a numerical answer to a question, assign a price to a given asset, identify optimal settings for controllable parameters, examine the effect of exogenous variables or identify which of several schemes is more efficient or more profitable. The variables that have the greatest effect on a system can be isolated. We can also use simulation to verify the results obtained from an analytic solution. For example many of the tractable models used in finance to select portfolios and price derivatives are wrong. They put too little weight on the extreme observations, the large positive and negative movements (crashes), which have the most dramatic effect on the results. Is this lack of fit of major concern when we use a standard model such as the Black-Scholes model to price a derivative? Questions such as this one can be answered in part by examining simulations which accord more closely with the real world, but which are intractable to mathematical analysis.

Simulation is also used to answer questions starting with "what if". For example, What would be the result if interest rates rose 3 percentage points over the next 12 months? In engineering, determining what would happen under more extreme circumstances is often referred to as stress testing and simulation is a particularly valuable tool here since the scenarios we are concerned about are

those that we observe too rarely to have a substantial experience of. Simulations are used, for example, to determine the effect of an aircraft of flying under extreme conditions and is used to analyse the flight data information in the event of an accident. Simulation often provides experience at a lower cost than the alternatives.

But these advantages are not without some sacrifice. Two individuals may choose to model the same phenomenon in different ways, and as a result, may have quite different simulation results. Because the output from a simulation is random, it is sometimes harder to analyze- some statistical experience and tools are a valuable asset. Building models and writing simulation code is not always easy. Time is required both to construct the simulation, validate it, and to analyze the results. And simulation does not render mathematical analysis unnecessary. If a reasonably simple analytic expression for a solution exists, it is always preferable to a simulation. While a simulation may provide an approximate numerical answer at one or more possible parameter values, only an expression for the solution provides insight to the way in which it responds to the individual parameters, the sensitivities of the solution.

In constructing a simulation, you should be conscious of a number of distinct steps;

1. Formulate the problem at hand. Why do we need to use simulation?

2. Set the objectives as specifically as possible. This should include what measures on the process are of most interest.

3. Suggest candidate models. Which of these are closest to the real-world? Which are fairly easy to write computer code for? What parameter values are of interest?

4. If possible, collect real data and identify which of the above models is most appropriate. Which does the best job of generating the general

characteristics of the real data?

5. Implement the model. Write computer code to run simulations.

6. Verify (debug) the model. Using simple special cases, insure that the code is doing what you think it is doing.

7. Validate the model. Ensure that it generates data with the characteristics of the real data.

8. Determine simulation design parameters. How many simulations are to be run and what alternatives are to be simulated?

9. Run the simulation. Collect and analyse the output.

10. Are there surprises? Do we need to change the model or the parameters? Do we need more runs?

11. Finally we document the results and conclusions in the light of the simulation results. Tables of numbers are to be avoided. Well-chosen graphs are often better ways of gleaning qualitative information from a simulation.

In this book, we will not always follow our own advice, leaving some of the above steps for the reader to fill in. Nevertheless, the importance of model validation, for example, cannot be overstated. Particularly in finance where data is often plentiful, highly complex mathematical models are too often applied without any evidence that they fit the observed data adequately. The reader is advised to consult and address the points in each of the steps above with each new simulation (and many of the examples in this text).

Example

Let us consider the following example illustrating a simple use for a simulation model. We are considering a buy-out bid for the shares of a company. Although the company's stock is presently valued at around $11.50 per share, a careful analysis has determined that it fits sufficiently well with our current

assets that if the buy-out were successful, it would be worth approximately $14.00 per share in our hands. We are considering only three alternatives, an immediate cash offer of $12.00, $13.00 or $14.00 per share for outstanding shares of the company. Naturally we would like to bid as little as possible, but we expect a competitor to virtually simultaneously make a bid for the company and the competitor values the shares differently. The competitor has three bidding strategies that we will simply identify as I, II, and III. There are costs associated with any pair of strategies (our bid-competitor's bidding strategy) including costs associated with losing a given bid to the competitor or paying too much for the company. In other words, the payoff to our firm depends on the amount bid by the competitor and the possible scenarios are as given in the following table.

			Competitor's	Strategy	
	Bid	I	II	III	
Your	12	3	2	-2	
Bid	13	1	-4	4	
	14	0	-5	5	

The payoffs to the competitor are somewhat different and given below

		Competitor's	Strategy	
		I	II	III
Your	12	-1	-2	3
Bid	13	0	4	-6
	14	0	5	-5

For example, the combination of your bid=$13 per share and your competitor's strategy II results in a loss of 4 units (for example four dollars per share) to you and a gain of 4 units to your competitor. However it is not always the case that the your loss is the same as your competitor's gain. A game with this property is called a *zero-sum game* and these are much easier to analyze analytically. Define the 3×3 matrix of payoffs to your company by A and the

payoff matrix to your competitor by B,

$$A = \begin{pmatrix} 3 & 2 & -2 \\ 1 & -4 & 4 \\ 0 & -5 & 5 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & -2 & 3 \\ 0 & 4 & -6 \\ 0 & 5 & -5 \end{pmatrix}.$$

Provided that you play strategy $i = 1, 2, 3$ (i.e. bid \$12,\$13,\$14 with probabilities p_1, p_2, p_3 respectively and the probabilities of the competitor's strategies are q_1, q_2, q_3. Then if we denote

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}, \text{ and } q = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix},$$

we can write the expected payoff to you in the form $\sum_{i=1}^{3} \sum_{j=1}^{3} p_i A_{ij} q_j$. When written as a vector-matrix product, this takes the form $p^T Aq$. This might be thought of as the average return to your firm in the long run if this game were repeated many times, although in the real world, the game is played only once. *If the vector q were known to you,* you would clearly choose $p_i = 1$ for the row i corresponding to the maximum component of Aq since this maximizes your payoff. Similarly if your competitor knew p, they would choose $q_j = 1$ for the column j corresponding to the maximum component of $p^T B$. Over the long haul, if this game were indeed repeated may times, you would likely keep track of your opponent's frequencies and replace the unknown probabilities by the frequencies. However, we assume that both the actual move made by your opponent and the probabilities that they use in selecting their move are unknown to you at the time you commit to your strategy. However, if the game is repeated many times, each player obtains information about their opponent's taste in moves, and this would seem to be a reasonable approach to building a simulation model for this game. Suppose the game is played repeatedly, with each of the two players updating their estimated probabilities using information gathered about their opponent's historical use of their available strategies. We

may record number of times each strategy is used by each player and hope that the relative frequencies approach a sensible limit. This is carried out by the following Matlab function;

```
function [p,q]=nonzerosum(A,B,nsim)
% A and B are payoff matrices to the two participants in a game.
Outputs
%mixed strategies p and q determined by simulation conducted nsim
times
n=size(A);                          % A and B have  the same size
p=ones(1,n(1)); q=ones(n(2),1); % initialize with positive weights
on all strategies
for i=1:nsim                        % runs the simulation nsim times
[m,s]=max(A*q);                     % s=index of optimal strategy for
us
[m,t]=max(p*B);                     % =index of optimal strategy for
competitor
p(s)=p(s)+1;                        % augment counts for us
q(t)=q(t)+1;                        % augment counts for competitor
end
p=p-ones(1,n(1)); p=p/sum(p);  %remove initial weights from counts
and then
q=q-ones(n(2),1); q=q/sum(q);  % convert counts to relative frequencies
```

The following output results from running this function for 50,000 simulations.

```
[p,q]=nonzerosum(A,B,50000)
```

This results in approximately $p' = [\frac{2}{3}\ 0\ \frac{1}{3}]$ and $q' = [0 \quad \frac{1}{2} \quad \frac{1}{2}]$ with an average payoff to us of 0 and to the competitor 1/3. This seems to indicate that the strategies should be "mixed" or random. You should choose a bid of \$12.00 with probability around 2/3, and \$14.00 with probability 1/3. It appears that

the competitor need only toss a fair coin and select between B and C based on its outcome. Why randomize your choice? The average value of the game to you is 0 if you use the probabilities above (in fact if your competitor chooses probabilities $q' = [0 \quad \frac{1}{2} \quad \frac{1}{2}]$ it doesn't matter what your frequencies are, your average is 0). If you were to believe a single fixed strategy is always your "best" then your competitor could presumably determine what your "best" strategy is and act to reduce your return (i.e. substantially less than 0) while increasing theirs. Only randomization provides the necessary insurance that neither player can guess the strategy to be employed by the other. This is a rather simple example of a two-person game with non-constant sum (in the sense that A+B is not a constant matrix). Mathematical analysis of such games can be quite complex. In such case, provided we can ensure cooperation, participants may cooperate for a greater total return.

There is no assurance that the solution above is optimal. In fact the above solution is worth an average of 0 per game to us and 1/3 to our competitor. If we revise our strategy to $p' = [\frac{2}{3} \ \frac{2}{9} \ \frac{1}{9}]$, for example, our average return is still 0 but we have succeeded in reducing that of our opponent to 1/9. The solution we arrived at in this case seems to be sensible solution, achieved with little effort. Evidently, in a game such as this, there is no clear definition of what an optimal strategy would be, since one might plan one's play based on the worst case, or the best case scenario, or something in between such as an average? Do you attempt to collaborate with your competitor for greater total return and then subsequently divide this in some fashion? This simulation has emulated a simple form of competitor behaviour and arrived at a reasonable solution, the best we can hope for without further assumptions.

There remains the question of how we actually select a bid with probabilities 2/3, 0 and 1/3 respectively. First let us assume that we are able to choose a "random number" U in the interval [0,1] so that the probability that it falls in any given subinterval is proportional to the length of that subinterval. This

means that the random number has a uniform distribution on the interval [0,1]. Then we could determine our bid based on the value of this random number from the following table;

If	$U < 2/3$		$2/3 \leq U < 1$
Bid	12	13	14

The way in which U is generated on a computer will be discussed in more detail in chapter 2, but for the present note that each of the three alternative bids have the correct probabilities.

Chapter 2

Some Basic Theory of Finance

Introduction to Pricing: Single Period Models

Let us begin with a very simple example designed to illustrate the no-arbitrage approach to pricing derivatives. Consider a stock whose price at present is $\$s$. Over a given period, the stock may move either up or down, up to a value su where $u > 1$ with probability p or down to the value sd where $d < 1$ with probability $1 - p$. In this model, these are the only moves possible for the stock in a single period. Over a longer period, of course, many other values are possible. In this market, we also assume that there is a so-called risk-free bond available returning a guaranteed rate of $r\%$ per period. Such a bond cannot default; there is no random mechanism governing its return which is known upon purchase. An investment of \$1 at the beginning of the period returns a guaranteed $\$(1 + r)$ at the end. Then a portfolio purchased at the beginning of a period consisting of y stocks and x bonds will return at the end of the period an amount $\$x(1 + r) + ysZ$ where Z is a random variable taking

values u or d with probabilities p and $1-p$ respectively. We permit owning a negative amount of a stock or bond, corresponding to shorting or borrowing the correspond asset for immediate sale.

An ambitious investor might seek a portfolio whose initial cost is zero (i.e. $x+ys=0$) such that the return is greater than or equal to zero with positive probability. Such a strategy is called an *arbitrage.* This means that the investor is able to achieve a positive probability of future profits with no down-side risk with a net investment of \$0. In mathematical terms, the investor seeks a point (x,y) such that $x+ys=0$ (net cost of the portfolio is zero) and

$$x(1+r)+ysu \geq 0,$$
$$x(1+r)+ysd \geq 0$$

with at least one of the two inequalities strict (so there is never a loss and a non-zero chance of a positive return). Alternatively, is there a point on the line $y=-\frac{1}{s}x$ which lies *above both* of the two lines

$$y=-\frac{1+r}{su}x$$
$$y=-\frac{1+r}{sd}x$$

and strictly above one of them? Since all three lines pass through the origin, we need only compare the slopes; an arbitrage will NOT be possible if

$$-\frac{1+r}{sd} \leq -\frac{1}{s} \leq -\frac{1+r}{su} \tag{2.1}$$

and otherwise there is a point (x,y) permitting an arbitrage. The condition for no arbitrage (2.1) reduces to

$$\frac{d}{1+r} < 1 < \frac{u}{1+r} \tag{2.2}$$

So the condition for no arbitrage demands that $(1+r-u)$ and $(1+r-d)$ have opposite sign or $d \leq (1+r) \leq u$. Unless this occurs, the stock *always* has either better or worse returns than the bond, which makes no sense in a

free market where both are traded without compulsion. Under a no arbitrage assumption since $d \leq (1+r) \leq u$, the bond payoff is a *convex combination* or a weighted average of the two possible stock payoffs; i.e. there are probabilities $0 \leq q \leq 1$ and $(1-q)$ such that $(1+r) = qu + (1-q)d$. In fact it is easy to solve this equation to determine the values of q and $1-q$.

$$q = \frac{(1+r)-d}{u-d}, \qquad \text{and} \quad 1-q = \frac{u-(1+r)}{u-d}.$$

Denote by Q the probability distribution which puts probabilities q and $1-q$ on these points su, sd. Then if S_1 is the value of the stock at the end of the period, note that

$$\frac{1}{1+r}E_Q(S_1) = \frac{1}{1+r}(qsu + (1-q)sd) = \frac{1}{1+r}s(1+r) = s$$

where E_Q denotes the expectation assuming that Q describes the probabilities of the two outcomes.

In other words, *if there is to be no arbitrage, there exists a probability measure Q such that the expected price of future value of the stock S_1 discounted to the present using the return from a risk-free bond is exactly the present value of the stock.* The measure Q is called the *risk-neutral* measure and the probabilities that it assigns to the possible outcomes of S are not necessarily those that determine the future behaviour of the stock. The risk neutral measure embodies both the current consensus beliefs in the future value of the stock and the consensus investors' attitude to risk avoidance. It is not usually true that $\frac{1}{1+r}E_P(S_1) = s$ with P denoting the actual probability distribution describing the future probabilities of the stock. Indeed it is highly unlikely that an investor would wish to purchase a risky stock if he or she could achieve exactly the same expected return with no risk at all using a bond. We generally expect that to make a risky investment attractive, its expected return should be greater than that of a risk-free investment. Notice in this example that the risk-neutral measure Q did not use the probabilities p, and $1-p$ that the stock would go

up or down and this seems contrary to intuition. Surely if a stock is more likely to go up, then a call option on the stock should be valued higher!

Let us suppose for example that we have a friend willing, in a private transaction with me, to buy or sell a stock at a price determined from his subjectively assigned distribution P, different from Q. The friend believes that the stock is presently worth

$$\frac{1}{1+r}E_P S_1 = \frac{psu + (1-p)sd}{1+r} \neq s \text{ since } p \neq q.$$

Such a friend offers their assets as a sacrifice to the gods of arbitrage. If the friend's assessed price is greater than the current market price, we can buy on the open market and sell to the friend. Otherwise, one can do the reverse. Either way one is enriched monetarily (and perhaps impoverished socially)!

So why should we use the Q measure to determine the price of a given asset in a market (assuming, of course, there is a risk-neutral Q measure and we are able to determine it)? Not because it precisely describes the future behaviour of the stock, *but because if we use any other distribution, we offer an intelligent investor (there are many!) an arbitrage opportunity, or an opportunity to make money at no risk and at our expense.*

Derivatives are investments which derive their value from that of a corresponding asset, such as a stock. A *European call option* is an option which permits you (but does not compel you) to purchase the stock at a fixed future date (the *maturity date*) or for a given predetermined price, the *exercise price* of the option). For example a call option with exercise price \$10 on a stock whose future value is denoted S_1, is worth on expiry $S_1 - 10$ if $S_1 > 10$ but nothing at all if $S_1 < 10$. The difference $S_1 - 10$ between the value of the stock on expiry and the exercise price of the option is your profit if you exercises the option, purchasing the stock for \$10 and sell it on the open market at $\$S_1$. However, if $S_1 < 10$, there is no point in exercising your option as you are not compelled to do so and your return is \$0. In general, your payoff from pur-

chasing the option is a simple function of the future price of the stock, such as $V(S_1) = \max(S_1 - 10,\, 0)$. We denote this by $(S_1 - 10)^+$. The future value of the option is a random variable but it derives its value from that of the stock, hence it is called a *derivative* and the stock is the *underlying.*

A function of the stock price $V(S_1)$ which may represent the return from a portfolio of stocks and derivatives is called a *contingent claim.* $V(S_1)$ represents the payoff to an investor from a certain financial instrument or derivative when the stock price at the end of the period is S_1. In our simple binomial example above, the random variable takes only two possible values $V(su)$ and $V(sd)$. We will show that there is a portfolio, called a *replicating* portfolio, consisting of an investment solely in the above stock and bond which reproduces these values $V(su)$ and $V(sd)$ exactly. We can determine the corresponding weights on the bond and stocks (x, y) simply by solving the two equations in two unknowns

$$x(1+r) + ysu = V(su)$$
$$x(1+r) + ysd = V(sd)$$

Solving: $y^* = \frac{V(su)-V(sd)}{su-sd}$ and $x^* = \frac{V(su)-y^*su}{1+r}$. By buying y^* units of stock and x^* units of bond, we are able to replicate the contingent claim $V(S_1)$ exactly- i.e. produce a portfolio of stocks and bonds with exactly the same return as the contingent claim. So in this case at least, there can be only one possible present value for the contingent claim and that is the present value of the replicating portfolio $x^* + y^*s$. If the market placed any other value on the contingent claim, then a trader could guarantee a positive return by a simple trade, shorting the contingent claim and buying the equivalent portfolio or buying the contingent claim and shorting the replicating portfolio. Thus this is the only price that precludes an arbitrage opportunity. There is a simpler

expression for the current price of the contingent claim in this case: Note that

$$\begin{aligned}\frac{1}{1+r}E_Q V(S_1) &= \frac{1}{1+r}(qV(su)+(1-q)V(sd)) \\ &= \frac{1}{1+r}(\frac{1+r-d}{u-d}V(su)+\frac{u-(1+r)}{u-d}V(sd)) \\ &= x^* + y^* s.\end{aligned}$$

In words, *the discounted expected value of the contingent claim is equal to the no-arbitrage price of the derivative where the expectation is taken using the Q-measure*. Indeed any contingent claim that is attainable must have its price determined in this way. While we have developed this only in an extremely simple case, it extends much more generally.

Suppose we have a total of N risky assets whose prices at times $t = 0, 1$, are given by $(S_0^j, S_1^j), j = 1, 2, ..., N$. We denote by S_0, S_1 the column vector of initial and final prices

$$S_0 = \begin{pmatrix} S_0^1 \\ S_0^2 \\ . \\ . \\ . \\ S_0^N \end{pmatrix}, S_1 = \begin{pmatrix} S_1^1 \\ S_1^2 \\ . \\ . \\ . \\ S_1^N \end{pmatrix}$$

where at time 0, S_0 is known and S_1 is random. Assume also there is a riskless asset (a bond) paying interest rate r over one unit of time. Suppose we borrow money (this is the same as shorting bonds) at the risk-free rate to buy w_j units of stock j at time 0 for a total cost of $\sum w_j S_0^j$. The value of this portfolio at time $t = 1$ is $T(w) = \sum w_j(S_1^j - (1+r)S_0^j)$. If there are weights w_j so that this sum is always non-negative, and $P(T(w) > 0) > 0$, then this is an arbitrage opportunity. Similarly, by replacing the weights w_j by their negative $-w_j$, there is an arbitrage opportunity if for some weights the sum is non-positive and negative with positive probability. In summary, there are *no arbitrage op-*

portunities if for all weights w_j $P(T(w) > 0) > 0$ and $P(T(w) < 0) > 0$ so $T(w)$ takes both positive and negative values. We assume that the moment generating function $M(w) = E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))]$ exists and is an analytic function of w.Roughly the condition that the moment generating function is analytic assures that we can expand the function in a series expansion in w. This is the case, for example, if the values of S_1, S_0 are bounded. The following theorem provides a general proof, due to Chris Rogers, of the equivalence of the no-arbitrage condition and the existence of an equivalent measure Q. Refer to the appendix for the technical definitions of an equivalent probability measure and the existence and properties of a moment generating function $M(w)$.

Theorem 2 *A necessary and sufficient condition that there be no arbitrage opportunities is that there exists a measure Q equivalent to P such that $E_Q(S_1^j) = \frac{1}{1+r}S_0^j$ for all $j = 1, ..., N$.*

Proof. *Define $M(w) = E\exp(T(w)) = E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))]$ and consider the problem*

$$\min_w \ln(M(w)).$$

The no-arbitrage condition implies that for each j there exists $\varepsilon > 0$,

$$P[S_1^j - (1+r)S_0^j > \varepsilon] > 0$$

and therefore as $w_j \to \infty$ while the other weights $w_k, k \neq j$ remain fixed,

$$M(w) = E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))] > C\exp(w_j\varepsilon)P[S_1^j - (1+r)S_0^j > \varepsilon] \to \infty \quad as\ w_j \to \infty.$$

Similarly, $M(w) \to \infty$ as $w_j \to -\infty$. From the properties of a moment generating function (see the appendix) $M(w)$ is convex, continuous, analytic and $M(0) = 1$. Therefore the function $M(w)$ has a minimum w^* satisfying $\frac{\partial M}{\partial w_j} = 0$ or

$$\frac{\partial M(w)}{\partial w_j} = 0 \quad or \tag{2.3}$$

$$E[S_1^j \exp(T(w))] = (1+r)S_0^j E[\exp(T(w))]$$

or

$$S_0^j = \frac{E[\exp(T(w))S_1^j]}{(1+r)E[\exp(T(w))]}.$$

Define a distribution or probability measure Q as follows; for any event A,

$$Q(A) = \frac{E_P[I_A \exp(w'S_1)]}{E_P[\exp(w'S_1)]}.$$

The Radon-Nikodym derivative (see the appendix) is

$$\frac{dQ}{dP} = \frac{\exp(w'S_1)]}{E_P[\exp(w'S_1)]}.$$

Since $\infty > \frac{dQ}{dP} > 0$, the measure Q is equivalent to the original probability measure P (in the intuitive sense that it has the same support). When we calculate expected values under this new measure, note that for each j,

$$\begin{aligned} E_Q(S_1^j) &= E_P[\frac{dQ}{dP}S_1^j] \\ &= \frac{E_P[S_1^j \exp(w'S_1)]}{E_P[\exp(w'S_1)]} \\ &= (1+r)S_0^j. \end{aligned}$$

or

$$S_0^j = \frac{1}{1+r}E_Q(S_1^j).$$

Therefore, the current price of each stock is the discounted expected value of the future price under this "risk-neutral" measure Q.

Conversely if

$$E_Q(S_1^j) = \frac{1}{1+r}S_0^j, \quad \textit{for all } j \tag{2.4}$$

holds for some measure Q then $E_Q[T(w)] = 0$ for all w and this implies that the random variable $T(w)$ is either identically 0 or admits both positive and negative values. Therefore the existence of the measure Q satisfying (2.4) implies that there are no arbitrage opportunities. ■

The so-called risk-neutral measure Q is constructed to minimize the cross-entropy between Q and P subject to the constraints $E(S_1 - (1+r)S_0) = 0$

where cross-entropy is defined in Section 1.5. If there N possible values of the random variables S_1 and S_0 then (2.3) consists of N equations in N unknowns and so it is reasonable to expect a unique solution. In this case, the Q measure is unique and we call the market *complete*.

The theory of pricing derivatives in a complete market is rooted in a rather trivial observation because in a complete market, the derivative can be replicated with a portfolio of other marketable securities. If we can reproduce exactly the same (random) returns as the derivative provides using a linear combination of other marketable securities (which have prices assigned by the market) then the derivative must have the same price as the linear combination of other securities. Any other price would provide arbitrage opportunities.

Of course in the real world, there are costs associated with trading, these costs usually related to a bid-ask spread. There is essentially a different price for buying a security and for selling it. The argument above assumes a frictionless market with no trading costs, with borrowing any amount at the risk-free bond rate possible, and a completely liquid market- any amount of any security can be bought or sold. Moreover it is usually assumed that the market is complete and it is questionable whether complete markets exist. For example if a derivative security can be perfectly replicated using other marketable instruments, then what is the purpose of the derivative security in the market? All models, excepting those on *Fashion File,* have deficiencies and critics. The merit of the frictionless trading assumption is that it provides an accurate approximation to increasingly liquid real-world markets. Like all useful models, this permits tentative conclusions that should be subject to constant study and improvement.

Multiperiod Models.

When an asset price evolves over time, the investor normally makes decisions about the investment at various periods during its life. Such decisions are made

with the benefit of current information, and this information, whether used or not, includes the price of the asset and any related assets at all previous time periods, beginning at some time $t = 0$ when we began observation of the process. We denote this information available for use at time t as H_t. Formally, H_t is what is called a *sigma-field* (see the appendix) generated by the past, and there are two fundamental properties of this sigma-field that will use. The first is that the sigma-fields increase over time. In other words, our information about this and related processes increases over time because we have observed more of the relevant history. In the mathematical model, we do not "forget" relevant information: this model fits better the behaviour of youthful traders than aging professors. The second property of H_t is that it includes the value of the asset price $S_\tau, \tau \leq t$ at all times $\tau \leq t$. In measure-theoretic language, S_t *is adapted to or measurable with respect to* H_t. Now the analysis above shows that when our investment life began at time $t = 0$ and we were planning for the next period of time, absence of arbitrage implies a risk-neutral measure Q such that $E_Q(\frac{1}{1+r}S_1) = S_0$. Imagine now that we are in a similar position at time t, planning our investment for the next unit time. All expected values should be taken in the light of our current knowledge, i.e. given the information H_t. An identical analysis to that above shows that under the risk neutral measure Q, if S_t represents the price of the stock after t periods, and r_t the risk-free one-period interest rate offered that time, then

$$E_Q(\frac{1}{1+r_t}S_{t+1}|H_t) = S_t. \tag{2.5}$$

Suppose we let B_t be the value of \$1 invested at time $t = 0$ after a total of t periods. Then $B_1 = (1 + r_0)$, $B_2 = (1 + r_0)(1 + r_1)$, and in general $B_t = (1+r_0)(1+r_1)...(1+r_{t-1})$. Since the interest rate per period is announced at the beginning of this period, the value B_t is known at time $t - 1$. If you owe exactly \$1.00 payable at time t, then to cover this debt you should have an

investment at time $t = 0$ of $\$E(1/B_t)$, which we might call the *present value* of the promise. In general, at time t, the present value of a certain amount $\$V_T$ promised at time T (i.e. the present value or the value discounted to the present of this payment) is

$$E(V_T \frac{B_t}{B_T} | H_t).$$

Now suppose we divide (2.5) above by B_t. We obtain

$$E_Q(\frac{S_{t+1}}{B_{t+1}}|H_t) = E_Q(\frac{1}{B_t(1+r_t)}S_{t+1}|H_t) = \frac{1}{B_t}E_Q(\frac{1}{1+r_t}S_{t+1}|H_t) = \frac{S_t}{B_t}. \tag{2.6}$$

Notice that we are able to take the divisor B_t outside the expectation since B_t is known at time t (in the language of Appendix 1, B_t is measurable with respect to H_{t+1}). This equation (2.6) describes an elegant mathematical property shared by all marketable securities in a complete market. Under the risk-neutral measure, the discounted price $Y_t = S_t/B_t$ forms a *martingale*. A *martingale* is a process Y_t for which the expectation of a future value given the present is equal to the present i.e.

$$E(Y_{t+1}|H_t) = Y_t \text{.for all } t. \tag{2.7}$$

Properties of a martingale are given in the appendix and it is easy to show that for such a process, when $T > t$,

$$E(Y_T|H_t) = E[...E[E(Y_T|H_{T-1})|H_{T-2}]...|H_t] = Y_t. \tag{2.8}$$

A martingale is a fair game in a world with no inflation, no need to consume and no mortality. Your future fortune if you play the game is a random variable whose expectation, given everything you know at present, is your present fortune.

Thus, under a risk-neutral measure Q in a complete market, all marketable securities discounted to the present form martingales. For this reason, we often refer to the risk-neutral measure as a martingale measure. The fact that prices of

marketable commodities must be martingales under the risk neutral measure has many consequences for the canny investor. Suppose, for example, you believe that you are able to model the history of the price process nearly perfectly, and it tells you that the price of a share of XXX computer systems increases on average 20% per year. Should you use this $P-$measure in valuing a derivative, even if you are confident it is absolutely correct, in pricing a call option on XXX computer systems with maturity one year from now? If you do so, you are offering some arbitrager another free lunch at your expense. The measure Q, not the measure P, determines derivative prices in a no-arbitrage market. This also means that there is no advantage, when pricing derivatives, in using some elaborate statistical method to estimate the expected rate of return because this is a property of P not Q.

What have we discovered? In general, prices in a market are determined as expected values, but expected values with respect to the measure Q. This is true in any complete market, regardless of the number of assets traded in the market. For any future time $T > t$, and for any derivative defined on the traded assets in a market whose value at time t is given by V_t, $E_Q(\frac{B_t}{B_T}V_T|H_t] = V_t =$ the market price of the derivative at time t. So in theory, determining a reasonable price of a derivative should be a simple task, one that could be easily handled by simulation. Suppose we wish to determine a suitable price for a derivative whose value is determined by some stock price process S_t. Suppose that at time $T > t$, the value of the derivative is a simple function of the stock price at that time $V_T = V(S_T)$. We may simply generate many simulations of the future value of the stock and corresponding value of the derivative $S_T, V(S_T)$ given the current store of information H_t. These simulations must be conducted under the measure Q. In order to determine a fair price for the derivative, we then average the discounted values of the derivatives, discounted to the present, over all the simulations. The catch is that the Q measure is often neither obvious from the present market prices nor statistically estimable from its past. It is given

implicitly by the fact that the expected value of the discounted future value of traded assets must produce the present market price. In other words, a first step in valuing any asset is to determine a measure Q for which this holds. Now in some simple models involving a single stock, this is fairly simple, and there is a unique such measure Q. This is the case, for example, for the stock model above in which the stock moves in simple steps, either increasing or decreasing at each step. But as the number of traded assets increases, and as the number of possible jumps per period changes, a measure Q which completely describes the stock dynamics and which has the necessary properties for a risk neutral measure becomes potentially much more complicated as the following example shows.

Solving for the Q Measure.

Let us consider the following simple example. Over each period, a stock price provides a return greater than, less than, or the same as that of a risk free investment like a bond. Assume for simplicity that the stock changes by the factor $u(1+r)$ (greater) or $(1+r)$ (the same) $d(1+r)$(less) where $u > 1 > d = 1/u$. The Q probability of increases and decreases is unknown, and may vary from one period to the next. Over two periods, the possible paths executed by this stock price process are displayed below assuming that the stock begins at time $t = 0$ with price $S_0 = 1$.

[FIGURE 2.1 ABOUT HERE]

In general in such a tree there are three branches from each of the nodes at times $t = 0, 1$ and there are a total of $1 + 3 = 4$ such nodes. Thus, even if we assume that probabilities of up and down movements do not depend on how the process arrived at a given node, there is a total of $3 \times 4 = 12$ unknown parameters. Of course there are constraints; for example the sum of the three probabilities on branches exiting a given node must add to one and the price

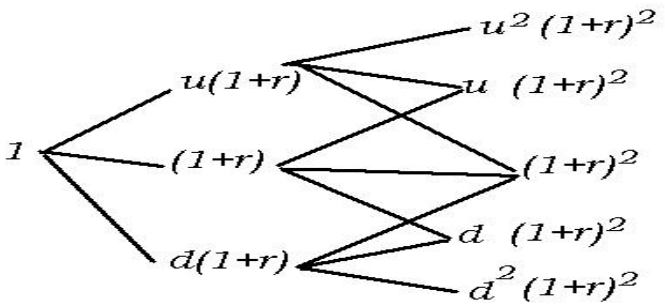


Figure 2.1: A Trinomial Tree for Stock Prices

process must form a martingale. For each of the four nodes, this provides two constraints for a total of 8 constraints, leaving 4 parameters to be estimated. We would need the market price of 4 different derivatives or other contingent claims to be able to generate 4 equations in these 4 unknowns and solve for them. Provided we are able to obtain prices of four such derivatives, then we can solve these equations. If we denote the risk-neutral probability of 'up' at each of the four nodes by p_1, p_2, p_3, p_4 then the conditional distribution of S_{t+1} given $S_t = s$ is:

Stock value	$su(1+r)$	$s(1+r)$	$sd(1+r)$
Probability	p_i	$1-\frac{u-d}{1-d}p_i = 1-kp_i$	$\frac{u-1}{1-d}p_i = cp_i$

Consider the following special case, with the risk-free interest rate per period r, $u = 1.089$, $S_0 = \$1.00$. We also assume that we are given the price of four call options expiring at time $T = 2$. The possible values of the price at time $T = 2$ corresponding to two steps up, one step up and one constant, one up one down, etc. are the values of $S(T)$ in the set

$$\{1.1859, 1.0890, 1.0000, 0.9183, 0.8432\}.$$

Now consider a "call option" on this stock expiring at time $T = 2$ with strike

price K. Such an option has value at time $t = 2$ equal to $(S_2 - K)$ if this is positive, or zero otherwise. For brevity we denote this by $(S_2 - K)^+$. The present value of the option is $E_Q(S_2 - K)^+$ discounted to the present, where K is the exercise price of the option and S_2 is the price of the stock at time 2. Thus the price of the call option at time 0 is given by

$$V_0 = E_Q(S_2 - K)^+/(1 + r)^2$$

Assuming interest rate $r = 1\%$ per period, suppose we have market prices of four call options with the same expiry and different exercise prices in the following table;

K =Exercise Price	T =Maturity	V_0 =Call Option Price
0.867	2	0.154
0.969	2	.0675
1.071	2	.0155
1.173	2	.0016

If we can observe the prices of these options only, then the equations to be solved for the probabilities associated with the measure Q equate the observed price of the options to their theoretical price $V_0 = E(S_2 - K)^+/(1 + r)^2$.

$$0.0016 = \frac{1}{(1.01)^2}(1.186 - 1.173)p_1p_2$$

$$0.0155 = \frac{1}{(1.01)^2}[(1.186 - 1.071)p_1p_2 + (1.089 - 1.071)\{p_1(1 - kp_2) + (1 - kp_1)p_2\}]$$

$$\begin{aligned} 0.0675 = \frac{1}{(1.01)^2}[&0.217p_1p_2 + 0.12\{p_1(1 - kp_2) + (1 - kp_1)p_2\} \\ &+ 0.031\{(1 - kp_1)(1 - kp_2) + cp_1p_2 + cp_1p_4)\} \end{aligned}$$

$$\begin{aligned} 0.154 = \frac{1}{(1.01)^2}[&0.319p_1p_2 + 0.222\{p_1(1 - kp_2) + (1 - kp_1)p_2\} \\ &+ 0.133\{(1 - kp_1)(1 - kp_2) + cp_1p_2 + cp_1p_4)\} \\ &+ 0.051\{\{cp_1(1 - kp_4) + (1 - kp_1)cp_3\}]. \end{aligned}$$

While it is not too difficult to solve this system in this case one can see that with more branches and more derivatives, this non-linear system of equations becomes difficult very quickly. What do we do if we observe market prices for only two derivatives defined on this stock, and only two parameters can be obtained from the market information? This is an example of what is called an incomplete market, a market in which the risk neutral distribution is not uniquely specified by market information. In general when we have fewer equations than parameters in a model, there are really only two choices

(a) Simplify the model so that the number of unknown parameters and the number of equations match.

(b) Determine additional natural criteria or constraints that the parameters must satisfy.

In this case, for example, one might prefer a model in which the probability of a step up or down depends on the time, but not on the current price of the stock. This assumption would force equal all of $p_2 = p_3 = p_4$ and simplify the system of equations above. For example using only the prices of the first two derivatives, we obtain equations, which, when solved, determine the probabilities on the other branches as well.

$$0.0016 = \frac{1}{(1.01)^2}(1.186 - 1.173)p_1p_2$$
$$0.0155 = \frac{1}{(1.01)^2}[(1.186 - 1.071)p_1p_2 + (1.089 - 1.071)\{p_1(1 - kp_2) + (1 - kp_1)p_2\}]$$

This example reflects a basic problem which occurs often when we build a reasonable and flexible model in finance. Frequently there are more parameters than there are marketable securities from which we can estimate these parameters. It is quite common to react by simplifying the model. For example, it is for this reason that binomial trees (with only two branches emanating from each node) are often preferred to the trinomial tree example we use above, even though they provide a worse approximation to the actual distribution of stock

returns.

In general if there are n different securities (excluding derivatives whose value is a function of one or more of these) and if each security can take any one of m different values, then there are a total of m^n possible states of nature at time $t = 1$. The Q measure must assign a probability to each of them. This results in a total of m^n unknown probability values, which, of course must add to one, and result in the right expectation for each of n marketable securities. To uniquely determine Q we would require a total of $m^n - n - 1$ equations or $m^n - n - 1$ different derivatives. For example for $m = 10$, $n = 100$, approximately one with a hundred zeros, a prohibitive number, are required to uniquely determine Q. In a complete market, Q is uniquely determined by marketed securities, but in effect no real market can be complete. In real markets, one asset is not perfectly replicated by a combination of other assets because there is no value in duplication. Whether an asset is a derivative whose value is determined by another marketed security, together with interest rates and volatilities, markets rarely permit exact replication. The most we can probably hope for in practice is to find a model or measure Q in a subclass of measures with desirable features under which

$$E_Q[\frac{B_t}{B_T}V(S_T)|H_t] \approx V_t \text{ for all marketable } V. \tag{2.9}$$

Even if we had equalities in (2.9), this would represent typically fewer equations than the number of unknown Q probabilities so some simplification of the model is required before settling on a measure Q. One could, at one's peril, ignore the fact that certain factors in the market depend on others. Similar stocks behave similarly, and none may be actually independent. Can we, with any reasonable level of confidence, accurately predict the effect that a lowering of interest rates will have on a given bank stock? Perhaps the best model for the future behaviour of most processes is the past, except that as we have seen the historical distribution of stocks do not generally produce a risk-neutral

measure. Even if historical information provided a flawless guide to the future, there is too little of it to accurately estimate the large number of parameters required for a simulation of a market of reasonable size. Some simplification of the model is clearly necessary. Are some baskets of stocks independent of other combinations? What independence can we reasonably assume over time?

As a first step in simplifying a model, consider some of the common measures of behaviour. Stocks can go up, or down. The drift of a stock is a tendency in one or other of these two directions. But it can also go *up and down*- by a lot or a little. The measure of this, the variance or variability in the stock returns is called the *volatility* of the stock. Our model should have as ingredients these two quantities. It should also have as much dependence over time and among different asset prices as we have evidence to support.

Determining the Process B_t.

We have seen in the last section that given the Q or risk-neutral measure, we can, at least in theory, determine the price of a derivative if we are given the price B_t of a risk-free investment at time t (in finance such a yardstick for measuring and discounting prices is often called a "numeraire"). Unfortunately no completely liquid risk-free investment is traded on the open market. There are government treasury bills which, depending on the government, one might wish to assume are almost risk-free, and there are government bonds, usually with longer terms, which complicate matters by paying dividends periodically. The question dealt with in this section is whether we can estimate or approximate an approximate risk-free process B_t given information on the prices of these bonds. There are typically too few genuinely risk-free bonds to get a detailed picture of the process $B_s, s > 0$. We might use government bonds for this purpose, but are these genuinely risk-free? Might not the additional use of bonds issued by other large corporations provide a more detailed picture of the bank account process B_s?

Can we incorporate information on bond prices from lower grade debt? To do so, we need a simple model linking the debt rating of a given bond and the probability of default and payoff to the bond-holders in the event of default. To begin with, let us assume that a given basket of companies, say those with a common debt rating from one of the major bond rating organisations, have a common distribution of default time. The thesis of this section is that even if no totally risk-free investment existed, we might still be able to use bond prices to estimate what interest rate such an investment would offer.

We begin with what we know. Presumably we know the current prices of marketable securities. This may include prices of certain low-risk bonds with face value F, the value of the bond on maturity at time T. Typically such a bond pays certain payments of value d_t at certain times $t < T$ and then the face value of the bond F at maturity time T, *unless the bond-holder defaults.* Let us assume for simplicity that the current time is 0. The current bond prices P_0 provide some information on B_t as well as the possibility of default. Suppose we let τ denote the random time at which default or bankruptcy would occur. Assume that the effect of possible default is to render the payments at various times random so for example d_t is paid provided that default has not yet occurred, i.e. if $\tau > t$, and similarly the payment on maturity is the face value of the Bond F if default has not yet occurred and if it has, some fraction of the face value pF is paid. When a real bond defaults, the payout to bondholders is a complicated function of the hierarchy of the bond and may occur before maturity, but we choose this model with payout at maturity in any case for simplicity. Then the current price of the bond is the expected discounted value of all future payments, so

$$\begin{aligned} P_0 &= E_Q\Big(\sum_{\{s;0<s<T\}} \frac{1}{B_s} d_s I(\tau > s) + \frac{pF}{B_T} I(\tau \le T) + \frac{F}{B_T} I(\tau > T)\Big) \\ &= \sum_{\{s;0<s<T\}} d_s E_Q[B_s^{-1} I(\tau > s)] + F E_Q[B_T^{-1}(p + (1-p) I(\tau > T))] \end{aligned}$$

The bank account process B_t that we considered is the compounded value at time of an investment of \$1 deposited at time 0. This value might be random but the interest rate is declared at the beginning of each period so, for example, B_t is completely determined at time $t-1$. In measure-theoretical language, B_t is H_{t-1} measurable for each t. With Q is the risk-neutral distribution

$$P_0 = E_Q\{\sum_{\{s;0<s<T\}} d_s B_s^{-1} Q(\tau > s|H_{s-1}) + FB_T^{-1}(p+(1-p)Q(\tau > T|H_{T-1}))\}.$$

This takes a form very similar to the price of a bond which does not default but with a different bank account process. Suppose we define a new bank account process $\widetilde{B_s}$, equivalent in expectation to the risk-free account, but that only pays if default does not occur in the interval. Such a process must satisfy

$$E_Q(\widetilde{B_s} I(\tau > s)|H_{s-1}) = B_s.$$

From this we see that the process $\widetilde{B_s}$ is defined by

$$\widetilde{B_s} = \frac{B_s}{Q[\tau > s|H_{s-1}]} \text{ on the set } Q[\tau > s|H_{s-1}] > 0.$$

In terms of this new bank account process, the price of the bond can be rewritten as

$$P_0 = E_Q\{\sum_{\{s;0<s<T\}} d_s \widetilde{B_s^{-1}} + (1-p)F\widetilde{B_T^{-1}} + pFB_T^{-1}\}.$$

If we subtract from the current bond price the present value of the guaranteed payment of pF, the result is

$$P_0 - pFE_Q(B_T^{-1}) = E_Q\{\sum_{\{s;0<s<T\}} d_s \widetilde{B_s^{-1}} + (1-p)F\widetilde{B_T^{-1}}\}.$$

This equation has a simple interpretation. The left side is the price of the bond reduced by the present value of the guaranteed payment on maturity Fp. The right hand side is the current value of a risk-free bond paying the same dividends, with interest rates increased by replacing B_s by $\widetilde{B_s}$ and with face value $F(1-p)$ all discounted to the present using the bank account process

$\widetilde{B_s}$. In words, *to value a defaultable bond, augment the interest rate using the probability of default in intervals, change the face value to the potential loss of face value on default and then add the present value of the guaranteed payment on maturity.*

Typically we might expect to be able to obtain prices of a variety of bonds issued on one firm, or firms with similar credit ratings. If we are willing to assume that such firms share the same conditional distribution of default time $Q[\tau > s|H_{s-1}]$ then they must all share the same process $\widetilde{B_s}$ and so each observed bond price P_0 leads to an equation of the form

$$P_0 = \sum_{\{s;0<s<T\}} d_s \widetilde{v_s} + (1-p)F\widetilde{B_T^{-1}} + pFv_T.$$

in the unknowns $\widetilde{v}_s = E_Q(\widetilde{B_s^{-1}}), ...s \leq T.$ and $v_T = E_Q(B_T^{-1})$. If we assume that the coupon dates of the bonds match, then k bonds of a given maturity T and credit rating will allow us to estimate the k unknown values of $\widetilde{v}_s$. Since the term v_T is included in all bonds, it can be estimated from all of the bond prices, but most accurately from bonds with very low risk.

Unfortunately, this model still has too many unknown parameters to be generally useful. We now consider a particular case that is considerably simpler. While it seems unreasonable to assume that default of a bond or bankruptcy of a firm is unrelated to interest rates, one might suppose some simple model which allows a form of dependence. For most firms, one might expect that the probability of survival another unit time is negatively associated with the interest rate. For example we might suppose that the probability of default in the next time interval conditional on surviving to the present is a function of the current interest rate, for example

$$h_t = Q(\tau = t|\tau \geq t, r_t) = \frac{a + (b-1)r_t}{1 + a + br_t}.$$

The quantity h_t is a more natural measure of the risk at time t than are other measures of the distribution of τ and the function h_t is called the hazard

function. If the constant $b > 1+a$, then the "hazard" h_t increases with increasing interest rates, otherwise it decreases. In case the default is independent of the interest rates, we may put $b = 1+a$ in which case the hazard is $a/(1+a)$. Then on the set $[\tau \geq s]$

$$\widetilde{B_s} = \frac{1+r_s}{1-h_s}\widetilde{B}_{s-1} = (1+a+br_s)\widetilde{B}_{s-1}$$

which means that the bond is priced using a similar bank account process but one for which the effective interest rate is not r_s but $a + br_s$. The difference $a + (b-1)r_s$ between the effective interest rate and r_s is usually referred to as the *spread* and this model justifies using a linear function to model this spread. Now suppose that default is assumed independent of the past history of interest rates under the risk-neutral measure Q. In this case, $b = 1 + a$ and the spread is $a(1 + r_s) \simeq a \simeq a/(1 + a)$ provided both a and r_s is small. So in this case the spread gives an approximate risk-neutral probability of default in a given time interval, conditional on survival to that time.

We might hope that the probabilities of default are very small and follow a relatively simple pattern. If the pattern is not perfect, then little harm results provided that indeed the default probabilities are small. Suppose for example that the time of default follows a geometric distribution so that the hazard is constant $h_t = h = a/(1 + a)$. Then

$$\widetilde{B_s} = (1+a)^s B_s \text{ for } s > 0.$$

$\widetilde{B_s}$ grows faster than B_s and it grows even faster as the probability of default h increases. The effective interest rate on this account is approximately a units per period higher.

Given only three bond prices with the same default characteristics, for example, and assuming constant interest rates so that $B_s = (1+r)^s$, we may solve for the values of the three unknown parameters (r, a, p) equations of the form

$$P_0 - pF(1+r)^{-T} = \sum_{0<s<T} (1+a+r+ar)^{-s} d_s + (1-p)F(1+a+r+ar)^{-T}.$$

Market prices for a minimum of three different bonds would allow us to solve for the unknowns (r, a, p) and these are obtainable from three different bonds.

Minimum Variance Portfolios and the Capital Asset Pricing Model.

Let us begin by building a model for portfolios of securities that captures many of the features of market movements. We assume that by using the methods of the previous section and the prices of low-risk bonds, we are able to determine the value B_t of a risk-free investment at time t in the future. Normally these values might be used to discount future stock prices to the present. However for much of this section we will consider only a single period and the analysis will be essentially the same with our without this discounting.

Suppose we have a number n of potential investments or securities, each risky in the sense that prices at future dates are random. Suppose we denote the price of these securities at time t by $S_i(t), i = 1, 2, ..., n$. There is a better measure of the value of an investment than the price of a security or even the change in the price of a security $S_i(t) - S_i(t-1)$ over a period because this does not reflect the cost of our initial investment. A common measure on investments that allows to obtain prices, but is more stable over time and between securities is the *return*. For a security that has prices $S_i(t)$ and $S_i(t+1)$ at times t and $t+1$, we define the return $R_i(t+1)$ on the security over this time interval by

$$R_i(t+1) = \frac{S_i(t+1) - S_i(t)}{S_i(t)}.$$

For example a stock that moved in price from \$10 per share to \$11 per share over a period of time corresponds to a return of 10%. Returns can be measured

in units that are easily understood (for example 5% or 10% per unit time) and are independent of the amount invested. Obviously the \$1 profit obtained on the above stock could has easily been obtained by purchasing 10 shares of a stock whose value per share changed from \$1.00 to \$1.10 in the same period of time, and the return in both cases is 10%. Given a sequence of returns and the initial value of a stock $S_i(0)$, it is easy to obtain the stock price at time t from the initial price at time 0 and the sequence of returns.

$$\begin{aligned} S_i(t) &= S_i(0)(1+R_i(1))(1+R_i(2))...(1+R_i(t)) \\ &= S_i(0)\Pi_{s=1}^{t}(1+R_i(s)). \end{aligned}$$

Returns are not added over time they are multiplied as above. A 10% return followed by a 20% return is not a 30% return but a return equal to $(1+.1)(1+.2)-1$ or 32%. When we buy a portfolio of stocks, the individual stock returns combine in a simple fashion to give the return on the whole portfolio. For example suppose that we wish to invest a total amount $\$I(t)$ at time t. The amounts will change from period to period because we may wish to reinvest gains or withdraw sums from the account. Suppose the proportion of our total investment in stock i at time t is $w_i(t)$ so that the amount invested in stock i is $w_i(t)I(t)$. Note that since $w_i(t)$ are proportions, $\sum_{i=1}^{n} w_i(t) = 1$. What is the return on this investment over the time interval from t to $t+1$? At the end of this period of time, the value of our investment is

$$I(t)\sum_{i=1}^{n} w_i(t)S_i(t+1).$$

If we now subtract the value invested at the beginning of the period and divide by the value at the beginning, we obtain

$$\frac{I(t)\sum_{i=1}^{n} w_i(t)S_i(t+1) - I(t)\sum_{i=1}^{n} w_i(t)S_i(t)}{I(t)\sum_{i=1}^{n} w_i(t)S_i(t)} = \sum_{i=1}^{n} w_i(t)R_i(t+1)$$

which is just a weighted average of the individual stock returns. Note that it does not depend on the initial price of the stocks or the total amount that we

invested at time t. The advantage in using returns instead of stock prices to assess investments is that the *return of a portfolio over a period is a value-weighted average of the returns of the individual investments.*

When time is measured continuously, we might consider defining returns by using the definition above for a period of length h and then reducing h. In other words we could define the instantaneous returns process as

$$\lim_{h\to 0} \frac{S_i(t+h) - S_i(t)}{S_i(t)}.$$

In most cases, the returns over shorter and shorter periods are smaller and smaller, and approach the limit zero so some renormalization is required above. It seems more sensible to consider returns per unit time and then take a limit i.e.

$$R_i(t) = \lim_{h\to 0} \frac{S_i(t+h) - S_i(t)}{hS_i(t)}.$$

Notice that by the definition of the derivative of a logarithm and assuming that this derivative is well-defined,

$$\begin{aligned} \frac{d\ln(S_i(t))}{dt} &= \frac{1}{S_i(t)}\frac{d}{dt}S_i(t) \\ &= \lim_{h\to 0} \frac{S_i(t+h) - S_i(t)}{hS_i(t)} \\ &= R_i(t) \end{aligned}$$

In continuous time, if the stock price process $S_i(t)$ is differentiable, the natural definition of the returns process is the derivative of the logarithm of the stock price. This definition needs some adjustment later because the most common continuous time models for asset prices does not result in a differentiable process $S_i(t)$. The solution we will use then will be to adopt a new concept of an integral and recast the above in terms of this integral.

The Capital Asset Pricing Model (CAPM)

We now consider a simplified model for building a portfolio based on quite basic properties of the potential investments. Let us begin by assuming a single period so that we are planning at time $t = 0$ investments over a period ending at time $t = 1$. We also assume that investors are interested in only two characteristics of a potential investment, the expected value and the variance of the return over this period. We have seen that the return of a portfolio is the value-weighted average of the returns of the individual investments so let us denote the return on stock i by

$$R_i = \frac{S_i(1) - S_i(0)}{S_i(0)},$$

and define $\mu_i = E(R_i)$ and w_i the proportion of my total investment in stock i at the beginning of the period. For brevity of notation, let $\mathbf{R}, \mathbf{w}$ and μ denote the column vectors

$$\mathbf{R} = \begin{pmatrix} R_1 \\ R_2 \\ . \\ . \\ . \\ R_n \end{pmatrix}, \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_n \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_n \end{pmatrix}.$$

Then the return on the portfolio is $\sum_i w_i R_i$ or in matrix notation $\mathbf{w}'\mathbf{R}$. Let us suppose that the covariance matrix of returns is the $n \times n$ matrix $\mathbf{\Sigma}$ so that

$$cov(R_i, R_j) = \Sigma_{ij}.$$

We will frequently use the following properties of expected value and covariance.

Lemma 3 *Suppose*

$$\mathbf{R} = \begin{pmatrix} R_1 \\ R_2 \\ . \\ . \\ . \\ R_n \end{pmatrix}$$

is a column vector of random variables R_i *with* $E(R_i) = \mu_i, i = 1, ..., n$ *and suppose* $\mathbf{R}$ *has covariance matrix* Σ. *Suppose* A *is a non-random vector or matrix with exactly* n *columns so that* $A\mathbf{R}$ *is a vector of random variables. Then* $A\mathbf{R}$ *has mean* $A\mu$ *and covariance matrix* $A\Sigma A'$.

Then it is easy to see that the expected return from the portfolio with weights w_i is $\sum_i w_i E(R_i) = \sum_i w_i \mu_i = \mathbf{w}'\mu$ and the variance is

$$var(\mathbf{w}'\mathbf{R}) = \mathbf{w}'\mathbf{\Sigma}\mathbf{w}.$$

We will need to assume that the covariance matrix $\mathbf{\Sigma}$ is non-singular, that is it has a matrix inverse $\mathbf{\Sigma}^{-1}$. This means, at least for the present, that our model covers only risky stocks for which the variance of returns is positive. If a risk-free investment is available (for example a secure bond whose return is known exactly in advance), this will be handled later.

In the Capital Asset Pricing model it is assumed at the outset that investors concentrate on two measures of return from a portfolio, the expected value and standard deviation. These expected values and variances are computed under the real-world probability distribution P not under some risk-neutral Q measure. Clearly investors prefer high expected return, wherever possible, associated with small standard deviation of return. As a first step in this direction suppose we plot the standard deviation and expected return for the n stocks, i.e. the n points $\{(\sigma_i, \mu_i), i = 1, 2, ..., n\}$ where $\mu_i = E(R_i)$ and $\sigma_i = \sqrt{var(R_i)} = \sqrt{\Sigma_{ii}}$. These n points do not consist of the set of all achievable values of mean and

standard of return, since we are able to construct a portfolio with a certain proportion of our wealth w_i invested in stock i.In fact the set of possible points consists of

$\{(\sqrt{\mathbf{w}'\Sigma\mathbf{w}}, \mathbf{w}'\mu)$ as the vector $\mathbf{w}$ ranges over all possible weights such that $\sum w_i = 1\}$.

The resulting set has a boundary as in Figure 2.2.

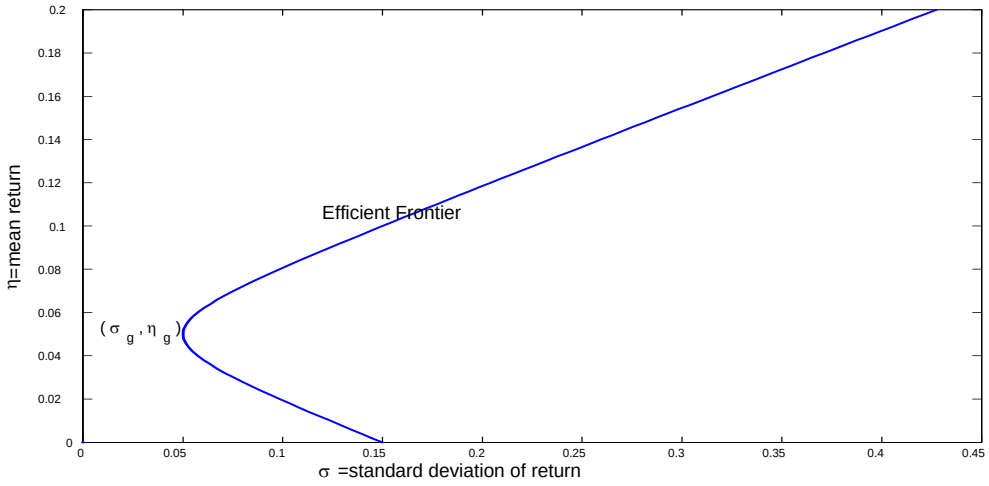


Figure 2.2: The Efficient Frontier

[FIGURE 2.2 ABOUT HERE]

Exactly what form this figure takes depends in part on the assumptions applied to the weights. Since they represent the proportion of our total investment in each of n stocks they must add to one. Negative weights correspond to selling short one stock so as to be able to invest more in another, and we may assume no limit on our ability to do so. In this case the only constraint on $\mathbf{w}$ is the constraint $\sum w_i = 1$. With this constraint alone, we can determine the boundary of the admissible set by fixing the vertical component (the mean return) of a portfolio at some value say η and then finding the minimum possible standard

deviation corresponding to that mean. This allows us to determine the leading edge or left boundary of the region. The optimisation problem is as follows

$$\min \sqrt{\mathbf{w}'\Sigma\mathbf{w}} \text{ subject to}$$

subject to the two constraints on the weights

$$\mathbf{w}'\mathbf{1} = 1$$

$$\mathbf{w}'\mu = \eta.$$

where $\mathbf{1}$ is the column vector of n ones. Since we will often make use of the method of Lagrange multipliers for constrained problems such as this one, we interject a lemma justifying the method. For details, consult Apostol (1973), Section 13.7 or any advanced calculus text.

Lemma 4 *Consider the optimisation problem*

$$\min\{f(w); w \in R^n\} \textit{ subject to } p \textit{ constraints} \tag{2.10}$$

$$\textit{of the form } g_1(w) = 0, g_2(w) = 0, ..., g_p(w) = 0.$$

Then provided the functions $f, g_1, ..., g_p$ are continuously differentiable, a necessary solution for a solution to (2.10) is that there is a solution in the $n + p$ variables $(w_1, ...w_n, \lambda_1, ..., \lambda_p)$ of the equations

$$\frac{\partial}{\partial w_i}\{f(w) + \lambda_1 g_1(w) + ... + \lambda_p g_p(w)\} = 0, i = 1, 2, ..., n$$

$$\frac{\partial}{\partial \lambda_j}\{f(w) + \lambda_1 g_1(w) + ... + \lambda_p g_p(w)\} = 0, j = 1, 2, ..., p.$$

This constants $\lambda_i$are called the Lagrange multipliers and the function that is differentiated, $\{f(w) + \lambda_1 g_1(w) + ... + \lambda_p g_p(w)\}$ is the Lagrangian.

Let us return to our original minimization problem with one small simplification. Since minimizing $\sqrt{\mathbf{w}'\Sigma\mathbf{w}}$ results in the same weight vector $\mathbf{w}$ as does minimizing $\mathbf{w}'\Sigma\mathbf{w}$ we choose the latter as our objective function.

We introduce Lagrange multipliers λ_1, λ_2 and we wish to solve

$$\frac{\partial}{\partial w_i}\{\mathbf{w}^{'}\Sigma\mathbf{w} + \lambda_1(w'\mathbf{1} - 1) + \lambda_2(w'\mu - \eta)\} = 0, i = 1, 2, ..., n$$
$$\frac{\partial}{\partial \lambda_j}\{\mathbf{w}^{'}\Sigma\mathbf{w} + \lambda_1(w'\mathbf{1} - 1) + \lambda_2(w'\mu - \eta)\} = 0, j = 1, 2.$$

The solution is obtained from the simple differentiation rule

$$\frac{\partial}{\partial \mathbf{w}}\mathbf{w}'\mathbf{\Sigma}\mathbf{w} = 2\mathbf{\Sigma}\mathbf{w} \text{ and } \frac{\partial}{\partial \mathbf{w}}\mu'\mathbf{w} = \mathbf{w}$$

and is of the form

$$\mathbf{w} = \lambda_1\Sigma^{-1}\mathbf{1} + \lambda_2\Sigma^{-1}\mu$$

with the Lagrange multipliers λ_1, λ_2chosen to satisfy the two constraints, i.e.

$$\lambda_1\mathbf{1}'\Sigma^{-1}\mu + \lambda_2\mathbf{1}'\Sigma^{-1}\mathbf{1} = 1$$
$$\lambda_1\mu'\Sigma^{-1}\mu + \lambda_2\mu'\Sigma^{-1}\mathbf{1} = \eta.$$

Suppose we define an $n \times 2$ matrix M with columns $\mathbf{1}$ and μ,

$$M = [\ \mathbf{1}\ \ \mu\ ]$$

and the 2×2 matrix $A = (M'\Sigma^{-1}M)^{-1}$, then the Lagrange multipliers are given by the vector

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = A \begin{bmatrix} 1 \\ \eta \end{bmatrix}$$

and the weights by the vector

$$\mathbf{w} = \Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta \end{bmatrix}. \tag{2.11}$$

We are now in a position to identify the boundary or the curve in Figure 2.2. As the mean of the portfolio η changes, the point takes the form $(\sqrt{\mathbf{w}'\Sigma\mathbf{w}}, \eta)$

with $\mathbf{w}$ given by (2.11). Notice that

$$\begin{aligned}
\mathbf{w}'\Sigma\mathbf{w} &= \begin{bmatrix} 1 & \eta \end{bmatrix} A'M'\Sigma^{-1}\Sigma\Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta \end{bmatrix} \\
&= \begin{bmatrix} 1 & \eta \end{bmatrix} A'M'\Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta \end{bmatrix} \\
&= \begin{bmatrix} 1 & \eta \end{bmatrix} A \begin{bmatrix} 1 \\ \eta \end{bmatrix} \\
&= A_{11} + 2A_{12}\eta + A_{22}\eta^2.
\end{aligned}$$

Therefore a point on the boundary $(\sigma, \eta) = (\sqrt{\mathbf{w}'\Sigma\mathbf{w}}, \eta)$ satisfies

$$\sigma^2 - A_{22}\eta^2 - 2A_{12}\eta - A_{11} = 0$$

or

$$\begin{aligned}
\sigma^2 &= A_{22}\eta^2 + 2A_{12}\eta + A_{11} \\
&= \sigma_g^2 + A_{22}(\eta - \eta_g)^2
\end{aligned}$$

where

$$\eta_g = -\frac{A_{12}}{A_{22}} = \frac{\mathbf{1}'\Sigma^{-1}\mu}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \tag{2.12}$$

$$\begin{aligned}
\sigma_g^2 &= A_{11} - \frac{A_{12}^2}{A_{22}} = \frac{|A|}{A_{22}} \\
&= \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}.
\end{aligned} \tag{2.13}$$

and the point (σ_g, μ_g) represents the point in the region corresponding to the minimum possible standard deviation over all portfolios. This is the most conservative investment portfolio available with this class of securities. What weights to do we need to put on the individual stocks to achieve this conservative portfolio? It is easy to see that the weight vector is given by

$$\mathbf{w}_g' = \frac{\mathbf{1}'\Sigma^{-1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \tag{2.14}$$

and since the quantity $\mathbf{1}'\Sigma^{-1}\mathbf{1}$ in the denominator is just a scale factor to insure that the weights add to one, the amount invested in stock i is proportional to the sum of the elements of the i'th row of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$.

An equation of the form

$$\sigma^2 - A_{22}(\eta - \eta_g)^2 = \sigma_g^2$$

represents a hyperbola since $A_{22} > 0$. Of course investors are presumed to prefer higher returns for a given value of the standard deviation of portfolio so it is only the upper boundary of this curve in Figure 2.2 that is efficient in the sense that there is no portfolio that is strictly better (better in the sense of higher return combined with standard deviation that is not larger).

Now let us return to a portfolio whose standard deviation and mean return lie on the efficient frontier. Let us call these *efficient portfolios*. It turns out that any portfolio on this efficient frontier has the same covariance with the minimum variance portfolio $\mathbf{w}_g'\mathbf{R}$ derived above.

Proposition 5 *Every efficient portfolio has the same covariance* $\frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$ *with the conservative portfolio* $\mathbf{w}_g'\mathbf{R}$.

Proof. We noted before that such a portfolio has mean return η and standard deviation σ which satisfy the relation

$$\sigma^2 - A_{22}\eta^2 - 2A_{12}\eta - A_{11} = 0.$$

Moreover the weights for this portfolio are described by

$$\mathbf{w} = \Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta \end{bmatrix}. \tag{2.15}$$

so the returns vector from this portfolio can be written as

$$\mathbf{w}'\mathbf{R} = [\ 1 \quad \eta\ ]AM'\mathbf{\Sigma}^{-1}\mathbf{R}.$$

It is interesting to observe that the covariance of returns between this efficient portfolio and the conservative portfolio $\mathbf{w}_g^{'}\mathbf{R}$ is given by

$$\begin{aligned} cov(\mathbf{w}_g^{'}\mathbf{R}, [\ 1 \quad \eta\ ]AM'\mathbf{\Sigma}^{-1}\mathbf{R}) &= [\ 1 \quad \eta\ ]AM'\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{w}_g \\ &= [\ 1 \quad \eta\ ]A \begin{bmatrix} \mathbf{1}' \\ \mu' \end{bmatrix} \Sigma^{-1}\mathbf{1}\frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \\ &= [\ 1 \quad \eta\ ]A \begin{bmatrix} \mathbf{1}'\Sigma^{-1}\mathbf{1} \\ \mu'\Sigma^{-1}\mathbf{1} \end{bmatrix} \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \\ &= [\ 1 \quad \eta\ ] \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \\ &= \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \end{aligned}$$

where we use the fact that, by the definition of A,

$$A \begin{bmatrix} \mathbf{1}'\Sigma^{-1}\mathbf{1} & \mu'\Sigma^{-1}\mathbf{1} \\ \mu'\Sigma^{-1}\mathbf{1} & \mu'\Sigma^{-1}\mu \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

■

Now consider two portfolios on the boundary in Figure 2.2. For each the weights are of the same form, say

$$\mathbf{w}_p = \Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta_p \end{bmatrix} \quad \text{and} \quad \mathbf{w}_q = \Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta_q \end{bmatrix} \tag{2.16}$$

where the mean returns are η_p and η_q respectively. Consider the covariance between these two portfolios

$$\begin{aligned} cov(\mathbf{w}_p'\mathbf{R}, \mathbf{w}_q'\mathbf{R}) &= \mathbf{w}_p'\mathbf{\Sigma}\mathbf{w}_q \\ &= [\ 1 \quad \eta_p\ ](M'\Sigma^{-1}M)^{-1} \begin{bmatrix} 1 \\ \eta_q \end{bmatrix} \\ &= A_{11} + A_{12}(\eta_p + \eta_q) + A_{22}\eta_p\eta_q \\ &= var(\mathbf{w}_p'\mathbf{R}) - [\ 1 \quad \eta_p\ ]A \begin{bmatrix} 0 \\ \eta_p - \eta_q \end{bmatrix} \end{aligned}$$

An interesting special portfolio that is a "zero-beta" portfolio, one that is perfectly uncorrelated with the portfolio with weights $\mathbf{w}_p'\mathbf{R}$. This is obtained by setting the above covariance equal to 0 and solving we obtain

$$\begin{aligned}\eta_q &= -\frac{A_{11}+A_{12}\eta_p}{A_{12}+A_{22}\eta_p}\\ &= \frac{\mu'\Sigma^{-1}\mu-(\mu'\Sigma^{-1}\mathbf{1})\eta_p}{\mu'\Sigma^{-1}\mathbf{1}-(\mathbf{1}'\Sigma^{-1}\mathbf{1})\eta_p}.\end{aligned}$$

There is a simple method for determining the point $(,\eta_q)$ graphically indicated in Figure **??**. From the equation relating points on the boundary,

$$\sigma^2 - A_{22}(\eta-\eta_g)^2 = \sigma_g^2$$

we obtain

$$\frac{\partial\eta}{\partial\sigma} = \frac{\sigma}{A_{22}(\eta-\eta_g)}$$

and so the tangent line at the point (σ_p,η_p) strikes the $\sigma=0$ axis at a point η_q which satisfies

$$\frac{\eta_p-\eta_q}{\sigma_p} = \frac{\sigma_p}{A_{22}(\eta_p-\eta_g)}$$

or

$$\begin{aligned}\eta_q &= \eta_p - \frac{\sigma_p^2}{A_{22}(\eta_p-\eta_g)}\\ &= \eta_p - \frac{A_{22}\eta_p^2+2A_{12}\eta_p+A_{11}}{A_{22}\eta_p+A_{12}}\\ &= -\frac{A_{11}+A_{12}\eta_p}{A_{12}+A_{22}\eta_p}.\end{aligned} \tag{2.17}$$

Note that this is exactly the same mean return obtained earlier for the portfolio which has zero covariance with $\mathbf{w}_p'\mathbf{R}$. This shows that we can find the standard deviation and mean of this uncorrelated portfolio by constructing the tangent line at the point (σ_p,η_p) and then setting η_q to be the y-coordinate of the point where this tangent line strikes the $\sigma=0$ axis as in Figure 2.3.

