MONTE CARLO METHODS IN FINANCE

Don L. McLeish

August, 2000

Contents

0 **i**

1 Some Basic Theory of Finance. **1**
1.1 Introduction to Pricing: Single Period Models. 1
1.2 Multiperiod Models. 5
1.3 Determining the Process B_t. 10
1.4 Minimum Variance Portfolios and the Capital Asset Pricing Model. 12
1.4.1 The Capital Asset Pricing Model (CAPM) 13
1.5 Entropy: choosing a Q measure 18
1.5.1 Cross Entropy 19
1.5.2 Maximum Entropy 21
1.6 Models in Continuous Time 22
1.7 Review Problems 41

2 Basic Monte Carlo Methods **45**
2.1 Simulation and Monte Carlo Methods 45
2.2 Uniform Random Number Generation 46
2.3 Apparent Randomness of Pseudo-Random Number Generators . 51
2.4 Non-Uniform Random Number Generation 55
2.5 Generating Discrete Distributions 72
2.6 Simulating Stochastic Partial Differential Equations. 77
2.7 Problems 81

3 Variance Reduction Techniques. **91**
3.1 Introduction. 91
3.2 Variance reduction for one-dimensional Monte-Carlo Integration. 92
3.3 Simulations from the Stationary Distribution of a Markov Chain. 107
3.4 Coupling and Perfect Simulations. 109
3.5 Some Multivariate Applications in Finance. 112
3.6 Problems 127

4 Quasi- Monte Carlo Multiple Integration **131**
4.1 Introduction 131
4.2 Errors in numerical Integration 132

4.2.1 Low discrepancy sequences . . . 133
4.2.2 Definition: Measures of Discrepancy. . . . 134
4.2.3 Definition: Total Variation . . . 135
4.2.4 Theorem: Koksma - Hlawka inequality . . . 135
4.2.5 Examples of low discrepancy sequences . . . 135
4.2.6 Halton Sequences. . . . 136
4.2.7 Sobol Sequence . . . 138
4.2.8 Definition: elementary interval . . . 138
4.2.9 Definition: (t, m, s) - net . . . 139
4.2.10 Definition: (t, s) - sequence . . . 139

5 Sensitivity Analysis, Estimating Derivatives and the Greeks. 141
5.1 Estimating Derivatives. . . . 142
5.1.1 Problem. . . . 143
5.1.2 Definition . . . 144
5.1.3 Example. . . . 145
5.1.4 Example. Estimating Vega. . . . 146
5.1.5 Gaussian Quadrature. . . . 147
5.2 Infinitesimal Perturbation Analysis: Pathwise differentiation. . . . 148
5.2.1 Sensitivity of the value of a spread option to the correlation. 151

6 Estimation and Calibration. 153
6.1 Using Historical Data for Diffusion Models. . . . 153
6.2 Estimating Volatility . . . 154

6.3 Estimating Hedge ratios and Correlation Coefficients. . . . 158
6.4 Estimation using the Term Structure of Interest rates . . . 165

7 Miscellany 167
7.1 Neural Nets . . . 167
7.2 Chaos, Long term dependence and non-linear Dynamics . . . 168
7.3 ARCH AND GARCH . . . 168
7.3.1 ARCH(1) . . . 169
7.3.2 Estimating Parameters . . . 169
7.3.3 Akaike's Information Criterion . . . 170
7.3.4 Testing for ARCH effects . . . 170
7.3.5 Example. Deutschmark Exchange . . . 171

8 Appendix A: Some Basic Theory of Probability 175
8.1 Probability Models. . . . 175
8.2 Independence and Conditional Probabilities. . . . 178
8.3 Expected Values, Mean, Variances . . . 184
8.4 Discrete Bivariate and Multivariate Distributions . . . 188
8.5 Continuous Distributions . . . 192
206
8.6 Stochastic Processes . . . 208

8.7 Conditional Expectation and Martingales 213
8.7.1 Conditional Expectation. 213
8.7.2 Martingales. 215
8.7.3 Martingales and Finance 217

9 Appendix B: Stochastic Integration and Continuous Time Models 219
9.1 Ordinary Differential Equations 220
9.2 Systems of Ordinary Differential Equations. 223

9.3 Partial Differential Equations 223

10 Appendix: Numerical Solutions of DE's and PDE's 231
10.0.1 Difference and Differential Operators and solving ODE's. 231
10.0.2 Numerical Methods for P.D.E.'s. Explicit Finite Difference Method. 235
10.0.3 Implicit Finite Difference Solutions to the Diffusion Equation. 238
10.0.4 The Crank-Nicolson Method 240
10.0.5 Stability and Consistency of the Crank-Nicolson. 241
10.0.6 The Method of Lines. 242
10.0.7 Finite Elements and the Galerkin Method. 243
10.0.8 Solution of the Diffusion Equation. 248
10.0.9 Black-Scholes with Transaction Costs. 253
10.0.10 Methods for American Options 254

11 Appendix C: Glossary. 257

Chapter 0

Introduction

This book concerns the analysis of models for financial markets, particularly the assets traded there. 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 at expiry 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 a 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, engineers is now available to all.

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. A stochastic model does not militate against some ability to forecast. It 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 attempt to model.

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. When compared with a purely mathematical analysis, models with more complexity and fewer assumptions, models that are closer to the real-world, 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 analysis- 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 analysis the results. And simulation does not render all 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. Unfortunately we expect a competitor to also make a bid for the company, and the competitor values the shares differently. Moreover there are costs associated with losing a given bid to the competitor. Suppose 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		
	Bid	A	B	C	
Your	12	3	2	-2	
Co.	13	1	-4	4	
	14	0	-5	5	

The payoffs to the competitor are somewhat different and given below

		Competitor	Bid	
		A	B	C
Your	12	-1	-2	3
Bid	13	0	4	-6
	14	0	5	-5

Define the 3×3 matrix of payoffs listed in Table 1 by A and that in Table 1 by B. 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 $q_j, j = 1, 2, 3$, then we can write the expected payoff to you in the form $\sum_{i=1}^{3} \sum_{j=1}^{3} p_i A_{ij} q_j.$, which, when written as a vector-matrix product, takes the form $p^T Aq$. This can be thought of as the average return to your firm in the long run if this game were repeated many times. In this case you would clearly choose $p_i = 1$ for the row i corresponding to the maximum component of Aq, if the vector q were known to you. Similarly your competitor 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. This would seem to be a reasonable approach to building a simulation model for this game. Play the game repeatedly with each of the two players updating the estimated probabilities with which their opponent uses their available strategies and record the number of times each strategy is used. In this case we run the 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
[m,s]=max(A*q); p(s)=p(s)+1;
[m,s]=max(p*B); q(s)=q(s)+1;
end
p=p-ones(1,n(1)); p=p/sum(p);
q=q-ones(n(2),1); q=q/sum(q);
```

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