[FIGURE 2.3 ABOUT HERE]

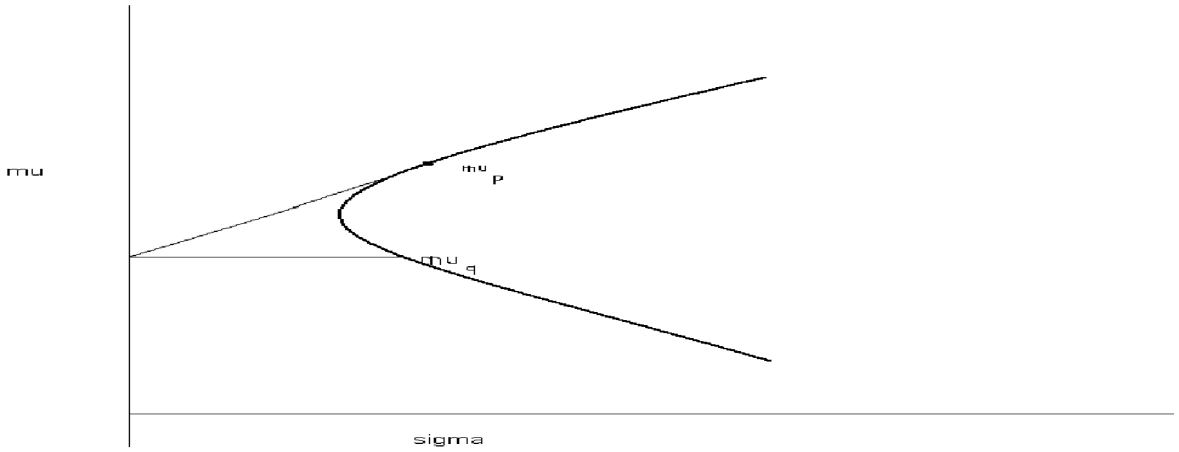


Figure 2.3: The tangent line at the point (σ_p, η_p)

Now suppose that there is available to all investors a risk-free investment. Such an investment typically has smaller return than those on the efficient frontier but since there is no risk associated with the investment, its standard deviation is 0. It may be a government bond or treasury bill yielding interest rate r so it corresponds to a point in Figure 2.4 at $(0, r)$. Since all investors are able to include this in their portfolio, the efficient frontier changes. In fact if an investor invests an amount β in this risk-free investment and amount $1 - \beta$ (this may be negative) in the risky portfolio with standard deviation and mean return (σ_p, η_p) then the resulting investment has mean return

$$E(\beta r + (1-\beta)\mathbf{w}_p'\mathbf{R}) = \beta r + (\mathbf{1} - \beta)\eta_p$$

and standard deviation of return

$$\sqrt{Var(\beta r + (1-\beta)\mathbf{w}_p'\mathbf{R})} = (1-\beta)\sigma_p.$$

This means that every point on a line joining $(0, r)$ to points in the risky portfolio are now attainable and so the new set of attainable values of (σ, η) consists of a cone with vertex at $(0, r)$,the region shaded in Figure 2.4. The efficient frontier

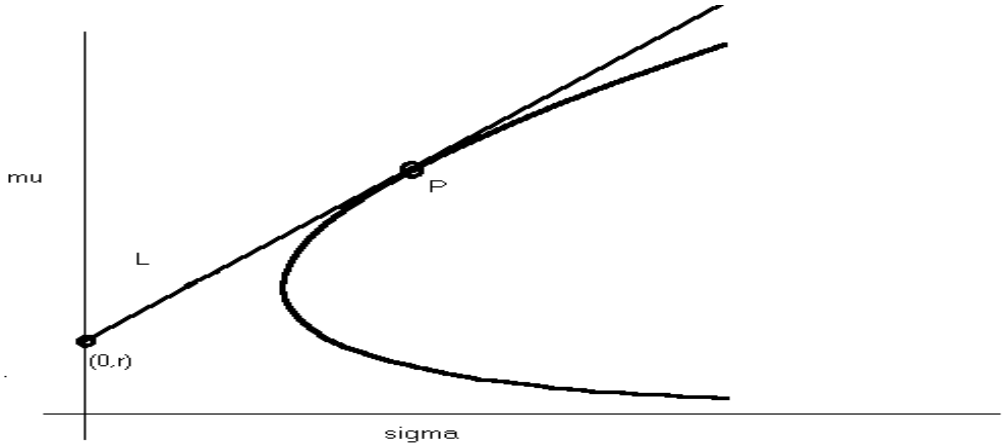


Figure 2.4: _____

is now the line L in Figure 2.4. The point m is the point at which this line is tangent to the efficient frontier determined from the risky investments. Under this theory, this point has great significance.

[FIGURE 2.4 ABOUT HERE]

Lemma 6 *The value-weighted market average corresponds to the point of tangency m of the line to the risky portfolio efficient frontier.*

From (2.17) the point m has standard deviation, mean return η_m which solves

$$\begin{aligned} r &= -\frac{A_{11} + A_{12}\eta_m}{A_{12} + A_{22}\eta_m} \\ &= \frac{\mu'\Sigma^{-1}\mu - (\mu'\Sigma^{-1}\mathbf{1})\eta_m}{\mu'\Sigma^{-1}\mathbf{1} - (\mathbf{1}'\Sigma^{-1}\mathbf{1})\eta_m} \end{aligned}$$

and this gives

$$\eta_m = \frac{\mu'\Sigma^{-1}\mu - r(\mu'\Sigma^{-1}\mathbf{1})}{\mu'\Sigma^{-1}\mathbf{1} - r(\mathbf{1}'\Sigma^{-1}\mathbf{1})}.$$

The corresponding weights on individual stocks are given by

$$\begin{aligned}
\mathbf{w}_m &= \Sigma^{-1} M A \begin{bmatrix} 1 \\ \eta_m \end{bmatrix}. \\
&= \Sigma^{-1}[\mathbf{1}\ \mu] \begin{bmatrix} A_{11} + A_{12}\eta_m \\ A_{12} + A_{22}\eta_m \end{bmatrix} \\
&= c\Sigma^{-1}[\mathbf{1}\ \mu] \begin{bmatrix} -r \\ 1 \end{bmatrix}, \quad \text{where } c = A_{12} + A_{22}\eta_m \\
&= c\Sigma^{-1}(\mu - r\mathbf{1}).
\end{aligned}$$

These market weights depend essentially on two quantities. If R denotes the correlation matrix

$$R_{ij} = \frac{\Sigma_{ij}}{\sigma_i \sigma_j}$$

where $\sigma_i = \sqrt{\Sigma_{ii}}$ is the standard deviation of the returns from stock i, and

$$\lambda_i = \frac{\mu_i - r}{\sigma_i}$$

is the standardized excess return or the price of risk, then the weight w_i on stock i is such that

$$w_i \sigma_i \propto R^{-1}\lambda \tag{2.18}$$

with λ the column vector of values of λ_i. For the purpose of comparison, recall that the conservative portfolio, one minimizing the variance over all portfolios of risky stocks, has weights

$$\mathbf{w}_g \propto \Sigma^{-1}\mathbf{1}$$

which means that the weight on stock i satisfies a relation exactly like (2.18) except that the mean returns μ_i have all been replaced by the same constant.

Let us suppose that stocks, weighed by their total capitalization in the market result in some weight vector $\mathbf{w} \neq \mathbf{w_m}$. When there is a risk-free investment, m is the only point in the risky stock portfolio that lies in the efficient frontier and so evidently if we are able to trade in a market index (a stock whose value

depends on the total market), we can find an investment which is a combination of the risk-free investment with that corresponding to m which has the same standard deviation as $\mathbf{w}'\mathbf{R}$ but higher expected return. By selling short the market index and buying this new portfolio, an arbitrage is possible. In other words, the market will not stay in this state for long.

If the market portfolio m has standard deviation σ_m and mean η_m, then the line L is described by the relation

$$\eta = r + \frac{\eta_m - r}{\sigma_m}\sigma.$$

For any investment with mean return η and standard deviation of return σ to be competitive, it must lie on this efficient frontier, i.e. it must satisfy the relation

$$\eta - r = \beta(\eta_m - r), \quad \text{where} \quad \beta = \frac{\sigma}{\sigma_m} \text{ or equivalently} \qquad (2.19)$$
$$\frac{\eta - r}{\sigma} = \frac{(\eta_m - r)}{\sigma_m}.$$

This is the most important result in the capital asset pricing model. The excess return of a stock $\eta - r$ divided by its standard deviation σ is supposed constant, and is called the *Sharpe ratio* or the *market price of risk*. The constant β called the *beta* of the stock or portfolio and represents the change in the expected portfolio return for each unit change in the market. It is also the ratio of the standard deviations of return of the stock and the market. Values of $\beta > 1$ indicate a stock that is more variable than the market and tends to have higher positive and negative returns, whereas values of $\beta < 1$ are investments that are more conservative and less volatile than the market as a whole.

We might attempt to use this model to simplify the assumed structure of the joint distribution of stock returns. One simple model in which (2.19) holds is one in which all stocks are linearly related to the market index through a simple linear regression. In particular, suppose the return from stock i, R_i, is

related to the return from the market portfolio R_m by

$$R_i - r = \beta_i(R_m - r) + \epsilon_i, \text{ where } \beta_i = \frac{\sigma_i}{\sigma_m}, \text{ and } \sigma_i^2 = \Sigma_{ii}.$$

The "errors" ϵ_i are assumed to be random variables, uncorrelated with the market returns R_m. This model is called the *single-index model* relating the returns from the stock R_i and from the market portfolio R_m.It has the merit that the relationship (2.19) follows immediately.

Taking variance on both sides, we obtain

$$var(R_i) = \beta_i^2 var(R_m) + var(\epsilon_i) = \sigma_i^2 + var(\epsilon) > \sigma_i^2$$

which contradicts the assumption that $var(R_i) = \sigma_i^2$. What is the cause of this contradiction? The relationship (2.19) assumes that the investment lies on the efficient frontier. Is this not a sufficient condition for investors to choose this investment? All that is required for rational investors to choose a particular stock is that it forms part of a portfolio which does lie on the efficient frontier.

Is every risk in an efficient market rewarded with additional expected return? We cannot expect the market to compensate us with a higher rate of return for additional risks that could be diversified away. Suppose, for example, we have two stocks with identical values of β. Suppose their returns $R_1$and R_2 both satisfy a linear regression relation above

$$R_i - r = \beta(R_m - r) + \epsilon_i, \ i = 1, 2,$$

where $cov(\epsilon_1, \epsilon_2) = 0$. Consider an investment of equal amounts in both stocks so that the return is

$$\frac{R_1 + R_2}{2} = \beta(R_m - r) + \frac{\epsilon_1 + \epsilon_2}{2}.$$

For simplicity assume that $\sigma_1 \leq \sigma_2$ and notice that the variance of this new investment is

$$\beta^2\sigma_m^2 + \frac{1}{4}[var(\epsilon_1) + var(\epsilon_2)] < var(R_2).$$

The diversified investment consisting of the average of the two results in the same mean return with smaller variance. Investors should not compensated for the additional risk in stock 2 above the level that we can achieve by sensible diversification. In general, by averaging or diversifying, we are able to provide an investment with the same average return characteristics but smaller variance than the original stock. We say that the risk (i.e. $var(\epsilon_i)$) associated with stock i which can be diversified away is the *specific risk*, and this risk is not rewarded with increased expected return. Only the so-called systematic risk σ_i which cannot by removed by diversification is rewarded with increased expected return with a relation like (2.19).

The covariance matrix of stock returns is one of the most difficult parameters to estimate in practice form historical data. If there are n stocks in a market (and normally n is large), then there are $n(n+1)/2$ elements of Σ that need to be estimated. For example if we assume all stocks in the TSE 300 index are correlated this results in a total of $(300)(301)/2 = 45,150$ parameters to estimate. We might use historical data to estimate these parameters but variances and covariances among stocks change over time and it is not clear over what period of time we can safely use to estimate these parameters. In spite of its defects, the single index model can be used to provide a simple approximate form for the covariance matrix Σ of the vector of stock returns. Notice that under the model, assuming uncorrelated random errors ϵ_i with $var(\epsilon_i) = \delta_i$,

$$R_i - r = \beta_i(R_m - r) + \epsilon_i,$$

we have

$$cov(R_i, R_j) = \beta_i\beta_j\sigma_m^2, i \neq j, \qquad var(R_i) = \beta_i^2\sigma_m^2 + \delta_i.$$

Whereas n stocks would otherwise require a total of $n(n+1)/2$ parameters in the covariance matrix Σ of returns, the single index model allows us to reduce this to the $n+1$ parameters σ_m^2, and δ_i, $i = 1, ..., n$. There is the disadvantage

in this formula however that every pair of stocks in the same market must be *positively correlated,* a feature that contradicts some observations of real market returns.

Suppose we use this form $\Sigma = \beta\beta'\sigma_m^2 + \Delta,$ to estimate weights on individual stocks, where Δ is the diagonal matrix with the δ_i along the diagonal and β is the column vector of individual stock betas. In this case $\Sigma^{-1} = \Delta^{-1} + c\Delta^{-1}\beta\beta'\Delta^{-1}$ where

$$c = \frac{-1}{\sigma_m^{-2} + \sum_i \beta_i^2/\delta_i} = -\sigma_m^2 \frac{1}{1 + \sum_i \beta_i^2\sigma_m^2/\delta_i}$$

and consequently the conservative investor by (2.14) invests in stock i proportionally to the components of $\Sigma^{-1}\mathbf{1}$

$$\text{or to } \frac{1}{\delta_i} + c\beta_i(\sum_j \beta_j/\delta_j)$$

$$\text{or proportional to } \beta_i + \frac{1}{c\delta_i(\sum_j \beta_j/\delta_j)}$$

The conditional variance of R_i given the market return R_m is δ_i. Let us call this the *excess volatility for stock i. Then the weights for the conservative portfolio are linear in the beta for the stock and the reciprocal of the excess volatility.*

The weights in the market portfolio are given by

$$\mathbf{w}_m = \Sigma^{-1}MA \begin{bmatrix} 1 \\ \eta_p \end{bmatrix} = (\Delta^{-1} + c\Delta^{-1}\beta\beta'\Delta^{-1})[\ \mathbf{1}\ \ \mu\ ](M'\Sigma^{-1}M)^{-1} \begin{bmatrix} 1 \\ \eta_p \end{bmatrix}$$

Minimum Variance under Q.

Suppose we wish to find a portfolios of securities which has the smallest possible variance under the risk neutral distribution Q. For example for a given set of weights $w_i(t)$ representing the number of shares held in security i at time t, define the portfolio $\Pi(t) = \sum w_i(t)S_i(t)$. Recall from Section 2.1 that under a risk neutral distribution, all stocks have exactly the same expected return as the risk-free interest rate so the portfolio $\Pi(t)$ will have exactly the same

conditional expected rate of return under Q as all the constituent stocks,

$$E_Q[\Pi(t+1)|H_t] = \sum_i w_i(t) E_Q[S_i(t+1)|H_t] = \sum_i w_i(t) \frac{B(t+1)}{B(t)} S_i(t) = \frac{B(t+1)}{B(t)} \Pi(t).$$

Since all portfolios have the same conditional expected return under Q, we might attempt to minimize the (conditional) variance of the portfolio return of the portfolio. The natural constraint is that the cost of the portfolio is determined by the amount $c(t)$ that we presently have to invest. We might assume a constant investment over time, for example $c(t) = 1$ for all t. Alternatively, we might wish to study a *self-financing portfolio* $\Pi(t)$, one for which past gains (or perish the thought, past losses) only are available to pay for the current portfolio so we neither withdraw from nor add money to the portfolio over its lifetime. I this case $c(t) = \Pi(t)$. We wish to minimise

$$var_Q[\Pi(t+1)|H_t] \text{ subject to the constraint } \sum_i w_i(t) S_i(t) = c(t).$$

As before, the solution is quite easy to obtain, and in fact the weights are given by the vector

$$\mathbf{w}(t) = \begin{pmatrix} w_1(t) \\ w_2(t) \\ \cdot \\ \cdot \\ \cdot \\ w_n(t) \end{pmatrix} = \frac{c(t)}{S'(t)\Sigma_t^{-1} S(t)} \Sigma_t^{-1} S(t).$$

where $\Sigma_t = var_Q(S(t+1)|H_t)$ is the instantaneous conditional covariance matrix of $S(t)$ under the measure Q. If my objective were to minimize risk under the Q measure, then this portfolio is optimal for fixed cost. The conditional variance of this portfolio is given by

$$var_Q(\Pi(t+1)|H_t) = \mathbf{w}'(t)\Sigma_t \mathbf{w}(t) = \frac{c^2(t)}{S'(t)\Sigma_t^{-1} S(t)}.$$

In terms of the portfolio return $R_\Pi(t+1) = \frac{\Pi(t+1)-\Pi(t)}{\Pi(t)}$, if the portfolio is self-financing so that $c(t) = \Pi(t)$, the above relation states that the conditional variance of the return $R_\Pi(t+1)$ given the past is simply

$$var_Q(R_\Pi(t+1)|H_t) = \frac{1}{S'(t)\Sigma_t^{-1}S(t)}$$

which is similar to the form of the variance of the conservative portfolio (2.13).

Similarly, covariances between returns for individual stocks and the return of the portfolio Π are given by exactly the same quantity, namely

$$cov(R_i(t+1), R_\Pi(t+1)|H_t) = \frac{1}{S'(t)\Sigma_t^{-1}S(t)}.$$

Let us summarize our findings so far. We assume that the conditional covariance matrix Σ_t of the vector of stock prices is non-singular. *Under the risk neutral measure, all stocks have exactly the same expected returns equal to the risk-free rate. There is a unique self-financing minimum-variance portfolio* $\Pi(t)$ *and all stocks have exactly the same conditional covariance* β *with* Π. *All stocks have exactly the same regression coefficient* β *when we regress on the minimum variance portfolio.*

Are other minimum variance portfolios conditionally uncorrelated with the portfolio we obtained above. Suppose we define $\Pi_2(t)$ similarly to minimize the variance subject to the condition that $Cov_Q(\Pi_2(t+1), \Pi(t+1)|H_t) = 0$. It is easy to see that this implies that the cost of such a portfolio at the beginning of each period is 0. This means that in this new portfolio, there is a perfect balance between long and short stocks, or that the value of the long and short stocks are equal.

The above analysis assumes that our objective is minimizing the variance of the portfolio under the risk-neutral distribution Q. Two objections could be made. First we argued earlier that the performance of an investment should be made through the *returns* , not through the stock *prices.* Since under the risk neutral measure Q, the expected return from every stock is the risk-free rate of

return, we are left with the problem of minimizing the variance of the *portfolio return.* By our earlier analysis, this is achieved when the proportion of our total investment at each time period in stock i is chosen as the corresponding component of the vector $\frac{\Sigma_t^{-1}1}{1'\Sigma_t^{-1}1}$ where now Σ_t is the conditional covariance matrix of the stock *returns.* This may appear to be a different criterion and hence a different solution, but because at each time step the stock price is a linear function of the return $S_i(t+1) = S_i(t)(1+R_i(t+1))$ the variance minimizing portfolios are essentially the same. There is another objection however to an analysis in the risk-neutral world of Q. This is a distribution which determines the value of options in order to avoid arbitrage in the system, not the actual distribution of stock prices. It is not clear what the relationship is between the covariance matrix of stock prices under the actual historical distribution and the risk neutral distribution Q, but observations seem to indicate a very considerable difference. Moreover, if this difference is large, there is very little information available for estimating the parameters of the covariance matrix under Q, since historical data on the fluctuations of stock prices will be of doubtful relevance.

Entropy: choosing a Q measure

Maximum Entropy

In 1948 in a fundamental paper on the transmission of information, C. E. Shannon proposed the following idea of *entropy.* The entropy of a distribution attempts to measure the expected number of steps required to determine a given outcome of a random variable with a given distribution when using a simple binary poll. For example suppose that a random variable X has distribution

given by

x	0	1	2
$P[X = x]$	.25	.25	.5

In this case, if we ask first whether the random variable is ≥ 2 and then, provided the answer is no, if it is ≥ 1, the expected number of queries to ascertain the value of the random variable is $1+1(1/2) = 1.5$. There is no more efficient scheme for designing this binary poll in this case so we will take 1.5 to be a measure of entropy of the distribution of X. For a discrete distribution, such that $P[X = x] = p(x)$, the entropy may be defined to be

$$H(p) = E\{-\ln(p(X))\} = -\sum_x p(x)\ln(p(x)).$$

More generally we define the entropy of an arbitrary distribution through the form for a discrete distribution. If P is a probability measure (see the appendix),

$$H(P) = \sup\{-\sum P(E_i)\ln(P(E_i))\}$$

where the supremum is taken over all finite partitions $(E_i\}$ of the space.

In the case of the above distribution, if we were to replace the natural logarithm by the log base 2, (ln and $\log_2$ differ only by a scale factor and are therefore the corresponding measures of entropy are equivalent up a constant multiple) notice that $-\sum_x p(x)\log_2(p(x)) = .5(1) + .5(2) = 1.5$, so this formula correctly measures the difficulty in ascertaining a random variable from a sequence of questions with yes-no or binary answers. This is true in general. The complexity of a distribution may be measured by the expected number of questions in a binary poll to determine the value of a random variable having that distribution, and such a measure results in the entropy $H(p)$ of the distribution.

Many statistical distributions have an interpretation in terms of maximizing entropy and it is often remarkable how well the maximum entropy principle reproduces observed distributions. For example, suppose we know that a discrete random variable takes values on a certain set of n points. What distribution p

on this set maximizes the entropy $H(p)$? First notice that if p is uniform on n points, $p(x) = 1/n$ for all x and so the entropy is $-\sum_x \frac{1}{n}\ln(\frac{1}{n}) = \ln(n)$. Now consider the problem of maximizing the entropy $H(p)$ for any distribution on n points subject to the constraint that the probabilities add to one. As in (2.10), the Lagrangian for this problem is $-\sum_x p(x)\ln(p(x)) - \lambda\{\sum_x p(x) - 1\}$ where λ is a Lagrange multiplier. Upon differentiating with respect to $p(x)$ for each x, we obtain $-\ln(p(x)) - 1 - \lambda = 0$ or $p(x) = e^{-(1+\lambda)}$. The probabilities evidently do not depend on x and the distribution is thus uniform. Applying the constraint that the sum of the probabilities is one results in $p(x) = 1/n$ for all x. *The discrete distribution on n points which has maximum entropy is the uniform distribution.* What if we repeat this analysis using additional constraints, for example on the moments of the distribution? Suppose for example that we require that the mean of the distribution is some fixed constant μ and the variance fixed at σ^2. The problem is similar to that treated above but with two more terms in the Lagrangian for each of the additional constraints. The Lagrangian becomes

$$-\sum_x p(x)\ln(p(x)) - \lambda_1\{\sum_x p(x) - 1\} - \lambda_2\{\sum_x xp(x) - \mu\} - \lambda_3\{\sum x^2 p(x) - \mu^2 - \sigma^2\}$$

whereupon setting the derivative with respect to $p(x)$ equal to zero and applying the constraints we obtain

$$p(x) = \exp\{-\lambda_1 - \lambda_2 x - \lambda_3 x^2\},$$

with constants $\lambda_1, \lambda_2, \lambda_3$ chosen to satisfy the three constraints. Since the exponent is a quadratic function of x, this is analogous to the normal distribution except that we have required that it be supported on a discrete set of points x. With more points, positioned more closely together, the distribution becomes closer to the normal. Let us call such a distribution the discrete normal distribution. For a simple example, suppose that we wish to use the maximum entropy principle to approximate the distribution of the sum of the values on

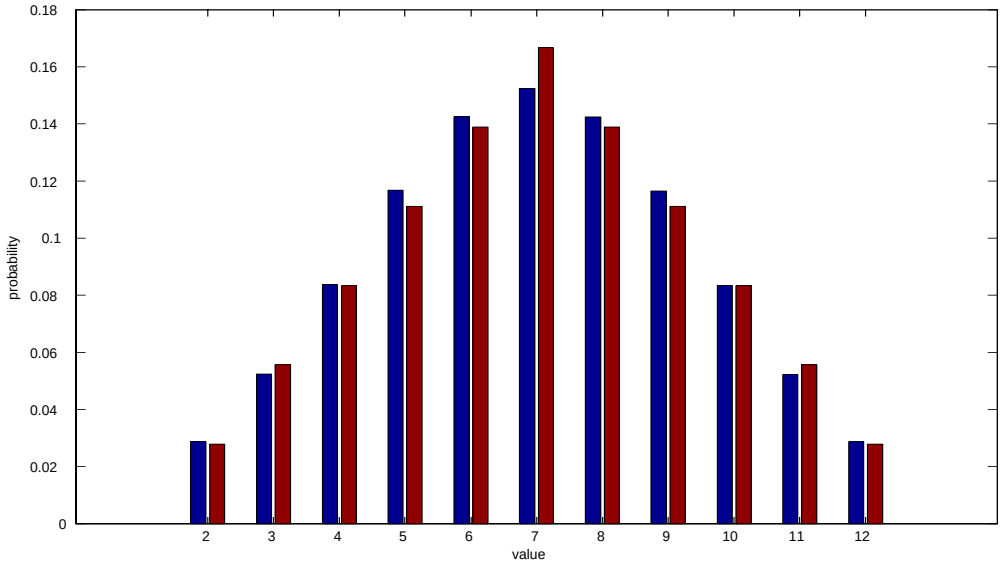


Figure 2.5: A discrete analogue of the normal distribution compared with the distribution of the sum of the values on two dice.

two dice. In this case the actual distribution is known to us as well as the mean and variance $E(X) = 7$, $var(X) = 35/6$;

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The maximum entropy distribution on these same points constrained to have the same mean and variance is very similar to this, the actual distribution. This can been seen in Figure 2.5.

[FIGURE 2.5 ABOUT HERE]

In fact if we drop the requirement that the distribution is discrete, or equivalently take a limit with an increasing number of discrete points closer and closer together, the same kind of argument shows that the maximum entropy distribution subject to a constraint on the mean and the variance is the normal distribution. So at least two well-known distributions arise out of maximum

entropy considerations. *The maximum entropy distribution on a discrete set of points is the uniform distribution. The maximum entropy subject to a constraint on the mean and the variance is a (discrete) normal distribution.* There are many other examples as well. In fact most common distributions in statistics have an interpretation as a maximum entropy distribution subject to some constraints.

Entropy has a number of properties that one would expect of a measure of the information content in a random variable. It is non-negative, and can in usual circumstances be infinite. We expect that the information in a function of X , say $g(X)$, is less than or equal to the information in X itself, equal if the function is one to one (which means in effect we can determine X from the value of $g(X)$). Entropy is a property of a distribution, not of a random variable. Nevertheless it is useful to be able to abuse the notation used earlier by referring to $H(X)$ as the entropy of the distribution of X. Then we have the following properties

Proposition 7 $H(X) \geq 0$

Proposition 8 $H(g(X)) \leq H(X)$ *for any function* $g(x)$..

The information or uncertainty in two random variables is clearly greater than that in one. The definition of entropy is defined in the same fashion as before, for discrete random variables (X, Y),

$$H(X, Y) = -E(\ln p(X, Y))$$

where $p(x, y)$ is the joint probability function

$$p(x, y) = P[X = x, Y = y].$$

If the two random variables are independent, then we expect that the uncertainty should add. If they are dependent, then the entropy of the pair (X, Y) is less than the sum of the individual entropies.

Proposition 9 *$H(X,Y) \leq H(X)+H(Y)$ with equality if and only if X and Y are independent.*

Let us now use the principle of maximum entropy to address an eminently practical problem, one of altering a distribution to accommodate a known mean value. Suppose we are interested in determining a risk-neutral distribution for pricing options at maturity T. Theorem 1 tells us that if there is to be no arbitrage, our distribution or measure Q must satisfy a relation of the form

$$E_Q(e^{-rT}S_T) = S_0$$

where r is the continuously compounded interest rate, S_0 is the initial (present) value of the underlying stock, and S_T is its value at maturity. Let us also suppose that we constraint the variance of the future stock price under the measure Q so that

$$var_Q(S_T) = \sigma^2 T.$$

Then from our earlier discussion, the maximum entropy distribution under constraints on the mean and variance is the normal distribution so that the probability density function of S_T is

$$f(s) = \frac{1}{\sigma\sqrt{2\pi T}} \exp\{-\frac{(s - e^{rT}S_0)^2}{2\sigma^2 T}\}.$$

If we wished a maximum entropy distribution which is compatible with a number of option prices, then we should impose these option prices as additional constraints. Again suppose the current time $t = 0$ and we know the prices $P_i, i = 1, ..., n$ of n different call options available on the market, all on the same security and with the same maturity T but with different strike prices K_i. The distribution Q we assign to S_T must satisfy the constraints

$$E(e^{-rT}(S_T - K_i)^+) = P_i, i = 1, ..., n \tag{2.20}$$

as well as the martingale constraint

$$E(e^{-rT}S_T) = S_0. \tag{2.21}$$

Once again introducing Lagrange multipliers, the probability density function of S_T will take the form

$$f(s) = k \exp\{e^{-rT} \sum_{i=1}^{n} \lambda_i (s - K_i)^+ + \lambda_0 s\}$$

where the parameters $\lambda_0, ..., \lambda_n$ are chosen to satisfy the constraints (2.20) and (2.21) and k so that the function integrates to 1. When fit to real option price data, these distributions typically resemble a normal density, usually however with some negative skewness and excess kurtosis. See for example Figure XXX. There are also"sawtooth" like appendages with teeth corresponding to each of the n options. Note too this density is strictly positive at the value $s = 0$, a feature that we may or may not wish to have. Because of the "teeth", a smoother version of the density is often used, one which may not perfectly reproduce option prices but is nevertheless appears to be more natural.

Minimum Cross-Entropy

Normally market information does not completely determine the risk-neutral measure Q . We will argue that while market data on derivative prices rather than historical data should determine the Q measure, historical asset prices can be used to fill in the information that is not dictated by no-arbitrage considerations. In order to relate the real world to the risk-free world, we need either sufficient market data to completely describe a risk-neutral measure Q (such a model is called a *complete market)* or we need to limit our candidate class of Q measures somewhat. We may either define the joint distributions of the stock prices or their returns, since from one we can pass to the other. For convenience, suppose we describe the joint distribution of the returns process. The conditions we impose on the martingale measure are the following;

1. Under Q, each normalized stock price $S_j(t)/B_t$ and derivative price V_t/B_t forms a martingale. Equivalently, $E_Q[S_i(t+1)|H_t] = S_i(t)(1+r(t))$

where $r(t)$ is the risk free interest rate over the interval $(t, t+1)$. (Recall that this risk-free interest rate $r(t)$ is defined by the equation $B(t+1) = (1+r(t))B(t)$.)

2. Q is a probability measure.

A slight revision of notation is necessary here. We will build our joint distributions conditionally on the past and if P denotes the joint distribution stock prices $S(1), S(2), ...S(T)$ over the whole period of observation $0 < t < T$ then P_{t+1} denotes the conditional distribution of $S(t+1)$ given H_t. Let us denote the conditional moment generating function of the vector $S(t+1)$ under the measure P_{t+1} by

$$m_t(u) = E_P[\exp(u'S(t+1)|H_t] = E_P[\exp(\sum_i u_i S_i(t+1))|H_t]$$

We implicitly assume, of course, that this moment generating function exists. Suppose, for some vector of parameters η we choose Q_{t+1} to be the exponential tilt of P_{t+1}, i.e.

$$dQ_{t+1}(s) = \frac{\exp(\eta' s)}{m_t(\eta)} dP_{t+1}(s)$$

The division by $m_t(\eta)$ is necessary to ensure that Q_{t+1} is a probability measure.

Why transform a density by multiplying by an exponential in this way? There are many reasons for such a transformation. Exponential families of distributions are built in exactly this fashion and enjoy properties of sufficiency, completeness and ease of estimation. This exponential tilt resulted from maximizing entropy subject to certain constraints on the distribution. But we also argue that the measure Q is the probability measure which is closest to P in a certain sense while still satisfying the required moment constraint. We first introduce cross-entropy which underlies considerable theory in Statistics and elsewhere in Science.

Cross Entropy

Consider two probability measures P and Q on the same space. Then the cross entropy or Kullbach-Leibler "distance" between the two measures is given by

$$H(Q,P) = \sup_{\{E_i\}} \sum Q(E_i) \log \frac{Q(E_i)}{P(E_i)}$$

where the supremum is over all finite partitions $\{E_i\}$ of the probability space. Various properties are immediate.

Proposition 10 *$H(Q,P) \geq 0$ with equality if and only if P and Q are identical.*

If Q is absolutely continuous with respect to P, that is if there is some density function $f(x)$ such that

$$Q(E) = \int_E f(x)dP \text{ for all } E$$

then provided that f is smooth, we can also write

$$H(Q,P) = E_Q \log(\frac{dQ}{dP}).$$

If Q is not absolutely continuous with respect to P then the cross entropy $H(Q,P)$ is infinite. We should also remark that the cross entropy is not really a distance in the usual sense (although we used the term "distance" in reference to it) because in general $H(Q,P) \neq H(P|Q)$. For a finite probability space, there is an easy relationship between entropy and cross entropy given by the following proposition. In effect the result tells us that maximizing entropy $H(Q)$ is equivalent to minimizing the cross-entropy $H(Q,P)$ where P is the uniform distribution.

Proposition 11 *If the probability space has a finite number n points, and P denotes the uniform distribution on these n points, then for any other probability measure Q,*

$$H(Q,P) = n - H(Q)$$

Now the following result asserts that the probability measure Q which is closest to P in the sense of cross-entropy but satisfies a constraint on its mean is generated by a so-called "exponential tilt" of the distribution of P.

Theorem 12 *: Minimizing cross-entropy.*

Let $f(X)$ be a vector valued function $f(X) = (f_1(X), f_2(X), ..., f_n(X))$ and $\mu = (\mu_1, ..., \mu_n)$. *Consider the problem*

$$\min_Q H(Q, P)$$

subject to the constraint $E_Q(f_i(X)) = \mu_i, i = 1, ..., n.$ *Then the solution, if it exists, is given by*

$$dQ = \frac{\exp(\eta' f(X))}{m(\eta)} dP = \frac{\exp(\sum_{i=1}^n \eta_i f_i(X))}{m(\eta)}$$

where $m(\eta) = E_P[\exp(\sum_{i=1}^n \eta_i f_i(X))]$ *and* η *is chosen so that* $\frac{\partial m}{\partial \eta_i} = \mu m(\eta)$.

The proof of this result, in the case of a discrete distribution P is a straightforward use of Lagrange multipliers (see Lemma 3). We leave it as a problem at the end of the chapter.

Now let us return to the constraints on the vector of stock prices. In order that the discounted stock price forms a martingale under the Q measure, we require that $E_Q[S(t+1)|H_t] = (1 + r(t))S(t)$. This is achieved if we define Q such that for any event $A \in H_t$,

$$Q(A) = \int_A Z_t dP \quad \text{where}$$

$$Z_s = k_t \exp(\sum_{t=1}^{s} \eta_t'(S_{t+1} - S_t)) \tag{2.22}$$

where k_t are H_t measurable random variables chosen so that Z_t forms a martingale

$$E(Z_{t+1}|H_t) = Z_t.$$

Theorem 9 shows that this exponentially tilted distribution has the property of being the closest to the original measure P while satisfying the condition that the normalized sequence of stock prices forms a martingale.

There is a considerable literature exploring the links between entropy and risk-neutral valuation of derivatives. See for example Gerber and Shiu (1994), Avellaneda et. al (1997), Gulko(1998), Samperi (1998). In a complete or incomplete market, risk-neutral valuation may be carried out using a martingale measure which maximizes entropy or minimizes cross-entropy subject to some natural constraints including the martingale constraint. For example it is easy to show that when interest rates r are constant, Q is the risk-neutral measure for pricing derivatives on a stock with stock price process $S_t, t = 0, 1, ...$ if and only if it is the probability measure minimizing $H(Q, P)$ subject to the martingale constraint

$$S_t = E_Q[\frac{1}{1+r} S_{t+1}]. \tag{2.23}$$

There is a continuous time analogue of (2.22) as well which we can anticipate by inspecting the form of the solution. Suppose that S_t denotes the stock price at time t where we now allow t to vary continuously in time. which we will discuss later but (2.22) can be used to anticipate it. Then an analogue of (2.22) could be written formally as

$$Z_s = \exp(\int_0^t \eta_t' dS_t - g_t)$$

where both processes η_t and g_t are "predictable" which loosely means that they are determined in advance of observing the increment $S_t, S_{t+\Delta t}$. Then the process Z_s is the analogue of the Radon-Nikodym derivative $\frac{dQ}{dP}$ of the processes restricted to the time interval $0 \leq t \leq s$. For a more formal definition, as well as an explanation of how we should interpret the integral, see the appendix. This process Z_s is, both in discrete and continuous time, a *martingale*.

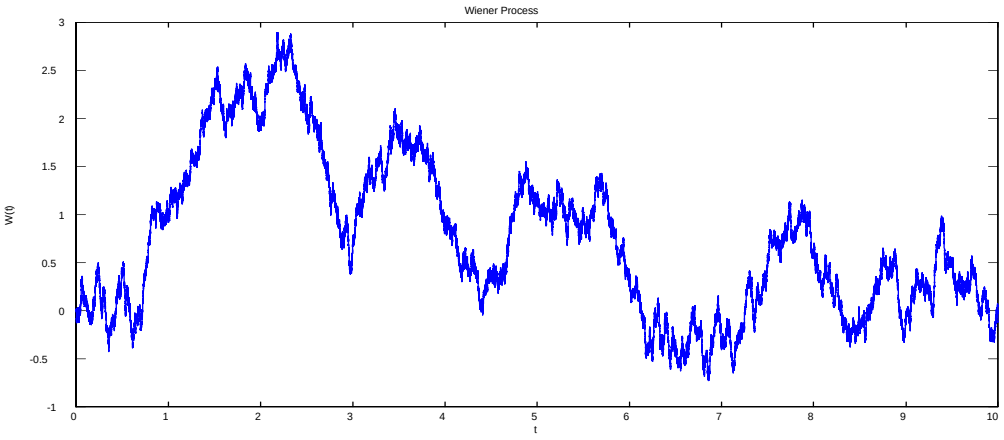


Figure 2.6: A sample path of the Wiener process

Models in Continuous Time

We begin with some oversimplified rules of stochastic calculus which can be omitted by those with a background in Brownian motion and diffusion. First, we define a stochastic process W_t called the *standard Brownian motion* or *Wiener process* having the following properties;

1. *For each* $h > 0$, *the increment* $W(t+h)-W(t)$ *has a* $N(0,h)$ *distribution and is independent of all preceding increments* $W(u)-W(v), t > u > v > 0$.

2. $W(0) = 0$.

[FIGURE 2.6 ABOUT HERE]

The fact that such a process exists is by no means easy to see. It has been an important part of the literature in Physics, Probability and Finance at least since the papers of Bachelier and Einstein, about 100 years ago. A Brownian motion process also has some interesting and remarkable theoretical properties; it is continuous with probability one but the probability that the process has finite

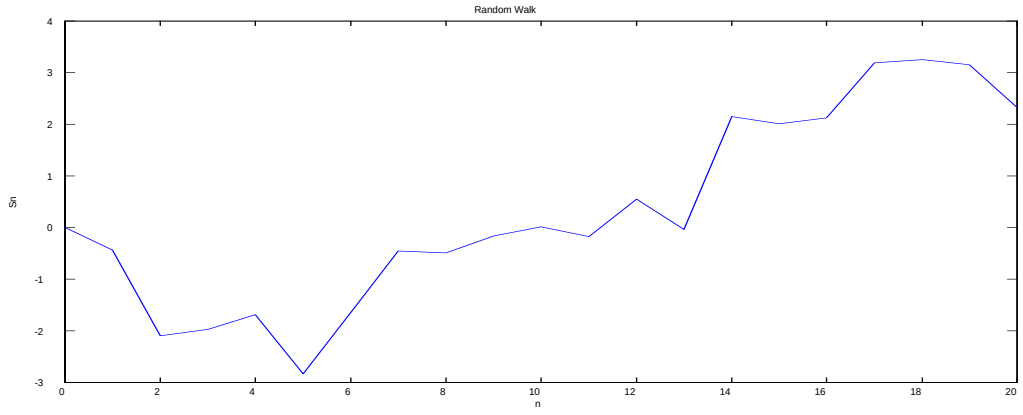


Figure 2.7: A sample path of a Random Walk

variation in any interval is 0. With probability one it is *nowhere differentiable*. Of course one might ask how a process with such apparently bizarre properties can be used to approximate real-world phenomena, where we expect functions to be built either from continuous and differentiable segments or jumps in the process. The answer is that a very wide class of functions constructed from those that are quite well-behaved (e.g. step functions) and that have independent increments converge as the scale on which they move is refined either to a Brownian motion process or to a process defined as an integral with respect to a Brownian motion process and so this is a useful approximation to a broad range of continuous time processes. For example, consider a random walk process $S_n = \sum_{i=1}^{n} X_i$ where the random variables X_i are independent identically distributed with expected value $E(X_i) = 0$ and $var(X_i) = 1$. Suppose we plot the graph of this random walk (n, S_n) as below. Notice that we have linearly interpolated the graph so that the function is defined for all n, whether integer or not.

[FIGURE 2.7 ABOUT HERE]

Now if we increase the sample size and decrease the scale appropriately on both axes, the result is, in the limit, a Brownian motion process. The vertical scale is to be decreased by a factor $1/\sqrt{n}$ and the horizontal scale by a factor n^{-1} . The theorem concludes that the sequence of processes

$$Y_n(t) = \frac{1}{\sqrt{n}} S_{nt}$$

converges weakly to a standard Brownian motion process as $n \to \infty$. In practice this means that a process with independent stationary increments tends to look like a Brownian motion process. As we shall see, there is also a wide variety of non-stationary processes that can be constructed from the Brownian motion process by integration. Let us use the above limiting result to render some of the properties of the Brownian motion more plausible, since a serious proof is beyond our scope. Consider the question of continuity, for example. Since $|Y_n(t+h) - Y_n(t)| \approx |\frac{1}{\sqrt{n}} \sum_{i=nt}^{n(t+h)} X_i|$ and this is the absolute value of an asymptotically normally$(0, h)$ random variable by the central limit theorem, it is plausible that the limit as $h \to 0$ is zero so the function is continuous at t. On the other hand note that

$$\frac{Y_n(t+h) - Y_n(t)}{h} \approx \frac{1}{h} \frac{1}{\sqrt{n}} \sum_{i=nt}^{n(t+h)} X_i$$

should by analogy behave like h^{-1} times a $N(0, h)$ random variable which blows up as $h \to 0$ so it would appear that the derivative at t does not exist. To obtain the total variation of the process in the interval $[t, t+h]$, consider the lengths of the segments in this interval, i.e.

$$\frac{1}{\sqrt{n}} \sum_{i=nt}^{n(t+h)} |X_i|$$

and notice that since the law of large numbers implies that $\frac{1}{nh} \sum_{i=nt}^{n(t+h)} |X_i|$ converges to a positive constant, namely $E|X_i|$, if we multiply by $\sqrt{n}h$ the limit must be infinite, so the total variation of the Brownian motion process is infinite.

Continuous time process are usually built one small increment at a time and defined to be the limit as the size of the time increment is reduced to zero. Let us consider for example how we might define a stochastic (*Ito*) integral of the form $\int_0^T h(t)dW_t$. An approximating sum takes the form

$$\int_0^T h(t)dW_t \approx \sum_{i=0}^{n-1} h(t_i)(W(t_{i+1}) - W(t_i)), 0 = t_0 < t_1 < ... < t_n = T.$$

Note that the function $h(t)$ is evaluated at the left hand end-point of the intervals $[t_i, t_{i+1}]$, and this is characteristic of the Ito calculus, and an important feature distinguishing it from the usual Riemann calculus studied in undergraduate mathematics courses. There are some simple reasons why evaluating the function at the left hand end-point is necessary for stochastic models in finance. For example let us suppose that the function $h(t)$ measures how many shares of a stock we possess and $W(t)$ is the price of one share of stock at time t. It is clear that we cannot predict precisely future stock prices and our decision about investment over a possibly short time interval $[t_i, t_{i+1}]$ must be made at the *beginning* of this interval, not at the end or in the middle. Second, in the case of a Brownian motion process $W(t)$, it *makes a difference* where in the interval $[t_i, t_{i+1}]$ we evaluate the function h to approximate the integral, whereas it makes no difference for Riemann integrals. As we refine the partition of the interval, the approximating sums $\sum_{i=0}^{n-1} h(t_{i+1})(W(t_{i+1}) - W(t_i))$, for example, approach a completely different limit. This difference is essentially due to the fact that $W(t)$, unlike those functions studied before in calculus, is of infinite variation. As a consequence, there are other important differences in the Ito calculus. Let us suppose that the increment dW is used to denote small increments $W(t_{i+1}) - W(t_i)$ involved in the construction of the integral. If we denote the interval of time $t_{i+1} - t_i$ by dt, we can loosely assert that dW has the normal distribution with mean 0 and variance dt. If we add up a large number of independent such increments, since the variances add, the sum has variance the sum of the values dt and standard deviation the square root. Very

roughly, we can assess the size of dW since its standard deviation is $(dt)^{1/2}$. Now consider defining a process as a function both of the Brownian motion and of time, say $V_t = g(W_t, t)$. If W_t represented the price of a stock or a bond, V_t might be the price of a derivative on this stock or bond. Expanding the increment dV using a Taylor series expansion gives

$$dV_t = \frac{\partial}{\partial W} g(W_t, t) dW + \frac{\partial^2}{\partial W^2} g(W_t, t) \frac{dW^2}{2} + \frac{\partial}{\partial t} g(W_t, t) dt \qquad (2.24)$$
$$+ (\text{stuff}) \times (dW)^3 + (\text{more stuff}) \times (dt)(dW)^2 + ...$$

Loosely, dW is normal with mean 0 and standard deviation $(dt)^{1/2}$ and so dW is non-negligible compared with dt as $dt \to 0$. We can define each of the differentials dW and dt essentially by reference to the result when we integrate both sides of the equation. If I were to write an equation in differential form

$$dX_t = h(t) dW_t$$

then this only has real meaning through its integrated version

$$X_t = X_0 + \int_0^t h(t) dW_t.$$

What about the terms involving $(dW)^2$? What meaning should we assign to a term like $\int h(t)(dW)^2$? Consider the approximating function $\sum h(t_i)(W(t_{i+1}) - W(t_i))^2$. Notice that, at least in the case that the function h is non-random we are adding up independent random variables $h(t_i)(W(t_{i+1}) - W(t_i))^2$ each with expected value $h(t_i)(t_{i+1} - t_i)$ and when we add up these quantities the limit is $\int h(t) dt$ by the law of large numbers. Roughly speaking, as differentials, we should interpret $(dW)^2$ as dt because that is the way it acts in an integral. Subsequent terms such as $(dW)^3$ or $(dt)(dW)^2$ are all $o(dt)$, i.e. they all approach 0 faster than does dt as $dt \to 0$. So finally substituting for $(dW)^2$ in 2.24 and ignoring all terms that are $o(dt)$, we obtain a simple version of *Ito's lemma*

$$dg(W_t,t) = \frac{\partial}{\partial W} g(W_t,t)dW + \{\frac{1}{2}\frac{\partial^2}{\partial W^2} g(W_t,t) + \frac{\partial}{\partial t} g(W_t,t)\}dt.$$

This rule results, for example, when we put $g(W_t,t) = W_t^2$ in

$$d(W_t^2) = 2W_t dW_t + dt$$

or on integrating both sides and rearranging,

$$\int_a^b W_t dW_t = \frac{1}{2}(W_b^2 - W_a^2) - \frac{1}{2}\int_a^b dt.$$

The term $\int_a^b dt$ above is what distinguishes the Ito calculus from the Riemann calculus, and is a consequence of the nature of the Brownian motion process, a continuous function of infinite variation.

There is one more property of the stochastic integral that makes it a valuable tool in the construction of models in finance, and that is that a stochastic integral with respect to a Brownian motion process is *always a martingale*. To see this, note that in an approximating sum

$$\int_0^T h(t)dW_t \approx \sum_{i=0}^{n-1} h(t_i)(W(t_{i+1}) - W(t_i))$$

each of the summands has conditional expectation 0 given the past, i.e.

$$E[h(t_i)(W(t_{i+1}) - W(t_i))|H_{t_i}] = h(t_i)E[(W(t_{i+1}) - W(t_i))|H_{t_i}] = 0$$

since the Brownian increments have mean 0 given the past and since $h(t)$ is measurable with respect to H_t.

We begin with an attempt to construct the model for an Ito process or diffusion process in continuous time. We construct the price process one increment at a time and it seems reasonable to expect that both the mean and the variance of the increment in price may depend on the current price but does not depend on the process before it arrived at that price. This is a loose description of a Markov property. The conditional distribution of the future of the process

depends only on the current time t and the current price of the process. Let us suppose in addition that the increments in the process are, conditional on the past, normally distributed. Thus we assume that for small values of h, conditional on the current time t and the current value of the process X_t, the increment $X_{t+h} - X_t$ can be generated from a normal distribution with mean $a(X_t, t)h$ and with variance $\sigma^2(X_t, t)h$ for some functions a and σ^2 called the drift and diffusion coefficients respectively. Such a normal random variable can be formally written as $a(X_t, t\ )dt + \sigma^2(X_t, t)dW_t$. Since we could express X_T as an initial price X_0 plus the sum of such increments, $X_T = X_0 + \sum_i (X_{t_{i+1}} - X_{t_i})$.

The single most important model of this type is called the *Geometric Brownian motion or Black-Scholes model.* Since the actual value of stock, like the value of a currency or virtually any other asset is largely artificial, depending on such things as the number of shares issued, it is reasonable to suppose that the *changes in a stock price* should be modeled relative to the current price. For example rather than model the increments, it is perhaps more reasonable to model the relative change in the process. The simplest such model of this type is one in which both the mean and the standard deviation of the increment in the price are linear multiples of price itself; viz. dX_t is approximately normally distributed with mean $aX_t dt$ and variance $\sigma^2 X_t^2 dt$. In terms of stochastic differentials, we assume that

$$dX_t = aX_t dt + \sigma X_t dW_t. \tag{2.25}$$

Now consider the relative return from such a process over the increment $dY_t = dX_t/X_t$. Putting $Y_t = g(X_t) = \ln(X_t)$ note that analogous to our derivation of Ito's lemma

$$\begin{aligned} dg(X_t) &= g'(X_t)dX_t + \frac{1}{2}g''(X_t)(dX)^2 + ... \\ &= \frac{1}{X_t}\{aX_t dt + \sigma X_t dW_t.\} - \frac{1}{2X_t^2}\sigma^2 X_t^2 dt \\ &= (a - \frac{\sigma^2}{2})dt + \sigma dW_t. \end{aligned}$$

which is a description of a general Brownian motion process, a process with increments dY_t that are normally distributed with mean $(a - \frac{\sigma^2}{2})dt$ and with variance $\sigma^2 dt$. This process satisfying $dX_t = aX_t dt + \sigma X_t dW_t$ is called the *Geometric Brownian motion* process (because it can be written in the form $X_t = e^{Y_t}$ for a Brownian motion process Y_t) or a Black-Scholes model.

Many of the continuous time models used in finance are described as Markov diffusions or Ito processes which permits the mean and the variance of the increments to depend more generally on the present value of the process and the time. The integral version of this relation is of the form

$$X_T = X_0 + \int_0^T a(X_t, t)dt + \int_0^T \sigma(X_t, t)dW_t.$$

We often write such an equation with differential notation,

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t. \tag{2.26}$$

but its meaning should always be sought in the above integral form. The coefficients $a(X_t, t)$ and $\sigma(X_t, t)$ vary with the choice of model. As usual, we interpret 2.26 as meaning that a small increment in the process, say $dX_t = X_{t+h} - X_t$ (h very small) is approximately distributed according to a normal distribution with conditional mean $a(X_t, t)dt$ and conditional variance given by $\sigma^2(X_t, t)var(dW_t) = \sigma^2(X_t, t)dt$. Here the mean and variance are conditional on H_t, the history of the process X_t up to time t.

Various choices for the functions $a(X_t, t), \sigma(X_t, t)$ are possible. For the Black-Scholes model or geometric Brownian motion, $a(X_t, t) = aX_t$ and $\sigma(X_t, t) = \sigma X_t$ for constant drift and volatility parameters a, σ. The *Cox-Ingersoll-Ross model,* used to model spot interest rates, corresponds to $a(X_t, t) = A(b - X_t)$ and $\sigma(X_t, t) = c\sqrt{X_t}$ for constants A, b, c. The Vasicek model, also a model for interest rates, has $a(X_t, t) = A(b - X_t)$ and $\sigma(X_t, t) = c$. There is a large number of models for most continuous time processes observed in finance which can be written in the form 2.26. So called multi-factor models are of similar form

where X_t is a vector of financial time series and the coefficient functions $a(X_t, t)$ is vector valued, $\sigma(X_t, t)$ is replaced by a matrix-valued function and dW_t is interpreted as a vector of independent Brownian motion processes. For technical conditions on the coefficients under which a solution to 2.26 is guaranteed to exist and be unique, see Karatzas and Shreve, sections 5.2, 5.3.

As with any differential equation there may be initial or boundary conditions applied to 2.26 that restrict the choice of possible solutions. Solutions to the above equation are difficult to arrive at, and it is often even more difficult to obtain distributional properties of them. Among the key tools are the *Kolmogorov differential equations* (see Cox and Miller, p. 215). Consider the transition probability kernel

$$p(s, z, t, x) = P[X_t = x | X_s = z]$$

in the case of a *discrete Markov Chain.* If the Markov chain is continuous (as it is in the case of diffusions), that is if the conditional distribution of X_t given X_s is absolutely continuous with respect to Lebesgue measure, then we can define $p(s, z, t, x)$ to be the *conditional probability density function* of X_t given $X_s = z$. The two equations, for a diffusion of the above form, are:

Kolmogorov's backward equation

$$\frac{\partial}{\partial s}p = -a(z, s)\frac{\partial}{\partial z}p - \frac{1}{2}\sigma^2(z, s)\frac{\partial^2}{\partial z^2}p \tag{2.27}$$

and the *forward equation*

$$\frac{\partial}{\partial t}p = -\frac{\partial}{\partial x}(a(x, t)p) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x, t)p) \tag{2.28}$$

Note that if we were able to solve these equations, this would provide the transition density function p, giving the conditional distribution of the process. It does not immediately provide other characteristics of the diffusion, such as the distribution of the maximum or the minimum, important for valuing various exotic options such as look-back and barrier options. However for a European

option defined on this process, knowledge of the transition density would suffice at least theoretically for valuing the option. Unfortunately these equations are often very difficult to solve explicitly.

Besides the Kolmogorov equations, we can use simple ordinary differential equations to arrive at some of the basic properties of a diffusion. To illustrate, consider one of the simplest possible forms of a diffusion, where $a(X_t, t) = \alpha(t)+\beta(t)X_t$ where the coefficients $\alpha(t), \beta(t)$ are deterministic (i.e. non-random) functions of time. Note that the integral analogue of 2.26 is

$$X_t = X_0 + \int_0^t a(X_s, s)ds + \int_0^t \sigma(X_s, s)dW_s \tag{2.29}$$

and by construction that last term $\int_0^t \sigma(X_s, s)dW_s$ is a zero-mean martingale. For example its small increments $\sigma(X_t, t)dW_s$ are approximately $N(0, \sigma(X_t, t)dt)$. Therefore, taking expectations on both sides conditional on the value of X_0, and letting $m(t) = E(X_t)$, we obtain:

$$m(t) = X_0 + \int_0^t [\alpha(s) + \beta(s)m(s)]ds \tag{2.30}$$

and therefore $m(t)$solves the ordinary differential equation

$$m'(t) = \alpha(t) + \beta(t)m(t). \tag{2.31}$$

$$m(0) = X_0 \tag{2.32}$$

Thus, in the case that the *drift term* a is a linear function of X_t, the mean or expected value of a diffusion process can be found by solving a similar ordinary differential equation, similar except that the diffusion term has been dropped.

These are only two of many reasons to wish to solve both ordinary and partial differential equations in finance. The solution to the Kolmogorov partial differential equations provides the conditional distribution of the increments of a process. And when the drift term $a(X_t, t\ )$ is linear in X_t, the solution of an ordinary differential equation will allow the calculation of the expected value of the process and this is the first and most basic description of its behaviour. The

appendix provides an elementary review of techniques for solving partial and ordinary differential equations.

However, that the information about a stochastic process obtained from a deterministic object such as a ordinary or partial differential equation is necessarily limited. For example, while we can sometimes obtain the marginal distribution of the process at time t it is more difficult to obtain quantities such as the joint distribution of variables which depending on the path of the process, and these are important in valuing certain types of exotic options such as lookback and barrier options. For such problems, we often use Monte Carlo methods.

The Black-Scholes Formula

Before discussing methods of solution in general, we develop the Black-Scholes equation in a general context. Suppose that a security price is an Ito process satisfying the equation

$$dS_t \ = a(S_t,\ t\ )\ dt \ +\ \sigma(S_t,\ t)\ dW_t \tag{2.33}$$

Assumed the market allows investment in the stock as well as a risk-free bond whose price at time t is B_t. It is necessary to make various other assumptions as well and strictly speaking all fail in the real world, but they are a reasonable approximation to a real, highly liquid and nearly frictionless market:

1. partial shares may be purchased
2. there are no dividends paid on the stock
3. There are no commissions paid on purchase or sale of the stock or bond
4. There is no possibility of default for the bond
5. Investors can borrow at the risk free rate governing the bond.
6. All investments are liquid- they can be bought or sold instantaneously.

Since bonds are assumed risk-free, they satisfy an equation

$$dB_t = r_t B_t dt$$

where r_t is the risk-free (spot) interest rate at time t.

We wish to determine $V(S_t, t)$, the value of an option on this security when the security price is S_t, at time t. Suppose the option has expiry date T and a general payoff function which depends only on S_T, the process at time T.

Ito's lemma provides the ability to translate an a relation governing the differential dS_t into a relation governing the differential of the process $dV(S_t, t)$. In this sense it is the stochastic calculus analogue of the chain rule in ordinary calculus. It is one of the most important single results of the twentieth century in finance and in science. The stochastic calculus and this mathematical result concerning it underlies the research leading to 1997 Nobel Prize to Merton and Scholes for their work on hedging in financial models. We saw one version of it at the beginning of this section and here we provide a more general version.

Ito's lemma.

Suppose S_t is a diffusion process satisfying

$$dS_t = a(S_t, t)dt + \sigma(S_t, t)dW_t$$

and suppose $V(S_t, t)$ is a smooth function of both arguments. Then $V(S_t, t)$ also satisfies a diffusion equation of the form

$$dV = [a(S_t, t)\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t, t)}{2}\ \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]dt + \sigma(S_t, t)\frac{\partial V}{\partial S}dW_t. \qquad (2.34)$$

Proof. The proof of this result is technical but the ideas behind it are simple. Suppose we expand an increment of the process $V(S_t, t)$ (we write V

in place of $V(S_t, t)$ omitting the arguments of the function and its derivatives. We will sometimes do the same with the coefficients a and σ.)

$$V(S_{t+h}, t+h) \approx V + \frac{\partial V}{\partial S}(S_{t+h} - S_t) + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}(S_{t+h} - S_t)^2 + \frac{\partial V}{\partial t}h \quad (2.35)$$

where we have ignored remainder terms that are $o(h)$. Note that substituting from 2.33 into 2.35, the increment $(S_{t+h} - S_t)$ is approximately normal with mean $a(S_t,\ t\ )\ h$ and variance $\sigma^2(S_t,\ t\ )\ h$. Consider the term $(S_{t+h} - S_t)^2$. Note that it is the square of the above normal random variable and has expected value $\sigma^2(S_t, t)h + a^2(S_t, t)h^2$. The variance of this random variable is $O(h^2)$ so if we ignore all terms of order $o(h)$ the increment $V(S_{t+h}, t+h) - V(S_t, t)$ is approximately normally distributed with mean

$$[a(S_t,\ t\ )\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t,\ t)}{2}\ \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]h$$

and standard deviation $\sigma(S_t,\ t)\frac{\partial V}{\partial S}\sqrt{h}$ justifying (but not proving!) the relation 2.34. ■

By Ito's lemma, provided $V$is smooth, it also satisfies a diffusion equation of the form 2.34. We should note that when V represents the price of an option, some lack of smoothness in the function V is inevitable. For example for a European call option with exercise price K, $V(S_T, T) = \max(S_T - K, 0)$ does not have a derivative with respect to S_T at $S_T = K$, the exercise price. Fortunately, such exceptional points can be worked around in the argument, since the derivative does exist at values of $t < T$.

The basic question in building a replicating portfolio is: for hedging purposes, is it possible to find a *self-financing* portfolio consisting only of the security and the bond which exactly replicates the option price process $V(S_t, t)$? The self-financing requirement is the analogue of the requirement that the net cost of a portfolio is zero that we employed when we introduced the notion of

arbitrage. The portfolio is such that no funds are needed to be added to (or removed from) the portfolio during its life, so for example any additional amounts required to purchase equity is obtained by borrowing at the risk free rate. Suppose the self-financing portfolio has value at time t equal to $V_t = u_t S_t + w_t B_t$ where the (predictable) functions u_t, w_t represent the number of shares of stock and bonds respectively owned at time t. Since the portfolio is assumed to be self-financing, all returns obtain from the changes in the value of the securities and bonds held, i.e. it is assumed that $dV_t = u_t dS_t + w_t dB_t$. Substituting from 2.33,

$$dV_t = u_t dS_t + w_t dB_t = [u_t a(S_t, t) + w_t r_t B_t]dt + u_t \sigma(S_t, t) dW_t \tag{2.36}$$

If V_t is to be exactly equal to the price $V(S_t, t\ )$ of an option, it follows on comparing the coefficients of dt and dW_t in 2.34 and 2.36, that $u_t = \frac{\partial V}{\partial S}$, called the *delta* corresponding to *delta hedging*. Consequently,

$$V_t = \frac{\partial V}{\partial S} S_t + w_t B_t$$

and solving for w_t we obtain:

$$w_t = \frac{1}{B_t}[V - \frac{\partial V}{\partial S} S_t].$$

The conclusion is that it is possible to dynamically choose a trading strategy, i.e. the weights w_t, u_t so that our portfolio of stocks and bonds perfectly replicates the value of the option. If we own the option, then by shorting (selling) delta= $\frac{\partial V}{\partial S}$ units of stock, we are **perfectly** hedged in the sense that our portfolio replicates a risk-free bond. Surprisingly, in this ideal word of continuous processes and continuous time trading commission-free trading, the perfect hedge is possible. In the real world, it is said to exist only in a Japanese garden. The equation we obtained by equating both coefficients in 2.34 and 2.36 is;

$$-r_t V + r_t S_t \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} + \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{2.37}$$

Rewriting this allows an interpretation in terms of our hedged portfolio. If we own an option and are short delta units of stock our net investment at time t is given by $(V - S_t \frac{\partial V}{\partial S})$ where $V = V_t = V(S_t, t)$. Our return over the next time increment dt if the portfolio were liquidated and the identical amount invested in a risk-free bond would be $r_t(V_t - S_t \frac{\partial V}{\partial S})dt$. On the other hand if we keep this hedged portfolio, the return over an increment of time dt is

$$\begin{aligned} d(V - S_t \frac{\partial V}{\partial S}) &= dV - (\frac{\partial V}{\partial S})dS \\ &= (\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 V}{\partial S^2} + a\frac{\partial V}{\partial S})dt + \sigma \frac{\partial V}{\partial S} dW_t \\ &- \frac{\partial V}{\partial S}[a dt + \sigma dW_t] \\ &= (\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 V}{\partial S^2})dt \end{aligned}$$

Therefore

$$r_t(V - S_t \frac{\partial V}{\partial S}) = \frac{\partial V}{\partial t} + \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 V}{\partial S^2}.$$

The left side $r_t(V - S_t \frac{\partial V}{\partial S})$ represents the amount made by the portion of our portfolio devoted to risk-free bonds. The right hand side represents the return on a hedged portfolio long one option and short delta stocks. Since these investments are at least in theory identical, so is their return. This fundamental equation is evidently satisfied by any option price process where the underlying security satisfies a diffusion equation and the option value at expiry depends only on the value of the security at that time. The type of option determines the terminal conditions and usually uniquely determines the solution.

It is extraordinary that this equation in no way depends on the drift coefficient $a(S_t, t)$. This is a remarkable feature of the arbitrage pricing theory. *Essentially, no matter what the drift term for the particular security is, in order to avoid arbitrage, all securities and their derivatives are priced as if they had as drift the spot interest rate. This is the effect of calculating the expected values under the martingale measure Q.*

This PDE governs most derivative products, European call options, puts,

futures or forwards. However, the boundary conditions and hence the solution depends on the particular derivative. The solution to such an equation is possible analytically in a few cases, while in many others, numerical techniques are necessary. One special case of this equation deserves particular attention. In the case of geometric Brownian motion, $a(S_t, t) = \mu S_t$ and $\sigma(S_t, t) = \sigma S_t$ for constants μ, σ. Assume that the spot interest rate is a constant $r$and that a constant rate of dividends D_0 is paid on the stock. In this case, the equation specializes to

$$-rV + \frac{\partial V}{\partial t} + (r - D_0)S\frac{\partial V}{\partial S} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 V}{\partial S^2} = 0. \tag{2.38}$$

Note that we have not used *any* of the properties of the particular derivative product yet, nor does this differential equation involve the drift coefficient μ. The assumption that there are no transaction costs is essential to this analysis, as we have assumed that the portfolio is continually rebalanced.

We have now seen two derivations of parabolic partial differential equations, so-called because like the equation of a parabola, they are first order (derivatives) in one variable (t) and second order in the other (x). Usually the solution of such an equation requires reducing it to one of the most common partial differential equations, the heat or diffusion equation, which models the diffusion of heat along a rod. This equation takes the form

$$\frac{\partial}{\partial t}u = k\frac{\partial^2}{\partial x^2}u \tag{2.39}$$

A solution of 2.39 with appropriate boundary conditions can sometime be found by the separation of variables. We will later discuss in more detail the solution of parabolic equations, both by analytic and numerical means. First, however, when can we hope to find a solution of 2.39 of the form $u(x,t) = g(x/\sqrt{t})$. By differentiating and substituting above, we obtain an ordinary differential equation of the form

$$g''(\omega) + \frac{1}{2k}\omega g'(\omega) = 0, \omega = x/\sqrt{t} \tag{2.40}$$

Let us solve this using MAPLE.

```
eqn := diff(g(w),w,w)+(w/(2*k))*diff(g(w),w)=0;
dsolve(eqn,g(w));
```

and because the derivative of the solution is slightly easier (for a statistician) to identify than the solution itself,

```
> diff(%,w);
```

giving

$$\frac{\partial}{\partial w}g(\omega) = C_2 \exp\{-w^2/4k\} = C_2 \exp\{-x^2/4kt\} \tag{2.41}$$

showing that a constant plus a constant multiple of the Normal $(0, 2kt)$ cumulative distribution function or

$$u(x,t) = C_1 + C_2 \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^{x} \exp\{-z^2/4kt\} dz \tag{2.42}$$

is a solution of this, the heat equation for $t > 0$. The role of the two constants is simple. Clearly if a solution to 2.39 is found, then we may add a constant and/or multiply by a constant to obtain another solution. The constant in general is determined by initial and boundary conditions. Similarly the integral can be removed with a change in the initial condition for if u solves 2.39 then so does $\frac{\partial u}{\partial x}$. For example if we wish a solution for the half real $x > 0$ with initial condition $u(x,0) = 0, u(0,t) = 1$ all $t > 1$, we may use

$$u(x,t) = 2P(N(0,2kt) > x) = \frac{1}{\sqrt{\pi kt}} \int_{x}^{\infty} \exp\{-z^2/4kt\} dz, t > 0, x \geq 0.$$

Let us consider a basic solution to 2.39:

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-x^2/4kt\} \tag{2.43}$$

This connection between the heat equation and the normal distributions is fundamental and the wealth of solutions depending on the initial and boundary conditions is considerable. We plot a fundamental solution of the equation as follows with the plot in Figure 2.8:

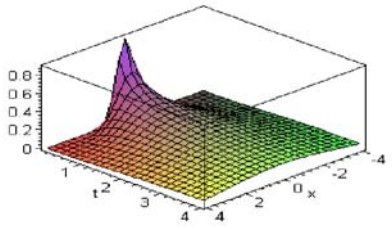


Figure 2.8: Fundamental solution of the heat equation

```
>u(x,t) := (.5/sqrt(Pi*t))*exp(-x^2/(4*t));

>plot3d(u(x,t),x=-4..4,t=.02..4,axes=boxed);
```

[FIGURE 2.8 ABOUT HERE]

As $t \to 0$, the function approaches a spike at $x = 0$, usually referred to as the "Dirac delta function" (although it is no function at all) and symbolically representing the derivative of the "Heaviside function". The Heaviside function is defined as $H(x) = 1, x \geq 0$ and is otherwise 0 and is the cumulative distribution function of a point mass at 0. Suppose we are given an initial condition of the form $u(x, 0) = u_0(x)$. To this end, it is helpful to look at the solution $u(x, t)$ and the initial condition $u_0(x)$ as a distribution or measure (in this case described by a density) over the space variable x. For example the density $u(x, t)$ corresponds to a measure for fixed t of the form $\nu_t(A) = \int_A u(x, t)dx$. Note that the initial condition compatible with the above solution 2.42 can be described somewhat clumsily as "$u(x, 0)$ corresponds to a measure placing all mass at $x = x_0 = 0$ ".In fact as $t \to 0$, we have in some sense the following convergence $u(x, t) \to \delta(x) = dH(x)$, the Dirac delta function. We could just as easily construct solve the heat equation with a more general initial condition of

the form $u(x,0) = dH(x - x_0)$ for arbitrary x_0 and the solution takes the form

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-(x - x_0)^2/4kt\}. \tag{1.22}$$

Indeed sums of such solutions over different values of x_0, or weighted sums, or their limits, integrals will continue to be solutions to 2.39. In order to achieve the initial condition $u_0(x)$ we need only pick a suitable weight function. Note that

$$u_0(x) = \int u_0(z) dH(z - x)$$

Note that the function

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^{\infty} \exp\{-(z - x)^2/4kt\} u_0(z) dz \tag{1.22}$$

solves 2.39 subject to the required boundary condition.

Solution of the Diffusion Equation.

We now consider the general solution to the diffusion equation of the form 2.37, rewritten as

$$\frac{\partial V}{\partial t} = r_t V - r_t S_t \frac{\partial V}{\partial S} - \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} \tag{2.44}$$

where $S_t$is an asset price driven by a diffusion equation

$$dS_t = a(S_t, t)dt + \sigma(S_t, t)dW_t, \tag{2.45}$$

$V(S_t, t)$ is the price of an option on that asset at time t, and $r_t = r(t)$ is the spot interest rate at time t. We assume that the price of the option at expiry $T$is a known function of the asset price

$$V(S_T, T) = V_0(S_T). \tag{2.46}$$

Somewhat strangely, the option is priced using a related but not identical process (or, equivalently, the same process under a different measure). Recall from the

backwards Kolmogorov equation 2.27 that if a related process $X_t$satisfies the stochastic differential equation

$$dX_t = r(X_t, t)X_t dt + \sigma(X_t, t)dW_t \tag{2.47}$$

then its transition kernel $p(t, s, T, z) = \frac{\partial}{\partial z}P[X_T \leq z | X_t = s]$ satisfies a partial differential equation similar to 2.44;

$$\frac{\partial p}{\partial t} = -r(s,t)s\frac{\partial p}{\partial s} - \frac{\sigma^2(s,t)}{2}\frac{\partial^2 p}{\partial s^2} \tag{2.48}$$

For a given process $X_t$this determines one solution. For simplicity, consider the case (natural in finance applications) when the spot interest rate is a function of time, not of the asset price; $r(s,t) = r(t)$. To obtain the solution so that terminal conditions is satisfied, consider a product

$$f(t, s, T, z) = p(t, s, T, z)q(t, T) \tag{2.49}$$

where

$$q(t,T) = exp\{-\int_t^T r(v)dv\}$$

is the discount function or the price of a zero-coupon bond at time t which pays 1$ at maturity.

Let us try an application of one of the most common methods in solving PDE's, the "lucky guess" method. Consider a linear combination of terms of the form 2.49 with weight function $w(z)$. i.e. try a solution of the form

$$V(s,t) = \int p(t, s, T, z)q(t, T)w(z)dz \tag{2.50}$$

for suitable weight function $w(z)$. In view of the definition of $p$as a transition probability density, this integral can be rewritten as a conditional expectation:

$$V(t,s) = E[w(X_T)q(t,T)|X_t = s] \tag{2.51}$$

the discounted conditional expectation of the random variable $w(X_T)$ given the current state of the process, where the process is assumed to follow (2.18). Note

that in order to satisfy the terminal condition 2.46, we choose $w(x) = V_0(x)$. Now

$$\begin{aligned}\frac{\partial V}{\partial t} &= \frac{\partial}{\partial t}\int p(t,s,T,z)q(t,T)w(z)dz \\ &= \int [-r(S_t,t)S_t\frac{\partial p}{\partial s} - \frac{\sigma^2(S_t,t)}{2}\frac{\partial^2 p}{\partial s}2]q(t,T)w(z)dz \\ &+ r(S_t,t)\ \int p(t,S_t,T,z)q(t,T)w(z)dz \text{ by 2.48} \\ &= -r(S_t,t)S_t\frac{\partial V}{\partial S} - \frac{\sigma^2(S_t,t)}{2}\frac{\partial^2 V}{\partial S^2} + r(S_t,t)V(S_t,t)\end{aligned}$$

where we have assumed that we can pass the derivatives under the integral sign. Thus the process

$$V(t,s) = E[V_0(X_T)q(t,T)|X_t = s] \tag{2.52}$$

satisfies both the partial differential equation 2.44 and the terminal conditions 2.46 and is hence the solution. Indeed it is the unique solution satisfying certain regularity conditions. The result asserts that the value of any European option is simply the conditional expected value of the *discounted payoff* (discounted to the present) assuming that the distribution is that of the process 2.47. This result is a special case when the spot interest rates are functions only of time of the following more general theorem.

Theorem 13 *(Feynman-Kac)*

Suppose the conditions for a unique solution to (2.44,2.46) (see for example Duffie, appendix E) are satisfied. Then the general solution to (2.15) under the terminal condition 2.46 is given by

$$V(S,t) = E[V_0(X_T)exp\{-\int_t^T r(X_v,v)dv\}|\ X_t = S] \tag{2.53}$$

This represents the discounted return from the option under the distribution of the process X_t. The distribution induced by the process X_t is referred to as the *equivalent martingale measure* or *risk neutral measure*. Notice that when the original process is a diffusion, the equivalent martingale measure shares the same diffusion coefficient but has the drift replaced by $r(X_t, t)X_t$. The option is priced as if the drift were the same as that of a risk-free bond i.e. as if the instantaneous rate of return from the security if identical to that of bond. Of course, in practice, it is not. A risk premium must be paid to the stock-holder to compensate for the greater risk associated with the stock.

There are some cases in which the conditional expectation 2.53 can be determined explicitly. In general, these require that the process or a simple function of the process is Gaussian.

For example, suppose that both $r(t)$ and $\sigma(t)$ are deterministic functions of time only. Then we can solve the stochastic differential equation (2.22) to obtain

$$X_T = \frac{X_t}{q(t,T)} + \int_t^T \frac{\sigma(u)}{q(u,T)} dW_u \tag{2.54}$$

The first term above is the conditional expected value of X_T given X_t. The second is the random component, and since it is a weighted sum of the normally distributed increments of a Brownian motion with weights that are non-random, it is also a normal random variable. The mean is 0 and the (conditional) variance is $\int_t^T \frac{\sigma^2(u)}{q^2(u,T)} du$. Thus the conditional distribution of X_T given X_t is normal with conditional expectation $\frac{X_t}{q(t,T)}$ and conditional variance $\int_t^T \frac{\sigma^2(u)}{q^2(u,T)} du$.

The special case of 2.53 of most common usage is the Black-Scholes model: suppose that $\sigma(S,t) = S\sigma(t)$ for $\sigma(t)$ some deterministic function of t. Then the distribution of X_t is not Gaussian, but fortunately, its logarithm is. In this case we say that the distribution of X_t is lognormal.

Lognormal Distribution

Suppose Z is a normal random variable with mean μ and variance σ^2. Then we say that the distribution of $X = e^Z$ is lognormal with mean $\eta = \exp\{\mu + \sigma^2/2\}$ and volatility parameter σ. The lognormal probability density function with mean $\eta > 0$ and volatility parameter $\sigma > 0$ is given by the probability density function

$$g(x|\eta,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(\log x - \log\eta - \sigma^2/2)^2/2\sigma^2\}. \tag{2.55}$$

The solution to (2.18) with non-random functions $\sigma(t), r(t)$ is now

$$X_T = X_t exp\{\int_t^T (r(u) - \sigma^2(u)/2)du + \int_t^T \sigma(u)dW_u\}. \tag{2.56}$$

Since the exponent is normal, the distribution of X_T is lognormal with mean $log(X_t) + \int_t^T (r(u) - \sigma^2(u)/2)du$ and variance $\int_t^T \sigma^2(u)du$. It follows that the conditional distribution is lognormal with mean $\eta = X_t q(t,T)$ and volatility parameter $\sqrt{\int_t^T \sigma^2(u)du}$.

We now derive the well-known Black-Scholes formula as a special case of 2.53. For a call option with exercise price E, the payoff function is $V_0(S_T) = max(S_T - E, 0)$. Now it is helpful to use the fact that for a standard normal random variable Z and arbitrary $\sigma > 0, -\infty < \mu < \infty$ we have the expected value of $max(e^{\sigma Z+\mu}, 0)$ is

$$e^{\mu+\sigma^2/2}\Phi(\frac{\mu}{\sigma} + \sigma) - \Phi(\frac{\mu}{\sigma}) \tag{2.57}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. As a result, in the special case that r and σ are constants, (2.53) results in the famous Black-Scholes formula which can be written in the form

$$V(S,t) = S\Phi(d_1) - Ee^{-r(T-t)}\Phi(d_2) \tag{2.58}$$

where

$$d_1 = \frac{\log(S/E) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, d_2 = d_1 - \sigma\sqrt{T-t}$$

are the values $\pm\sigma^2(T-t)/2$ standardized by adding $log(S/E)+r(T-t)$ and dividing by $\sigma\sqrt{T-t}$. This may be derived by the following device; Assume (i.e. pretend) that, given current information, the distribution of $S(T)$ at expiry is lognormally distributed with the mean $\eta = S(t)e^{r(T-t)}$.

The mean of the log-normal in the risk neutral world $S(t)e^{r(T-t)}$ is exactly the future value of our current stocks $S(t)$ if we were to sell the stock and invest the cash in a bank deposit. Then the future value of an option with payoff function given by $V_0(S_T)$ is the expected value of this function against this lognormal probability density function, then discounted to present value

$$e^{-r(T-t)}\int_0^\infty V_0(x)g(x|S(t)e^{r(T-t)},\ \sigma\sqrt{T-t})dx. \tag{2.59}$$

Notice that the Black-Scholes derivation covers any diffusion process governing the underlying asset which is driven by a stochastic differential equation of the form

$$dS = a(S)dt + \sigma S dW_t \tag{2.60}$$

regardless of the nature of the drift term $a(S)$. For example a non-linear function $a(S)$ can lead to distributions that are not lognormal and yet the option price is determined as if it were.

Example: Pricing Call and Put options.

Consider pricing an index option on the S&P 500 index an January 11, 2000 (the index SPX closed at 1432.25 on this day). The option SXZ AE-A is a January call option with strike price 1425. The option matures (as do equity options in general) on the third Friday of the month or January 21, a total of 7 trading days later. Suppose we wish to price such an option using the Black-Scholes model. In this case, $T-t$ measured in years is $7/252 = 0.027778$. The annual volatility of the Standard and Poor 500 index is around 19.5 percent or 0.195 and assume the very short term interest rates approximately 3%. In *Matlab* we can value this option using

[CALL,PUT] = BLSPRICE(1432.25,1425,0.03,7/252,0.195,0)

CALL = 23.0381

PUT = 14.6011

Arguments of the function BLSPRICE are, in order, the current equity price, the strike price, the annual interest rate r, the time to maturity $T-t$ in years, the annual volatility σ and the last argument is the dividend yield in percent which we assumed 0. Thus the Black-Scholes price for a call option on SPX is around 23.03. Indeed this call option did sell on Jan 11 for $23.00. and the put option for $14 5/8. From the put call parity relation (see for example Wilmott, Howison, Dewynne, page 41) $S + P - C = Ee^{-r(T-t)}$ or in this case $1432.25 + 14.625 - 23 = 1425e^{-r(7/252)}$. We might solve this relation to obtain the spot interest rate r. In order to confirm that a different interest rate might apply over a longer term, we consider the September call and put options (SXZ) on the same day with exercise price 1400 which sold for $152 and 71$ respectively. In this case there are171 trading days to expiry and so we need to solve $1432.25 + 71 - 152 = 1400e^{-r(171/252)}$, whose solution is $r = 0.0522$. This is close to the six month interest rates at the time, but 3% is low for the very short term rates. The discrepancy with the actual interest rates is one of several modest failures of the Black-Scholes model to be discussed further later. The low implied interest rate is influenced by the cost of handling and executing an option, which are non-negligible fractions of the option prices, particularly with short term options such as this one. An analogous function to the Matlab function above which provides the Black-Scholes price in Splus or **R** is given below:

```
blsprice=function(So,strike,r,T,sigma,div){
d1<-(log(So/strike)+(r-div+(sigma^2)/2)*T)/(sigma*sqrt(T))
d2<-d1-sigma*sqrt(T)
call<-So*exp(-div*T)*pnorm(d1)-exp(-r*T)*strike*pnorm(d2)
put=call-So+strike*exp(-r*T)
```

```
c(call,put)}
```

Problems

1. It is common for a stock whose price has reached a high level to *split* or issue shares on a two-for-one or three-for-one basis. What is the effect of a stock split on the price of an option?

2. If a stock issues a dividend of exactly D (known in advance) on a certain date, provide a no-arbitrage argument for the change in price of the stock at this date. Is there a difference between deterministic D and the case when D is a random variable with known distribution but whose value is declared on the dividend date?

3. Suppose Σ is a positive definite covariance matrix and η a column vector. Show that the set of all possible pairs of standard deviation and mean return $(\sqrt{w^T \Sigma w}, \eta^T w)$ for weight vector w such that $\sum_i w_i = 1$ is a convex region with a hyperbolic boundary.

4. The current rate of interest is 5% per annum and you are offered a random bond which pays either \$210 or \$0 in one year. You believe that the probability of the bond paying \$210 is one half. How much would you pay now for such a bond? Suppose this bond is publicly traded and a large fraction of the population is risk averse so that it is selling now for \$80. Does your price offer an arbitrage to another trader? What is the risk-neutral measure for this bond?

5. Which would you prefer, a gift of \$100 or a 50-50 chance of making \$200? A fine of \$100 or a 50-50 chance of losing \$200? Are your preferences self-consistent and consistent with the principle that individuals are risk-averse?

6. Compute the stochastic differential dX_t (assuming W_t is a Wiener process) when

   (a) $X_t = \exp(rt)$

   (b) $X_t = \int_0^t h(t)dW_t$

   (c) $X_t = X_0 \exp\{at + bW_t\}$

   (d) $X_t = \exp(Y_t)$ where $dY_t = \mu dt + \sigma dW_t$.

7. Show that if X_t is a geometric Brownian motion, so is X_t^β for any real number β.

8. Suppose a stock price follows a geometric Brownian motion process

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

   Find the diffusion equation satisfied by the processes (a) $f(S_t) = S_t^n$,(b) $\log(S_t)$, (c) $1/S_t$. Find a combination of the processes S_t and $1/S_t$ that does not depend on the drift parameter μ. How does this allow constructing estimators of σ that do not require knowledge of the value of μ?

9. Consider an Ito process of the form

$$dS_t = a(S_t)dt + \sigma(S_t)dW_t$$

   Is it possible to find a function $f(S_t)$ which is also an Ito process but with zero drift?

10. Consider an Ito process of the form

$$dS_t = a(S_t)dt + \sigma(S_t)dW_t$$

    Is it possible to find a function $f(S_t)$ which has constant diffusion term?

11. Consider approximating an integral of the form $\int_0^T g(t)dW_t \approx \sum g(t)\{W(t+h) - W(t)\}$ where $g(t)$ is a non-random function and the sum is over values of $t = nh, n = 0, 1, 2, ...T/h - 1$. Show by considering the distribution

of the sum and taking limits that the random variable $\int_0^T g(t)dW_t$ has a normal distribution and find its mean and variance.

12. Consider two geometric Brownian motion processes X_t and Y_t both driven by the same Wiener process

$$dX_t = aX_t dt + bX_t dW_t$$
$$dY_t = \mu Y_t dt + \sigma Y_t dW_t.$$

Derive a stochastic differential equation for the ratio $Z_t = X_t/Y_t$. Suppose for example that X_t models the price of a commodity in \$C and Y_t is the exchange rate ($\$C/\$US$) at time t. Then what is the process Z_t? Repeat in the more realistic situation in which

$$dX_t = aX_t dt + bX_t dW_t^{(1)}$$
$$dY_t = \mu Y_t dt + \sigma Y_t dW_t^{(2)}$$

and $W_t^{(1)}, W_t^{(2)}$ are correlated Brownian motion processes with correlation ρ.

13. Prove the *Shannon inequality* that

$$H(Q,P) = \sum q_i \log(\frac{q_i}{p_i}) \geq 0$$

for any probability distributions P and Q with equality if and only if all $p_i = q_i$.

14. Consider solving the problem

$$\min_q H(Q,P) = \sum q_i \log(\frac{q_i}{p_i})$$

subject to the constraints $\sum_i q_i = 1$ and $E_Q f(X) = \sum q_i f(i) = \mu$. Show that the solution, if it exists, is given by

$$q_i = \frac{\exp(\eta f(i))}{m(\eta)} p_i$$

where $m(\eta) = \sum_i p_i \exp(\eta f(i))]$ and η is chosen so that $\frac{m'(\eta)}{m(\eta)} = \mu$. (This shows that the closest distribution to P which satisfies the constraint is obtained by a simple "exponential tilt" or Esscher transform so that $\frac{dQ}{dP}(x)$ is proportional to $\exp(\eta f(x))$ for a suitable parameter η).

15. Let Q^* minimize $H(Q,P)$ subject to a constraint

$$E_Q g(X) = c. \tag{2.61}$$

Let Q be some other probability distribution satisfying the same constraint. Then prove that

$$H(Q,P) = H(Q,Q^*) + H(Q^*,P).$$

16. Let I_1, I_2 ,... be a set of constraints of the form

$$E_Q g_i(X) = c_i \tag{2.62}$$

and suppose we define P_n^* as the solution of

$$\max_P H(P)$$

subject to the constraints $I_1 \cap I_2 \cap ...I_n$. Then prove that

$$H(P_n^*, P_1^*) = H(P_n^*, P_{n-1}^*) + H(P_{n-1}^*, P_{n-2}^*) + ... + H(P_2^*, P_1^*).$$

17. Consider a defaultable bond which pays a fraction of its face value Fp on maturity in the event of default. Suppose the risk free interest rate continuously compounded is r so that $B_s = \exp(sr)$. Suppose also that a constant coupon $\$d$ is paid at the end of every period $s = t+1, ..., T-1$. Then show that the value of this bond at time t is

$$P_t = d\frac{\exp\{-(r+k)\} - \exp\{-(r+k)\{T-t)\}}{1-\exp\{-(r+k)\}}$$
$$+ pF\exp\{-r(T-t)\} + (1-p)F\exp\{-(r+k)(T-t)\}$$

18. (a) Show that entropy is always positive and if $Y = g(X)$ is a function of X then Y has smaller entropy than X, i.e. $H(p_Y) \leq H(p_X)$.

    (b) Show that if X has any discrete distribution over n values, then its entropy is $\leq log(n)$.

Chapter 3

Basic Monte Carlo Methods

Consider as an example the following very simple problem. We wish to price a European call option with exercise price \$22 and payoff function $V(S_T) = (S_T - 22)^+$. Assume for the present that the interest rate is 0% and S_T can take only the following five values with corresponding risk neutral (Q) probabilities

s	20	21	22	23	24
$Q[S_T = s]$	1/16	4/16	6/16	4/16	1/16

In this case, since the distribution is very simple, we can price the call option explicitly;

$$E_Q V(S_T) = E_Q(S_T - 22)^+ = (23 - 22)\frac{4}{16} + (24 - 22)\frac{1}{16} = \frac{3}{8}.$$

However, the ability to value an option explicitly is a rare luxury. An alternative would be to generate a large number (say $n = 1000$) independent simulations of the stock price S_T under the measure Q and average the returns from the option. Say the simulations yielded values for S_T of $22, 20, 23, 21, 22, 23, 20, 24, ...$ then

the estimated value of the option is

$$\begin{aligned}\overline{V(S_T)} &= \frac{1}{1000}[(22-22)^+ + (20-22)^+ + (23-22)^+ + ...]. \\ &= \frac{1}{1000}[0+0+1+....]\end{aligned}$$

The law of large numbers assures us for a large number of simulations n, the average $\overline{V(S_T)}$ will approximate the true expectation $E_Q V(S_T)$. Now while it would be foolish to use simulation in a simple problem like this, there are many models in which it is much easier to randomly generate values of the process S_T than it is to establish its exact distribution. In such a case, simulation is the method of choice.

Randomly generating a value of S_T for the discrete distribution above is easy, provided that we can produce independent random uniform random numbers on a computer. For example, if we were able to generate a random number Y_i which has a uniform distribution on the integers $\{0, 1, 2, ..., 15\}$ then we could define S_T for the i'th simulation as follows:

If Y_i is in set	$\{0\}$	$\{1,2,3,4\}$	$\{5,6,7,8,9,10\}$	$\{11,12,13,14\}$	$\{15\}$
define $S_T =$	20	21	22	23	24

Of course, to get a reasonably accurate estimate of the price of a complex derivative may well require a large number of simulations, but this is decreasingly a problem with increasingly fast computer processors. The first ingredient in a simulation is a stream of uniform random numbers Y_i used above. In practice all other distributions are generated by processing discrete uniform random numbers. Their generation is discussed in the next section.

Uniform Random Number Generation

The first requirement of a stochastic model is the ability to generate "random" variables or something resembling them. Early such generators attached to computers exploited physical phenomena such as the least significant digits in

an accurate measure of time, or the amount of background cosmic radiation as the basis for such a generator, but these suffer from a number of disadvantages. They may well be "random" in some more general sense than are the pseudo-random number generators that are presently used but their properties are difficult to establish, and the sequences are impossible to reproduce. The ability to reproduce a sequence of random numbers is important for debugging a simulation program and for reducing its variance.

It is quite remarkable that some very simple recursion formulae define sequences that behave like sequences of independent random numbers and *appear* to more or less obey the major laws of probability such as the law of large numbers, the central limit theorem, the Glivenko-Cantelli theorem, etc. Although computer generated *pseudo random numbers* have become more and more like independent random variables as the knowledge of these generators grows, the main limit theorems in probability such as the law of large numbers and the central limit theorem still do not have versions which directly apply to dependent sequences such as those output by a random number generator. The fact that certain pseudo-random sequences appear to share the properties of independent sequences is still a matter of observation rather than proof, indicating that many results in probability hold under much more general circumstances than the relatively restrictive conditions under which these theorems have so far been proven. One would intuitively expect an enormous difference between the behaviour of independent random variables X_n and a deterministic (i.e. non-random) sequence satisfying a recursion of the form $x_n = g(x_{n-1})$ for a simple function g. Surprisingly, for many carefully selected such functions g it is quite difficult to determine the difference between such a sequence and an independent sequence. Of course, numbers generated from a simple recursion such as this are neither random, nor are x_{n-1} and x_n independent. We sometimes draw attention to this by referring to such a sequence as *pseudo-random numbers.* While they are in no case independent, we will nevertheless attempt to find

simple functions g which provide behaviour *similar* to that of independent uniform random numbers. The search for a satisfactory random number generator is largely a search for a suitable function g, possibly depending on more than one of the earlier terms of the sequence, which imitates in many different respects the behaviour of independent observations with a specified distribution.

Definition: reduction modulo m. For positive integers x and m, the value $a \bmod m$ is the remainder (between 0 and $m-1$) obtained when a is divided by m. So for example $7 \bmod 3 = 1$ since $7 = 2 \times 3 + 1$.

The single most common class of random number generators are of the form

$$x_n := (ax_{n-1} + c) \mod m$$

for given integers $a, c,$ and m which we select in advance. This generator is initiated with a "seed" x_0 and then run to produce a whole sequence of values. When $c = 0$, these generators are referred to as *multiplicative congruential generators* and in general as *mixed or linear congruential generators*. The "seed", x_0, is usually updated by the generator with each call to it. There are two common choices of m, either m prime or $m = 2^k$ for some k (usually 31 for 32 bit machines).

Example: Mixed Congruential generator

Define $x_n = (5x_{n-1} + 3) \bmod 8$ and the seed $x_0 = 3$. Note that by this recursion

$$x_1 = (5 \times 3 + 3) \bmod 8 = 18 \bmod 8 = 2$$

$$x_2 = 13 \bmod 8 = 5$$

$$x_3 = 28 \bmod 8 = 4$$

and $x_4, x_5, x_{6,} x_7, x_8 = 7, 6, 1, 0, 3$ respectively

and after this point (for $n > 8$) the recursion will simply repeat again the pattern already established, $3, 2, 5, 4, 7, 6, 1, 0, 3, 2, 5, 4, ...$

The above repetition is inevitable for a linear congruential generator. There are at most m possible numbers after reduction mod m and once we arrive back at the seed the sequence is destined to repeat itself. In the example above, the sequence cycles after 8 numbers. The length of one cycle, before the sequence begins to repeat itself again, is called the *period* of the generator. For a mixed generator, the period must be less than or equal to m. For multiplicative generators, the period is shorter, and often considerably shorter.

Multiplicative Generators.

For multiplicative generators, $c = 0$. Consider for example the generator $x_n = 5x_{n-1} \; mod \; 8$ and $x_0 = 3$. This produces the sequence $3, 7, 3, 7, ...$ In this case, the period is only 2, but for general m, it is clear that the maximum possible period is $m-1$ because it generates values in the set $\{1, ..., m-1\}$. The generator cannot generate the value 0 because if it did, all subsequent values generated are identically 0. Therefore the maximum possible period corresponds to a cycle through non-zero integers exactly once. But in the example above with $m = 2^k$, the period is far from attaining its theoretical maximum, $m - 1$. The following Theorem shows that the period of a multiplicative generator is maximal when m is a prime number and a satisfies some conditions.

Theorem 14 *(period of multiplicative generator).*
If m is prime, the multiplicative congruential generator $x_n = ax_{n-1}(mod\, m)$, $a \neq 0$, has maximal period $m - 1$ if and only if $a^i \neq 1(mod\ m)$ for all $i = 1, 2, ..., m - 1$.

If m is a prime, and if the condition $a^{m-1} = 1(mod\ m)$ and $a^i \neq 1(mod\ m)$ for all $i < m - 1$ holds, we say that a is a *primitive root* of m, which means

that the powers of a generate all of the possible elements of the multiplicative group of integers mod m. Consider the multiplicative congruential generator $x_n = 2x_{n-1} \bmod 11$. It is easy to check that $2^i \bmod 11 = 2, 4, 8, 5, 10, 9, 7, 3, 6, 1$ as $i = 1, 2, ...10$. Since the value $i = m-1$ is the first for which $2^i \bmod 11 = 1$, 2 is a primitive root of 11 and this is a maximal period generator having period 10. When $m = 11$, only the values $a = 2, 6, 7, 8$ are primitive roots and produce full period (10) generators.

One of the more common moduli on 32 bit machines is the Mersenne prime $m = 2^{31} - 1$. In this case, the following values of a (among many others) all produce full period generators:

$$a = 7, 16807, 39373, 48271, 69621, 630360016, 742938285, 950706376,$$

$$1226874159, 62089911, 1343714438$$

Let us suppose now that m is prime and a_2 is the multiplicative inverse (mod m) of a_1 by which we mean $(a_1 a_2) \bmod m = 1$. When m is prime, the set of integers $\{0, 1, 2, ..., m-1\}$ together with the operations of addition and multiplication $\bmod m$ forms what is called a *finite field.* This is a finite set of elements together with operations of addition and multiplication such as those we enjoy in the real number system. For example for integers $x_1, a_1, a_2 \in \{0, 1, 2, ..., m-1\}$, the product of a_1 and x_1 can be defined as $(a_1 x_1) \bmod m = x_2$, say. Just as non-zero numbers in the real number system have multiplicative inverses, so too do non=zero elements of this field. Suppose for example a_2 is the muultiplicative inverse of a_1 so that $a_2 a_1 \bmod m = 1$. If we now multiply x_2 by a_2 we have

$$(a_2 x_2) \bmod m = (a_2 a_1 x_1) \bmod m = (a_2 a_1 \bmod m)(x_1 \bmod m) = x_1.$$

This shows that $x_1 = (a_2 x_2) \bmod m$ is equivalent to $x_2 = (a_1 x_1) \bmod m$. In other words, using a_2 the multiplicative inverse of $a_1 \bmod m$, the multiplicative generator with multiplier a_2 generates exactly the same sequence as that with

multiplier a_1 except in reverse order. Of course if a is a primitive root of m, then so is its multiplicative inverse.

Theorem 15 *(Period of Multiplicative Generators with $m = 2^k$)*
If $m = 2^k$ with $k \geq 3$, and if $a \bmod 8 = 3$ or 5 and x_0 is odd, then the multiplicative congruential generator has maximal period $= 2^{k-2}$.

For the proof of these results, see Ripley(1987), Chapter 2. The following simple Matlab code allows us to compare linear congruential generators with small values of m. It generates a total of n such values for user defined a, c, m, x_0 =seed. The efficient implementation of a generator for large values of m depends very much on the architecture of the computer. We normally choose m to be close to the machine precision (e.g. 2^{32} for a 32 bit machine.

```
function x=lcg(x0,a,c,m,n)
y=x0;      x=x0;
for i=1:n   ;      y=rem(a*y+c,m);      x=[x y];      end
```

The period of a linear congruential generator varies both with the multiplier a and the constant c. For example consider the generator

$$x_n = (ax_{n-1} + 1) \bmod 2^{10}$$

for various multipliers a. When we use an even multiplier such as $a = 2, 4, ...$(using seed 1) we end up with a sequence that eventually locks into a specific value. For example with $a = 8$ we obtain the sequence 1,9,73,585,585,....never changing beyond that point. The periods for odd multipliers are listed below (all started with seed 1)

a	1	3	5	7	9	11	13	15	17	19	21	23	25
Period	1024	512	1024	256	1024	512	1024	128	1024	512	1024	256	1024

The astute reader will notice that the only full-period multipliers a are those which are multipliers of 4. This is a special case of the following theorem.

Theorem 16 *(Period of Mixed or Linear Congruential Generators.)*
The Mixed Congruential Generator,

$$x_n = (ax_{n-1} + c) \bmod m \tag{3.1}$$

has full period m *if and only if*
(i) c *and* m *are relatively prime.*
(ii) Each prime factor of m *is also a factor of* $a - 1$.
(iii) If 4 *divides* m *it also divides* $a - 1$.

When m is prime, (ii) together with the assumption that $a < m$ implies that m must divide $a - 1$ which implies $a = 1$. So for prime m the only full-period generators correspond to $a = 1$. Prime numbers m are desirable for long periods in the case of multiplicative generators, but in the case of mixed congruential generators, only the trivial one $x_n = (x_{n-1} + c)(mod\ m)$ has maximal period m when m is prime. This covers the popular Mersenne prime $m = 2^{31} - 1$..

For the generators $x_n = (ax_{n-1} + c) \bmod 2^k$ where $m = 2^k, k \geq 2$, the condition for full period 2^k requires that c is odd, and $a = 4j + 1$ for some integer j.

Some of the linear or multiplicative generators which have been suggested are the following:

m	a	c	
$2^{31}-1$	$7^5=16807$	0	Lewis,Goodman, Miller (1969)IBM,
$2^{31}-1$	630360016	0	Fishman (Simscript II)
$2^{31}-1$	742938285	0	Fishman and Moore
2^{31}	65539	0	RANDU
2^{32}	69069	1	Super-Duper (Marsaglia)
2^{32}	3934873077	0	Fishman and Moore
2^{32}	3141592653	1	DERIVE
2^{32}	663608941	0	Ahrens (C-RAND)
2^{32}	134775813	1	Turbo-Pascal,Version 7(period= 2^{32})
2^{35}	5^{13}	0	APPLE
$10^{12}-11$	427419669081	0	MAPLE
2^{59}	13^{13}	0	NAG
$2^{61}-1$	$2^{20}-2^{19}$	0	Wu (1997)

Table 3.1: Some Suggested Linear and Multiplicative Random Number Generators

Other Random Number Generators.

A generalization of the linear congruential generators which use a k-dimensional vectors X has been considered, specifically when we wish to generate correlation among the components of X. Suppose the components of X are to be integers between 0 and $m-1$ where m is a power of a prime number. If A is an arbitrary $k \times k$ matrix with integral elements also in the range $\{0, 1, ..., m-1\}$ then we begin with a vector-valued seed X_0, a constant vector C and define recursively

$$X_n := (AX_{n-1} + C) \bmod m$$

Such generators are more common when C is the zero vector and called *matrix multiplicative congruential generators*. A related idea is to use a higher order

recursion like

$$x_n = (a_1 x_{n-1} + a_2 x_{n-2} + .. + a_k x_{n-k}) \bmod m, \tag{3.2}$$

called a *multiple recursive generator*. L'Ecuyer (1996,1999) combines a number of such generators in order to achieve a period around $2^{319}$and good uniformity properties. When a recursion such as (3.2) with $m = 2$ is used to generate pseudo-random bits $\{0, 1\}$, and these bits are then mapped into uniform (0,1) numbers, it is called a *Tausworthe* or *Feedback Shift Register* generators. The coefficients a_i are determined from *primitive polynomials* over the *Galois Field*.

In some cases, the uniform random number generator in proprietary packages such as *Splus* and *Matlab* are not completely described in the package documentation. This is a further recommendation of the transparency of packages like **R**. Evidently in *Splus*, the multiplicative congruential generator is used, and then the sequence is " shuffled" using a Shift-register generator (a special case of the matrix congruential generator described above). This secondary processing of the sequence can increase the period but it is not always clear what other effects it has. In general, shuffling is conducted according to the following steps

1. Generate a sequence of pseudo-random numbers x_i using $x_{i+1} = a_1 x_i (\bmod m_1)$.
2. For fixed k put $(T_1, \ldots, T_k) = (x_1, \ldots, x_k)$.
3. Generate, using a different generator, a sequence $y_{i+1} = a_2 y_i (\bmod m_2)$.
4. Output the random number T_I where $I = \lceil Y_i k/m_2 \rceil$.
5. Increment i, replace T_I by the next value of x, and return to step 3.

One generator is used to produce the sequence x, numbers needed to fill k holes. The other generator is then used select which hole to draw the next number from or to "shuffle" the x sequence.

Example: A shuffled generator

Consider a generator described by the above steps with $k = 4, x_{n+1} = (5x_n)(\bmod_1 19)$ and $y_{n+1} = (5y_n)(\bmod\ 29)$

$x_n =$ 3 15 18 14 13 8 2

$y_n =$ 3 15 17 27 19 8 11

We start by filling four pigeon-holes with the numbers produced by the first generator so that $(T_1, \ldots, T_4) = (3, 15, 18, 14)$. Then use the second generator to select a random index I telling us which pigeon-hole to draw the next number from. Since these holes are numbered from 1 through 4, we use $I = \lceil 4 \times 3/29 \rceil = 1$. Then the first number in our random sequence is drawn from box 1, i.e. $z_1 = T_1 = 3$, so $z_1 = 3$. This element T_1 is now replaced by 13, the next number in the x sequence. Proceeding in this way, the next index is $I = \lceil 4 \times 15/29 \rceil = 3$ and so the next number drawn is $z_2 = T_3 = 18$. Of course, when we have finished generating the values $z_1, z_2, ...$ all of which lie between 1 and $m_1 = 18$, we will usually transform them in the usual way (e.g. z_i/m_1) to produce something approximating continuous uniform random numbers on [0,1]. When m_1, is large, it is reasonable to expect the values z_i/m_1 to be approximately continuous and uniform on the interval $[0, 1]$. One advantage of shuffling is that the period of the generator is usually greatly extended. Whereas the original x sequence had period 9 in this example, the shuffled generator has a larger period or around 126.

There is another approach, summing pseudo-random numbers, which is also used to extend the period of a generator. This is based on the following theorem (see L'Ecuyer (1988)). For further discussion of the effect of taking linear combinations of the output from two or more random number generators, see Fishman (1995, Section 7.13).

Theorem 17 *(Summing mod m)*
If X is random variable uniform on the integers $\{0, \ldots, m-1\}$ and if Y is any integer-valued random variable independent of X, then the random variable $W = (X + Y)(mod\ m)$ is uniform on the integers $\{0, \ldots, m-1\}$.

Theorem 18 *(Period of generator summed mod m_1)*

If $x_{i+1} = a_1 x_i \bmod m_1$ *has period* $m_1 - 1$ *and* $y_{i+1} = a_2 y_i \bmod m_2$ *has period* $m_2 - 1$*, then* $(x_i + y_i) (\bmod m_1)$ *has period the least common multiple of* $(m_1 - 1, m_2 - 1)$.

Example: summing two generators

If $x_{i+1} = 16807 x_i mod(2^{31} - 1)$ and $y_{i+1} = 40692 y_i mod(2^{31} - 249)$, then the period of $(x_i + y_i) mod(2^{31} - 1)$ is

$$\frac{(2^{31} - 2)(2^{31} - 250)}{2 \times 31} \approx 7.4 \times 10^{16}$$

This is much greater than the period of either of the two constituent generators.

Other generators.

One such generator, the "Mersenne-Twister", from Matsumoto and Nishimura (1998) has been implemented in **R** and has a period of $2^{19937} - 1$. Others use a non-linear function g in the recursion $x_{n+1} = g(x_n) (\bmod\ m)$ to replace a linear one For example we might define $x_{n+1} = x_n^2 (\bmod m)$ (called a *quadratic residue generator*) or $x_{n+1} = g(x_n) \bmod m$ for a quadratic function or some other non-linear function g. Typically the function g is designed to result in large values and thus more or less random low order bits. I*nversive congruential generators* generate x_{n+1} using the $(\bmod\ m)$ inverse of x_n.

Other generators which have been implemented in R include: the *Wichmann-Hill* (1982,1984) generator which uses three multiplicative generators with prime moduli 30269, 30307, 30323 and has a period of $\frac{1}{4}(30268 \times 30306 \times 30322)$. The outputs from these three generators are converted to $[0, 1]$ and then summed mod 1. This is similar to the idea of Theorem 17, but the addition takes place after the output is converted to [0,1]. See *Applied Statistics* (1984), **33**, 123. Also implemented are *Marsaglia's Multicarry* generator which has a period of more than 2^{60} and reportedly passed all tests (according to Marsaglia), Marsaglia's "Super-Duper", a linear congruential generator listed in Table 1,

and two generators developed by Knuth (1997,2002) the *Knuth-TAOCP* and *Knuth-TAOCP-2002*.

Conversion to Uniform $(0,1)$ **generators:**

In general, random integers should be mapped into the unit interval in such a way that the values 0 and 1, each of which have probability 0 for a continuous distribution are avoided. For a multiplicative generator, since values lie between 1 and $m-1$, we may divide the random number by m. For a linear congruential generator taking possible values $x \in \{0, 1, ..., m-1\}$, it is suggested that we use $(x + 0.5)/m$.

Apparent Randomness of Pseudo-Random Number Generators

Knowing whether a sequence behaves in all respects like independent uniform random variables is, for the statistician, pretty close to knowing the meaning of life. At the very least, in order that one of the above generators be reasonable approximations to independent uniform variates it should satisfy a number of statistical tests. Suppose we reduce the uniform numbers on $\{0, 1, ..., m-1\}$ to values approximately uniformly distributed on the unit interval $[0, 1]$ as described above either by dividing through by m or using $(x + 0.5)/m$. There are many tests that can be applied to determine whether the hypothesis of independent uniform variates is credible (not, of course, whether the hypothesis is *true*. We know by the nature of all of these pseudo-random number generators that it is not!).

Runs Test

We wish to test the hypothesis H_0 that a sequence$\{U_i, i = 1, 2, ..., n\}$ consists of n independent identically distributed random variables under the assumption that they have a continuous distribution. The *runs test* measures runs, either in the original sequence or in its differences. For example, suppose we denote a positive difference between consecutive elements of the sequence by $+$ and a negative difference by $-$. Then we may regard a sequence of the form $.21, .24, .34, .37, .41, .49, .56, .51, .21, .25, .28, .56, .92, .96$ as unlikely under independence because the corresponding differences $+++++++--++++++$ have too few "runs" (the number of runs here is $R = 3$). Under the assumption that the sequence $\{U_i, i = 1, 2, ..., n\}$ is independent and continuous, it is possible to show that $E(R) = \frac{2n-1}{3}$and $var(R) = \frac{3n-5}{18}$. The proof of this result is a problem at the end of this chapter. We may also approximate the distribution of R with the normal distribution for $n \geq 25$. A test at a 0.1% level of significance is therefore: reject the hypothesis of independence if

$$\left| \frac{R - \frac{2n-1}{3}}{\sqrt{\frac{3n-5}{18}}} \right| > 3.29,$$

where 3.29 is the corresponding $N(0, 1)$ quantile. A more powerful test based on runs compares the lengths of the runs of various lengths (in this case one run up of length 7, one run down of length 3, and one run up of length 6) with their theoretical distribution.

Another test of independence is the *serial correlation test*. The runs test above is one way of checking that the pairs $(U_n, U_{n+1})$are approximately uniformly distributed on the unit square. This could obviously be generalized to pairs like (U_i, U_{i+j}). One could also use the sample correlation or covariance as the basis for such a test. For example, for $j \geq 0$,

$$C_j = \frac{1}{n}(U_1U_{1+j} + U_2U_{2+j} + ..U_{n-j}U_n + U_{n+1-j}U_1 + ... + U_nU_j) \qquad (3.3)$$

The test may be based on the normal approximation to the distribution of C_j with mean $E(C_0) = 1/3$ and $E(C_j) = 1/4$ for $j \geq 1$. Also

$$var(C_j) = \begin{cases} \frac{4}{45n} & \text{for} \quad j = 0 \\ \frac{13}{144n} & \text{for} \quad j \geq 1,\, j \neq \frac{n}{2} \\ \frac{7}{72n} & \text{for} \quad j = \frac{n}{2} \end{cases}$$

Such a test, again at a 0.1% level will take the form: reject the hypothesis of independent uniform if

$$\left| \frac{C_j - \frac{1}{4}}{\sqrt{\frac{13}{144n}}} \right| > 3.29.$$

for a particular preselected value of j (usually chosen to be small, such as $j = 1, ...10$).

Chi-squared test.

The chi-squared test can be applied to the sequence in any dimension, for example $k = 2$. Suppose we have used a generator to produce a sequence of uniform$(0, 1)$ variables, $U_j, j = 1, 2, ...2n$, and then, for a partition $\{A_i; i = 1, ..., K\}$ of the unit square, we count N_i, the number of pairs of the form $(U_{2j-1}, U_{2j}) \in A_i$. See for example the points plotted in Figure 3.1. Clearly this should be related to the area or probability $P(A_i)$ of the set A_i. Pearson's chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{[N_i - nP(A_i)]^2}{nP(A_i)} \tag{3.4}$$

which should be compared with a chi-squared distribution with degrees of freedom $K-1$ or one less than the number of sets in the partition. Observed values of the statistic that are unusually large for this distribution should lead to rejection of the uniformity hypothesis. The partition usually consists of squares of identical area but could, in general, be of arbitrary shape.

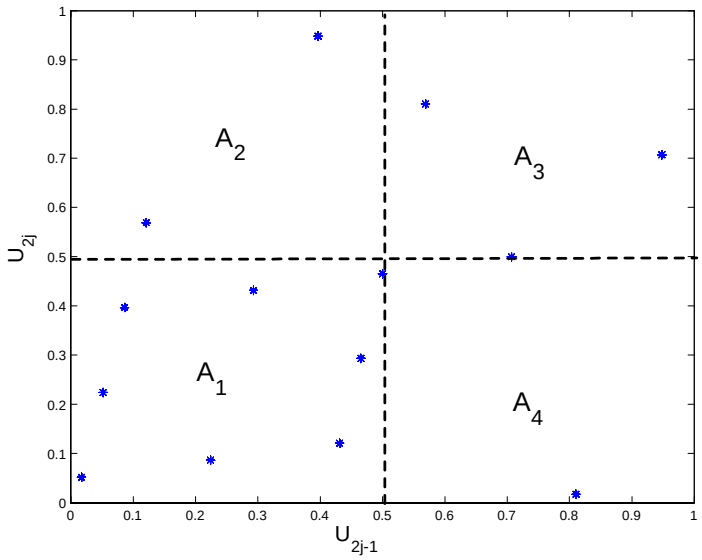


Figure 3.1: The Chi-squared Test

Spectral Test

Consecutive values plotted as pairs (x_n, x_{n+1}), when generated from a multiplicative congruential generator $x_{n+1} = ax_n (\mathrm{mod}\, m)$ fall on a *lattice*. A lattice is a set of points of the form $t_1 e_1 + t_2 e_2$ where t_1, t_2 range over all integers and e_1, e_2 are vectors, (here two dimensional vectors since we are viewing these points in pairs of consecutive values (x_n, x_{n+1})) called the "basis" for the lattice. A given lattice, however, has many possible different bases, and in order to analyze the lattice structure, we need to isolate the most "natural" basis, e.g. the one that we tend to see in viewing a lattice in two dimensions. Consider, for example, the lattice formed by the generator $x_n = 23x_{n-1} \bmod 97$. A plot of adjacent pairs (x_n, x_{n+1}) is given in Figure 3.2. For basis vectors we could use $e_1 = (1, 23)$ and $e_2 = (4, -6)$, or we could replace e_1 by $(5, 18)$or $(9, 13)$ etc. Beginning at an arbitrary point O on the lattice as origin (in this case, since the original point (0,0) is on the lattice, we will leave it unchanged), we choose an unambiguous

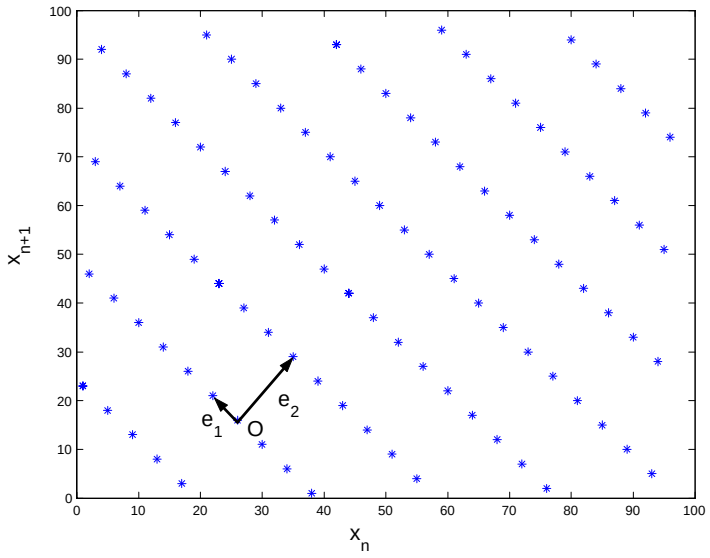


Figure 3.2: The Spectral Test

definition of e_1 to be the *shortest* vector in the lattice, and then define e_2 as the shortest vector in the lattice which is not of the form te_1 for integer t. Such a basis will be called a *natural basis*. The best generators are those for which the cells in the lattice generated by the 2 basis vectors e_1, e_2 or the parallelograms with sides parallel to e_1, e_2 are as close as possible to squares so that e_1 and e_2 are approximately the same length. As we change the multiplier a in such a way that the random number generator still has period $\simeq m$, there are roughly m points in a region above with area approximately m^2 and so the area of a parallelogram with sides e_1 and e_2 is approximately a constant (m) whatever the multiplier a. In other words a longer e_1 is associated with a shorter vector e_2 and therefore for an ideal generator, the two vectors of reasonably similar length. A poor generator corresponds to a basis with e_2 much longer than e_1. The *spectral test statistic* ν is the renormalized length of the first basis vector $||e_1||$. The extension to a lattice in k-dimensions is done similarly. All linear

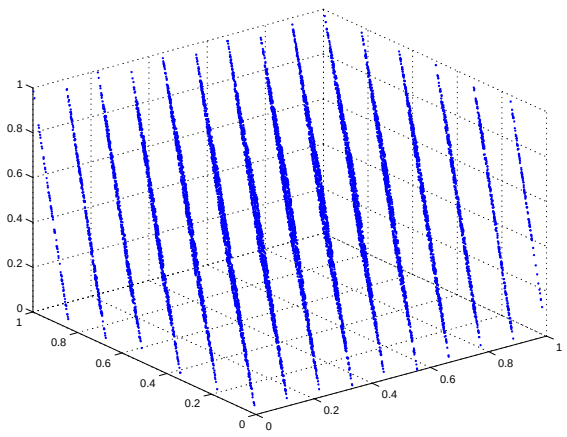


Figure 3.3: Lattice Structure of Uniform Random Numbers generated from RANDU

congruential random number generators result in points which when plotted as consecutive k-tuples lie on a lattice. In general, for k consecutive points, the spectral test statistic is equal to $min(b_1^2 + b_2^2 + . + . + . + b_k^2)^{1/2}$ under the constraint $b_1 + b_2 a + ...b_k a^{k-1} = mq, q \neq 0$. Large values of the statistic indicate that the generator is adequate and Knuth suggests as a minimum threshold the value $\pi^{-1/2}[(k/2)!m/10]^{1/k}$.

One of the generators that fails the spectral test most spectacularly with $k = 3$ is the generator RANDU, $x_{n+1} = 65539\ x_n (\text{mod}\ 2^{31})$. This was used commonly in simulations until the 1980's and is now notorious for the fact that a small number of hyperplanes fit through all of the points (see Marsaglia, 1968). For RANDU, successive triplets tend to remain on the plane $x_n = 6x_{n-1} - 9x_{n-2}$. This may be seen by rotating the 3-dimensional graph of the sequence of triplets of the form $\{(x_{n-2}, x_{n-1}, x_n); n = 2, 3, 4, ...N\}$ as in Figure 3.3

As another example, in Figure 3.4 we plot 5000 consecutive triplets from a linear congruential random number generator with $a = 383, c = 263, m =$

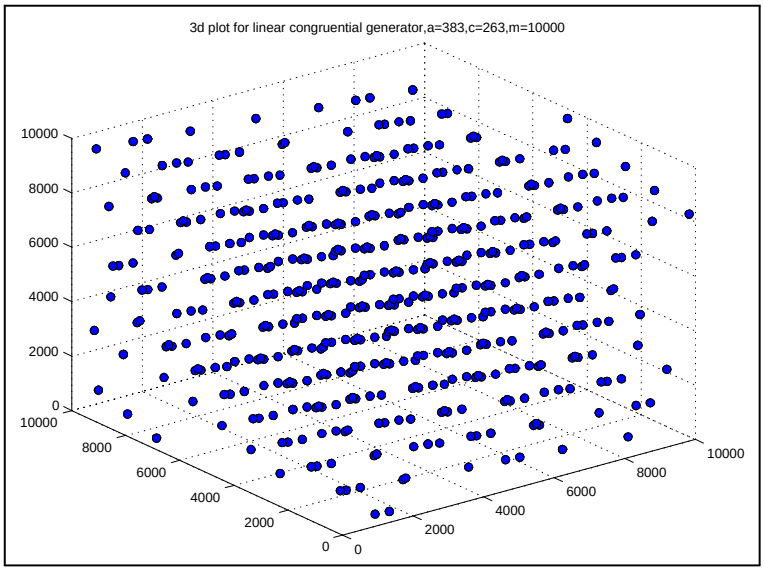


Figure 3.4: The values (x_i, x_{i+1}, x_{i+2}) generated by a linear congruential generator $x_{n+1} = (383x_n + 263)(\text{mod } 10000)$

$10,000$.

Linear planes are evident from some angles in this view, but not from others. In many problems, particularly ones in which random numbers are processed in groups of three or more, this phenomenon can lead to highly misleading results. The spectral test is the most widely used test which attempts to insure against lattice structure. tABLE 3.2 below is taken from Fishman(1996) and gives some values of the spectral test statistic for some linear congruential random number generators in dimension $k \leq 7$.

m	a	c	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$2^{31}-1$	7^5	0	0.34	0.44	0.58	0.74	0.65	0.57
$2^{31}-1$	630360016	0	0.82	0.43	0.78	0.80	0.57	0.68
$2^{31}-1$	742938285	0	0.87	0.86	0.86	0.83	0.83	0.62
2^{31}	65539	0	0.93	0.01	0.06	0.16	0.29	0.45
2^{32}	69069	0	0.46	0.31	0.46	0.55	0.38	0.50
2^{32}	3934873077	0	0.87	0.83	0.83	0.84	0.82	0.72
2^{32}	663608941	0	0.88	0.60	0.80	0.64	0.68	0.61
2^{35}	5^{13}	0	0.47	0.37	0.64	0.61	0.74	0.68
2^{59}	13^{13}	0	0.84	0.73	0.74	0.58	0.64	0.52

TABLE 3.2. Selected Spectral Test Statistics

The unacceptably small values for RANDU in the case $k = 3$ and $k = 4$ are highlighted. On the basis of these values of the spectral test, the multiplicative generators

$$x_{n+1} = 742938285x_n (\text{mod } 2^{31} - 1)$$

$$x_{n+1} = 3934873077x_n (\text{mod } 2^{32})$$

seem to be recommended since their test statistics are all reasonably large for $k = 2, ..., 7$.

Generating Random Numbers from Non-Uniform Continuous Distributions

By far the simplest and most common method for generating non-uniform variates is based on the inverse cumulative distribution function. For arbitrary cumulative distribution function $F(x)$, define $F^{-1}(y) = min\{x; F(x) \geq y\}$. This defines a pseudo-inverse function which is a real inverse (i.e. $F(F^{-1}(y)) =$

$F^{-1}(F(y)) = y$) only in the case that the cumulative distribution function is continuous and strictly increasing. However, in the general case of a possibly discontinuous non-decreasing cumulative distribution function the function continues to enjoy some of the properties of an inverse. Notice that $F^{-1}(F(x)) \leq x$ and $F(F^{-1}(y)) \geq y$ but $F^{-1}(F(F^{-1}(y))) = F^{-1}(y)$ and $F(F^{-1}(F(x))) = F(x)$. In the general case, when this pseudo-inverse is easily obtained, we may use the following to generate a random variable with cumulative distribution function $F(x)$.

Theorem 19 *(inverse transform) If F is an arbitrary cumulative distribution function and U is uniform$[0,1]$ then $X = F^{-1}(U)$ has cumulative distribution function $F(x)$.*

Proof. The proof is a simple consequence of the fact that

$$[U < F(x)] \subset [X \leq x] \subset [U \leq F(x)] \text{ for all } x, \tag{3.5}$$

evident from Figure 3.5. Taking probabilities throughout (3.5), and using the continuity of the distribution of U so that $P[U = F(x)] = 0$, we obtain

$$F(x) \leq P[X \leq x] \leq F(x).$$

■

Examples of Inverse Transform

Exponential (θ)

This distribution, a special case of the gamma distributions, is common in most applications of probability. For example in risk management, it is common to model the time between defaults on a contract as exponential (so the default times follow a Poisson process). In this case the probability density function is

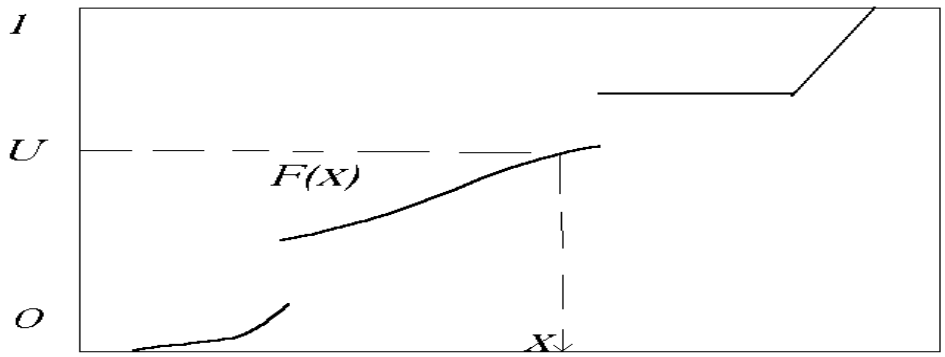


Figure 3.5: The Inverse Transform generator

$f(x) = \frac{1}{\theta} e^{-x/\theta}, x \geq 0$ and $f(x) = 0$ for $x < 0$. The cumulative distribution function is $F(x) = 1 - e^{-x/\theta}, x \geq 0$. Then taking its inverse,

$$X = -\theta \ln(1 - U) \text{ or equivalently}$$

$$X = -\theta \ln U \text{ since } U \text{ and } 1 - U \text{ have the same distribution.}$$

In *Matlab*, the exponential random number generators is called *exprnd* and in *Splus* or *R* it is *rexp.*

Cauchy (a, b)

This distribution is a member of the *stable family* of distributions which we discuss later. It is similar to the normal only substantially more peaked in the center and with more area in the extreme tails of the distribution. The probability density function is

$$f(x) = \frac{b}{\pi(b^2 + (x - a)^2)}, -\infty < x < \infty.$$

See the comparison of the probability density functions in Figure 3.6. Here we have chosen the second (scale) parameter b for the Cauchy so that the two densities would match at the point $x = a = 0$.

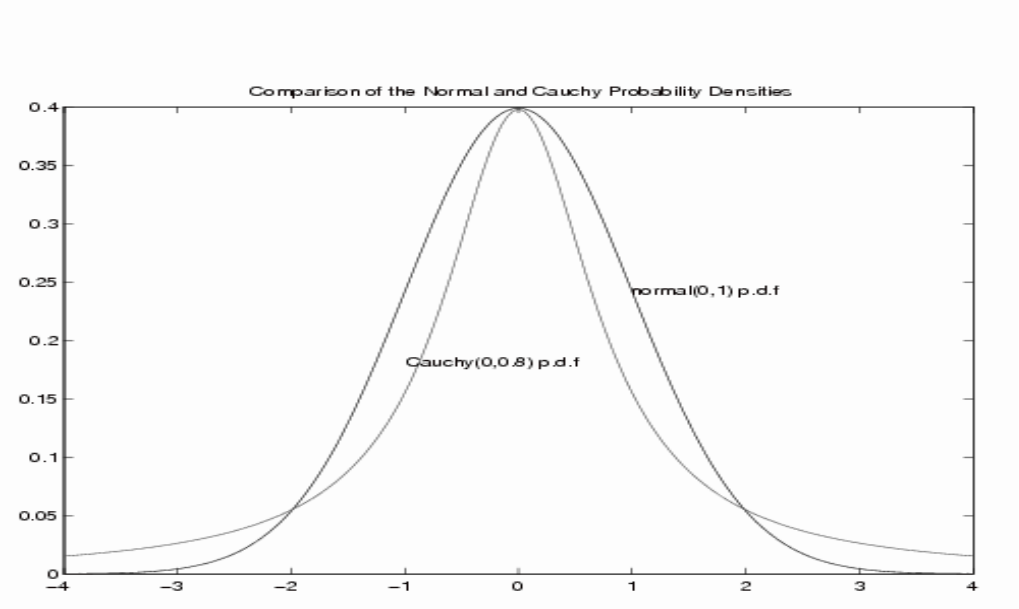


Figure 3.6: The Normal and the Cauchy Probability Density Functions

The cumulative distribution function is $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{x-a}{b})$. Then the inverse transform generator is, for U uniform on [0,1],

$$X = a + b\tan\{\pi(U - \frac{1}{2})\} \text{ or equivalently } X = a + \frac{b}{\tan(\pi U)}$$

where the second expression follows from the fact that $\tan(\pi(x-\frac{1}{2})) = (\tan \pi x)^{-1}$.

Geometric (p)

This is a discrete distribution which describes the number of (independent) trials necessary to achieve a single success when the probability of a success on each trial is p. The probability function is

$$f(x) = p(1-p)^x, x = 1, 2, 3, ...$$

and the cumulative distribution function is

$$F(x) = P[X \leq x] = 1 - (1-p)^{[x]}, x \geq 0$$

where $[x]$ denotes the integer part of x. To invert the cumulative distribution function of a discrete distribution like this one, we need to refer to a graph of the cumulative distribution function analogous to Figure 3.5. We wish to output an integer value of x which satisfies the inequalities

$$F(x-1) < U \leq F(x).$$

Solving these inequalities for **integer** x,we obtain

$$1-(1-p)^{x-1} < U \leq 1-(1-p)^x$$
$$(1-p)^{x-1} > 1-U \geq (1-p)^x$$
$$(x-1)\ln(1-p) > \ln(1-U) \geq x\ln(1-p)$$
$$(x-1) < \frac{\ln(1-U)}{\ln(1-p)} \leq x$$

Note that changes of direction of the inequality occurred each time we multiplied or divided by negative quantity. We should therefore choose the smallest integer for X which is greater than or equal to $\frac{\ln(1-U)}{\ln(1-p)}$ or equivalently,

$$X = 1 + [\frac{log(1-U)}{log(1-p)}] \text{ or} 1 + [\frac{-E}{log(1-p)}]$$

where we write $-\log(1-U) = E$, an exponential(1) random variable. In *Matlab*, the geometric random number generators is called *geornd* and in R or *Splus* it is called *rgeom*.

Pareto (a, b)

This is one of the simpler families of distributions used in econometrics for modeling quantities with lower bound b.

$$F(x) = 1 - (\frac{b}{x})^a, \text{ for } x \geq b > 0.$$

Then the probability density function is

$$f(x) = \frac{ab^a}{x^{a+1}}$$

and the mean is $E(X) =$. The inverse transform in this case results in

$$X = \frac{b}{(1-U)^{1/a}} \quad \text{or} \quad \frac{b}{U^{1/a}}$$

The special case $b = 1$ is often considered in which case the cumulative distribution function takes the form

$$F(x) = 1 - \frac{1}{x^a}$$

and the inverse

$$X = (1-U)^{1/a}.$$

Logistic

This is again a distribution with shape similar to the normal but closer than is the Cauchy. Indeed as can be seen in Figure 3.7, the two densities are almost indistinguishable, except that the logistic is very slightly more peaked in the center and has slightly more weight in the tails. Again in this graph, parameters have been chosen so that the densities match at the center.

The logistic cumulative distribution function is

$$F(x) = \frac{1}{1 + exp\{-(x-a)/b\}}.$$

and on taking its inverse, the logistic generator is

$$X = a + b\ln(U/(1-U)).$$

Extreme Value

This is one of three possible distributions for modelling extreme statistics such as the largest observation in a very large random sample. As a result it is relevant to risk management. The cumulative distribution function is for parameters $-\infty < a < \infty$ and $b > 0$,

$$F(x) = 1 - \exp\{-\exp(\frac{x-a}{b})\}.$$

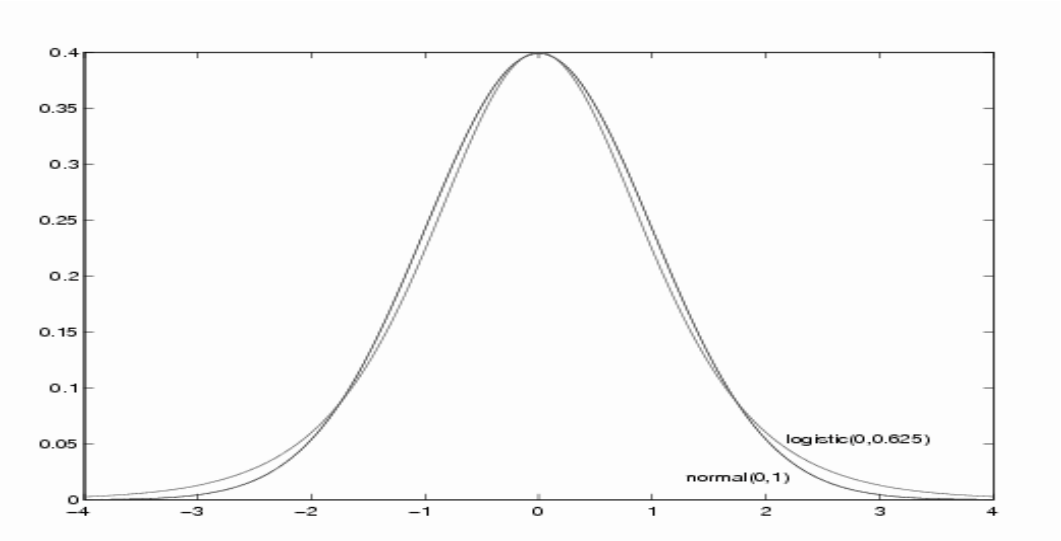


Figure 3.7: Comparison of the Standard Normal and Logistic(0.625) Probability density functions.

The corresponding inverse is

$$X = a + b\ln(\ln(U)).$$

Weibull Distribution

In this case the parameters a, b are both positive and the cumulative distribution function is

$$F(x) = 1 - exp\{-ax^b\} \text{ for } x \geq 0.$$

The corresponding probability density function is

$$f(x) = abx^{b-1}\exp\{-ax^b\}.$$

Then using inverse transform we may generate X as

$$X = \left\{\frac{-\ln(1-U)}{a}\right\}^{1/b}.$$

Student's t.

The Student t distribution is used to construct confidence intervals and tests for the mean of normal populations. It also serves as a wider-tailed alternative to the normal, useful for modelling returns which have moderately large outliers. The probability density function takes the form

$$f(x) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\Gamma(v/2)}(1+\frac{x^2}{v})^{-(v+1)/2}, -\infty < x < \infty.$$

The case $v = 1$ corresponds to the Cauchy distribution. There are specialized methods of generating random variables with the Student t distribution we will return to later. In MATLAB, the student's t generator is called *trnd*. In general, *trnd(v,m,n)* generates an $m \times n$ matrix of student's t random variables having v degrees of freedom.

The generators of certain distributions are as described below. In each case a vector of length n with the associated parameter values is generated.

DISTRIBUTION	R and SPLUS	MATLAB
normal	rnorm(n, μ, σ)	normrnd$(\mu, \sigma, 1, n)$ or randn$(1, n)$ if $\mu = 1, \sigma = 1$
Student's t	rt(n, ν)	trnd$(\nu, 1, n)$
exponential	rexp(n, λ)	exprnd$(\lambda, 1, n)$
uniform	runif(n, a, b)	unifrnd$(a, b, 1, n)$ or rand$(1, n)$ if $a = 0, b = 1$
Weibull	rweibull(n, a, b)	weibrnd$(a, b, 1, n)$
gamma	rgamma(n, a, b)	gamrnd$(a, b, 1, n)$
Cauchy	rcauchy(n, a, b)	a+b*trnd$(1, 1, n)$
binomial	rbinom(n, m, p)	binornd$(m, p, 1, n)$
Poisson	rpois(n, λ)	poissrnd$(\lambda, 1, n)$

TABLE 3.3: Some Random Number Generators in R, SPLUS and MATLAB

Inversion performs reasonably well for any distribution for which *both the*

cumulative distribution function and its inverse can be found in closed form and computed reasonably efficiently. This includes the Weibull, the logistic distribution and most discrete distributions with a small number possible values. However, for other distributions such as the Normal, Student's t, the chi-squared, the Poisson or Binomial with large parameter values, other more specialized methods are usually used, some of which we discuss later.

When the cumulative distribution function is known but not easily inverted, we might attempt to invert it by numerical methods. For example, using the Newton-Ralphson method, we would iterate until convergence the equation

$$X = X - \frac{F(X) - U}{f(X)} \tag{3.6}$$

with $f(X) = F'(X)$, beginning with a good approximation to X. For example we might choose the initial value of $X = X(U)$ by using an easily inverted approximation to the true function $F(X)$. The disadvantage of this approach is that for each X generated, we require an iterative solution to an equation and this is computationally very expensive.