```
[p,q]=nonzerosum(A,B,50000)
p =  0.6667    0.2222     0.1111
q = 0  0.5000   0.5000
```

and this seems to indicate that the strategies will be "mixed" or random. You should choose a bid of \$12.00 with probability around 2/3, \$13.00 with probability about 2/9 and \$14.00 with probability 1/9. 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? If you were to choose a single strategy as your "best" then your competitor could presumably determine what your "best" strategy is and act to reduce your return 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. Participants may compete or cooperate for a greater total return.

While there is no assurance that the solution is optimal, it is in this case easily seen to be sensible solution, achieved with little effort. Indeed in a game such as this, there is no clear definition of what an optimal strategy would be. Do you plan your play based on the worst case or the best case scenario or something in between such as some form of average? Do you attempt to collaborate with your competitor for greater total return and then subsequently divide this in some fashion? Here the simulation has emulated a simple form of competitor behaviour and arrived at a reasonable solution.

There remains the question of how we actually select a bid with probabilities 2/3, 2/9 and 1/9 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	$U < 2/3$	$2/3 \leq U < 8/9$	$8/9 \leq U < 1$
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 probability.

Chapter 1

Some Basic Theory of Finance.

1.1 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* since with a net investment of \$0, we are able to achieve non-negative return (i.e. the possibility of future profits with no down-side risk). In mathematical terms, the investor seeks a point (x, y) such that $x + ys = 0$ (net cost of the portfolio is zero) and

$$\begin{aligned} x(1+r) + ysu &\geq 0, \\ x(1+r) + ysd &\geq 0 \end{aligned}$$

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

$$\begin{aligned} y &= -\frac{1+r}{su}x \\ y &= -\frac{1+r}{sd}x \end{aligned}$$

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

$$-\frac{1+r}{sd} \leq -\frac{1}{s} \leq -\frac{1+r}{su} \tag{1.1}$$

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

$$\frac{d}{1+r} < 1 < \frac{u}{1+r} \tag{1.2}$$

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

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

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 necessarily 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 to value a stock using the actual distribution P different from Q. Then discounted to the present, the friend believes that the stock is worth

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

Such a friend offers their assets as a sacrifice to any investor. If the friend's assessed price is greater than the current market price, buy on the market and sell to the friend. Otherwise, do the reverse. In any case, you are richer (except of course by the loss of one friend)!

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 future time 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 $S_1 > 10$. This is the difference between the value of the stock on expiry and the exercise price of the option or your profit on purchasing the stock for \$10 and selling it on the open market at $\$S_1$. However, if $S_1 < 10$, there is no point whatever in exercising your option as you are not compelled to do so and your return is \$0. In general, your payoff from purchasing the option is a simple function of the future price of the stock, such as $V(S_1) = \max(S_1 - 10,\ 0)$. The future value of the option is itself a random variable but it derives its value from that of the stock, hence it is called a *derivative.*

Now consider an arbitrary function of the stock, $X = V(S_1)$ representing the payoff to an investor from a certain financial instrument (derivative) when the stock price at the end of the period is S_1. Any function of the stock price is called a *contingent claim.* In our 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

$$\begin{aligned} x(1+r) + ysu &= V(su) \\ x(1+r) + ysd &= V(sd) \end{aligned}$$

Solving: $y^* = \frac{V(su)-V(sd)}{su-sd}$ and $x^* = \frac{V(su)-y^*su}{1+r}$. Upon solving these two equations we are able to replicate the contingent claim $V(S_1)$ exactly- i.e. buy x^* units of bond and y^* of stock, we 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 more generally to complete markets, or markets in which any contingent claim is attainable by an investment in other marketable instruments. The following theorem provides a more general proof of this result, the proof due to Chris Rogers.

Suppose we hae a total of N risky assets whose prices at times $t = 0, 1$, are given by $(S_0^j(\omega), S_1^j(\omega)), j = 1, 2, ..., N$ for possible states ω. For simplicity assume that these states have positive probability $P(\omega) > 0$. Assume also a riskless asset (a bond) paying interest rate r over one unit of time. Suppose we borrow (i.e. short bonds) at the risk-free rate to buy w_j units of stock j at time 0 for a total cost of $\sum w_j S_0^j(\omega)$. The value of this portfolio at time $t = 1$ is $\sum w_j(S_1^j(\omega) - (1+r)S_0^j(\omega))$. Then we say there are *no arbitrage opportunities if* for all weights w_j this is either identically 0 for all ω or it takes both positive and negative values.

Theorem 1 *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 $j = 1, ..., N$.*

Problem 2 Proof. *Define $M(w) = E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))]$ and consider the problem*

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

If there is no arbitrage opportunity then there is a unique minimum satisfying $\frac{\partial M}{\partial w_j} = 0$ or

$$E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))S_1^j] = (1+r)S_0^j E[\exp(\sum w_j(S_1^j - (1+r)S_0^j))].$$

or

$$S_o^j = \frac{E[\exp(\sum w_j S_1^j)S_1^j]}{(1+r)E[\exp(\sum w_j S_1^j)]}.$$

Define a measure Q equivalent to the original probability measure such that

$$Q(\omega) = P(\omega)\frac{\exp(\sum w_j S_1^j(\omega))]}{E[\exp(\sum w_j S_1^j)]}$$

Then note that for each j,

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

so the current price of each stock is the discounted expected value of the future price under the risk-neutral measure Q. Conversely if

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

holds for some measure Q then $E_Q[\sum w_j(S_1^j - (1+r)S_0^j)] = 0$ for all w_j and this implies that the random variable $\sum w_j(S_1^j - (1+r)S_0^j)$ is either identically 0 or admits both posiive and negative values. Therefore the existence of the measure Q implies that there are not arbitrage opportunities. ■

So the theory of pricing derivatives in a complete market is based on a rather trivial observation. If we can reproduce exactly the same (random) returns as the derivative provides using a linear combination of other marketable securities (which have prices signed 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 represented by a bid-ask spread. In other words there is 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 assumes a complete market and it is questionable whether such a market can 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? So like all models, this one has its deficiencies and its critics. Like all reasonable models, its merit is that it provides an approximation to a real-world phenomenon, one that permits further study and improvement.

1.2 Multiperiod Models.

When an asset price evolves over time, we normally allow the investor to make decisions about an investment at various periods during the life of that investment. Such decisions are made with the benefit of 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* 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, though regrettably this is an increasingly important factor in real-life. The second property of H_t that we use is that it includes the value of every asset price at time t, or, 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, there was defined 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

$$E^Q(\frac{1}{1+r_t}S_{t+1}|H_t) = S_t. \tag{1.3}$$

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 $B_t = (1 + r_0)(1 + r_1)...(1 + r_{t-1})$. If I were to promise you exactly \$1.00 payable at time t (and if you believed me), then to cover this promise I would require an investment at time $t = 0$ of $\$1/B_t$, which we might call (at least in

the case when the interest rates r_t are known) the *present value* of the promise. In general, at time t, the present value of a certain amount $\$V$ promised at time T (i.e. the present value or the value discounted to the present of this payment) is $V\frac{B_t}{B_T}$. Now suppose we divide (**??**) above by B_t. We obtain

$$E^Q(\frac{S_{t+1}}{B_{t+1}}|H_t) = E^Q(\frac{1}{B_t(1+r_t)}S_{t+1}|H_t) = \frac{1}{B_t}E^Q(\frac{1}{1+r_t}S_{t+1}|H_t) = \frac{S_t}{B_t}. \tag{1.4}$$

Notice that we are able to take the divisor B_t outside the expectation since B_t is known at time t and therefore a constant with respect to the history H_t. This equation (**??**) 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.

$$E(Y_T|H_t) = Y_t \text{ for all } T > t. \tag{1.5}$$

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. Now the condition (**??** implies of the process $Y_t = S_t/B_t$ that $E(Y_{t+1}|H_t) = Y_t$ for all $T > t$ and this implies the martingale condition (**??**) since $E(Y_T|H_t) = E[...E[E(Y_T|H_{T-1})|H_{T-2}]...|H_t] = Y_t$.

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. And the fact that prices of marketable commodities must be martingales under the risk neutral measure has many consequences to 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 2% 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.

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, $E^Q(\frac{B_t}{B_T}V(S_T)|H_t] = V_t =$ the market price of the asset 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. In order to

determine a suitable price for a derivative simply generate many simulations of the future value $V(S_T)$ of the derivative given the current store of information H_t, simulations conducted under the measure Q, and then average the values, discounted to the present, over all simulations. The catch is that the Q measure is 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 which has this property. Now in a simple model involving a single stock, this is a fairly simple job, 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.

Solving for the Q Measure.

Let us consider the following simple example. Over each period, a stock price provides greater or less or the same return as a risk free investment. Assume for simplicity that increases are by the factor $u(1+r)$ and decreases by factor $(1+r)/u$ where $u > 1$ and otherwise the stock price increases by the risk free rate factor $(1+r)$. 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.

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 zero and the price 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. Consider the following special case, with the risk-free interest rate per period=$r = 1\%$, $u = 1.089$, $s_0 = \$1.00$. 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, are the values of $S(T)$ in the set

$\{\ 1.1859, 1.0890, 1.0000, 0.9183, 0.8432\}$. Recall that a call option expiring at time $T = 2$ has price the value $E^Q(S_2 - K)^+$ 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. We have market prices $E^Q(S_2 - K)^+/(1 + r)^2$ of four call options with the same expiry and different exercise prices in the following table;

The price of the call options are in the following table

Exercise Price	Maturity	Call Option Price
0.867	2	0.154
0.969	2	.0675
1.071	2	.0155
1.173	2	.0016

Since in general the price of a call option with exercise price K and maturity date $T = 2$ is given by $E^Q(S_2-K)^+/(1+r)^2$, The equations to be solved equate the observed price of the options to their theoretical price $E(S_2 - K)^+/(1 + r)^2$ and are as follows;

$$
\begin{aligned}
0.0016 &= \frac{1}{(1.01)^2}(1.186 - 1.173)p_1p_2 \\
0.0155 &= \frac{1}{(1.01)^2}[(1.186 - 1.071)p_1p_2 + (1.089 - 1.071)\{p_1(1 - 2p_2) + (1 - 2p_1)p_2\}] \\
0.0675 &= \frac{1}{(1.01)^2}[0.217p_1p_2 + 0.12\{p_1(1 - 2p_2) + (1 - 2p_1)p_2\} \\
&\quad +0.031\{(1 - 2p_1)(1 - 2p_2) + p_1p_2 + p_1p_4)\} \\
0.154 &= \frac{1}{(1.01)^2}[0.319p_1p_2 + 0.222\{p_1(1 - 2p_2) + (1 - 2p_1)p_2\} \\
&\quad +0.133\{(1 - 2p_1)(1 - 2p_2) + p_1p_2 + p_1p_4)\} \\
&\quad +0.051\{\{p_1(1 - 2p_4) + (1 - 2p_1)p_3\}]
\end{aligned}
$$

While it is not too difficult to solve this system in this particular case (in this case the solution is given by $p_1 = 0.2, p_2 = 0.22, p_3 = 0.2, p_4 = 0.3$) 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 only have market prices for two derivatives defined on this stock, and therefore only two parameters which 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 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.

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

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. 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 substantially 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. But in a complete market, Q is uniquely determined by marketable securities. No real market can be complete. And in real markets, one asset is not perfectly replicated by a combination of other assets. This is true whether one asset is a derivative defined as a function of another marketed security (and interest rates and volatilities). 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

$$E^Q(\frac{B_t}{B_T}V(S_T)|H_t] \approx V_t \text{ for all marketable } V$$

Now if these were equalities, this would represent a number of equations in the unknown Q probabilities, typically fewer equations than unknowns 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, few are really 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.

1.3 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 "numeraire" or the price B_t of a risk-free investment at time t. Unfortunately no such 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 the process B_t given information on the prices of these bonds.

We begin with what we know. We assume we know the current prices, components of the vector, S_t of marketable securities. We also know the price of certain risk-free bonds with face value F , the value of the bond on maturity at time T. These pricesP_t provide some information on the bank account process B_t. In particular since a dividend-paying bond is a linear combination of payments at certain times $t < T$ plus a final payment of F, each current bond

price provides a value of the form $P_t = \sum_{T>s>t} d_s B_t/B_s + FB_t/B_T$. This can be written as a system of linear equations

$$P_t/B_t = \sum_{T>s>t} d_s/B_s + F/B_T$$

and provided that we have a sufficient number of bond prices P_t, possibly with different maturities, this system permits solving for certain values of $1/B_s$, $s > t$. Now the catch here is that there are typically too few risk-free bond maturities to get a detailed picture of the process $1/B_s, s > 0$. We could use government bonds for this purpose. But are these genuinely risk-free? Might not the additional use of bonds in large highly rated companies provide a more detailed picture of the bank account process B_s.

Can 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. We will also assume in this preliminary model that once default occurs, provided it occurs before the maturity date T of the bond, the payoff is a constant proportion p of the principal amount F owing. Then if τ denotes the time of default, a bond with face value $\$F$ which promises dividend payments d_s at time $s < T$ has price at time t given by

$$\begin{aligned} P_t &= \sum_{t<s<T} \frac{B_t}{B_s} d_s P(\tau > s|\tau > t) + \frac{pFB_t}{B_T} P(\tau \le T|\tau > t) + \frac{FB_t}{B_T} P(\tau > T|\tau > t) \\ &= \sum_{t<s<T} \frac{B_t}{B_s} d_s P(\tau > s|\tau > t) + \frac{pFB_t}{B_T} + \frac{(1-p)FB_t}{B_T} P(\tau > T|\tau > t). \end{aligned}$$

All probabilities are conditional on the event $[\tau > t]$ because unless this is true the debt has already defaulted and therefore its value is known. Unknowns in this equation are $P(\tau > s|\tau > t)/B_s$, $t < s < T$, $\frac{1}{B_T}$ and $pP(\tau > T|\tau > t)$. Now if we died and went to investors' heaven, a bond of every maturity T would be sold and we could solve this system of equations simply using the given bond prices. We might also hope that the probabilities of default are very small and follow a 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 or exponential distribution so that the probability of a default occurring in any period of fixed length is constant. Then $P(\tau > s|\tau > t) = \exp\{-k(s-t)\}$ for some $k > 0$. Suppose we define a new bank account process

$$\widetilde{B_s} = \frac{B_s}{P[\tau > s]} = B_s \exp\{ks\} \text{ for } s > t.$$

Clearly this bank account grows faster than the original, and it grows faster as the probability of default increases. The effective interest rate on this account is k units per period higher. Then rewriting the above equation for P_t,

$$P_t - \frac{pFB_t}{B_T} = \sum_{t<s<T} \frac{\widetilde{B_t}}{\widetilde{B_s}} d_s + (1-p)F\frac{\widetilde{B_t}}{\widetilde{B_T}}$$

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 augmented by k and with face value $F(1-p)$. *So to value a defaultable bond, augment the interest rate, 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.* Given only three bond prices with the same default characteristics, for example, and assuming constant interest rates so that $B_s = \exp(rs\}$, we may solve for the values of the three unknown parameters (r, k, p).

1.4 Minimum Variance Portfolios and the Capital Asset Pricing Model.

Let us begin by building a model for portfolios of securities that capture more or less of the major market movements. We have solved above for the values of $1/B_s$ only for certain values of s, but let us assume for the present that by interpolation, B_s is known for all $s > t$.

To begin with, define a common measure on investments that is equivalent to price, but from many perspectives, more convenient and stable statistically. For a security that has price $S(t)$ and $S(t+1)$ at times t and $t+1$, we define the return $R_i(t+1)$ on the security over this increment by

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

Returns can be measured in units that are easily understood (for example 5% or 10% per unit time) and independent of the amount invested. It is also easy to obtain the price at time t from the initial price at time 0 and the sequence of returns.

$$S_i(t) = S_i(0)(1 + R_i(1))(1 + R_i(2))...(1 + R_i(t)).$$

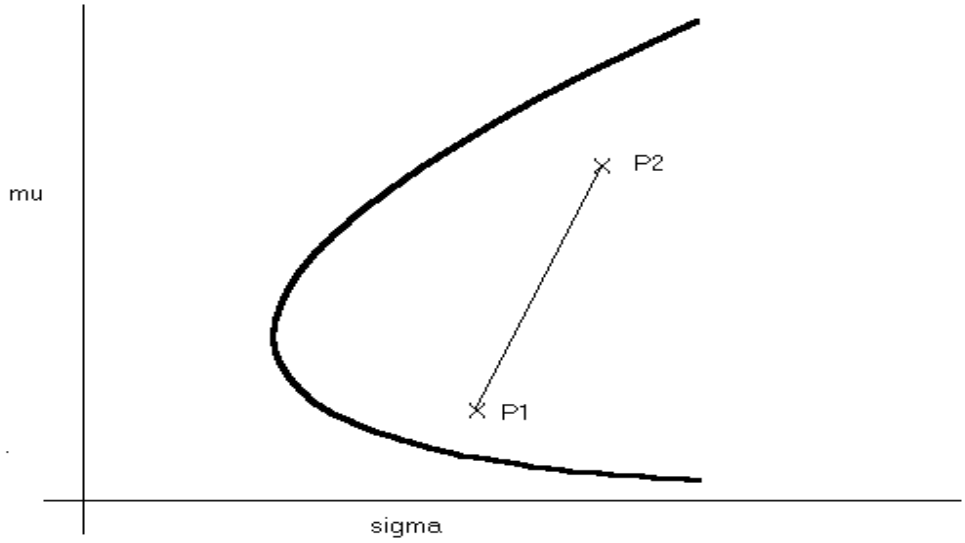
When we buy a portfolio the return on the portfolio is simply a weighted average of the individual stock returns and the weights are the relative amounts invested in each stock. For example if our portfolio is such that $w_i(t)$ is the proportion of our total investment of $\$I$ invested in stock i at time t then the number of shares of stock i purchased is $\frac{w_i(t)}{S_i(t)}I$ and the portfolio return over the next period is

$$\frac{\sum \frac{w_i(t)}{S_i(t)} I[S_i(t+1) - S_i(t)]}{\sum w_i(t) I} = \sum w_i(t) R_i(t+1).$$

When time is measured continuously, the instantaneous return process might be defined by a limit of the above form $R_i(t) = \lim_{h\to 0} \frac{S_i(t+h)-S_i(t)}{hS_i(t)}$ or formally $dS_i/S_i = d(\ln(S_i(t))$ provided these are well-defined. More generally, the returns process is a process whose product integral results in the original stock price. process.

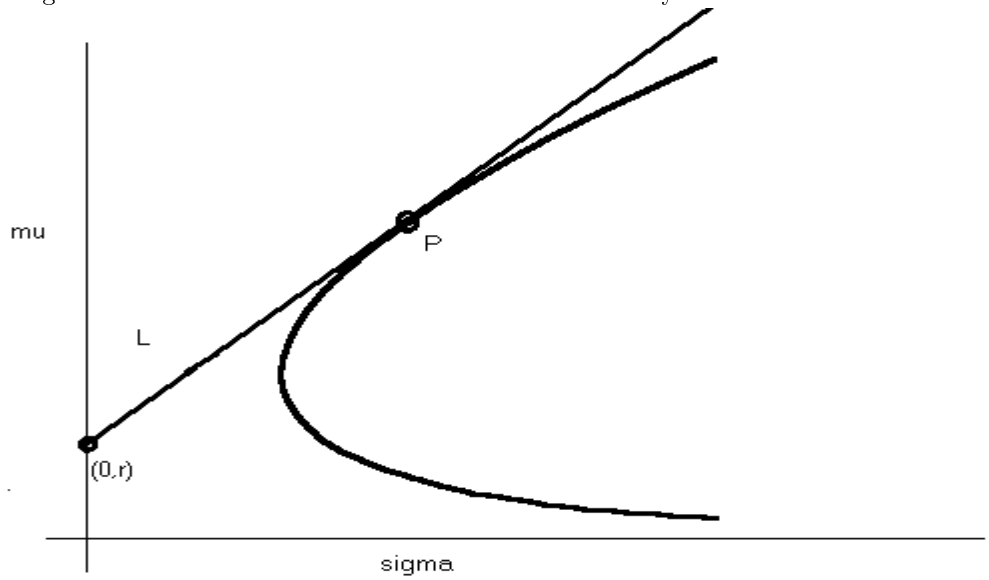
1.4.1 The Capital Asset Pricing Model (CAPM)

In the Capital Asset Pricing model it is assumed at the outset that investors concentrate on two measures of return, its expected value μ and its standard deviation σ. We consider the expected values and variances under the real-world probability measure P not under the risk-neutral Q measure. Suppose, for example we were to plot the set of all possible pairs (σ, μ) for portfolios of risky stocks. Let us assume for the present that the vector of all stock returns has mean return column vector given by η and covariance matrix of returns given by Σ. If a portfolio is such that we invest proportion w_i of our wealth in stock i, then defining $w = (w_1, ..., w_n)^T$, the total return on the portfolio has mean $\eta^T w$ and covariance matrix $w^T \Sigma w$. The set of all possible pairs of standard deviation and mean return $(\sqrt{w^T \Sigma w}, \eta^T w)$ has a semi-elliptical boundary as in the following figure.



Now assuming that investors prefer higher expected return for the same standard deviation, only the upper envelope or roof of this region is efficient in the sense that no other portfolio has higher expected return for the same standard deviation. This is called the efficient frontier. Now the picture changes substantially if there is also a risk-free investment that all investors are able to include, because in this case there is also a point on the $\mu-$axis corresponding to

$\sigma = 0$. If this point is added, then the efficient frontier is now the region below the line L in the following figure. The point $(0, r)$ corresponds to the risk free investment whose return is r, and the point P is the point at which this line is tangent to the efficient frontier determined from the risky investments.



If all investors have access to the same risk free rate, then the line L is the unique efficient frontier for all investors, and P is the only efficient point in the risky portfolio. It must therefore represent the total market in the sense that the proportions βw_i represents the proportion of the total market invested in stock i. Suppose the market portfolio P has standard deviation σ_P and mean μ_P. Then this line is described by the relation

$$\mu = r + \frac{\mu_P - r}{\sigma_P}\sigma.$$

For any stock i with mean and standard deviation of return (μ_i, σ_i) to be competitive, it must lie on this efficient frontier, i.e. it must satisfy the relation

$$\mu_i - r = \beta_i(\mu_P - r), \quad \text{where} \quad \beta_i = \frac{\sigma_i}{\sigma_P}.$$

This is the most important result in the capital asset pricing model. The constant β_i called the *beta* of a stock is both the change in the expected stock return for each unit change in the market expected return and also the ratio of the standard deviations of return of the stock and the market. Such a model is a reasonable basis for simplifying the covariance structure of stock returns to some manageable number of parameters. Analogous to the relation above describing the expected returns is a regression model relating the returns from the stock R_i and from the market portfolio R_P.

$$R_i - r = \beta_i(R_P - r) + \epsilon_i$$

where ϵ_i is a zero-mean random error uncorrelated with the market return. Taking variance on both sides, we obtain

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

which appears to indicate that the variance of the return from stock i is greater than the value σ_i^2 assumed earlier. What is the cause of this contradiction? We assumed earlier that the stock i lay on the efficient frontier, but this is not a necessary condition for investors to choose it. All that is required is that it form a part of a portfolio which lies on the efficient frontier. We cannot expect a higher rate of return to compensate for additional risk that can be diversified away. In an efficient market, not all risk need be rewarded with additional return. Suppose, for example, we had many stocks with similar β, then we could presumably invest equally in all of them and end up with the average of many returns $\bar{R}$. Notationally,

$$\bar{R} - r = \beta(R_P - r) + \bar{\epsilon}$$

where, provided that we have sufficiently many such investments to average over, $\bar{\epsilon}$ has mean zero and variance close to 0 and is itself almost 0. By averaging or diversifying, we are able to provide an investment with the same average return characteristics but smaller variance than the original stock, implying that the original stocks could not have been on the efficient frontier. We say that the specific risk (i.e. $var(\epsilon_i)$) associated with stock i can be diversified away, and should not therefore be rewarded with increased return. Only the systematic risk σ_i is rewarded with increased expected return.

The capital asset pricing model provides a simplified form of the covariance matrix Σ of the vector of stock returns. Notice that under the model

$$R_i - r = \beta_i(R_P - r) + \epsilon_i, \qquad var(\epsilon_i) = \delta_i$$

we have

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

Whereas N stocks would otherwise require a total of $N(N+1)/2$ parameters in the covariance matrix Σ of returns, the Capital Asset Pricing Model allows us to reduce this to the $N+1$ parameters σ_P^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 seems to contradict some observations at least over substantial periods of time.

Minimum Variance under Q.

Suppose we wish to find a portfolios of securities which has the smallest possible variance under Q. . For example for a given set of weights, define the portfolio

$\Pi(t) = \sum w_i(t)S_i(t)$. For any weights $w_i(t)$ this produces a portfolio with exactly the same conditional rate of return under the Q measure as any one of the constituent stocks, for

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

In terms of returns, this equation says that all securities have the same expected return under Q as does the portfolio Π. If we are interested in hedging our investment, we wish to minimize the instantaneous variance of the return of the portfolio since the conditional mean is unaffected by the choice of weights. The natural constraint is that the cost of the portfolio is determined by the amount $c(t)$, say, that we presently have to invest. Several choices of the function $c(t)$ present themselves as possible. For example you might wish to compare, at time t, the benefits of liquidating a risk free investment of $B(t)$ for investment in this portfolio, in which case we could take $c(t) = B(t)$. Since $c(t)$ is purely a matter of rescaling the results, one might also scale so that $c(t) = 1$. 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. In this case $c(t) = \Pi(t)$. We wish to minimise

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

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

$$W_1(t) = \begin{pmatrix} w_1(t) \\ w_2(t) \\ \cdot \\ \cdot \\ \cdot \\ w_n(t) \end{pmatrix} = \frac{c(t)}{S'(t)\Sigma_t^{-1}S(t)}\Sigma_t^{-1}S(t).$$

where $\Sigma_t = var_Q(S(t+1)|H_t)$ is the instantaneous conditional covariance 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

$$var_Q(\Pi(t+1)|H_t) = W_1'(t)\Sigma_t W_1(t) = \frac{c^2(t)}{S'(t)\Sigma_t^{-1}S(t)}.$$

Once again we express this relation in terms of the return $R_\Pi(t+1) = \frac{\Pi(t+1)-\Pi(t)}{\Pi(t)}$ from the portfolio Π. Assume for the present that the portfolio is self-financing so that $c(t) = \Pi(t)$. Then the above relation states that the conditional variance of the return $R_\Pi(t+1)$ given the past is simply

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

The covariances between returns for individual stocks and the portfolio Π as it turns out are given by exactly the same quantity, namely

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

By this formula, note that all such covariances are positive and therefore in the definition of $\Pi(t)$, all weights are positive and so no stocks are shorted. Now suppose that we wish to model the individual stock returns using a model for the minimum variance portfolio Π. Under the risk neutral measure Q, recall that every stock (and portfolio) has the risk-free rate of return $r(t+1) = \frac{B(t+1)-B(t)}{B(t)}$. Then we have by standard regression formulae,

$$R_i(t+1) - r(t+1) = \beta(R_\Pi(t+1) - r(t+1)) + \varepsilon_i(t+1)$$

where $\varepsilon_i(t+1)$ is a (random) error whose expected values under Q are all zero. The regression coefficient $\beta = \beta(t) = \frac{1}{S'(t)\Sigma_t^{-1}S(t)}$ is the same for all stocks. Of course the usual "beta" in the CAPM model in finance is the regression coefficient with a market rate of return under the actual measure P and these do vary from stock to stock, as do the expected returns.

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.*

The question arises whether there are other minimum variance portfolios uncorrelated with this one. Suppose we define $\Pi_2(t)$ similarly to minimize the variance subject to the condition that the weights $W_2(t)$ satisfy $W_2'(t)\Sigma_t W_1(t) = 0$. This implies that the corresponding portfolio $\Pi_2(t)$ satisfies $Cov_Q(\Pi_2(t+1), \Pi(t+1)|H_t) = 0$. In view of the definition of $W_1(t)$, this implies that $W_2'(t)S(t) = 0$ or that the cost of such a portfolio at the beginning of periods is 0. This means that the portfolio is such that there is a perfect balance between long and short stocks, or that the value of the long and short stocks are equal. Subject to this restriction it is possible to make the conditional variance as large or as small as we wish, simply by scaling up or down by a common factor the amount of our investments.

The above analysis assumes that our objective is minimizing the variance of the portfolio. However, under the risk neutral measure Q, every stock has the same risk-free rate of return and so it is equally logical to minimize the variance of the *portfolio return.* By the same analysis as above, this is achieved when the proportion of our total investment at each time period in stock i is chosen as components 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 since at each step the stock price is a linear

function of the return $S_i(t+1) = S_i(t)(1+R_i(t))$ the variance minimizing portfolios are essentially the same.

Before we continue, let us examine how practical the above decomposition is for a large market. Note that it requires knowledge of Σ_t and worse still, it requires inverting this matrix. We will try to avoid estimating all $\frac{n(n+1)}{2}$ parameters in Σ_t by using the form implied by the Capital Asset Pricing Model $\Sigma_t = \beta\beta'\sigma_P^2(t) + \Delta(t)$, where $\sigma_P^2(t)$ is the market volatility at time t, $\Delta(t)$ is the diagonal matrix with the $\delta_i(t)$ along the diagonal and β is the vector of individual stock betas. In this case $\Sigma_t^{-1} = \Delta^{-1} + c\Delta^{-1}\beta\beta'\Delta^{-1}$ where

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

and consequently the optimal investment in each stock i is proportional to

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

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

The conditional variance of $R_i(t+1)$ given the market at time $t+1$ is δ_i. Let us call this the excess volatility for stock i. *Then the weights for the optimal portfolio are linear in the beta for the stock and the reciprocal of the excess volatility.*

1.5 Entropy: choosing a Q measure

Typically market information does not completely determine the risk-neutral measure Q . We will argue that while the historical data should not strictly determine the Q measure, it should 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, $E^Q[S_i(t+1)|H_t] = S_i(t)r(t+1)$ where $r(t+1)$ is the risk free interest rate over the interval $(t, t+1)$. (Recall that this risk-free interest rate $r(t+1)$ is defined by the equation $B(t+1) = (1+r(t+1))B(t)$.)

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. 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 must begin with the notion of entropy which underlies considerable theory in Statistics and elsewhere in Science.

1.5.1 Cross Entropy

Consider two probability measures P and Q. 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. It is not hard to show that this measure is always non-negative and if Q is absolutely continuous with respect to P this can be rewritten in the form

$$H(Q|P) = E^Q \log(\frac{dQ}{dP}).$$

If, however, Q is not absolutely continuous with respect to P then the cross entropy is infinite. We should also remark that the cross entropy is not really a metric in the usual sense (although we unashamedly use the term distance in reference to it) since in general $H(Q|P) \neq H(P|Q)$. Now the following result asserts that the probability measure Q which is closest to P but satisfies a constraint on its mean is generated by an exponential tilt of the distribution of P.

Theorem 3 *: Minimizing cross-entropy.*

Consider the problem

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

subject to the constraint $E^Q(f(X)) = \mu$. *Then the solution, if it exists, is given by*

$$dQ = \frac{\exp(\eta' f(X))}{m(\eta)} dP$$

where $m(\eta) = E^P[\exp(\eta' f(X))]$ *and* η *is chosen so that* $\frac{m'(\eta)}{m(\eta)} = \mu$.

The proof of this result, in the case of a discrete distribution P is a straightforward use of Lagrange multipliers. 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 $E^Q[S(t+1)|H_t] = (1 + r(t+1))S(t)$ we require that $\eta = \eta_t$ be chosen so that

$$\int s \frac{\exp(\eta' s)}{m_t(\eta)} dP_{t+1}(s) = \frac{m_t'(\eta)}{m_t(\eta)} = \frac{d}{d\eta} \log(m_t(\eta)) = (1 + r(t+1))S(t).$$

Of course, the parameter $\eta = \eta_t$ is dependent on time since it depends on the conditional distribution given H_t. Now consider a measure Q given by

$$dQ = \frac{\exp(\sum_t \eta_t' dx_t)}{\Pi_t m_t(\eta_t)} dP$$

Theorem 1.4.5 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 continuous time analogue of this result, which, for completeness, we simply state below and reserve the proof for the problem set. The statement requires understanding of the continuous time models of the next section, and so can be skipped on first reading.

Theorem 4 .

Suppose under a probability measure P, the stock price process S_t satisfies an Ito equation of the form

$$dS_t = \mu_t dt + \sigma_t dW_t$$

Then the measure Q which satisfies

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

subject to the constraint that the drift of the process is given by the risk free interest rate, (formally written $E^Q[dS_t|H_t] = r_t S_t dt$) is such that

$$dQ = \exp(\int \eta_t dS_t - \int (\eta_t \mu_t + \eta_t^2 \sigma_t^2/2)dt)dP$$

where

$$\eta_t = \frac{S_t r_t - \mu_t}{\sigma_t^2}$$

The effect of this change of measure is such that under Q, S_t satisfies an Ito equation with the same diffusion term and drift determined by the risk-free interest rate

$$dS_t = r_t S_t dt + \sigma_t dW_t$$

1.5.2 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. In general for a discrete distribution, such that $P[X = x] = p(x)$, the entropy may be defined to be

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

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 scale 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 in fact. The complexity of a distribution as measured by the expected number of questions in a binary poll to determine the value of a random variable with that distribution can be measured by the

Many statistical distributions have an interpretation in terms of maximizing entropy. For example, what discrete distribution p has values on a certain set and yet 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, of course, that the probabilities add to one). 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 $p(x)$ we obtain $-\ln(p(x)) - 1 - \lambda = 0$ or $p(x) = e^{-(1+\lambda)}$. Applying the constraint that the sum of the probabilities is one results in $p(x) = 1/n$ for all x. This shows that the discrete distribution with 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. Let us call such a distribution the discrete normal distribution. In fact if we drop the requirement that the distribution is discrete, the same kind of argument shows that the maximum entropy distribution is the normal distribution.

So here, at least, are two simple distributions arising 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.*

1.6 Models in Continuous Time

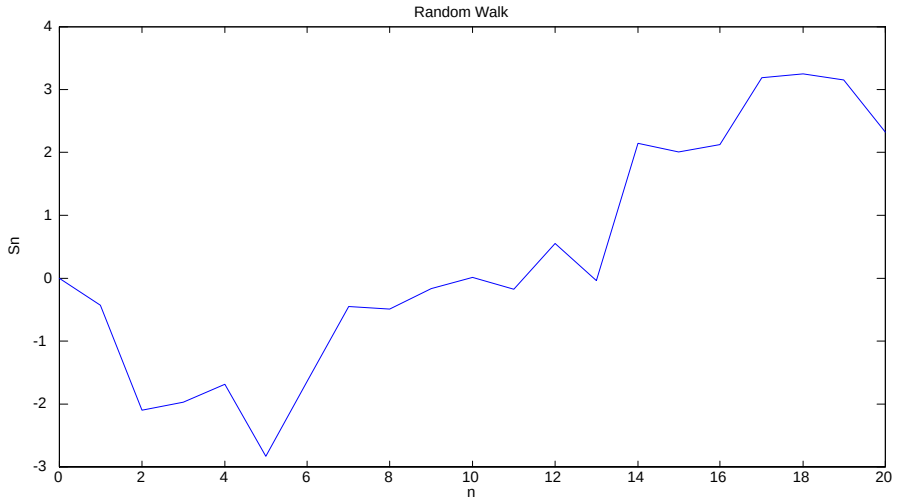
We begin with some oversimplified rules of stochastic calculus which can be omitted by those with a background in Brownian motion and diffusion. First, we define a stochastic process W_t called the *standard Brownian motion* or *Wiener process* having the following properties;

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. $W(0) = 0$.

The Standard Brownian Motion Process

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 Einsten, 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 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.



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

$$\begin{aligned} dV_t &= \frac{\partial}{\partial W}g(W_t,t)dW + \frac{\partial^2}{\partial W^2}g(W_t,t)\frac{dW^2}{2} + \frac{\partial}{\partial t}g(W_t,t)dt \qquad (1.6) \\ &\quad +(\text{stuff}) \times (dW)^3 + (\text{more stuff}) \times (dt)(dW)^2 + ... \end{aligned}$$

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 **??** 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

$$dX_t = aX_t dt + \sigma X_t dW_t. \tag{1.7}$$

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,

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t. \tag{1.8}$$

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 **??** 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 $a(X_t, t)dt = a(X_t, t)h$ and conditional variance given by $\sigma^2(X_t, t)var(dW_t) = \sigma^2(X_t, t)var(W_{t+h} - W_t) = \sigma^2(X_t, t)h$. 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,* usesd 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 **??**. 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 **??** 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 **??** 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

$$\frac{\partial}{\partial s}p = -a(z,s)\frac{\partial}{\partial z}p - \frac{1}{2}\sigma^2(z,s)\frac{\partial^2}{\partial z^2}p \tag{1.9}$$

and the *forward equation*

$$\frac{\partial}{\partial t}p = -\frac{\partial}{\partial x}(a(x,t)p) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x,t)p) \tag{1.10}$$

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 **??** is

$$X_t = X_0 + \int_0^t a(X_s,s)ds + \int_0^t \sigma(X_s,s)dW_s \tag{1.11}$$

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:

$$m(t) = X_0 + \int_0^t [\alpha(s) + \beta(s)m(s)]ds \tag{1.12}$$

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

$$m'(t) = \alpha(t) + \beta(t)m(t). \tag{1.13}$$

$$m(0) = X_0 \tag{1.14}$$

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

$$dS_t = a(S_t,\ t\ )\ dt\ +\ \sigma(S_t,\ t)\ dW_t \tag{1.15}$$

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 resonable 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 Black 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*

$$dV = [a(S_t, t)\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]dt + \sigma(S_t, t)\frac{\partial V}{\partial S}dW_t. \qquad (1.16)$$

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.)

$$V(S_{t+h}, t+h) \approx V + \frac{\partial V}{\partial S}(S_{t+h} - S_t) + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}(S_{t+h} - S_t)^2 + \frac{\partial V}{\partial t}h \qquad (1.17)$$

where we have ignored remainder terms that are $o(h)$. Note that substituting from **??** into **??**, 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 **??**. ■

By Ito's lemma, provided V is smooth, it also satisfies a diffusion equation of the form **??**. 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 **??**,

$$dV_t = u_t dS_t + w_t dB_t = [u_t a(S_t, t) + w_t r_t B_t]dt + u_t \sigma(S_t, t) dW_t \tag{1.18}$$

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 **??** and **??**, 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 **??** and **??** is;

$$-r_t V + r_t S_t \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} + \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{1.19}$$

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

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

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

$$-rV + \frac{\partial V}{\partial t} + (r - D_0)S\frac{\partial V}{\partial S} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 V}{\partial S^2} = 0. \tag{1.20}$$

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

$$\frac{\partial}{\partial t}u = k\frac{\partial^2}{\partial x^2}u \tag{1.21}$$

A solution of **??** 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 **??** of the form $u(x,t) = g(x/\sqrt{t})$. By differentiating and substituting above, we obtain an ordinary differential equation of the form

$$g''(\omega) + \frac{1}{2k}\omega g'(\omega) = 0, \omega = x/\sqrt{t} \tag{1.22}$$

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

$$\frac{\partial}{\partial w}g(\omega) = C_2 \exp\{-w^2/4k\} = C_2 \exp\{-x^2/4kt\} \tag{1.23}$$

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

$$u(x,t) = C_1 + C_2\frac{1}{2\sqrt{\pi kt}}\int_{-\infty}^{x} \exp\{-z^2/4kt\}dz \tag{1.24}$$

is a solution of this, the heat equation for $t > 0$. The role of the two constants is simple. Clearly if a solution to **??** 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 **??** 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

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

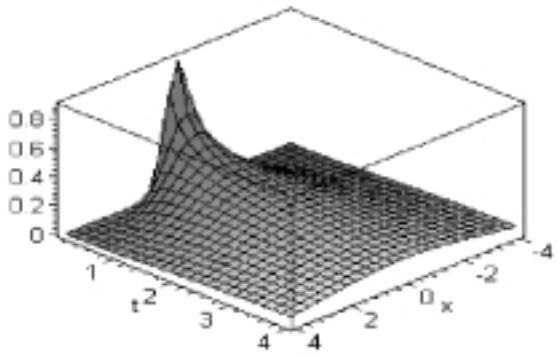


Figure 1.1: The function $u(x,t)$

Let us consider a basic solution to **??**:

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-x^2/4kt\} \tag{1.25}$$

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:

```
> u(x,t) := (.5/sqrt(Pi*t))*exp(-x^2/(4*t));
> plot3d(u(x,t),x=-4..4,t=.02..4,axes=boxed);
```

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 **??** 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 **??**. 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 **??** subject to the required boundary condition.

Solution of the Diffusion Equation.

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

$$\frac{\partial V}{\partial t} = r_t V - r_t S_t \frac{\partial V}{\partial S} - \frac{\sigma^2(S_t,t)}{2} \frac{\partial^2 V}{\partial S^2} \tag{1.26}$$

where S_t is an asset price driven by a diffusion equation

$$dS_t = a(S_t,t)dt + \sigma(S_t,t)dW_t, \tag{1.27}$$

$V(S_t,t)$ 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

$$V(S_T,T) = V_0(S_T). \tag{1.28}$$

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 **??** that if a related process X_t satisfies the stochastic differential equation

$$dX_t = r(X_t,t)X_t dt + \sigma(X_t,t)dW_t \tag{1.29}$$

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

$$\frac{\partial p}{\partial t} = -r(s,t)s \frac{\partial p}{\partial s} - \frac{\sigma^2(s,t)}{2} \frac{\partial^2 p}{\partial s^2} \tag{1.30}$$

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

$$f(t,s,T,z) = p(t,s,T,z)q(t,T) \tag{1.31}$$

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 **??** with weight function $w(z)$. i.e. try a solution of the form

$$V(s,t) = \int p(t,s,T,z)q(t,T)w(z)dz \tag{1.32}$$

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:

$$V(t,s) = E[w(X_T)q(t,T)|X_t = s] \tag{1.33}$$

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 **??**, we choose $w(x) = V_0(x)$. Now

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

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

$$V(t,s) = E[V_0(X_T)q(t,T)|X_t = s] \tag{1.34}$$

satisfies both the partial differential equation **??** and the terminal conditions **??** 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 **??**. This result is a special case when the spot interest rates are functions only of time of the following more general theorem.

Theorem 5 *(Feynman-Kac)*

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

$$V(S,t) = E[V_0(X_T)exp\{-\int_t^T r(X_v, v)dv\} \mid X_t = S] \tag{1.35}$$

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 **??** 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

$$X_T = \frac{X_t}{q(t,T)} + \int_t^T \frac{\sigma(u)}{q(u,T)} dW_u \tag{1.36}$$

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 **??** 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

$$g(x|\eta,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(log\ x - log\ \eta\ - \sigma^2/2)^2/2\sigma^2\}. \tag{1.37}$$

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

$$X_T = X_t exp\{\int_t^T (r(u) - \sigma^2(u)/2)du + \int_t^T \sigma(u)dW_u\}. \tag{1.38}$$

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 **??**. 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

$$e^{\mu+\sigma^2/2}\Phi(\frac{\mu}{\sigma} + \sigma) - \Phi(\frac{\mu}{\sigma}) \tag{1.39}$$

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

$$V(S,t) = S\Phi(d_1) - Ee^{-r(T-t)}\Phi(d_2) \tag{1.40}$$

where $d_1 < d_2$ are the values $\pm\sigma^2/2$ 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

$$e^{-r(T-t)} \int_0^\infty V_0(x) g(x|S(t)e^{r(T-t)},\ \sigma\sqrt{T-t})dx. \tag{1.41}$$

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

$$dS = a(S)dt + \sigma S dW_t \tag{1.42}$$

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.

1.7 Review 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.

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 an elliptical 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. Give an example of a function $g(t, W_t)$ such that the random variable $\int_0^1 g(t, W_t)dW_t$ does not have a normal distribution but has larger tails than does the normal distribution.

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

$$\begin{aligned} dX_t &= aX_t dt + bX_t dW_t \\ dY_t &= \mu Y_t dt + \sigma Y_t dW_t. \end{aligned}$$

Derive a stochastic differential equation for the ratio $Z_t = X_t/Y_t$. Suppose for example that X_t models the price of Telecom stock in \$NZ and Y_t is the exchange rate ($\$NZ/\$US$) at time t. Then what is the process Z_t?

14. Verify that for any pair of constants $a \neq 0$ and $b > 0$

$$dX_t = (X_t^{-1} + ab)X_t dt + bX_t dW_t$$

does not have a solution in the form $X_t = f(t, Y_t)$, where $f(t, y)$ is, say, a real function and Y_t is a Gaussian process.

15. Consider solving the problem

$$\min_q \sum q_i \log(\frac{q_i}{p_i})$$

subject to the constraints $\sum_i q_i = 1$ and $\sum q_i f(i) = \mu$. 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 $\frac{m'(\eta)}{m(\eta)} = \mu$.

16. 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

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

17. (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 2

Basic Monte Carlo Methods

2.1 Simulation and 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 (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 estimator $\overline{V(S_T)}$ will approximate the true expectation $E^Q V(S_T)$. Now while it would be foolish to use simulation 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 in the above discrete distribution 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.

2.2 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 in the amount of background cosmic radiation as the basis for such a generator, but these suffer from a number of disadvantages. While they may well be "random" in some more general sense than are the pseudo-random number generators that we use now, their properties are difficult to establish, and the sequences are impossible to reproduce, the latter being important for debugging a simulation program and for reducing the variance therein. Quite remarkably, it was discovered that very simple recursion formulae defined sequences that for practical purposes looked like sequences of independent random numbers and seemed (although the theorems rarely allow a proof of this fact) 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. This would seem to indicate that the conclusions of probability hold under much more general circumstances than the relatively restrictive conditions on these theorems indicate. Indeed, one would intuitively expect an enormous difference between the behaviour of independent random variables X_n and a sequence satisfying a recursion of the form $x_n = g(x_{n-1})$ for a simple function g and so it is surprising that for a large class of such functions g it is quite difficult to determine the difference between such a sequence and an independent sequence. Of course, any sequence of numbers generated from a simple recursion such as this is neither random, nor are x_{n-1} and x_n independent. Often we will emphasize this failure by referring to the sequence of *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.

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 $7 mod(3) = 1$ 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 congruential generators* 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 $x_n = (5x_{n-1} + 3) \ mod \ 8$, and the seed $x_0 = 3$. Note that by this recursion

$$\begin{aligned} x_1 &= (5 \times 3 + 3) \bmod 8 = 18 \bmod 8 = 2 \\ x_2 &= 13 \bmod 8 = 5 \\ x_3 &= 28 \bmod 8 = 4 \\ \text{and } x_4, x_5, x_6, x_7, x_8 &= 7, 6, 1, 0, 3 \text{ respectively} \end{aligned}$$

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 in general it is clear that the maximal 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 here, the period is far from attaining its theoretical maximum, $m - 1$. When is this maximal period achieved? 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 6 *: 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 $a^{m-1} = 1(mod\ m)$ and $a^i \neq 1(mod\ m)$ for all $i < m-1$.

Consider the multiplicative congruential generator $x_n = 2x_{n-1} mod 11$. 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 $2^i \bmod 11 = 1$, this is a maximal period generator having period 10. When $m = 11$, only the values $a = 2, 6, 7, 8$ produce full period (10) generators.

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

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

$$1226874159,\ 62089911,\ 1343714438,\ 39373$$

Let us suppose now that m is prime and a_2 is the multiplicative inverse $(\mathrm{mod}(m))$ 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 $\mathrm{mod}(m)$ forms a finite field and this allows essentially the same operations as we enjoy in the real number system. Suppose for two integers $x_1, x_2 \in \{0, 1, 2, ..., m-1\}$, $x_2 = a_1 x_1 \bmod(m)$. Then multiplying by a_2 we have

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

and this shows that $\mathrm{x}_1 = a_2 x_2 \bmod(m)$. In other words, if a_2 is the multiplicative inverse of $a_1(\mathrm{mod}\ m)$, then the multiplicative generator with multiplier a_2 generates exactly the same sequence as that with multiplier a_1 except in reverse order.

Theorem 7 *Period of Multiplicative Generators with $m = 2^k$*

If $m = 2^k \geq 8$, the multiplicative congruential generator has maximal period $m/4$ if $a \, mod \, 8 = 3$ or 5 and if x_0 is odd. T

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 $a, c, m, x0 = seed$. The efficient implementation of a generator for large values of m of approximately the same size as the machine precision depends very much on the architecture of the computer.

```
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
```

Theorem 8 *Period of Mixed Congruential Generators.*

The Mixed Congruential Generator,

$$x_n = (ax_{n-1} + c) mod(m) \tag{2.1}$$

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* $a-1$.

Consider $m = 2^{31} - 1$ which is a prime. When m is prime, by (i) together with the assumption that $a < m$, m must divide $a-1$ which implies $a = 1$. So for prime m the only full-period generators correspond to $a = 1$.

Also by Theorem 3.2.5, if $m = 2^k, k \geq 2$, the conditions become that c is odd, and 4 divides $a-1$. Then the generator $x_n = (ax_{n-1} + c) mod(m)$ has full period 2^k.

Among the common linear or multiplicative generators 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}	65539	0	RANDU
2^{32}	69069	1	Super-Duper (Marsaglia)
2^{35}	$5^{13} = 1220703125$	0	APPLE
2^{32}	134775813	1	Turbo-Pascal,Version 7.(period=2^{32})
2^{59}	13^{13}	0	NAG
$2^{31}-1$	630360016	0	Fishman (Simscript II)
$2^{31}-1$	742938285	0	Fishman and Moore
2^{32}	3934873077		Fishman and Moore
$10^{12}-11$	427419669081	0	MAPLE
2^{32}	3141592653	1	DERIVE
2^{32}	663608941	0	Ahrens (C-RAND)

Matrix Congruential Generators.

We consider a generator of $k-$dimensional vectors 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 one simple generator is to begin with a seed X_0 a vector, a constant vector C and define recursively

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

Such generators are most common when C = the zero vector and called *matrix multiplicative congruential generators* .

In many cases, the uniform random number generator in packages such as *Splus* and *Matlab* are not completely described in the package documentation For example, 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. In general, shuffling is conducted according to the following steps

1. Generate a sequence X_i using $X_{i+1} = a_1 X_i (mod\ m_1)$.
2. For fixed k put $(T_1, \ldots, T_k) = (X_1, \ldots, X_k)$.
3. Generate, using a different generator, a sequence $Y_{i+1} = a_2 Y_i (mod m_2)$.
4. Output the random number T_I where $I = ceiling(Y_i k/m_2)$.
5. Increment i, replace T_I by the next value of X, and return to 3.

One generator is used to produce the sequence as numbers are needed to fill k holes. The other generator is then used select which hole to draw the next number from.

Example: A shuffled generator

Consider a generator described by the above steps with $k = 4, m_1 = 19, m_2 = 29$

```
X(i) = 2   14   7   13   11   18   6   3   9

Y(i) = 11  25  17    4   16   28  14   0   3
```

We start by filling four pigeon-holes with the numbers produced by the first generator so that $(T_1, \ldots, T_4) = (2, 14, 7, 13)$. 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 $I = \lceil 4 \times 11/29 \rceil = 2$. Then the first number in our random sequence is drawn from box 2, i.e. $z_1 = T_2 = 14$, so $z_1 = 14$. This element of the vector is now replaced by 11, the next number in the X sequence. Proceeding in this way, the next index is $I = \lceil 4 \times 25/29 \rceil = 4$ and so the next number drawn is $z_2 = T_4 = 13$. Of course, when we have finished generating the values $z_1, z_2, \ldots$ all of which lie between 0 and $m_1 = 19$, 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]. Because of the small value we chose for m_1, this approximation will not be very good in this case. But the advantage of shuffling is that the period of the generator is greatly extended.

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 9 *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 10 *(period of generator summed mod m_1)*

If $X_{i+1} = a_1 X_i (mod\ m_1)$ has period $m_1 - 1$ and $Y_{i+1} = a_2 Y_i (mod\ m_2)$ has period $m_2 - 1$, then $(X_i + Y_i)(mod\ m_1)$ has period the least common multiple of $(m_1 - 1, m_2 - 1)$.

Example: a shuffled generator

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.

There is, in addition to those mentioned above, a wide variety of generators in the literature that have been proposed. Some, like the *Tausworthe* generators, generate pseudo-random bits $\{0, 1\}$ according based on a primitive polynomial over a Galois Field and then map these bits into uniform (0,1) numbers. Others use a non-linear map to replace a linear one. For example we might define $x_{n+1} = x_n^2 \bmod(m)$ (called a *quadratic residue generator*) or $x_{n+1} = h(x_n) \bmod(m)$ for any function h designed to result in large values and more or less random low order bits.

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 $\{0, 1, ..., m-1\}$, it is suggested that we use $(x + 0.5)/m$.

2.3 Apparent Randomness of Pseudo-Random Number Generators

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 $\{0, 1, \ldots m-1\}$ to values approximately uniformly distributed on the unit interval $[0, 1]$ by dividing through by m (it

may sometimes be better to eliminate 0 by first adding 1/2 and then dividing through by m but for a multiplicative generator, since 0 does not occur, this is unnecessary). There is a large number of tests that can be applied to determine whether the hypothesis of independent uniform variates is credible (not, of course, whether the hypothesis is true since we know in advance it is not!).

Runs Test

Consider the hypothesis $H_0 : \{U_i, \quad i = 1, 2, ...\}$ are independent identically distributed random variables. The *runs test* seeks 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 regard may regard a sequence of the form .21 .24 .34 .37 .41 .49 .56 .51 .21 .25 .28 .56 .92 .96 as unlikely because the corresponding differences $+ + + + + + - - + + + + +$ has too few "runs" (here $R = 3$). Under the assumption of independence, $E(R) = \frac{2n-1}{3}$ and $var(R) = \frac{16n-29}{90}$ and we may approximate the distribution of R with the normal distribution for $n \geq 25$. We can look for runs of length n in batches in a longer sequence of variates.

Another alternative is the *serial correlation test*. The above test checks for the uniformity of the marginal distribution of $(x_n \ , \ x_{n+1})$ and this could obviously be generalized to variables separated by any number; say $(x_i, \ x_{i+j})$. One could also use the sample correlation or covariance as the basis for such a test. For example, for $j \geq 0$,

$$C_j = \frac{1}{n}(x_1 x_{1+j} + x_2 x_{2+j} + \ . \ . \ x_{n-j} x_n + x_{n+1-j} x_1 + \ . \ . \ . \ x_n x_j) \qquad (2.2)$$

The test may be based on the normal approximation to the distribution of C_j with mean $E(C_j) = 1/3, \ j = 0, \ 1/4$ for $j \geq 1$. Also $var(C_j) = \frac{13}{144n}, \ j \geq 1, \quad var(C_0) = 4/45n$.

Chi-squared test.

The chi-squared test can be applied to the sequence in any dimension, for example $k = 2$. We use the 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 $(U_{2j-1}, U_{2j}) \in A_i$. Clearly this should be related to the area or probability $P(A_i)$ of the set A_i. Pearson's chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{[N_i - nP(A_i)]^2}{nP(A_i)} \qquad (2.3)$$

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

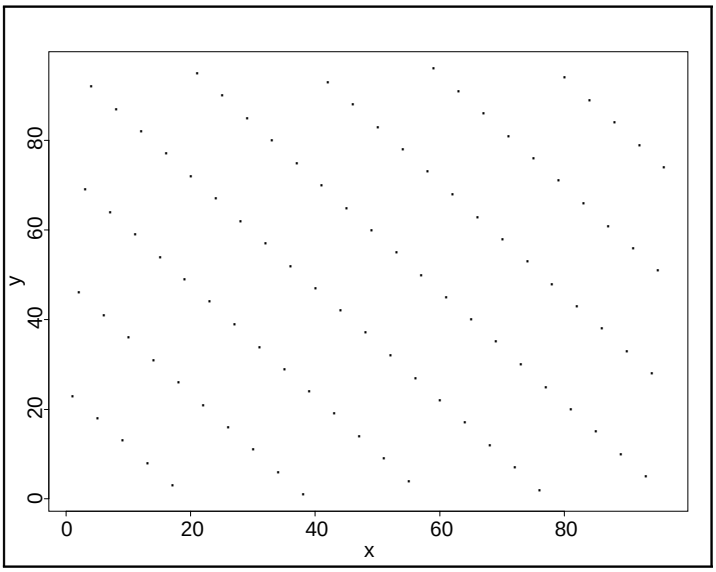


Figure 2.1:

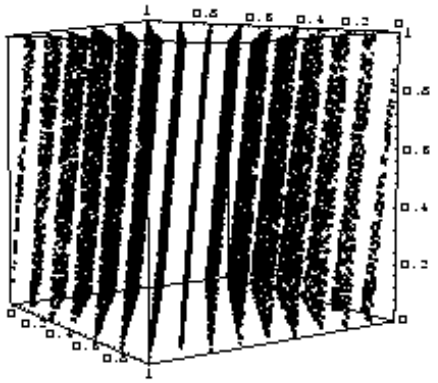
of the statistic that are unusually large for this distribution should lead to rejection of the uniformity hypothesis. The partition usually consists of squares of identical area but could, in general, be of arbitrary shape.

Spectral Test

Consecutive values plotted as pairs (x_n, x_{n+1}) , when generated from a multiplicative congruential generator $x_{n+1} = ax_n \; mod \; m$ fall on a *lattice*. A lattice is a set of points of the form $t_1e_1 + t_2e_2$ 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 $x_n = 23x_{n-1} mod \; 97$ when we plot adjacent pairs in 2-dimensional space. This plot is given below in Figure ??.

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 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 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 extension to a lattice in$k$-dimensions is done similarly. All linear

congruential random number generators result in points which when plotted as consecutive k-tuples lie on a lattice. The best generators are those for which the cells in the lattice are as close as possible to squares so that e_1 and e_2 are approximately the same length. Note that the area of the parallelogram with sides e_1 and e_2 is approximately a constant $(1/m)$ whatever the multiplier a so that a longer vector e_1 is associated with a shorter vector e_2 and therefore the two vectors of reasonably similar length. The *spectral test statistic* ν is the renormalized length of the first basis vector $||e_1||$. In general, for k consecutive points, it is equal to $min\ (b_1^2 + b_2^2 + . + . + . + b_k^2)^{1/2}$ under the constraint $b_1 + b_2 a + \ldots b_k a^{k-1} = mq, \quad q \neq 0$. 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 this test most spectacularly with $k = 3$ is the generator RANDU, used commonly in simulations until the 1980's and notorious for the fact that very few 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 a 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 **??**



Lattice Structure of RANDU

For example consider the following plot of 5000 consecutive triplets from a linear congruential random number generator with $a = 383, c = 263, m = 10,000$.

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. The table below gives some values of the spectral test statistic for some linear congruential random number generators in dimension $k \leq 7$.

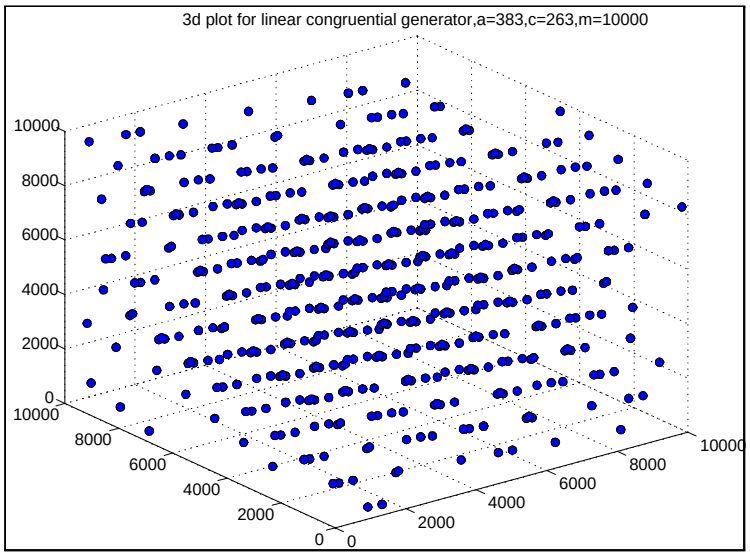


Figure 2.2:

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

2.4 Non-Uniform Random Number Generation

By far the simplest and most common method for generating non-uniform variates is based on the inverse cumulative distribution function. For arbitrary c.d.f. $F(x)$, 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 c.d.f. is continuous and strictly increasing. However, in the general case of a possibly discontinuous non-decreasing c.d.f. the function continues to enjoy some of the properties of an inverse. In particular, in the general case,

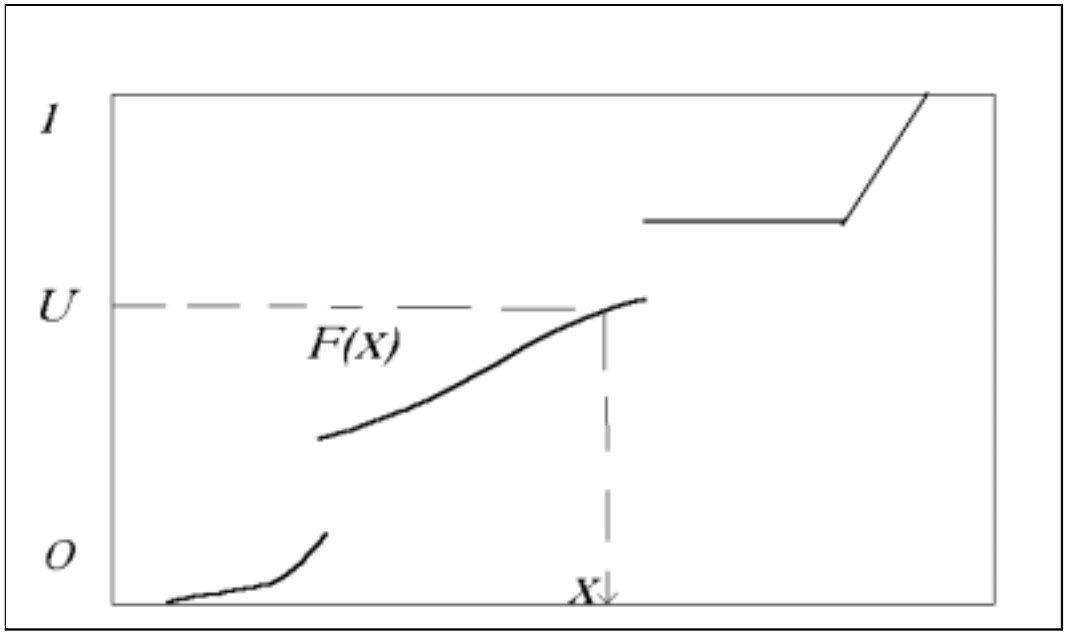


Figure 2.3:

Theorem 11 *(: inverse transform)*

If F is an arbitrary c.d.f. and U is uniform$[0,1]$ then $X = F^{-1}(U)$ has c.d.f. $F(x)$.

Proof:

The proof is a simple consequence of the fact that

$$[U < F(x)] \subset [X \leq x] \subset [U \leq F(x)] \text{ for all x,} \tag{2.4}$$

evident from Figure **??**. Taking probabilities throughout **??**, 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

$f(x) = \frac{1}{\theta}e^{-x/\theta}, x \geq 0$ 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,

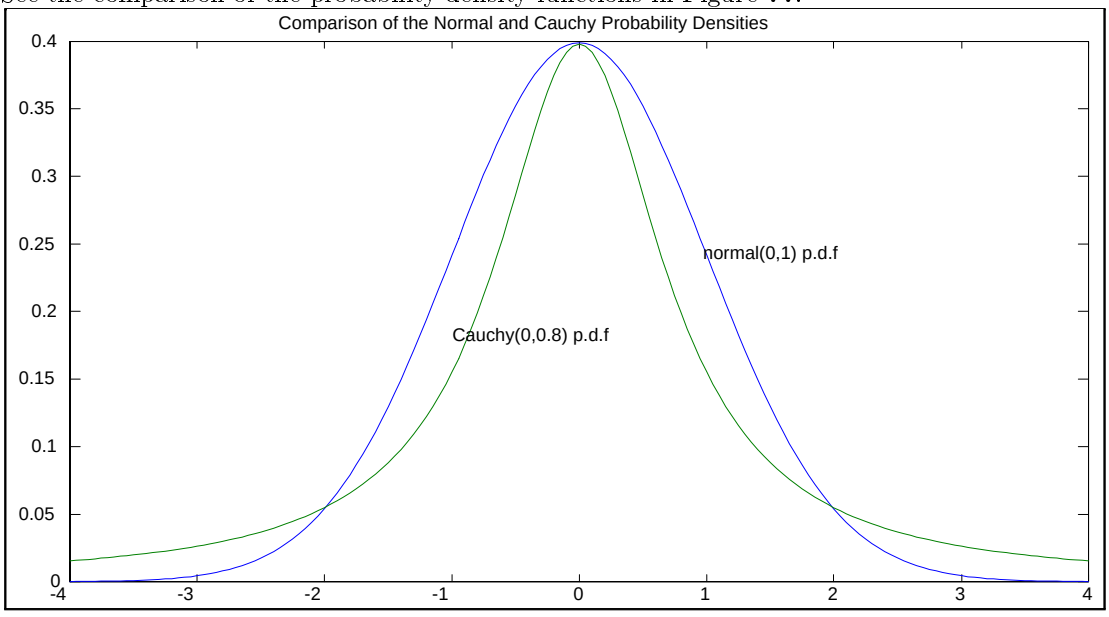
$$\begin{aligned} X &= -\theta ln(1-U) \quad \text{or} \\ X &= -\theta lnU \text{ since } U \text{ and } 1-U \text{ have the same distribution.} \end{aligned}$$

Cauchy (a, b)

This distribution is a member of the "stable family " of distributions. 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 **??**.



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

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

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

Geometric (p)

The probability function is $f(x) = p(1-p)^x, x = 1, 2, 3, ...$ and the cumulative distribution function is $F(x) = 1 - (1-p)^{[x]}$, $x \geq 0$ where $[x]$ denotes the

integer part of x. Then

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

where we write $\log(1-U) = E$, an exponential(1) random variable.

Pareto (a, b)

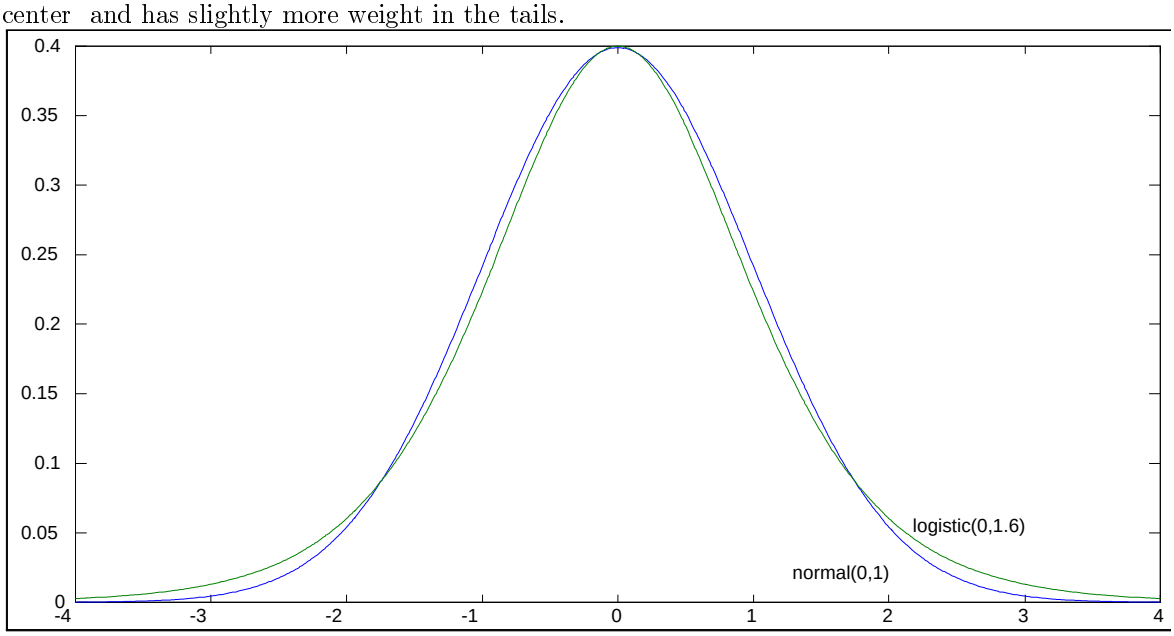
This is one of the simpler families of distributions used in econometrics for modeling quantities with lower bound b (often equal to 0).

$F(x) \;=\; 1 - \left[\frac{b}{x}\right]^a, \quad x \geq b > 0$. Then

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

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 **??**, the two densities are almost indistinguishable, except that the logistic is very slightly more peaked i the center and has slightly more weight in the tails.



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

$F(x) = 1 - exp\{-exp[(x-a)/b]\}$. Then

$$X = a + b\ log\ log(U)$$

In *Matlab*, the exponential and geometric random number generators are called *exprnd, geornd* respectively and the Cauchy distribution can be generated using the central student's t generator *trnd*. For example, *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 we produce a vector of length n with the associated parameter values.

DISTRIBUTION	SPLUS	MATLAB
normal	rnorm(n, μ, σ)	normrnd$(\mu, \sigma, 1, n)$ or randn$(1, n)$ if $\mu = 1, \sigma = 1$
Student's t	$rt(n, \nu)$	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)	
binomial	rbinom(n, m, p)	binornd$(m, p, 1, n)$
Poisson	rpois(n, λ)	poissrnd$(\lambda, 1, n)$

Inversion performs reasonably well for any distribution for which the cumulative distribution function and its inverse can be found in closed form and computed reasonably efficiently. This includes, as well as the distributions above, the Weibull, the logistic distribution and most discrete distributions which are reasonably well concentrated about the mode. However, for other distributions such as the Poisson with large mean, or the normal, chi-squared, beta etc. other methods need to be used.

In some circumstances, when both the c.d.f. and the probability density are known, we might attempt to invert the c.d.f. by numerical methods. For example, if we use the Newton-Ralphson method, we would iterate until convergence the equation

$$X = X - \frac{F(X) - U}{f(X)} \tag{2.5}$$

beginning with a good approximation to X. For example we might choose the initial value of $X = X(U)$ by using an easily inverted c.d.f. to approximate the true c.d.f.

Suppose $F(x)$ is a cumulative distribution function and $f(x)$ is the corresponding probability density function. Consider the transformation

$$x(u,v) = F^{-1}(u), \quad y(u,v) = vf(F^{-1}(u)), 0 < u < 1, 0 < v < 1$$

This maps a pair of random variables (U, V) which is uniform on the unit square into points uniformly distributed under the graph of the probability density f. This is most easily seen by examining the inverse transformation: $U = F(X), V = Y/f(X)$. The variate V is not needed here if our only objective is producing X with given c.d.f. This is standard inversion. Nevertheless, the fact that the point (X, Y) is uniform under the density underlies 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 simple result.

Theorem 12 . *(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

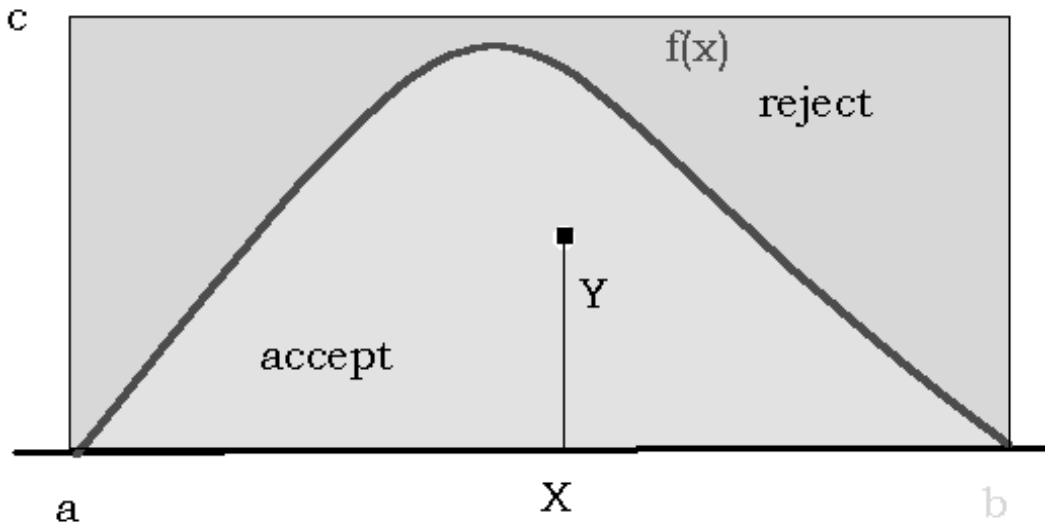
If a point (X, Y) is uniformly distributed under the graph of $f(x)$ notice that the probability $P[a < X < b]$ is proportional to the area under the graph betwen the lines at $x = a$ and $x = b$. In other words $P[a < X < b]$ is proportional to $\int_a^b f(x)dx$. This implies that $f(x)$ is proportional 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.

The *acceptance-rejection method* works as follows. Suppose $g(x)$ is some easy density for which we can generate variates say by inversion. Suppose we wish to generate a variate X from the harder density $f(x)$ where $f(x) \leq cg(x)$ for some $c > 1$ and for all x. We generate a point uniformly under the graph of $cg(x)$ and then accept that point (in particular X , the x-coordinate of the point) if it turns out to be also below the graph of $f(x)$. Otherwise generate a new point (X, Y), repeating until the condition is satisfied. See Figure **??** where it is assumed that $g(x)$ is the uniform probability density function on the interval $[a, b]$. In algorithmic form, the acceptance-rejection method is;

REPEAT: Generate independent random variables X, U where X has density g and U is uniform on $[0, 1]$.

UNTIL: $U \leq \frac{f(X)}{cg(X)}$

THEN RETURN X.



The Acceptance-Rejection Algorithm

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 a point satisfies the condition 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. We have seen that 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 the unit square $[0,1]^2$. Notice that any such transformation defines a random number generator for the density $f(x)$ since we can generate a uniform variable in the set A by rejection and when the point falls inside A apply the inverse transformation T^{-1} to this point. The first coordinate X will then have density f. We can think of inversion as a mapping on $[0,1]$ and acceptance-rejection algorithms as an area preserving mapping on $[0,1]^2$.

The most common distribution required for simulations in finance and elsewhere is the *normal distribution*. Recall that if (X,Y) are independent standard normal variates, then expressed in polar coordinates,

$$(R,\Theta) = (\sqrt{X^2+Y^2},\ arctan(Y/X)) \tag{2.6}$$

are distributed as independent variates. $\sqrt{X^2+Y^2}$ has the distribution of the square root of a chi-squared(2) or exponential(2) variable. The distribution of $arctan(Y/X))$ is uniform 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* $[0, 2\pi]$. This result is left as a problem.

This observation is the basis of two of the most common normal generators. The *Box-Muller* algorithm uses two uniform$[0, 1]$ variates U, V to generate R and Θ with the above distributions as

$$R = \{-2ln(U)\}^{1/2}, \quad \Theta = 2\pi V \tag{2.7}$$

and then defines two independent normal(0,1) variates as

$$(X, Y) = R(cos\Theta, sin\Theta) \tag{2.8}$$

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 13 *(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* $U[0, 2\pi]$ *distribution. Then* $(X, Y) = R(cos\ \Theta, sin\ \Theta)$ *is distributed as a pair of independent normal variates.*

Proof. Since R^2 has an exponential distribution, R has probability density function

$$f_R(r) = \frac{d}{dr}(1 - e^{-r^2/2}) = re^{-r^2/2}, \quad \text{for } r > 0.$$

and Θ has probability density function $f_\Theta(\theta) = \frac{1}{2\pi}$ for $0 < \theta < 2\pi$. The Jacobian of the transformation is

$$|\frac{\partial(r, \theta)}{\partial(x, y)}| = \begin{vmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} = (x^2 + y^2)^{-1/2}.$$

Consequently the joint probability density function of (X, Y) is given by

$$f_\Theta(\arctan(y/x)) f_R((x^2 + y^2)^{-1/2}) |\frac{\partial(r, \theta)}{\partial(x, y)}| = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

and this is joint probability density function of two independent standard normal random variables. ■

An alternative algorithm for generating standard normal variates is the *Marsaglia polar* method. This is a modification of the Box-Muller generator and avoids the calculation of *sin* or *cos*. Here we generate a point (Z_1, Z_2) from the uniform distribution on the unit circle. This is done by rejection, generating the point initially from the square $-1 \le z_1 \le 1, \quad -1 \le z_2 \le 1$ and accepting it when it falls in the unit circle or if $z_1^2 + z_2^2 \le 1$. Note that we

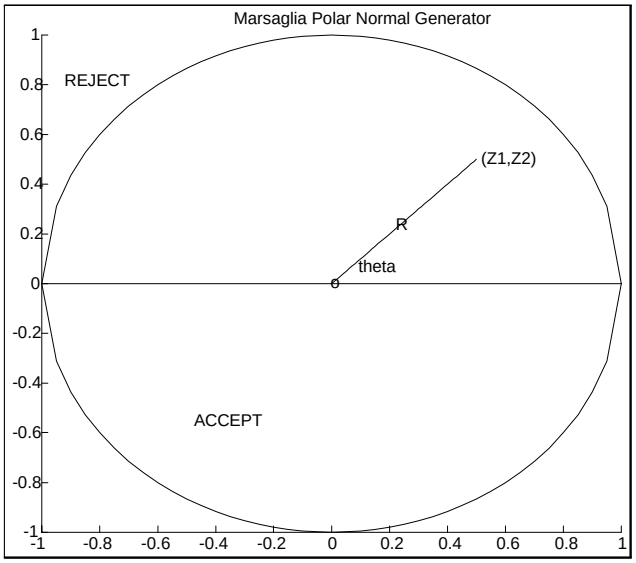


Figure 2.4:

can replace R^2 by $-2log(Z_1^2+Z_2^2)$ and $cos(\Theta), sin(\Theta)$ by $\frac{Z_1}{\sqrt{Z_1^2+Z_2^2}}$ and $\frac{Z_2}{\sqrt{Z_1^2+Z_2^2}}$ respectively. Thus the pair of independent standard normal variables is given by

$$(X,Y) = \sqrt{-2log(Z_1^2+Z_2^2)}(\frac{Z_1}{\sqrt{Z_1^2+Z_2^2}}, \frac{Z_2}{\sqrt{Z_1^2+Z_2^2}}) \text{ when } Z_1^2+Z_2^2 < 1$$

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 **??**:

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 deviation σ and *rand(m,n)* generates an $m \times n$ matrix of standard normal random numbers.

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 $\eta = \exp\{\mu + \sigma^2/2\}$ and volatility parameter σ. Note that a random variable with a lognormal distribution is strictly positive, making it a good candidate for modelling stock prices. The lognormal probability density function with mean $\eta > 0$ and volatility parameter $\sigma > 0$ is given by the probability density function

$$g(x|\eta,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(log\ x - log\ \eta\ - \sigma^2/2)^2/2\sigma^2\}. \tag{2.9}$$

A random variable with a lognormal distribution is easily generated by generating an appropriate normal random variable Z and then exponentiating.

A Discrete Time Black-Scholes Model

Suppose that a stock price $S_t, t = 1, 2, 3, ...$ has a lognormal distribution such that the returns over non-overlapping independent periods are independent. Let us assume that there is a total of N such periods in a year. In other words, we require that $S_t = S_0 \exp\{\sum_{i=1}^t Z_i\}$ for independent normal random variables Z_i which have an expected value which may depend on i. We assume that $var(Z_i) = \sigma^2/N$ so that the parameter σ^2 represents the volatility parameter of the stock price after one year. This is a fairly natural model since $S_t = S_0 \Pi_{i=1}^t e^{Z_i}$ is the product of independent returns over the t periods. 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 such that the stock price discounted to the present forms a martingale, and let us assume for the present that under the risk neutral measure, the stock price process has a similar lognormal representation $S_t = S_0 \exp\{\sum_{i=1}^t Z_i\}$ for independent normal random variables Z_i where the Z_i may have a different mean. Full justification of this process really relies on the continuous time version of the Black Scholes described in Section 1.6. Note that if the process

$$e^{-rt/N} S_t = S_0 \exp\{\sum_{i=1}^t (Z_i - \frac{r}{N})\}$$

is to form a martingale, it is necessary that

$$\begin{aligned} E^Q[S_0 \exp\{\sum_{i=1}^{t+1}(Z_i - \frac{r}{N})\}|H_t] &= S_0 \exp\{\sum_{i=1}^t (Z_i - \frac{r}{N})\} E^Q[\exp\{Z_{t+1} - \frac{r}{N}\}] \\ &= S_0 \exp\{\sum_{i=1}^t (Z_i - \frac{r}{N})\} \end{aligned}$$

and so $\exp\{Z_{t+1} - \frac{r}{N}\}$ must have a lognormal distribution with expected value 1. In other words, for each i the expected value of $Z_i - \frac{r}{N}$ is, under Q, equal to

$-\sigma^2/2N$ and variance σ^2/N. Notice then that under Q, S_T has a lognormal distribution with mean

$$S_0 e^{rT/N}$$

and volatility parameter $\sigma\sqrt{T/N}$. We can price a call option with maturity T periods from now by generating the random path $S_t, t = 1, 2, ...T$ using the lognormal distribution for S_T and then discounting the payoffs to the present and then averaging the results; i.e. the value of a call with exercise price K is an average of simulated values of

$$e^{-rT/N}(S_0 \exp\{\sum_{t=1}^{T} Z_i\} - K)^+, \text{ where } Z_i \text{ are independent } N(\frac{r}{N} - \sigma^2/(2N), \sigma^2/N).$$

The following function simulates the stock price over the whole period until maturity and then values a European call option on the stock by averaging.

Example 14 *(simulating a call option)*

Consider simulating a call option on a stock whose current value is \$1.00. The option expires in T days and the strike price is K. We assume constant spot interest rate r and the stock price follows a lognormal distribution with annual volatility 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 days from now,
% current stock price=$1.
% r = annual spot interest rate
% sigma=annual vol. K= strike price.
N=250 ; % N is the assumed number of periods in a year.
s = sigma/sqrt(N); %s is volatility per period
mn = r/N - s^2/2; % mean of the normal increments per period
y=exp(cumsum(normrnd(mn,s,T,1)));
y=[1 y'];
x = (0:T)/N;
plot(x,y,'-',x,K*ones(1,T+1),'y')
xlabel('time (in years)')
ylabel('value of stock')
title('SIMULATED RETURN FROM CALL OPTION')
z = exp(-r*T/N)*max(y(T+1)-K, 0);        % payoff from option discounted to present
```