The Acceptance-Rejection Method

Suppose $F(x)$ is a cumulative distribution function and $f(x)$ is the corresponding probability density function. In this case F is continuous and strictly increasing wherever f is positive and so it has a well-defined inverse F^{-1}. Consider the transformation of a point (u, v) in the unit square defined by

$$\begin{aligned} x(u,v) &= F^{-1}(u) \\ y(u,v) &= vf(F^{-1}(u)) = vf(x) \\ &\text{for } 0 < u < 1, \quad 0 < v < 1 \end{aligned}$$

This maps a random point (U, V) uniformly distributed on the unit square into a point (X, Y) uniformly distributed under the graph of the probability density

f. The fact that X has cumulative distribution function F follows from its definition as $X = F^{-1}(U)$ and the inverse transform theorem. By the definition of $Y = Vf(X)$ with V uniform on $[0, 1]$ we see that the conditional distribution of Y given the value of X, is uniform on the interval $[0, f(X)]$. Suppose we seek a random number generator for the distribution of X but we are unable to easily invert the cumulative distribution function We can nevertheless use the result that the point (X, Y) is uniform under the density as the basis for one of the simplest yet most useful methods of generating non-uniform variates, the *rejection* or acceptance-rejection method. It is based on the following very simple relationship governing random points under probability density functions.

Theorem 20 *(Acceptance-Rejection)* (X, Y) *is uniformly distributed in the region between the probability density function* $y = f(x)$ *and the axis* $y = 0$ *if and only if the marginal distribution of* X *has density* $f(x)$ *and the conditional distribution of* Y *given* X *is uniform on* $[0, f(X)]$.

Proof. *If a point* (X, Y) *is uniformly distributed under the graph of* $f(x)$ *notice that the probability* $P[a < X < b]$ *is proportional to the area under the graph between vertical lines at* $x = a$ *and* $x = b$. *In other words* $P[a < X < b]$ *is proportional to* $\int_a^b f(x)dx$. *This implies that* $f(x)$ *is proportional to the probability density function of* X *and provided that* $\int_{-\infty}^{\infty} f(x)dx = 1$, $f(x)$ *is the probability density function of* X. *The converse and the rest of the proof is similar.* ■

Even if the scaling constant for a probability density function is unavailable, in other words if we know $f(x)$ only up to some unknown scale multiple, we can still use Theorem 19 to generate a random variable with probability density f because the X coordinate of a random point uniform under the graph of a constant$\times f(x)$ is the same as that of a random point uniformly distributed under the graph of $f(x)$. The *acceptance-rejection method* works as follows. We wish to generate a random variable from the probability density function $f(x)$.

We need the following ingredients:

- A probability density function $g(x)$ with the properties that

  1. the corresponding cumulative distribution function $G(x) = \int_{-\infty}^{x} g(z)dz$ is easily inverted to obtain $G^{-1}(u)$.

  2. $$\sup\{\frac{f(x)}{g(x)}; -\infty < x < \infty\} < \infty. \tag{3.7}$$

For reasonable efficiency we would like the supremum in (3.7) to be as close as possible to one (it is always greater or equal to one).

The condition (3.7) allows us to find a constant $c > 1$ such that $f(x) \leq cg(x)$ for all x. Suppose we are able to generate a point (X, Y) uniformly distributed under the graph of $cg(x)$. This is easy to do using Theorem 19. Indeed we can define $X = G^{-1}(U)$ and $Y = V \times cg(X)$ where U and V are independent $U[0, 1]$. Can we now find a point (X, Y) which is uniformly distributed under the graph of $f(x)$? Since this is a subset of the original region, this is easy. We simple test the point we have already generated to see if it is in this smaller region and if so we use it. If not start over generating a new pair (X, Y), and repeating this until the condition $Y \leq f(X)$ is eventually satisfied, (see Figure **??**).The simplest version of this algorithm corresponds to the case when $g(x)$ is a uniform density on an interval $[a, b]$. In algorithmic form, the acceptance-rejection method is;

1. Generate a random variables $X = G^{-1}(U)$, where U where U is uniform on $[0, 1]$.

2. Generate independent $V \sim U[0, 1]$

3. If $V \leq \frac{f(X)}{cg(X)}$, then return X and exit

4. ELSE go to step 1.

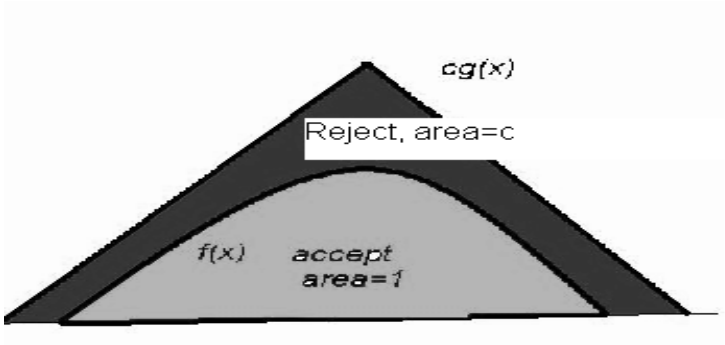


Figure 3.8: The acceptance-Rejection Method

The rejection method is useful if the density g is considerably simpler than f both to evaluate and to generate distributions from and if the constant c is close to 1. The number of iterations through the above loop until we exit at step 3 has a geometric distribution with parameter $p = 1/c$ and mean c so when c is large, the rejection method is not very effective.

Most schemes for generating non-uniform variates are based on a transformation of uniform with or without some rejection step. The rejection algorithm is a special case. Suppose, for example, that $T = (u(x, y), v(x, y))$ is a one-one area-preserving transformation of the region $-\infty < x < \infty, 0 < y < f(x)$ into a subset A of a square in R^2 as is shown in Figure 3.9.

Notice that any such transformation defines a random number generator for the density $f(x)$. We need only generate a point (U, V) uniformly distributed in the set A by acceptance-rejection and then apply the inverse transformation T^{-1} to this point, defining $(X, Y) = T^{-1}(U, V)$. Since the transformation is area-preserving, the point (X, Y) is uniformly distributed under the probability density function $f(x)$ and so the first coordinate X will then have density f. We can think of inversion as a mapping on $[0, 1]$ and acceptance-rejection algorithms

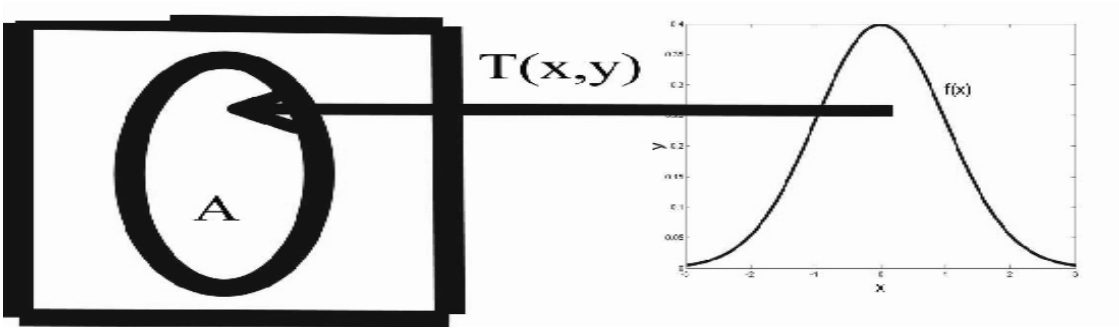


Figure 3.9: $T(x, y)$ is an area Preserving invertible map $f(x, y)$ from the region under the graph of f into the set A, a subset of a rectangle.

as an area preserving mapping on $[0, 1]^2$.

The most common distribution required for simulations in finance and elsewhere is the *normal distribution*. The following theorem provides the simple connections between the normal distribution in Cartesian and in polar coordinates.

Theorem 21 *If* (X, Y) *are independent standard normal variates, then expressed in polar coordinates,*

$$(R, \Theta) = (\sqrt{X^2 + Y^2}, arctan(Y/X)) \tag{3.8}$$

are independent random variables. $R = \sqrt{X^2 + Y^2}$ *has the distribution of the square root of a chi-squared(2) or exponential(2) variable.* $\Theta = \arctan(Y/X))$ *has the uniform distribution on* $[0, 2\pi]$.

It is easy to 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* $[0, 2\pi]$. The proof of this result is left as a problem.

This observation is the basis of two related popular normal pseudo-random number generators. The *Box-Muller* algorithm uses two uniform$[0, 1]$ variates

U, V to generate R and Θ with the above distributions as

$$R = \{-2\ln(U)\}^{1/2}, \Theta = 2\pi V \tag{3.9}$$

and then defines two independent normal(0,1) variates as

$$(X, Y) = R(\cos\Theta, \sin\Theta) \tag{3.10}$$

Note that normal variates must be generated in pairs, which makes simulations involving an even number of normal variates convenient. If an odd number are required, we will generate one more than required and discard one.

Theorem 22 *(Box-Muller Normal Random Number generator)*

Suppose (R, Θ) *are independent random variables such that* R^2 *has an exponential distribution with mean* 2 *and* Θ *has a* Uniform$[0, 2\pi]$ *distribution. Then* $(X, Y) = (R\cos\Theta, R\sin\Theta)$ *is distributed as a pair of independent normal variates.*

Proof. Since R^2 has an exponential distribution, R has probability density function

$$\begin{aligned} f_R(r) &= \frac{d}{dr}P[R \le r] \\ &= \frac{d}{dr}P[R^2 \le r^2] \\ &= \frac{d}{dr}(1 - e^{-r^2/2}) \\ &= re^{-r^2/2}, \quad \text{for } r > 0. \end{aligned}$$

and Θ has probability density function $f_\Theta(\theta) = \frac{1}{2\pi}$ for $0 < \theta < 2\pi$. Since $r = r(x, y) = \sqrt{x^2 + y^2}$ and $\theta(x, y) = \arctan(y/x)$, the Jacobian of the trans-

formation is

$$
\begin{aligned}
|\frac{\partial(r,\theta)}{\partial(x,y)}| &= \begin{vmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} \\
&= \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix} \\
&= \frac{1}{\sqrt{x^2+y^2}}
\end{aligned}
$$

Consequently the joint probability density function of (X, Y) is given by

$$
\begin{aligned}
f_\Theta(\arctan(y/x)) f_R(\sqrt{x^2+y^2})|\frac{\partial(r,\theta)}{\partial(x,y)}| &= \frac{1}{2\pi} \times \sqrt{x^2+y^2} e^{-(x^2+y^2)/2} \times \frac{1}{\sqrt{x^2+y^2}} \\
&= \frac{1}{2\pi} e^{-(x^2+y^2)/2} \\
&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}
\end{aligned}
$$

and this is joint probability density function of two independent standard normal random variables. ■

The tails of the distribution of the pseudo-random numbers produced by the Box-Muller method are quite sensitive to the granularity of the uniform generator. For this reason although the Box-Muller is the simplest normal generator it is not the method of choice in most software. A related alternative algorithm for generating standard normal variates is the *Marsaglia polar* method. This is a modification of the Box-Muller generator designed to avoid the calculation of sin or cos. Here we generate a point $(Z_1, Z_2)$from the uniform distribution on the unit circle by rejection, generating the point initially from the square $-1 \leq z_1 \leq 1, -1 \leq z_2 \leq 1$ and accepting it when it falls in the unit circle or if $z_1^2 + z_2^2 \leq 1$. Now suppose that the points (Z_1, Z_2) is uniformly distributed

inside the unit circle. Then for $r > 0$,

$$P[\sqrt{-2\log(Z_1^2 + Z_2^2)} \leq r] = P[Z_1^2 + Z_2^2 \geq \exp(-r^2/2)]$$
$$\frac{1 - \text{area of a circle of radius } \exp(-r^2/2)}{\text{area of a circle of radius 1}}$$
$$= 1 - e^{-r^2/2}.$$

This is exactly the same cumulative distribution function as that of the random variable R in Theorem 21. It follows that we can replace R^2 by $-2log(Z_1^2+Z_2^2)$. Similarly, if (Z_1, Z_2) is uniformly distributed inside the unit circle then the angle subtended at the origin by a line to the point (X, Y) is random and uniformly$[0, 2\pi]$ distributed and so we can replace $\cos\Theta$, and $\sin\Theta$ by $\frac{Z_1}{\sqrt{Z_1^2+Z_2^2}}$ and $\frac{Z_2}{\sqrt{Z_1^2+Z_2^2}}$ respectively. The following theorem is therefore proved.

Theorem 23 *If the point (Z_1, Z_2) is uniformly distributed in the unit circle $Z_1^2 + Z_2^2 \leq 1$, then the pair of random variables defined by*

$$X = \sqrt{-2log(Z_1^2 + Z_2^2)}\frac{Z_1}{\sqrt{Z_1^2 + Z_2^2}}$$
$$Y = \sqrt{-2log(Z_1^2 + Z_2^2)}\frac{Z_2}{\sqrt{Z_1^2 + Z_2^2}}$$

are independent standard normal variables.

If we use acceptance-rejection to generate uniform random variables Z_1, Z_2 inside the unit circle, the probability that a point generated inside the square falls inside the unit circle is $\pi/4$,so that on average around $4/\pi \approx 1.27$ pairs of uniforms are needed to generate a pair of normal variates.

The speed of the Marsaglia polar algorithm compared to that of the Box-Muller algorithm depends on the relative speeds of generating uniform variates versus the sine and cosine transformations. The Box-Muller and Marsaglia polar method are illustrated in Figure 3.10:

Unfortunately the speed of these normal generators is not the only consideration. If we run a linear congruential generator through a full period we

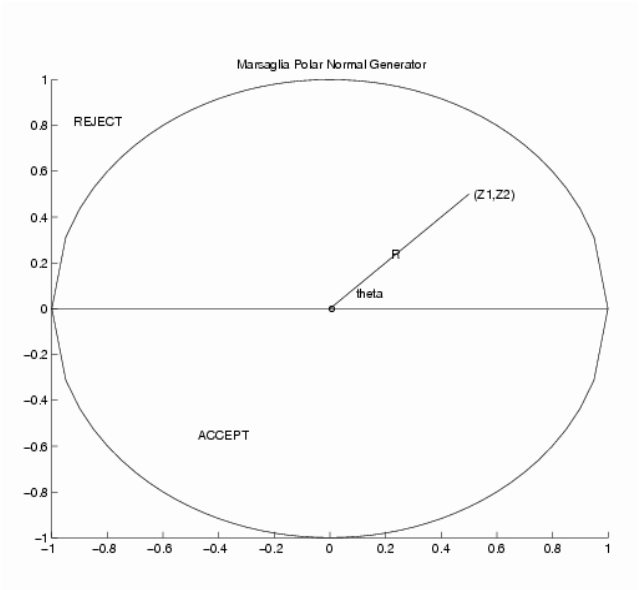


Figure 3.10: Marsaglia's Method for Generating Normal Random Numbers

have seen that the points lie on a lattice, doing a reasonable job of filling the two dimensional rectangle. Transformations like (3.10) are highly non-linear functions of (U, V) stretching the space in some places and compressing it in others. It would not be too surprising if, when we apply this transformation to our points on a lattice, they do not provide the same kind of uniform coverage of the space. In Figure 3.11 we see that the lattice structure in the output from the linear congruential generator results in an interesting but alarmingly non-normal pattern, particularly sparse in the tails of the distribution. Indeed, if we use the full-period generator $x_n = 16807x_{n-1} \bmod (2^{31} - 1)$ the smallest possible value generated for y is around -4.476 although in theory there should be around 8,000 normal variates generated below this.

The normal random number generator in *Matlab* is called *normrnd* or for standard normal *randn*. For example $normrnd(\mu, \sigma, m, n)$ generates a matrix of $m \times n$ pseudo-independent normal variates with mean μ and standard devia-

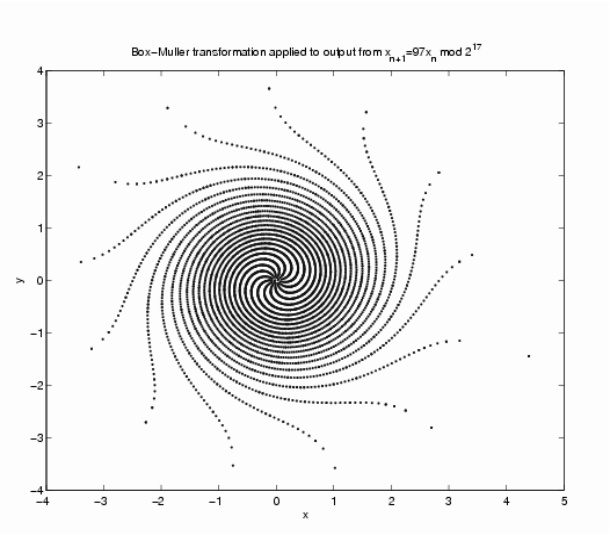


Figure 3.11: Box Muller transformation applied to the output to $x_n = 97x_{n-1} \bmod 2^{17}$

tion σ and *rand(m,n)* generates an $m \times n$ matrix of standard normal random numbers. A more precise algorithm is to use inverse transform and a highly refined rational approximation to normal inverse cumulative distribution function available from P.J. Acklam (2003). The Matlab implementation of this inverse c.d.f. is called *ltqnorm* after application of a refinement, achieves full machine precision. In R or Splus, the normal random number generator is called *rnorm.* The inverse random number function in Excel has been problematic in many versions. These problems appear to have been largely corrected in *Excel 2002*, although there is still significant error (roughly in the third decimal) in the estimation of lower and upper tail quantiles. The following table provides a comparison of the *normsinv* function in Excel and the Matlab inverse normal *norminv.* The "exact" values agree with the values generated by Matlab *norminv* to the number of decimals shown.

p	Excel 2002	Exact
10^{-1}	-1.281551939	-1.281551566
10^{-2}	-2.326347	-2.326347874
10^{-3}	-3.090252582	-3.090232306
10^{-4}	-3.719090272	-3.719016485
10^{-5}	-4.265043367	-4.264890794
10^{-6}	-4.753672555	-4.753424309
10^{-7}	-5.199691841	-5.199337582
10^{-8}	-5.612467211	-5.612001244
10^{-9}	-5.998387182	-5.997807015
10^{-10}	-6.362035677	-6.361340902

The Lognormal Distribution

If Z is a normal random variable with mean μ and variance σ^2, then we say that the distribution of $X = e^Z$ is lognormal with mean $E(X) = \eta = \exp\{\mu+\sigma^2/2\} > 0$ and parameter $\sigma > 0$. Because a lognormal random variable is obtained by *exponentiating* a normal random variable it is strictly positive, making it a reasonable candidate for modelling quantities such as stock prices, exchange rates, lifetimes, though in a fools paradise in which stock prices and lifetimes are never zero. To determine the lognormal probability density function, notice that

$$\begin{aligned} P(X \le x] &= P[e^Z \le x] \\ &= P[Z \le \ln(x)] \\ &= \Phi(\frac{\ln(x) - \mu}{\sigma}) \quad \text{with } \Phi \text{ the standard normal c.d.f.} \end{aligned}$$

and differentiating to obtain the probability density function $g(x|\eta,\sigma)$ of X, we obtain

$$\begin{aligned} g(x|\eta,\sigma) &= \frac{d}{dx}\Phi(\frac{\ln(x)-\mu}{\sigma}) \\ &= \frac{1}{x\sigma\sqrt{2\pi}}\exp\{-(\ln(x)-\mu)^2/2\sigma^2\} \\ &= \frac{1}{x\sigma\sqrt{2\pi}}\exp\{-(\ln(x)-\ln(\eta)+\sigma^2/2)^2/2\sigma^2\} \end{aligned}$$

A random variable with a lognormal distribution is easily generated by generating an appropriate normal random variable Z and then exponentiating. We may use either the parameter μ, the mean of the random variable Z in the exponent or the parameter η, the expected value of the lognormal. The relationship is not as simple as a naive first impression might indicate since

$$E(e^Z) \neq e^{E(Z)}.$$

Now is a good time to accommodate to this correction factor of $\sigma^2/2$ in the exponent

$$\eta = E(e^Z) = e^{E(Z)+\sigma^2/2} = e^{\mu+\sigma^2/2} \text{ or,}$$
$$E(e^{Z-\mu-\sigma^2/2}) = 1$$

since a similar factor appears throughout the study of stochastic integrals and mathematical finance. Since the lognormal distribution is the one most often used in models of stock prices, it is worth here recording some of its conditional moments used in the valuation of options. In particular if X has a lognormal distribution with mean $\eta = e^{\mu+\sigma^2/2}$ and volatility parameter σ, then for any p

and $l > 0$,

$$\begin{aligned}
E[X^p I(X > l)] &= \frac{1}{\sigma\sqrt{2\pi}} \int_l^\infty x^{p-1} \exp\{-(\ln(x) - \mu)^2/2\sigma^2\} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{\ln(l)}^\infty e^{zp} \exp\{-(z - \mu)^2/2\sigma^2\} dz \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{p\mu + p^2\sigma^2/2} \int_{\ln(l)}^\infty \exp\{-(z - \xi)^2/2\sigma^2\} dz \quad \text{where } \xi = \mu + \sigma^2 p \\
&= e^{p\mu + p^2\sigma^2/2} \Phi(\frac{\xi - \ln(l)}{\sigma}) \\
&= \eta^p \exp\{-\frac{\sigma^2}{2} p(1-p)\} \Phi(\sigma^{-1} \ln(\eta/l) + \sigma(p - \frac{1}{2})) \qquad (3.11)
\end{aligned}$$

where Φ is the standard normal cumulative distribution function.

Application: A Discrete Time Black-Scholes Model

Suppose that a stock price $S_t, t = 1, 2, 3, ...$ is generated from an independent sequence of returns Z_1, Z_2 over non-overlapping time intervals. If the value of the stock at the end of day $t = 0$ is S_0, and the return on day 1 is Z_1 then the value of the stock at the end of day 1 is $S_1 = S_0 e^{Z_1}$. There is some justice in the use of the term "return" for Z_1 since for small values Z_1, $S_0 e^{Z_1} \simeq S_0(1 + Z_1)$ and so Z_1 is roughly $\frac{S_1 - S_0}{S_1}$. Assume similarly that the stock at the end of day i has value $S_i = S_{i-1} \exp(Z_i)$. In general for a total of j such periods (suppose there are n such periods in a year) we assume that $S_j = S_0 \exp\{\sum_{i=1}^j Z_i\}$ for independent random variables Z_i all have the same normal distribution. Note that in this model the returns over non-overlapping independent periods of time are independent. Denote $var(Z_i) = \sigma^2/N$ so that

$$var(\sum_{i=1}^N Z_i) = \sigma^2$$

represents the squared annual volatility parameter of the stock returns. Assume that the annual interest rate on a risk-free bond is r so that the interest rate per period is r/N.

Recall that the risk-neutral measure Q is a measure under which the stock price, discounted to the present, forms a martingale. In general there may be many such measures but in this case there is only one under which the stock price process has a similar lognormal representation $S_j = S_0 \exp\{\sum_{i=1}^{j} Z_i\}$ for independent normal random variables Z_i. Of course under the risk neutral measure, the normal random variables Z_i may have a different mean. Full justification of this model and the uniqueness of the risk-neutral distribution really relies on the continuous time version of the Black Scholes described in Section 2.6. Note that if the process

$$e^{-rt/N} S_j = S_0 \exp\{\sum_{i=1}^{j} (Z_i - \frac{r}{N})\}$$

is to form a martingale under Q, it is necessary that

$$\begin{aligned} E_Q[S_{j+1}|H_t] &= S_j \quad \text{or} \\ E_Q[S_j \exp\{Z_{j+1} - \frac{r}{N}\}|H_j] &= S_j E_Q[\exp\{Z_{j+1} - \frac{r}{N}\}] \\ &= S_j \end{aligned}$$

and so $\exp\{Z_{j+1} - \frac{r}{N}\}$ must have a lognormal distribution with expected value 1. Recall that, from the properties of the lognormal distribution,

$$E_Q[\exp\{Z_{t+1} - \frac{r}{N}\}] = \exp\{E_Q(Z_{t+1}) - \frac{r}{N} + \frac{\sigma^2}{2N}\}$$

since $var_Q(Z_{t+1}) = \frac{\sigma^2}{N}$. In other words, for each i the expected value of Z_i is, under Q, equal to $\frac{r}{N} - \frac{\sigma^2}{2N}$. So under Q, S_j has a lognormal distribution with mean

$$S_0 e^{rj/N}$$

and volatility parameter $\sigma\sqrt{j/N}$.

Rather than use the Black-Scholes formula of Section 2.6, we could price a call option with maturity $j = NT$ periods from now by generating the random path $S_i, i = 1, 2, ...j$ using the lognormal distribution for S_j and then averaging

the returns discounted to the present. The value at time $j = 0$ of a call option with exercise price K is an average of simulated values of

$$e^{-rj/N}(S_j - K)^+ = e^{-rj/N}(S_0 \exp\{\sum_{i=1}^{T} Z_i\} - K)^+,$$

with the simulations conducted under the risk-neutral measure Q with initial stock price the current price S_0. Thus the random variables Z_i are independent $N(\frac{r}{N} - \frac{\sigma^2}{2N}, \frac{\sigma^2}{N})$. The following *Matlab* function simulates the stock price over the whole period until maturity and then values a European call option on the stock by averaging the discounted returns.

Example 24 *(simulating the return from a call option)*

Consider simulating a call option on a stock whose current value is $S_0 =$ \$1.00. The option expires in j days and the strike price is $K =$ \$1.00. We assume constant spot (annual) interest rate r and the stock price follows a lognormal distribution with annual volatility parameter σ. The following Matlab function provides a simple simulation and graph of the path of the stock over the life of the option and then outputs the discounted payoff from the option.

function z=plotlogn(r,sigma,T, K)

% outputs the discounted simulated return on expiry of a call option (per dollar pv of stock).

% Expiry =T years from now, $(T = j/N)$

% current stock price=\$1. $(= S_0)$, r = annual spot interest rate, sigma=annual volatility $(=\sigma)$,

% K= strike price.

N=250 ; % N is the assumed number of business days in a year.

j=N*T; % the number of days to expiry

s = sigma/sqrt(N); % s is volatility per period

```
mn = r/N - s^2/2;                % mn= mean of the normal increments per period
y=exp(cumsum(normrnd(mn,s,j,1)));
y=[1 y'];                        % the value of the stock at times 0,...,
x = (0:j)/N;                     % the time points i
plot(x,y,'-',x,K*ones(1,j+1),'y')
xlabel('time (in years)')
ylabel('value of stock')
title('SIMULATED RETURN FROM CALL OPTION')
z = exp(-r*T)*max(y(j+1)-K, 0);  % payoff from option  discounted to
present
```

Figure 3.12 resulted from one simulation run with $r = .05, j = 63$ (about 3 months), $\sigma = .20, K = 1$.

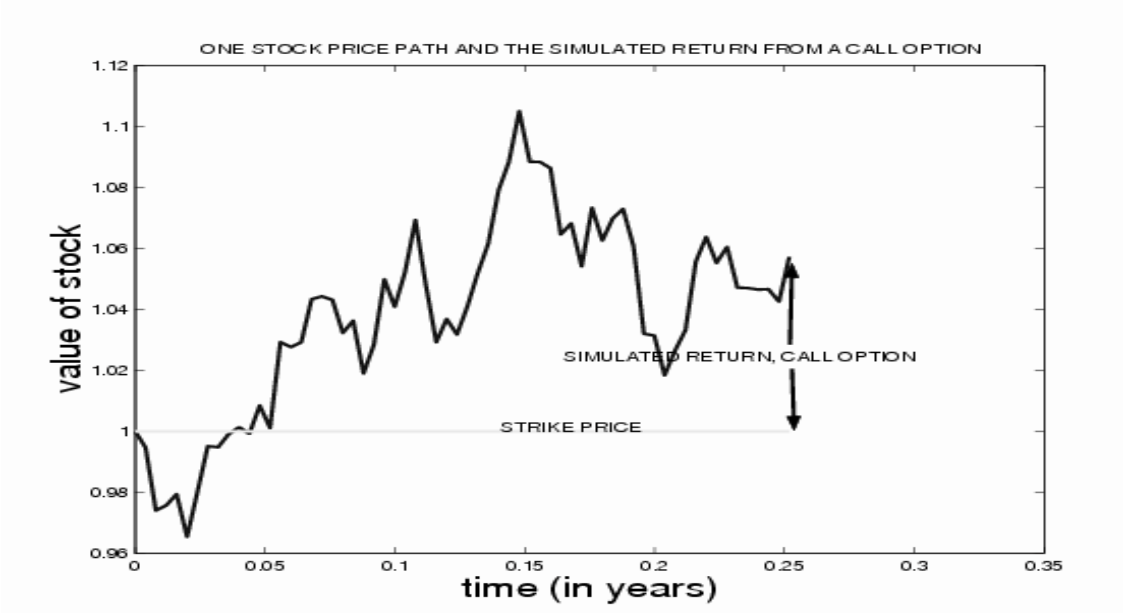


Figure 3.12: One simulation of the return from a call option with strike price $1.00

The return on this run was the discounted difference between the terminal value of the stock and the strike price or around 0.113. We may repeat this many times, averaging the discounted returns to estimate the present value of the option.

For example to value an at the money call option with exercise price=the initial price of the stock=\$1, 5% annual interest rate, 20% annual volatility and maturity 0.25 years from the present, we ran this function 100 times and averaged the returns to estimate the option price as *0.044978.* If we repeat the identical statement, the output is different, for example *option val= 0.049117* because each is an average obtained from only 100 simulations. Averaging over more simulations would result in greater precision, but this function is not written with computational efficiency in mind. We will provide more efficient simulations for this problem later. For the moment we can compare the price of this option as determined by simulation with the exact price according to the Black-Scholes formula. This formula was developed in Section 2.6. The price of a call option at time $t = 0$ given by

$$V(S_T, T) = S_T \Phi(d_1) - K e^{-rT/N} \Phi(d_2)$$

where

$$d_1 = \frac{\log(S_T/K) + (r + \frac{\sigma^2}{2})T/N}{\sigma\sqrt{T/N}} \quad \text{and} \quad d_2 = \frac{\log(S_T/K) + (r - \frac{\sigma^2}{2})T/N}{\sigma\sqrt{T/N}}$$

and the Matlab function which evaluates this is the function *blsprice* which gives, in this example, and exact price on entering *[CALL,PUT] =BLSPRICE(1,1,.05,63/250,.2,0)* which returns the value *CALL=0.0464.* With these parameters, 4.6 cents on the dollar allows us to lock in any anticipated profit on the price of a stock (or commodity if the lognormal model fits) for a period of about three months. The fact that this can be done cheaply and with ease is part of the explanation for the popularity of derivatives as tools for hedging.

Algorithms for Generating the Gamma and Beta Distributions

We turn now to algorithms for generating the *Gamma distribution* with density

$$f(x|a,b) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, \text{ for } x > 0, a > 0, b > 0. \tag{3.12}$$

The exponential distribution ($a = 1$) and the chi-squared (corresponding to $a = \nu/2, b = 2$, for ν integer) are special cases of the Gamma distribution. The gamma family of distributions permits a wide variety of shapes of density functions and is a reasonable alternative to the lognormal model for positive quantities such as asset prices. In fact for certain parameter values the gamma density function is very close to the lognormal. Consider for example a typical lognormal random variable with mean $\eta = 1.1$ and volatility $\sigma = 0.40$.

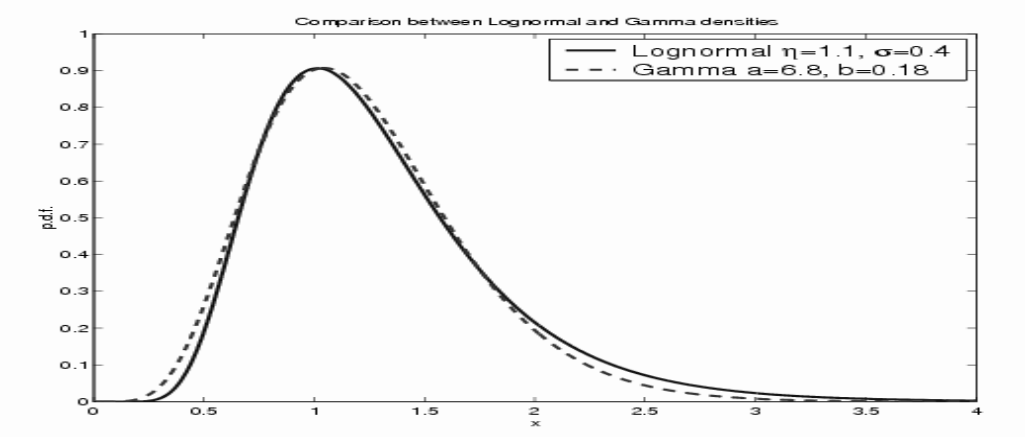


Figure 3.13: Comparison between the Lognormal and the Gamma densities

The probability density functions can be quite close as in Figure 3.13. Of course the lognormal, unlike the gamma distribution, has the additional attractive feature that a product of independent lognormal random variables also has a lognormal distribution.

Another common distribution closely related to the gamma is the *Beta distribution* with probability density function defined for parameters $a, b > 0$,

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, 0 \leq x \leq 1. \tag{3.13}$$

The beta density obtains for example as the distribution of order statistics in a sample from independent uniform $[0,1]$ variates. This is easy to see. For example if $U_1, ..., U_n$ are independent uniform random variables on the interval $[0,1]$ and if $U_{(k)}$ denotes the $k'th$ largest of these n values, then

$$\begin{aligned} P[U_{(k)} < x] &= P[\text{there are } k \text{ or more values less than } x] \\ &= \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}. \end{aligned}$$

Differentiating we find the probability density function of $U_{(k)}$ to be

$$\begin{aligned} \frac{d}{dx} \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j} &= \sum_{j=k}^{n} \binom{n}{j} \{j x^{j-1}(1-x)^{n-j} + (n-j)x^j(1-x)^{n-j-1}\} \\ &= k\binom{n}{k} x^{k-1}(1-x)^{n-k} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{n-k} \end{aligned}$$

and this is the beta density with parameters $a = k-1$, $b = n-k+1$. Order statistics from a Uniform sample therefore have a beta distribution with the k'th order statistic having the Beta$(k-1, n-k+1)$ distribution. This means that order statistics from more general continuous distributions can be easily generated using the inverse transform and a beta random variable. For example suppose we wish to simulate the largest observation in a normal(μ, σ^2) sample of size 100. Rather than generate a sample of 100 normal observations and take the largest, we can simulate the value of the largest uniform order statistic $U_{(100)} \sim Beta(99, 1)$ and then $\mu + \sigma\Phi^{-1}(U_{(100)})$ (with Φ^{-1} the standard normal inverse cumulative distribution function) is the required simulated value. This may be used to render simulations connected with risk management more efficient.

The following result lists some important relationships between the Gamma and Beta distributions. For example it allows us to generate a Beta random variable from two independent Gamma random variables.

Theorem 25 *(Gamma distribution) If X_1, X_2 are independent Gamma (a_1, b) and Gamma (a_2, b) random variables, then $Z = \frac{X_1}{X_1+X_2}$ and $Y = X_1 + X_2$ are independent random variables with Beta (a_1, a_2) and Gamma $(a_1 + a_2, b)$ distributions respectively. Conversely, if (Z, Y) are independent variates with Beta (a_1, a_2) and the Gamma $(a_1 + a_2, b)$ distributions respectively, then $X_1 = YZ$, and $X_2 = Y(1 - Z)$ are independent and have the Gamma (a_1, b) and Gamma (a_2, b) distributions respectively.*

Proof. *Assume that X_1, X_2 are independent Gamma (a_1, b) and Gamma (a_2, b) variates. Then their joint probability density function is*

$$f_{X_1X_2}\ (x_1, x_2) = \frac{1}{\Gamma(a_1)\Gamma(a_2)} x_1^{a_1-1} x_2^{a_2-1} e^{-(x_1+x_2)/b},\ \text{for } x_1 > 0, x_2 > 0.$$

Consider the change of variables $x_1(z, y) = zy, x_2(z, y) = (1 - z)y$. Then the Jacobian of this transformation is given by

$$\begin{vmatrix} \frac{\partial x_1}{\partial z} & \frac{\partial x_1}{\partial y} \\ \frac{\partial x_2}{\partial z} & \frac{\partial x_2}{\partial y} \end{vmatrix} = \begin{vmatrix} y & z \\ -y & 1-z \end{vmatrix}$$

$$= y.$$

Therefore the joint probability density function of (z, y) is given by

$$f_{z,y}(z, y) = f_{X_1X_2}\ (zy, (1 - z)y) \begin{vmatrix} \frac{\partial x_1}{\partial z} & \frac{\partial x_1}{\partial y} \\ \frac{\partial x_2}{\partial z} & \frac{\partial x_2}{\partial y} \end{vmatrix}$$

$$= \frac{1}{\Gamma(a_1)\Gamma(a_2)} z^{a_1-1}(1 - z)^{a_2-1} y^{a_1+a_2-1} e^{-y/b}, \text{for } 0 < z < 1, y > 0$$

$$= \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} z^{a_1-1}(1 - z)^{a_2-1} \times \frac{1}{\Gamma(a_1 + a_2)} y^{a_1+a_2-1} e^{-y/b}, \text{for } 0 < z < 1, y > 0$$

and this is the product of two probability density functions, the Beta(a_1, a_2) density for Z and the Gamma$(a_1 + a_2, b)$ probability density function for Y. The converse holds similarly. ■

This result is a basis for generating gamma variates with integer value of the parameter a (sometimes referred to as the shape parameter). According to the theorem, if a is integer and we sum a independent Gamma(1,b) random variables the resultant sum has a Gamma(a, b) distribution. Notice that $-b\log(U_i)$ for uniform$[0, 1]$ random variable U_i is an exponential or a Gamma$(1, b)$ random variable. Thus $-b\log(\prod_{i=1}^{n} U_i)$ generates a gamma (n, b) variate for independent uniform U_i. The computation required for this algorithm, however, increases linearly in the parameter $a = n$, and therefore alternatives are required, especially for large a. Observe that the *scale parameter* b is easily handled in general: simply generate a random variable with scale parameter 1 and then multiply by b. Most algorithms below, therefore, are only indicated for $b = 1$.

For large a Cheng (1977) uses acceptance-rejection from a density of the form

$$g(x) = \lambda\mu\frac{x^{\lambda-1}}{(\mu + x^{\lambda})^2}dx \quad , x > 0 \tag{3.14}$$

called the *Burr XII distribution.* The two parameters μ and λ of this density (μ is not the mean) are chosen so that it is as close as possible to the gamma distribution. We can generate a random variable from (3.14) by inverse transform as $G^{-1}(U) = \{\frac{\mu U}{1-U}\}^{1/\lambda}$.

A much simpler function for dominating the gamma densities is a minor extension of that proposed by Ahrens and Dieter (1974). It corresponds to using as a dominating probability density function

$$g(x) = \begin{cases} \frac{x^{a-1}}{k^{a-1}(\frac{k}{a}+\exp(-k))} & 0 \le x \le k \\ \frac{k^{a-1}e^{-x}}{k^{a-1}(\frac{k}{a}+\exp(-k))} & x > k \end{cases} \quad , x > k \tag{3.15}$$

Other distributions that have been used as dominating functions for the Gamma are the Cauchy (Ahrens and Dieter), the Laplace (Tadakamalla), the exponential (Fishman), the Weibull, the relocated and scaled t distribution with 2 degrees of freedom (Best), a combination of normal density (left part) and exponential density (right part) (Ahrens and Dieter), and a mixture of two

Erlang distributions (Gamma with integral shape parameter α).

Best's algorithm generates a Student's t_2 variate as

$$Y = \frac{\sqrt{2}(U - 1/2)}{\sqrt{U(1-U}} \tag{3.16}$$

where $U \sim U[0,1]$. Then Y has the Students t distribution with 2 degrees of freedom having probability density function

$$g(y) = \frac{1}{(2+y^2)^{3/2}}. \tag{3.17}$$

We then generate a random variable $X = (a-1) + Y\sqrt{3a/2 - 3/8}$ and apply a rejection step to X to produce a Gamma random variable. See Devroye (p. 408) for details.

Most of the above algorithms are reasonably efficient only for $a > 1$ with the one main exception being the combination of power of x and exponential density suggested by Ahrens and Dieter above. Cheng and Feast (1979) also suggest a ratio of uniforms algorithm for the gamma distribution, $a > 1$.

A final fast and simple procedure for generating a gamma variate with $a > 1$ is due to Marsaglia and Tsang (2000) and generates a gamma variate as the cube of a suitably scaled normal. Given a fast generator of the Normal to machine precision, this is a highly efficient rejection technique. We put $d = a - \frac{1}{3}$ and generate a standard normal random variable X and a uniform variate U until, with $V = (1 + \frac{X}{\sqrt{9d}})^3$, the following inequality holds:

$$\ln(U) < \frac{X^2}{2} + d - dV + d\ln(V).$$

When this inequality is satisfied, we accept the value $d \times V$ as obtained from the Gamma$(a, 1)$ distribution. As usual multiplication by b results in a Gamma(a, b) random variable. The efficiency of this algorithm appears to be very high (above 96% for $a > 1$).

In the case $0 < a < 1$, Stuart's theorem below allows us to modify a Gamma variate with $a > 1$ to one with $a < 1$. We leave the proof of the theorem as an exercise.

Theorem 26 *(Stuart) Suppose U is uniform $[0,1]$ and X is Gamma $(a+1,1)$ independent of U. Then $XU^{1/a}$ has a gamma $(a,1)$ distribution*

The Matlab function *gamrnd* uses Best's algorithm and acceptance rejection for $\alpha > 1$. For $\alpha < 1$, it uses Johnk's generator, which is based on the following theorem.

Theorem 27 *(Johnk) Let U and V be independent Uniform[0,1] random variables. Then the conditional distribution of*

$$X = \frac{U^{1/\alpha}}{U^{1/\alpha} + V^{1/(1-\alpha)}}$$

given that the denominator $U^{1/\alpha} + V^{1/(1-\alpha)} < 1$ is Beta$(\alpha, 1-\alpha)$.

Multiplying this beta random variable by an independent exponential (1) results in a Gamma$(\alpha, 1)$ random variable.

Toward generating the *beta distribution,* use of Theorem 24 and the variable $Z = \frac{X_1}{X_1+X_2}$ with X_1, X_2 independent gamma variates is one method of using a gamma generator to produce beta variates, and this is highly competitive as long as the gamma generator is reasonably fast. The MATLAB generator is *betarnd(a,b,1,n)* Alternatives are, as with the gamma density, rejection from a Burr XII density (Cheng, 1978) and use of the following theorem as a generator (due to Johnk). This a more general version of the theorem above.

Theorem 28 *(Beta distribution)*

Suppose U, V are independent uniform$[0,1]$ variates. Then the conditional distribution of

$$X = \frac{U^{1/a}}{U^{1/a} + V^{1/b}} \tag{3.18}$$

given that $U^{1/a} + V^{1/b} \leq 1$ is Beta (a,b). Similarly the conditional distribution of $U^{1/a}$ given that $U^{1/a} + V^{1/b} \leq 1$ is Beta $(a+1,b)$.

Proof. *Define a change of variables*

$$X = \frac{U^{1/a}}{U^{1/a} + V^{1/b}}, Y = U^{1/a} + V^{1/b}$$

$$\text{or } U = (YX)^a \text{ and } V = [(1-X)Y]^b$$

so that the joint probability density function of (X, Y) *is given by*

$$f_{X,Y}(x,y) = f_{U,V}((yx)^a, [(1-x)y]^b) \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix}$$

$$= aby^{a+b-1}x^{a-1}(1-x)^{b-1}$$

provided either $(0 < x < 1$ and $y < 1)$ or $(1 - \frac{1}{y} < x < \frac{1}{y}$ and $1 < y < 2)$. *Notice that in the case* $y < 1$, *the range of values of* x *is the unit interval and does not depend on* y *and so the conditional probability density function of* X *given* $Y = y$ *is a constant times* $x^{a-1}(1-x)^{b-1}$, *i.e. is the Beta*(a, b) *probability density function. The rest of the proof is similar.* ■

A generator exploiting this theorem produces pairs (U, V) until the condition is satisfied and then transforms to the variable X. However, the probability that the condition is satisfied is $\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+1)}$ which is close to 0 unless a, b are small, so this procedure should be used only for small values of both parameters. Theorems 24 and 25 together provide an algorithm for generating Gamma variates with non-integral a from variates with integral ones. For example if X is Gamma$(4, 1)$and Z is independent Beta $(3.4, .6)$then XZ is Gamma $(3.4, 1)$.

There are various other continuous distributions commonly associated with statistical problems. For example the *Student's t-distribution* with ν degrees of freedom is defined as a ratio $\sqrt{\frac{2\nu}{X}}Z$ where Z is standard normal and X is gamma $(\frac{\nu}{2}, 2)$. Alternatively, we may use $\sqrt{\nu}\frac{X-1/2}{\sqrt{X(1-X)}}$where X is generated as a symmetric beta$(\nu/2, \nu/2)$ variate.

Example 29 *(some alternatives to lognormal distribution)*

The assumption that stock prices, interest rates, or exchange rates follow a lognormal distribution is a common exercise in wishful thinking. The lognormal

distribution provides a crude approximation to many financial time series, but other less theoretically convenient families of distributions sometimes provide a better approximation. There are many possible alternatives, including the students t distribution and the stable family of distributions discussed later. Suppose, for the present, we modify the usual normal assumption for stock returns slightly by assuming that the log of the stock price has a distribution "close" to the normal but with somewhat more weight in the tails of the distribution. Specifically assume that under the Q measure, $S_T = S_0 \exp\{\mu + cX\}$ where X has cumulative distribution function $F(x)$. Some constraint is to be placed on the constant c if we are to compare the resulting prices with the Black-Scholes model and it is natural to require that both models have identical volatility, or identical variance of returns. Since the variance of the return in the Black Scholes model over a period of length T is $\sigma^2 T$ where σ is the annual volatility, we therefore require that

$$var(cX) = \sigma^2 T \text{ or } \quad c = \sqrt{\frac{\sigma^2 T}{var(X)}}.$$

The remaining constraint is required of all option pricing measures is the martingale constraint and this implies that the discounted asset price is a martingale, and in consequence

$$e^{-rT} E_Q S_T = S_0. \tag{3.19}$$

Letting the moment generating function of X be

$$m(s) = Ee^{sX},$$

the constraint (3.19) becomes

$$e^{\mu - rT} m(c) = 1$$

and solving for μ, we obtain

$$\mu = rT - \ln(m(c)).$$

Provided that we can generate from the cumulative distribution function of X, the price of a call option with strike price K under this returns distribution can be estimated from N simulations by the average discounted return from N options,

$$\begin{aligned} e^{-rT}\frac{1}{N}\sum_{i=1}^{N}(S_{Ti}-K)^{+} &= e^{-rT}\frac{1}{N}\sum_{i=1}^{N}(S_0e^{\mu+cX_i}-K)^{+} \\ &= e^{-rT}\frac{1}{N}\sum_{i=1}^{N}(S_0e^{rT-\ln(m(c))+cX_i}-K)^{+} \\ &= \frac{1}{N}\sum_{i=1}^{N}(S_0\frac{e^{cX_i}}{m(c)}-e^{-rT}K)^{+} \end{aligned}$$

A more precise calculation is the difference between the option price in this case and the comparable case of normally distributed returns. Suppose we use inverse transform together with a uniform[0,1] variate to generate both the random variable $X_i = F^{-1}(U_i)$ and the corresponding normal return $Z_i = rT + \sigma\sqrt{T}\Phi^{-1}(U_i)$. Then the difference is estimated by

$$\begin{aligned} &\text{option price under } F - \text{option price under } \Phi \\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\{(S_0\frac{e^{cF^{-1}(U_i)}}{m(c)}-e^{-rT}K)^{+}-(S_0e^{\sigma\sqrt{T}\Phi^{-1}(U_i)-\sigma^2T/2}-e^{-rT}K)^{+}\} \end{aligned}$$

If necessary, in case the moment generating function of X is unknown, we can estimate it and the variance of X using sample analogues over a large number N of simulations. In this case c is estimated by

$$\sqrt{\frac{\sigma^2T}{\widehat{var}(X)}}$$

with $\widehat{var}$ representing the sample variance and $m(c)$ estimated by

$$\frac{1}{N}\sum_{i=1}^{N}e^{cF^{-1}(U_i)}.$$

To consider a specific example, the logistic$(0, 0.522)$ distribution is close to the normal, except with slightly more weight in the tails. The scale parameter in

this case was chosen so that the logistic has approximate unit variance. The cumulative distribution function is $F(x) = \frac{1}{1+exp\{-x/b\}}$ and its inverse is $X = b\ln(U/(1-U))$. The moment generating function is $m(s) = \Gamma(1-bs)\Gamma(1+bs)$, $s < 1/b$. The following function was used to compare the price of a call option when stock returns have the logistic distribution(i.e. stock prices have the "loglogistic" distribution) with the prices in the Black-Scholes model.

```
function [re,op1,opbs]=diffoptionprice(n,So,strike,r,sigma,T)
%estimates the relative error in the BS option price and price under
% logistic returns distribution . Runs n simulations.
u=rand(1,n);
x=log(u./(1-u));                                    % generates standard
logistic*
z=sigma*sqrt(T)*norminv(u)-sigma^2*T/2;
c=sigma*sqrt(T/var(x));
mc=mean(exp(c*x));
re=[]; op1=[]; opbs=[];
for i=1:length(strike)
op1=[op1 mean(max(exp(c*x)*So/mc-exp(-r*T)*strike(i),0))];  % price under F
opbs=[opbs mean(max(So*exp(z)-exp(-r*T)*strike(i),0))];      % price under BS
end
dif=op1-opbs;
re=[re dif./(dif+BLSPRICE(So,strike,r,T,sigma,0))];
plot(strike/So,re)
xlabel('Strike price/initial price')
ylabel('relative error in Black Scholes formula')
```

The relative error in the Black-Scholes formula obtained from a simulation of 100,000 is graphed in Figure 3.14. The logistic distribution differs only slightly

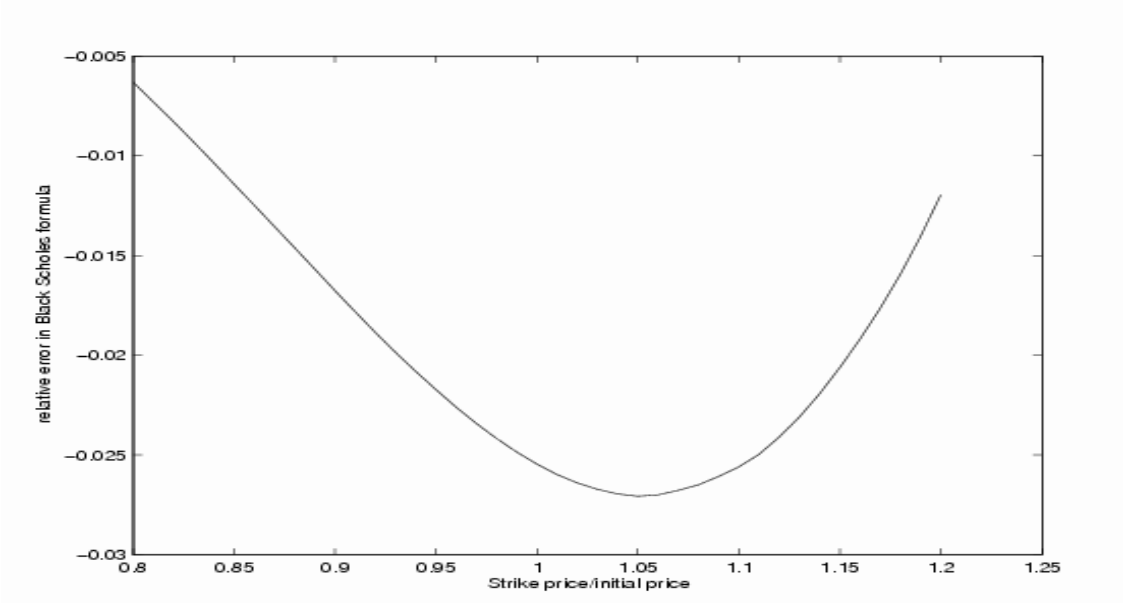


Figure 3.14: Relative Error in Black-Scholes price when asset prices are loglogistic, $\sigma = .4$, $T = .75$, $r = .05$

from the standard normal, and the primary difference is in the larger kurtosis or weight in the tails. Indeed virtually any large financial data set will differ from the normal in this fashion; there may be some skewness in the distribution but there is often substantial kurtosis. How much difference does this slightly increased weight in the tails make in the price of an option? Note that the Black-Scholes formula overprices all of the options considered by up to around 3%. The differences are quite small, however and there seems to be considerable robustness to the Black-Scholes formula at least for this type of departure in the distribution of stock prices.

A change in the single line x=log(u./(1-u)) in the above function permits revising the returns distribution to another alternative. For example we might

choose the double exponential or Laplace density

$$f(x) = \frac{1}{2}\exp(-|x|)$$

for returns, by replacing this line by $x = (u < .5)\log(2 * u) - (u > .5)\log(2 * (1 - u))$. The resulting Figure 3.15 shows a similar behaviour but more substantial pricing error, in this case nearly 10% for an at-the-money option.

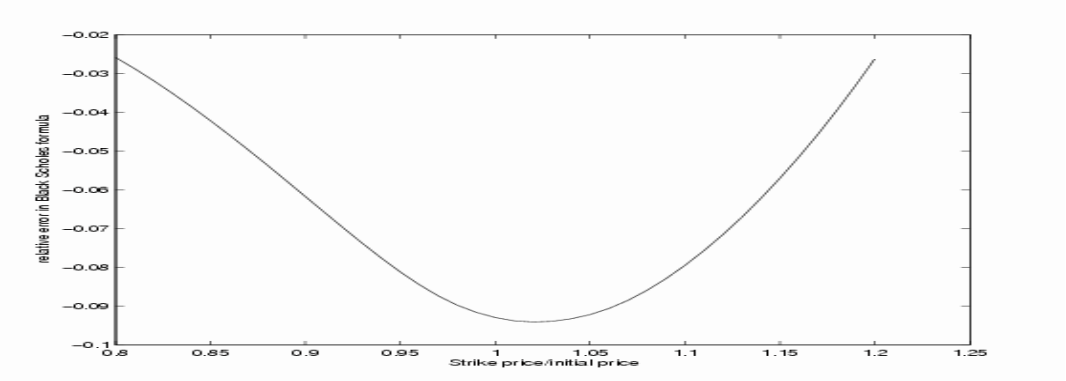


Figure 3.15: Relative pricing error in Black Scholes formula when returns follow the Laplace distribution

Another possible distribution of stock returns which can be used to introduce some skewness to the returns distribution is the loggamma or extreme value distribution whose probability density function takes the form

$$f(x) = \frac{1}{\Gamma(a)}\exp\{-e^{(x-c)} + (x-c)a\}, -\infty < x < \infty.$$

We can generate such a distribution as follows. Suppose Y is a random variable with gamma(a, e^c) distribution and probability density function

$$g(y) = \frac{y^{a-1}e^{-ca}}{\Gamma(a)}e^{-ye^{-c}}.$$

and define $X = \ln(Y)$. Then X has probability density function

$$f(x) = g(e^x)|\frac{d}{dx}e^x| = \frac{1}{\Gamma(a)}\exp\{(x(a-1) - ca - e^{x-c}\}e^x$$
$$= \frac{1}{\Gamma(a)}\exp\{-e^{x-c} + (x-c)a\}, -\infty < x < \infty.$$

As an example in Figure 3.16 we plot this density in the case $a = 2, c = 0$ This distribution is negatively skewed, a typical characteristic of risk-neutral distributions of returns. The large left tail in the risk-neutral distribution of returns reflects the fact that investors have an aversion to large losses and consequently the risk-neutral distribution inflates the left tail.

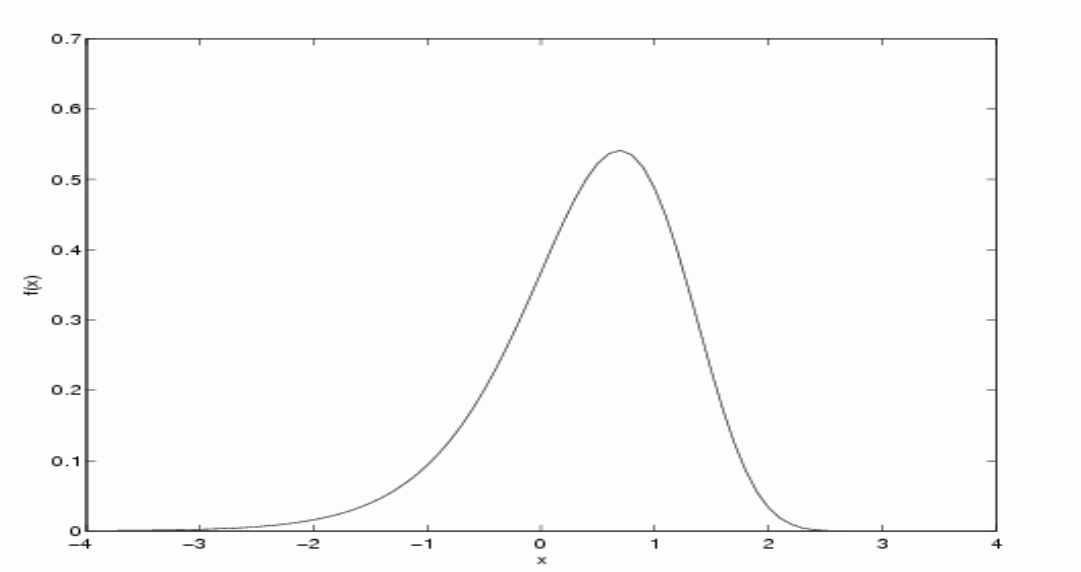


Figure 3.16: The probability density function e^{-e^x+2x}

Introducing a scale parameter ν, the probability density function of $\nu\ln(Y) = \ln(Y^\nu)$ where Y has a Gamma(2,1) distribution is

$$f(x) = \nu e^{-e^{(\nu x-c)}+2(\nu x-c)}.$$

The mean is approximately 0 and variance approximately σ^2 when we choose $c = -.42278$ and $\nu = .80308/\sigma$ and so this distribution is analogous to the

standard normal. However, the skewness is -0.78 and this negative skewness is more typical of risk neutral distributions of stock returns. We might ask whether the Black-Scholes formula is as robust to the introduction of skewness in the returns distribution as to the somewhat heavier tails of the logistic distribution. For comparison with the Black-Scholes model we permitted adding a constant and multiplying the returns by a constant which, in this case, is equivalent to assuming under the risk neutral distribution that

$$S_T = S_0 e^{\alpha} Y^{\nu}, Y \text{ is Gamma(2,1)}$$

where the constants α and ν are chosen so that the martingale condition is satisfied and the variance of returns matches that in the lognormal case. With some integration we can show that this results in the equations

$$\alpha = -\ln(E(Y^{\nu})) = -\ln(\Gamma(2+\nu))$$
$$\nu^2 var(\ln(Y)) = \nu^2 \psi'(2) = \sigma^2 T$$

where $\psi'(\alpha)$ is the *trigamma function* defined as the second derivative of $\ln(\Gamma(\alpha))$, and evaluated fairly easily using the series $\psi'(\alpha) = \sum_{k=0}^{\infty} \frac{1}{(k+\alpha)^2}$. For the special cases required here, $\psi'(2) \approx .6449$ so $\nu \approx \sigma\sqrt{T}/.8031$ and $\alpha = -\log(\Gamma(2+\sigma\sqrt{T}/.8031))$. Once again replacing the one line marked with a * in the function diffoptionprice by x=log(gaminf(u,2,1); permits determining the relative error in the Black-Scholes formula. There is a more significant pricing error in the Black-Scholes formula now, more typical of the relative pricing error that is observed in practice. Although the graph can be shifted and tilted somewhat by choosing different variance parameters, the shape appears to be a consequence of assuming a symmetric normal distribution for returns when the actual risk-neutral distribution is skewed. It should be noted that the practice of obtaining implied volatility parameters from options with similar strike prices and maturities is a partial, though not a compete, remedy to the substantial pricing errors caused by using a formula derived from a frequently ill-fitting

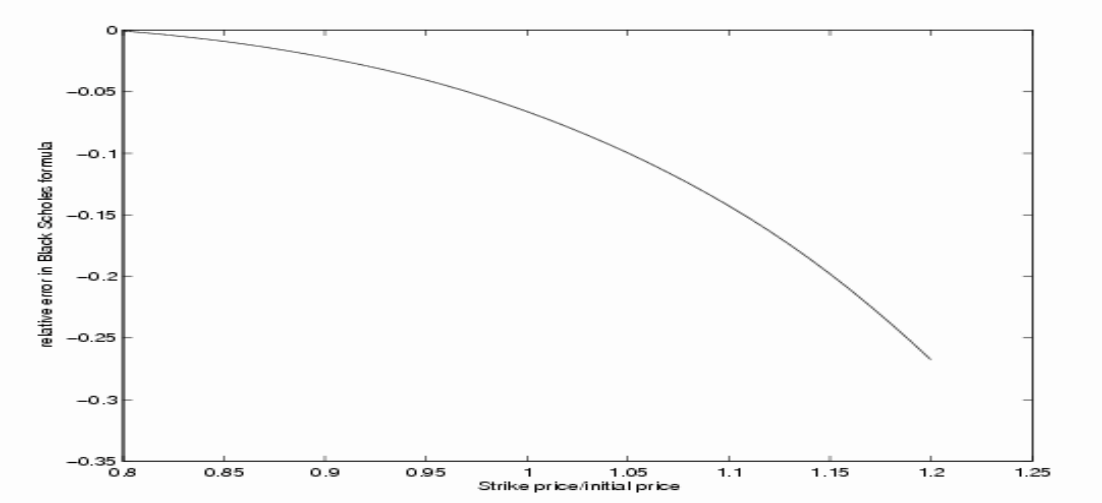


Figure 3.17: Relative Error in Black-Scholes formula when Asset returns follow extreme value

Black_Scholes model.

The Symmetric Stable Laws

A final family of distributions of increasing importance in modelling is the *stable family* of distributions. The stable cumulative distribution functions F are such that if two random variables X_1 and X_2 are independent with cumulative distribution function $F(x)$ then so too does the sum $X_1 + X_2$ after a change in location and scale. More generally the cumulative distribution function F of independent random variables X_1, X_2 is said to be *stable* if for each pair of constants a and b, there exist constants c and m such that

$$\frac{a\,X_1 + b\,X_2 - m}{c}$$

has the same cumulative distribution function F. A stable random variable X is most easily characterized through its characteristic function

$$Ee^{iuX} = \begin{cases} \exp(iu\theta - |u|^\alpha c^\alpha (1 - i\beta(sign\ u)\tan\frac{\pi\alpha}{2}) & \text{for} \quad \alpha \neq 1 \\ \exp(iu\theta - |u|c(1 + i\beta(sign\ \mathrm{u})\ln|u|)\frac{2}{\pi}) & \text{if} \quad \alpha = 1 \end{cases}$$

where i is the complex number $i^2 = -1$, θ is a location parameter of the distribution, and c is a scale parameter. The parameter $0 < \alpha \leq 2$ is the index of the stable distribution and governs the tail behavior and $\beta \in [-1, 1]$ governs the skewness of the distribution. In the case $\beta = 0$, we obtain the symmetric stable family of distributions, all unimodal densities, symmetric about their mode, and roughly similar in shape to the normal or Cauchy distribution (both special cases). They are of considerable importance in finance as an alternative to the normal distribution, in part because they tend to fit observations better in the tail of the distribution than does the normal, and in part because they enjoy theoretical properties similar to those of the normal family: sums of independent stable random variables are stable. Unfortunately, this is a more complicated family of densities to work with; neither the density function nor the cumulative distribution function can be expressed in a simple closed form. Both require a series expansion. The parameter $0 < \alpha \leq 2$ indicates what moments exist. Except in the special case $\alpha = 2$ (the normal distribution) or the case $\beta = -1$, moments of order less than α exist while moments of order α or more do not. This is easily seen because the tail behaviour is, when $\alpha < 2$,

$$\lim_{x\to\infty} x^\alpha P[X > x] = K_\alpha \frac{1+\beta}{2} c^\alpha$$
$$\lim_{x\to\infty} x^\alpha P[X < -x] = K_\alpha \frac{1-\beta}{2} c^\alpha$$

for constant K_α depending only on α. Of course, for the normal distribution, moments of all orders exist. The stable laws are useful for modelling in situations in which variates are thought to be approximately normalized sums of independent identically distributed random variables. To determine robustness

against heavy-tailed departures from the normal distribution, tests and estimators can be computed with data simulated from a symmetric stable law with α near 2. The probability density function does not have a simple closed form except in the case $\alpha = 1$ (Cauchy) and $\alpha = 2$ (Normal) but can be expressed as a series expansion of the form

$$f_c(x) = \frac{1}{\pi\,\alpha\,c}\sum_{k=0}^{\infty}(-1)^k\frac{\Gamma\left(\frac{2k+1}{\alpha}\right)}{(2k)!}(\frac{x}{c})^k$$

where c is the scale parameter (and we have assumed the mode is at 0). Especially for large values of x, this probability density function converges extremely slowly. However, Small (2003) suggests using an Euler transformation to accelerate the convergence of this series, and this appears to provide enough of an improvement in the convergence to meet a region in which a similar tail formula (valid for large x) provides a good approximation. According to Chambers, Mallows and Stuck, (1976), when $1 < \alpha < 2$, such a variate can be generated as

$$X = c\sin(\alpha U)\left[\frac{\cos(U(1-\alpha))}{E}\right]^{\frac{1}{\alpha}-1}(\cos U)^{-1/\alpha} \tag{3.20}$$

where U is uniform $[-\pi/2, \pi/2]$ and E, standard exponential are independent. The case $\alpha = 1$ and $X = \tan(U)$ is the Cauchy. It is easy to see that the Cauchy distribution can also be obtained by taking the ratio of two independent standard normal random variables and $\tan(U)$ may be replaced by Z_1/Z_2 for independent standard normal random variables Z_1, Z_2 produced by Marsaglia's polar algorithm. Equivalently, we generate $X = V_1/V_2$ where $V_i \sim U[-1, 1]$ conditional on $V_1^2 + V_2^2 \leq 1$ to produce a standard Cauchy variate X.

Example: Stable random walk.

A stable random walk may be used to model a stock price but the closest analogy to the Black Scholes model would be a logstable process S_t under which the distribution of $\ln(S_t)$ has a symmetric stable distribution. Unfortunately, this specification renders impotent many of our tools of analysis, since except in

the case $\alpha = 2$ or the case $\beta = -1$, such a stock price process S_t has no finite moments at all. Nevertheless, we may attempt to fit stable laws to the distribution of $\ln(S_t)$ for a variety of stocks and except in the extreme tails, symmetric stable laws with index $\alpha \simeq 1.7$ often provide a reasonably good fit. To see what such a returns process looks like, we generate a random walk with 10,000 time steps where each increment is distributed as independent stable random variables having parameter 1.7. The following *Matlab* function was used

```
function s=stabrnd(a,n)
u=(unifrnd(0,1,n,1)*pi)-.5*pi;
e = exprnd(1,n,1);
s=sin(a*u).*(cos((1-a)*u)./e).^(1/a-1).*(cos(u)).^(-1/a)
```

Then the command

```
plot(1:10000, cumsum(stabrnd(1.7,10000)));
```

resulted in the Figure 3.18. Note the occasional very large jump(s) which dominates the history of the process up to that point, typical of random walks generated from the stable distributions with $\alpha < 2$.