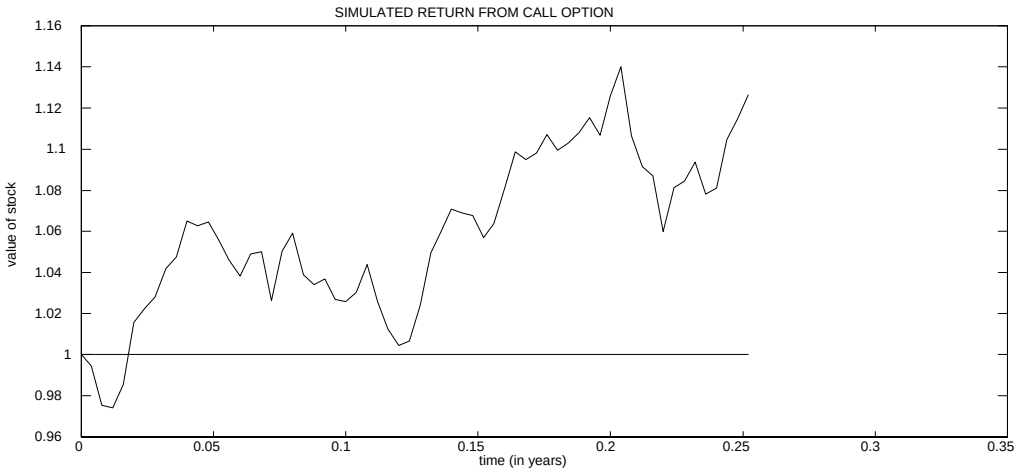


Figure 2.5:

Figure **??** resulted from one simulation run with $r = .05, T = 63$ (about 3 months), $\sigma = .20, K = 1$.

The return on this run was the discounted difference between the terminal value of the stock and the strike price or 0.113. We may repeat this many times, averaging the result and then discounting these returns to the present as in the following function.

```
function z = simcall(r, sigma, T, K, m)
% repeats plotlogn a total of m times and averages discounted return to the present.
z=[];
hold on
for i=1:m
z = [z plotlogn( r, sigma, T, K )];
end
optionval=mean(z);
disp([' option val= 'num2str(optionval)])
hold off
```

For example to value an at the money call option with exercise price=the initial price of the stock=\$1, we type

simcall(.05,.2,63,1,100); and obtain the output *option val= 0.044978.* If we repeat the identical statement, the output is different, i.e. *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 1.6. The price of a call option at time $t = 0$ given by

$$V(S_T, T) = S_T\Phi(d_1) - Ke^{-rT/N}\Phi(d_2)$$

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)

of CALL=0.0464. With these parameters, 4.6 cents on the dollar allows us to lock in the present 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.

We turn now to algorithms for generating the *Gamma distribution* with density

$$f(x|a, b) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, \quad x > 0 \tag{2.10}$$

Recall that the exponential ($a = 1$) and the chi-squared ($a = \nu/2$, $b = 2$, ν *integer*) are special cases of the Gamma distribution. The following result lists some of the properties of the Gamma distributions.

Theorem 15 *(Gamma distribution)*

If X_1, X_2 are independent Gamma (a_1, b) and Gamma (a_2, b) variates, then $Z = \frac{X_1}{X_1+X_2}$ and $Y = X_1 + X_2$ are independent variates with the beta (a_1, a_2) and the Gamma $(a_1 + a_2,\ b)$ distributions respectively. Conversely, if $(Z,\ Y)$ are independent variates with the latter pair of distributions, then $X_1 = YZ$, $X_2 = Y(1 - Z)$ have the indicated Gamma distributions.

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*

$$f_{X_1X_2}(x_1, x_2) = kx_1^{a_1-1}x_2^{a_2-1}e^{-(x_1+x_2)/b}, x_1 > 0, x_2 > 0$$

where k is the constant $[\Gamma(a_1)\Gamma(a_2)]^{-1}$. Consider the change of variables $x_1 = zy, x_2 = (1 - z)y$. Then the Jacobian $\begin{vmatrix} \frac{\partial x_1}{\partial z} & \frac{\partial x_1}{\partial y} \\ \frac{\partial x_2}{\partial z} & \frac{\partial x_2}{\partial y} \end{vmatrix} = y$. The joint probability density function of (z, y) is given by

$$\begin{aligned} f_{z,y}(z, y) &= f_{X_1X_2}(zy, (1 - z)y) \begin{vmatrix} \frac{\partial x_1}{\partial z} & \frac{\partial x_1}{\partial y} \\ \frac{\partial x_2}{\partial z} & \frac{\partial x_2}{\partial y} \end{vmatrix} \\ &= kz^{a_1-1}(1 - z)^{a_2-1}y^{a_1+a_2-1}e^{-y/b}, 0 < z < 1, y > 0. \end{aligned}$$

and this is easily seen to be 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. ■

This result is a basis for generating gamma variates with integral shape parameter a since this can be done by adding independent exponential variates. Thus $-log(\prod_{i=1}^{n} U_i)$ generates a gamma $(n, 1)$ 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 .

For large a one successful algorithm is due to Cheng (1977) and involves rejection from the *Burr XII* density of the form

$$g(x) = \lambda\mu \frac{x^{\lambda-1}}{(\mu + x^{\lambda})^2} \tag{2.11}$$

generated by inverse transform as $\{\frac{\mu U}{1-U}\}^{1/\lambda}$. Assume that the scale parameter of the Gamma $b = 1$. Matching the modes of these two distributions for large a results in choosing $\mu = a^{\lambda}$ and choosing λ to minimize $max\{f(x)/g(x)| - \infty < x < \infty\}$ results in $\lambda = \sqrt{2a-1}$. In this case, $c = \frac{4a^a e^{-a}}{\lambda\Gamma(a)}$ and this approaches $\sqrt{4/\pi}$ as $a \to \infty$.

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

$$cg(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)} , \quad x \le b \tag{2.12}$$

and

$$cg(x) = \frac{b^{\alpha-1}e^{-x}}{\Gamma(\alpha)} , \quad x > b \tag{2.13}$$

where the efficiency is determined by $c = \int cg(x) = \frac{b^{\alpha}}{\Gamma(\alpha+1)} + f(b)$ and we would clearly try to choose b corresponding as closely as possible to a minimum of this quantity. Ahrens and Dieter use $b = 1$. 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 variate as

$$Y = \frac{\sqrt{2}(U - 1/2)}{\sqrt{U(1-U)}} \tag{2.14}$$

where $U \sim U[0,1]$. Then Y has the Student(2) density

$$g(y) = \frac{1}{(2+y^2)^{3/2}}. \tag{2.15}$$

We then generate X as $(\alpha-1)+Y\sqrt{3\alpha/2 - 3/8}$ and apply a rejection step to X . See Devroye (p. 408) for details.

Most of the above algorithms are reasonably efficient only for $\alpha > 1$ 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, $\alpha > 1$.

A final alternative for the case $\alpha < 1$ is the use of Stuart's theorem which states that $XU^{1/\alpha}$ *has a gamma* $(\alpha,1)$ *distribution when* U *is uniform* $[0,1]$ *and* X *is Gamma* $(\alpha+1,1)$. The Matlab function *gamrnd* uses Best's algorithm and acceptance rejection for $\alpha > 1$ and for $\alpha < 1$, it uses Johnk's generator. This consists of generating U and V both independent U[0,1] and setting

$$X = \frac{U^{1/\alpha}}{U^{1/\alpha}+V^{1/(1-\alpha)}}$$

conditional on the denominator $U^{1/\alpha}+V^{1/(1-\alpha)} < 1$. Multiplying by an independent exponential (1) results in a Gamma$(\alpha,1)$ random variable.

We now turn to generating the *beta distribution* which has density given by

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}, \quad 0 \le x \le 1 \tag{2.16}$$

The beta density obtains as a transformation of an F-distribution (the basis of the transformation is the theorem on page ..), or as the distribution of order statistics in a sample from independent uniform $[0,1]$ variates. The variable $Z = \frac{X_1}{X_1+X_2}$ indicates 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).

Theorem 16 *(Beta distribution)*

Suppose U, V *are independent uniform* $[0,1]$ *variates. Then the conditional distribution of*

$$X = \frac{U^{1/a}}{U^{1/a}+V^{1/b}} \tag{2.17}$$

given that $U^{1/a}+V^{1/b} \le 1$ *is Beta* (a,b). *Similarly the conditional distribution of* $U^{1/a}$ *given that* $U^{1/a}+V^{1/b} \le 1$ *is Beta* $(a+1,b)$.

Proof. *Define a change of variables*

$$\begin{aligned} X &= \frac{U^{1/a}}{U^{1/a}+V^{1/b}}, Y = U^{1/a}+V^{1/b} \\ \text{or } U &= (YX)^a \text{ and } V=[(1-X)Y]^b \end{aligned}$$

so that the joint probability density function of (X,Y) *is given by*

$$\begin{aligned} f_{X,Y}(x,y) &= f_{U,V}((yx)^a,[(1-x)y]^b)\left|\begin{array}{cc}\frac{\partial u}{\partial x} & \frac{\partial u}{\partial y}\\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y}\end{array}\right| \\ &= aby^{a+b-1}x^{a-1}(1-x)^{b-1} \quad \text{provided } 0<x<1 \text{ when } y<1 \\ \text{or provided } 1-\frac{1}{y} &< x < \frac{1}{y} \text{ when } 1<y<2. \end{aligned}$$

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 3.3.12 and 3.3.15 together provide an algorithm for generating Gamma variates with non-integral α 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.

The Symmetric Stable Laws.

A final family of distributions of increasing importance in modelling is the *symmetric stable family*. These are 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, because they tend to fit observations better in the tail of the distribution than does the normal. However, 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. They are most easily described by their *characteristic function*, which, upon setting location equal to 0 and scale equal to 1 is $Ee^{iXt} = e^{-|t|^\alpha}$ where here i is the complex number $i^2=-1$. The parameter

$0 < \alpha \leq 2$ indicates what moments exist, for except in the special case $\alpha = 2$ (the normal distribution), moments of order less than α exists while moments of order α or more do not. 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 $\alpha = 2$ but can be determined from the series expansion of the probability density

$$f_c(x) = \sum_{k=0}^{\infty} (-1)^k \frac{\Gamma((k+1)/2)}{\pi c \alpha k!} \cos(\frac{k\pi}{c})(\frac{x}{c})^k$$

where c is a scale parameter. Especially for large values of x, this probability density function converges extremely slowly. According to Chambers, Mallows and Stuck, (1976), such a variate can be generated in the case $\alpha \neq 1$,

$$X = sin(\alpha U) \left[\frac{cos(U(1-\alpha))}{E} \right]^{\frac{1}{\alpha}-1} (cos\ U\ )^{-1/\alpha} \tag{2.18}$$

where U is uniform $[-\pi/2,\ \pi/2]$ and E is standard exponential and independent. The case $\alpha = 1$ is the Cauchy $X = tan(U)$. Since the tan is a relatively slow operation, this is sometimes replaced by the ratio of Normal variates produced by Marsaglia's polar algorithm. In other words we use $X = V_1/V_2$ where $V_i \sim U[-1, 1]$ conditional on $V_1^2 + V_2^2 \leq 1$ as a standard Cauchy variate.

Example: Stable random walk.

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 **??**. Note the occasional very large jump which dominates the history of the process up to that point.

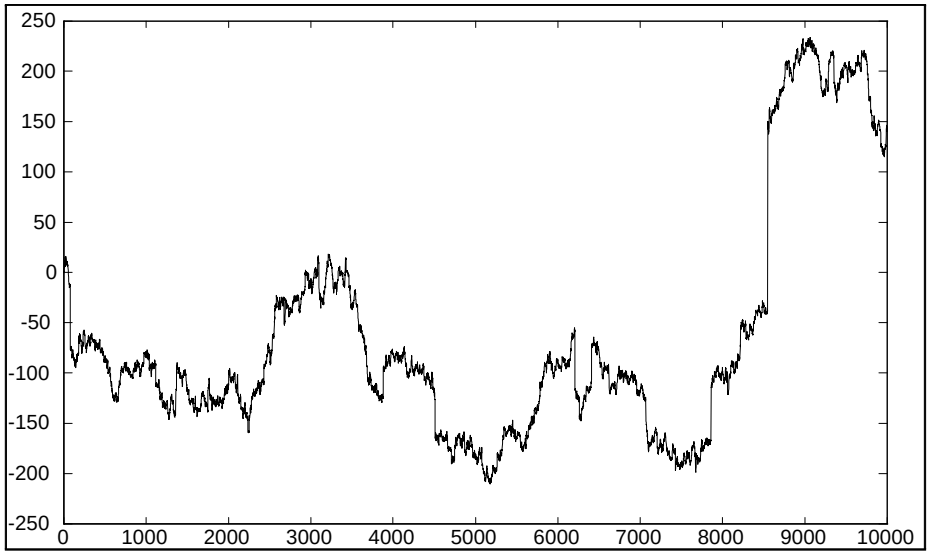


Figure 2.6:

2.5 Generating Discrete Distributions

Many of the methods described above such as inversion and rejection for generating continuous distributions work well for discrete random variables. For example, if X is a discrete distribution taking values on the integers with probability function $f(x)$, $x = 0, 1, \ldots$ we may use rejection to generate a continuous variate Y which has the same cumulative distribution function at the integers $F_Y(j) = F_X(j)$ and then set $X = \lfloor Y \rfloor$ the integer part of Y.

Inversion for discrete variates often requires, for reasonable efficiency, some setup costs. For example if X has c.d.f. $F(x)$, $x = 0, 1, \ldots$ we wish to output an integer X satisfying $F(X-1) < U \leq F(X)$ and the most obvious technique for finding such a value of X is to search sequentially through the potential values $0, 1, 2, \ldots$. Figure **??** is the search tree for inversion for the distribution on the integers $1, \ldots 5$ given by $(p_1, p_2, p_3, p_4, p_5) = (0.11, 0.30, 0.25, 0.21, 0.13)$ 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. 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 the median m (or mode or mean) of the distribution, searching to the left or right depending on the value of U as in Figure **??**.

This results in searching an average of $E[|X + 1 - m|]$ before obtaining the generated variable often substantially smaller than $E(X)$ when the mean is

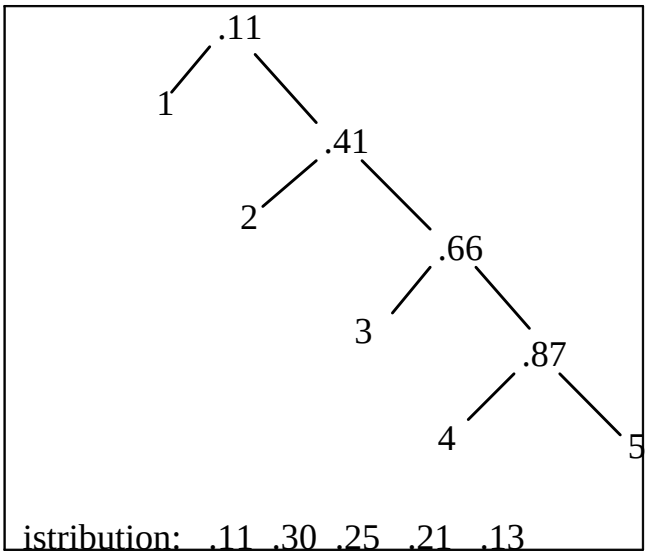


Figure 2.7:

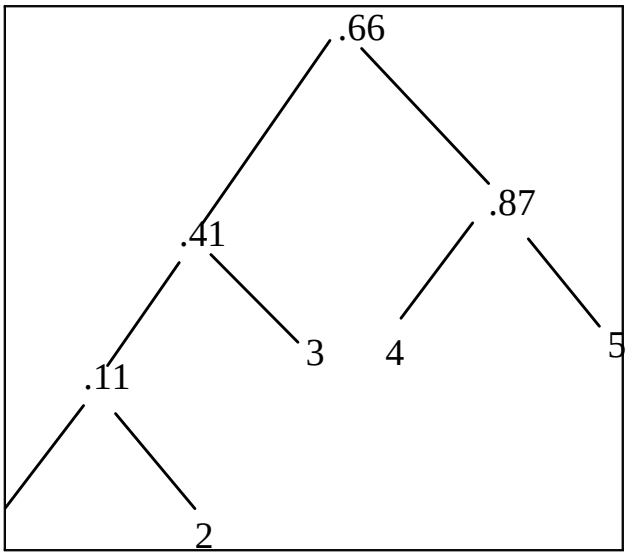


Figure 2.8:

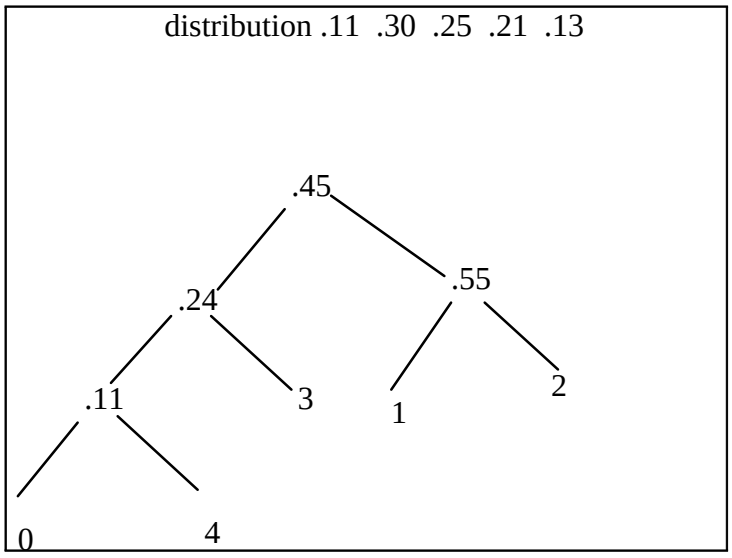


Figure 2.9:

large but still unacceptably large when the distribution has large variance. An optimal binary search tree for this distribution is graphed in Figure **??**. This tree has been constructed by joining the two smallest probabilities to form a new node with weight a combination of the two, and hence working from the leaves to the root of the tree. Equivalently, we use inversion after a re-ordering of the values from those with smallest to those with largest probability.

The leaves of the tree are the individual p_i and the internal nodes are sums of the weights of the children. If D_i represents the depth of the $i'th$ leaf, then the average number of comparisons to generate a single X is $\sum_i p_i D_i$ and the procedure for constructing this tree provides an optimal algorithm in the sense that this quantity is minimized. It is interesting to note 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*. 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, unlike inversion or binary search which increase as $E(X)$ increases. The idea is to reduce any discrete distribution to a uniform mixture of two-point distributions. For a discrete distribution of the form $p_1, p_2, \ldots p_K$ on the integers $1,\ 2,\ \ldots K$, we seek a table of values of $A(i)$ and associated alias probabilities $q(i)$ so that the following algorithm generates the desired discrete distribution.

GENERATE I UNIFORM ON $\{1, \ldots K\}$.

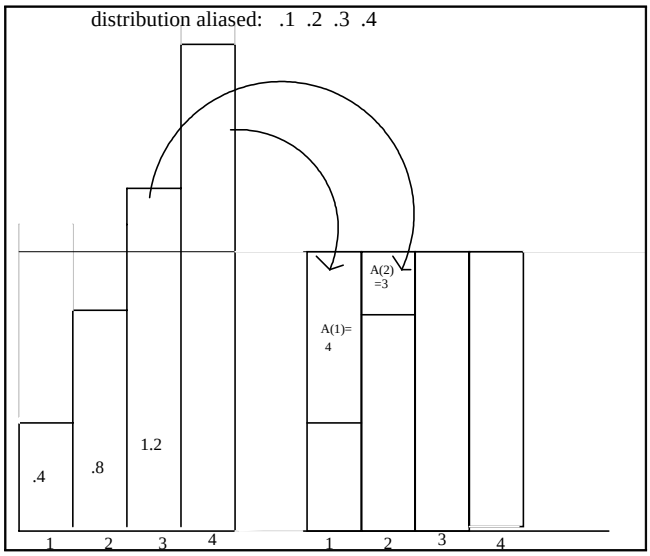


Figure 2.10:

WITH PROBABILITY $q(I)$, OUTPUT $X = I$, OTHERWISE, $X = A(I)$.

An algorithm for producing these values of $A(i)$, $q(i)$ is suggested by Walker(1977) and proceeds by reducing the number of non-zero probabilities one at a time.

1. Put $q_i = Kp_i$ for all i.

2. LET m be the index of $min\{q_i; q_i > 0\}$ and let M be the index of the maximum.

3. SET $A(m) = M$ and fix $q(m)$.

4. Replace $(q_1, \ldots q_K)$ by $(q_1, ..., q_{m-1}, q_{m+1}, ...q_M - (1 - q_m)\ldots$ (so the component with index m is removed).

5. Return to 2 unless all remaining $q_i = 1$ or the vector of q_i's is empty.

Figure **??** shows the way in which aliasing iteratively adjusts a probability histogram to form a rectangle with base K. We construct the vector of aliasing probabilities and aliases for the distribution vector

$$p_i = (.1,\ .2,\ .3,\ .4), \quad i = 1, \ldots 4.$$

This results in $A(i) = (4, 3, x, x)$ and $q_i = (.4, .8, 1, 1)$ respectively.

The Poisson Distribution.

Consider the *Poisson distribution*

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \; x = 0, 1, \ldots \tag{2.19}$$

The simplest generator is to use the Poisson process. Recall that if points are distributed on the line in such a way that the spacings between consecutive points are independent exponential(1), then the resulting process is a Poisson process with rate 1. Thus, the number of points in an interval of length λ has the desired Poisson (λ) distribution. So for small λ, we could use $inf\{X; \sum_{i=1}^{X+1}(-ln\ U_i) > \lambda\}$ or equivalently

$$inf\{X; \prod_{i=1}^{X+1} U_i < e^{-\lambda}\} \tag{2.20}$$

Once again this generator requires time which grows linearly with λ and so an alternative for large λ is to use rejection. Various possibilities have been suggested for dominating function, from the logistic distribution(Atkinson (1979)) to a normal distribution with exponential right tail (cf. Devroye, lemma 3.8, page 509). A simple all-purpose alternative seems to be a table-mountain function (cf. Stadlober (1989)), essentially a function with a flat top and tails that decrease as $1/x^2$.

A 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

$$g(x|a, b) = \frac{c_0}{b^2 + (x-a)^2}, \quad x > 0 \tag{2.21}$$

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 $poissrnd(\lambda, m, n)$ which generates an $m \times n$ matrix of Poisson(λ) variables. This uses the simple generator **??** and is not computationally efficient for large values of λ.

The Binomial Distribution

For the *Binomial* distribution, we may use any one of the following alternatives:

(1) $\sum_{i=1}^{n} I(U_i < p), \quad U_i \sim uniform[0, 1]$

(2) $inf\{X; \sum_{i=1}^{X+1} G_i > n\}, \quad G_i \sim Geometric(p)$

(3) $inf\{X; \sum_{i=1}^{X+1} \frac{E_i}{n-i+1} > -log(1-p)\}, \quad E_i \sim Exponential(1).$

Methods (1) and (2) are more efficient when n is large but np fairly small. Of course for large n and np sufficiently small (e.g. <1), we can replace the Binomial distribution by its Poisson ($\lambda = np$) approximation. For large mean, a rejection method is required. Again we may use rejection beginning with a Lorentzian distribution, choosing $a = np$, $b = \sqrt{2np(1-p)}$ 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.

2.6 Simulating Stochastic Partial Differential Equations.

Consider a derivative product whose underlying asset has price X_t which satisfies a diffusion equation. Then if the derivative payoff function depends only on the current time and the current value of the asset, the Feynman- Kac theorem indicates that its value is an expectation of the form

$$V(S,t) = E[V_0(X_T)exp\{-\int_t^T r(X_v, v)dv\} \mid X_t = S] \tag{2.22}$$

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 careful simulation. The simplest version requires simulating a number of sample paths for the process X_v , evaluating **??** and averaging the results over all simulations. We begin by discussing the simulation of the process X_v.

Many of the stochastic models in finance reduce to simple diffusion equation (which may have more than one *factor* or dimension). Since most of the models in finance are Markovian, we restrict to the Markov diffusion model of the form

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t \tag{2.23}$$

with initial value for X_0 where 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 to adopt the crudest interpretation of the above equation, 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 (3.2) 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 $t_i = i\Delta x$, the equation **??** in integral form implies

$$X_{t_{i+1}} = X_{t_i} + \int_{t_i}^{t_{i+1}} a(X_s, s)ds + \int_{t_i}^{t_{i+1}} \sigma(X_s, s)dW_s \tag{2.24}$$

Ito's lemma can be written in terms of two operators on twice (with respect to x) differentiable functions f: In particular,

$$df(X_t, t) = L^o f dt + L^1 f dW_t \quad \text{where}$$

$$L^o = a\frac{\partial}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial t},$$

and

$$L^1 = \sigma\frac{\partial}{\partial x}.$$

Then for any twice differentiable function f,

$$f(X_{t_{i+1}}, t_{i+1}) = f(X_{t_i}, t_i) + \int_{t_i}^{t_{i+1}} L^0 f(X_s, s)ds + \int_{t_i}^{t_{i+1}} L^1 f(X_s, s)dW_s. \tag{2.25}$$

By substituting in each of the integrands in **??** using the above identity and iterating this process we arrive at the Ito-Taylor expansions (e.g. Kloeden and Platen, 1992). For example,

$$\int_{t_i}^{t_{i+1}} a(X_s, s)ds = \int_{t_i}^{t_{i+1}} \{a(X_{t_i}, t_i) + \int_{t_i}^{s} L^0 a(X_u, u)du + \int_{t_i}^{s} L^1 a(X_u, u)dW_u\}ds$$

$$\approx a(X_{t_i},t_i)\Delta t + L^0 a(X_{t_i},t_i)\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} duds + L^1 a(X_{t_i},t_i)\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} dW_u ds \tag{2.26}$$

The first term in **??**, $a(X_{t_i})\Delta t$, is an initial approximation to the desired integral. The rest, we may regard as an error term for the moment, and it has smaller order. For example it is easy to see that the second term is $O(\Delta t)^2$ because the integral $\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} duds = (\Delta t)^2/2$. The third term in **??** is $O(\Delta t)^{3/2}$ 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 $\int_{t_i}^{t_{i+1}}(t_{i+1}-u)^2 du = (\Delta t)^3/3$. 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

$$\int_{t_i}^{t_{i+1}} \sigma(X_s,s)dW_s = \int_{t_i}^{t_{i+1}} \{\sigma(X_{t_i},t_i) + \int_{t_i}^{s} L^0\sigma(X_u,u)du + \int_{t_i}^{s} L^1\sigma(X_u,u)dW_u\}dW_s$$

$$\approx \sigma(X_{t_i},t_i)\Delta W_t + L^0\sigma(X_{t_i},t_i)\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} dudW_s + L^1\sigma(X_{t_i},t_i)\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} dW_u dW_s$$

$$= \sigma(X_{t_i},t_i)\Delta W_t + \frac{\sigma(X_{t_i},t_i)\frac{\partial}{\partial x}\sigma(X_{t_i},t_i)}{2}[(\Delta W_t)^2 - \Delta t] + O(\Delta t)^{3/2}$$

Since the integral $\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} dW_u dW_s = \frac{1}{2}[(\Delta W_t)^2 - \Delta t]$. and $\int_{t_i}^{t_{i+1}}\int_{t_i}^{s} dudW_s = O(\Delta t)^{3/2}$. Putting these terms together, we arrive at an approximation to the increment of the form

$$\Delta X_t = a(X_{t_i},t_i)\Delta t + \sigma(X_{t_i},t_i)\Delta W_t + \frac{\sigma(X_{t_i},t_i)\frac{\partial}{\partial x}\sigma(X_{t_i},t_i)}{2}[(\Delta W_t)^2 - \Delta t] + O(\Delta t)^{3/2} \tag{2.27}$$

which allow an explicit representation of the increment in the process X in terms of the increment of a Brownian motion process $\Delta W_t \sim N(0,\Delta t)$. This is called the *Milstein approximation*. After the Euler scheme, 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 $O(\Delta t)^{3/2}$ *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 for details.

Example: Down-and-out-Call.

Consider an asset whose price follows a geometric Brownian motion process

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{2.28}$$

for a standard Brownian motion process W_t . A down-and-out call option with exercise price E provides the usual payment of a call option if the asset never falls below a given *out barrier* b . Use simulation to price such an option with exercise price E , current asset price S , time to maturity $T - t$, and out barrier $b < S$. Assume constant interest rate r, and evaluate the expected return from the *risk neutral process*, the one with the drift term replaced by $rS_t dt$. Then discount this return to the present by multiplying by $e^{-r(T-t)}$. 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

$$dS_t = rS_t dt + \sigma S_t dW_t \tag{2.29}$$

then $Y_t = log(S_t)$ satisfies

$$dY_t = (r - \sigma^2/2)dt + \sigma dW_t. \tag{2.30}$$

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 $min\{Y_s; t < s < T\} \geq log(b)$. The following *Matlab* functions were used.

```
function z=barrier(r,sigma,dt,T, e,b)
n=T/dt;
sigma = sigma*sqrt(dt);
mn = r*dt - sigma^2/2;
y=exp(cumsum(normrnd(mn,sigma,n,1)));
x = (1:n).*dt;
plot(x,y,'-',x,e*ones(1,n),'y',x,b*ones(1,n),'b')
z = max(y(n)-e, 0);
if min(y) < b
  z=0;
end;

function z = simbarr(r, sigma, dt,T, e,b, m)
hold on
for i=1:m
   z(i) = barrier( r, sigma,dt,T,e,b);
end
disp([‘mean value=’  num2str(exp(-r*T)*sum(z)/m)])
hold off
```

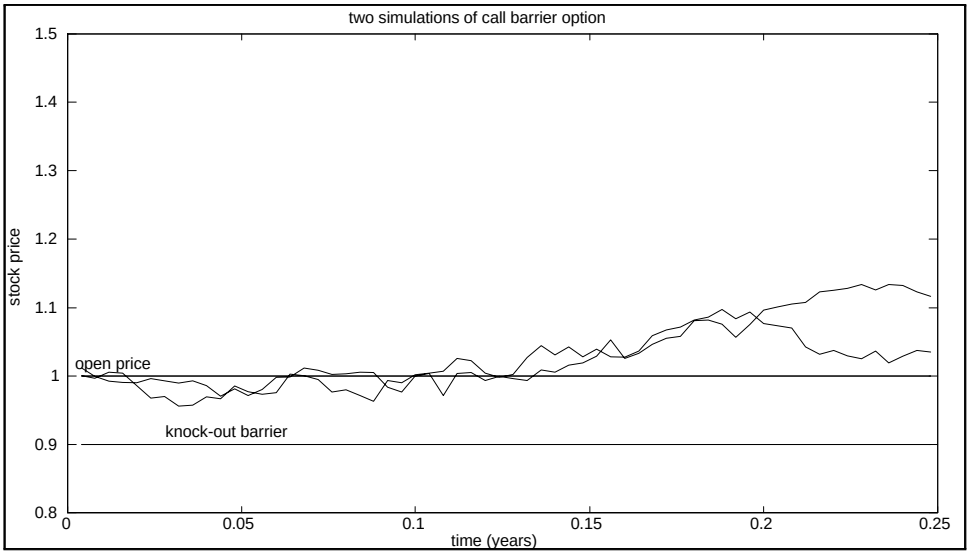


Figure 2.11:

For example when $r = .05, \sigma = .2, \Delta t = .004, T - t = .25, E = 1, b = .9$, this is run with the command

```
simbarr(.05, .2, .004, .25, 1, .9, 2)
```

for a total of 2 simulations, both, in this case, (see Figure **??**) ending in the money.

2.7 Problems

1. Consider the mixed generator $x_n = (ax_{n-1} + 1) mod(m)$ with $m = 64$. What values of a result in the maximum possible period and which generators appears more random?

2. Consider the multiplicative generator with $m = 64$. What values of a result in the maximum possible period and which generators appears more random?

3. Consider a shuffled described in section XXX with $k = 4, m_1 = 19, m_2 = 29$ X(i) = 2 14 7 13 11 18 6 3 9, Y(i) = 11 25 17 4 16 28 14 0 3 .

   We start by filling four pigeon-holes with the numbers produced by the first generator so that $(T_1, \ldots, T_4) = (2, 14, 7, 13)$. 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 numberd from 1 through 4, we use $I = \lceil 4 \times 11/29 \rceil = 2$. Then the first number in our random

sequence is drawn from box 2, i.e. $z_1 = T_2 = 14$, so $z_1 = 14$. This element of the vector is now replaced by 11, the next number in the X sequence. Proceeding in this way, the next index is $I = \lceil 4 \times 25/29 \rceil = 4$ and so the next number drawn is $z_2 = T_4 = 13$. Of course, when we have finished generating the values $z_1, z_2, ...$ all of which lie between 0 and $m_1 = 19$, 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]. Because of the small value we chose for m_1, this approximation will not be very good in this case. But the advantage of shuffling is that the period of the generator is greatly extended. Determine the period of the shuffled random number generator above and compare with the periods of the two constituent generators.

4. Prove: *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 $\{0, \ldots, m-1\}$.*

5. Consider the above quadratic residue generator $x_{n+1} = x_n^2 \bmod(m)$ with $m = 4783 \times 4027$. What is the period of the generator starting with seed $x_0 = 196$, or with seed $x_0 = 400$?

6. Verify that for the serial correlation statistic C_j, $var(C_j) = \frac{13}{144n}$, $j \geq 1$, $var(C_0) = 4/45n$.

7. Consider the turbo-pascal generator $x_{n+1} = (134775813x_n + 1) \bmod 2^{32}$. Generate a sequence of length 5000 and apply the serial correlation test. Is there evidence of dependence?

8. Generate 1000 daily "returns" $X_i, i = 1, 2, ..., 1000$ from each of the two distributions, the Cauchy and the logistic. In each case, assume that $a = 0, b = 0.1$. 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 average daily return.

9. Briefly indicate an efficient algorithm for generating one random variable from each of the following distributions and generate such a random variable using one or more of the uniform[0,1] random numbers.

U_i:	0.794	0.603	0.412	0.874	0.268	0.990	0.059	0.112	0.395

   (a) X$\sim U[-1, 2]$.
   (b) a random variable X with probability density function $f(x) = \frac{3}{16}x^{1/2}$, $0 < x < 4$
   (c) A discrete random number X having probability function $P[X = x] = (1-p)^x p$, $x = 0, 1, 2, ..., p = 0.3$.
   (d) A random variable X with the normal distribution, mean 1 and *variance* 4.

(e) A random variable X with probability density function

$$f(x) = cx^2e^{-x}, \quad 0 \leq x < 1$$

for constant $c = 1/(2 - 5e^{-1})$.

(f) A random variable X with the following probability function:

x	0	1	2	3
$P[X = x]$	0.1	0.2	0.3	0.4

10. Consider the multiplicative pseudo-random number generator

$$x_{n+1} = ax_n \bmod(150)$$

starting with seed $x_0 = 7$. Try various values of the multiplier a and determine for which values the period of the generator appears to be maximal.

11. 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 $c = 0$.

12. Evaluate the following integral by simulation:

$$\int_0^1 (1 - x^2)^{3/2} dx.$$

13. 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]. The interval $[a, b]$?