The Normal Inverse Gaussian Distribution

There is a very substantial body of literature that indicates that the normal distribution assumption for returns is a poor fit to data, in part because the observed area in the tails of the distribution is much greater than the normal distribution permits. One possible remedy is to assume an alternative distribution for these returns which, like the normal distribution, is infinitely divisible, but which has more area in the tails. A good fit to some stock and interest rate data has been achieved using the Normal Inverse Gaussian (NIG) distribution (see for example Prausse, 1999). To motivate this family of distributions, let us

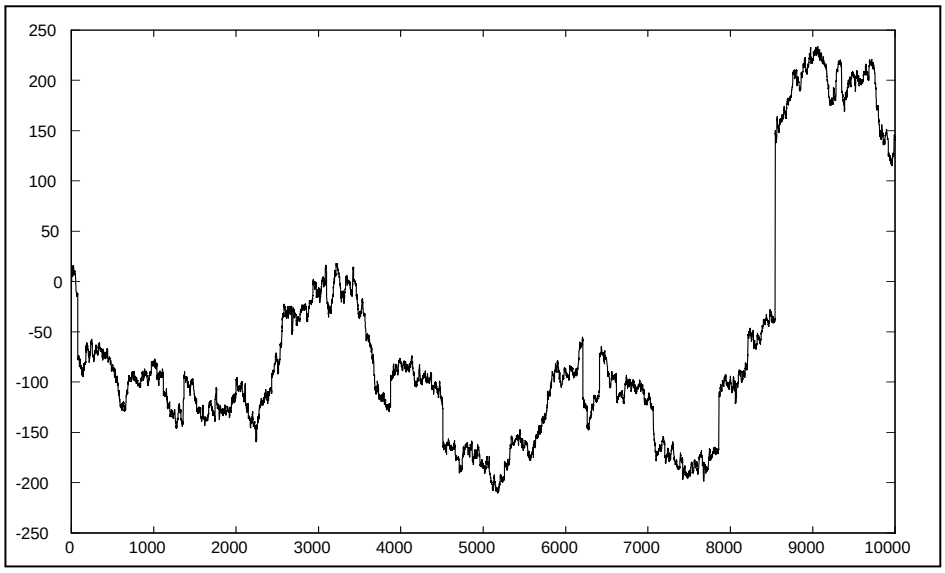


Figure 3.18: A Symmetric Stable Random Walk with index $\alpha = 1.7$

suppose that stock returns follow a Brownian motion process but with respect to a random time scale possibly dependent on volume traded and other external factors independent of the Brownian motion itself. After one day, say, the return on the stock is the value of the Brownian motion process at a random time, τ, independent of the Brownian motion. Assume that this random time has the *Inverse Gaussian* distribution having probability density function

$$g(t) = \frac{\theta}{c\sqrt{2\pi t^3}} \exp\{-\frac{(\theta - t)^2}{2c^2 t}\} \tag{3.21}$$

for parameters $\theta > 0, c > 0$. This is the distribution of a first passage time for Brownian motion. In particular consider a Brownian motion process $B(t)$ having drift 1 and diffusion coefficient c. Such a process is the solution to the stochastic differential equation

$$dB(t) = dt + cdW(t), B(0) = 0.$$

Then the first passage of the Brownian motion to the level θ is $T = \inf(t; B(t) = \theta\}$ and this random variable has probability density function (3.21). The mean

of such a random variable is θ and with variance θc^2. These can be obtained from the moment generating function of the distribution with probability density function (3.21),

$$g^*(s) = \exp\{-\theta(\frac{-1+\sqrt{1-2sc}}{c^2})\}.$$

Expanding this locally around $c = 0$ we obtain

$$g^*(s) = \exp\{\theta s + \frac{1}{2}\theta s^2 c^2 + O(c^4)\}$$

and by comparing this with the moment generating function of the normal distribution, as $c \to 0$, the distribution of

$$\frac{T-\theta}{c\sqrt{\theta}}$$

approaches the standard normal or, more loosely, the distribution (3.21) approaches Normal$(\theta, \theta c^2)$.

Lemma 30 *Suppose $X(t)$ is a Brownian motion process with drift β and diffusion coefficient* 1*, hence satisfying*

$$dX_t = \beta dt + dW_t, \quad X(0) = \mu.$$

Suppose a random variable T has probability density function (3.21) and is independent of X_t. Then the probability density function of the randomly stopped Brownian motion process is given by

$$f(x;\alpha,\beta,\delta,\mu) = \frac{\alpha\delta}{\pi}\exp(\delta\sqrt{\alpha^2-\beta^2}+\beta(x-\mu))\frac{K_1(\alpha\sqrt{\delta^2+(x-\mu)^2})}{\sqrt{\delta^2+(x-\mu)^2}} \quad (3.22)$$

with

$$\delta = \frac{\theta}{c}, \quad \text{and } \alpha = \sqrt{\beta^2+\frac{1}{c^2}}$$

and the function $K_\lambda(x)$ is the modified Bessel function of the second kind defined by

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp(-\frac{x}{2}(y+y^{-1}))dy, \text{ for } x > 0.$$

Proof. The distribution of the randomly stopped variable $X(T)$ is the same as that of the random variable

$$X = \mu + \beta T + \sqrt{T}Z$$

where Z is $N(0,1)$ independent of T. Conditional on the value of T the probability density function of X is

$$f(x|T) = \frac{1}{\sqrt{2\pi T}} \exp(-\frac{1}{2T}(x-\mu-\beta T)^2)$$

and so the unconditional distribution of X is given by

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi t}} \exp(-\frac{1}{2t}(x-\mu-\beta t)^2) \frac{\theta}{c\sqrt{2\pi t^3}} \exp(-\frac{(\theta-t)^2}{2c^2 t}) dt \\
&= \frac{\theta}{2\pi c} \int_0^\infty t^{-2} \exp(-\frac{1}{2t}(x-\mu-\beta t)^2 - \frac{(\theta-t)^2}{2c^2 t}) dt \\
&= \frac{\theta}{2\pi c} \int_0^\infty t^{-2} \exp(-\frac{1}{2t}(x^2 - 2x\mu + \mu^2 + \theta^2) + (\beta(x-\mu) + \frac{\theta}{c^2}) - \frac{t}{2}(\beta^2 + \frac{1}{c^2})) dt \\
&= \frac{\theta}{2\pi c} \exp(\beta(x-\mu) + \frac{\theta}{c^2}) \int_0^\infty t^{-2} \exp(-\frac{1}{2t}((x-\mu)^2 + \theta^2) - \frac{t}{2}(\beta^2 + \frac{1}{c^2})) dt \\
&= \frac{\alpha\delta}{\pi} \exp(\delta\sqrt{\alpha^2-\beta^2} + \beta(x-\mu)) \frac{K_1(\alpha\sqrt{\delta^2 + (x-\mu)^2})}{\sqrt{\delta^2 + (x-\mu)^2}}.
\end{aligned}$$

■

The modified Bessel function of the second kind $K_\lambda(x)$ is given in MATLAB by *besselk*(ν, x) and in **R** by besselK(x,ν,expon.scaled=FALSE). The distribution with probability density function given by (3.22) is called the *normal inverse Gaussian* distribution with real-valued parameters x, μ, $0 \leq \delta$ and $\alpha \geq |\beta|$. The tails of the normal inverse Gaussian density are substantially heavier than those of the normal distribution. In fact up to a constant

$$f(x; \alpha, \beta, \delta, \mu) \sim |x|^{-3/2} \exp((\mp\alpha + \beta)x) \text{ as } x \to \pm\infty.$$

The moments of this distribution can be obtained from the moment generating function

$$M(s) = e^{\mu s} \left[\frac{\alpha^2 - (\beta+s)^2}{\alpha^2 - \beta^2}\right]^{1/4} \exp\{\delta(\alpha^2 - \beta^2)^{1/2} - \delta(\alpha^2 - (s+\beta)^2)^{1/2}\} \text{ for } |\beta+s| < \alpha. \tag{3.23}$$

These moments are:

$$E(X) = \mu + \delta\beta(\alpha^2 - \beta^2)^{-1/2}$$
$$var(X) = \delta\alpha^2(\alpha^2 - \beta^2)^{-3/2}$$

and the skewness and kurtosis:

$$\text{skew} = 3\beta\alpha^{-1}\delta^{-1/2}(\alpha^2 - \beta^2)^{-1/4}$$
$$\text{kurtosis} = 3\delta^{-1}\alpha^{-2}(\alpha^2 + 4\beta^2)(\alpha^2 - \beta^2)^{-1/2}.$$

One of the particularly attractive features of this family of distributions, shared by the normal and the stable family of distributions, is that it is closed under convolutions. This is apparent from the moment generating function (3.23) since

$$M^N(s)$$

gives a moment generating function of the same form but with μ replaced by μN and δ by δN. In Figure 3.19 we plot the probability density function of a member of this family of distributions.

Note the similarity to the normal density but with a modest amount of skewness and increased weight in the tails. We can generate random variables from this distribution as follows:

Sample T from an inverse Gaussian distribution (3.21)

Return $X = \mu + \beta T + N(0, T)$

where $N(0, T)$ is a normal random variable with mean 0 and variance T.

We sample from the inverse Gaussian by using a property of the distribution that if T has density of the form (3.21) then

$$\frac{(T-\theta)^2}{c^2 T} \tag{3.24}$$

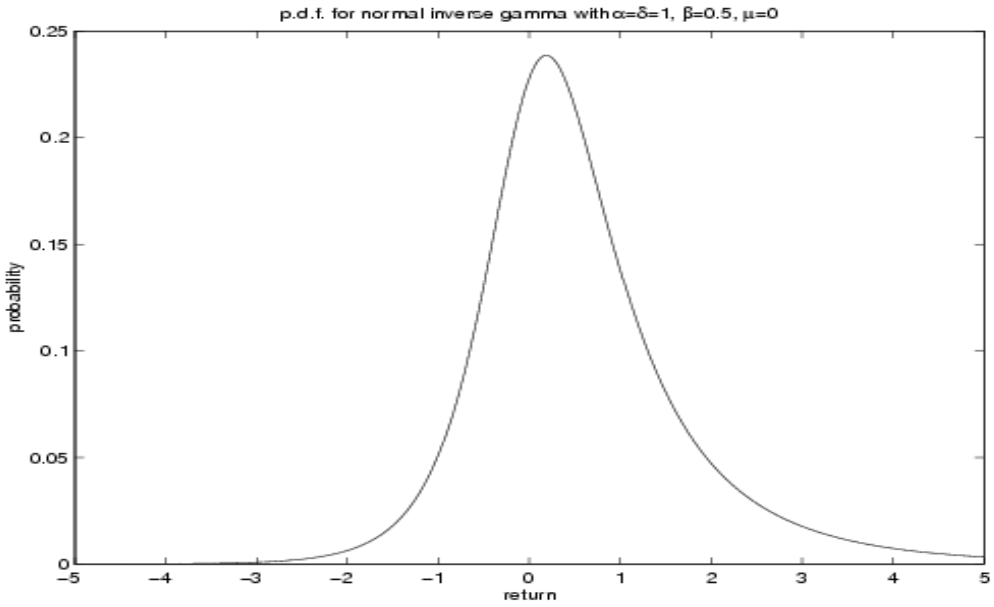


Figure 3.19: Normal Inverse Gaussian probability density function with $\alpha = \delta = 1$, $\beta = \frac{1}{2}, \mu = 0$

has a chi-squared distribution with one degree of freedom (easily generated as the square of a standard normal random variable). The algorithm is (see Michael, Shucany, Hass (1976));

1. For
$$c = \frac{1}{\sqrt{\alpha^2 - \beta^2}}, \text{ and } \theta = \delta c,$$
generate G_1 from the Gamma$(\frac{1}{2}, \frac{c}{\delta})$ distribution. Define
$$Y_1 = 1 + G_1(1 - \sqrt{1 + \frac{2}{G_1}})\}.$$

2. Generate $U_2 \sim U[0,1]$. If $U_2 \leq \frac{1}{1+Y_1}$ then output $T = \theta Y_1$

3. Otherwise output $T = \theta Y_1^{-1}$.

The two values θY_1, and θY_1^{-1} are the two roots of the equation obtained by setting (3.24) equal to a chi-squared variate with one degree of freedom and the relative values of the probability density function at these two roots are $\frac{1}{1+Y_1}$ and $1 - \frac{1}{1+Y_1}$.

Finally to generate from the normal inverse Gaussian distribution (3.22) we generate an inverse gamma random variable above and then set $X = \mu + \beta T + N(0, T)$. Prause (1999) provides a statistical evidence that the Normal Inverse Gaussian provides a better fit than does the normal itself. For example we fit the normal inverse gamma distribution to the S&P500 index returns over the period Jan 1, 1997-Sept 27, 2002. There were a total of 1442 values over this period. Figure 3.20 shows a histogram of the daily returns together with the normal and the NIG fit to the data. The mean return over this period is 8×10^{-5} and the standard deviation of returns 0.013. If we fit the normal inverse Gaussian distribution to these returns we obtain parameter estimates

$$\alpha = 95.23, \beta = -4.72, \delta = 0.016, \mu = 0.0009$$

and the Q-Q plots in Figure 3.21 . Both

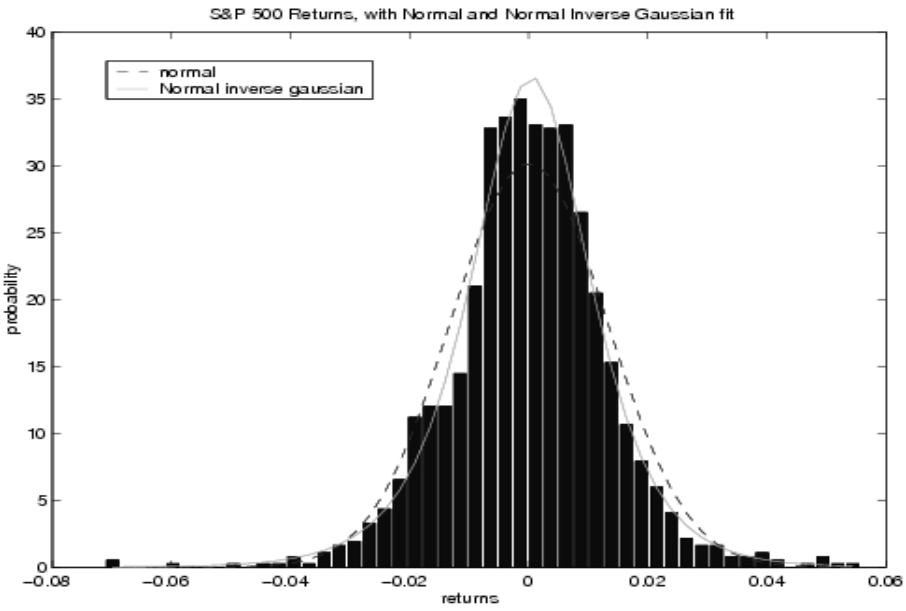


Figure 3.20: The Normal and the Normal inverse Gaussian fit to the S&P500 Returns

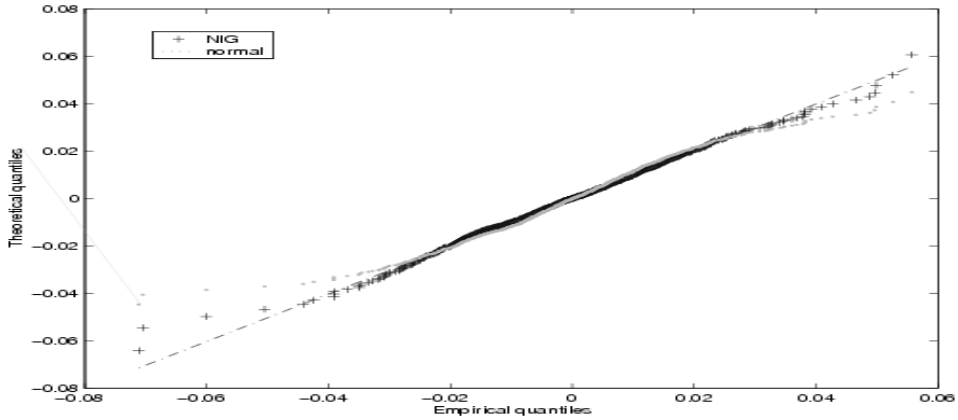


Figure 3.21: QQ plots showing the Normal Inverse Gaussian and the Normal fit to S&P 500 data, 1997-2002

indicate that the normal approximation fails to properly fit the tails of the distribution but that the NIG distribution is a much better fit. This is similar to the conclusion in Prause using observations on the Dow Jones Industrial Index.

Generating Random Numbers from Discrete Distributions

Many of the methods described above such as inversion and acceptance-rejection for generating continuous distributions work as well for discrete random variables. Suppose for example X is a discrete distribution taking values on the integers with probability function $P[X = x] = f(x)$, for $x = 0, 1, 2, ...$ Suppose we can find a continuous random variable Y which has exactly the same value of its cumulative distribution function at these integers so that $F_Y(j) = F_X(j)$ for all $j = 1, 2, ...$ Then we may generate the continuous random variable Y, say by inversion or acceptance-rejection and then set $X = \lfloor Y \rfloor$ the integer part of Y. Clearly X takes integer values and since $P[X \leq j] = P[Y \leq j] = F_X(j)$ for all $j = 0, 1, ...,$ then X has the desired distribution. The continuous exponential distribution and the geometric distribution are linked in this way. If X has a geometric(p) distribution and Y has the exponential distribution with parameter $\lambda = -\ln(1-p)$, then X has the same distribution as $\lceil Y \rceil$ or $\lfloor Y \rfloor + 1$.

Using the inverse transform method for generating discrete random variables is usually feasible but for random variables with a wide range of values of reasonably high probability, it often requires some setup costs to achieve reasonable efficiency. For example if X has cumulative distribution function $F(x), x = 0, 1, ...$inversion requires that we output an integer $X = F^{-1}(U)$,an integer X satisfying $F(X-1) < U \leq F(X)$. The most obvious technique for finding such a value of X is to search sequentially through the potential values $x = 0, 1, 2, ...$ Figure 3.22 is the search tree for inversion for the distribution on

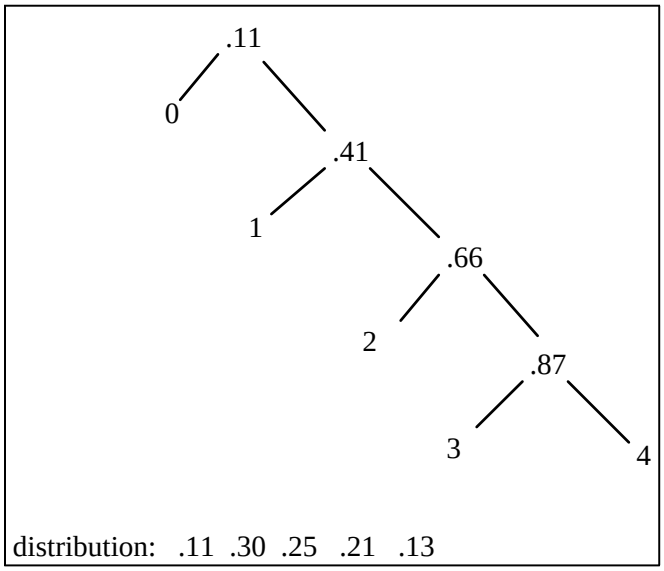


Figure 3.22: Sequential Search tree for Inverse Transform with root at $x = 0$

the integers $0, \ldots 4$ given by

x	0	1	2	3	4
$f(x)$	0.11	0.30	0.25	0.21	0.13

We generate an integer by repeatedly comparing a uniform [0,1] variate U with the value at each node, taking the right branch if it is greater than this threshold value, the left if it is smaller. If X takes positive integer values $\{1, 2, ..., N\}$,the number of values searched will average to $E(X)$ which for many discrete distributions can be unacceptably large.

An easy alternative is to begin the search at a value m which is near the median (or mode or mean) of the distribution. For example we choose $m = 2$ and search to the left or right depending on the value of U in Figure 3.23.

If we assume for example that we root the tree at m then this results in searching roughly an average of $E[|X - m + 1|]$ before obtaining the generated variable. This is often substantially smaller than $E(X)$ especially when $E(X)$

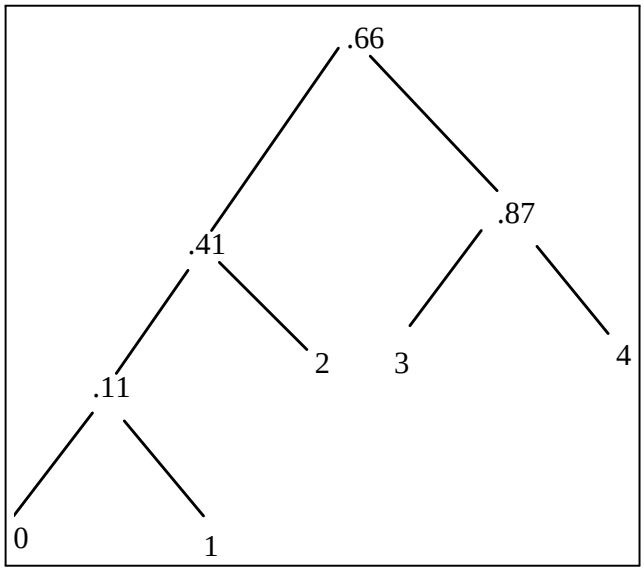


Figure 3.23: Search tree rooted near the median

is large but still unacceptably large when the distribution has large variance.

An optimal binary search tree for this distribution is graphed in Figure 3.24. This tree has been constructed from the bottom up as follows. We begin by joining the two smallest probabilities $f(4)$ and $f(0)$ to form a new node with weight $f(0) + f(4) = 0.24$. Since we take the left path (towards $X = 0$ rather than towards $X = 4$) if U is smaller than the value .11 labelling the node at the intersection of these two branches. We now regard this pair of values as a unit and continue to work up the tree from the leaves to the root. The next smallest pair of probabilities are $\{0, 1\}$ and $\{3\}$ which have probabilities 0.24 and 0.21 respectively so these are the next to be joined hence working from the leaves to the root of the tree. This optimal binary search tree provides the minimum expected number of comparisons and is equivalent to sorting the values in order of largest to smallest probability, in this case $1, 2, 3, 4, 0$, relabelling them or coding them $\{0, 1, 2, 3, 4\}$ and then applying the inverse transform method starting at 0.

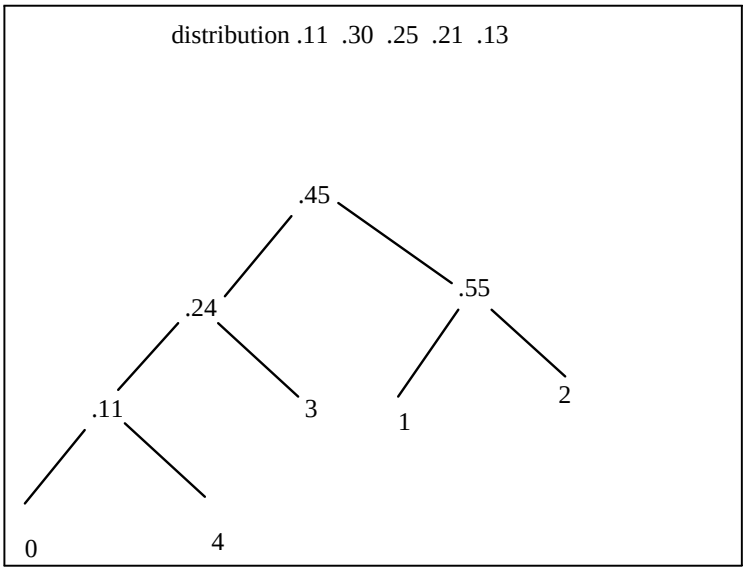


Figure 3.24: Optimal Binary Search Tree

The leaves of the tree are the individual probabilities $f(j)$ and the internal nodes are sums of the weights or probabilities of the "children", the values $f(j)$ for j on paths below this node. Let D_i represents the depth of the $i'th$ leaf so for example the depth of leaf 0 in Figure 3.24 is $D_0 = 3$. Then the average number of comparisons to generate a single random variable Xstarting at the root is $\sum_i f(i)D_i$. The procedure for constructing the last tree provides an optimal algorithm in the sense that this quantity is minimized. It is possible to show that an optimal binary search tree will reduce the average number of comparisons from $E(X)$ for ordinary inversion to less than $1 + 4\,[log_2(1 + E(X))]$.

Another general method for producing variates from a discrete distribution was suggested by Walker (1974, 1977) and is called the *alias method*. This is based on the fact that *every discrete distribution is a uniform mixture of two-point distributions*. Apart from the time required to set up an initial table of aliases and aliasing probabilities, the time required to generate values from a discrete distribution with K supporting points is bounded in K, whereas methods

such as inverse transform have computational time which increase proportionally with $E(X)$.

Consider a discrete distribution of the form with probability function $f(j)$ on K integers $j = 1, 2, ...K$. We seek a table of values of $A(i)$ and associated "alias" probabilities $q(i)$ so that the desired discrete random variable can be generated in two steps, first generate one of the integers $\{1, 2, ..., K\}$ at random and uniformly, then if we generated the value I, say, replace it by an "alias" value $A(I)$ with alias probability $q(I)$. These values $A(I)$ and $q(I)$ are determined below. The algorithm is:

GENERATE I UNIFORM ON $\{1, ...K\}$.

WITH PROBABILITY $q(I)$, OUTPUT $X = I$, OTHERWISE, $X = A(I)$.

An algorithm for producing these values of $(A(i), q(i)), i = 1, ..., K\}$ is suggested by Walker(1977) and proceeds by reducing the number of non-zero probabilities one at a time.

1. Put $q(i) = Kf(i)$ for all $i = 1, ..., K$.

2. LET m be the index so that $q(m) = \min\{q(i); q(i) > 0\}$ and let $q(M) = \max\{q(i); q(i) > 0\}$.

3. SET $A(m) = M$ and fix $q(m)$ (it is no longer is subject to change).

4. Replace $q(M)$ by $q(M) - (1 - q(m))$

5. Replace $(q(1), ...q(K))$by $(q(1), ..., q(m-1), q(m+1), .q(M)$ (so the component with index m is removed).

6. Return to 2 unless all remaining $q_i = 1$ or the vector of q_i's is empty.

Note that on each iteration of the steps above, we fix one of components $q(m)$ and remove it from the vector and adjust one other, namely $q(M)$. Since we always fix the smallest $q(m)$ and since the average $q(i)$ is one, we always

obtain a probability, i.e. fix a value $0 < q(m) \leq 1$. Figure 3.25 shows the way in which this algorithm proceeds for the distribution

$x =$	1	2	3	4
$f(x) =$	.1	.2	.3	.4

We begin with $q(i) = 4 \times f(i) = .4, .8, 1.2, 1.6$ for $i = 1, 2, 3, 4$. Then since $m = 1$ and $M = 4$ these are the first to be adjusted. We assign $A(1) = 4$ and $q(1) = 0.4$. Now since we have reassigned mass $1 - q(1)$ to $M = 4$ we replace $q(4)$ by $1.6 - (1 - 0.4) = 1$. We now fix and remove $q(1)$ and continue with $q(i) = .8, 1.2, 1.0$ for $i = 2, 3, 4$. The next step results in fixing $q(2) = 0.8$, $A(2) = 3$ and changing $q(3)$ to $q(3) - (1 - q(2)) = 1$. After this iteration, the remaining $q(3)$, $q(4)$ are both equal to 1, so according to step 6 we may terminate the algorithm. Notice that we terminated without assigning a value to $A(3)$ and $A(4)$. This assignment is unnecessary since the probability the alias $A(i)$ is used is $(1 - q(i))$ which is zero in these two cases. The algorithm therefore results in aliases $A(i) = 4, 3, i = 1, 2$ and $q(i) = .4, .8, 1, 1$, respectively for $i = 1, 2, 3, 4$. Geometrically, this method iteratively adjusts a probability histogram to form a rectangle with base K as in Figure 3.25.

Suppose I now wish to generate random variables from this discrete distribution. We simply generate a random variable uniform on the set $\{1, 2, 3, 4\}$ and if 1 is selected, we replace it by $A(1) = 4$ with probability $1 - q(1) = 0.6$. If 2 is selected it is replaced by $A(2) = 3$ with probability $1 - q(2) = 0.2$.

Acceptance-Rejection for Discrete Random Variables

The acceptance-rejection algorithm can be used both for generating discrete and continuous random variables and the geometric interpretation in both cases is essentially the same. Suppose for example we wish to generate a discrete random variable X having probability function $f(x)$ using as a dominating function a multiple of $g(x)$ the probability density function of a *continuous* random variable. Take for example the probability function

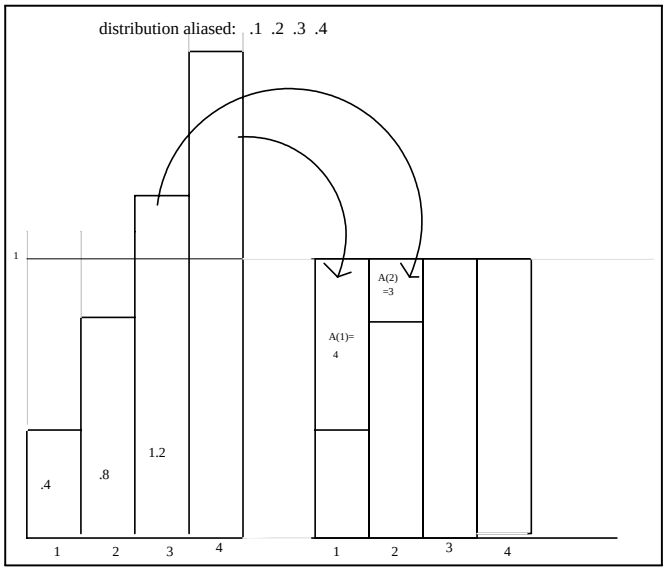


Figure 3.25: The alias method for generating from the distribution 0.1 0.2 0.3 0.4

$x =$	1	2	3	4
$f(x) =$	.1	.3	.4	.2

using the dominating function $2g(x) = 0.1 + 0.2(x - 0.5)$ for $0.5 < x < 4.5$. It is easy to generate a continuous random variable from the probability density function $g(x)$ by inverse transform. Suppose we generate the value X. Then if this value is under the probability histogram graphed in Figure 3.26 we accept the value (after rounding it to the nearest integer to conform the discreteness of the output distribution) and otherwise we reject and repeat.

We may also dominate a discrete distribution with another discrete distribution in which case the algorithm proceeds as in the continuous case but with the probability density functions replaced by probability functions.

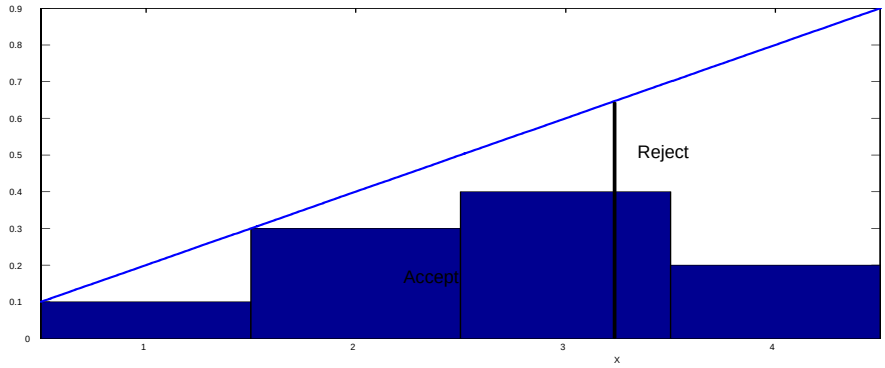


Figure 3.26: Acceptance-Rejection with for Discrete Distribution with continuous dominating function.

The Poisson Distribution.

Consider the probability function for a *Poisson distribution with parameter* λ

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, ... \tag{3.25}$$

The simplest generator is to use the Poisson process. Recall that a Poisson process with rate 1 on the real line can be described in two equivalent ways:

1. Points are distributed on the line in such a way that the spacings between consecutive points are independent exponential(λ) random variables. Then the resulting process is a Poisson process with rate λ.

2. The number of points in an interval of length h has a Poisson (λh) distribution. Moreover the numbers of points in non-overlapping intervals are independent random variables.

The simplest generator stems from this equivalence. Suppose we use the first specification to construct a Poisson process with rate parameter 1 and then examine X = the number of points occurring in the interval $[0, \lambda]$. This is the number of partial sums of exponential(1) random variables that are less

than or equal to λ

$$X = \inf\{n; \sum_{i=1}^{n+1}(-lnU_i) > \lambda\}$$

or equivalently

$$X = \inf\{n; \prod_{i=1}^{n+1} U_i < e^{-\lambda}\} \tag{3.26}$$

This generator requires CPU time which grows linearly with λ since the number of exponential random variables generated and summed grows linearly with λ and so an alternative for large λ is required. Various possibilities of acceptance-rejection algorithms have been suggested including dominating the Poisson probability function with multiples of the logistic probability density function (Atkinson (1979)), the normal density with exponential right tail (cf. Devroye, lemma 3.8, page 509). A simple all-purpose dominating function is the so-called table-mountain function (cf. Stadlober (1989)), essentially a function with a flat top and tails that decrease as $1/x^2$. Another simple alternative for generating Poisson variates that is less efficient but simpler to implement is to use the *Lorentzian*, or truncated Cauchy distribution with probability density function

$$g(x|a,b) = \frac{c_0}{b^2 + (x-a)^2}, x > 0 \tag{3.27}$$

where c_0 is the normalizing constant. A random variable is generated from this distribution using the inverse transform method; $X = a + b\tan(\pi U),$, where $U \sim U[0,1]$. Provided that we match the modes of the distribution $a = \lambda$ and put $b = \sqrt{2\lambda}$, this function may be used to dominate the Poisson distribution and provide a simple rejection generator. The *Matlab* Poisson random number generator is *poissrnd*(λ, m, n) which generates an $m \times n$ matrix of Poisson(λ) variables. This uses the simple generator (3.26) and is not computationally efficient for large values of λ.In **R** the command *rpois(n,*λ) generates a vector of n Poisson variates.

The Binomial Distribution

For the *Binomial* distribution, we may use any one of the following alternatives:

(1) $X = \sum_{i=1}^{n} I(U_i < p), U_i \sim$ independent uniform$[0,1]$

(2) $X = \inf\{x; \sum_{i=1}^{x+1} G_i > n\}$, where $G_i \sim$independent Geometric(p)

(3) $X = \inf\{x; \sum_{i=1}^{x+1} \frac{E_i}{n-i+1} > -log(1-p)\}$, where $E_i \sim$independent Exponential(1).

Method (1) obtains from the definition of the sum of independent Bernoulli random variables since each of the random variables $I(U_i < p)$ are independent, have values 0 and 1 with probabilities $1-p$ and p respectively. The event $(U_i < p)$ having probability p is typically referred to as a "success". Obviously this method will be slow if n is large. For method (2), recall that the number of trials necessary to obtain the first success, G_1, say, has a geometric distribution. Similarly, G_2 represents the number of additional trials to obtain the second success. So if $X = j$, the number of trials required to obtain $j+1$ successes was greater than n and to obtain j successes, less than or equal to n. In other words there were exactly j successes in the first n trials. When n is large but np fairly small, method (2) is more efficient since it is proportional to the number os successes rather than the total number of trials. Of course for large n and np sufficiently small (e.g. <1), we can also replace the Binomial distribution by its Poisson ($\lambda = np$) approximation. Method (3) is clearly more efficient if $-log(1-p)$ is not too large so that p is not too close to 1, because in this case we need to add fewer exponential random variables.

For large mean np and small $n(1-p)$ we can simply reverse the role of successes and failures and use method (2) or (3) above. But if both np and $n(1-p)$ are large, a rejection method is required. Again we may use rejection beginning with a Lorentzian distribution, choosing $a = np$, and $b = \sqrt{2np(1-p)}$ in the case $p < 1/2$. When $p > 1/2$, we simply reverse the roles of "failures" and "successes". Alternatively, a dominating table-mountain function may be

used (Stadlober (1989)). The binomial generator in *Matlab* is the function *binornd(n,p,j,k)* which generates an $n \times k$ matrix of binomial(n, p) random variables. This uses the simplest form (1) of the binomial generator and is not computationally efficient for large n. In **R**, *rbinom(m,n,p)* will generate a vector of length m of Binomial(n, p) variates.

Random Samples Associated with Markov Chains

Consider a finite state *Markov Chain,* a sequence of (discrete) random variables $X_1, X_2, \ldots$each of which takes integer values $1, 2, \ldots N$ (called *states*). The number of states of a Markov chain may be large or even infinite and it is not always convenient to label them with the positive integers and so it is common to define the *state space* as the set of all possible states of a Markov chain, but we will give some examples of this later. For the present we restrict attention to the case of a finite state space. The *transition probability matrix* is a matrix P describing the conditional probability of moving between possible states of the chain, so that

$$P[X_{n+1} = j | X_n = i] = P_{ij}, i = 1, \ldots N, j = 1, \ldots N.$$

where $P_{ij} \geq 0$ for all i, j and $\sum_j P_{ij} = 1$ for all i. A *limiting distribution* of a Markov chain is a vector ($\underline{\pi}$ say) of long run probabilities of the individual states with the property that

$$\pi_i = lim_{t \to \infty} P[X_t = i].$$

A *stationary distribution* of a Markov chain is the column vector ($\underline{\pi}$ say) of probabilities of the individual states such that

$$\underline{\pi}' P = \underline{\pi}'. \tag{3.28}$$

$\underline{\pi}'P = \underline{\pi}'$. For a Markov chain, every limiting distribution is in fact a stationary distribution. For the basic theory of Markov Chains, see the Appendix. Roughly, a Markov chain which eventually "forgets" the states that were occupied in the distant path, in other words for which the probability of the current states does not vary much as we condition on different states in the distant past, is called ergodic. A Markov chain which simply cycles through three states $1 \to 2 \to 3 \to 1 \to ...$ is an example of a periodic chain, and is not ergodic.

It is often the case that we wish to simulate from a finite ergodic Markov chain when it has reached equilibrium or stationarity, which is equivalent to sampling from the distribution of X_n assuming that the distribution of X_0 is given by the stationary distribution $\underline{\pi}$. In a few cases, we can obtain this stationary distribution directly from (3.28) but when N is large this system of equations is usually not feasible to solve and we need to find another way to sample from the probability vector π. Of course we can always begin the Markov chain in some arbitrary initial state and run it waiting for Hele to freeze over (it does happen since Helle is in Devon) until we are quite sure that the chain has essentially reachedequilibrium, and then use a subsequent portion of this chain, discarding this initial period, sometimes referred to as the "initial transient".

Clearly this is often not a very efficient method, particularly in cases in which the chain mixes or forgets its past very slowly for in this case the required initial transient is long. On the other hand if we shortened it, we run the risk of introducing bias into our simulations because the distribution generated is too far from the equilibrium distribution π. There are a number of solutions to this problem proposed in a burgeoning literature. Here we limit ourselves to a few of the simpler methods.

Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm is a method for generating random variables from a distribution π that applies even in the case of an infinite number of states or a continuous distribution π. It is assumed that π is known up to some multiplicative constant. Roughly, the method consists of using a convenient "proposal" Markov chain with transition matrix Q to generate transitions, but then only "accept" the move to these new states with probability that depends on the distribution π. The idea resembles that behind importance sampling. The basic result on which the Metropolis-Hastings algorithm is pinned is the following theorem.

Theorem 31 *Suppose Q_{ij} is the transition matrix of a Markov chain. Assume that g is a vector of non-negative values such that $\sum_{i=1}^{N} g_i = G$ and*

$$|\frac{g_j}{Q_{ij}}| \leq K < \infty \text{ for all } i, j$$

for some finite value K. Define

$$\rho_{ij} = \min(1, \frac{g_j Q_{ji}}{g_i Q_{ij}})$$

Then the Markov Chain with transition probability matrix

$$P_{ij} = Q_{ij}\rho_{ij}, \text{ for } i \neq j \tag{3.29}$$

has stationary distribution $\pi_i = \frac{g_i}{G}$.

Proof. The proof consists of showing that the so-called "detailed balance condition" is satisfied, i.e. with $\pi_i = \frac{g_i}{G}$, that

$$\pi_i P_{ij} = \pi_j P_{ji}, \text{ for all } i, j. \tag{3.30}$$

This condition implies that when the chain is operating in equilibrium,

$$P[X_n = i, X_{n+1} = j] = P[X_n = j, X_{n+1} = i]$$

reflecting a cavalier attitude to the direction in which time flows or reversibility of the chain. Of course (3.30) is true automatically if $i = j$ and for $i \neq j$,

$$\begin{aligned}\pi_i P_{ij} &= \frac{g_i}{G} Q_{ij} \min(1, \frac{g_j Q_{ji}}{g_i Q_{ij}}) \\ &= \frac{1}{G} \min(g_i Q_{ij}, g_j Q_{ji}) \\ &= \pi_j P_{ji}\end{aligned}$$

by the symmetry of the function $\frac{1}{G} \min(g_i Q_{ij}, g_j Q_{ji})$. Now the detailed balance condition (3.30) implies that π is a stationary distribution for this Markov chain since

$$\begin{aligned}\sum_{i=1}^{N} \pi_i P_{ij} &= \sum_{i=1}^{N} \pi_j P_{ji} \\ &= \pi_j \sum_{i=1}^{N} P_{ji} \\ &= \pi_j \text{ for each } j = 1, ..., N.\end{aligned}$$

■

Provided that we are able to generate transitions for the Markov Chain with transition matrix Q, it is easy to generate a chain with transition matrix P in (3.29). If we are currently in state i, generate the next state with probability Q_{ij}. If $j = i$ then we stay in state i. If $j \neq i$, then we "accept" the move to state j with probability ρ_{ij}, otherwise we stay in state i. Notice that the Markov Chain with transition matrix P tends to favour moves which increase the value of π. For example if the proposal chain is as likely to jump from i to j as it is to jump back so that $Q_{ij} = Q_{ji}$, then if $\pi_j > \pi_i$ the move to j is always accepted whereas if $\pi_j < \pi_i$ the move is only accepted with probability $\frac{\pi_j}{\pi_i}$. The assumption $Q_{ij} = Q_{ji}$ is a common and natural one, since in applications of the Metropolis-Hastings algorithm, it is common to choose j "at random" (i.e. uniformly distributed) from a suitable neighborhood of i.

The above proof only provides that π is a stationary distribution of the Markov Chain associated with P, not that it is necessarily the limiting distrib-

ution of this Markov chain. For this to follow we need to know that the chain is ergodic. Various conditions for ergodicity are given in the literature. See for example Robert and Casella (1999, Chapter 6) for more detail.

Gibbs Sampling

There is one simple special case of the Metropolis-Hastings algorithm that is particularly simple, common and compelling. To keep the discussion simple, suppose the possible states of our Markov Chain are points in two-dimensional space (x, y). We may assume both components are discrete or continuous. Suppose we wish to generate observations from a stationary distribution which is proportional to $g(x, y)$ so

$$\pi(x, y) = \frac{g(x, y)}{\sum_x \sum_y g(x, y)} \tag{3.31}$$

defined on this space but that the form of the distribution is such that directly generating from this distribution is difficult, perhaps because it is difficult to obtain the denominator of (3.31). However there are many circumstances where it is much easier to obtain the value of the conditional distributions

$$\pi(x|y) = \frac{\pi(x, y)}{\sum_z \pi(z, y)} \text{ and}$$
$$\pi(y|x) = \frac{\pi(x, y)}{\sum_z \pi(x, z)}$$

Now consider the following algorithm: begin with an arbitrary value of y_0 and generated x_1 from the distribution $\pi(x|y_0)$ followed by generating y_1 from the distribution $\pi(y|x_1)$. It is hard to imagine a universe in which iteratively generating values x_{n+1} from the distribution $\pi(x|y_n)$ and then y_{n+1} from the distribution $\pi(y|x_{n+1})$ does not, at least asymptotically as $n \to \infty$, eventually lead to a draw from the joint distribution $\pi(x, y)$. Indeed that is the case since the transition probabilities for this chain are given by

$$P(x_{n+1}, y_{n+1}|x_n, y_n) = \pi(x_{n+1}|y_n)\pi(y_{n+1}|x_{n+1})$$

and it is easy to show directly from these transition probabilities that

$$\begin{aligned}
&\sum_{(x,y)} P(x_1, y_1|x, y)\pi(x, y) \\
&= \pi(y_1|x_1) \sum_y \pi(x_1|y) \sum_x \pi(x, y) \\
&= \pi(y_1|x_1) \sum_y \pi(x_1, y) \\
&= \pi(x_1, y_1).
\end{aligned}$$

Of course the real power of Gibbs Sampling is achieved in problems that are not two-dimensional such as the example above, but have dimension sufficiently high that calculating the sums or integrals in the denominator of expressions like (3.31) is not computationally feasible.

Coupling From the Past: Sampling from the stationary distribution of a Markov Chain

All of the above methods assume that we generate from the stationary distribution of a Markov chain by the "until Hele freezes over" method, i.e. wait until run the chain from an arbitrary starting value and then delete the initial transient. An alternative elegant method that is feasible at least for some finite state Markov chains is the method of "coupling from the past" due to Propp and Wilson (1996).

We assume that we are able to generate transitions in the Markov Chain. In other words if the chain is presently in state i at time n we are able to generate a random variable X_{n+1} from the distribution proportional to $P_{ij}, j = 1, ...K$. Suppose $F(x|i)$ is the cumulative distribution function $P(X_{n+1} \leq x|X_n = i)$ and let us denote its inverse by $F^{-1}(y|i)$. So if we wish to generate a random variable X_{n+1} conditional on X_n, we can use the inverse transform $X_{n+1} = F^{-1}(U_{n+1}|X_n)$ applied to the Uniform[0,1] random variable U_{n+1}. Notice that a starting value say X_{-100} together with the sequence of uniform[0,1] variables

$(U_{-99}, ..., U_0)$ determines the chain completely over the period $-100 \leq t \leq 0$. If we wish to generated the value of X_t given $X_s, s < t$, then we can work this expression backwards

$$\begin{aligned} X_t &= F^{-1}(U_{t-1}|X_{t-1}) \\ &= F^{-1}(U_{t-1}|F^{-1}(U_{t-2}|X_{t-2})) \\ &= F^{-1}(U_{t-1}|F^{-1}(U_{t-2}|...F^{-1}(U_{t-1}|F^{-1}(U_s|i)))) \\ &= F_s^t(X_s), \text{ say.} \end{aligned}$$

Now imagine an infinite sequence $\{U_t, t = ..., -3, -2, -1\}$ of independent uniform[0,1] random variables that was used to generate the state X_0 of a chain at time 0. Let us imagine for the moment that there is a value of M such that $F_{-M}^0(i)$ is a constant function of i. This means that *for this particular draw of uniform random numbers,* whatever the state i of the system at time $-M$, the same state $X_0 = F_{-M}^0(i)$ is generated to time 0. All chains, possibly with different behaviour prior to time $-M$ are "coupled" at time $-M$ and identical from then on. In this case we say that *coalescence* has occurred in the interval $[-M, 0]$. No matter where we start the chain at time $-M$ it ends up in the same state at time 0, so it is quite unnecessary to simulate the chain over the whole infinite time interval $-\infty < t \leq 0$. N*o matter what state is occupied at time* $t = -M$, *the chain ends up in the same state at time* $t = 0$. When coalescence has occurred, 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 $t = -\infty$.

There is sometimes an easy way to check whether coalescence has occurred in an interval, if the state space of the Markov chain is suitably ordered. For example suppose the states are numbered $1, 2, ..., N$. Then it is sometimes possible to relabel the states so that the conditional distribution functions $F(x|i)$ are stochastically ordered, or equivalently that $F^{-1}(U|i)$ is monotonic (say monotonically increasing) in i for each value of U. This is the case for example

provided that the partial sums $\sum_{l=1}^{j} P_{il}$ are increasing functions of i for each $j = 1, 2, ..., N$. If follows that the functions $F^0_{-M}(i)$ are all monotonic functions of i and so

$$F^0_{-M}(1) \leq F^0_{-M}(2) \leq ...F^0_{-M}(N).$$

Therefore, if $F^0_{-M}(1) = F^0_{-M}(N)$, then $F^0_{-M}(i)$ must be a constant function. Notice also that if there is any time in an interval $[s, t]$ at which coalescence occurs so that $F^t_s(i)$ is a constant function of i, then for any interval $[S, T]$ containing it $[S, T] \supset [s, t]$, $F^T_S(i)$ is also a constant function of i.

It is easy to prove that coalescence occurs in the interval $[-M, 0]$ for sufficiently large M. For an ergodic finite Markov chain, there is some step size τ such that every transition has positive probability $P[X_{t+\tau} = j | X_t = i] > \epsilon$ for all i, j. Consider two independent chains, one beginning in state i and the other in state i' at time $t = 0$. Then the probability that they occupy the same state j at time $t = \tau$ is at least ϵ^2. It is easy to see that if we use inverse transform to generate the transitions and if they are driven by common random numbers then this can only increase the probability of being in the same state, so the probability these two chains are coupled at time τ is at least ϵ^2. Similarly for N possible states, the probability of coalescence in an interval of length τ is at least $\varepsilon^N > 0$. Since there are infinitely many intervals disjoint of length τ in $[-\infty, 0]$ and the events that there is a coalescence in each interval are independent, the probability that coalescence occurs somewhere in $[-\infty, 0]$ is 1.

We now detail the Propp Wilson algorithm

1. Set $M = 1, X_U = N, X_L = 1$
2. Generate $U_{-M}....U_{-M/2+1}$ all independent $Uniform[0, 1]$.
3. For $t = -M$ to -1 repeat
   (a) obtain $X_L = F^{-1}(U_{t-1}|X_L)$ and $X_U = F^{-1}(U_{t-1}|X_U)$.

(b) If $X_L = X_U$ stop and output $X(0) = X_L$

4. Otherwise, set $M = 2M$ and go to step 2.

This algorithm tests for coalescence repeatedly by starting on the intervals

$$[-1, 0], [-2, -1], [-4, -2], [-8, -4].$$

We are assured that with probability one, the process will terminate with coalescence after a finite number of steps. Moreover, in this algorithm that the random variable U_t once generated is NOT generated again on a subsequent pass when M is doubled. The generated U_t is reused at each pass until coalescence occurs. If U_t were regenerated on subseuqent passes, this would lead to bias in the algorithm.

It may well be that this algorithm needs to run for a very long time before achieving coalescence and an impatient observer who interrupts the algorithm prior to coalescence and starts over will bias the results. Varous modifications have been made to speed up the algorithm (e.g. Fill, 1998).

Sampling from the Stationary Distribution of a Diffusion Process

A basic Ito process of the form

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

is perhaps the simplest extension of a Markov chain to continuous time, continuous state-space. It is well-known that under fairly simple conditions, there is a unique (strong) solution to this equation and that the limiting distribution of X_T as $T \to \infty$ has stationary distribution with probability density function

$$f(x) = c\frac{1}{\sigma^2(x)} \exp\{2\int_0^x \frac{a(z)}{\sigma^2(z)}dz\}$$

where the constant c is chosen so that the integral of the density is 1. To be able to do this we need to assume that

$$\int_{-\infty}^{\infty} \frac{1}{\sigma^2(x)} \exp\{2\int_0^x \frac{a(z)}{\sigma^2(z)}dz\}dx < \infty. \tag{3.32}$$

In order to generate from this stationary distribution, we can now start the process at some arbitrary value X_0 and run it for a very long time T, hoping that this is sufficiently long that the process is essentially in its stationary state, or try to generate X_0 more directly from (3.32) in which case the process is beginning (and subsequently running) with its stationary distribution.

For an example, let us return to the CIR process

$$dX_t = k(b - X_t)dt + \sigma X_t^{1/2} dW_t. \tag{3.33}$$

In this case

$$a(x) = k(b - x), \text{ for } x > 0,$$
$$\sigma^2(x) = \sigma^2 x, \text{ for } x > 0.$$

Notice that

$$\frac{1}{\sigma^2 x} \exp\{2\int_\varepsilon^x \frac{k(b-z)}{\sigma^2 z}dz\} = \frac{1}{\sigma^2} x^{-1} \exp\{\frac{2kb}{\sigma^2}\ln(x/\varepsilon) - \frac{k}{\sigma^2}(x-\varepsilon)\}$$

is proportional to

$$x^{2kb/\sigma^2 - 1} \exp\{-kx/\sigma^2\}$$

and the integral of this function, a Gamma function, will fail to converge unless $2kb/\sigma^2 - 1 > -1$ or $2kb > \sigma^2$. Under this condition the stationary distribution of the CIR process is Gamma$(2kb/\sigma^2, \frac{\sigma^2}{k})$. If this condition fails and $2kb < \sigma^2$, then the process X_t is absorbed at 0. If we wished to simulate a CIR process in equilibrium, we should generate starting values of X_0 from the Gamma distribution. More generally for a CEV process satisfying

$$dX_t = k(b - X_t)dt + \sigma X_t^{\gamma/2} dW_t \tag{3.34}$$

a similar calculation shows that the stationary density is proportional to

$$x^{-\gamma}\exp\{-\frac{2kb}{\sigma^2}\frac{1}{x^{\gamma-1}(\gamma-1)}-\frac{k}{\sigma^2\gamma}x^{\gamma}\},\ \text{for } \gamma>1.$$

Simulating Stochastic Partial Differential Equations.

Consider a derivative product whose underlying asset has price X_t which follows some model. Suppose the derivative pays an amount $V_0(X_T)$ on the maturity date T. Suppose that the value of the derivative depends only on the current time t and the current value of the asset S, then its current value is the discounted future payoff, an expectation of the form

$$V(S,t)=E[V_0(X_T)exp\{-\int_t^T r(X_v,v)dv\}|X_t=S] \tag{3.35}$$

where $r(X_t,t)$ is the current spot interest rate at time t. In most cases, this expectation is impossible to evaluate analytically and so we need to resort to numerical methods. If the spot interest rate is function of *both arguments* (X_v,v) and not just a function of time, then this integral is over the *whole joint distribution of the process* $X_v, 0<v<T$ and simple one-dimensional methods of numerical integration do not suffice. In such cases, we will usually resort to a Monte-Carlo method. The simplest version requires simulating a number of sample paths for the process X_v starting at $X_t=S$, evaluating $V_0(X_T)exp\{-\int_t^T r(X_v,v)dv\}$ and averaging the results over all simulations. We begin by discussing the simulation of the process X_v required for integrations such as this.

Many of the stochastic models in finance reduce to simple diffusion equation (which may have more than one *factor* or dimension). Most of the models in finance are Markovian in the sense that at any point t in time, the future evolution of the process depends only on the current state X_t and not on the

past behaviour of the process $X_s, s < t$. Consequently we restrict to a "Markov diffusion model" of the form

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t \tag{3.36}$$

with some initial value X_0 for X_t at $t = 0$. Here W_t is a driving standard Brownian motion process. Solving deterministic differential equations can sometimes provide a solution to a specific problem such as finding the arbitrage-free price of a derivative. In general, for more complex features of the derivative such as the distribution of return, important for considerations such as the *Value at Risk*, we need to obtain a solution $\{X_t, 0 < t < T\}$to an equation of the above form which is a stochastic process. Typically this can only be done by simulation. One of the simplest methods of simulating such a process is motivated through a crude interpretation of the above equation in terms of discrete time steps, that is that a small increment $X_{t+h} - X_t$in the process is approximately normally distributed with mean given by $a(X_t, t)h$and variance given by $\sigma^2(X_t, t)h$. We generate these increments sequentially, beginning with an assumed value for X_0, and then adding to obtain an approximation to the value of the process at discrete times $t = 0, h, 2h, 3h, \ldots$. Between these discrete points, we can linearly interpolate the values. Approximating the process by assuming that the conditional distribution of $X_{t+h} - X_t$ is $N(a(X_t, t)h, \sigma^2(X_t, t)h)$ is called *Euler's method* by analogy to a simple method by the same name for solving ordinary differential equations. Given simulations of the process satisfying (3.36) together with some initial conditions, we might average the returns on a given derivative for many such simulations, (provided the process is expressed with respect to the risk-neutral distribution), to arrive at an arbitrage-free return for the derivative.

In this section we will discuss the numerical solution, or simulation of the solution to stochastic differential equations.

Letting $t_i = i\Delta x$, Equation (3.36) in integral form implies

$$X_{t_{i+1}} = X_{t_i} + \int_{t_i}^{t_{i+1}} a(X_s, s)ds + \int_{t_i}^{t_{i+1}} \sigma(X_s, s)dW_s \tag{3.37}$$

For the following lemma we need to introduce O_p or "order in probability", notation common in mathematics and probability. A sequence indexed by Δt, say $Y_{\Delta t} = O_p(\Delta t)^k$ means that when we divide this term by $(\Delta t)^k$ and then let $\Delta t \to 0$, the resulting sequence is bounded in probability or that for each ε there exists $K < \infty$ so that

$$P[|\frac{Y_{\Delta t}}{\Delta t^k}| > K] < \varepsilon$$

whenever $|\Delta t| < \varepsilon$. As an example, if W is a Brownian motion, then $\Delta W_t = W(t+\Delta t) - W(t)$ has a Normal distribution with mean 0 and standard deviation $\sqrt{\Delta t}$ and is therefore $O_p(\Delta t)^{1/2}$. Similarly Then we have two very common and useful approximations to a diffusion given by the following lemma.

Lemma 32 *If X_t satisfies a diffusion equation of the form (3.36) then*

$$X_{t_{i+1}} = X_{t_i} + a(X_{t_i}, t_i)\Delta t + \sigma(X_{t_i}, t_i)\Delta W_t + O_p(\Delta t) \quad \textit{(Euler approximation)}$$

$$X_{t_{i+1}} = X_{t_i} + a(X_{t_i}, t_i)\Delta t + \sigma(X_{t_i}, t_i)\Delta W_t + \frac{\sigma(X_{t_i}, t_i)\frac{\partial}{\partial x}\sigma(X_{t_i}, t_i)}{2}[(\Delta W_t)^2 - \Delta t] + O_p(\Delta t)^{3/2} \quad \textit{(Milstei}$$

Proof. Ito's lemma can be written in terms of two operators on functions f for which the derivatives below exist;

$$df(X_t, t) = L^o f dt + L^1 f dW_t \quad \text{where}$$

$$L^0 = a\frac{\partial}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial t}, \text{ and}$$
$$L^1 = \sigma\frac{\partial}{\partial x}.$$

Integrating, this and applying to twice differentiable functions a and σ and $s > t_i$,

$$a(X_s, s) = a(X_{t_i}, t_i) + \int_{t_i}^{s} L^0 a(X_u, u)du + \int_{t_i}^{s} L^1 a(X_u, u)dW_u$$
$$\sigma(X_s, s) = \sigma(X_{t_i}, t_i) + \int_{t_i}^{s} L^0 \sigma(X_u, u)du + \int_{t_i}^{s} L^1 \sigma(X_u, u)dW_u.$$

By substituting in each of the integrands in 3.37 using the above identity and iterating this process we arrive at the Ito-Taylor expansions (e.g. Kloeden and Platen, 1992). For example,

$$\int_{t_i}^{t_{i+1}} a(X_s, s)ds = \int_{t_i}^{t_{i+1}} \{a(X_{t_i}, t_i) + \int_{t_i}^{s} L^0 a(X_u, u)du + \int_{t_i}^{s} L^1 a(X_u, u)dW_u\}ds$$
$$\approx a(X_{t_i}, t_i)\Delta t + L^0 a(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} duds + L^1 a(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dW_u ds$$

The first term $a(X_{t_i}, t_i)\Delta t$, is an initial approximation to the desired integral and the rest is a lower order correction that we may regard as an error term for the moment. For example it is easy to see that the second term $L^0 a(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} duds$ is $O_p(\Delta t)^2$ because the integral $\int_{t_i}^{t_{i+1}} \int_{t_i}^{s} duds = (\Delta t)^2/2$ and $L^0 a(X_{t_i}, t_i)$ is bounded in probability. The third term $L^1 a(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dW_u ds$ is $O_p(\Delta t)^{3/2}$ since $\int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dW_u ds = \int_{t_i}^{t_{i+1}} (t_{i+1} - u)dW_u$ and this is a normal random variable with mean 0 and variance $\int_{t_i}^{t_{i+1}} (t_{i+1} - u)^2 du = (\Delta t)^3/3$. We can write such a normal random variable as $3^{-1/2}(\Delta t)^{3/2} Z$ for Z a standard normal random variable and so this is obviously $O_p(\Delta t)^{3/2}$. Thus the simplest *Euler approximation* to the distribution of the increment assumes that ΔX has conditional mean $a(X_{t_i}, t_i)\Delta t$. Similarly

$$\int_{t_i}^{t_{i+1}} \sigma(X_s, s)dW_s = \int_{t_i}^{t_{i+1}} \{\sigma(X_{t_i}, t_i) + \int_{t_i}^{s} L^0 \sigma(X_u, u)du + \int_{t_i}^{s} L^1 \sigma(X_u, u)dW_u\}dW_s$$
$$\approx \sigma(X_{t_i}, t_i)\Delta W_t + L^0 \sigma(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dudW_s + L^1 \sigma(X_{t_i}, t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dW_u dW_s$$
$$= \sigma(X_{t_i}, t_i)\Delta W_t + \frac{\sigma(X_{t_i}, t_i)\frac{\partial}{\partial x}\sigma(X_{t_i}, t_i)}{2}[(\Delta W_t)^2 - \Delta t] + O_p(\Delta t)^{3/2}$$

since $\int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dW_u dW_s = \frac{1}{2}[(\Delta W_t)^2 - \Delta t]$, $L^0 \sigma(X_u, u) = \sigma(X_{t_i}, t_i) + O_p(\Delta t)^{1/2}$, $L^1 \sigma(X_u, u) = \sigma(X_u, u)\frac{\partial}{\partial x}\sigma(X_u, u)$ and $\int_{t_i}^{t_{i+1}} \int_{t_i}^{s} dudW_s = O_p(\Delta t)^{3/2}$. Putting

these terms together, we arrive at an approximation to the increment of the form

$$\Delta X_t = a(X_{t_i}, t_i)\Delta t + \sigma(X_{t_i}, t_i)\Delta W_t + \frac{\sigma(X_{t_i}, t_i)\frac{\partial}{\partial x}\sigma(X_{t_i}, t_i)}{2}[(\Delta W_t)^2 - \Delta t] + O_p(\Delta t)^{3/2} \tag{3.38}$$

which allow an explicit representation of the increment in the process X in terms of the increment of a Brownian motion process $\Delta W_t \sim N(0, \Delta t)$. ■

The approximation (3.38) is called the *Milstein approximation,* a refinement of the first, the *Euler approximation.* It is the second Ito-Taylor approximation to a diffusion process. Obviously, the increments of the process are quadratic functions of a normal random variable and are no longer normal. The error approaches 0 at the rate $O_p(\Delta t)^{3/2}$ *in probability* only. This does not mean that the trajectory is approximated to this order but that the difference between the Milstein approximation to a diffusion and the diffusion itself is bounded in probability when divided by $(\Delta t)^{3/2}$ and as we let $\Delta t \to 0$. Higher order Taylor approximations are also possible, although they grow excessively complicated very quickly. See the book by Kloeden and Platten(1992) for details.

There remains the question of how much difference it makes which of these approximations we employ for a particular diffusion. Certainly there is no difference at all between the two approximations in the case that the diffusion coefficient $\sigma(X_t, t)$ does not depend at all on X_t. In general, the difference is hard to assess but in particular cases we can at least compare the performance of the two methods. The approximations turn out to be very close in most simple cases. For example consider the stock price path in Figure 3.27. The dashed line corresponds to a Milstein approximation whereas the piecewise continuous line corresponds to the Euler approximation. In this case the Milstein appears to be a little better, but if I run a number of simulations and compare the sum of the squared errors (i.e. squared differences between the approximate value of X_t and the true value of X_t) we find that the improvement is only about

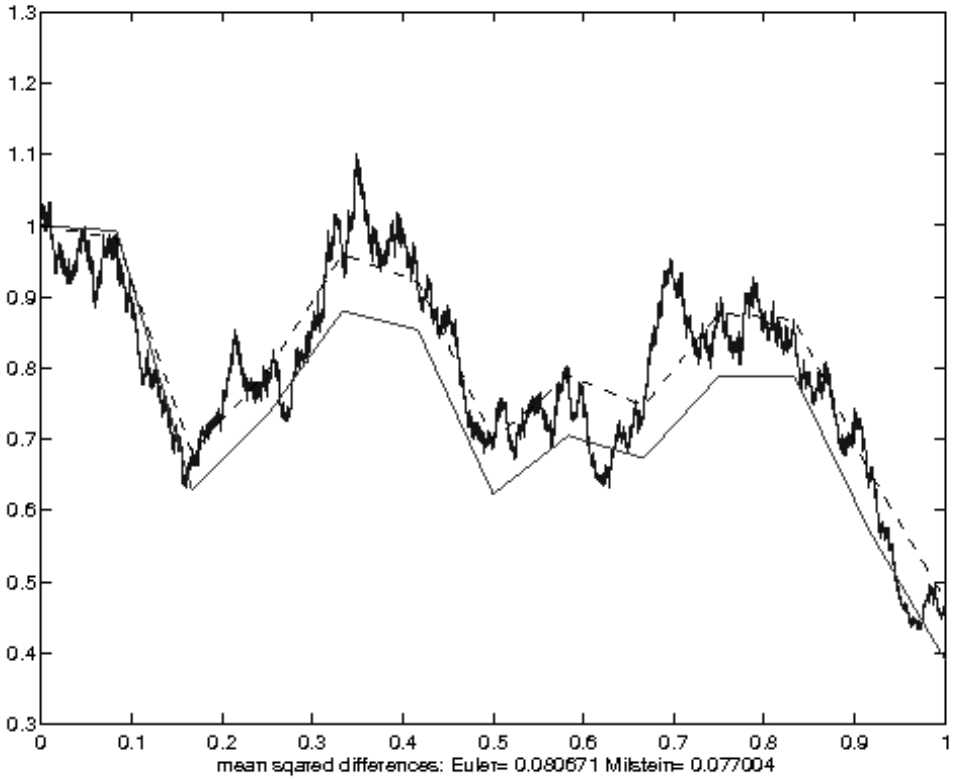


Figure 3.27: Comparison of Milstein and Euler approximation to stock with $\Delta t = 1/12$ year.

two percent of the difference. The same is true even if I change the value of Δt from 1/12 (i.e. one month) to 1/52 (i.e. one week). Unlike the behaviour of higher order approximations to deterministic functions, there appears to be little advantage in using a higher order approximation, at least in the case of diffusions with smooth drift and diffusion coefficients.

We can compare using Milstein approximation on the original process and using Euler's approximation on a transformation of the process in the case that the diffusion term depends only on the state of the process (not time). In other

words, suppose we have an Ito process of the form

$$dX_t = a(X_t, t)dt + \sigma(X_t)dW_t \tag{3.39}$$

where W_t is an ordinary Wiener measure. A simple transformation reduces this to a problem with constant diffusion term. Suppose $\sigma(x) > 0$ for all x and let

$$s(x) = \int_0^x \frac{1}{\sigma(z)}dz, \text{ for } x \geq 0$$
$$s(x) = -\int_x^0 \frac{1}{\sigma(z)}dz \text{ for } x < 0$$

where we assume these integrals are well defined. Let g be the inverse function of s. This inverse exists since the function is continuous monotonically increasing. Suppose we apply Ito's lemma to the transformed process $Y_t = s(X_t)$. We obtain

$$\begin{aligned} dY_t &= \{a(X_t, t)s'(X_t) + \frac{1}{2}\sigma^2(X_t)s''(X_t)\}dt + \sigma(X_t)s'(X_t)dW_t \\ &= \{\frac{a(X_t, t)}{\sigma(X_t)} - \frac{1}{2}\sigma^2(X_t)\frac{\sigma'(X_t)}{\sigma^2(X_t)}\}dt + dW_t \\ &= \mu(Y_t, t)dt + dW_t \end{aligned}$$

where

$$\mu(Y_t, t) = \frac{a(g(Y_t), t)}{\sigma(g(Y_t))} - \frac{1}{2}\sigma'(g(Y_t)).$$

In other words, Yt satisfies an Ito equation with constant diffusion term. Suppose we generate an increment in Y_t using Euler's method and then solve for the corresponding increment in X_t.Then using the first two terms in the Taylor series expansion of g,

$$\begin{aligned} \Delta X_t &= g'(Y_t)\Delta Y_t + \frac{1}{2}g''(Y_t)(\Delta Y_t)^2 \\ &= g'(Y_t)(\mu(Y_{t_i}, t_i)\Delta t + \Delta W_t) + \frac{1}{2}\sigma'(g(Y_t))\sigma(g(Y_t))(\Delta Y_t)^2 \\ &= \{a(g(Y_t), t) - \frac{1}{2}\sigma(g(Y_t))\sigma'(g(Y_t))\}\Delta t + \sigma(g(Y_t))\Delta W_t + \frac{1}{2}\sigma'(g(Y_t))\sigma(g(Y_t))(\Delta Y_t)^2 \end{aligned}$$

since

$$g'(Y_t) = \frac{1}{s'(g(Y_t))} = \sigma(g(Y_t)) \text{ and}$$
$$g''(Y_t) = \sigma'(g(Y_t))\sigma(g(Y_t)).$$

But since $(\Delta Y_t)^2 = (\Delta W_t)^2 + o(\Delta t)$ it follows that

$$\Delta X_t = \{a(X_t,t) - \frac{1}{2}\sigma(X_t)\sigma'(X_t)\}\Delta t + \sigma(X_t)\Delta W_t + \frac{1}{2}\sigma'(X_t)\sigma(X_t)(\Delta W_t)^2 + o(\Delta t)$$

and so the approximation to this increment is identical, up to the order considered, to the Milstein approximation. For most processes, it is preferable to apply a diffusion stabilizing transformation as we have here, prior to discretizing the process. For the geometric Brownian motion process, for example, the diffusion-stabilizing transformation is a multiple of the logarithm, and this transforms to a Brownian motion, for which the Euler approximation gives the exact distribution.

Example: Down-and-out-Call.

Consider an asset whose price under the risk-neutral measure Q follows a constant elasticity of variance (CEV) process

$$dS_t = rS_t dt + \sigma S_t^{\gamma} dW_t \tag{3.40}$$

for a standard Brownian motion process W_t. A down-and-out call option with exercise price K provides the usual payment $(S_T - K)^+$ of a European call option on maturity T if the asset never falls below a given *out barrier* b. The parameter $\gamma > 0$ governs the change in the diffusion term as the asset price changes. We wish to use simulation to price such an option with current asset price S_0, time to maturity T, out barrier $b < S_0$ and constant interest rate r and compare with the Black-Scholes formula as $b \to 0$.

A geometric Brownian motion is most easily simulated by taking logarithms.

For example if S_t satisfies the risk-neutral specification

$$dS_t = rS_t dt + \sigma S_t dW_t \tag{3.41}$$

then $Y_t = \log(S_t)$ satisfies

$$dY_t = (r - \sigma^2/2)dt + \sigma dW_t. \tag{3.42}$$

This is a Brownian motion and is simulated with a normal random walk. Independent normal increments are generated $\Delta Y_t \sim N((r - \sigma^2/2)\Delta t, \sigma^2 \Delta t)$ and their partial sums used to simulate the process Y_t. The return for those options that are *in the money* is the average of the values of $(e^{Y_T} - E)^+$ over those paths for which $\min\{Y_s; t < s < T\} \geq \ln(b)$. Similarly the transformation of the CEV process which provides a constant diffusion term is determined by

$$\begin{aligned} s(x) &= \int_0^x \frac{1}{\sigma(z)} dz \\ &= \int_0^x z^{-\gamma} dz = \begin{cases} \frac{x^{1-\gamma}}{1-\gamma} & \text{if} \quad \gamma \neq 1 \\ \ln(x) & \text{if} \quad \gamma = 1 \end{cases} . \end{aligned}$$

Assuming $\gamma \neq 1$, the inverse function is

$$g(y) = cy^{1/(1-\gamma)}$$

for constant c and the process $Y_t = (1-\gamma)^{-1} S_t^{1-\gamma}$ satisfies an Ito equation with constant diffusion coefficient;

$$\begin{aligned} dY_t &= \{\frac{r}{\sigma} S_t^{1-\gamma} - \frac{1}{2}\gamma\sigma S_t^{\gamma-1}\}dt + dW_t \\ dY_t &= \{\frac{r}{\sigma}(1-\gamma)Y_t - \frac{\gamma\sigma}{2(1-\gamma)Y_t}\}dt + dW_t. \end{aligned} \tag{3.43}$$

After simulating the process Y_t we invert the relation to obtain $S_t = ((1-\gamma)Y_t)^{1/(1-\gamma)}$. There is one fine point related to simulating the process (3.43) that we implemented in the code below. The equation (3.40) is a model for a non-negative asset price S_t but when we simulate the values Y_t from (3.43) there is nothing to prevent the process from going negative. Generally if $\gamma \geq 1/2$

and if we increment time in sufficiently small steps Δt, then it is *unlikely* that a negative value of Y_t will obtain, but when it does, we assume absorption at 0 (analogous to default or bankruptcy). The following *Matlab* function was used to simulate sample paths from the CEV process over the interval $[0, T]$.

```
function s=simcev(n,r,sigma,So,T,gam)
% simulates n sample paths of a CEV process on the interval [0,T] all with
% the same starting value So. assume gamma != 1.
Yt=ones(n,1)*(So^(1-gam))/(1-gam);  y=Yt;
dt=T/1000; c1=r*(1-gam)/sigma; c2=gam*sigma/(2*(1-gam));
dw=normrnd(0,sqrt(dt),n,1000);
for i=1:1000
    v=find(Yt);   % selects positive components of Yt for update
    Yt=max(0,Yt(v)+(c1.*Yt(v)-c2./Yt(v))*dt+dw(v,i));
    y=[y Yt];
end
s=((1-gam)*max(y,0)).^(1/(1-gam));  %transforms to St
```

For example when $r = .05, \sigma = .2, \Delta t = .00025, T = .25$, $\gamma = 0.8$ we can generate 1000 sample paths with the command

```
s=simcev(1000,.05,.2,10,.25,.8);
```

In order to estimate the price of a barrier option with a down-and-out barrier at b and exercise price K, capture the last column of s,

```
ST=s(:,1001);
```

then value a European call option based on these sample paths

```
v=exp(-r*T)*max(ST-K,0);
```

finally setting the values equal to zero for those paths which breached the lower barrier and then averaging the return from these 1000 replications;

```
v(min(s')<=9)=0;
mean(v);
```

which results in an estimated value for the call option of around \$0.86. Although the standard error is still quite large (0.06), we can compare this with the Black-Scholes price with similar parameters. [CALL,PUT] = BLSPRICE(10,10,.05,.25,.2,0) which gives a call option price of \$0.4615. Why such a considerable difference? Clearly the down-and-out barrier can only reduce the value of a call option. Indeed if we remove the down-and-out feature, the European option is valued closer to \$1.28 so the increase must be due to the differences betwen the CEV process and the geometric Brownian motion. We can confirm this by simulating the value of a barrier option in the Black_Scholes model later on.

Problems

1. Consider the mixed generator $x_n = (ax_{n-1} + 1)mod(m)$ with $m = 64$. What values of a results in the maximum possible period. Can you indicate which generators appears more and less random?

2. Consider a shuffled generator described in Section 3.2 with $k = 3, m_1 = 7, m_2 = 11$.

   Determine the period of the shuffled random number generator above and compare with the periods of the two constituent generators.

3. Consider the quadratic residue generator $x_{n+1} = x_n^2 \bmod m$ with $m = 4783 \times 4027$. Write a program to generate pseudo-random numbers from this generator. Use this to determine the period of the generator starting with seed $x_0 = 196$, and with seed $x_0 = 400$.

4. Consider a sequence of independent $U[0,1]$ random variables $U_1, ..., U_n$. Define indicator random variables

   $S_i = 1$ if $U_{i-1} < U_i$ and $U_i > U_{i+1}$ for $i = 2, 3, ..., n-1$, otherwise $S_i = 0$,

   $T_i = 1$ if $U_{i-1} > U_i$ and $U_i < U_{i+1}$ for $i = 2, 3, ..., n-1$, otherwise $T_i = 0$.

   Verify the following:

   (a)
   $$R = 1 + \sum (S_i + T_i)$$

   (b)
   $$E(T_i) = E(S_i) = \frac{1}{3} \text{ and } E(R) = \frac{2n-1}{3}$$

   (c) $cov(T_i, T_j) = cov(S_i, S_j) = -\frac{1}{9}$ if $|i-j| = 1$ and it equals 0 if $|i-j| > 1$.

   (d) $cov(S_i, T_j) = \begin{cases} \frac{5}{24} - \frac{1}{9} = \frac{7}{72} & \text{if} \quad |i-j| = 1 \\ -\frac{1}{9} & \text{if} \quad i = j \\ 0 & \text{if} \quad |i-j| > 1 \end{cases}$.

   (e) $var(R) = 2(n-2)\frac{1}{3}(\frac{2}{3}) + 4(n-3)(-\frac{1}{9}) + 4(n-3)(\frac{7}{72}) + 2(n-2)(-\frac{1}{9}) = \frac{3n-5}{18}$.

   (f) Confirm these formulae for mean and variance of R in the case $n = 3, 4$.

5. Generate 1000 daily "returns" $X_i, i = 1, 2, ..., 1000$ from each of the two distributions, the Cauchy and the logistic. Choose the parameters so that the median is zero and $P[|X_i| < .06] = .95$. Graph the total return over an n day period versus n. Is there a qualitative difference in the two graphs? Repeat with a graph of the daily return averaged over days $1, 2, ..., n$.

6. Consider the linear congruential generator

   $$x_{n+1} = (ax_n + c) \bmod 2^8$$

What is the maximal period that this generator can achieve when $c = 1$ and for what values of a does this seem to be achieved? Repeat when $c = 0$.

7. Let U be a uniform random variable on the interval [0,1]. Find a function of U which is uniformly distributed on the interval [0,2]. Repeat for the interval $[a, b]$.

8. Evaluate the following integral by simulation:

$$\int_0^2 x^{3/4}(4-x)^{1/3}dx.$$

9. Evaluate the following integral by simulation:

$$\int_{-\infty}^{\infty} e^{-x^4}dx.$$

(Hint: Rewrite this integral in the form $2\int_0^{\infty} e^{-x^4}dx$ and then change variables to $y = x/(1+x)$)

10. Evaluate the following integral by simulation:

$$\int_0^1 \int_0^1 e^{(x+y)^4}dxdy.$$

(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).

11. Find the covariance $cov(e^U, e^{-U})$ by simulation where U is uniform[0,1] and compare the simulated value to the true value. Compare the actual error with the standard error of your estimator.

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

Estimate $E(N)$ by simulation. Repeat for larger and larger numbers of simulations. Guess on the basis of these simulations what is the value of $E(N)$. Can you prove your hypothesis concerning the value of $E(N)$?

13. Give an algorithm for generating observations from a distribution which has cumulative distribution function $F(x) = \frac{x+x^3+x^5}{3}, 0 < x < 1$. Record the time necessary to generate the sample mean of 100,000 random variables with this distribution. (Hint: Suppose we generate X_1 with cumulative distribution function $F_1(x)$ and X_2 with cumulative distribution function $F_2(x)$, X_3 with cumulative distribution function $F_3(x)$ We then generate $J = 1, 2,$ or 3 such that $P[J = j] = p_j$ and output the value X_J. What is the cumulative distribution function of the random variable output?)

14. Consider independent random variables X_i $i = 1, 2, 3$ with cumulative distribution function

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

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

(Hint: consider the cumulative distribution function of the minimum and maximum).

15. Suppose we wish to estimate a random variable X having cumulative distribution function $F(x)$ 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 cumulative distribution function $F(x)$.

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

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

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

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

18. Obtain generators for the following distributions:

    (a) *Rayleigh*

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, x \geq 0 \tag{3.44}$$

    (b) *Triangular*

$$f(x) = \frac{2}{a}(1 - \frac{x}{a}), 0 \leq x \leq a \tag{3.45}$$

19. 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* $[0, 2\pi]$.

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

$$(X, Y) = R(cos\Theta, sin\Theta) \tag{3.46}$$

    where we use a random number generator with poor lattice properties such as the generator $x_{n+1} = (383x_n + 263) \bmod 10000$ to generate our uniform random numbers. Use this generator together with the Box-Mueller algorithm to generate 5,000 pairs of independent random normal numbers. Plot the results. Do they appear independent?

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

$$g(x|\eta, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(logx - log\eta + \sigma^2/2)^2/2\sigma^2\}. \tag{3.47}$$

22. (*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),* or in **R**, A= chol(sigma, pivot = FALSE, LINPACK = pivot) 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.

23. *(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?

24. Suppose interest rates follow the constant elasticity of variance process of the form

$$dr_t = k(b - r_t) + \sigma |r_t|^\gamma dW_t$$

for parameters value $\gamma, b, k > 0$. For various values of the parameters k, γ and for $b = 0.04$ use both Euler and Milsten to generate paths from this process. Draw conclusions about the following:

(a) When does the marginal distribution of r_t appear to approach a steady state solution. Plot the histogram of this steady state distribution.

(b) Are there simulations that result in a negative value of r? How do you rectify this problem?

(c) What does the parameter σ represent? Is it the annual volatility of the process?

25. Consider a sequence of independent random numbers $X_1, X_2, ...$with a continuous distribution and let M be the first one that is less than its predecessor:

$$M = \min\{n; X_1 \leq X_2 \leq ... \leq X_{n-1} > X_n\}$$

(a) Use the identity $E(M) = \sum_{n=0}^{\infty} P[M > n\}$ to show $E(M) = e$.

(b) Use 100,000 simulation runs and part a to estimate e with a 95% confidence interval.

(c) How many simulations are required if you wish to estimate e within 0.005 (using a 95% confidence interval)?

Chapter 4

Variance Reduction Techniques

Introduction

In this chapter we discuss techniques for improving on the speed and efficiency of a simulation, usually called "variance reduction techniques".

Much of the simulation literature concerns *discrete event simulations* (DES), simulations of systems that are assumed to change instantaneously in response to sudden or discrete events. These are the most common in operations research and examples are simulations of processes such as networks or queues. Simulation models in which the process is characterized by a state, with changes only at discrete time points are DES. In modeling an inventory system, for example, the arrival of a batch of raw materials can be considered as an event which precipitates a sudden change in the state of the system, followed by a demand some discrete time later when the state of the system changes again. A system driven by differential equations in continuous time is an example of a DES because the changes occur continuously in time. One approach to DES is *future event*

simulation which schedules one or more future events at a time, choosing the 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 with binomial or trinomial trees.

Often 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, customer waiting times for a bank teller scheduling system, delays or throughput for computer networks, response times for the location of fire stations or supply depots, etc. Performance measures again are important in engineering examples or in operations research, but less common in finance. They may be used to calibrate a simulation model, however. For example our performance measure might be the average distance between observed option prices on a given stock and prices obtained by simulation from given model parameters. In all cases, the *performance measure* is usually the expected value of a complicated function of many variables, often expressible only by a computer program with some simulated random variables as input. Whether these input random variables are generated by inverse transform, or acceptance-rejection or some other method, they are ultimately a function of uniform[0,1] random variables $U_1, U_2, ...$ These uniform random variables determine such quantities as the normally distributed increments of the logarithm of the stock price. In summary, the simulation is used simply to estimate a multidimensional integral of the form

$$E(g(U_1, ..., U_d)) = \int \int .. \int g(u_1, u_2, ...u_d) du_1 du_2 \dots du_d \tag{4.1}$$

over the unit cube in d dimensions where often d is large.

As an example in finance, suppose that we wish to price a European option on a stock price under the following *stochastic volatility* model.

Example 33 *Suppose the daily asset returns under a risk-neutral distribution is assumed to be a variance mixture of the Normal distribution, by which we mean that the variance itself is random, independent of the normal variable and follows a distribution with moment generating function* $s(s)$. *More specifically assume under the* Q *measure that the stock price at time* $n\Delta t$ *is determined from*

$$S_{(n+1)\Delta t} = S_{n\Delta t}\frac{\exp\{r\Delta t + \sigma_{n+1}Z_{n+1}\}}{m(\frac{1}{2})}$$

where, under the risk-neutral distribution, the positive random variables σ_i^2 *are assumed to have a distribution with moment generating function* $m(s) = E\{\exp(s\sigma_i)\}$, Z_i *is standard normal independent of* σ_i^2 *and both* (Z_i, σ_i^2) *are independent of the process up to time* $n\Delta t$. *We wish to determine the price of a European call option with maturity* T, *and strike price* K.

It should be noted that the rather strange choice of $m(\frac{1}{2})$ in the denominator above is such that the discounted process is a martingale, since

$$\begin{aligned} E\left[\frac{\exp\{\sigma_{n+1}Z_{n+1}\}}{m(\frac{1}{2})}\right] &= E\{E\left[\frac{\exp\{\sigma_{n+1}Z_{n+1}\}}{m(\frac{1}{2})}|\sigma_{n+1}\right]\} \\ &= E\{\frac{\exp\{\sigma_{n+1}^2/2\}}{m(\frac{1}{2})}\} \\ &= 1. \end{aligned}$$

There are many ways of simulating an option price in the above example, some much more efficient than others. We might, for example, simulate all of the $2n$ random variables $\{\sigma_i, Z_i, i = 1, ..., n = T/\Delta t\}$ and use these to determine the simulated value of S_T, finally averaging the discounted payoff from the option in this simulation, i.e. $e^{-rT}(S_T - K)^+$. The price of this option at time 0 is the average of many such simulations (say we do this a total of N times) discounted to present,

$$\overline{e^{-rT}(S_T - K)^+}$$

where $\overline{x}$ denotes the average of the $x's$ observed over all simulations. This is

a description of a crude and inefficient method of conducting this simulation. Roughly the time required for the simulation is proportional to $2Nn$, the total number of random variables generated. This chapter discusses some of the many improvements possible in problems like this. Since each simulation requires at least $d = 2n$ independent uniform random variables to generate the values $\{\sigma_i, Z_i, i = 1, ..., n\}$ then we are trying to estimate a rather complicated integral of the form 4.1 of high dimension d. In this case, however, we can immediately see some obvious improvements. Notice that we can rewrite S_T in the form

$$S_T = S_0 \frac{\exp\{rT + \sigma Z\}}{m^n(\frac{1}{2})} \tag{4.2}$$

where the random variable $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ has moment generating function $m^n(s)$ and Z is independent standard normal. Obviously, if we can simulate σ directly, we can avoid the computation involved in generating the individual σ_i. Further savings are possible in the light of the Black-Scholes formula which provides the price of a call option when a stock price is given by (4.2) and the volatility parameter σ is non-random. Since the expected return from the call under the risk-neutral distribution can be written, using the Black-Scholes formula,

$$E(e^{-rT}(S_T - K)^+) = E\{E[e^{-rT}(S_T - K)^+|\sigma]\}$$
$$= e^{-rT} E\{S_0 \Phi(\frac{\log(S_0/K) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}) - Ke^{-rT}\Phi(\frac{\log(S_0/K) + (r - \frac{\sigma^2}{2})T}{\sigma\sqrt{T}})\}$$

which is now a one-dimensional integral over the distribution of σ. This can now be evaluated either by a one-dimensional numerical integration or by repeatedly simulating the value of σ and averaging the values of

$$e^{-rT} S_0 \Phi(\frac{\log(S_0/K) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}) - Ke^{-rT}\Phi(\frac{\log(S_0/K) + (r - \frac{\sigma^2}{2})T}{\sigma\sqrt{T}})$$

obtained from these simulations. As a special case we might take the distribution of σ_i^2 to be Gamma$(\alpha\Delta t, \beta)$ with moment generating function

$$m(s) = \frac{1}{(1 - \beta s)^{\alpha\Delta t}}$$

in which case the distribution of σ^2 is Gamma$(\alpha T, \beta)$. This is the so-called "variance-gamma" distribution investigated extensively by ..... and originally suggested as a model for stock prices by .....Alternatively many other wider-tailed alternatives to the normal returns model can be written as a variance mixture of the normal distribution and option prices can be simulated in this way. For example when the variance is generated having the distribution of the *reciprocal of a gamma* random variable, the returns have a student's t distribution. Similarly, the stable distributions and the Laplace distribution all have a representation as a variance mixture of the normal.

The rest of this chapter discusses "variance reduction techniques" such as the one employed above for evaluating integrals like (4.1), beginning with the much simpler case of an integral in one dimension.

Variance reduction for one-dimensional Monte-Carlo Integration.

We wish to evaluate a one-dimensional integral $\int_0^1 f(u)du$, which we will denote by θ using by Monte-Carlo methods. We have seen before that whatever the random variables that are input to our simulation program they are usually generated using uniform[0,1] random variables U so without loss of generality we can assume that the integral is with respect to the uniform[0,1] probability density function, i.e. we wish to estimate

$$\theta = E\{f(U)\} = \int_0^1 f(u)du.$$

One simple approach, called *crude Monte Carlo* is to randomly sample $U_i \sim Uniform[0,1]$ and then average the values of $f(U_i)$ obtain

$$\hat{\theta}_{CR} = \frac{1}{n}\sum_{i=1}^{n} f(U_i).$$

It is easy to see that $E(\hat{\theta}_{CR}) = \theta$ so that this average is an *unbiased estimator* of the integral and the variance of the estimator is

$$var(\hat{\theta}_{CR}) = var(f(U_1))/n.$$

Example 34 *A crude simulation of a call option price under the Black-Scholes model:*

For a simple example that we will use throughout, consider an integral used to price a call option. We saw in Section 3.8 that if a European option has payoff $V(S_T)$ *where* S_T *is the value of the stock at maturity* T*, then the option can be valued at present* $(t = 0)$ *using the discounted future payoff from the option under the risk neutral measure;*

$$e^{-rT}E[V(S_T)] = e^{-rT}E[V(S_0e^X)]$$

where, in the Black-Scholes model, the random variable $X = \ln(S_T/S_0)$ *has a normal distribution with mean* $rT - \sigma^2T/2$ *and variance* σ^2T*. A normally distributed random variable* X *can be generated by inverse transform and so we can assume that* $X = \Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2T)$ *is a function of a uniform*$[0, 1]$ *random variable* U *where* $\Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2T)$ *is the inverse of the normal* $(rT - \sigma^2T/2,\ \sigma^2T)$ *cumulative distribution function. Then the value of the option can be written as an expectation over the distribution of the uniform random variable* U*,*

$$E\{f(U)\} = \int_0^1 f(u)du$$

$$\text{where } f(u) = e^{-rT}V(S_0 \exp\{\Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2T)\})$$

This function is graphed in Figure 4.1 in the case of a simple call option with strike price K*, with payoff at maturity* $V(S_T) = (S_T - K)^+$*, the current stock price* $S_0 = \$10$*, the exercise price* K *is* $\$10$*, the annual interest rate* $r = 5\%$*, the maturity is three months or one quarter of year* $T = 0.25$*, and the annual volatility* $\sigma = 0.20$*.*

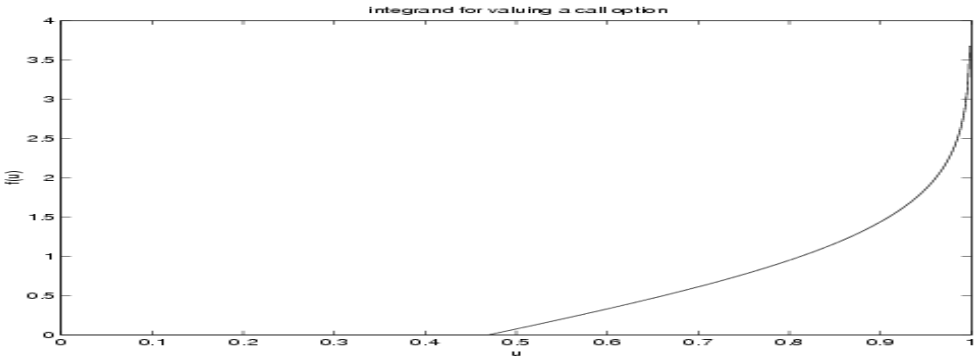


Figure 4.1: The function $f(u)$ whose integral provides the value of a call option

A simple crude Monte Carlo estimator corresponds to evaluating this function at 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 accepts a vector of inputs $u = (U_1, ..., U_n)$ assumed to be Uniform[0,1], outputs the values of $f(U_1), ...f(U_n)$ which can be averaged to give $\hat{\theta}_{CR} = \frac{1}{n}\sum_{i=1}^{n} f(U_i)$.

function v=fn(u)

% value of the integrand for a call option with exercise price ex, r=annual interest rate,

%sigma=annual vol, S0=current stock price.

% u=vector of uniform (0,1) inputs to

%generate normal variates by inverse transform. T=maturity

S0=10 ;K=10;r=.05; sigma=.2 ;T=.25 ; % Values of parameters

ST=S0*exp(norminv(u,r*T-sigma^2*T/2,sigma*sqrt(T)));

% ST $=S_0 \exp\{\Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2 T)\}$ is stock price at time T

v=exp(-r*T)*max((ST-ex),0); % v is the discounted to present payoffs from the call option

and the analogous function in **R,**

fn<-function(u,So,strike,r,sigma,T){

value of the integrand for a call option with exercise price=strike, r=annual interest rate,

sigma=annual volatility, So=current stock price, u=uniform (0,1) input to generate normal variates

by inverse transform. T=time to maturity. For Black-Scholes price, integrate over (0,1).

x<-So*exp(qnorm(u,mean=r*T-sigma^2*T/2,sd=sigma*sqrt(T)))

v<-exp(-r*T)*pmax((x-strike),0)

v}

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=rand(1,500000); mean(fn(u))

and in **R**,

mean(fn(runif(500000),So=10,strike=10,r=.05,sigma=.2,T=.25))

and this provides an approximate value of the option of $\hat{\theta}_{CR} = 0.4620$. The standard error of this estimator, computed using the formula (**??**) below, is around $\sqrt{8.7 \times 10^{-7}}$. We may confirm with the black-scholes formula, again in *Matlab*,

[CALL,PUT] = BLSPRICE(10,10,0.05,0.25,0.2,0).

The arguments are, in order $(S_0.K, r, T, \sigma, q)$ where the last argument (here $q = 0$) is the annual dividend yield which we assume here to be zero. Provided that no dividends are paid on the stock before the maturity of the option, this is reasonable. This Matlab command provides the result *CALL = 0.4615 and* PUT = 0.3373 indicating that our simulated call option price was reasonably accurate- out by 1 percent or so. The *put option* is an option to sell the stock at the specified price \$10 at the maturity date and is also priced by this same function.

One of the advantages of Monte Carlo methods over numerical techniques is that, because we are using a sample mean, we have a simple estimator of accuracy. In general, when n simulations are conducted, the accuracy is measured by the standard error of the sample mean. Since

$$var(\hat{\theta}_{CR}) = \frac{var(f(U_1))}{n},$$

the standard error of the sample mean is the standard deviation or

$$SE(\hat{\theta}_{CR}) = \frac{\sigma_f}{\sqrt{n}}. \tag{4.3}$$

where $\sigma_f^2 = var(f(U))$. As usual we estimate σ_f^2 using the sample standard deviation. Since fn(u) provides a whole vector of estimators $(f(U_1), f(U_2), ..., f(U_n))$ then sqrt(var(fn(u))) is the sample estimator of σ_f so the standard error $SE(\hat{\theta}_{CR})$ is given by

```
Sf=sqrt(var(fn(u)));
Sf/sqrt(length(u))
```

giving an estimate 0.6603 of the standard deviation σ_f or standard error $\sigma_f/\sqrt{500000}$ or 0.0009. Of course parameters in statistical problems are usually estimated using an interval estimate or a *confidence interval,* an interval constructed using a method that guarantees capturing the true value of the parameter under similar circumstances with high probability (the confidence coefficient, often taken to be 95%). Formally,

Definition 35 *A 95% confidence interval for a parameter θ is an interval $[L, U]$ with random endpoints L, U such that the probability $P[L \leq \theta \leq U] = 0.95$.*

If we were to repeat the experiment 100 times, say by running 100 more similar independent simulations, and in each case use the results to construct a 95% confidence interval, then this definition implies that roughly 95% of the intervals constructed will contain the true value of the parameter (and of course

roughly 5% will not). For an approximately Normal(μ_X, σ_X^2) random variable X, we can use the approximation

$$P[\mu_X - 2\sigma_X \leq X \leq \mu_X + 2\sigma_X] \approx 0.95 \tag{4.4}$$

(i.e. approximately normal variables are within 2 standard deviations of their mean with probability around 95%) to build a simple confidence interval. Strictly, the value $2\sigma_X$ should be replaced by $1.96\sigma_X$ where 1.96 is taken from the Normal distribution tables. The value 2 is very close to correct for a t distribution with 60 degrees of freedom. In any case these confidence intervals which assume approximate normality are typically too short (i.e. contain the true value of the parameter less frequently than stated) for most real data and so a value marginally larger than 1.96 is warranted. Replacing σ_X above by the standard deviation of a sample mean, (4.4) results in the approximately 95% confidence interval

$$\hat{\theta}_{CR} - 2\frac{\sigma_f}{\sqrt{n}} \leq \theta \leq \hat{\theta}_{CR} + 2\frac{\sigma_f}{\sqrt{n}}$$

for the true value θ. With confidence 95%, the true price of the option is within the interval $0.462 \pm 2(0.0009)$. As it happens in this case this interval does capture the true value *0.4615* of the option.

So far Monte Carlo has not told us anything we couldn't obtain from the Black-Scholes formula, but what is we used a distribution other than the normal to generate the returns? This is an easy modification of the above. For example suppose we replace the standard normal by a logistic distribution which, as we have seen, has a density function very similar to the standard normal if we choose $b = 0.625$. Of course the Black-Scholes formula does not apply to a process with logistically distributed returns. We need only replace the standard normal inverse cumulative distribution function by the corresponding inverse for the logistic,

$$F^{-1}(U) = b \ln\left(\frac{U}{1-U}\right)$$

and thus replace the Matlab code, "`norminv(u,T*(r-sigma^2/2),sigma*sqrt(T))''` by `''T*(r-sigma^2/2)+sigma*sqrt(T)*.625*log(u./(1-u))''`. This results in a slight increase in option value (to 0.504) and about a 50% considerable increase in the variance of the estimator.

We will look at the efficiency of various improvements to crude Monte Carlo, and to that end, we record the value of the variance of the estimator based on a single uniform variate in this case;

$$\sigma^2_{crude} = \sigma^2_f = var(f(U)) \approx 0.436.$$

Then the crude Monte Carlo estimator using n function evaluations or n uniform variates has variance approximately $0.436/n$. If I were able to adjust the method so that the variance σ^2_f based on a single evaluation of the function f 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 computational effort 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.

$$\text{Efficiency} = \frac{\text{variance of Crude Monte Carlo Estimator}}{\text{Variance of new estimator}} \tag{4.5}$$

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 of 100 would indicate that the crude Monte Carlo estimator would require 100 times the number of function evaluations to achieve the same variance or standard error of estimator.

Consider a crude estimator obtained from five $U[0,1]$ variates,

$$U_i = 0.1, 0.3, 0.5, 0.6, 0.8, i = 1, ..., 5.$$

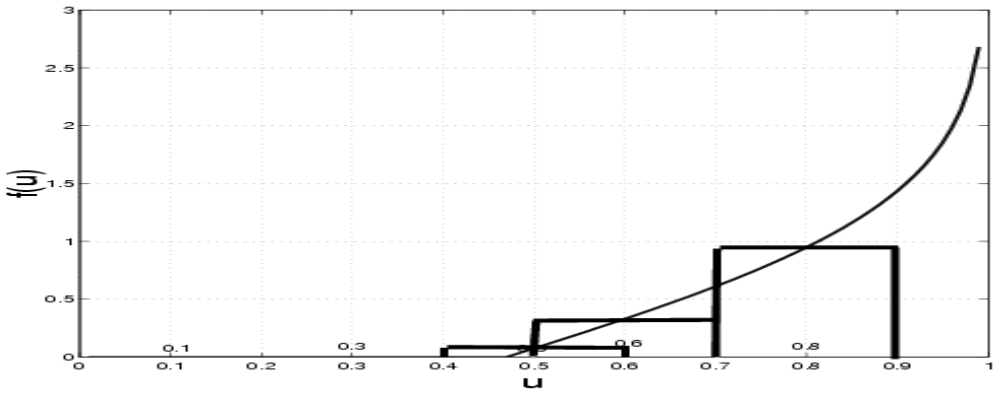


Figure 4.2: Crude Monte Carlo Estimator based on 5 observations $U_i = 0.1, 0.3, 0.5, 0.6, 0.8$

The crude Monte Carlo estimator in the case $n = 5$ is displayed in Figure 3.1, the estimator being the sum of the areas of the marked rectangles. Only three of the five points actually contribute to this area since for this particular function

$$f(u) = e^{-rT}(S_0 \exp\{\Phi^{-1}(u; rT - \frac{\sigma^2}{2}T, \sigma^2 T)\} - K)^+ \tag{4.6}$$

and the parameters chosen, $f(0.1) = f(0.3) = 0$. Since these two random numbers contributed 0 and the other three appear to be on average slightly too small, the sum of the area of the rectangles appears to underestimate of the integral. Of course another selection of five uniform random numbers may prove to be even more badly distributed and may result in an under or an overestimate.

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 $\hat{\theta}_{CR}$, are weighted averages of the values of the function at certain points U_i. What if we evaluated the function at non-random points, chosen to attempt reasonable balance between locations where the function is large and small? Numerical integration techniques and quadrature methods choose both points at which we evaluate the function and

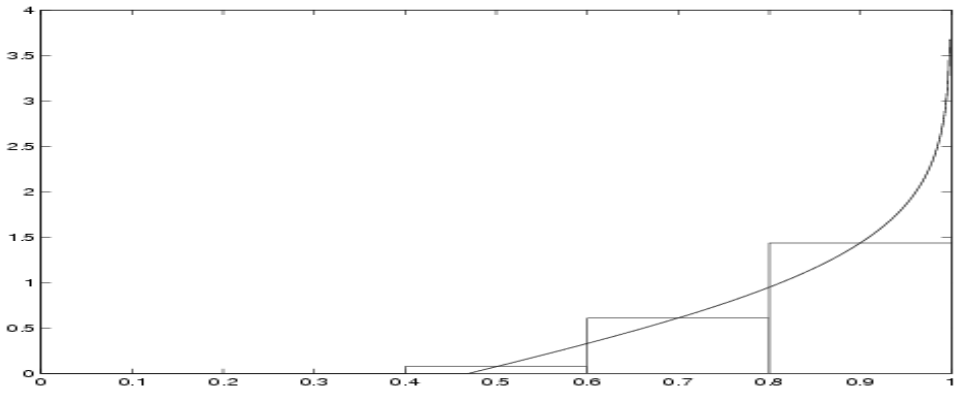


Figure 4.3: Graphical illustration of the trapezoidal rule (4.8)

weights that we attach to these points to provide accurate approximations for polynomials of certain degree. For example, suppose we insist on evaluating the function at equally spaced points, for example the points $0, 1/n, 2/n, ..., (n-1)/n, 1$. In some sense these points are now "more uniform" than we are likely to obtain from $n+1$ randomly and independently chosen points $U_i, i = 1, 2, ..., n$. The trapezoidal rule corresponds to using such equally spaced points and equal weights (except at the boundary) so that the "estimator" of the integral is

$$\hat{\theta}_{TR} = \frac{1}{2n}\{f(0) + 2f(1/n) + \ldots + 2f(1 - \frac{1}{n}) + f(1)\} \tag{4.7}$$

or the simpler and very similar alternative in our case, with $n = 5$,

$$\hat{\theta}_{TR} = \frac{1}{5}\{f(0.1) + f(0.3) + f(0.5) + f(0.7) + f(0.9)\} \tag{4.8}$$

A reasonable balance between large and small values of the function is almost guaranteed by such a rule, as shown in Figure 4.8 with the observations equally spaced.

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

integral is estimated with

$$\hat{\theta}_{SR} = \frac{1}{3n}\{f(0) + 4f(1/n) + 2f(2/n) + \ldots + 4f(\frac{n-1}{n}) + f(1)\}. \qquad (4.9)$$

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

These one-dimensional numerical integration rules provide some insight into how to achieve lower variance in Monte Carlo integration. It illustrates some options for increasing accuracy over simple random sampling. We may either vary the weights attached to the individual points or vary the points (the U_i) themselves or both. Notice that as long as the U_i individually have distributions that are $Uniform[0,1]$, we can introduce any degree of dependence among them in order to come closer to the equal spacings characteristic of numerical integrals. Even if the U_i are dependent U[0,1], an estimator of the form

$$\frac{1}{n}\sum_{i=1}^{n} f(U_i)$$

will continue to be an unbiased estimator because each of the summands continue to satisfy $E(f(U_i)) = \theta$. Ideally if we introduce dependence among the various U_i and the expected value remains unchanged , we would wish that the variance

$$var(\frac{1}{n}\sum_{i=1}^{n} f(U_i))$$

is reduced over independent uniform. The simplest case of this idea is the use of antithetic random variables.

Antithetic Random Numbers.

Consider first the simple case of $n = 2$ function evaluations at possibly dependent points. Then the estimator is

$$\hat{\theta} = \frac{1}{2}\{f(U_1) + f(U_2)\}$$

with expected value $\theta = \int_0^1 f(u)du$ and variance given by

$$var(\hat{\theta}) = \frac{1}{2}\{var(f(U_1)) + cov[f(U_1), f(U_2)]\}$$

assuming both U_1, U_2 are uniform[0,1]. In the independent case the covariance term disappears and we obtain the variance of the crude Monte-Carlo estimator

$$\frac{1}{2}var(f(U_1)).$$

Notice, however, that if we are able to introduce a *negative covariance*, the resulting variance of $\hat{\theta}$ will be smaller than that of the corresponding crude Monte Carlo estimator, so the question is how to generate this negative covariance. *Suppose for example that f is monotone (increasing or decreasing).* Then $f(1-U_1)$ decreases whenever $f(U_1)$ increases, so that substituting $U_2 = 1 - U_1$ has the desired effect and produces a negative covariance(in fact we will show later that we cannot do any better when the function f is monotone). Such a choice of $U_2 = 1 - U_1$ which helps reduce the variability in $f(U_1)$, 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

$$\frac{1}{2}\{var(f(U_1)) + cov[f(U_1), f(U_2)]\} < \frac{1}{2}var(f(U_1)).$$

that is, the variance is decreased over simple random sampling. Of course in practice our sample size is much greater than $n = 2$, but we still enjoy the benefits of this argument if we generate the points in antithetic pairs. For example, to determine the extent of the variance reduction using antithetic random numbers, suppose we generate $500,000$ uniform variates U and use as well the values of $1 - U$ as (for a total of $1,000,000$ function evaluations as before).

```
F=(fn(u)+fn(1-u))/2;
```

This results in *mean(F)=0.46186* and *var(F)=0.1121.* The standard error of the estimator is

$$\sqrt{\frac{0.1121}{length(F)}} = \sqrt{}{2.24 \times 107.}$$

Since each of the 500,000 components of F obtains from two function evaluations, the variance should be compared with a crude Monte Carlo estimator with the same number 1000000 function evaluations, $\sigma^2_{crude}/1000000 = 4.35 \times 10^{-7}$. The efficiency gain due to the use of antithetic random numbers is $4.35/2.24$ or about two, so roughly half as many function evaluations using antithetic random numbers provide the same precision as a crude Monte Carlo estimator. There is the additional advantage that only half as many uniform random variables are required. The introduction of antithetic variates has had the same effect on precision as increasing the sample size under crude Monte Carlo by a factor of approximately 2.

We have noted that antithetic random numbers improved the efficiency whenever the function being integrated is monotone in u. What if it is not. For example suppose we use antithetic random numbers to integrate the function $f(u) = u(1-u)$ on the interval $0 < u < 1$? Rather than balance large values with small values and so reduce the variance of the estimator, in this case notice that $f(U)$ and $f(1-U)$ are strongly *positively* correlated, in fact are equal, and so the argument supporting the use of antithetic random numbers for monotone functions will show that in this case they increase the variance over a crude estimator with the same number of function evaluations. Of course this problem can be remedied if we can identify intervals in which the function is monotone, e.g. in this case use antithetic random numbers in the two intervals $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$, so for example we might estimate $\int_0^1 f(u)du$ by an average of terms like

$$\frac{1}{4}\{f(\frac{U_1}{2}) + f(\frac{1-U_1}{2}) + f(\frac{1+U_2}{2}) + f(\frac{2-U_2}{2})\}$$

for independent $U[0,1]$ random variables U_1, U_2.

Stratified Sample.

One of the reasons for the inaccuracy of the crude Monte Carlo estimator in the above example is the large interval, evident in Figure 4.1, in which the function is zero. Nevertheless, both crude and antithetic Monte Carlo methods sample in that region, this portion of the sample contributing nothing to our integral. Naturally, we would prefer to concentrate our sample in the region where the function is positive, and where the function is more variable, use larger sample sizes. One method designed to achieve this objective is the use of a *stratified sample*. Once again for a simple example we choose $n = 2$ function evaluations, and with $V_1 \sim U[0, a]$ and $V_2 \sim U[a, 1]$ define an estimator

$$\hat{\theta}_{st} = af(V_1) + (1 - a)f(V_2).$$

Note that this is a weighted average of the two function values with weights a and $1 - a$ proportional to the length of the corresponding intervals. It is easy to show once again that the estimator $\hat{\theta}_{st}$ is an unbiased estimator of θ, since

$$\begin{aligned} E(\hat{\theta}_{st}) &= aEf(V_1) + (1 - a)Ef(V_2) \\ &= a\int_0^a f(x)\frac{1}{a}dx + (1 - a)\int_a^1 f(x)\frac{1}{1 - a}dx \\ &= \int_0^1 f(x)dx. \end{aligned}$$

Moreover,

$$var(\hat{\theta}_{st}) = a^2 var[f(V_1)] + (1 - a)^2 var[f(V_2)] + 2a(1 - a)cov[f(V_1), f(V_2)]. \tag{4.10}$$

Even when $V_1, V_2$are independent, so we obtain $var(\hat{\theta}_{st}) = a^2 var[f(V_1)] + (1 - a)^2 var[f(V_2)]$, there may be a dramatic improvement in variance over crude Monte Carlo provided that the variability of f in each of the intervals $[0, a]$ and $[a, 1]$ is substantially less than in the whole interval $[0, 1]$.

Let us return to the call option example above, with f defined by (4.6).

Suppose for simplicity we choose independent values of V_1, V_2. In this case

$$var(\hat{\theta}_{st}) = a^2 var[f(V_1)] + (1-a)^2 var[f(V_2)]. \tag{4.11}$$

For example for $a = .7$, this results in a variance of about 0.046 obtained from the following

```
F=a*fn(a*rand(1,500000))+(1-a)*fn(a+(1-a)*rand(1,500000));
var(F)
```

and the variance of the sample mean of the components of the vector F is var(F)/length(F) or around 9.2×10^{-8}. Since each component of the vector above corresponds to two function evaluations we should compare this with a crude Monte Carlo estimator with $n = 1000000$ having variance $\sigma_f^2 \times 10^{-6} = 4.36 \times 10^{-7}$. This corresponds to an efficiency gain of .43.6/9.2 or around 5. We can afford to use one fifth the sample size by simply stratifying the sample into two strata. The improvement is somewhat 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 $0 = x_0 < x_1 < ...x_k = 1$, and then select n_i random variables uniform on the corresponding interval $V_{ij} \sim U[x_{i-1}, x_i], j = 1, 2, ..., n_i$. Then the estimator of θ is

$$\hat{\theta}_{st} = \sum_{i=1}^{k} (x_i - x_{i-1}) \frac{1}{n_i} \sum_{j=1}^{n_i} f(V_{ij}). \tag{4.12}$$

Once again the weights $(x_i - x_{i-1})$ on the average of the function in the $i'th$ interval are proportional to the lengths of these intervals and the estimator $\hat{\theta}_{st}$

is unbiased;

$$\begin{aligned} E(\hat{\theta}_{st}) &= \sum_{i=1}^{k}(x_i - x_{i-1})E\{\frac{1}{n_i}\sum_{j=1}^{n_i} f(V_{ij})\} \\ &= \sum_{i=1}^{k}(x_i - x_{i-1})Ef(V_{i1}) \\ &= \sum_{i=1}^{k}(x_i - x_{i-1})\int_{x_{i-1}}^{x_i} f(x)\frac{1}{x_i - x_{i-1}}dx \\ &= \int_0^1 f(x)dx = \theta. \end{aligned}$$

In the case that all of the V_{ij} are independent, the variance is given by:

$$var(\hat{\theta}_{st}) = \sum_{i=1}^{k}(x_i - x_{i-1})^2\frac{1}{n_i}var[f(V_{i1})]. \tag{4.13}$$

Once again, if we choose our intervals so that the variation within intervals $var[f(V_{i1})]$ is small, this provides a substantial improvement over crude Monte Carlo. Suppose we wish to choose the sample sizes so as to minimize this variance. Obviously to avoid infinite sample sizes and to keep a ceiling on costs, we need to impose a constraint on the total sample size, say

$$\sum_{i}^{k} n_i = n. \tag{4.14}$$

If we treat the parameters n_i as continuous variables we can use the method of Lagrange multipliers to solve

$$\min_{\{n_i\}} \sum_{i=1}^{k}(x_i - x_{i-1})^2\frac{1}{n_i}var[f(V_{i1})]$$

subject to constraint (4.14).

It is easy to show that the optimal choice of sample sizes within intervals are

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

or more precisely that

$$n_i = n\frac{(x_i - x_{i-1})\sqrt{var[f(V_{i1})]}}{\sum_{j=1}^{k}(x_j - x_{j-1})\sqrt{var[f(V_{j1})]}}. \tag{4.15}$$

In practice,of course, this will not necessarily produce an integral value of n_i and so we are forced to round to the nearest integer. For this optimal choice of sample size, the variance is now given by

$$var(\hat{\theta}_{st}) = \frac{1}{n}\{\sum_{j=1}^{k}(x_j - x_{j-1})\sqrt{var[f(V_{j1})]}\}^2$$

The term $\sum_{j=1}^{k}(x_j - x_{j-1})\sqrt{var[f(V_{j1})]}$ is a weighted average of the standard deviation of the function f within the interval (x_{i-1}, x_i) and it is clear that, at least for a continuous function, these standard deviations can be made small simply by choosing k large with $|x_i - x_{i-1}|$ small. In other words if we ignore the fact that the sample sizes must be integers, at least for a continuous function f, we can achieve arbitrarily small $var(\hat{\theta}_{st})$ using a fixed sample size n simply by stratifying into a very large number of (small) strata. 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 summary, *optimal sample sizes are proportional to the lengths of intervals times the standard deviation of function evaluated at a uniform random variable on the interval. For sufficiently small strata we can achieve arbitrarily small variances.* The following function was designed to accept the strata $x_1, x_2, ..., x_k$ and the desired sample size n as input, and then determine optimal sample sizes and the stratified sample estimator as follows:

1. Initially sample sizes of 1000 are chosen from each stratum and these are used to estimate $\sqrt{var[f(V_{i1})]}$

2. Approximately optimal sample sizes n_i are then calculated from (4.15).

3. Samples of size n_i are then taken and the stratified sample estimator (4.12), its variance (4.13) and the sample sizes n_i are output.

```
function [est,v,n]=stratified(x,nsample)
% function for optimal sample size stratified estimator on call option price example
```

```
%[est,v,n]=stratified([0 .6 .85 1],100000)  uses three strata (0,.6),(.6 .85),(.85 1)
and total sample size 100000
est=0;
n=[];
m=length(x);
for i=1:m-1                          % the preliminary sample of size 1000
 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));   %calculation of the optimal sample sizes, rounded
down
v=0;
for i=1:m-1
 F=callopt2(unifrnd(x(i),x(i+1),1,n(i)),10,10,.05,.2,.25);    %evaluate the function
f at n(i) uniform points in interval
 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 three strata[0, 0.6], (0.6, 0.85], and (0.8, 1] and outputs the estimate $est = 0.4617$, its variance $v = 3.5 \times 10^{-7}$ 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 $\sigma_f^2/(99998 \times 3.5 \times 10^{-7}) = 12.8$. Evidently this stratified random sample can account for an improvement in efficiency of about a factor of 13. Of course there is a little setup cost here (a preliminary sample of size 3000) which we have not included in our calculation but the results of that preliminary sample could have been combined with the main sample for a very slight decrease in variance as well). For comparison, the function call

[est,v,n]=stratified([.47 .62 .75 .87 .96 1],1000000)

uses five strata [.47 .62],[.62 .75], [.75, .87], [.87, .96], [.96, 1] and gives a variance of the estimator of 7.4×10^{-9}. Since a crude sample of the same size has variance around 4.36×10^{-7} the efficiency is about 170. This stratified sample is as good as a crude Monte Carlo estimator with 170 million simulations! By introducing more strata, we can increase this efficiency as much as we wish.

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, or if we believe that the function is increasing across adjacent strata, we can introduce antithetic pairs between two intervals. For example, we may generate $U \sim Uniform[0,1]$ and then sample the point $V_{ij} = x_{i-1} + (x_i - x_{i-1})U$ from the interval (x_{i-1}, x_i) as well as the point $V_{(i+1)j} = x_{i+1} - (x_{i+1} - x_i)U$ from the interval (x_i, x_{i+1}) to obtain antithetic pairs between intervals. For a simple example of this applied to the above call option valuation, consider the estimator based on three strata [0,.47),[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, so the sample size here is set to 0. Then using antithetic random numbers within each of the two strata [0.47 0.84),[0.84 1], and $U \sim Uniform[0,1]$ we obtain the estimator

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

To assess this estimator,
we evaluated, for U a vector of 1000000 uniform,

```
U=rand(1,1000000);
F=.37*.5*(fn(.47+.37*U)+fn(.84-.37*U))+.16*.5*(fn(.84+.16*U)+fn(1-.16*U));
mean(F)                    % gives 0.4615
var(F)/length(F)                    %  gives 1.46×10^-9
```

This should be compared with the crude Monte-Carlo estimator having the same number $n = 4\times 10^6$ function evaluations as each of the components of the vector $F : \sigma^2_{crude}/(4\times 10^6) = 1.117\times 10^{-7}$. The gain in efficiency is therefore 1.117/.0146 or approximately 77. The above stratified-antithetic simulation with 1,000,000 input variates and 4,000,000 function evaluations is equivalent to a crude Monte Carlo simulation with sample size 308 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. However on a Pentium IV 2.2GHZ laptop it took approximately 58 seconds to run.

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*, based on the trivial identity

$$\int f(u)du = \int g(u)du + \int (f(u) - g(u))du. \tag{4.16}$$

for an arbitrary function $g(u)$. Assume that the integral of g is known, so we can substitute its known value for the first term above. The second integral we assume is more difficult and we estimate it by crude Monte Carlo, resulting in estimator

$$\hat{\theta}_{cv} = \int g(u)du + \frac{1}{n}\sum_{i=1}^{n}[f(U_i) - g(U_i)]. \tag{4.17}$$

This estimator is clearly unbiased and has variance

$$\begin{aligned} var(\hat{\theta}_{cv}) &= var\{\frac{1}{n}\sum_{i=1}^{n}[f(U_i) - g(U_i)]\} \\ &= \frac{var[f(U) - g(U)]}{n} \end{aligned}$$

so the variance is reduced over that of crude Monte Carlo estimator having the same sample size n by a factor

$$\frac{var[f(U)]}{var[f(U) - g(U)]} \quad \text{for } U \sim U[0,1]. \tag{4.18}$$

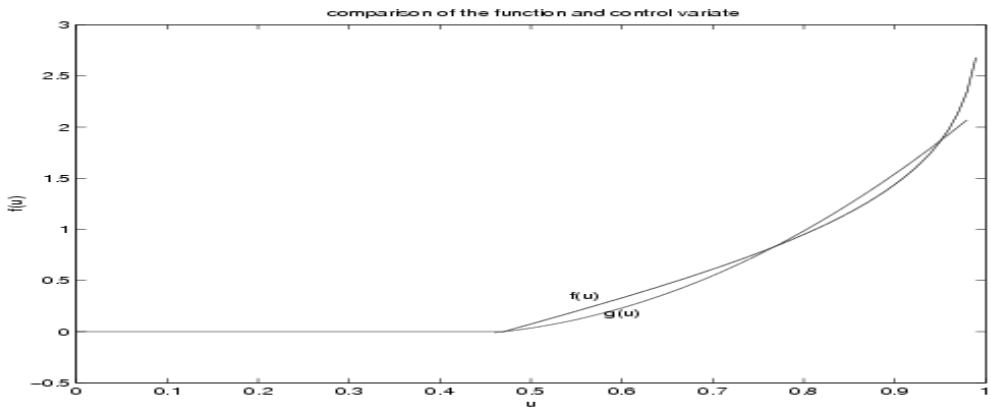


Figure 4.4: Comparison of the function $f(u)$ and the control variate $g(u)$

Let us return to the example of pricing a call option. By some experimentation, which could involve a preliminary crude simulation or simply evaluating the function at various points, we discovered that the function

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

provided a reasonable approximation to the function $f(u)$. The two functions are compared in Figure 4.4. Moreover, the integral $2 \times 0.53^2 + \frac{1}{2}0.53^3$ of the function $g(.)$ is easy to obtain.

It is obvious from the figure that since $f(u) - g(u)$ is generally much smaller and less variable than is $f(u)$, $var[f(U) - g(U)] < var(f(U))$. The variance of the crude Monte Carlo estimator is determined by the variability in the function $f(u)$ over its full range. The variance of the control variate 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, the first to generate the function $g(u)$ and the second to determine the efficiency gain of the control variate estimator;

```
function g=GG(u)          %  this is the function g(u), a control variate for fn(u)
u=max(0,u-.47);
g=6*u.^2+u;
```

```
function [est,var1,var2]=control(f,g,intg,n)
%  run using a statement like [est,var1,var2]=control('fn','GG',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=∫_0^1 g(u)du
%      outputs estimator est and variances var1,var2, variances with and without control variate.
U=unifrnd(0,1,1,n);
FN=eval(strcat(f,'(U)'));          % evaluates f(u)  for vector u
CN=eval(strcat(g,'(U)'));          % evaluates  g(u)
est=intg+mean(FN-CN);
var1=var(FN);
var2=var(FN-CN);
```

Then the call *[est,var1,var2]=control('fn','GG',2*(.53)^3+(.53)^2/2,1000000)* yields the estimate 0.4616 and variance=1.46×10^{-8} for an efficiency gain over crude Monte Carlo of around 30.

This elementary form of control variate suggests using the estimator

$$\int g(u)du + \frac{1}{n}\sum_{i=1}^{n}[f(U_i) - g(U_i)]$$

but it may well be that $g(U)$ is not the best estimator we can imagine for $f(U)$. We can often find a linear function of $g(U)$ which is better by using regression. Since elementary regression yields

$$f(U) - E(f(U)) = \beta(g(U) - E(g(U))) + \epsilon \tag{4.19}$$

where

$$\beta = \frac{cov(f(U), g(U))}{var(g(U))} \tag{4.20}$$

and the errors ϵ have expectation 0, it follows that $E(f(U)) + \epsilon = f(U) - \beta[g(U) - E(g(U))]$ and so$f(U) - \beta[g(U) - E(g(U))]$ is an unbiased estimator of $E(f(U))$. For a sample of n uniform random numbers this becomes

$$\hat{\theta}_{cv} = \beta E(g(U)) + \frac{1}{n}\sum_{i=1}^{n}[f(U_i) - \beta g(U_i)]. \tag{4.21}$$

Moreover this estimator having smallest variance among all linear combinations of $f(U)$and $g(U)$. Note that when $\beta = 1$ (4.21) reduces to the simpler form of the control variate technique (4.17) discussed above. However, the latter is generally better in terms of maximizing efficiency. Of course in practice it is necessary to estimate the covariance and the variances in the definition of β from the simulations themselves by evaluating f and g at many different uniform random variables $U_i, i = 1, 2, ..., n$ and then estimating β using the standard least squares estimator

$$\widehat{\beta} = \frac{n\sum_{i=1}^{n} f(U_i)g(U_i) - \sum_{i=1}^{n} f(U_i)\sum_{i=1}^{n} g(U_i)}{n\sum_{i=1}^{n} g^2(U_i) - (\sum_{i=1}^{n} g(U_i))^2}.$$

Although in theory the substitution of an estimator $\widehat{\beta}$ for the true value β results in a small bias in the estimator, for large numbers of simulations n our estimator $\widehat{\beta}$ is so close to the true value that this bias can be disregarded.

Importance Sampling.

A second technique that is similar is that of *importance sampling*. Again we depend on having a reasonably simple function g that after muultiplication by some constant, is similar to f. However, rather than attempt to minimize the difference $f(u) - g(u)$ between the two functions, we try and find $g(u)$ such that $f(u)/g(u)$ is nearly a constant. We also require that g is non-negative

and can be integrated so that, after rescaling the function, it integrates to one, i.e. it is a probability density function. Assume we can easily generate random variables from the probability density function $g(z)$. The distribution whose probability density function is $g(z), z \in [0,1]$ is the *importance distribution.* Note that if we generate a random variable Z having the probability density function $g(z), z \in [0,1]$ then

$$\begin{aligned} \int f(u)du &= \int_0^1 \frac{f(z)}{g(z)} g(z)dz \\ &= E\left[\frac{f(Z)}{g(Z)}\right]. \end{aligned} \tag{4.22}$$

This can therefore be estimated by generating independent random variables Z_i with probability density function $g(z)$ and then setting

$$\hat{\theta}_{im} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(Z_i)}{g(Z_i)}. \tag{4.23}$$

Once again, according to (4.22), this is an unbiased estimator and the variance is

$$var\{\hat{\theta}_{im}\} = \frac{1}{n} var\{\frac{f(Z_1)}{g(Z_1)}\}. \tag{4.24}$$

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 $g(u) = 2(0.53)^{-2}(u - .47)^{+}$ having cumulative distribution function $G(u) = (0.53)^2\ [(u-.47)^{+}]^2$ and inverse cumulative distribution function $G^{-1}(u) = 0.47 + 0.53\sqrt{u}$. In this case we generate Z_i using $Z_i = G^{-1}(U_i)$ for $U_i \sim Uniform[0,1]$. The following function simulates an importance sample estimator:

```
function [est,v]=importance(f,g,Ginv,u)
```

%runs a simulation on the function 'f" using importance density "g"(both character strings) and inverse c.d.f. "Ginverse"

% outputs all estimators (should be averaged) and variance.

% IM is the inverse cf of the importance distribution c.d.f.

% run e.g.

% [est,v]=importance('fn','2*(IM-.47)/(.53)^2;',.47+.53*sqrt(u);',rand(1,1000));

IM= eval(Ginv); %=.47+.53*sqrt(u);

%IMdens is the density of the importance sampling distribution at IM

IMdens=eval(g); %2*(IM-.47)/(.53)^2;

FN=eval(strcat(f,'(IM)'));

est=FN./IMdens; % mean(est) prrovides the estimator

v=var(FN./IMdens)/length(IM); % this is the variance of the estimator per simulation

The function was called with *[est,v]=importance('fn','2*(IM-.47)/(.53)^2;',.47+.53*sqrt(u);',rand(1,*

giving an estimate mean(*est)* = *0.4616* with variance 1.28×10^{-8} for an efficiency gain of around 35 over crude Monte Carlo.

Example 36 *(Estimating Quantiles using importance sampling.) Suppose we are able to generate random variables X from a probability density function of the form*

$$f_\theta(x)$$

and we wish to estimate a quantile such as VAR, i.e. estimate x_p such that

$$P_{\theta_0}(X \leq x_p) = p$$

for a certain value θ_0 of the parameter.

As a very simple example suppose S is the sum of 10 independent random variables having the exponential distribution with mean θ, and $f_\theta(x_1, ..., x_{10})$ is the joint probability density function of these 10 observations. Assume $\theta_0 = 1$

and $p = .999$ so that we seek an extreme quantile of the sum, i.e. we want to determine $x_p$such that $P_{\theta_0}(S \leq x_p) = p$. The equation that we wish to solve for x_p is

$$E_{\theta_0}\{I(S \leq x_p)\} = p. \tag{4.25}$$

The crudest estimator of this is obtained by generating a large number of independent observations of S under the parameter value $\theta_0 = 1$ and finding the p'th quantile, i.e. by defining the empirical c.d.f.. We generate independent random vectors $X_i = (X_{i1}, ..., X_{i10})$ from the probability density $f_{\theta_0}(x_1, ..., x_{10})$ and with $S_i = \sum_{j=1}^{10} X_{ij}$,define

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(S_i \leq x). \tag{4.26}$$

Invert it (possibly with interpolation) to estimate the quantile

$$\widehat{x_p} = \widehat{F}^{-1}(p). \tag{4.27}$$

If the true cumulative distribution function is differentiable, the variance of this quantile estimator is asymptotically related to the variance of our estimator of the cumulative distribution function,

$$var(\widehat{x_p}) \simeq \frac{var(\widehat{F}(x_p))}{(F'(x_p))^2},$$

so any variance reduction in the estimator of the c.d.f. us reflected, at least asymptotically, in a variance reduction in the estimator of the quantile. Using importance sampling (4.25) is equivalent to the same technique but with

$$\widehat{F}_I(x) = \frac{1}{n}\sum_{i=1}^{n} W_i I(S_i \leq x) \text{ where} \tag{4.28}$$
$$W_i = \frac{f_{\theta_0}(X_{i1}, ..., X_{i10})}{f_{\theta}(X_{i1}, ..., X_{i10})}$$

Ideally we should choose the value of θ so that the variance of $\widehat{x}_p$ or of

$$W_i I(S_i \leq x_p)$$

is as small as possible. This requires a wise guess or experimentation with various choices of θ. For a given θ we have another choice of empirical cumulative distribution function

$$\widehat{F}_{I2}(x) = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} W_i I(S_i \leq x). \tag{4.29}$$

Both of these provide fairly crude estimates of the sample quantiles when observations are weighted and, as one does with the sample median, one could easily interpolate between adjacent values around the value of x_p.

The alternative (4.29) is motivated by the fact that the values W_i appear as weights attached to the observations S_i and it therefore seems reasonable to divide by the sum of the weights. In fact the expected value of the denominator is

$$E_\theta\{\sum_{i=1}^{n} W_i\} = n$$

so the two denominators are similar. In the example where the X_{ij} are independent exponential(1) let us examine the weight on S_i determined by $X_i = (X_{i1}, ..., X_{i10})$,

$$W_i = \frac{f_{\theta_0}(X_{i1}, ..., X_{i10})}{f_\theta(X_{i1}, ..., X_{i10})} = \prod_{j=1}^{10} \frac{\exp(-X_{ij})}{\theta^{-1}\exp(-X_{ij}/\theta)} = \theta^{10}\exp\{-S_i(1-\theta^{-1})\}.$$

The renormalized alternative (4.29) might be necessary for estimating extreme quantiles when the number of simulations is small but only the first provides an completely unbiased estimating function. In our case, using (4.28) with $\theta = 2.5$ we obtained an estimator of $F(x_{0.999})$ with efficiency about 180 times that of a crude Monte Carlo simulation. There is some discussion of various renormalizations of the importance sampling weights in Hesterberg(1995).

Importance Sampling, the Exponential Tilt and the Saddlepoint Approximation

When searching for a convenient importance distribution, particularly if we wish to increase or decrease the frequency of observations in the tails, it is

quite common to embed a given density in an exponential family. For example suppose we wish to estimate an integral

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

where $f(x)$ is a probability density function. Suppose $K(s)$ denotes the cumulant generating function (the logarithm of the moment generating function) of the density $f(x)$,i.e. if

$$\exp\{K(s)\} = \int e^{xs} f(x)dx.$$

The cumulant generating function is a useful summary of the moments of a distribution since the mean can be determined as $K'(0)$ and the variance as $K''(0)$. From this single probability density function, we can now produce a whole (exponential) family of densities

$$f_\theta(x) = e^{\theta x - K(\theta)} f(x) \tag{4.30}$$

of which $f(x)$ is a special case corresponding to $\theta = 0$. The density (4.30) is often referred to as an exponential tilt of the original density function and increases the weight in the right tail for $\theta > 0$, decreases it for $\theta < 0$.

This family of densities is closely related to the saddlepoint approximation. If we wish to estimate the value of a probability density function $f(x)$ at a particular point x, then note that this could be obtained from (4.30) if we knew the probability density function $f_\theta(x)$. On the other hand a normal approximation to a density is often reasonable at or around its mode, particularly if we are interested in the density of a sum or an average of independent random variables. The cumulant generating function of the density $f_\theta(x)$ is easily seen to be $K(\theta + s)$ and the mean is therefore $K'(\theta)$. If we choose the parameter $\theta = \theta(x)$ so that

$$K'(\theta) = x \tag{4.31}$$

then the density f_θ has mean x and variance $K''(\theta)$. How do we know for a given value of x there exists a solution to (4.31)? From the properties of cumulant

generating functions, $K(t)$ is convex, increasing and $K(0) = 0$. This implies that as t increases, the slope of the cumulant generating function $K'(t)$ is non-decreasing. It therefore approaches a limit $x_{\max}$(finite or infinite) as $t \to \infty$ and as long as we restrict the value of x in (4.31) to the interval $x < x_{\max}$ we can find a solution. The value of the $N(x, K''(\theta))$ at the value x is

$$f_\theta(x) \approx \sqrt{\frac{1}{2\pi K''(\theta)}}$$

and therefore the approximation to the density $f(x)$ is

$$f(x) \approx \sqrt{\frac{1}{2\pi K''(\theta)}} e^{K(\theta)-\theta x}. \tag{4.32}$$

$$\text{where } \theta = \theta(x) \text{ satisfies } K'(\theta) = x.$$

This is the saddlepoint approximation, discovered by Daniels (1954, 1980), and usually applied to the distribution of sums or averages of independent random variables because then the normal approximation is better motivated. Indeed, the saddlepoint approximation to the distribution of the sum of n independent identically distributed random variables is accurate to order $O(n^{-1})$ and if we renormalize it to integrate to one, accuracy to order $O(n^{-3/2})$ is possible, substantially better than the order $O(n^{-1/2})$ of of the usual normal approximation.

Consider, for example, the saddlepoint approximation to the Gamma$(\alpha, 1)$ distribution. Because the moment generating function of the Gamma$(\alpha, 1)$ distribution is

$$m(t) = \frac{1}{(1-t)^\alpha}, t < 1,$$

the cumulant generating function is

$$\begin{aligned} K(t) &= \ln(m(t)) = -\alpha \ln(1-t), \\ K'(\theta) &= x \text{ implies } \theta(x) = 1 - \frac{\alpha}{x} \text{ and} \\ K''(\theta) &= \frac{\alpha}{(1-\theta)^2} \text{ so that } K''(\theta(x)) = \frac{x^2}{\alpha}. \end{aligned}$$

Therefore the saddlepoint approximation to the probability density function is

$$\begin{aligned} f(x) &\simeq \sqrt{\frac{\alpha}{2\pi x^2}}\exp\{-\alpha\ln(\alpha/x) - x(1-\frac{\alpha}{x})\} \\ &= \sqrt{\frac{1}{2\pi}}\alpha^{\frac{1}{2}-\alpha}e^{\alpha}x^{\alpha-1}\exp(-x). \end{aligned}$$

This is exactly the gamma density function with Stirling's approximation replacing $\Gamma(\alpha)$ and after renormalization this is exactly the Gamma density function.

Since it is often computationally expensive to generate random variables whose distribution is a convolution of known densities, it is interesting to ask whether (4.32) makes this any easier. In many cases the saddlepoint approximation can be used to generate a random variable whose distribution is close to this convolution with high efficiency. For example suppose that we wish to generate the random variable $S_n = \sum_{i=1}^{n} X_i$ where each random variable X_i has the *non-central chi-squared* distribution with cumulant generating function

$$K(t) = \frac{2\lambda t}{1-2t} - \frac{p}{2}\ln(1-2t). \tag{4.33}$$

The parameter λ is the *non-centrality parameter* of the distribution and p is the *degrees of freedom.* Notice that the cumulant generating function of the sum takes the same form but with (λ, p) replaced by $(n\lambda, np)$ so in effect we wish to generate a random variable with cumulant generating function (4.33) for large values of the parameters (λ, p). In stead we generate from the saddlepoint approximation (4.32) to this distribution and in fact we do this indirectly. If we change variable in (4.32) to determine the density of the new random variable Θ which solves the equation

$$K'(\Theta) = X$$

then the saddlepoint approximation (4.32) is equivalent to specifying a probability density for this variable,

$$\begin{aligned} f_{\Theta}(\theta) &= f(K'(\theta))\frac{dx}{d\theta} \\ &= \text{constant} \times \sqrt{K''(\theta)}e^{K(\theta)-\theta K'(\theta)}. \end{aligned} \tag{4.34}$$

In general, this probability density function can often be bounded above by some density over the range of possible values of θ allowing us to generate Θ by acceptance rejection. Then the value of the random variable is $X = K'(\Theta)$. In the particular case of the non-central chi-squared example above, swe may take the dominating density to be the $U[0, \frac{1}{2}]$ density since (4.34) is bounded.

Combining Monte Carlo Estimators.

We have now seen a number of different variance reduction techniques and there are many more possible. With many of these methods such as importance and stratified sampling are associated parameters which may be chosen in different ways. The variance formula may be used as a basis of choosing a "best" method but these variances and efficiencies must also estimated from the simulation and it is rarely 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

$$\hat{\theta}_{a1} = \frac{1}{2}[f(U) + f(1-U)], \quad U \sim U[0,1] \tag{4.35}$$

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

$$\hat{\theta}_{a2} = \frac{1}{4}[f(U/2) + f((1-U)/2) + f((1+U)/2) + f(1-U/2)]. \tag{4.36}$$

Notice that any weighted average of these two unbiased estimators of θ would also provide an unbiased estimator of θ. The large number of potential variance reduction techniques is an embarrassment of riches. Which variance reduction methods we should use and how will we know whether it is better than the competitors? Fortunately, the answer is often to use "all of the methods" (within reason of course); that choosing a single method is often neither necessary nor desirable. Rather it is preferable to use a weighted average of the available estimators with the optimal choice of the weights provided by regression.

Suppose in general that we have k estimators or statistics $\widehat{\theta}_i, i = 1, ..k$, all unbiased estimators of the same parameter θ so that $E(\widehat{\theta}_i) = \theta$ for all i . In vector notation, letting $\mathbf{\Theta}' = (\widehat{\theta}_1, ..., \widehat{\theta}_k)$, we write $E(\Theta) = \mathbf{1}\theta$ where $\mathbf{1}$ is the k-dimensional column vector of ones so that $\mathbf{1}' = (1, 1, ..., 1)$. Let us suppose for the moment that we know the variance-covariance matrix V of the vector $\mathbf{\Theta}$, defined by

$$V_{ij} = cov(\widehat{\theta}_i, \widehat{\theta}_j).$$

Theorem 37 *(best linear combinations of estimators)*
The linear combination of the $\widehat{\theta}_i$ which provides an unbiased estimator of θ and has minimum variance among all linear unbiased estimators is

$$\widehat{\theta}_{blc} = \sum_i b_i \widehat{\theta}_i \tag{4.37}$$

where the vector $\mathbf{b} = (b_1, ..., b_k)'$ is given by

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

The variance of the resulting estimator is

$$var(\widehat{\theta}_{blc}) = \mathbf{b}^t V \mathbf{b} = 1/(\mathbf{1}^t V^{-1} \mathbf{1})$$

Proof. The proof is straightforward. It is easy to see that for any linear combination (4.37) the variance of the estimator is

$$\mathbf{b}^t V \mathbf{b}$$

and we wish to minimize this quadratic form as a function of $\mathbf{b}$ subject to the constraint that the coefficients add to one, or that

$$\mathbf{b}'\mathbf{1} = 1.$$

Introducing the Lagrangian, we wish to set the derivatives with respect to

the components b_i equal to zero

$$\frac{\partial}{\partial \mathbf{b}}\{\mathbf{b}^t V\mathbf{b} + \lambda(\mathbf{b}'\mathbf{1} - 1)\} = \mathbf{0} \text{ or}$$

$$2V\mathbf{b} + \lambda\mathbf{1} = \mathbf{0}$$

$$\mathbf{b} = \text{constant} \times V^{-1}\mathbf{1}$$

and upon requiring that the coefficients add to one, we discover the value of the constant above is $(\mathbf{1}^t V^{-1}\mathbf{1})^{-1}$. ■

This theorem indicates that the ideal linear combination of estimators has coefficients proportional to the *row sums of the inverse covariance matrix*. Notably, the variance of a particular estimator $\widehat{\theta}_i$ is an ingredient in that sum, but one of many. In practice, of course, we almost never know the variance-covariance matrix V of a vector of estimators Θ. However, when we do simulation evaluating these estimators using the same uniform input to each, we obtain independent replicated values of Θ. This permits us to estimate the covariance matrix V and since we typically conduct many simulations this estimate can be very accurate. Let us suppose that we have n simulated values of the vectors Θ, and call these $\Theta_1, ..., \Theta_n$. As usual we estimate the covariance matrix V using the sample covariance matrix

$$\widehat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(\Theta_i - \overline{\Theta})(\Theta_i - \overline{\Theta})'$$

where

$$\overline{\Theta} = \frac{1}{n}\sum_{i=1}^{n}\Theta_i.$$

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

$\widehat{\theta}_1 = \frac{0.53}{2}[f(.47 + .53u) + f(1 - .53u)]$ an antithetic estimator,

$\widehat{\theta}_2 = \frac{0.37}{2}[f(.47 + .37u) + f(.84 - .37u)] + \frac{0.16}{2}[f(.84 + .16u) + f(1 - .16u)]$,

$\widehat{\theta}_3 = 0.37[f(.47 + .37u)] + 0.16[f(1 - .16u)]$, (stratified-antithetic)

$\widehat{\theta}_4 = \int g(x)dx + [f(u) - g(u)]$, (control variate)

$\widehat{\theta}_5 = \hat{\theta}_{im}$, the importance sampling estimator (4.23).

Then $\widehat{\theta}_2$, and $\widehat{\theta}_3$ are both stratified-antithetic estimators, $\widehat{\theta}_4$ is a control variate estimator and $\widehat{\theta}_5$ the importance sampling estimator discussed earlier, all obtained from a single input uniform random variate U. 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
%  input U  a row vector of U[0,1] random numbers
T1=(.53/2)*(fn(.47+.53*U)+fn(1-.53*U));
T2=.37*.5*(fn(.47+.37*U)+fn(.84-.37*U))+.16*.5*(fn(.84+.16*U)+fn(1-.16*U));
T3=.37*fn(.47+.37*U)+.16*fn(1-.16*U);
intg=2*(.53)^3+.53^2/2;
T4=intg+fn(U)-GG(U);
T5=importance('fn',U);
X=[T1' T2' T3' T4' T5'];   % matrix whose columns are replications of the same
estimator, a row=5 estimators using same U
mean(X)
V=cov(X);  % this estimates the covariance matrix V
```

```
on=ones(5,1);
V1=inv(V);                    % the inverse of the covariance matrix
b=V1*on/(on'*V1*on);   % vector of coefficients of the optimal linear combination
o=mean(X*b);                  % vector of the optimal linear combinations
v=1/(on'*V1*on);              % variance of the optimal linear combination based on
a single U
```

One run of this estimator, called with *[o,v,b,V]= optimal(unifrnd(0,1,1,1000000))* yields

$$o = 0.4615$$

$$b' = [-0.5499 \quad 1.4478 \quad 0.1011 \quad 0.0491 \quad -0.0481].$$

The estimate 0.4615 is accurate to at least four decimals which is not surprising since the variance per uniform random number input is $v = 1.13 \times 10^{-5}$. In other words, the variance of the mean based on 1,000,000 uniform input is 1.13×10^{-10}, the standard error is around .00001 so we can expect accuracy to at least 4 decimal places. Note that some of the weights are negative and others are greater than one. Do these negative weights indicate estimators that are worse than useless? The effect of some estimators may be, on subtraction, to render the remaining function more linear and more easily estimated using another method and negative coefficients are quite common in regression generally. The efficiency gain over crude Monte Carlo is an extraordinary 40,000. However since there are 10 function evaluations for each uniform variate input, the efficiency when we adjust for the number of function evaluations is 4,000. This simulation using 1,000,000 uniform random numbers and taking a 63 seconds on a Pentium IV (2.4 GHz) (including the time required to generate all five estimators) is equivalent to *forty billion simulations by crude Monte Carlo, a major task on a supercomputer!*

If we intended to use this simulation method repeatedly, we 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 $1/(\mathbf{1}^t V^{-1}\mathbf{1})$, we might use this to attempt to select one of the estimators for deletion. Notice that it is not so much the covariance of the estimators V which enters into Theorem 35 but its inverse $\mathbf{J} = V^{-1}$ which we can consider a type of information matrix by analogy to maximum likelihood theory. For example we could choose to delete the $i'th$ estimator, i.e. delete the $i'th$ row and column of V where i is chosen to have the smallest effect on $1/(\mathbf{1}^t V^{-1}\mathbf{1})$ or its reciprocal $\mathbf{1}^t\mathbf{J}\mathbf{1} = \sum_i \sum_j \mathbf{J}_{ij}$. In particular, if we let $V_{(i)}$ be the matrix V with the $i'th$ row and column deleted and $\mathbf{J}_{(i)} = \mathbf{V}_{(i)}^{-1}$, then we can identify $\mathbf{1}^t\mathbf{J}\mathbf{1} - \mathbf{1}^t\mathbf{J}_{(i)}\mathbf{1}$ 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

$$\min_i \{\frac{\mathbf{1}^t\mathbf{J}\mathbf{1} - \mathbf{1}^t\mathbf{J}_{(i)}\mathbf{1}}{FE(i)}\}.$$

We should drop this $i'th$ estimator if the minimum is less than the information per function evaluation in the combined estimator, because this means we will increase the information available in our simulation per function evaluation. In the above example with all five estimators included, $\mathbf{1}^t\mathbf{J}\mathbf{1} = 88757$ (with 10 function evaluations per uniform variate) so the information per function evaluation is $8,876$.

i	$\mathbf{1}^t\mathbf{J}\mathbf{1} - \mathbf{1}^t\mathbf{J}_{(i)}\mathbf{1}$	$FE(i)$	$\frac{\mathbf{1}^t\mathbf{J}\mathbf{1}-\mathbf{1}^t\mathbf{J}_{(i)}\mathbf{1}}{FE(i)}$
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 eliminate one of the estimators, our choice would

likely be number 3 since it contributes the least information per function evaluation. However, since all contribute 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 antithetic variates called *common random numbers*, used for example whenever we wish to estimate the difference in performance between two systems or any other variable involving a difference such as a slope of a function.