14. Evaluate the following integral by simulation:

$$\int_0^2 x^{3/2}(4 - x)^{1/2} dx.$$

15. Evaluate the following integral by simulation:

$$\int_{-\infty}^{\infty} e^{-x^2} dx.$$

(Hint: Rewrite this integral in the form $2\int_0^{\infty} e^{-x^2} dx$ and then change variables to $y = x/(1 + x)$)

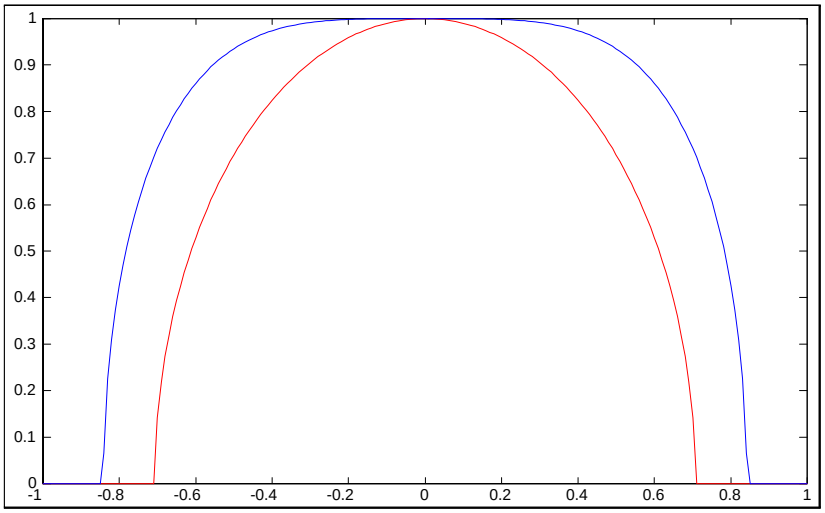


Figure 2.12:

16. Evaluate the following integral by simulation:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Discuss the efficiency of your approach.

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

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

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

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

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

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

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

(a) . LET $I = 0$

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

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

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

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

36. .Obtain generators for the following distributions:

(a) *Rayleigh*

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

(b) *Triangular*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

    (a) GENERATE U, V

    (b) PUT $X = V/U$

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

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

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

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

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

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

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

Chapter 3

Variance Reduction Techniques.

3.1 Introduction.

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

Example.

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

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

and the magnitude of the "shocks". The simulation is being used to estimate an integral

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

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

Discrete Event Simulation.

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

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

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

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

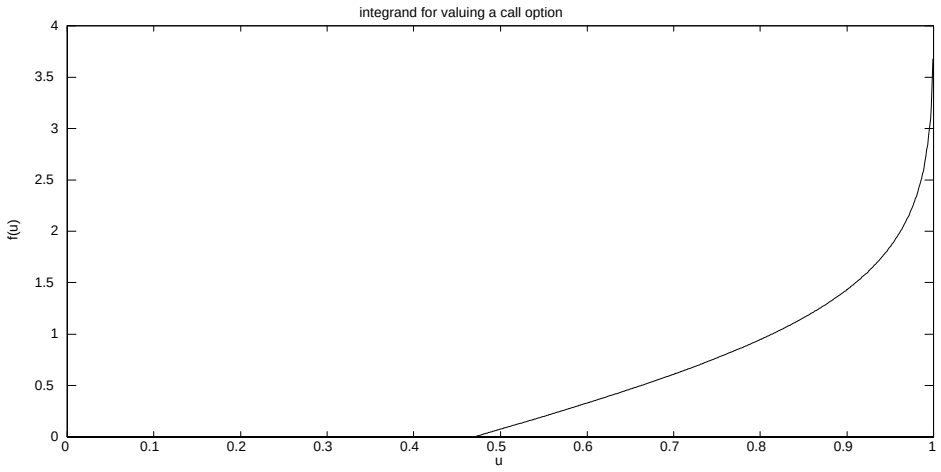


Figure 3.1:

Example: A crude simulation

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

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

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

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

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

```
%generate normal variate by inverse transform. T=maturity
%For Black-Scholes, integrate over (0,1).
x=S0*exp(norminv(u,r*T-sigma^2*T/2,sigma*sqrt(T)));   % stock price at time
```
T=$S_0 \exp\{\Phi^{-1}(U; rT - \frac{\sigma^2}{2}T, \sigma^2 T)\}$
```
v=exp(-r*T)*max((x-ex),0);                    % This is the discounted to
present value of the call option
```

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

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

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

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

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

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

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

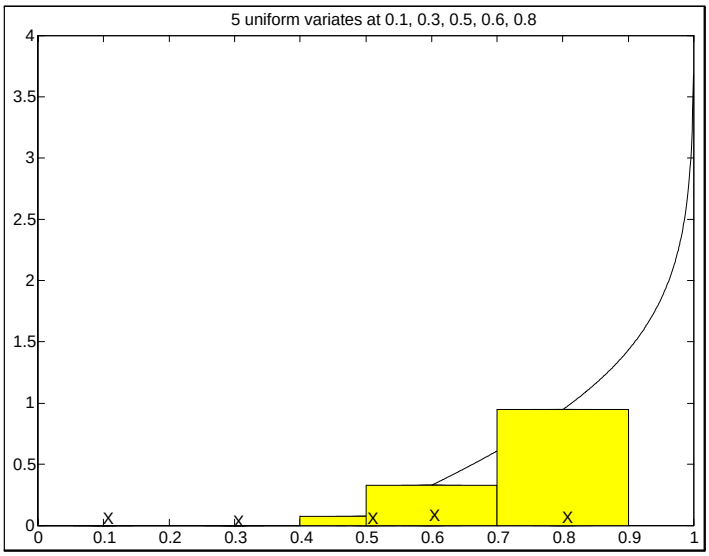


Figure 3.2:

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

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

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

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

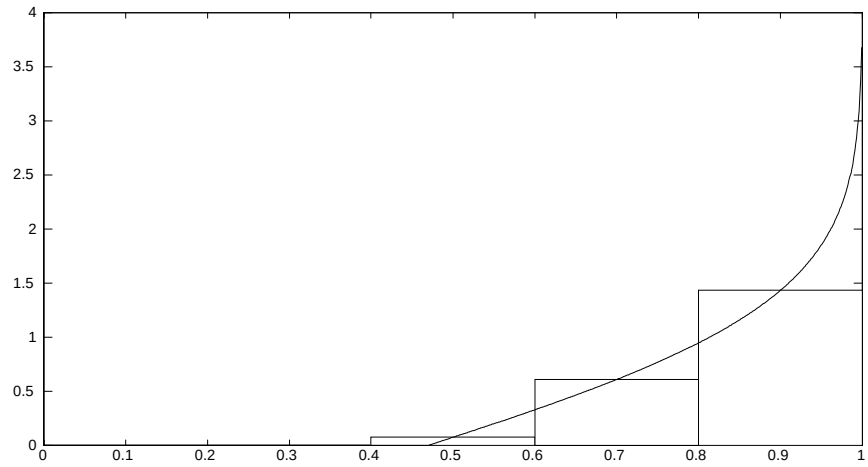


Figure 3.3:

"estimator" of the integral is

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

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

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

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

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

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

Using Antithetic Variates.

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

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

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

Stratified Sample.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

to obtain the result 0.46156. and the variance of the same vector is 0.00146. This should be compared with the crude Monte-Carlo estimator having the same number $n = 4$ of function evaluations as each of the components of the vector F : $\sigma^2_{crude}/4 = 0.4467/4 = .1117$. The gain in efficiency is therefore .1117/.0014

or approximately 80. The above stratified-antithetic simulation with 100,000 input variates and 400,000 function evaluations is equivalent to a crude Monte Carlo simulation with sample size 32 million! Variance reduction makes the difference between a simulation that is feasible on a laptop and one that would require a very long time on a mainframe computer..

Control Variates.

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

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

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

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

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

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

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

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

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

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

```
function g=GG(u)            %  the function g(u)
% control variate for callopt2.
u=max(0,u-.47);
g=6*u.^2+u;
function [est,var1,var2]=control(f,g,intg,n)
%[est,var1,var2]=control(f,g,intg,n)
%runs a simulation on the function f using control variate g (both character strings) n times.
% intg is the integral of g                % intg=∫_0^1 g(u)du
```

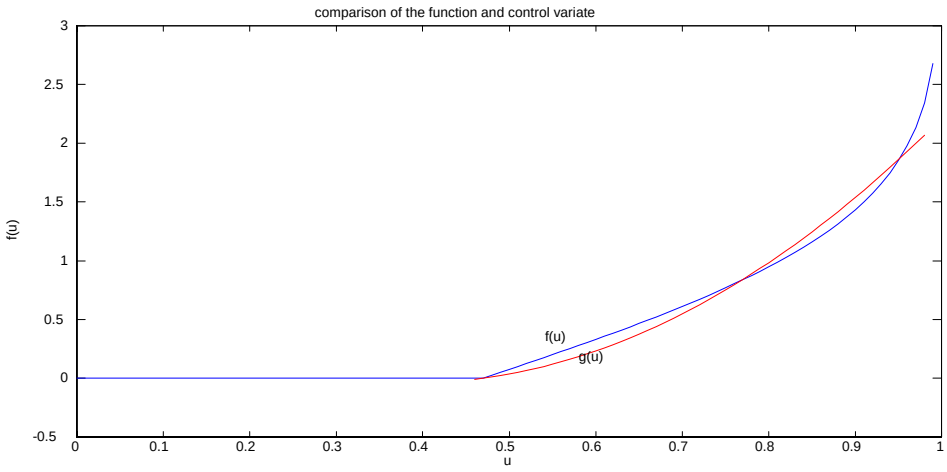


Figure 3.4:

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

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

and then the call *[est,var1,var2]=control('fn','GG',2*(.53)^3+(.53)^2/2,100000)* yielding the estimate 0.4602 and variances var1 = 0.4371, var2 = 0.0138 for an efficiency gain of around 32.

Importance Sampling.

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

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

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

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

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

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

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

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

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

Combining Monte Carlo Estimators.

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

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

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

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

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

Theorem 17 *(best linear combinations of estimators)*

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

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

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

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

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

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

Y_2 is an antithetic and stratified estimator, Y_3 a simpler version of a stratified-antithetic estimator, Y_4 is a control variate estimator and Y_5 the importance

sampling estimator all for a single input random variate. In order to determine the optimal linear combination we need to generate simulated values of all 5 estimators using the same uniform random numbers as inputs. We determine the best linear combination of these estimators using

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

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

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

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

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

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

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

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

Common Random Numbers.

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

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

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

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

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

Theorem 18 *(maximum/minimum covariance)*

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

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

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

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

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

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

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

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

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

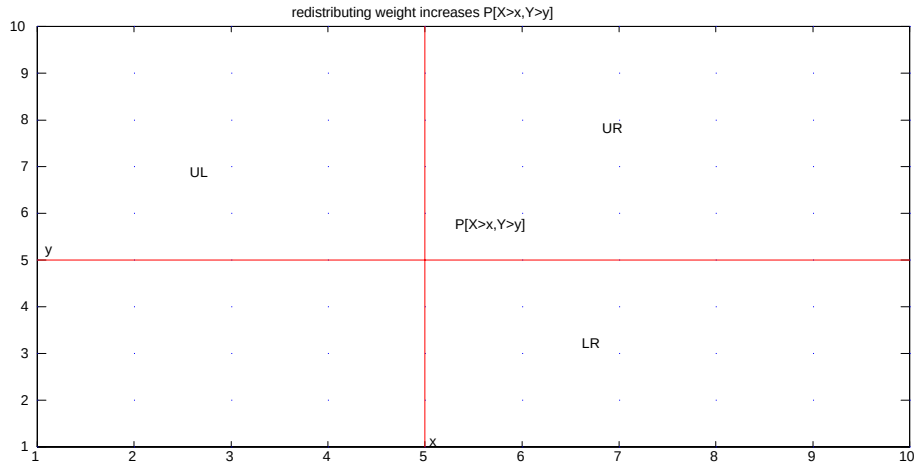


Figure 3.5:

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

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

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

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

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

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

3.3 Simulations from the Stationary Distribution of a Markov Chain.

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

It is easy to prove that coalescense occurs for sufficiently large M. For an ergodic finite Markov chain, there is some step size L such that every

transition has positive probability $P[X_{t+L} = j | X_t = i] > \epsilon$ for all i, j. Then the probability of coalescence in an interval of length L is at least $\varepsilon^{K+1} > 0$ and since there are infinitely many intervals disjoint of length L in $[-\infty, 0]$ and the event that there is a coalescense in each are independent, the probability that coalescense occurs somewhere in $[-\infty, 0]$ is 1..

We now detail the Propp Wilson algorithm

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

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

3.4 Coupling and Perfect Simulations.

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

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

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

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

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

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

FIGURE

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

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

(FINISH.....)

3.5 Some Multivariate Applications in Finance.

3.5.1 Asian Options.

Consider as an example a discretely sampled Asian call option on an asset with price process $S(t)$. An Asian option is like a European option but with the value function dependent not on the closing price of the underlying but on an average. An Asian call options pays an amount equal to $max(0, \bar{S}_k - K)$ where $\bar{S}_k = \frac{1}{k}\sum_{i=1}^{k} S(iT/k)$. Here k depends on the frequency of sampling (e.g. if $T = .25$ and sampling is weekly, then $k = 13$). If $S(t)$ follows a geometric Brownian motion, this is the sum of lognormally distributed random variables (rather than normally distributed ones) and as a result the distribution of the partial sums is very difficult to obtain. However, the distribution of the *geometric mean* is relatively simpler where the geometric mean of n values $x_1, ..., x_n$ is $(x_1 x_2 ... x_n)^{1/n}$. Our objective is to determine the value of the option $E(V_1) = E\{e^{-rT} max(0, \bar{S}_k - K)\}$. Since we expect geometric means to be close to arithmetic means, we may use as a control variate the random variable $V_2 = e^{-rT} max(0, \tilde{S}_k - K)$ where $\tilde{S}_k = \{\prod_{i=1}^{k} S(iT/k)\}^{1/k}$. Assume that V_1 and V_2 obtain from the same simulation and are therefore possibly correlated. Of course V_2 is only useful as a control variate if its expected value can be determined analytically or numerically more easily than V_1. Fortunately, in this case, we may use the relation between a geometric Brownian motion and the normal random walk to determine the distribution of the geometric mean. Since $S(t) = S_0 e^{Y(t)}$ where $Y(t)$ is a Brownian motion with drift $r - \sigma^2/2$

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

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

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

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

then

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

and

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

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

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

In fact

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

with

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

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

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

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

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

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

where

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

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

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

Table 4.1. Standard Errors for Arithmetic Average Asian Options.

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

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

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

For example we might confirm the last row of the above table using the command

asian(.1,100/1.1,.4,.2,100,50,10000);.

3.5.2 Girsanov's Lemma.

In the above, we implemented just one variance reduction scheme. There are many other possibilities. We expect the option to have a payoff closely related to the closing value of the stock $S(T)$. It might be reasonable to stratify the sample; i.e. sample more often when $S(T)$ is large, and there are several ways to implement this. One is to use importance sampling and generate $S(T)$ from a geometric Brownian motion with drift larger than rS_t so that it is more likely that $S(T) > K$. But if we do this we need to then multiply by the ratio of the

two probability density functions or the density of one process with respect to the other. This density is given by a result called Girsanov's lemma and a very simple form of this lemma appears below.

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

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

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

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

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

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

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

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

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

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

The ratio of these two probability density functions is

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

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

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

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

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

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

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

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

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

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

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

Call option with stochastic interest rates.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.6 Simulating Barrier and lookback options

Suppose we observe a stochastic process X_t over the interval $0 \le t \le T$. As is often done with financial time series we record the initial value or *open* of the time series $O = X_0$ the terminal value or *close* $C = X_T$, the maximum over the period or the *high* $H = max\{X_t;\ 0 \le t \le T\}$ and the minimum or the *low* $L = min\{X_t; 0 \le t \le T\}$. The recording of all four variables is common in practice but the use of all is rare. For example, the variance or volatility parameter is commonly estimated using only the open and close O, C although the information available for this parameter in the four random variables O, C, H, L is about seven times as great for Brownian motion. More commonly, in fact, volatility is determined as the "implied volatility" from the price of a derivative sold on the open market which has X_t as the underlying asset price. The implied volatility is the value of the volatility parameter which produces the market price of a given option (usually a heavily-traded or benchmark option). For example suppose a particular option with strike price K, maturity T, and initial value of the stock S_0 is traded on the market at a price given by V. Then we may solve the equation $BS(S_0, K, r, T, \sigma) = V$ for the implied volatility parameter σ. This estimate of volatiliy may differ

substantially from the historical volatility obtained in the Black-Scholes model by computing the sample variance of the returns $\log(S_{t+1}/S_t)$. Nevertheless since it agrees with the market price of the option it expected to more closely reflect the risk-neutral distribution Q and is therefore used. The disadvantage of this method of calibrating the parameter is that its value will depend on the strike price of the option, as well as the time and maturity parameters.

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

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

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

number of papers devoted to the valuation of such options. For details, see the references in Broadie, Glasserman and Kou (1996). We begin with the result on the distribution of highs and lows.

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

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

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

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

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

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

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

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

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

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

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

Therefore it is sufficient to prove the result for a Brownian motion process with T replaced by $\tau^{-1}(T)$. Assume without loss of generality that $Z_0 = 0$. Now let P, E denote probabilities and expectations in model (4.35) and P_0, E_0 be probabilities and expectations in the same model with zero drift i.e. $\nu = 0$. Assume without loss of generality that $Z_0 = 0$. Now a Brownian motion process can be considered as a limit of a sequence of simple random walks so the first

step is to verify a result for simple random walks, one in which the process jumps up or down with equal probability 1/2. Suppose that Figure 3.6 is a rescaled sample path from such a process. Consider two values z and s both possible values for the process. Notice that for each sample path ending at s which passes above a barrier at the point z there is a corresponding path ending at $2z - s$ obtained by reflecting the original path from the first time it crosses the barrier at z. In fact there is a one-one correspondence between such paths. It follows that there is the same number of paths ending at $s < z$ and with maximum $\geq z$ as paths ending at the reflection of s,namely $2z - s$. Since for a simple random walk all paths have the same probability, it follows that for a simple random walk

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

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

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

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

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

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

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

The results for the low follow by symmetry. ■

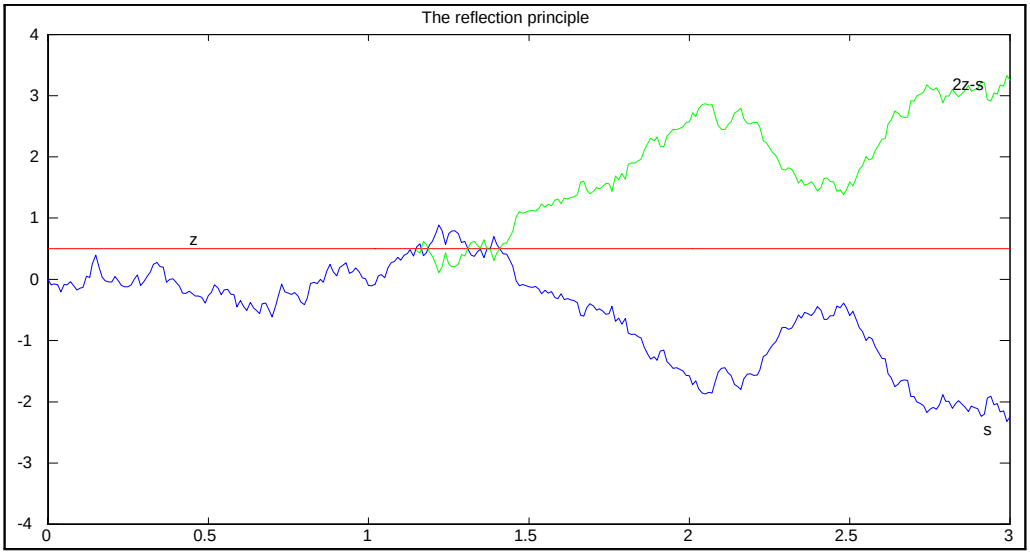


Figure 3.6:

Corollary 22 *For a Brownian motion process,*

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

Corollary 23 *For a Geometric Brownian motion process,*

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

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

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

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

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

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

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

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

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

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

Note that by symmetry,

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

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

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

$$\frac{H - O}{2H - O - C} \tag{3.42}$$

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

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

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

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

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

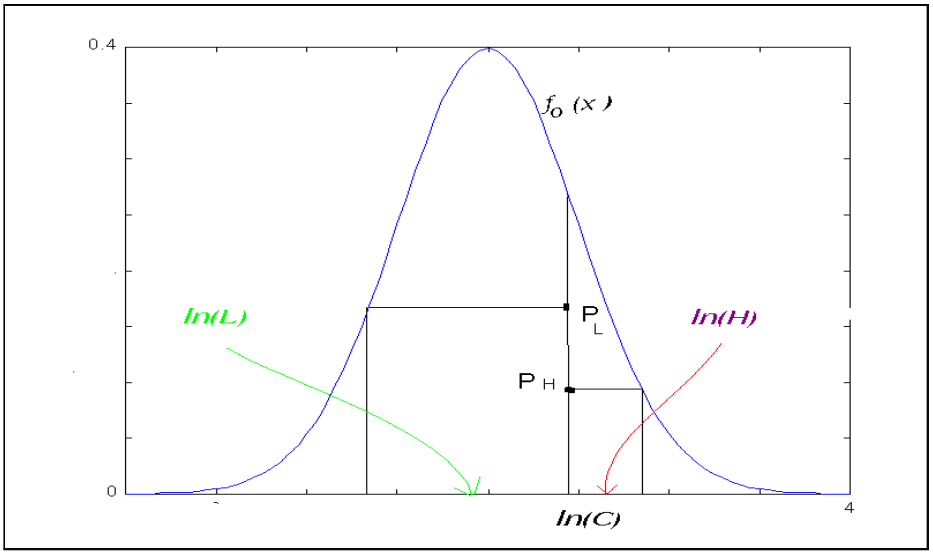


Figure 3.7:

3.6.1 One Barrier

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

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

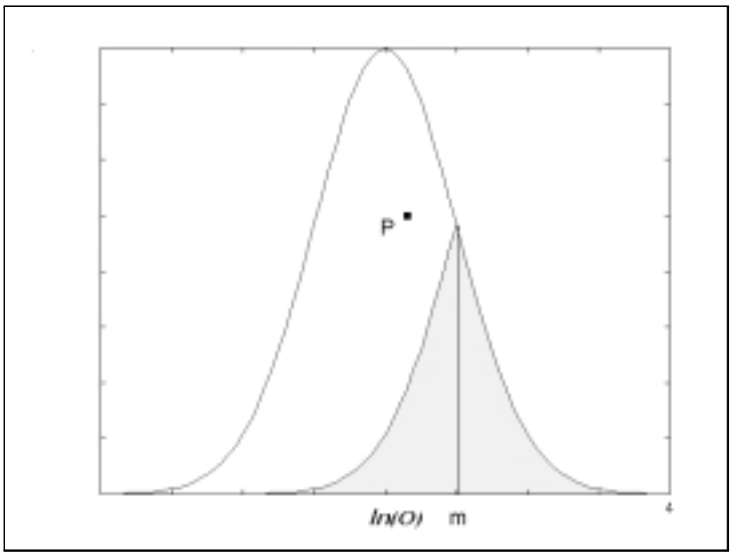


Figure 3.8:

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

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

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

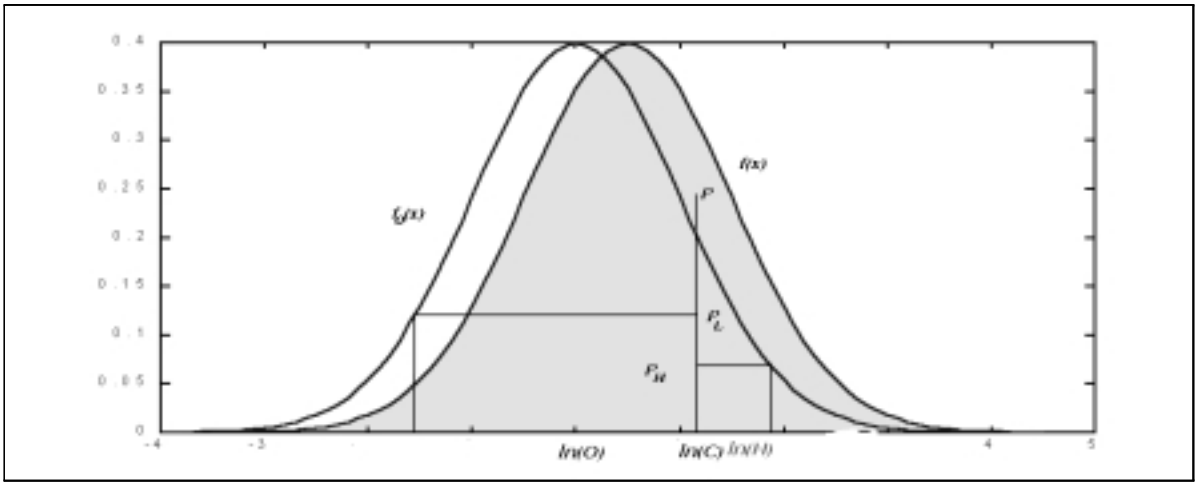


Figure 3.9:

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

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

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

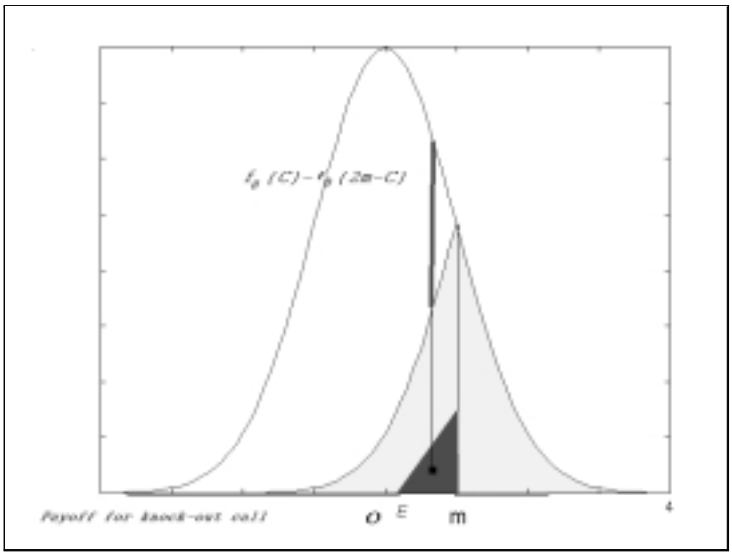


Figure 3.10:

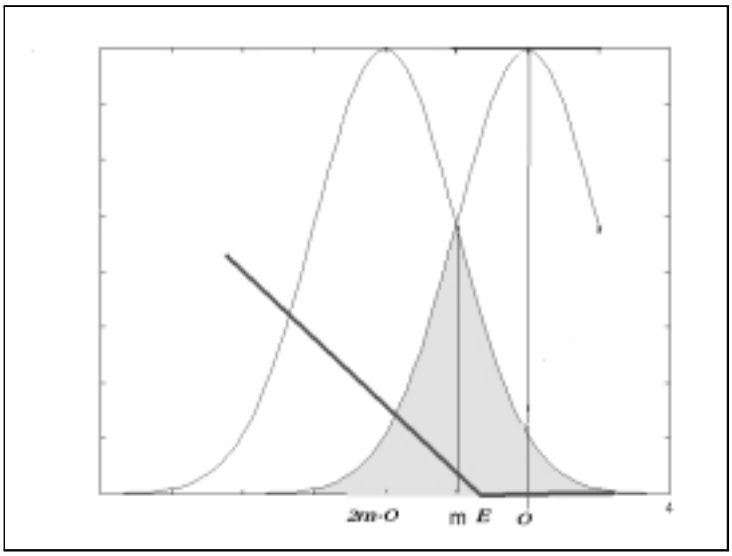


Figure 3.11:

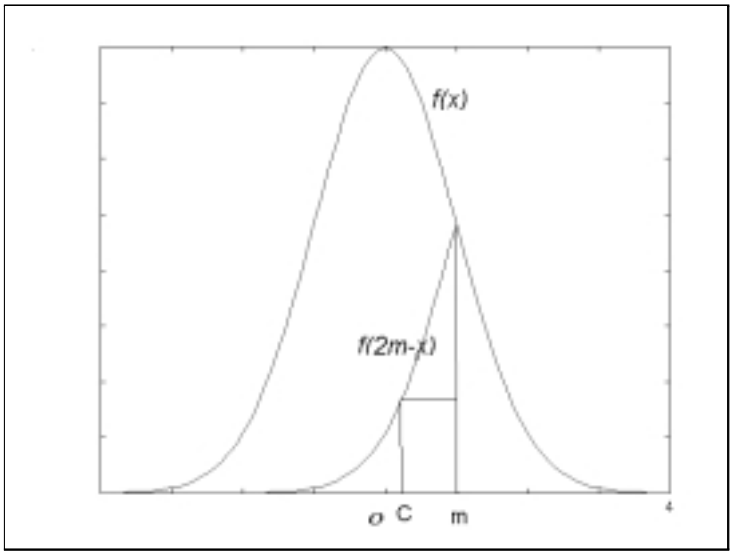


Figure 3.12:

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

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

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

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

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

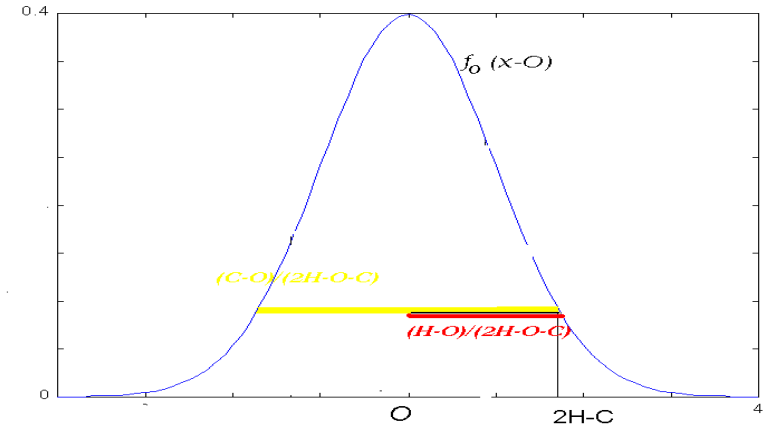
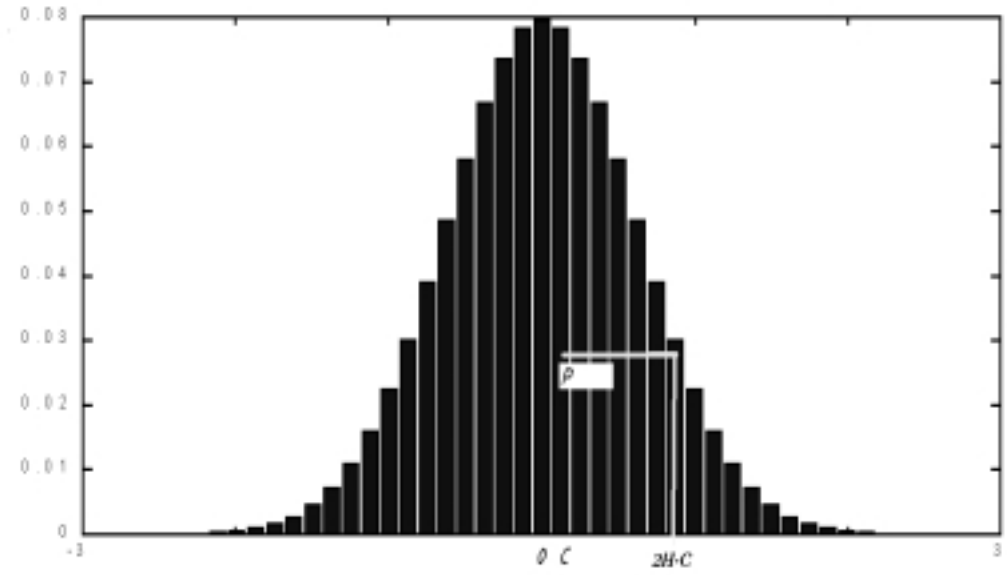


Figure 3.13:



Once again consider a point chosen uniformly and at random under this histogram. The closest point on the lattice to the x-coordinate of this point can be considered a generated value of the close. Moreover, if we run a horizontal line to the right, then the last bar, say y, passed through before it leaves the histogram corresponds, with one small adjustment, to the generated value of $2H - C$ as in the continuous case. The adjustment stems from the fact that the

difference between this point and the close is $2(H-C)$ which is an even multiple of the size of a jump. Thus, if we define $H = \Delta[\frac{y-C}{2\Delta}]$, where the square brackets indicate "the integer part of ", then H is a simulated value of the high.

3.6.2 One Factor, Two barriers

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

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

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

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

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

$-a < u < b$,

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

Note that for $n > 0$,

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

and for $n = -m < 0$,

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

With these substitutions,

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

since

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

■

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

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

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

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

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

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

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

provides an unbiased estimator of $E[V(C,H,L,O)|C]$ for each C. These random weights may be replaced by an average of a number of such randomly generated weights for each value of C. The weights can be negative- for example $(1 - W_a^+(C) - W_b^+(C))$ can equal -1. These values are adjustment for a small degree of overcounting that occurs when both barriers are crossed. We have

found that this scheme is particularly efficient if H and L are generated using antithetic uniform random numbers; for example solving for $H > \max(C, O)$ the equation

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

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

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

and similarly

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

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

3.7 Problems

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

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

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

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

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

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

5. Use a stratified random sample to integrate the function

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

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

1. Use a combination of stratified random sampling and an antithetic random number in the form

$$\frac{1}{2}[f(U/2) + f(1 - U/2)]$$

to integrate the function

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

What is the efficiency gain?

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

3. In the case $f(x) = \frac{e^x - 1}{e-1}$, consider using $g(x) = x$ as a control variate to integrate over [0,1]. Note that regression of $f(U)$ on $g(U)$ yields $f(U) - E(f(U)) = \beta[g(U) - Eg(U)] + \varepsilon$ where the error term ε has mean 0 and is uncorrelated with $g(U)$ and $\beta = cov(f(U), g(U))/var(g(U)$. Therefore, taking expectations on both sides and reorganising the terms, $E(f(U)) = f(U) - \beta[g(U) - E(g(U))]$. The Monte-Carlo estimator

$$\frac{1}{n}\sum_{i=1}^{n}\{f(U_i) - \beta[g(U_i) - E(g(U_i))]\}$$

is an improved control variate estimator, equivalent to the one discussed above in the case $\beta = 1$. Determine how much better this estimator is than the basic contol variate case $\beta = 1$ by performing simulations. Show that the variance is reduced by a factor of approximately 60. Is there much additional improvement if we use a more general quadratic function of x?

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

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

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

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

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

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

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

Chapter 4

Quasi- Monte Carlo Multiple Integration

4.1 Introduction

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

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

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

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

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

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

4.2 Errors in numerical Integration

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.1 Low discrepancy sequences

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

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

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

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

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

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

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

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

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

4.2.2 Definition: Measures of Discrepancy.

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

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

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

The *star discrepancy*:

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

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

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

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

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

4.2.3 Definition: Total Variation

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

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

where

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

We have the following inequality:

4.2.4 Theorem: Koksma - Hlawka inequality

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

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

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

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

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

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

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

4.2.5 Examples of low discrepancy sequences

In higher dimensions there exist several constructions:

4.2.6 Halton Sequences.

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

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

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

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

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

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

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

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

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

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

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

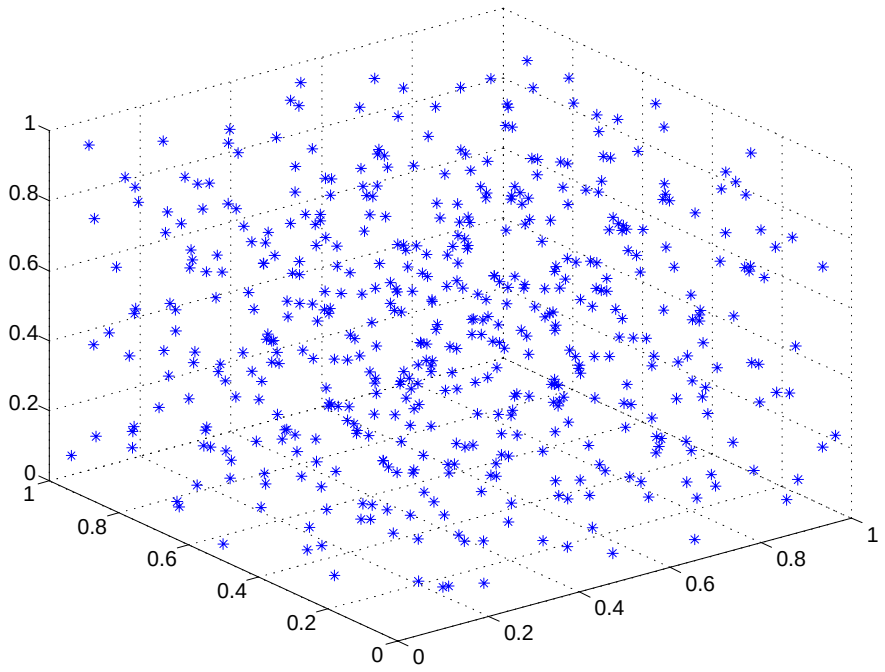


Figure 4.1:

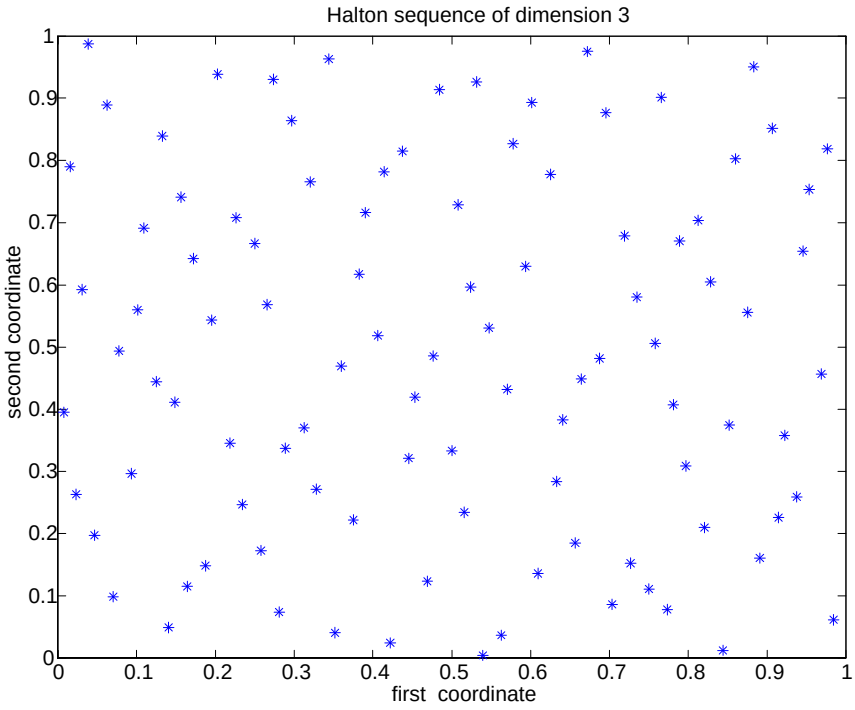

Halton sequence of dimension 3
second coordinate
first coordinate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4.2:

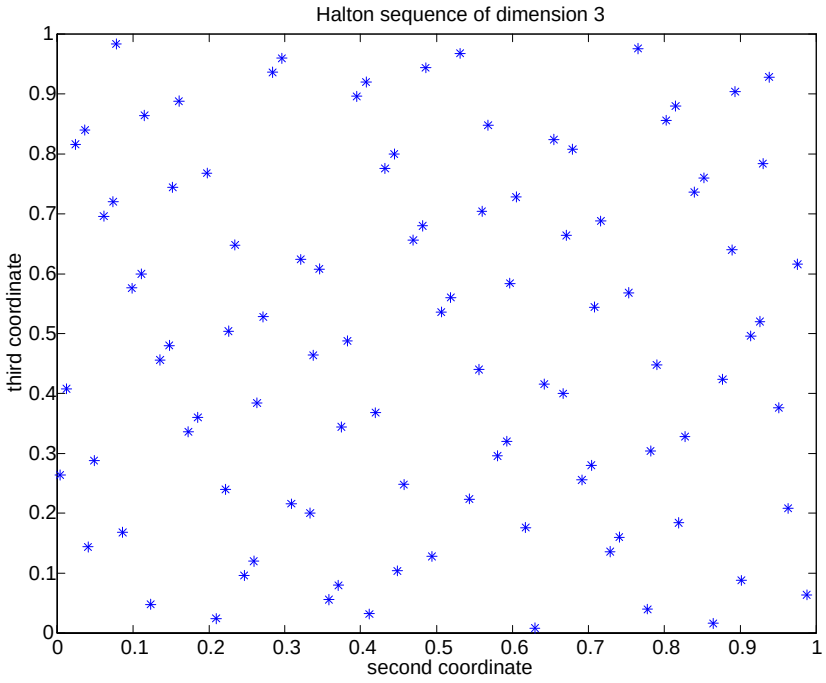

Halton sequence of dimension 3
third coordinate
second coordinate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4.3:

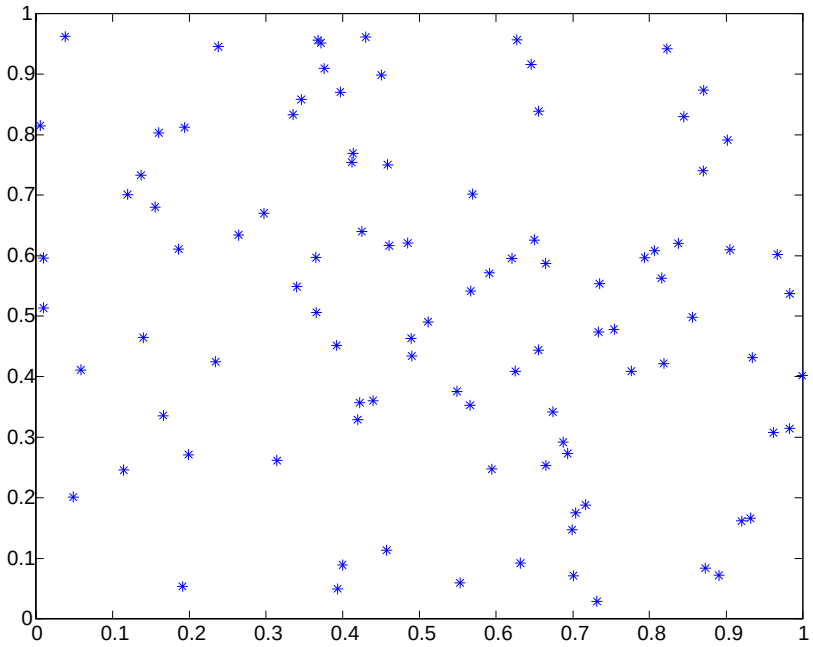


Figure 4.4: 100 independent $U[0,1]$ pairs

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

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

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

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

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

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

4.2.7 Sobol Sequence

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

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

is the recursion

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

where the addition is carried out using binary arithmetic. For the Sobol sequence, we then replace the binary digit a_k by $a_k v_k$.
The Sobel and Faure sequences are particular cases of $(t, s) - nets$. In order to define then we need the concept of an elementary interval.

4.2.8 Definition: elementary interval

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

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

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

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

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

4.2.10 Definition: (t, s) - sequence

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

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

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

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

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

Chapter 5

Sensitivity Analysis, Estimating Derivatives and the Greeks.

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

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

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

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

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

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

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

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

$$P(S,t) + x_S S + x_o V(S,t).$$

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

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

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

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

and the hedged portfolio has value

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

The availability of two instruments, the stock and a single option on the underlying S allow us to adjust a portfolio so that the first two derivatives of its value function with respect to S are both zero. The portfolio is therefore protected against reasonably small changes in S. Similarly, with more options on the same stock, one could arrange that the portfolio is immunized or protected against

adverse movements in the other parameters as well, including the interest rate and the volatility parameter. This hedged portfolio clearly requires derivatives of the value function, and for more complicated models than the Black-Scholes, we require simulation methods not only for valuing options and portfolios, but also for determining these derivatives with respect to underlying parameters.

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

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

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

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

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

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

5.1 Estimating Derivatives.

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

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

but this crude estimator suffers from a number of disadvantages;

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

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

5.1.1 The Score Function Estimator.

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

called the *Cramér conditions* we may interchange the integral and derivative (for example as required by the Cramer-Rao inequality)

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

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

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

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

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

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

Example. A Monte-Carlo Estimator of rho.

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

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

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

This score function estimator can be expressed as a limit of likelihood ratio estimators However, the score function is more stable than is the likelihood ratio for large sample size because its moment behaviour is, unlike that of the

likelihood ratio, similar to its behaviour in probability. Under the standard regularity conditions referred to above, the score function $S(\theta) = S_n(\theta)$ for an independent sample of size n satisfies a law of large numbers

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

and a central limit theorem;

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

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

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

Conditioning the Score Function Estimator.

Note that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.1.2 Example.

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

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

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

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

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

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

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

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

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

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

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

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

5.1.3 Example. Estimating Vega.

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

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

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

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

For brevity, write

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

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

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

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

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

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

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

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

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

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

5.1.4 Gaussian Quadrature.

We consider general integrals of the form

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2 Infinitesimal Perturbation Analysis: Pathwise differentiation.

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

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

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

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

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

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

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

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

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

1. write the expected value we wish to determine in terms of the parameters (as explicit arguments) and random variables whose distribution does not depend on these parameters (e.g. $U[0,1]$ or $N(0,1)$.) The simplest way to do this may be to use the inverse transform.

2. Differentiate this expected value with respect to the parameter of interest, passing the derivative under the expected value sign.

3. Simulate or numerically determine this expected value.

5.2.1 Example. IPA estimate of Vega.

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

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

and S_T, the terminal value of the asset, has a lognormal distribution with mean $S_0 e^{rT}$ and volatility parameter $\sigma^2 T$. We wish to write S_T in terms of random variables with distributions that do not depend on the parameters. Recall that

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

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

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

This can be simulated by generating values of Z and then $S_T = S_0 exp\{rT - \sigma^2 T/2 + \sigma\sqrt{T}Z\}$ and averaging the values of $e^{-rT} v'(S_T) S_T(\sqrt{T}Z - \sigma T)$. Alternatively, since this is a one dimensional integral, we can integrate the function against the standard normal p.d.f.ϕ i.e.

$$e^{-rT} \int_{-\infty}^{\infty} v'(S_0 e^{rT - \sigma^2 T/2 + \sigma\sqrt{T}z}) S_0 e^{rT - \sigma^2 T/2 + \sigma\sqrt{T}z} (\sqrt{T}z - \sigma T)\phi(z) dz.$$

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

call option. The following function was used to compare the estimators and their standard errors.

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

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

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

we obtain

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

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

The Multivariate Case.

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

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

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

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

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

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

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

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

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

$$= E_\theta V(X_1, ...x_i, ...X_n, \theta)\frac{\partial}{\partial \theta}F_{i\theta}(x_i)dx_i|_{-\infty}^{\infty} - E_\theta \int V^{(i)}(X_1, ...x_i, ...X_n, \theta)\frac{\partial}{\partial \theta}F_{i\theta}(x_i)dx_i$$

$$= -E_\theta\{V^{(i)}(X,\theta)\frac{\partial F_{i\theta}(X_i)/\partial \theta}{f_{i\theta}(X_i)}\}.$$

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

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

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

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

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

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

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

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

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

Consider two stocks or asset prices with closing values $S_1(T)$ and $S_2(T)$ jointly lognormally distributed with volatility parameters $\sigma_1, \sigma_2,$ and correlation ρ. Of course all of the parameters governing this distribution are subject to change in the market, including the correlation ρ. We are interested in the price of a European call option on the spread in price between the two stocks, and in particular, the sensitivity of this price to changes in the correlation. Let the payoff function be

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

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

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

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

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

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

where 1_A is the indicator function of the set

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

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

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

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

5.3 Problems.

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

Chapter 6

Estimation and Calibration.

6.1 Using Historical Data for Diffusion Models.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.2 Estimating Volatility

While the Euler method permits estimation of parameters in the diffusion coefficient as well as those in the drift, there is no likelihood argument based on continuous time observations which allows estimation of diffusion coefficient parameters. This is because the infinite variation of a diffusion process in arbitrarily small time intervals for continuous time observations theoretically permit *exact* estimation of parameters in the diffusion coefficient with an arbitrarily short observed trajectory of the process. In other words information for estimating the diffusion coefficient obtains much more rapidly than for the drift, and in this respect the continuous time processes are quite different than their discrete analogues. Two diffusions (or even Brownian motion processes) with different diffusion coefficients are mutually singular so that we can theoretically determine from a single sample path the exact diffusion term. Practice is considerably different for several reasons. First, we never observe a process in continuous time. Second, processes like security prices, interest rates and exchange rates are only similar to certain diffusion processes when viewed over a longer time interval than a single day or week. Their local behaviour is very different; for example they evolve through a series of jumps of varying magnitudes. Third, there is information on any process for which derivatives are sold in the derivative market. The usual estimate of volatility is the "implied volatility"

or the variance parameter which would, in the Black-Scholes formula, make the theoretical derivative prices equal to their observed values. While this is not exactly the historical volatility, it (provided that the model holds) is identical to the instantaneous value of the volatility since the risk neutral measure has the same diffusion coefficient as does the Ito process we assume derives the asset price.

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

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

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

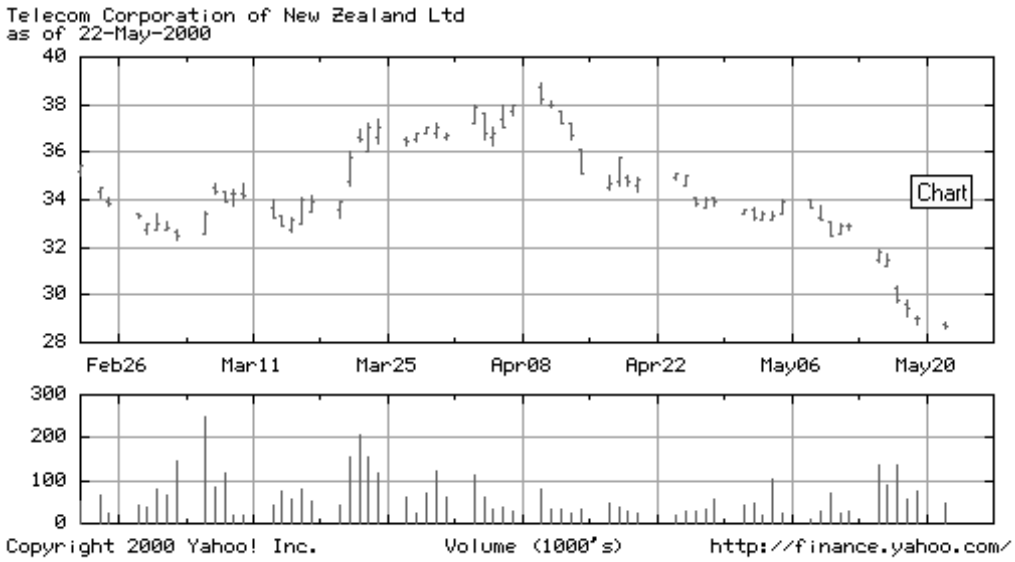
[CALL,PUT] = BLSPRICE(28.875,30,.06,42/252,σ,0).