Example 38 *For a simple example suppose we have two estimators $\widehat{\theta}_1, \widehat{\theta}_2$ of the "center" of a symmetric distribution. We would like to know which of these estimators is better in the sense that it has smaller variance when applied to a sample from a specific distribution symmetric about its median. If both estimators are unbiased estimators of the median, then the first estimator is better if*

$$var(\widehat{\theta}_1) < var(\widehat{\theta}_2)$$

and so we are interested in estimating a quantity like

$$Eh_1(X) - Eh_2(X)$$

where X is a vector representing a sample from the distribution and $h_1(X) = \widehat{\theta}_1^2, h_2(X) = \widehat{\theta}_2^2$. There are at least two ways of estimating these differences;

1. *Generate samples and hence values of $h_1(X_i), i = 1, ..., n$ and $Eh_2(X_j), j = 1, 2, ..., m$ independently and use the estimator*
$$\frac{1}{n}\sum_{i=1}^{n} h_1(X_i) - \frac{1}{m}\sum_{j=1}^{m} h_2(X_j).$$
2. *Generate samples and hence values of $h_1(X_i), h_2(X_i), i = 1, ..., n$ independently and use the estimator*
$$\frac{1}{n}\sum_{i=1}^{n}(h_1(X_i) - h_2(X_i)).$$

It seems intuitive that the second method is preferable since it removes the variability due to the particular sample from the comparison. This is a common type of problem in which we want to estimate the difference between two expected values. For example 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 between the new system and the old. In general, suppose that we wish to estimate the difference between two expectations, say

$$Eh_1(X) - Eh_2(Y) \tag{4.38}$$

where the random variable or vector X has cumulative distribution function F_X and Y has cumulative distribution function F_Y. Notice that the variance of a Monte Carlo estimator

$$var[h_1(X) - h_2(Y)] = var[h_1(X)] + var[h_2(Y)] - 2cov\{h_1(X), h_2(Y)\} \tag{4.39}$$

is *small* if we can induce a high degree of *positive correlation* between the generated random variables 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 lemma is due to Hoeffding (1940) and provides a useful bound on the joint cumulative distribution function of two random variables X and Y. Suppose X, Y have cumulative distribution functions $F_X(x)$ and $F_Y(y)$ respectively and joint cumulative distribution function $G(x, y) = P[X \leq x, Y \leq y]$.

Lemma 39 *(a) The joint cumulative distribution function G of (X, Y) always satisfies*

$$(F_X(x) + F_Y(y) - 1)^+ \leq G(x, y) \leq \min(F_X(x), F_Y(y)) \tag{4.40}$$

for all x, y .

(b) Assume that F_X *and* F_Y *are continuous functions. In the case that* $X = F_X^{-1}(U)$ *and* $Y = F_Y^{-1}(U)$ *for* U *uniform on* $[0,1]$, *equality is achieved on the right* $G(x,y) = \min(F_X(x), F_Y(y))$. *In the case that* $X = F_X^{-1}(U)$ *and* $Y = F_Y^{-1}(1-U)$ *there is equality on the left;* $(F_X(x) + F_Y(y) - 1)^+ = G(x,y)$.

Proof. *(a) Note that*

$$\begin{aligned} P[X \le x, Y \le y] &\le P[X \le x] \text{ and similarly} \\ &\le P[Y \le y]. \end{aligned}$$

This shows that

$$G(x,y) \le \min(F_X(x), F_Y(y)),$$

verifying the right side of (4.40). Similarly for the left side

$$\begin{aligned} P[X \le x, Y \le y] &= P[X \le x] - P[X \le x, Y > y] \\ &\ge P[X \le x] - P[Y > y] \\ &= F_X(x) - (1 - F_Y(y)) \\ &= (F_X(x) + F_Y(y) - 1). \end{aligned}$$

Since it is also non-negative the left side follows.

For (b) suppose $X = F_X^{-1}(U)$ *and* $Y = F_Y^{-1}(U)$, *then*

$$\begin{aligned} P[X \le x, Y \le y] &= P[F_X^{-1}(U) \le x, F_Y^{-1}(U) \le y] \\ &= P[U \le F_X(x), U \le F_Y(y)] \end{aligned}$$

since $P[X = x] = 0$ *and* $P[Y = y] = 0$.

But

$$P[U \le F_X(x), U \le F_Y(y)] = \min(F_X(x), F_Y(y))$$

verifying the equality on the right of (4.40) for common random numbers. By

a similar argument,

$$\begin{aligned} P[F_X^{-1}(U) \le x, F_Y^{-1}(1-U) \le y] &= P[U \le F_X(x), 1-U \le F_Y(y)] \\ &= P[U \le F_X(x), U \ge 1-F_Y(y)] \\ &= (F_X(x) - (1-F_Y(y)))^+ \end{aligned}$$

verifying the equality on the left. ■

The following theorem supports the use of common random numbers to maximize covariance and antithetic random numbers to minimize covariance.

Theorem 40 *(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 when the distributions are all continuous and h_1, h_2 are differentiable. Define $G(x,y) = P[X \le x, Y \le y]$. The following representation of covariance is useful: define

$$\begin{aligned} H(x,y) &= P(X > x, Y > y) - P(X > x)P(Y > y) \\ &= G(x,y) - F_X(x)F_Y(y). \end{aligned} \tag{4.41}$$

Notice that, using integration by parts,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)h_1'(x)h_2'(y)dxdy \\
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial}{\partial x}H(x,y)h_1(x)h_2'(y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial^2}{\partial x \partial y}H(x,y)h_1(x)h_2(y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(x)h_2(y)g(x,y)dxdy - \int_{-\infty}^{\infty} h_1(x)f_X(x)dx \int_{-\infty}^{\infty} h_2(y)f_Y(y)dy \\
&= cov(h_1(X), h_2(Y)) \qquad (4.42)
\end{aligned}$$

where $g(x,y), f_X(x), f_Y(y)$ denote the joint probability density function, the probability density function of X and that of Y respectively. In fact this result holds in general even without the assumption that the distributions are continuous. The covariance between $h_1(X)$ and $h_2(Y)$, for h_1 and h_2 differentiable functions, is

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

The formula shows that to maximize the covariance, if h_1, h_2 are both increasing or both decreasing functions, it is sufficient to maximize $H(x,y)$ for each x, y since $h_1'(x), h_2'(y)$ are both non-negative. Since we are constraining the marginal cumulative distribution functions F_X, F_Y, this is equivalent to maximizing $G(x,y)$ subject to the constraints

$$\lim_{y\to\infty} G(x,y) = F_X(x)$$

$$\lim_{x\to\infty} G(x,y) = F_Y(y).$$

Lemma 37 shows that the maximum is achieved when common random numbers are used and the minimum achieved when we use antithetic random numbers.

■

We can argue intuitively for the use of common random numbers in the case of a discrete distribution with probability on the points indicated in Figure 4.5.

This figure corresponds to a joint distribution with the following probabilities, say

x	0	0.25	0.25	0.75	0.75	1
y	0	0.25	0.75	0.25	0.75	1
$P[X = x, Y = y]$	.1	.2	.2	.1	.2	.2

Suppose we wish to maximize $P[X > x, Y > y]$ subject to the constraint that the probabilities $P[X > x]$ and $P[Y > y]$ are fixed. We have indicated arbitrary fixed values of (x, y) in the figure. Note that if there is any weight attached to the point in the lower right quadrant (labelled "P_2"), some or all of this weight can be reassigned to the point P_3 in the lower left quadrant provided there is an equal movement of weight from the upper left P_4 to the upper right P_1. Such a movement of weight will increase the value of $G(x, y)$ without affecting $P[X \leq x]$ or $P[Y \leq y]$. The weight that we are able to transfer in this example is 0.1, the minimum of the weights on P_4 and P_2 . In general, this continues until there is no weight in one of the off-diagonal quadrants for every choice of (x, y). The resulting distribution in this example is given by

x	0	0.25	0.25	0.75	0.75	1
y	0	0.25	0.75	0.25	0.75	1
$P[X = x, Y = y]$	.1	.3	0	.1	.3	.2

and it is easy to see that such a joint distribution can be generated from common random numbers $X = F_X^{-1}(U), Y = F_Y^{-1}(U)$.

Conditioning

We now consider a simple but powerful generalization of control variates. Suppose that we can decompose a random variable T into two components T_1, ε

$$T = T_1 + \varepsilon \tag{4.43}$$

so that T_1, ε are uncorrelated

$$cov(T_1, \varepsilon) = 0.$$

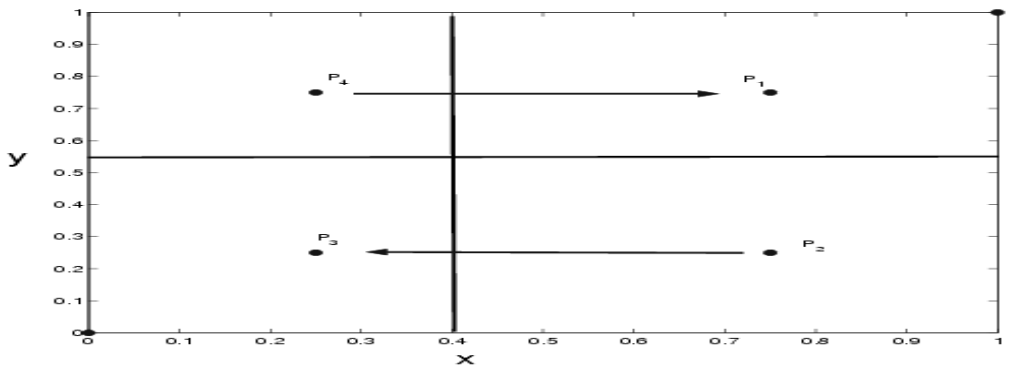


Figure 4.5: Changing weights on points to maximize covariance

Assume as well that $E(\varepsilon) = 0$. Regression is one method for determining such a decomposition and the error term ε in regression satisfies these conditions. Then T_1 has the same mean as T and it is easy to see that

$$var(T) = var(T_1) + var(\varepsilon)$$

so T_1 as smaller variance than T (unless $\varepsilon = 0$ with probability 1). This means that if we wish to estimate the common mean of T or T_1, the estimator T_1 is preferable, since it has the same mean with smaller variance.

One special case is variance reduction by *conditioning*. For the standard definition and properties of conditional expectation see the appendix. One common definition of $E[X|Y]$ is the unique (with probability one) function $g(y)$ of Y which minimizes $E\{X - g(Y)\}^2$. This definition only applies to random variables X which have finite variance and so this definition requires some modification when $E(X^2) = \infty$, but we will assume here that all random variables, say X, Y, Z have finite variances. We can define conditional covariance using conditional expectation as

$$cov(X, Y|Z) = E[XY|Z] - E[X|Z]E[Y|Z]$$

and conditional variance:

$$var(X|Z) = E(X^2|Z) - (E[X|Z])^2.$$

The variance reduction through conditioning is justified by the following well-known result:

Theorem 41 *(a)* $E(X) = E\{E[X|Y]\}$
(b) $cov(X,Y) = E\{cov(X,Y|Z)\} + cov\{E[X|Z], E[Y|Z]\}$
(c) $var(X) = E\{var(X|Z)\} + var\{E[X|Z]\}$

This theorem is used as follows. Suppose we are considering a candidate estimator $\hat{\theta}$, an unbiased estimator of θ. We also have an arbitrary random variable Z which is somehow related to $\hat{\theta}$. Suppose that we have chosen Z carefully so that we are able to calculate the conditional expectation $T_1 = E[\hat{\theta}|Z]$. Then by part (a) of the above Theorem, T_1 is also an unbiased estimator of θ. Define

$$\varepsilon = \hat{\theta} - T_1.$$

By part (c),

$$var(\hat{\theta}) = var(T_1) + var(\varepsilon)$$

and $var(T_1) = var(\hat{\theta}) - var(\varepsilon) < var(\hat{\theta})$. In other words, for any variable Z, $E[\hat{\theta}|Z]$ has the same expectation as does $\hat{\theta}$ but smaller variance and the decrease in variance is largest if Z and $\hat{\theta}$ are nearly independent, because in this case $E[\hat{\theta}|Z]$ is close to a constant and its variance close to zero. In general the search for an appropriate Z so as to reducing the variance of an estimator by conditioning requires searching for a random variable Z such that:

1. the conditional expectation $E[\hat{\theta}|Z]$ with the original estimator is computable

2. $var(E[\hat{\theta}|Z])$ is substantially smaller than $var(\hat{\theta})$.

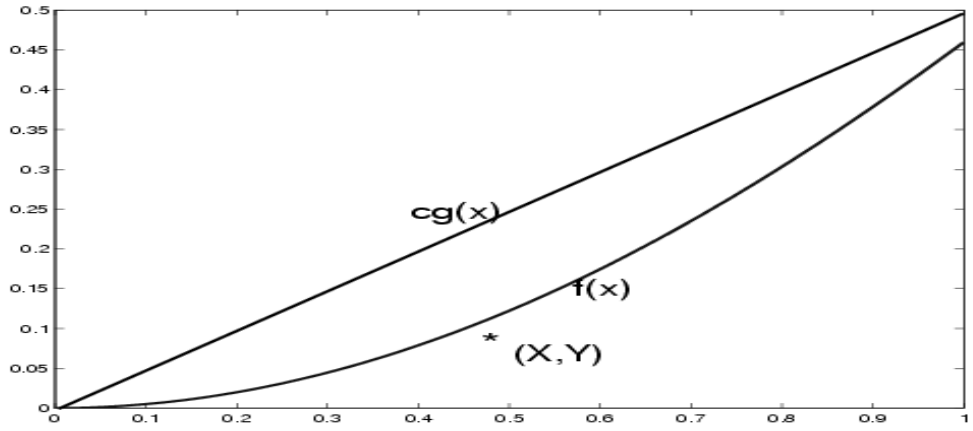


Figure 4.6: Example of the Hit and Miss Method

Example 42 *(hit or miss)*

Suppose we wish to estimate the area under a certain graph $f(x)$ by the hit and miss method. A crude method would involve determining a multiple c of a probability density function $g(x)$ which dominates $f(x)$ so that $cg(x) \geq f(x)$ for all x. We can generate points (X, Y) at random and uniformly distributed under the graph of $cg(x)$ by generating X by inverse transform $X = G^{-1}(U_1)$ where $G(x)$ is the cumulative distribution function corresponding to density g and then generating Y from the Uniform$[0, cg(X)]$ distribution, say $Y = cg(X)U_2$. An example, with $g(x) = 2x, 0 < x < 1$ and $c = 1/4$ is given in Figure 4.6.

The hit and miss estimator of the area under the graph of f obtains by generating such random points (X, Y) and counting the proportion that fall under the graph of g, i.e. for which $Y \leq f(X)$. This proportion estimates the probability

$$\begin{aligned} P[Y \leq f(X)] &= \frac{\text{area under } f(x)}{\text{area under } cg(x)} \\ &= \frac{\text{area under } f(x)}{c} \end{aligned}$$

since $g(x)$ is a probability density function. Notice that if we define

$$W = \begin{cases} c & \text{if} \quad Y \leq f(X) \\ 0 & \text{if} \quad Y > f(X) \end{cases}$$

then

$$\begin{aligned} E(W) &= c \times \frac{\text{area under } f(x)}{\text{area under } cg(x)} \\ &= \text{area under } f(x) \end{aligned}$$

so W is an unbiased estimator of the parameter that we wish to estimate. We might therefore estimate the area under $f(x)$ using a Monte Carlo estimator $\hat{\theta}_{HM} = \frac{1}{n}\sum_{i=1}^{n} W_i$ based on independent values of W_i.This is the "hit-or-miss" estimator. However, in this case it is easy to find a random variable Z such that the conditional expectation $E(Z|W)$ can be determined in closed form. In fact we can choose $Z = X$, we obtain

$$E[W|X] = \frac{f(X)}{g(X)}.$$

This is therefore an unbiased estimator of the same parameter and it has smaller variance than does W. For a sample of size n we should replace the crude estimator $\hat{\theta}_{cr}$ by the estimator

$$\begin{aligned} \hat{\theta}_{Cond} &= \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{2X_i} \end{aligned}$$

with X_i generated from $X = G^{-1}(U_i) = \sqrt{U_i}, i = 1, 2, ..., n$ and $U_i \sim$ Uniform[0,1]. In this case, the conditional expectation results in a familiar form for the estimator $\hat{\theta}_{Cond}$. This is simply an importance sampling estimator with $g(x)$ the importance distribution. However, this derivation shows that the estimator $\hat{\theta}_{Cond}$ has smaller variance than $\hat{\theta}_{HM}$.

Problems

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

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

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

   (b) 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%?

2. 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 $var(f(U) + f(1 - U)) = 0$?

3. Suppose that $F(x)$ is the normal(μ, σ^2) cumulative distribution function, Prove that $F^{-1}(1 - U) = 2\mu - F^{-1}(U)$ and therefore, if we use antithetic random numbers to generate two normal random variables X_1, X_2, having mean μ and variance σ^2, this is equivalent to setting $X_2 = 2\mu - X_1$. In other words, if we wish to use antithetic random numbers for normal variates, it is not necessary to generate the normal random variables using the inverse transform method.

4. Show that the variance of a weighted average

$$var(\alpha X + (1 - \alpha)W)$$

   is minimized over α when

$$\alpha = \frac{var(W) - cov(X, W)}{var(W) + var(X) - 2cov(X, W)}$$

   Determine the resulting minimum variance. What if the random variables X, W are independent?

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?

6. 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?

7. 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?

8. The second version of the control variate Monte-Carlo estimator

$$\widehat{\theta}_{cv} = \frac{1}{n} \sum_{i=1}^{n} \{f(U_i) - \beta[g(U_i) - E(g(U_i))]\}$$

is an improved control variate estimator, is equivalent to the first version in the case $\beta = 1$. 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]. Determine how much better $\widehat{\theta}_{cv}$ is than the basic control variate ($\beta = 1$) by performing simulations. Show that the variance is reduced by a factor of approximately 60 over crude Monte Carlo. Is there much additional improvement if we use a more general quadratic function of x for $g(x)$.

9. 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 Value at Risk, $Var_{.95}$, a value e^v such that $P[\sum_{i=1}^{5} X_i < v] = 0.95$. If we wish to do this by simulation, suggest an appropriate method involving importance sampling. Implement and estimate the variance reduction.

10. Suppose three different simulation estimators Y_1, Y_2, Y_3 have means which depend on two unknown parameters θ_1, θ_2 so that Y_1, Y_2, Y_3, are unbiased estimators of $\theta_1, \theta_1 + \theta_2, \theta_2$ respectively. Assume that $var(Y_i) = 1, cov(Y_i, Y_j) = -1/2$ an we want to estimate the parameter θ_1. Should we 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.

11. 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 (4.35) and (4.36), an importance sampling estimator and the control variate estimator above. What is the efficiency gain over crude Monte-Carlo?

12. Show that the Jacobian of the transformation used in the proof of Theorem 23; $(x, m) \to (x, y)$ where $y = \exp(-(2m-x)^2/2)$ is given by $\frac{\sqrt{1}}{2y\sqrt{-2\ln(y)}}$.

Chapter 5

Simulating the Value of Options

Asian Options

An Asian option, at expiration T, has value determined not by the closing price of the underlying asset as for a European option, but on an average price of the asset over an interval. For example a *discretely sampled Asian call option* on an asset with price process $S(t)$ pays an amount on maturity equal to $\max(0, \bar{S}_k - K)$ where $\bar{S}_k = \frac{1}{k}\sum_{i=1}^{k} S(iT/k)$ is the average asset price at k equally spaced time points in the time interval $(0, T)$. Here, k depends on the frequency of sampling (e.g. if $T = .25$ (years) and sampling is weekly, then $k = 13$. If $S(t)$ follows a geometric Brownian motion, then $\bar{S}_k$ is the sum of lognormally distributed random variables and the distribution of the sum or average of lognormal random variables is very difficult to express analytically. For this reason we will resort to pricing the Asian option using simulation. Notice, however that in contrast to the arithmetic average, the distribution of the *geometric average* has a distribution which can easily be obtained. The geometric

mean of n values $X_1, ..., X_n$ is $(X_1X_2...X_n)^{1/n} = \exp\{\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)\}$ and if the random variables X_n were each lognormally distributed then this results adding the normally distributed random variables $\ln(X_i)$ in the exponent, a much more familiar operation. In fact the sum in the exponent $\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ is normally distributed so the geometric average will have a lognormal distribution.

Our objective is to determine the value of the Asian option $E(V_1)$ with

$$V_1 = e^{-rT}max(0, \bar{S}_k - K)$$

Since we expect geometric means to be close to arithmetic means, a reasonable control variate is the random variable $V_2 = e^{-rT}max(0, \tilde{S}_k - K)$ where $\tilde{S}_k = \{\prod_{i=1}^{k} S(iT/k)\}^{1/k}$ is the geometric mean. 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 that of V_1 but in view of the fact that V_2 has a known lognormal distribution, the prospects of this are excellent. Since $S(t) = S_0e^{Y(t)}$ where $Y(t)$ is a Brownian motion with $Y(0) = 0$, drift $r - \sigma^2/2$ and diffusion σ, it follows that $\tilde{S}_k$ has the same distribution as does

$$S_0 \exp\{\frac{1}{k}\sum_{i=1}^{k} Y(iT/k)\}. \tag{5.1}$$

The exponent is a weighted average of the independent normal increments of the process and therefore normally distributed. In particular if we set

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

then we can find the mean and variance of $\bar{Y}$,

$$\begin{aligned} E(\bar{Y}) &= \frac{r-\sigma^2/2}{k}\sum_{i=1}^{k} iT/k \\ &= (r-\frac{\sigma^2}{2})\frac{k+1}{2k}T \\ &= \widetilde{\mu}T, \text{ say,} \end{aligned}$$

and

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

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. In fact

$$\begin{aligned} E(V_2) &= E\{e^{-rT}(S_0 e^{\overline{Y}} - K)^+\}, \text{ where } \overline{Y} \sim N(\widetilde{\mu}, \widetilde{\sigma}^2 T) \text{ so} \\ &= E[e^{-rT+\widetilde{\mu}T} S_0 e^{\overline{Y}-\widetilde{\mu}T} - e^{-rT}K]^+ \\ &= E[S_0 e^{(-r+\widetilde{\mu}+\frac{1}{2}\widetilde{\sigma}^2)T} \exp\{\overline{Y} - \widetilde{\mu}T - \frac{1}{2}\widetilde{\sigma}^2 T\} - e^{-rT}K]^+ \\ &= E[S_0 e^{(-r+\widetilde{\mu}+\frac{1}{2}\widetilde{\sigma}^2)T} \exp\{N(-\frac{\widetilde{\sigma}^2 T}{2}, \widetilde{\sigma}^2 T)\} - Ke^{-rT}]^+. \end{aligned}$$

where we temporarily denote a random variable with the Normal(μ, σ^2) distribution by $N(\mu, \sigma^2)$. Recall that the Black-Scholes formula gives the price at time $t = 0$ of a European option with exercise price K, initial stock price S_0,

$$BS(S_0, K, r, T, \sigma) = E(e^{-rT}(S_0 \exp\{N((r-\frac{\sigma^2}{2})T, \sigma^2 T)\} - K)^+ \tag{5.2}$$

$$= E(S_0 \exp\{N(-\frac{\sigma^2 T}{2}, \sigma^2 T)\} - Ke^{-rT})^+ \tag{5.3}$$

$$= S_0 \Phi(d_1) - Ee^{-rT}\Phi(d_2) \tag{5.4}$$

where

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}, d_2 = d_1 - \sigma\sqrt{T}.$$

Thus $E(V_2)$ is given by the Black-Scholes formula with S_0 replaced by

$$\widetilde{S_0} = S_0 \exp\{T(\frac{\widetilde{\sigma}^2}{2} + \widetilde{\mu} - r)\} = S_0 \exp\{-rT(1 - \frac{1}{k}) - \frac{\sigma^2 T}{12}(1 - \frac{1}{k^2})\}$$

and σ^2 by $\widetilde{\sigma}^2$. 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 at time T. For $k > 1$ the effective initial stock price is reduced $\widetilde{S_0} < S_0$ and the volatility parameter is also smaller $\widetilde{\sigma}^2 < \sigma^2$. With lower initial stock price and smaller volatility the price of a European call will decrease, indicating that an Asian option priced using a geometric mean has price lower than a similar European option on the same stock.

Recall from our discussion of a control variate estimators that we can estimate $E(V_1)$ unbiasedly using

$$V_1 - \beta(V_2 - E(V_2)) \tag{5.5}$$

where

$$\beta = \frac{cov(V_1, V_2)}{var(V_2)}. \tag{5.6}$$

In practice, of course, we simulate many values of the random variables V_1, V_2 and replace V_1, V_2 by their averages $\overline{V_1}, \overline{V}_2$ so the resulting estimator is

$$\overline{V_1} - \beta(\overline{V}_2 - E(V_2)). \tag{5.7}$$

Table 4.1 is similar to that in Boyle, Broadie and Glasserman(1997) and compares the variance of the crude Monte Carlo estimator with that of an estimator using a simple control variate,

$$E(V_2) + \overline{V_1} - \overline{V}_2,$$

a special case of (5.7) with $\beta = 1$. We chose $K = 100, k = 50, r = 0.10, T = 0.2$, a variety of initial asset prices S_0 and two values for the volatility parameter

$\sigma = 0.2$ and $\sigma = 0.4$. The efficiency depends only on S_0 and K through the ratio K/S_0 or alternatively the *moneyness* of the option, the ratio $e^{rT}S_0/K$ of the value on maturity of the current stock price to the strike price. Standard errors are estimated from $n = 10,000$ 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. Of course using the control variate estimator (5.7) described above could only improve the efficiency further.

Table 4.1. Standard Errors for Arithmetic Average Asian Options.

σ	Moneyness=$e^{rT}S_0/K$	STANDARD ERROR USING CRUDE MC	STANDARD ERROR USING CONTROL VARIATE
0.2	1.13	0.0558	0.0007
	1.02	0.0334	0.00064
	0.93	0.00636	0.00046
0.4	1.13	0.105	0.00281
	1.02	0.0659	0.00258
	0.93	0.0323	0.00227

The following function implements the control variate for an Asian option and was used to produce the above table.

```
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 if we use $K = 100$, we might confirm the last row of the above table using the command

asian(.1,100.93*exp(-r*T),.4,.2,100,50,10000)*;

Asian Options and Stratified Sampling

For many options, the terminal value of the stock has a great deal of influence on the option price. Although it is difficult in general to stratify samples of stock prices, it fairly easy to stratify along a single dimension, for example the dimension defined by the stock price at time T. In particular we may stratify the generation of

$$S_t = S_0 \exp(Z_t)$$

where Z_t can be written in terms of a standard Brownian motion

$$Z_t \ = \mu t + \sigma W_t, \quad \text{with } \mu = r - \sigma^2/2.$$

To stratify into K strata of equal probability for S_T we may generate Z_T using

$$Z_T = rT + \sqrt{rT - \sigma^2 T/2}\ \Phi^{-1}(i - 1 + \frac{U_i}{K}), i = 1, 2, ...K$$

for Uniform[0,1] random variables U_i and then randomly generate the rest of the path interpolating the value of S_0 and S_T using Brownian Bridge interpolation. To do this we use the fact that for a standard Brownian motion and $s < t < T$ we have that the conditional distribution of W_t given W_s, W_T is normal with mean a weighted average of the value of the process at the two endpoints

$$\frac{T-t}{T-s}W_s + \frac{t-s}{T-s}W_T$$

and variance

$$\frac{(t-s)(T-t)}{T-s}.$$

Thus given the value of S_T (or equivalently the value of $W(T)$) the increments of the process at times $\varepsilon, 2\varepsilon, ...N\epsilon = T$ can be generated sequentially so that the j'th increment $W(j\varepsilon) - W((j-1)\varepsilon)$ conditionally on the value of $W((j-1)\varepsilon)$ and of $W(T)$ has a Normal distribution with mean

$$(\frac{N-j}{N}W((j-1)\varepsilon) + \frac{j}{N}W(T)$$

and with variance

$$\frac{N-j}{N-j+1}.$$

Use of Girsanov's Lemma.

There are many other variance reduction schemes that one can apply to valuing an Asian Option. However prior to attacking this problem by other methods, let us consider a simpler example.

Importance Sampling and Pricing a European Call Option

Suppose we wish to estimate the value of a call option using Monte Carlo methods which is well "out of the money", one with a strike price K far above the current price of the stock S_0. If we were to attempt to evaluate this option using crude Monte Carlo, the majority of the values randomly generated for S_T would fall below the strike and contribute zero to the option price. One possible remedy is to generate values of S_T under a distribution that is more likely to exceed the strike, but of course this would increase the simulated value of the option. We can compensate for changing the underlying distibution by multiplying by a factor adjusting the mean as one does in importance sampling.

More specifically, we wish to estimate

$$E_Q[e^{-rT}(S_0e^{Z_T} - K)^+], \text{ where } Z_T \sim N(rT - \sigma^2T/2, \sigma^2T)$$

where E_Q indicates that the expectation is taken under a risk neutral distribution or probability measure Q for and K is large. Suppose that we modify the underlying probability measure of Z_T to Q_0, say a normal distribution with mean value $\ln(K/S_0) - \sigma^2 T/2$ but the same variance $\sigma^2 T$. Then the expected stock price under this new distribution

$$E_{Q_0} S_0 e^{Z_T} = S_0 \exp(E_{Q_0} Z_T + \sigma^2 T/2) = K$$

so there is a much greater probability (roughly 1/2) that the strike price is attained. The importance sampling adjustment that insures that the estimator continues to be an unbiased estimator of the option price is the ratio of two probability densities. Denote the normal probability density function by

$$\varphi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}.$$

Then the Radon-Nikodym derivative

$$\frac{dQ}{dQ_0}(z_T) = \frac{\varphi(z_t; rT - \frac{\sigma^2 T}{2}, \sigma^2 T)}{\varphi(z_t; \ln(K/S_0) - \frac{\sigma^2 T}{2}, \sigma^2 T)}$$

is simply the ratio of the two normal density functions with the two different means, and

$$\begin{aligned} E_Q[e^{-rT}(S_T - K)^+] &= E_{Q_0}[e^{-rT}(S_T - K)^+ \frac{dQ}{dQ_0}(Z_T)] \\ &= E_{Q_0}[e^{-rT}(S_0 e^{Z_T} - K)^+ \frac{\varphi(Z_T; rT - \frac{\sigma^2 T}{2}, \sigma^2 T)}{\varphi(Z_T; \ln(K/S_0) - \frac{\sigma^2 T}{2}, \sigma^2 T)}] \end{aligned}$$

so the importance sample estimator is the average of terms of the form

$$e^{-rT}(S_0 e^{Z_T} - K)^+ \frac{\varphi(Z_T; rT - \frac{\sigma^2 T}{2}, \sigma^2 T)}{\varphi(Z_T; \ln(K/S_0) - \frac{\sigma^2 T}{2}, \sigma^2 T)}, \text{ where } Z_T \sim N(\ln(K/S_0) - \frac{\sigma^2 T}{2}, \sigma^2 T).$$

The new simulation generates paths that are less likely to produce options expiring with zero value, and in a sense has thus eliminated some computational waste. What gains in efficiency result from this use of importance sampling? Let us consider a three month ($T = 0.25$) call option with $S_0 = 10$, $K = 15$,

$\sigma = 0.2$, $r = .05$. We determined the efficiency of the importance sampling estimator relative to using crude Monte Carlo in this situation using the function below. Running this using the command [eff,m,v]=importance2(10,.05,15,.2,.25) shows an efficiency gain of around 73, in part because very few of the crude estimates of S_T exceed the exercise price.

```
function [eff,m,v]=importance2(S0,r,K,sigma,T,N)
% simple importance sampling estimator of call option price
% outputs efficiency relative to crude. Run using
% [eff,m,v]=importance2(10,.05,15,.2,.25)
Z=randn(1,N);
%first do crude
ZT=(r-.5*sigma^2)*T+sigma*sqrt(T).*Z;
est1=exp(-r*T)*max(0,S0*exp(ZT)-K);
% now do importance
ZT=(log(K/S0)-.5*sigma^2)*T+sigma*sqrt(T).*Z;
ST=S0*exp(ZT);
est2=exp(-r*T)*max(0,ST-K).*normpdf(ZT,(r-.5*sigma^2)*T,sigma*sqrt(T))./normpdf(ZT,(log(K/S0)-
.5*sigma^2)*T,sigma*sqrt(T));
v=[var(est1) var(est2)];
m=[mean(est1) mean(est2)];
eff=v(1)/v(2);
```

Importance Sampling and Pricing an Asian Call Option

Let us now return to pricing an Asian option. We wish to use a variety of variance reduction techniques including importance sampling as in the above example, but in this case the relevant observation is not a simple stock price at one instant, but the whole stock price history from time 0 to T. An Asian option should nevertheless have payoff correlated with the value of the stock on

maturity $S(T)$. It might be reasonable to stratify the sample; i.e. sample more often when $S(T)$ is large or 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 $\dot{S(T)} > K$. As before if we do this we need to then multiply by the ratio of the two probability density functions, (the Radon Nikodym derivative of one process with respect to the other). This density is given by a result called Girsanov's Theorem (see Appendix B). The idea is as follows: Suppose P is the probability measure induced on the paths on $[0, T]$ by an Ito process

$$dS_t = \mu(S_t)dt + \sigma(S_t)dW_t, S_0 = s_0. \tag{5.8}$$

Similarly suppose P_0 is the probability measure on path generated by a similar process with the same diffusion term but different drift term

$$dS_t = \mu_0(S_t)dt + \sigma(S_t)dW_t, S_0 = s_0. \tag{5.9}$$

Note that in both cases, the process starts at the same initial value s_0.Then the "likelihood ratio" or the Radon-Nikodym $\frac{dP}{dP_0}$ of P with respect to P_0 is

$$\frac{dP}{dP_0} = \exp\{\int_0^T \frac{\mu(S_t) - \mu_0(S_t)}{\sigma^2(S_t)} dS_t - \int_0^T \frac{\mu^2(S_t) - \mu_0^2(S_t)}{2\sigma^2(S_t)} dt\} \tag{5.10}$$

We do not attempt to give a technical proof of this result, either here or in the appendix, since real "proofs" can be found in a variety of texts, including Steele (2004) and Karatzas and Shreve, (xxx). Instead we provide heuristic justification of (5.10). Let us consider the conditional distribution of a small increment dS_t in the process S_t under the model (5.8). Since this distribution is conditionally normal distributed it has conditional probability density function given the past

$$\frac{1}{\sqrt{2\pi dt}} \exp\{-(dS_t - \mu(S_t)dt)^2/(2\sigma^2(S_t)dt) \tag{5.11}$$

and under the model (5.9), it has the conditional probability density

$$\frac{1}{\sqrt{2\pi dt}} \exp\{-(dS_t - \mu_0(S_t)dt)^2/(2\sigma^2(S_t)dt) \tag{5.12}$$

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

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

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, we wish to determine by simulation the expected value of $V(r_T)$ for an interest rate model

$$dr_t = \mu(r_t)dt + \sigma dW_t \tag{5.13}$$

for some choice of function $\mu(r_t)$. Then according to Girsanov's theorem, we may simulate r_t under the Brownian motion model $dr_t = \mu_0 dt + \sigma dW_t$ (having the same initial value r_0 as in our original simulation) and then average the values of

$$V(r_T)\frac{dP}{dP_0} = V(r_T)\exp\{\int_0^T \frac{\mu(r_t)-\mu_0}{\sigma^2}dr_t - \int_0^T \frac{\mu^2(r_t)-\mu_0^2}{2\sigma^2}dt\} \tag{5.14}$$

So far, the constant μ_0 has been arbitrary and we are free to choose it in order to achieve as much variance reduction as possible. Ideally we do not want to get

too far from the original process so μ_0 should not be too far from the values of $\mu(r_t)$. In this case we hope that the term $\frac{dP}{dP_0}$ is not too variable (note that $c\frac{dP}{dP_0}$ would be the estimator if $V(S_T) = c$ were constant). On the other hand, the term $V(r_T)$ cannot generally be ignored, and there is no formula or simple rule for choosing parameters which minimize the variance of $V(r_T)\frac{dP}{dP_0}$. Essentially we need to resort to choosing μ_0 to minimize the variance of $V(r_T)\frac{dP}{dP_0}$ by experimentation, usually using some preliminary simulations.

Pricing a Call option under stochastic interest rates.

(REVISE MODEL)Again we consider pricing a call option, but this time under a more realistic model which permits stochastic interest rates. We will use the method of conditioning, although there are many other potential variance reduction tools here. Suppose the asset price, (under the risk-neutral probability measure, say) follows a geometric Brownian motion model of the form

$$dS_t = r_t S_t dt + \sigma S_t dW_t^{(1)} \tag{5.15}$$

where r_t is the spot interest rate. We assume r_t is stochastic and follows the Brennan-Schwartz model,

$$dr_t = a(b - r_t)dt + \sigma_0 r_t dW_t^{(2)} \tag{5.16}$$

where $W_t^{(1)}, W_t^{(2)}$are both Brownian motion processes and usually assumed to be correlated with correlation coefficient ρ. The parameter b in (5.16) can be understood to be the long run average interest rate (the value that it would converge to in the absence of shocks or resetting mechanisms) and the parameter $a > 0$ governs how quickly reversion to b occurs.

It would be quite remarkable if a stock price is completely independent of interest rates, since some of the same factors influence both. However we begin

by assuming something a little less demanding, that the random noise processes driving the asset price and stock price are independent or that $\rho = 0$.

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 assumption that $\rho = 0$ so that 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 so that in the conditional model it is known). While it may be difficult to obtain the value of an option under the model (5.15), (5.16) it is usually much easier under a 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

$$dZ_t = cZ_t dt + \sigma Z_t dW_t^{(1)}, Z_0 = S_0 \tag{5.17}$$

model "0" and denote the probability measure and expectations under this distribution by P_0 and E_0 respectively. 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

$$\ln(Z_T/Z_0) \text{ has a } N((c - \frac{\sigma^2}{2})T, \sigma^2 T) \quad \text{distribution}$$

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

$$E_0[exp\{-\int_0^T r_t dt\}(Z_T - K)^+ | r_s, 0 < s < T] \tag{5.18}$$
$$= EE_0[(S_0 e^{(c-\overline{r})T} e^W - e^{-\overline{r}T} K)^+ | r_s, 0 < s < T],$$
where W has a $N(-\sigma^2 T/2, \sigma^2 T)$
$$= E[BS(S_0 e^{(c-\overline{r})T}, K, \overline{r}, T, \sigma)], \text{ with } \overline{r} = \frac{1}{T}\int_0^T r_t dt.$$

Here, $\overline{r}$ is the average interest rate over the period and the function BS is the Black-Scholes formula (5.2). In other words by replacing 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 asset driven by (5.17) conditional on the value of $\overline{r}$. The constant interest rate model is a useful control variate for the more general model (5.16). The expected value $E[BS(S_0 e^{(c-\overline{r})T}, K, \overline{r}, T, \sigma)]$ 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 under (5.15),(5.16) 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 using common random numbers.

We are still able to make a choice of the constant c. One simple choice is $c \approx E(\overline{r})$ since this means that the second term is approximately $E\{BS(S_0, K, \overline{r}, T, \sigma)\}$. Alternatively we can again experiment with small numbers of test simulations and various values of c in an effort to roughly minimize the variance

$$var(exp\{-\int_0^T r_t dt\}[(S_T - K)^+ - (Z_T - K)^+]\}).$$

Evidently it is fairly easy to arrive at a solution in the case $\rho = 0$ since we really only need to average values of the Black Scholes price under various randomly generated interest rates. This does not work in the case $\rho \neq 0$ because

the conditioning involved in (5.18) does not result in the Black Scholes formula. Nevertheless we could still use common random numbers to generate two interest rate paths, one corresponding to $\rho = 0$ and the other to $\rho \neq 0$ and use the former as a control variate in the estimation of an option price in the general case.

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. By Girsanov's Theorem, $E\{V(S_T,T)\} = E_0\{V(S_T,T)\frac{dP}{dP_0}\}$ where P is the measure on the set of stock price paths corresponding to (5.15),(5.16) and P_0 that measure corresponding to (5.17). The required Radon-Nykodym derivative is

$$\frac{dP}{dP_0} = exp\{\int_0^T \frac{(r_t - c)S_t}{S_t^2\sigma^2}dS_t - \int_0^T \frac{(r_t^2 - c^2)S_t^2}{2\sigma^2 S_t^2}dt\} \tag{5.19}$$

$$= exp\{\int_0^T \frac{r_t - c}{S_t\sigma^2}dS_t - \int_0^T \frac{r_t^2 - c^2}{2\sigma^2}dt\} \tag{5.20}$$

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

$$V(S_T,T)exp\{-\int_0^T r_t dt + \int_0^T \frac{r_t - c}{\sigma^2 r_t}dS_t - \int_0^T \frac{r_t^2 - c^2}{2\sigma^2}dt\} \tag{5.21}$$

where the trajectories r_t are simulated under interest rate model (5.16).

As discussed before, we can attempt to choose the drift parameter c to approximately minimize the variance of the estimator (5.21).

Simulating Barrier and lookback options

For a financial times series X_t observed over the interval $0 \leq t \leq T$, what is recorded in newspapers is often just 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* $L = min\{X_t; 0 \leq t \leq T\}$. Very few uses of the highly informative variables H and L are made, partly becuase their distribution is a bit more complicated than that of the normal distribution of returns. Intuitively, however, the difference between H and L should carry a great deal of information about one of the most important parameters of the series, its volatility. Estimators of volatility obtained from the range of prices $H - L$ or H/L will be discussed in Chapter 6. In this section we look at how simple distributional properties of H and L can be used to simulate the values of certain exotic path-dependent options.

Here we consider options such as barrier options, lookback options and hindsight options whose value function depends only on the four variables (O, H, L, C) for a given process. Barrier options include knock-in and knock-out call options and put options. Barrier options are simple call or put options with a feature that should the underlying cross a prescribed barrier, the option is either knocked out (expires without value) or knocked in (becomes a simple call or put option). Hindsight options, also called fixed strike lookback options are like European call options in which we may use any price over the interval $[0, T]$ rather than the closing price in the value function for the option. Of course for a call option, this would imply using the high H and for a put the low L. A few of these path-dependent options are listed below.

Option	Payoff
Knock-out Call	$(C - K)^+ I(H \leq m)$
Knock-in Call	$(C - K)^+ I(H \geq m)$
Look-back Put	$H - C$
Look-Back Call	$C - L$
Hindsight (fixed strike lookback) Call	$(H - K)^+$
Hindsight (fixed strike lookback) put	$(K - L)^+$

Table 5.1: Value Function for some exotic options

For further details, see Kou et. al. (1999) and the references therein.

Simulating the High and the Close

All of the options mentioned above are functions of two or three variables O, C, and H or O, C, and L and so our first challenge is to obtain in a form suitable to calculation or simulation the joint distribution of these three variables. Our argument will be based on one of the simplest results in combinatorial probability, the reflection principle. We would like to be able to handle more than just a Black-Scholes model, both discrete and continuous distributions, and we begin with the simple discrete case. Much of the material in this section can be found in McLeish(2002).

In the real world, the market does not rigorously observe our notions of the passage of time. Volatility and volume traded vary over the day and by day of the week. A successful model will permit some variation in clock speed and volatility, and so we make an attempt to permit both in our discrete model.

In the discrete case, we will assume that the stock price S_t forms a trinomial tree, taking values on a set set $D = \{\cdots d_{-1} < d_0 < d_1 \cdots\}$... At each time point t, the stock may either increase, decrease, or stay in the same place and the probability of these movements may depend on the time. Specifically we assume that if $S_t = d_i$, then for some parameters $\theta, p_t, t = 1, 2, ...,$

$$P(S_{t+1} = d_j | S_t = d_i) = \frac{1}{k_t(\theta)} \times \begin{cases} p_t e^{\theta} & \text{if} \quad j = i+1 \\ 1 - 2p_t & \text{if} \quad j = i \\ p_t e^{-\theta} & \text{if} \quad j = i-1 \\ 0 & \text{otherwise} \end{cases} \tag{5.22}$$

where $k_t(\theta) = 1 + p_t(e^{\theta} + e^{-\theta} - 2)$ and $p_t \leq \frac{1}{2}$ for all t. If we choose all $p_t = \frac{1}{2}$, then this is a model of a simple binomial tree which either steps up or down at each time point. The increment in this process $X_{t+1} = S_{t+1} - S_t$ has mean

which depends on the time t except in the case $\theta = 0$

$$E(X_{t+1}|X_t = d_i) = \frac{p_t}{k_t(\theta)}\{(d_{i+1} - d_i)e^{\theta} - (d_i - d_{i-1})e^{-\theta}),$$

and variance, also time-dependent except in the case $\theta = 0$. The parameter θ describes one feature of this process which is not dependent on the time or the location of the process, since the log odds of a move up versus a move down is

$$2\theta = \log[\frac{P[\text{UP}]}{P[\text{DOWN}]}].$$

Suppose we label the states of the process so that $S_0 = d_0$ and there is a barrier at the point d_m where $m > 0$. The main result conerning the distribution of the high (or low) is the following:

Proposition 43 *Suppose a stock price S_t has dynamics determined by (5.22), and $S_0 = d_0$. Define*

$$H = \max_{0 \leq t \leq T} S_t \text{ and } C = S_T$$

Then for $u < m$,

$$P_\theta(H \geq d_m | C{=}d_u){=}\frac{P_0[C = d_{2m-u}]}{P_0[C = d_u]}, \text{ for } min(u, 0) < m, \tag{5.23}$$

$$= 1, \text{ for } min(u,0) \geq m$$

Proof. We wish to count the number of paths over an interval of time $[0, T]$ which touch or cross this barrier and end at a particular point $d_u, u < m$. Such a path is shown as a solid line in Figure 5.1 in the case that the points d_i are all equally spaced. Such a path has a natural "reflection" about the horizontal line at d_m. The reflected path is identical up to the first time τ that the original path touches the point d_m, and after this time, say at time $t > \tau$, the relected path takes the value d_{2m-i} where $S_t = d_i$. This path is the dotted line in Figure 5.1. Notice that if the original path ends at $d_u < d_m$, below the barrier, the reflected path ends at $d_{2m-u} > d_m$ or above the barrier. Each path touching

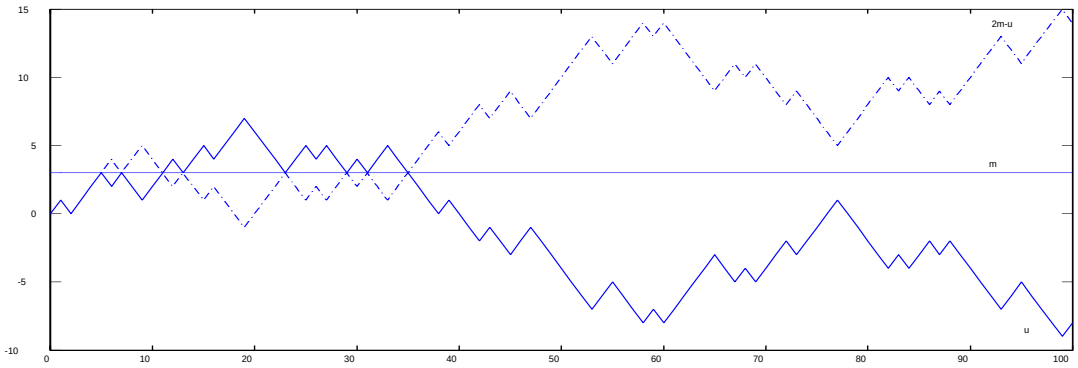


Figure 5.1: Illustration of the Reflection Principle

the barrier at least once and ending below it at d_u has a reflected path ending above it at d_{2m-u}, and of course each path that ends above the barrier must touch the barrier for a first time at some point and has a reflection that ends below the barrier. This establishes a one-one correspondence useful for counting these paths. Let us denote by the symbol "#" the "number of paths such that". Then:

$$\#\{H \geq d_m \text{ and } C = d_u < d_m\} = \#\{C = d_{2m-u}\}.$$

Now consider the probability of any path ending at a particular point d_u,

$$(S_0 = d_0, S_1, ..., S_T = d_u).$$

To establish this probability, each time the process jumps up in this interval we must multiply by the factor $\frac{p_t e^{\theta}}{k_t(\theta)}$ and each time there is a jump down we multiply by $\frac{p_t e^{-\theta}}{k_t(\theta)}$. If the process stays in the same place we multiply by $\frac{1-2p_t}{k_t(\theta)}$. The reflected path has exactly the same factors except that after the time τ at which the barrier is touched, the "up" jumps are replaced by "down" jumps and vice versa. For an up jump in the original path multiply by $e^{-2\theta}$. For a down jump in the original path, multiply by $e^{2\theta}$. Of course this allows us to compare path probabilities for an arbitrary value of the parameter θ, say with P_0, the

probability under $\theta = 0$ since, if the path ends at $C = d_u$,

$$\begin{aligned} P_\theta(\text{path}) &= \frac{e^{N_U\theta}e^{-N_D\theta}}{\prod_t k_t(\theta)} P_0[\text{path}] \\ &= \frac{e^{u\theta}}{\prod_t k_t(\theta)} P_0[\text{path}] \end{aligned} \tag{5.24}$$

where N_U and N_D are the number of up jumps and down jumps in the path. Note that we have subscripted the probability measure by the assumed value of the parameter θ. This makes it easy to compare the probabilities of the original and the reflected path, since

$$\frac{P_\theta[\text{original path}]}{P_\theta[\text{reflected path}]} = e^{-2\theta N_U}e^{2\theta N_D}$$

where now the number of up and down jumps N_U and N_D are counted following time τ. However, since $S_T = d_u$ and $S_\tau = d_m$, it follows that $N_D - N_U = m - u$ and that

$$\frac{P[\text{original path}]}{P[\text{reflected path}]} = e^{2\theta(u-m)}$$

which is completely independent of how that path arrived at the closing value d_u, depending only on the close. This makes it easy to establish the probability of paths having the property that $H \geq d_m$ and $C = d_u < d_m$ since there are exactly the same number of paths such that $C = d_{2m-u}$ and the probabilities of these paths differ by a constant factor $e^{2\theta(u-m)}$. Finally this provides the useful result for $u < m$.

$$P_\theta[H \geq d_m \text{ and } C = d_u] = e^{2\theta(u-m)}P_\theta[C = d_{2m-u}],$$

or, on division by $P[C = d_u]$,

$$\begin{aligned} P_\theta[H \geq d_m | C = d_u] &= \frac{e^{2\theta(u-m)}P_\theta[C = d_{2m-u}]}{P_\theta[C = d_u]} \\ &= \frac{e^{2\theta(u-m)}e^{\theta(2m-u)}P_0[C = d_{2m-u}]}{e^{\theta u}P_0[C = d_u]} \\ &= \frac{P_0[C = d_{2m-u}]}{P_0[C = d_u]} \end{aligned}$$

where we have used (5.24). This rather simple formula completely descibes the conditional distribution of the high under an arbitrary value of the parameter θ in terms of the value of the close under parameter value $\theta = 0$. ■

Thus, the conditional distribution of the high of a process given the open and close can be determined easily without knowledge of the underlying parameter and is related to the distribution of the close when the "drift" $\theta = 0$. This result also gives the expected value of the high in fairly simple form if the points d_j are equally spaced. Suppose $d_j = j\Delta$ for $j = 0, \pm 1, \pm 2, ...$ Then for $u = j\Delta$,with $j \geq 0$ and $k \geq 1$, (see Problem 1)

$$P_\theta[H - C \geq k\Delta | C = j\Delta] =$$
$$E[H|C = u] = u + \Delta \frac{P[C > u \text{ and } \frac{C-u}{\Delta} \text{ is even}]}{P[C = u]}.$$

Roughly, (5.23) indicates that if you can simulate the close under θ, then you use the properties of the close under $\theta = 0$ to simulate the high of the process. Consider the problem of simulating the high for a given value of the close $C = S_T = d_u$ and again assuming that $S_0 = d_0$. Suppose we use inverse transform from a uniform random variable U to solve the inequalities

$$P_\theta(\max_{0 \leq t \leq T} S_t \geq d_{m+1} | S_T = d_u) < U \leq P_\theta(\max_{0 \leq t \leq T} S_t \geq d_m | S_T = d_u)$$

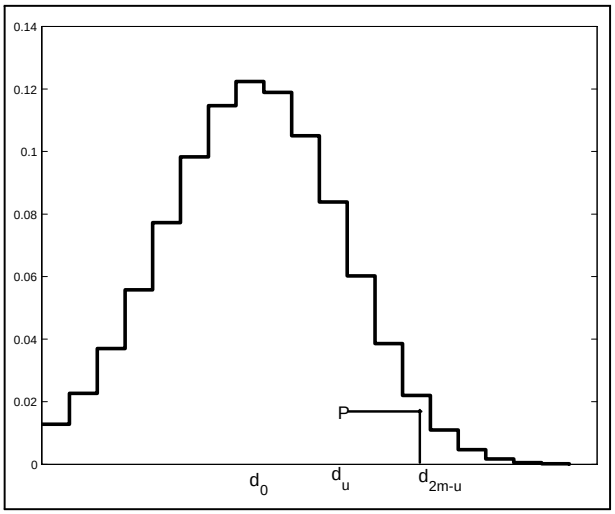
for the value of d_m. In this case the value of

$$d_m = \sup\{d_j; UP_0[S_T = d_u] \leq P_0[S_T = d_{2j-u}]\}$$

is the generated value of the high. This inequality is equivalent to

$$P_0[S_T = d_{2m+2-u}] < UP_0[S_T = d_u] \leq P_0[S_T = d_{2m-u}].$$

Graphically this inequality is demonstrated in Figure **??** which shows the probability histogram of the distribution S_T under $\theta = 0$. The value $UP_0[S_T = d_u]$ is the y-coordinate of a point P



Generating a High for a discrete distribution

randomly chosen from the bar corresponding to the point d_u. The high d_m is generated by moving horizontally to the right an even number of steps until just before exiting the histogram. This is above the value d_{2m-u} and d_m is between d_u and d_{2m-u}.

A similar result is available for Brownian motion and Geometric Brownian motion. A justification of these results can be made by taking a limit in the discrete case as the time steps and the distances $d_j - d_{j-1}$ all approach zero. If we do this, the parameter θ is analogous to the drift of the Brownian motion. The result for Brownian motion is as follows:

Theorem 44 *Suppose S_t is a Brownian motion process*

$$dS_t = \mu dt + \sigma dW_t,$$

$$S_0 = 0,\ S_T = C,$$

$$H = \max\{S_t; 0 \le t \le T\} \text{ and}$$

$$L = \min\{S_t; 0 \le t \le T\}.$$

If f_0 denotes the Normal(0,$\sigma^2 T$) probability density function, the distribution of

C under drift $\mu = 0$, then

$$U_H = \frac{f_0(2H - C)}{f_0(C)} \text{ is distributed as } U[0,1] \text{ independently of } C,$$

$$U_L = \frac{f_0(2L - C)}{f_0(C)} \text{ is distributed as } U[0,1] \text{ independently of } C.$$

$$Z_H = H(H - C) \text{ is distributed as Exponential } (\frac{1}{2}\sigma^2 T) \text{ independently of } C,$$

$$Z_L = L(L - C) \text{ is distributed as Exponential } (\frac{1}{2}\sigma^2 T) \text{independently of } C.$$

We will not prove this result since it is a special case of Theorem 46 below. However it is a natural extension of Proposition 43 in the special case that $d_j = j\Delta$ for some Δ and so we will provide a simple sketch of a proof using this proposition. Consider the ratio

$$\frac{P_0[C = d_{2m-u}]}{P_0[C = d_u]}$$

on the right side of (5.23). Suppose we take the limit of this as $\Delta \to 0$ and as $m\Delta \to h$ and $u\Delta \to c$. Then this ratio approaches

$$\frac{f_0(2h - c)}{f_0(c)}$$

where f_0 is the probability density function of C under $\mu = 0$. This implies for a Brownian motion process,

$$P[H \geq h | C = c] = \frac{f_0(2h - c)}{f_0(c)} \text{ for } h \geq c. \tag{5.25}$$

If we temporarily denote the cumulative distribution function of H given $C = c$ by $G_c(h)$ then (5.25) gives an expression for $1 - G_c(h)$ and recall that since the sumulative distribution function is continuous, when we evaluate it at the observed value of a random variable we obtain a $U[0,1]$ random variable e.g. $G_c(H) \sim U[0,1]$. In other words conditional on $C = c$ we have

$$\frac{f_0(2H - c)}{f_0(c)} \sim U[0,1].$$

This result verifies a simple geometric procedure, directly analogous to that in Figure 5.2, for generating H for a given value of $C = c$. Suppose we generate a point $P_H = (c, y)$ under the graph of $f_0(x)$ and uniformly distributed

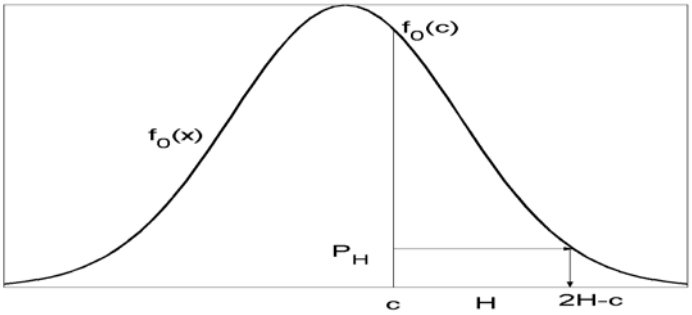


Figure 5.2: Generating H for a fixed value of C for a Brownian motion.

on $\{(c, y); 0 \leq y \leq f_0(c)\}$. This point is shown in Figure **??**. We regard the $y-$coordinate of this point as the generated value of $f_0(2H - c)$. Then H can be found by moving from P_H horizontally to the right until we strike the graph of f_0 and then moving vertically down to the axis (this is now the point $2H - c$) and finally taking the midpoint between this coordinate $2H - c$ and the close c to obtain the generated value of the high H. The low of the process can be generated in the same way but with a different point P_L uniform on the set $\{(c, y); 0 \leq y \leq f_0(c)\}$. The algorithm is the same in this case except that we move horizontally to the left.

There is a similar argument for generating the high under a geometric Brownian motion as well, since the logarithm of a geometric Brownian motion is a Brownian motion process.

Corollary 45 *For a Geometric Brownian motion process*

$$dS_t = \mu S_t dt + \sigma S_t dW_t,$$
$$S_0 = O \text{ and } S_T = C$$

with f_0 *the normal* $(0, \sigma^2 T)$ *probability density function, we have*

$$\ln(H/O)\ln(H/C) \sim \exp(\frac{1}{2}\sigma^2 T) \quad \textit{independently of } O, C \textit{ and}$$

$$\ln(L/O)\ln(L/C) \sim \exp(\frac{1}{2}\sigma^2 T) \quad \textit{independently of } O, C.$$

$$U_H = \frac{f_0(\ln(H^2/OC))}{f_0(\ln(C/O))} \sim U[0,1] \textit{ independently of } O, C \textit{ and}$$

$$U_L = \frac{f_0(\ln(L^2/OC))}{f_0(\ln(C/O))} \sim U[0,1] \textit{ independently of } O, C.$$

Both of these results are special cases of the following more general Theorem. We refer to McLeish(2002) for the proof. As usual, we consider a price process S_t and define the high $H = max\{S_t; 0 \leq t \leq T\}$, and the open and close $O = S_0$, $C = S_T$.

Theorem 46 *Suppose the process* S_t *satisfies the stochastic differential equation:*

$$dS_t = \{\nu + \frac{1}{2}\sigma'(S_t)\}\sigma(S_t)\lambda^2(t)dt + \sigma(S_t)\lambda(t)dW_t \tag{5.26}$$

where $\sigma(x) > 0$ *and* $\lambda(t)$ *are positive real-valued functions such that* $g(x) = \int^x \frac{1}{\sigma(y)}dy$ *and* $\theta = \int_0^T \lambda^2(s)ds < \infty$ *are well defined on* $\Re^+$.

(a) Then with f_0 *the* $N(0, \theta)$ *probability density function 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* $\frac{1}{2}\theta$.

A similar result holds for the low of the process over the interval, namely that

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

and $Z_H = \{g(L) - g(O)\}\{g(L) - g(C)\}$ is independent of O, C, and has an exponential distribution with mean $\frac{1}{2}\theta$.

Before we discuss the valuation of various options, we examing the significance of the ratio appearing in on the right hand side of (5.25) a little more closely. Recall that f_0 is the $N(0, \sigma^2 T)$ probability density function and so we can replace it by

$$\frac{f_0(2h-c)}{f_0(c)} = \frac{\exp\{-\frac{(2h-c)^2}{2\sigma^2 T}\}}{\exp\{-\frac{c^2}{2\sigma^2 T}\}} = \exp\{-2\frac{z_h}{\sigma^2 T}\} \tag{5.27}$$

where $z_h = h(h-c)$ or in the more general case where $S(0) = O$ may not be equal to zero,

$$z_h = (h - O)(h - c). \tag{5.28}$$

This ratio $\frac{f_0(2h-c)}{f_0(c)}$ represents the probability that a particular process with close c breaches a barrier at h and so the exponent

$$2\frac{z_h}{\sigma^2 T}$$

in the right hand side of (5.27) controls the probability of this event.

Of course we can use the above geometric algorithm for Brownian motion to generated highs and closing prices for a geometric Brownian motion, for example, S_t satisfying $d\ln(S_t) = \sigma dW_t$ (minor adjustments required to accommodate nonzero drift). The graph of the normal probability density function $f_0(x)$ of $\ln(C)$ is shown in Figure **??**.

If a point P_H is selected at random uniformly distributed in the region below the graph of this density, then, by the usual arguments supporting the acceptance rejection method of simulation, the x-coordinate of this point is a variate generated from the probability density function $f_0(x)$, that is, a simulated value from the distribution of $\ln(C)$. The y-coordinate of such a randomly selected point also generates the value of the high as before.If we extend a line horizontally to the right from P_H until it strikes the graph of the probability density and

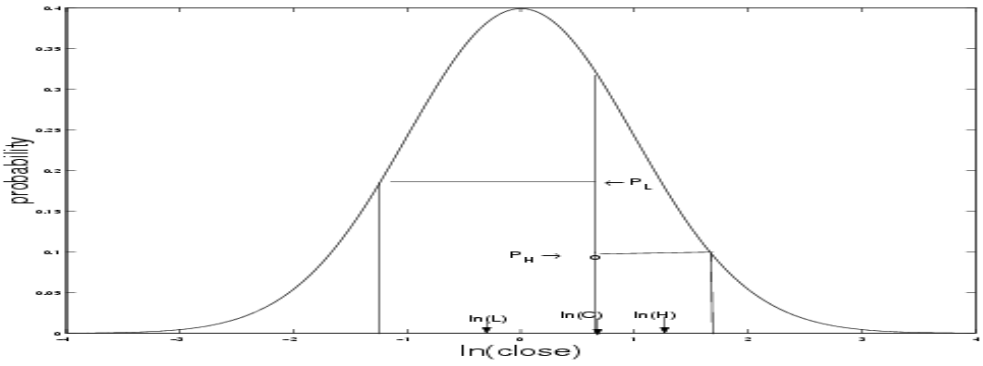


Figure 5.3: Simulating from the joint distribution of (H, C) or from (L, C)

then consider the abscissa, of this point, this is the simulated value of $\ln(H^2/C)$, and $\ln(H)$ the average the simulated values of $\ln(H^2/C)$ and $\ln(C)$.

A similar point P_L uniform under the probability density function f_0 can be used to generate the low of the process if we extend the line from P_L to the left until it strikes the density. Again the abscissa of this point is $\ln(L^2/C)$ and the average with $\ln(C)$ gives a simulated value of $\ln(L)$. Although the $y-$coordinate of both P_H and P_L are uniformly distributed on $[0, f_0(C)]$ conditional on the value of C they are not independent.

Suppose now we wish to price a barrier option whose payoff on maturity depends on the value of the close C but provided that the high H did not exceed a certain value, the barrier. This is an example of an knock-out barrier but other types are similarly handled. Once again we assume the simplest form of the geometric Brownian motion $d\ln(S_t) = \sigma dW_t$ and assume that the upper barrier is at the point Oe^b so that the payoff from the option on maturity T is

$$\psi(C)I(H < Oe^b)$$

for some function ψ. It is clear that the corresponding value of H does not exceed a boundary at Oe^b if and only if the point P_H is below the graph of the

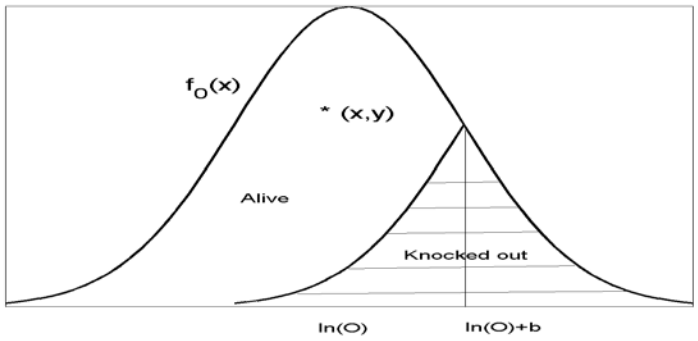


Figure 5.4: Simulating a knock-out barrier option with barrier at Oe^b

probability density function but **not** in the shaded region obtained by reflecting the right hand tail of the density about the vertical line $x = b - \ln(O)$ in Figure 5.4. To simulate the value of the option, choose points uniformly under the graph of the probability density $f_0(x)$. For those points in the non-shaded region under f_0 (the x-coordinate of these points are simulated values $\psi(C)$of $\ln(C)$ under the condition that the barrier is not breached) we average the values of $\psi(C)$ and for those in the shaded region we average 0.

Equivalently,

$$E\psi(C)I(H < Oe^b) = E\psi^*(C)$$

where

$$\psi^*(C) = \begin{cases} \psi(C) & \text{for} \quad C \leq Oe^b \\ -\psi(2b + \ln(O^2/C)) & \text{for} \quad C > Oe^b \end{cases} .$$

and so the barrier option can be priced as if it were a vanilla European option with payoff function $\psi^*(C)$.

Any option whose value depends on the high and the close of the process (or (L, C)) can be similarly valued as a European option. If an option becomes worthless whenever an upper boundary at Oe^b is breached, we need only

multiply the payoff from the option ignoring the boundary by the factor

$$1 - \exp\{-2\frac{z_h}{\sigma^2 T}\}$$

with

$$z_h = b(b + \ln(O/C))$$

to accommodate the filtering effect of the barrier and then value the option as if it were a European option.

There is a a variety of distributional results related to H, some used by Redekop (1995) to test the local Brownian nature of various financial time series. These are easily seen in Figure 5.5. For example, for a Brownian motion process with sero drift, suppose we condition on the value of $2H - O - C$. Then the point P_H must lie (uniformly distributed) on the line $\mathcal{L}_1$ and therefore the point H lies uniformly on this same line but to the right of the point O. This shows that conditional on $2H - O - C$ the random variable $H - O$ is uniform or,

$$\frac{H - O}{2H - O - C} \sim U[0, 1].$$

Similarly, conditional on the value of H, the point P_H must fall somewhere on the curve labelled $\mathcal{C}_2$ whose y-coordinate is uniformly distributed showing that

$$\frac{C - O}{2H - O - C} \sim U[0, 1].$$

Redekop shows that for a Brownian motion process, the statistic

$$\frac{H - O}{2H - O - C} \tag{5.29}$$

is supposed to be uniformly $[0, 1]$ distributed but when evaluated using real financial data, is far too often close to or equal the extreme values 0 or 1.

The joint distribution of (C, H) can also be seen from Figure 5.6. Note that the rectangle around the point (x, y) of area $\Delta x \Delta y$ under the graph of the density, when mapped into values of the high results in an interval of values for

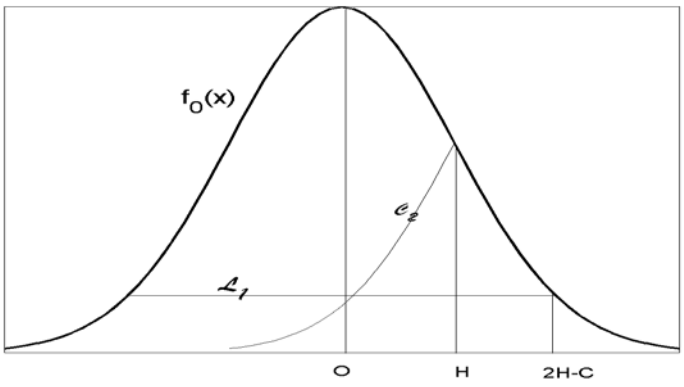


Figure 5.5: Some uniformly distributed statistics for Brownian Motion

$(2H - C)$ of width $-\Delta y/\phi'(2y - x)$ where ϕ' is the derivative of the standard normal probability density function (the minus sign is to adjust for the negative slope of the density here). This interval is labelled $\Delta(2H - C)$. This, in turn generates the interval ΔH of possible values of H, of width exactly half this, or

$$\frac{-\Delta y}{2\phi'(2y - x)}.$$

Inverting this relationship between (x, y) and (H, C),

$$P[H \in \Delta H, C \in \Delta C] = -2\phi'(2y - x)\Delta x \Delta y$$

confirming that the joint density of (H, C) is given by $-2\phi'(2y - x)$ for $x < y$.

In order to get the joint density of the High and the Close when the drift is non-zero, we need only multiply by the ratio of the two normal density functions of the close

$$\frac{f_\mu(x)}{f_0(x)}$$

and this gives the more general result in the table below.

The table below summarizes many of our distributional results for a Brownian motion process with drift on the interval $[0, 1]$,

$$dS_t = \mu dt + \sigma dW_t, \text{ with } S_0 = O.$$

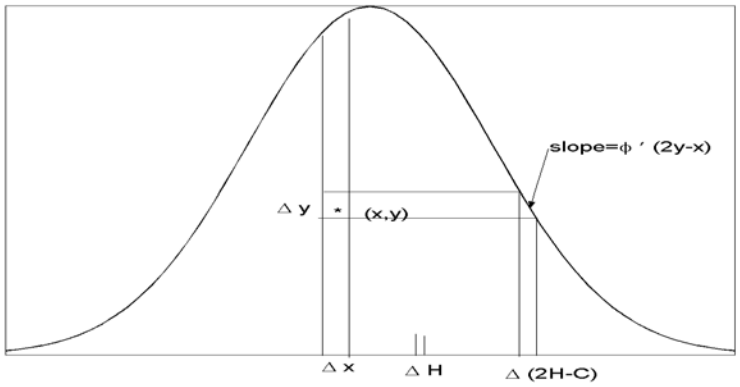


Figure 5.6: Confirmation of the joint density of (H, C)

Statistic	Density	Conditions
$X = C - O$, $Y = H - O$	$f(y,x) = -2\phi'(2y-x)\exp(\mu x - \mu^2/2)$	$-\infty < x < y$, and $y > 0, \sigma = 1$ given O
$Y\|X$	$f_{Y\|X}(y\|x) = 2(2y-x)e^{-2y(y-x)}$	$y > x, \sigma = 1$
$Z = Y(Y-X)$	$\exp((\sigma^2/2)$	given O, X
$(L-O)(L-C)$	$\exp(\sigma^2/2)$	given (O, C)
$(H-O)(H-C)$	$\exp(\sigma^2/2)$	given (O, C)
$\frac{H-O}{2H-O-C}$	$U[0,1]$	drift $\nu = 0$, given $O, 2H-O-C$
$\frac{L-O}{2L-O-C}$	$U[0,1]$	drift $\nu = 0$, given $O, 2L-O-C$
$\frac{C-O}{2H-O-C}$	$U[-1,1]$	drift $\nu = 0$, given H, O

TABLE 5.2: Some distributional results for High, Close and Low.

We now consider briefly the case of non-zero drift for a geometric Brownian motion. Fortunately, all that needs to be changed in the results 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. Suppose an option has payoff on

maturity $\psi(C)$ if an upper barrier at level $Oe^b, b > 0$ is not breached. We have already seen that to accommodate the filetering effect of this knock-out barrier we should determine, numerically or by simulation, the expected value

$$E[\psi(C)(1 - \exp\{-2\frac{b(b + \ln(O/C))}{\sigma^2 T}\})]$$

the expectation conditional (as always) on the value of the open O. The effect of a knock-out lower barrier at Oe^{-a} is essentially the same but with b replaced by a, namely

$$E[\psi(C)(1 - \exp\{-2\frac{a(a + \ln(C/O))}{\sigma^2 T}\})].$$

In the next section we consider the effect of two barriers, both an upper and a lower barrier.

One Process, 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 $P(H \leq b, L \geq a, C = u|O = 0]$ A first step in this direction is the the following result, obtained from the reflection principle with two barriers.

Theorem 47 *For a Brownian motion process*

$$dS_t = \mu dt + dW_t, S_0 = 0$$

defined on $[0, 1]$ *and for* $-a < u < b$,

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

where ϕ *is the* $N(0, 1)$ *probability density function.*

Proof. The proof is a well-known application of the reflection principal. It is sufficient to prove the result in the case $\mu = 0$ since the conditional distribution of L, H given C does not depend on μ (A statistician would say that C is a sufficient statistic for the drift parameter). Denote the following paths determined by their behaviour on $0 < t < 1$. All paths are assumed to end at $C = u$.

A_{+1} = $H > b$ (path goes above b)

A_{+2} = path goes above b and then falls below $-a$

A_{+3} = goes above b then falls below $-a$ then rises above b

etc.

A_{-1} = $L < -a$

A_{-2} = path falls below $-a$ then rises above b

A_{-3} = falls below $-a$ then rises above b then falls below $-a$

etc.

For an arbitrary event A, denote by $P(A|u)$ probability of the event conditional on $C = u$. Then according to the reflection principal the probability that the Brownian motion leaves the interval $[-a, b]$ is given from an inclusion-exclusion argument by

$$P(A_{+1}|u) - P(A_{+2}|u) + P(A_{+3}|u) - \cdots \tag{5.30}$$
$$+P(A_{-1}|u) - P(A_{-2}|u) + P(A_{-3}|u) \cdots$$

This can be verified by considering the paths in Figure 5.7. (It should be noted here that, as in our application of the reflection principle in the one-barrier case, the reflection principle allows us to show that the number of paths in two sets is the same, and this really only translates to probability in the case of a discrete sample space, for example a simple random walk that jumps up or down by a fixed amount in discrete time steps. This result for Brownian motion obtains if we take a limit over a sequence of simple random walks approaching a Brownian motion process.)

Note that

$$P(A_{+1}|u) = \frac{\phi(2b-u)}{\phi(u)}$$
$$P(A_{+2n}|u) = \frac{\phi\{2n(a+b)+u\}}{\phi(u)}$$
$$P(A_{+(2n-1)}|u) = \frac{\phi\{2n(a+b)-2a-u\}}{\phi(u)}$$

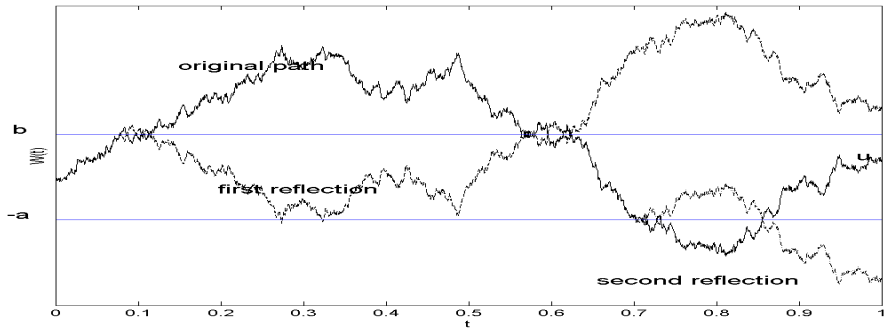


Figure 5.7: The Reflection principle with Two Barriers

and

$$P(A_{-1}|u) = \frac{\phi(-2a-u)}{\phi(u)}$$
$$P(A_{-2n}|u) = \frac{\phi\{-2n(b+a)+u\}}{\phi(u)}$$
$$P(A_{-(2n+1)}|u) = \frac{\phi\{2n(b+a)+2a+u\}}{\phi(u)}.$$

The result then obtains from substitution in (5.30). ■

As a consequence of this result we can obtain an expression for $P(a < L \leq H < b, u < C < v)$ (see also Billingsley, (1968), p. 79) for a Brownian motion on $[0,1]$ with zero drift:

$$\begin{aligned} P(a,b,u,v) &= P(a < L \leq H < b, u < C < v) \\ &= \sum_{k=-\infty}^{\infty} \Phi[v+2k(b-a)] - \Phi[u+2k(b-a)] \\ &- \sum_{k=-\infty}^{\infty} \Phi[2b-u+2k(b-a)] - \Phi[2b-v+2k(b-a)]. \end{aligned} \tag{5.31}$$

where Φ is the standard normal cumulative distribution function. From (5.31) we derive the joint density of (L,H,C) by taking the limit $P(a,b,u,u+\delta)/\delta$ as

$\delta \to 0$, and taking partial derivatives with respect to a and b:

$$\begin{aligned} f(a,b,u) &= 4\sum_{k=-\infty}^{\infty} k^2\phi''[u+2k(b-a)] - k(1+k)\phi''[2b-u+2k(b-a)] \\ &= 4\sum_{k=1}^{\infty} k^2\phi''[u+2k(b-a)] - k(1+k)\phi''[2b-u+2k(b-a)] \\ &\quad + k^2\phi''[u-2k(b-a)] + k(1-k)\phi''[2b-u-2k(b-a)] \end{aligned} \tag{5.32}$$

for $a < u < b$.

From this it is easy to see that the conditional cumulative distribution function of L given $C = u, H = b$ is given by on $a \leq u \leq b$ (where $-2\phi'(2b-u)$ is the joint p.d.f. of H, C) by

$$\begin{aligned} F(a|b,u) &= 1 + \frac{\frac{\partial^2}{\partial b \partial v} P(a,b,u,v)|_{v=u}}{2\phi'(2b-u)} \\ &= \frac{-1}{\phi'(2b-u)} \sum_{k=1}^{\infty} \{-k\phi'[u+2k(b-a)] + (1+k)\phi'[2b-u+2k(b-a)] \\ &\quad + k\phi'[u-2k(b-a)] + (1-k)\phi'[2b-u-2k(b-a)]\} \end{aligned} \tag{5.33}$$

This allows us to simulate both the high and the low, given the open and the close by first simulating the high and the close using $-2\phi'(2b-u)$ as the joint p.d.f. of (H, C) and then simulating the low by inverse transform from the cumulative distribution function of the form (5.33).

Survivorship Bias

It is quite common for retrospective studies in finance, medicine and to be subject to what is often called "survivorship bias". This is a bias due to the fact that only those members of a population that remained in a given class (for example the survivors) remain in the sampling frame for the duration of the study. In general, if we ignore the "drop-outs" from the study, we do so at risk of introducing substantial bias in our conclusions, and this bias is the survivorship bias.

Suppose for example we have hired a stable of portfolio managers for a large pension plan. These managers have a responsibility for a given portfolio over a period of time during which their performance is essentially under continuous review and they are subject to one of several possible decisions. If returns below a given threshhold, they are deemed unsatisfactory and fired or converted to another line of work. Those with exemplary performance are promoted, usually to an administrative position with little direct financial management. And those between these two "absorbing" barriers are retained. After a period of time, T, an amibitious graduate of an unnamed Ivey league school working out of head office wishes to compare performance of those still employed managing portfolios. How are should the performance measures reflect the filtering of those with unusually good or unusually bad performance? This is an example of a process with upper and lower absorbing barriers, and it is quite likely that the actual value of these barriers differs from one employee to another, for example the son-in-law of the CEO has a substantially different barriers than the math graduate fresh out of UW. However, let us ignore this difference, at least for the present, and concentrate on a difference that is much harder to ignore in the real world, the difference between the volatility parameters of portfolios, possibly in different sectors of the market, controlled by different managers. For example suppose two managers were responsible for funds that began and ended the year at the same level and had approximately the same value for the lower barrier as in Table 5.2. For each the value of the volatility parameter σ was estimated using individual historical volatilities and correlations of the component investments.

Portfolio	Open price	Close Price	Lower Barrier	Volatility
1	40	$56\frac{5}{8}$	30	.5
2	40	$56\frac{1}{4}$	30	.2

Table 5.3

Suppose these portfolios (or their managers) have been selected retrospectively from a list of "survivors" which is such that the low of the portfolio value never crossed a barrier at $l = Oe^{-a}$ (bankruptcy of fund or termination or demotion of manager, for example) and the high never crossed an upper barrier at $h = Oe^{b}$. However, for the moment let us assume that the upper barrier is so high that its influence can be neglected, so that the only absorbtion with any substantial probability is at the lower barrier. We interested in the estimate of return from the two portfolios, and a preliminary estimate indicates a continuously compounded rate of return from portfolio 1 of $R_1 = \ln(56.625/40) = 35\%$ and from portfolio two of $R_2 = \ln(56.25/40) = 34\%$. Is this difference significant and are these returns reasonably accurate in view of the survivorship bias?

We assume a geometric Brownian motion for both portfolios,

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \tag{5.34}$$

$$\text{and define } O = S(0), \qquad C = S(T),$$

$$H = \max_{0 \leq t \leq T} S(t), \quad L = \min_{0 \leq t \leq T} S(t)$$

with parameters μ, σ possibly different.

In this case it is quite easy to determine the expected return or the value of any performance measure dependent on C conditional on survival, since this is essentially the same as a problem already discussed, the valuation of a barrier option. According to (5.27), the probability that a given Brownian motion process having open 0 and close c strikes a barrier placed at $l < \min(0, c)$ is

$$\exp\{-2\frac{z_l}{\sigma^2 T}\}$$

with

$$z_l = l(l - c).$$

Converting this statement to the Geometric Brownian motion (5.34), the probability that a geometric Brownian motion process with open O and close c

breaches a lower barrier at l is

$$P[L \leq l|O, C] = \exp\{-2\frac{z_l}{\sigma^2 T}\}$$

with

$$z_l = \ln(O/l)\ln(C/l) = a(a + \ln(C/O)).$$

Of course the probability that a particular path with this pair of values (O, C) is a "survivor" is 1 minus this or

$$1 - \exp\{-2\frac{z_l}{\sigma^2 T}\}. \tag{5.35}$$

When we observe the returns or the closing prices C of survivors only, the results have been filtered with probability (5.35). In other words if the probability density function of C without any barriers at all is $f(c)$ (in our case this is a lognormal density with parameters that depend on μ and σ) then the density function of C of the survivors in the presence of a lower barrier is proportional to

$$\begin{aligned} &f(c)[1 - \exp\{-2\frac{\ln(O/l)\ln(c/l)}{\sigma^2 T}\}] \\ &= f(c)(1 - (\frac{l}{c})^\lambda), \text{ with } \lambda = \frac{2\ln(O/l)}{\sigma^2 T} = \frac{2a}{\sigma^2 T} > 0. \end{aligned}$$

It is interesting to note the effect of this adjustment on the moments of C for various values of the parameters. For example consider the expected value of C conditional on survival

$$\begin{aligned} E(C|L \geq l] &= \frac{\int_l^\infty cf(c)(1 - (\frac{l}{c})^\lambda)dc}{\int_l^\infty f(c)(1 - (\frac{l}{c})^\lambda)dc} \\ &= \frac{E[CI(C \geq l)] - l^\lambda E[C^{1-\lambda} I(C \geq l)]}{P[C \geq l] - l^\lambda E[C^{-\lambda} I(C \geq l)]} \end{aligned} \tag{5.36}$$

and this is easy to evaluate in the case of interest in which C has a lognormal distribution. In fact the same kind of calculation is used in the development of the Black-Scholes formula. In our case $C = \exp(Z)$ where Z is $N(\mu T, \sigma^2 T)$

and so for any p and $l > 0$, we have from (3.11), using the fact that $E(C|O) = O\exp\{\mu T + \sigma^2 T/2\}$, (and assuming O is fixed),

$$E[C^p I(C > l)] = O^p \exp\{p\mu T + p^2\sigma^2 T/2\}\Phi(\frac{1}{\sigma\sqrt{T}}(a + \mu T) + \sigma\sqrt{T}p)$$

To keep things slightly less combersome, let us assume that we observe the geometric Brownian motion for a period of $T = 1$. Then (5.36) results in

$$\frac{Oe^{\mu+\sigma^2/2}\Phi(\frac{1}{\sigma}(a+\mu)+\sigma) - Oe^{-a\lambda+(1-\lambda)\mu+(1-\lambda)^2\sigma^2/2}\Phi(\frac{1}{\sigma}(a+\mu)+\sigma(1-\lambda))}{\Phi(\frac{1}{\sigma}(a+\mu)) - e^{-\lambda a-\lambda\mu+\lambda^2\sigma^2/2}\Phi(\frac{1}{\sigma}(a+\mu)-\sigma\lambda)}$$

Let there be no bones about it. At first blush this is still a truly ugly and opaque formula. We can attempt to beautify it by re-expressing it in terms more like those in the Black-Scholes formula, putting

$$d_2(\lambda) = \frac{1}{\sigma}(\mu - a), \text{and } d_2(0) = \frac{1}{\sigma}(a + \mu),$$
$$d_1(\lambda) = d_2(\lambda) + \sigma, \quad d_1(0) = d_2(0) + \sigma.$$

These are analogous to the values of $d_{1,}d_2$ in the Black-Scholes formula in the case $\lambda = 0$. Then

$$E[C|L \geq l] = O\frac{e^{\mu+\sigma^2/2}\Phi(d_1(0)) - e^{-\lambda a+(1-\lambda)\mu+(1-\lambda)^2\sigma^2/2}\Phi(d_1(\lambda))}{\Phi(d_2(0)) - e^{-\lambda a-\lambda\mu+\lambda^2\sigma^2/2}\Phi(d_2(\lambda))}. \quad (5.37)$$

What is interesting is how this conditional expectation, the expected close for the survivors, behaves as a function of the volatility parameter σ. Although this is a rather complicated looking formula, we can get a simpler picture (Figure 5.8) using a graph with the drift parameter μ chosen so that $E(C) = 56.25$ is held fixed. We assume $a = -\ln(30/40)$ (consistent with Table 5.2)and vary the value of σ over a reasonable range from $\sigma = 0.1$ (a very stable investment) through $\sigma = .8$ (a highly volatile investment). In Figure 5.8 notice that for small volatility, e.g. for $\sigma \leq 0.2$, the conditional expectation $E[C|L \geq 30]$ remains close to its unconditional value $E(C)$ but for $\sigma \geq 0.3$ it increases almost linearly in σ to around 100 for $\sigma = 0.8$. The intuitive reason for this dramatic increase is quite simple. For large values of σ the process fluctuates more, and only those

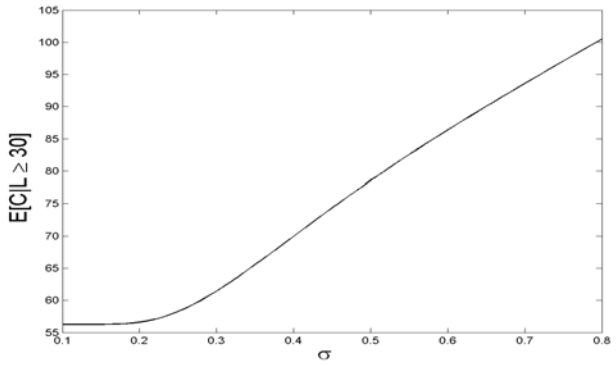


Figure 5.8: $E[C|L \geq 30]$ for various values of (μ, σ) chosen such that $E(C) = 56.25$.

paths with very large values of C have abeen able to avoid the absorbing barrier at $l = 30$. Two comparable portfolios with unconditional return about 40% will show radically different apparent returns in the presence of an absorbing barrier. If $\sigma = 20\%$ then the survivor's return will still average around 40%, but if $\sigma = 0.8$, the survivor's returns average close to 150%. The practical implications are compelling. *If there is any form of survivorship bias (as there usually is), no measure of performance should be applied to the returns from different investments, managers, or portfolios without an adjustment for the risk or volatility.*

In the light of this discussion we can return to the comparison of the two portfolios in Table 5.3. Evidently there is little bias in the estimate of returns for portfolio 2, since in this case the volatility is small $\sigma = 0.2$. However there is very substantial bias associated with the estimate for portfolio 1, $\sigma = 0.5$. In fact if we repeat the graph of Figure 5.8 assuming that the unconditional return is around 8% we discover that $E[C|L \geq 30]$ is very close to $56\frac{5}{8}$ when $\sigma = 0.5$ indicating that this is a more reasonable estimator of the performance of portfolio 1.

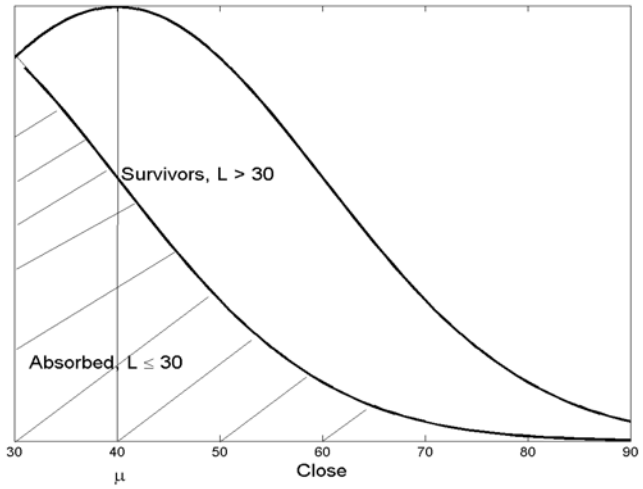


Figure 5.9: The Effect of Surivorship bias for a Brownian Motion

For a Brownian motion process it is easy to demonstrate graphically the nature of the surivorship bias. In Figure 5.9, the points under the graph of the probability density which are shaded correspond to those which whose low fell below the absorbing barrier at $l = 30$. The points in the unshaded region correspond to the survivors. The expected value of the return conditional on survival is the mean return (x-cooredinate of the center of mass) of those points chosen uniformly under the density but above the lower curve, in the region labelled "survivors". Note that if the mean μ of the unconditional density approaches the barrier (here at 30) , this region approaches a narrow band along the top of the curve and to the right of 30. Similarly if the unconditional standard deviation or volatility increases, the unshaded region stretches out to the right in a narrow band and the conditional mean increases.

We arrive at the following seemingly paradoxical conclusions which make it imperative to adjust for survivorship bias: *the conditional mean, conditional on survivorship, may increase as the volatility increases even if the unconditional*

mean decreases.

Let us return to the problem with both an upper and lower barrier and consider the distribution of returns conditional on the low never passing a barrier Oe^{-a} and the high never crossing a barrier at Oe^{b} (representing a fund buyout, recruitment of manager by competitor or promotion of fund manager to Vice President). It is common in process control to have an upper and lower barrier and to intervene if either is crossed, so we might wish to study those processes for which no intervention was required. Similarly, in a retrospective study we may only be able to determine the trajectory of a particle which has not left a given region and been lost to us. Again as an example, we use the following data on two portfolio managers, both observed conditional on survival, for a period of one year.

Portfolio	Open price	Close Price	Lower Barrier	Upper Barrier	Volatility
1	40	$56\frac{5}{8}$	30	100	.5
2	40	$56\frac{1}{4}$	30	100	.2

If ϕ denotes the standard normal p.d.f., then the conditional probability density function of $\ln(C/O)$ given that $Oe^{-a} < L < H < Oe^{b}$ is proportional to $\frac{1}{\sigma}\phi(\frac{u-\mu}{\sigma})w(u)$ where, as before

$$w(u) = 1 - e^{-2b(b-u)/\sigma^2} + e^{-2(a+b)(a+b-u)/\sigma^2} - e^{-2a(a+u)/\sigma^2} + e^{-2(a+b)(a+b+u)/\sigma^2} - E(W),$$

where $W = I[frac1(\frac{\ln(H)}{a+b}) > \frac{b}{a+b}] + I[frac1(\frac{-\ln(L)}{a+b}) > \frac{a}{a+b}]$, and

$$b = \ln(100/40), \quad a = -ln(30/40).$$

The expected return conditional on survival when the drift is μ is given by

$$E(\ln(C/O)|30 < L < H < 100) = \frac{1}{\sigma}\int_{-a}^{b} uw(u)\phi(\frac{u-\mu}{\sigma})du.$$

where $w(u)$ is the weight function above. Therefore a moment estimator of the drift for the two portfolios is determined by setting this expected return equal to the observed return, and solving for μ_i the equation

$$\frac{1}{\sigma_i}\int_{-a}^{b} uw(u)\phi(\frac{u-\mu_i}{\sigma_i})du = R_i, \qquad i = 1, 2.$$

The solution is, for portfolio 1, $\mu_1 = 0$ and for portfolio 2, $\mu_2 = 0.3$. Thus the observed values of C are completely consistent with a drift of 30% per annum for portfolio 2 and a zero drift for portfolio 1. The bias again very strongly effects the portfolio with the greater volatility and estimators of drift should account for this substantial bias. Ignoring the survivorship bias has led in the past to some highly misleading conclusions about persistence of skill among mutual funds.

Problems

1. If the values of d_j are equally spaced, i.e. if $d_j = j\Delta, j = ..., -2, -1, 0, 1, ...$and with $S_0 = 0, S_T = C$ and $M = \max(S_0, S_T)$, show that

$$E[H|C = u] = M + \Delta\frac{P[C > u \text{ and } \frac{C-M}{\Delta} \text{ is even}]}{P[C = u]}.$$

2. Let $W(t)$ be a standard Brownian motion on $[0, 1]$ with $W_0 = 0$. Define $C = W(1)$ and $H = \max\{W(t); 0 \le t \le 1\}$. Show that the joint probability density function of (C, H) is given by

$$f(c, h) = 2\phi(c)(2h - c)e^{-2h(h-c)}, \text{ for } h > \max(0, c)$$

   where $\phi(c)$ is the standard normal probability density function.

3. Use the results of Problem 2 to show that the joint probability density function of the random variables

$$Y = \exp\{-(2H - C)^2/2\}$$

   and C is a uniform density on the region $\{(x, y); y < \exp(x^2/2)\}$.

4. Let $X(t)$ be a Brownian motion on $[0,1]$, i.e. X_t satisfies

$$dX_t = \mu dt + \sigma dW_t, \text{ and } X_0 = 0.$$

Define $C = X(1)$ and $H = \max\{X(t); 0 \leq t \leq 1\}$. Find the joint probability density function of (C, H).

Chapter 6

Quasi- Monte Carlo Multiple Integration

Introduction

In some sense, this chapter fits within Chapter 4 on variance reduction; in some sense it is stratification run wild. Quasi-Monte Carlo methods are purely deterministic, numerical analytic methods in the sense that they do not even attempt to emulate the behaviour of *independent* uniform random variables, but rather cover the space in d dimensions with fewer gaps than independent random variables would normally admit. Although these methods are particularly when evaluating integrals in moderate dimensions, we return briefly to the problem of evaluating a one-dimension integral of the form

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

The simplest numerical approximation to this integral consists of choosing a point x_j in the interval $[\frac{j}{N}, \frac{j+1}{N}], j = 0, 1, ..., N-1$, perhaps the midpoint of the

interval, and then evaluating the average

$$\frac{1}{N}\sum_{j=0}^{N-1} f(x_j). \tag{6.1}$$

If the function f has one continuous derivative, such a numerical method with N equally or approximately equally spaced points will have bias that approaches 0 at the rate $1/N$ because, putting $M = \sup\{|f'(z)|; 0 < z < 1\}$,

$$\int_{j/N}^{(j+1)/N} f(x)dx - \frac{1}{N}f(x_j) \leq \frac{1}{N^2}M \tag{6.2}$$

and so summing both sides over j gives

$$|\int_0^1 f(x)dx - \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)| \leq \frac{1}{N}M.$$

We will refer to the error in the numerical integral in this case

$$\varepsilon_N = |\int_0^1 f(x)dx - \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)|$$

as $O(N^{-1})$ which means that the sequence of errors ε_N satisfies

$$\limsup_{N\to\infty} N^{-1}\varepsilon_N < \infty$$

or intuitively that the errors are bounded by a constant times N^{-1}.

If the function f is known to have bounded derivatives of second or third order, then integrals can be approximated to an even higher degree of precision. For example various numerical quadrature formulae permit approximating an integral of the form $\int_0^1 f(x)w(x)dx$ with a weighted average of N points

$$\sum_{j=1}^{N} w_j f(x_j) \tag{6.3}$$

in such a way that if $f(x)$ is a polynomial of degree $2N-1$ or less, the approximation is *exact*. Here the function $w(x)$ is typically some density such as the uniform, exponential or normal density and the optimal placement of the points x_j as well as the weights w_j depends on $w(x)$. Of course a smooth function

can be closely approximated with a polynomial of high degree and so numerical quadrature formulae of the form (6.3) permit approximating a one-dimension integral arbitrarily closely provided that the function is sufficiently smooth, i.e. it has bounded derivatives of sufficiently high order. We should note that in this case, the weights w_j and the points x_j are both deterministic. By contrast, the Monte Carlo integral

$$\widehat{\theta}_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(U_i)$$

with N points places these points at random or pseudo-random locations, has zero bias but the standard deviation of the estimator $\sqrt{var(\widehat{\theta}_{MC})}$ is a constant multiple of $1/\sqrt{N}$. The Central Limit theorem assures us that

$$N^{1/2}(\widehat{\theta}_{MC} - \int_0^1 f(x)dx)$$

converges to a normal distribution which means that the error is order (in probability) $N^{-1/2}$. Note that there is a change in our measure of the size of an error, since only the variance or standard deviation of a given term in the sequence of errors is bounded, not the whole sequence of errors ε_N. In particular if a pseudo-random estimator $\widehat{\theta}$ satisfies

$$E(\widehat{\theta} - \int_0^1 f(x)dx)^2 = O(N^{-2k})$$

then we say that the error is $O_P(N^{-k})$ where O_P denotes "order in probability". This is clearly a weaker notion than $O(N^{-k})$. Even the simplest numerical integral (6.1) has a faster rate of convergence then that of the Monte Carlo integral with or without use of the variance reduction techniques of Chapter 4. This is a large part of the reason numerical integration is usually preferred to Monte Carlo methods in one dimension, at least for smooth functions, but it also indicates that for regular integrands, there is room for improvement over Monte Carlo in higher dimensions as well.

The situation changes in 2 dimensions. Suppose we wish to distribute N points over a uniform lattice in some region such as the unit square. One

possible placement is to points of the form

$$(\frac{i}{\sqrt{N}}, \frac{j}{\sqrt{N}}), i, j = 1, 2, ... \sqrt{N}$$

assuming for convenience of notation that $\sqrt{N}$ is integer. The distance between adjacent points is of order $1/\sqrt{N}$ and by an argument akin to (6.2), the bias in a numerical integral is order $1/\sqrt{N}$. This is the now same order as the standard deviation of a Monte Carlo integral, indicating that the latter is already, in two dimensions, competitive. When the dimension $s \geq 3$, a similar calculation shows that the standard deviation of the Monte-Carlo method is strictly smaller order than the error of a numerical integral with weights at lattice points. Essentially, the placement of points on a lattice for evaluating a $d-$dimensional integral is far from optimal when $d \geq 2$. Indeed various deterministic alternatives called quasi-random samples provide substantially better estimators especially for smooth functions of several variables. Quasi-random samples are analogous to equally spaced points in one dimension and are discussed at length by Niederreiter (1978), where it is shown that for sufficiently smooth functions, one can achieve rates of convergence close to the rate $1/N$ for the one-dimensional case.

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 $d-$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.

Errors in numerical Integration

We consider the problem of numerical integration in d dimensions. For $d = 1$ classical integration methods, like the trapezoidal rule, are weighted averages of

the value of the function at equally spaced points;

$$\int_0^1 f(u)du \approx \sum_{n=0}^{m} w_n f(\frac{n}{m}), \tag{6.4}$$

where $w_0 = w_m = 1/(2m)$, and $w_n = 1/m$ for $1 \leq n \leq m-1$. The trapezoidal rule is exact for any function that is linear (or piecewise linear between grid-points) 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 $(\frac{j+1}{m}, f(\frac{j+1}{m}))$. Assume

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

If the function has a continuous second derivative, we have by Taylor's Theorem that the difference between the function and its linear interpolant is of order $O(x - \frac{j}{m})^2$, i.e.

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

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, $d \geq 2$. Suppose we evaluate the function at all of the $(m+1)^d$ 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 a Cartesian product of one-dimensional integration rules. For example, the d-fold Cartesian product of the trapezoidal rule is

$$\int_{[0,1]^d} f(\mathbf{u})d\mathbf{u} \approx \sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} w_{n_1} \cdots w_{n_s} f(\frac{n_1}{m}, \ldots, \frac{n_s}{m}), \tag{6.5}$$

where $[0,1]^d$ 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 is $O(m^{-2})$, provided that the second partial derivatives of f are continuous on $[0,1]^d$. We know that the error cannot be smaller because when the function depends on only one variable and is constant in the others, the one-dimensional result is a special case. In terms of the number N of nodes or function evaluations, since $m = O(N^{1/d})$, the error is $O(N^{-2/d})$, which, with increasing dimension d, changes dramatically. For example if we required $N = 100$ nodes to achieve a required precision in the case $d = 1$, to achieve the same precision for a $d = 5$ dimensional integral using this approach we would need to evaluate the function at a total of $100^d = 10^{10}$= *ten billion nodes.* As the dimension increases, the number of function evaluations or computation required for a fixed precision increases exponentially. This phenomena is often called the "curse of dimensionality", exorcised in part at least by quasi or regular Monte Carlo methods.

The ordinary Monte Carlo method based on simple random sampling is free of the curse of dimensionality. 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 d. 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 O_P which measures the average or probabilistic behaviour of the error.*

Some of the oft-cited deficiencies of the Monte Carlo method limiting its

usefulness are:

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. Regularity of the integrand is not exploited even when it is available. 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 or smoothness of the integrand. For example the estimator is no more precise if we know that the function f has several continuous derivatives. In cases when we do not know whether the integrand is smooth or differentiable, it may be preferable to use Monte Carlo since it performs reasonably well without this assumption.

3. Genuine Monte Carlo is not feasible anyway since generating truly independent random numbers is virtually impossible. In practice we use pseudo-random numbers to approximate independence.

Theory of 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. Quasi-random sequences of low discrepancy sequences would fail all of the tests applied to a pseudo-random number generate except those testing for uniformity of the marginal distribution because the sequence is, by construction, autocorrelated. Our objective is to approximate an integral using a average of the function at N points, and we may adjust the points so that the approximation is more accurate. Ideally we would prefer these sequences to be self-avoiding, so that as the sequence is generated, holes are filled. As usual

we will approximate the integral with an average;

$$\int_{[0,1]^d} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}). \tag{6.6}$$

Quasi Monte-Carlo is able to achieve a deterministic error bound $O((logN)^d/N)$ for suitably chosen sets of nodes and for integrands with a relatively low degree of regularity, much better than the rate $O(N^{-1/2})$ achieved by Monte Carlo methods. Even smaller error bounds can be achieved for sufficiently regular integrands. There are several algorithms or quasi-Monte-Carlo sequences which give rise to this level of accuracy.

Suppose, as with a crude Monte Carlo estimate, we approximate the integral with (6.6) with $\mathbf{x_1}, \ldots, \mathbf{x_N} \in [0,1]^d$. The sequence $\mathbf{x_1}, \ldots, \mathbf{x_N}$,... is deterministic (as indeed are the pseudo-random sequences we used for Crude Monte-Carlo), but they are now chosen so as to guarantee a small error. Points are chosen so as to achieve the maximal degree of *uniformity* or a *low degree of discrepancy* with a uniform distribution. A first requirement for a low discrepancy sequence is that we obtain convergence of the sequence of averages so that:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}) = \int_{[0,1]^d} f(\mathbf{u})\mathbf{du},$$

and this should hold for a reasonably large class of integrands. This suggests that the most desirable sequences of nodes $\mathbf{x_1}, \ldots, \mathbf{x_N}$ are "evenly distributed" over $[0,1]^d$. Various notions of discrepancy have been considered as quantitative measures for the deviation from the uniform distribution but we will introduce only one here, the so-called "star-discrepancy". The star discrepancy is perhaps the more 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. Suppose we construct

$$\widehat{F}_N(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} I(\mathbf{x_n} \leq \mathbf{x}),$$

the empirical cumulative distribution function of the points $\mathbf{x_1}, \ldots, \mathbf{x_N}$, and compare it with

$$F(\mathbf{x}) = F(x_1, ...x_d) = \min(1, x_1 x_2 ... x_d) \text{ if all } x_i \geq 0$$

the theoretical uniform distribution on $[0,1]^d$. While any measure of the difference could be used, the star discrepancy is simply the Kolmogorov-Smirnov distance between these two cumulative distribution functions

$$D_N^* = \sup_{\mathbf{x}} |\widehat{F}_N(\mathbf{x}) - F(\mathbf{x})| = \sup_B |\frac{\#\text{ of points in B}}{N} - \lambda(B)|,$$

where the supremum is taken over all rectangles B of the form $[0, x_1] \times [0, x_2] \times ... \times [0, x_d]$ and where $\lambda(B)$ denotes the Lebesgue measure of B in $\mathcal{R}^d$.

It makes intuitive sense that we should choose points $\{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ such that the discrepancy is small for each N. This intuition is supported by a large number of theoretical results, at least in the case of smooth integrands with smooth partial derivatives. The smoothness is measured using $V(f)$, a "total variation" in the sense of Hardy and Krause, intuitively the length of the monotone segments of f. For a one dimensional function with a continuous first derivative it is simply

$$V(f) = \int_0^1 |f'(x)| dx.$$

In higher dimensions, the Hardy Krause variation may be defined in terms of the integral of partial derivatives;

Definition 48 *Hardy and Krause Total Variation*

If f is sufficiently differentiable then the variation of f on $[0,1]^d$ in the sense of Hardy and Krause is

$$V(f) = \sum_{k=1}^{s} \sum_{1 \leq i_1 < \cdots < i_k \leq s} V^{(k)}(f; i_1, \ldots, i_k), \tag{6.7}$$

where

$$V^{(k)}(f; i_1, \ldots, i_k) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^s f}{\partial x_{i_1} \cdots \partial x_{i_k}} \right|_{x_j = 1, j \neq i_1, \ldots, i_k} dx_{i_1} \cdots dx_{i_k}. \tag{6.8}$$

The precision in our approximation to an integral as an average of function values is closely related to the discrepancy measure as the following result shows. Indeed the mean of the function values differs from the integral of the function by an error which is bounded by the product of the discrepancy of the sequence and the measure $V(f)$ of smoothness of the function.

Theorem 49 *(Koksma-Hlawka inequality)*

If f has bounded variation $V(f)$ on $[0,1]^d$ in the sense of Hardy and Krause, then, for any $\mathbf{x_1}, \ldots, \mathbf{x_N} \in [0,1]^d$, we have

$$|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x_n}) - \int_{I^s} f(\mathbf{u})\mathbf{du}| \leq V(f)D_N^*. \tag{6.9}$$

We do not normally use this inequality as it stands since the evaluation of the error bound on the right hand side requires determining $V(f)$, typically a very difficult task. However this bound allows a separation between the regularity properties of the integrand and the degree of uniformity of the sequence. We can guarantee a reasonable approximation for any function f with bounded total variation $V(f)$ by ensuring that the discrepancy of the sequence D_N^* is small. For this reason, the discrepancy is central to quasi-Monte Carlo integration. Sequences with small star discrepancy are called low-discrepancy sequences. In fact since a variety of sequences exist with discrepancy of order

$$\frac{(\log N)^d}{N}$$

as $N \to \infty$, the term "low-discrepancy" is often reserved for these.

Examples of low discrepancy sequences

Van der Corput Sequence.

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**, obtained by

reversing the digits in the representation of some sequence of integers in a given base. Consider one-dimensional case $d = 1$ 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 $[0, 1]$ so that the integer $\sum_{k=0}^{t} a_k b^k$ is mapped into the point $\sum_{k=0}^{t} a_k b^{-k-1}$. 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. $1011 = 1(\frac{1}{2}) + 0(\frac{1}{4}) + 1(\frac{1}{8}) + 1(\frac{1}{16}) = \frac{11}{16}$.

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. The intervals are recursively split in half in the sequence $1/2, 1/4, 3/4, 1/8, 5/8, 3/8, 7/8, ...$ and the points are fairly evenly spaced for any value for the number of nodes N, and perfectly spaced if N is of the form $2^k - 1$. The star discrepancy of this sequence is

$$D_N^* = O(\frac{\log N}{N})$$

which matches the best that is attained for infinite sequences.

The Halton Sequence

This is simply the multivariate extension of the Van der Corput sequence. In higher dimensions, say in d dimensions, we choose d distinct primes, $b_1, b_2, ...b_d$ (usually the smallest primes) and generate, from the same integer m , the d components of the vector using the method described for the Van der Corput

sequence. For example, we consider the case $d = 3$ and use bases $b_1 = 2$, $b_2 = 3, b_3 = 5$ because these are the smallest three prime numbers. 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 6.1 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 6.2 and 6.3.

These figures can be compared with the plot of 100 pairs of independent uniform random numbers in Figure 6.4, which seems to show 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=[];
```

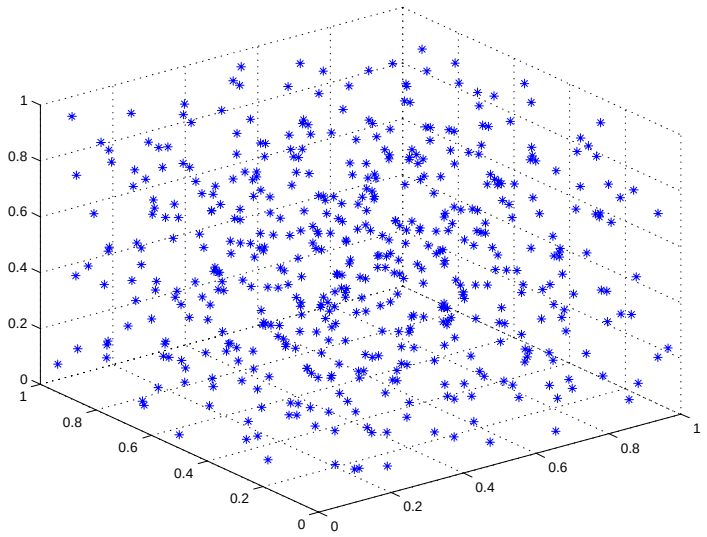


Figure 6.1: 500 points from a Halton seqnece of dimension 3

```
for i=1:s
 x=[x (corput(n,p(i)))'];
end
function x=corput(n,b)
% converts integers 1:n to from van der 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];
end
x=((1./b').^(1:(m+1)))*A;
```

The Halton sequence is a genuine low discrepancy sequence in the sense that

$$D_N^* = O(\frac{(\log N)^d}{N})$$

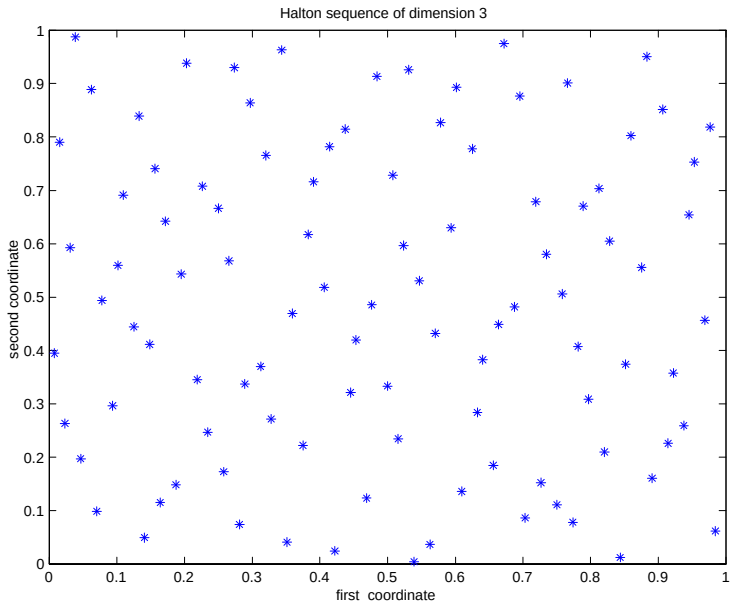


Figure 6.2: The first and second coordinate of 100 points from the Halton sequence of dimension 3

and the coverage of the unit cube is reasonably uniform for small dimensions. Unfortunately the notation $O()$ hides a constant multiple, one which, in this case, depends on the dimension d. Roughly (Niedereiter, 1992), this constant is asymptotic to d^d which grows extremely fast in d. This is one indicator that for large d, the uniformity of the points degrades rapidly, largely because the relative sparseness of the primes means that the $d'th$ prime is very large for d large. This results in larger holes or gaps in that component of the vector than we would like. This is evident for example in Figure6.5 where we plot the last two coordinates of the Halton sequence of dimension 15.

The performance of the Halton sequence is considerably enhanced by permuting the coefficients a_k prior to mapping into the unit interval as is done by the Faure sequence.

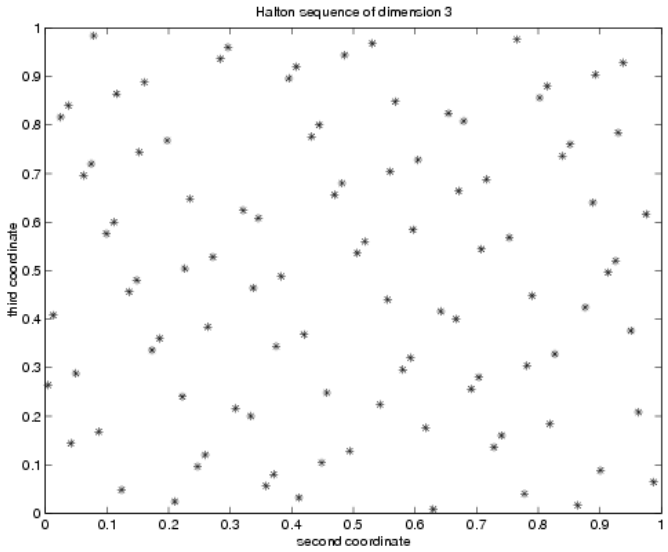


Figure 6.3: The second and third coordinate of 100 points from the Halton sequence of dimension 3

Faure Sequence

The **Faure** sequence is similar to the Halton sequence in that each dimension is a permutation of a van der Corput sequence; however, the same prime is used as the base b for each of the components of the vector, and is usually chosen to be the smallest prime greater than or equal to the dimension (Fox, 1996).

In the Van der Corput sequence we wrote the natural numbers in the form $\sum_{k=0}^{t} a_k b^k$ which was then mapped into the point $\sum_{k=0}^{t} a_k b^{-k-1}$in the unit interval. For the Faure sequence we use the same construction but we use different permutations of the coefficients a_k for each of the coordinates. In particular in order to generate the i'th coordinate we generate the point

$$\sum_{k=0}^{t} c_k b^{-k-1}$$

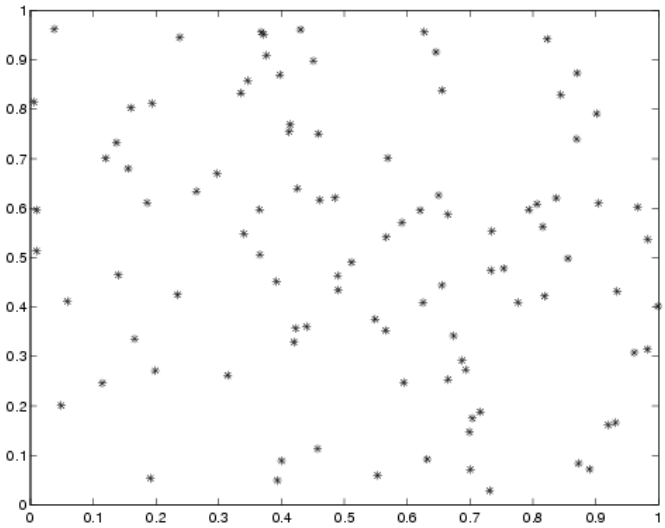


Figure 6.4: 100 independent $U[0,1]$ pairs

where

$$c_k = \sum_{m=k}^{t} \binom{m}{k} (i-1)^{m-k} a_m \bmod b$$

Notice that only the last $t-k+1$ values of a_i are used to generate c_k. For example consider the case $d=2, b=2$. Then the first 10 Faure numbers are

$$\begin{array}{cccccccccc} 0 & 1/2 & 1/4 & 3/4 & 1/8 & 5/8 & 3/8 & 7/8 & 1/16 & 9/16 \\ 0 & 1/2 & 3/4 & 1/4 & 5/8 & 1/8 & 3/8 & 7/8 & 15/16 & 7/16 \end{array}$$

The first row corresponds to the Van der Corput numbers and the second row of obtained from the first by permuting the values with the same denominator.

The Faure sequence has better regularity properties than does the Halton sequence above particularly in high dimensions. However the differences are by no means evident from a graph when the dimension is moderate. For example we plot in Figure 6.6 the 14'th and 15'th coordinates of 1000 points from the Faure sequence of dimension $d=15$ for comparison with Figure 6.5.

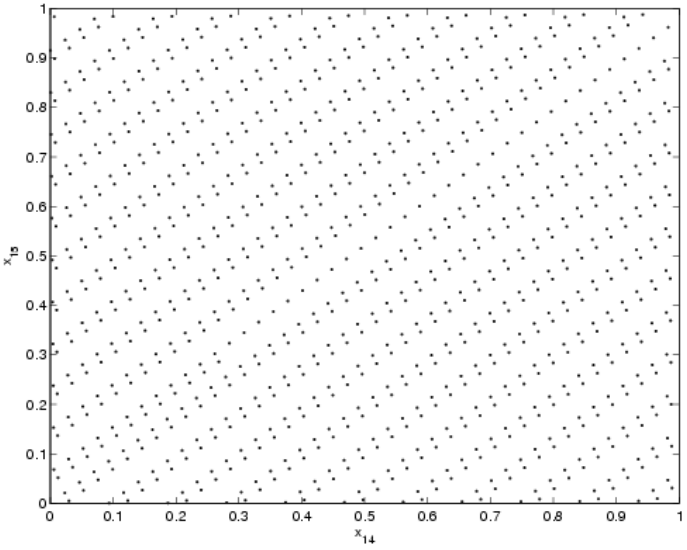


Figure 6.5: The 14'th and 15'th coordinates of the first 1000 of a Halton sequence $d = 15$

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 $d = 10$.

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

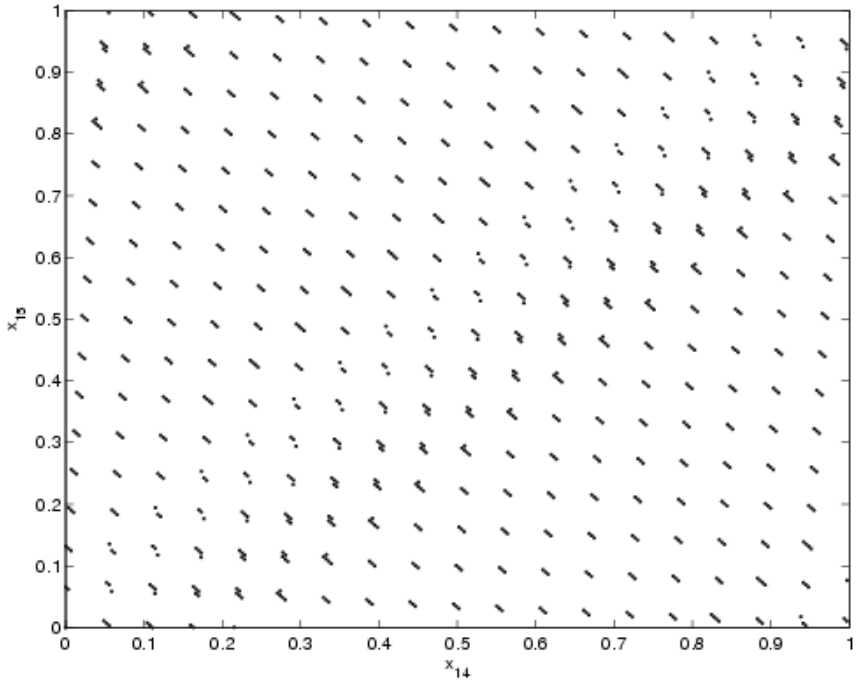


Figure 6.6: The last two coordinates of the first 1000 Faure points of dimension $d = 15$.

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

we obtain an estimate of 0.4614 very close to the correct value. I cannot compare these estimators using the notion of efficiency that we used there, however, because these low-discrepancy sequences are not random and do not even attempt to emulate random numbers. Though unable to compare performance with the variance of an estimator, we can look at the Mean squared error (see for example Figure 6.8). which shows a faster rate of convergence for Quasi Monte Carlo equivalent to variance reduction in excess of 100). Galanti & Jung (1997), report that the Faure sequence suffers from the problem of start-up

and especially in high-dimensions and the Faure numbers can exhibit clustering about zero. In order to reduce this problem, Faure suggests discarding the first $b^4 - 1$ points.

Sobol Sequence

The Sobol sequence is generated using a set of so-called *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.* A primitive polynomial is irreducible (i.e. cannot be factored into polynomials of smaller degree) and does not divide the polynomial $x^r + 1$ for $r < 2^p - 1$. For example the polynomial $x^2 + x + 1$ has no non-trival factors over the *Galois Field of order 2* and it does divide $x^3 + 1$ but not $x^r + 1$ for $r < 3$. 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 $a_k v_k$.

In the case $d = 2$, the first 10 Sobol numbers are, using irreducible polynomials $x + 1$ and $x^3 + x + 1$

0	1/2	1/4	3/4	3/8	7/8	1/8	5/8	5/16	13/16
0	1/2	1/4	3/4	1/8	5/8	3/8	7/8	11/16	3/16

Again we plot the last two coordinates for the first 1000 points from a Sobol sequence of dimension $d = 15$ in Figure 6.7 for comparison with Figures 6.5 and 6.6.

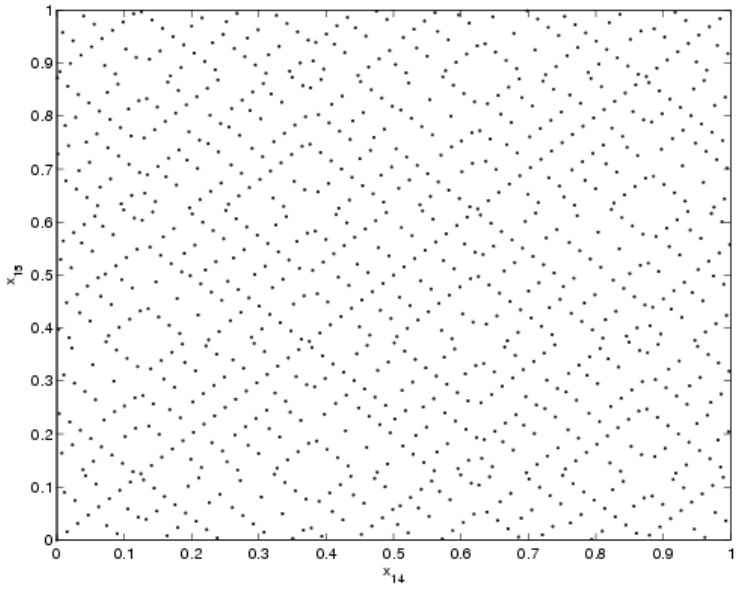


Figure 6.7: The last two coordinates of the first 1000 points from a Sobel sequence of dimension 15

Although there is a great deal of literature espousing the use of one quasi-Monte Carlo sequence over another, most results from a particular application and there is not strong evidence at least that when the dimension of the problem is moderate (for example $d \leq 15$) it makes a great deal of difference whether we use Halton, Faure or Sobol sequences. There is evidence that the starting values for the Sobol sequences have an effect on the speed of convergence, and that Sobol sequences can be generated more quickly than Faure Moreover neither the Faure nor Sobol sequence provides a "black-box" method because both are sensitive to intitialization. I will not attempt to adjudicate the considerable literature on this topic here, but provide only a fragment of evidence that, at least in the kind of example discussed in the variance reduction chapter, there is little to choose between the various methods. Of course this integral, the

discounted payoff from a call option as a function of the uniform input, is a one-dimensional integral so the Faure, Halton and Van der Corput sequences are all the same thing in this case. In Figure 6.8 we plot the (expected) squared error as a function of sample size for $n = 1, ..., 100000$ for crude Monte Carlo (the dashed line) and the Van der Corput sequence. The latter, although it oscillates somewhat, is substantially better at all sample sizes, and its mean squared error is equivalent to a variance reduction of around 1000 by the time we reach $n = 100,000$. The different slope indicates an error approaching zero at rate close to n^{-1} rather than the rate $n^{-1/2}$ for the Crude Monte Carlo estimator. The Sobol sequence, although highly more variable as a function of sample size, appears to show even more rapid convergence along certain subsequences.

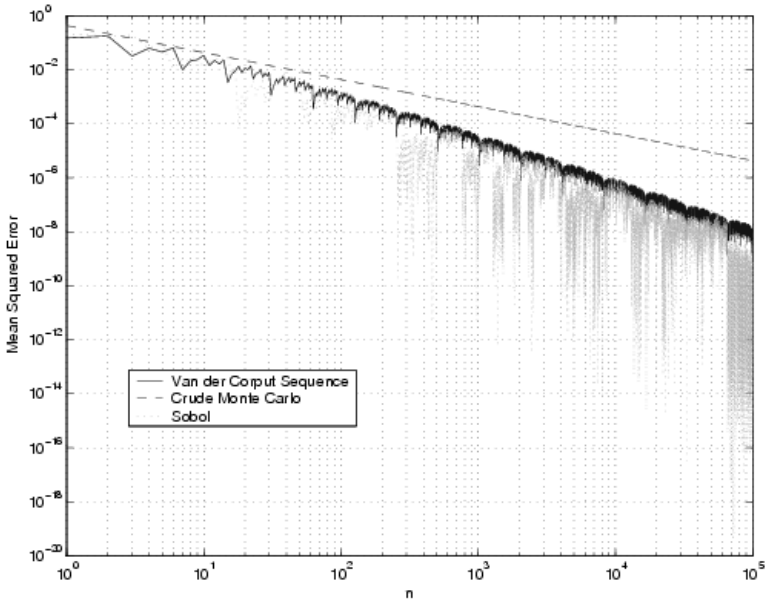


Figure 6.8: (Expected) squared error vs. sample size in the estimation of an Call option price for Crude MC and Van der Corput sequence.

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

Elementary Intervals and Nets

Definition: elementary interval

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

$$E = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \frac{(a_j + 1)}{b^{d_j}} \right), \tag{6.10}$$

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

In other words an elementary interval is a multidemsional generalization of a rectangle with sides of length b^d parallel to the axes. A net is a finite sequence which is perfectly balanced in the sense that certain elementary intervals all have exactly the same number of elements of the sequence.

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.

Definition: (t, s) - sequence

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

It is known that for a (t, s)-sequence in base b, we can obtain an upper bound for the star discrepancy of the form:

$$D_N^* \leq C \frac{(\log N)^s}{N} + O(\frac{(\log N)^{s-1}}{N}). \tag{6.11}$$

Special constructions of such sequences for $s \geq 2$ have the smallest discrepancy that is currently known (Niederreiter, 1992).

Tan(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 in a particular problem, important for financial computation snce 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).

In summary, Quasi-Monte Carlo frequently generates estimates superior to Monte-Carlo methods in many problems of low or intermediate effective dimension. If the dimension d is large, but a small number of variables determine most of the variability in the simulation, then we might expect Quasi Monte-Carlo methods to continue to perform well. Naturally we pay a price for the smaller error often associated with quasi Monte-Carlo methods and other numerical techniques or, in some cases any technique which other than a crude simulation of the process. Attempts to increase the efficiency for the estimation of a particular integral work by sacrificing information on the *distribution* of other functionals of the process of interest. If there are many objectives to a simulation, including establishing the distribution of a large number of different variables (some of which are necessarily not smooth), often only a crude Monte Carlo simulation will suffice. In addition, the theory supporting low-discrepancy sequences, both the measures of discrepancy themselves and the variation measure $V(f)$ are artificiall tied to the arbitrary direction of the axes. For example if $f(x)$ represents the indicator function of a square with sides parallel to the axes in dimension $d = 2$, then $V(f) = 0$. However, if we rotate this rectangle

by 45 degrees, the variation becomes infinite, indicating that functions with steep isoclines at a 45 degree angle to the axes may be particularly difficult to integrate using Quasi Monte Carlo.

Problems

1. Use 3-dimensional Halton sequences to integrate the function

$$\int_0^1 \int_0^1 \int_0^1 f(x,y,z)dxdydz$$

where $f(x,y,z) = 1$ if $x < y < z$ and otherwise $f(x,y,z) = 0$. Compare your answer with the true value of the integral and with crude Monte Carlo integral of the same function.

2. Use your program from Question 1 to generate 50 points uniformly distributed in the unit cube. Evaluate the Chi-squared statistic χ^2_{obs} for a test that these points are independent uniform on the cube where we divide the cube into 8 subcubes, each having sides of length 1/2. Carry out the test by finding $P[\chi^2 > \chi^2_{obs}]$ where χ^2 is a random chi-squared variate with the appropriate number of degrees of freedom. This quantity $P[\chi^2 > \chi^2_{obs}]$ is usually referrred to as the "significance probability" or "p-value" for the test. If we suspected *too much uniformity* to be consistent with assumption of independent uniform, we might use the other tail of the test, i.e. evaluate $P[\chi^2 < \chi^2_{obs}]$. Do so and comment on your results.