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

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

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

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



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

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

Figure 6.1:

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

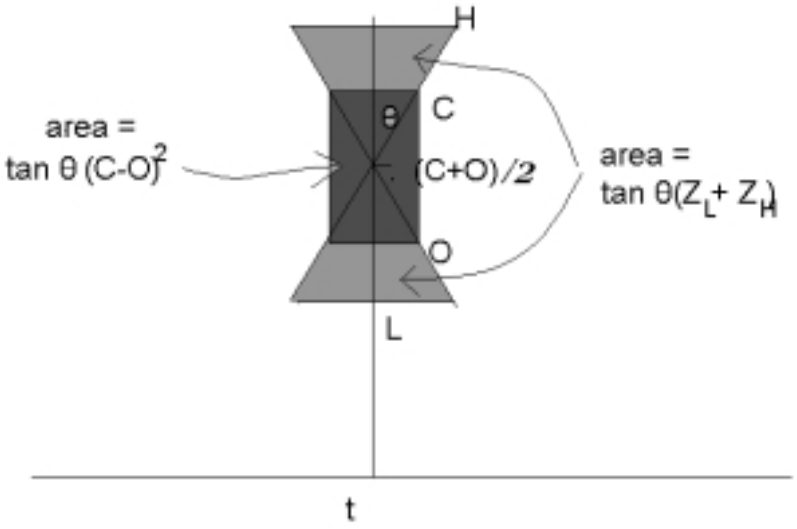
May 23,2000 @ 19:34 ET (Data 20 Minutes Delayed) **Bid** N/A **Ask** N/A **Size** N/AxN/A **Vol** 107200

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

Figure 6.2:

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

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



former.

These estimators are areas illustrated in Figure xx. Consider the plot corresponding to time t. The vertical scale is logarithmic so that logs are plotted. This plot is constructed using an arbitrarily chosen angle θ from the four values (O, C, H, L) using two lines $\ell_1,\ \ell_2$ through the point $(t, \frac{1}{2}(\log(O) + \log(C)))$ with slopes $\pm tan(\theta)$. Horizontal lines are drawn at the ordinate values $\log(H), \log(L), \log(O), \log(C)$ and using the points where $\log(O)$and $\log(C)$ strike the two lines as corners, a rectangle is constructed. The area of this rectangle $tan(\theta)(\log(C/O))^2$ is an unbiased estimator of $tan(\theta)\sigma^2\Delta t$ provided the Brownian motion has no drift. The second region consists of "wings" generated by the four points at which the horizontal line at $\log(H), \log(L)$ strike the lines $\ell_1,\ \ell_2$. The total area of this region (both wings) is $tan(\theta)(Z_L + Z_H)$ which is another unbiased estimator of $tan(\theta)\sigma^2 T$ independent of the first, and also independent of whether or not the underlying Brownian motion has drift. By comparing these areas, we can detect abnormal changes in the volatility, or changes in the drift of the process that will increase the observed value of $(\log(C/O))^2$ while leaving the second estimator unchanged. Because each estimator is based only on a single period, it is useful to provide as well a plot indicating whether there is a persistent change in either or both of the two estimators of volatility.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

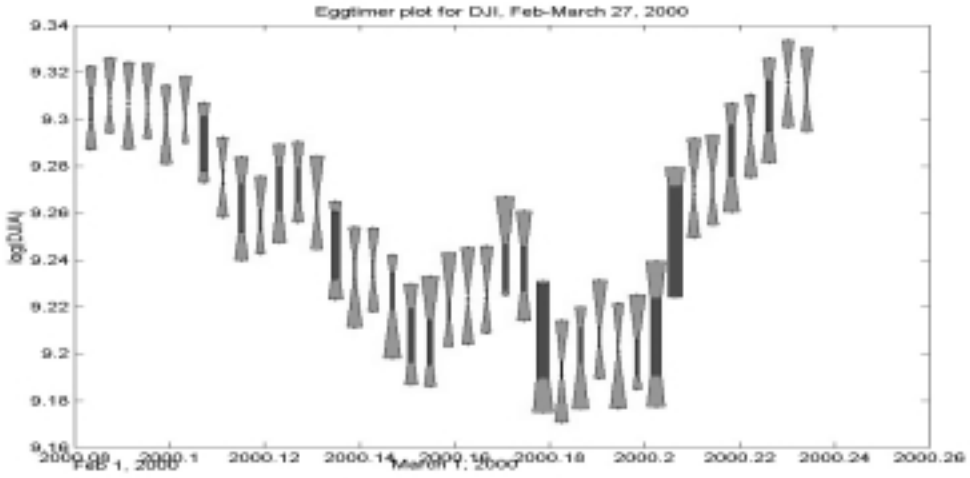


Figure 6.3:

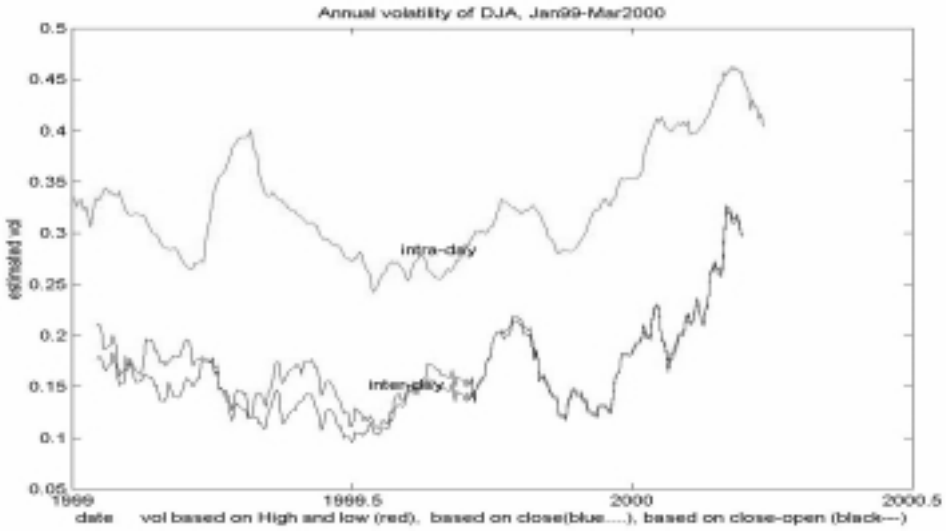


Figure 6.4:

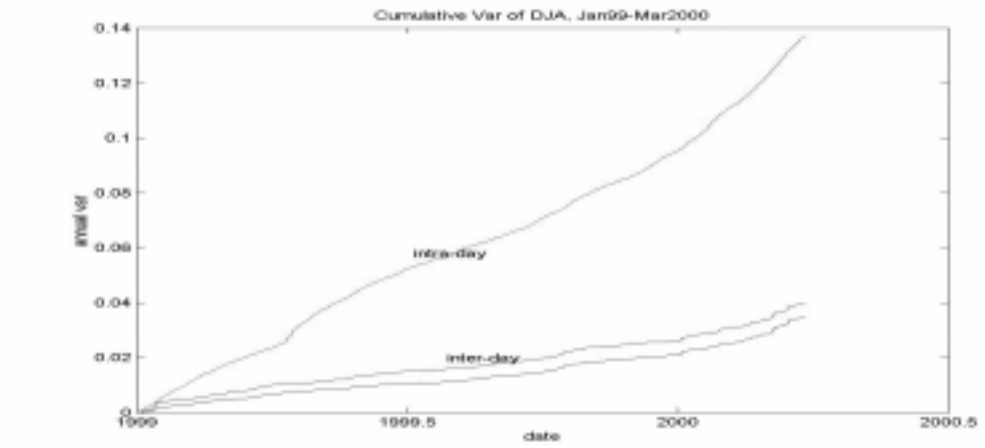


Figure 6.5:

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

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

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

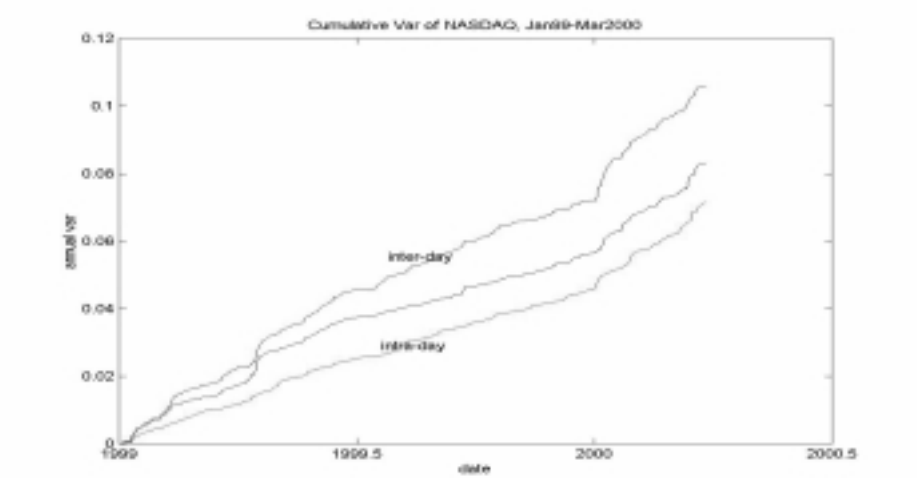


Figure 6.6:

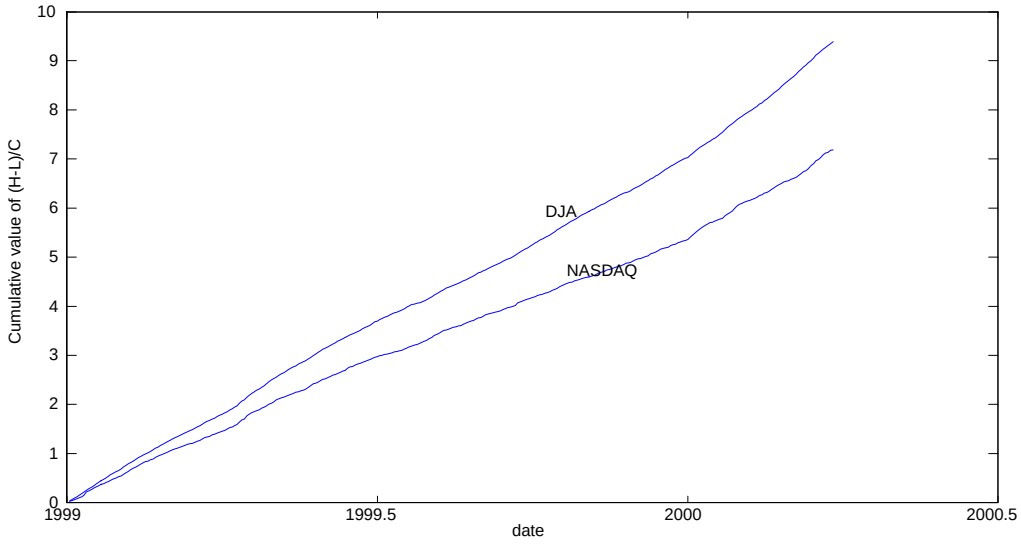


Figure 6.7:

6.3 Estimating Hedge ratios and Correlation Coefficients.

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

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

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

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

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

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

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

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

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

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

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

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

$$(C_2 - O_2 - \mu_2 T)[\ C_1 - O_1 - \mu_1 T - \frac{\rho\sigma_1}{\sigma_2}(C_2 - O_2 - \mu_2 T)] = 0$$

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

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

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

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

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

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

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

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

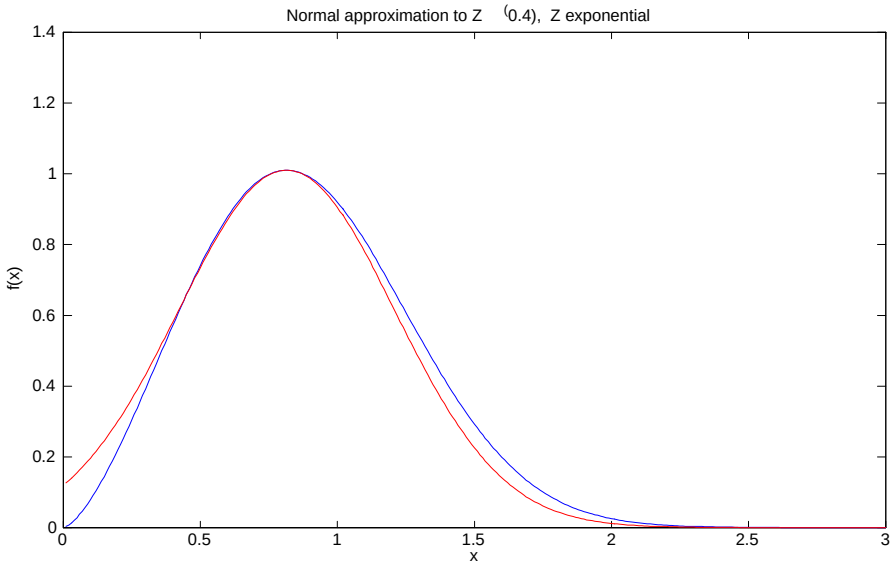
and the information in the optimal linear combination is given by the sum of the two informations

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

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

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

Normal approximation to $Z^{0.4}$



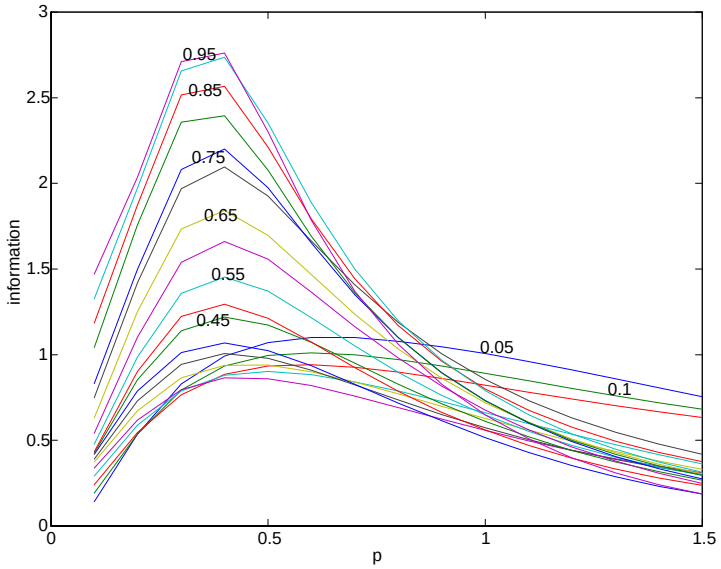
Standardization is achieved using the mean and variance

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

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

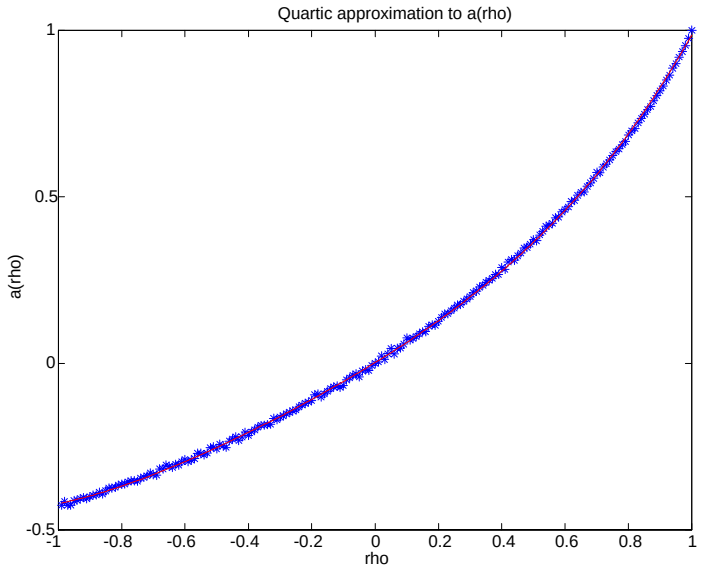
We begin by choosing a value of p which results in approximate maximization of the information in (**??**) . Since $a(\rho)$ is unknown, evaluating its derivative in the numerator of (**??**) must be done using either simulation or numerical

methods. Our approximation leads to the following plot of the information for versus p for various values of ρ.



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

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



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

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

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

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

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

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

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

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

where

$$B(\rho) = \begin{bmatrix} a(\rho) & a(-\rho) \\ a(-\rho) & a(\rho) \end{bmatrix},$$
$$\text{and } A = \begin{bmatrix} 1 & a(-1) \\ a(-1) & 1 \end{bmatrix} = B(1).$$

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

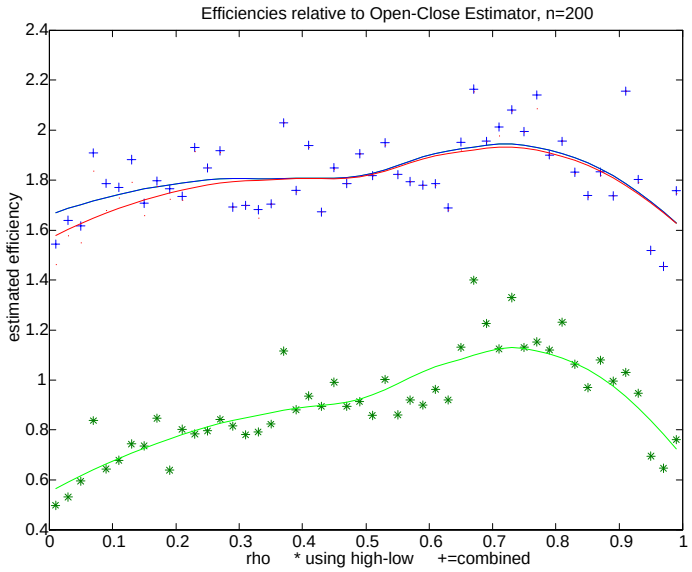
$$var\{g(Z_{H1})g(Z_{H2}) + g(Z_{L1})g(Z_{L2})\} = 2[1 + a^2(-1) + a^2(\rho) + a^2(-\rho)]$$

and (**??**) becomes

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

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

FIGURE XX



6.4 Estimation using the Term Structure of Interest rates

Largely for accounting reasons, and for the avoidance of arbitrage, financial analysts commonly estimate parameters of a model not from historical data but from some derivative. For example, the volatility parameter of an equity is often derived from the Black-Scholes price of options on the equity and the parameters of a diffusion model for interest rates from the term structure of interest rates. In general, of course, the decision whether to use implied values for parameters of efficient statistical estimators based on historical data is analogous to the choice between the real-world probability measure and the risk-neutral one. If one is interested in valuing derivatives employing a no-arbritrage principle, then the use of the risk-neutral measure is required. On the other hand, if one wishes to model the real-world behaviour of a given process, often statistical estimators, though complicated by the fact that parmaeters may change dynamically over time, provide a better fit to observed data and may be more useful in predicting the future.

It is common, for example, to assume a diffusion model for interest rates that permits time-varying coefficient;

$$dr_t = a(r_t, t)dt + \sigma(r_t, t)dW_t.$$

Consider a 0-coupon bond which, if invested today at time t returns 1\$ at time T. Then, if the current short rate is r_t, the value of this bond can be

written as a function

$$f(r_t, t) = E^Q[exp\{-\int_t^T r_s ds\}]$$

where E^Q denotes expectation under the risk-neutral measure. The *yield curve* describes the current expectations for average interest rates;

$$Yield(T-t) = -\frac{log(f(r_t, t))}{T-t}$$

For a given diffusion model, the function f can be determined by solving the PDE

$$\frac{\partial^2}{\partial t^2} f + a(x,t)\frac{\partial^2}{\partial t \partial x} f + \frac{1}{2}\sigma^2(x,t) f - xf = 0$$

subject to the boundary condition $f(x,T) = 1, \quad all \quad x \in \Re$. The more common models such as the Vasicek, the CIR and the Merton models for interest rate structure are such that the yield curve is *affine* or a linear function of the interest rate. In this case $f(x,t) = exp\{c(T-t)+d(T-t)x\}$ for some functions $c(.),\ d(.)$. Generally this linearity occurs provided that both the drift term and the square of the diffusion coefficient $\sigma^2(x,t)$ are linear in x.

One of the most popular current approaches to interest rates is the *HJM* or Heath-Jarrow-Morton model, which consists of modeling the *instantaneous forward rate* $f(t,u), \quad u \geq t$. Essentially, this function is assumed to follow a multidimensional diffusion. For details, see Chapter 8 of Duffie (1996).

Chapter 7

Miscellany

There are many othe models proposed for financial data that find support in some communities, and the debate about which are appropriate shows no sign of early resolution. The field of *artificial intelligence* offers *Neural Nets*, a locally simple model originally suggested as a design for the brain.

7.1 Neural Nets

A basic premise of much of modern research is that many otherwise extremely complex phenomena are much simpler when viewed locally. On this local scale, structures and organisation are substantially simpler. Complex societies of insects, for example, are organized with very simple interactions. Even differential equations like

$$\frac{dy}{dx} = ay$$

are used to describe the simple local structure of the more complicated exponential function.

Neural Nets are suggested as devices for processing information as it passes through a network. For example binary bits b_1, b_2, b_3 entering a given node j are processed with a very simple processor $g_j(b_1, b_2, b_3)$ which outputs a bit and then transmits it to another node. Thus a neural net consists of a description of the processors (usually simple functions weighted averages), an architecture describing the routing, and a procedure for estimating the parameters (for example the weights in the weighted average). They have the advantage of generality and flexibility- they can probably be modified to handle nearly any problem with some success. However, in specific models for which there are statistically motivated alternatives, they do not usually perform as well as a method designed for that model. Nevertheless, their generality makes them a popular research topic in finance.

7.2 Chaos, Long term dependence and non-linear Dynamics

Another topic, popularized in finance by books by Peters (...) and Gliek (1987), is *chaos*. Chaotic systems are generally purely deterministic systems that may resemble random or stochastic ones. For example if we define a sequence by a recursion of the form $x_t = f(x_{t-1})$ for some non-linear function f, the resulting system may have many of the apparent properties of a random sequence. Depending on the nature of the function f, the seqeunce may or may not appear "chaotic". Compare for example the bahaviour of the above recursion when $f(x) = ax(1-x), \quad 0 < x < 1, \quad , a < 4$ and a is small or a is near 4.

Similarly, the recursion

$$x_t = 1 - ax_{t-1}^2 + bx_{t-2}, \quad a = 1.4, \quad b = 0.3$$

describes a *bivariate* chaotic system, which, like an autoregressive process of order 2, requires two predecessors to define the current value. In general, a system might define x_t as a non-linear function of n predecessors. Then detecting chaos(or lack therof) is equivalent to determining whether the sequences $(x_t, x_{t+1}, \ldots x_{t+n}), \ t = 1, 2,$ fill $n+1$ dimensional space.

Tests designed to test whether a given sequence of stock returns are independent identically distributed generally result in rejecting this hypothesis but the most plausible explanation of this is not so clear. For example Hsieh (1991) tests for both chaotic behaviour and for arch-garch effects (predictable variance changes) and concludes that the latter is the most likely cause of apparent dependence in the data.

7.3 ARCH AND GARCH

One of the first noticable failures in the application of time series models to financial data such as a security price is the failure to adequately represent extended observed periods of high and low volatility. The innovations are supposed in the conventional ARMA models to be independent with 0 mean and constant variance σ^2 and the squared innovations should therefore be approximately independent (uncorrelated) variates.

The time series models discussed so far basically model the expected value of the series given the past observations, assuming that the conditional variance is constant. GARCH, or *Generalized Autoregressive Conditional Heteroscedasticity* takes this one step further, allowing this conditional variance to also be modeled by a time series. In particular, suppose that the innovations in an ARMA model are normally distributed given the past

$$a_t \sim N(0, h_t)$$

where the conditional variance h_t satisfies some ARMA relationship with

the squared innovations posing as the new innovations process.

$$\beta(B)h_t \;=\; \alpha_0 + \alpha(B)a_t^2$$

where $\beta(B) \;=\; 1 - \beta_1 B - \ldots \beta_r B^r$ and $\alpha(B) \;=\; \alpha_1 B + \ldots + \alpha_s B^s$.

The case $r \;=\; 0$ is the original ARCH *Autoregressive Conditional Heteroscedasticity* model, and the most common model takes $r \;=\; 0, s \;=\; 1$ so $h_t \;=\; \alpha_0 + \alpha_1 a_{t-1}^2$. For ARCH and GARCH models the parameters must be estimated using both the models for the conditional mean and the conditional variance and diagnostics apply to both models. The advantages of these models are that they provide both for some dependence among the observations through volatility rather than through the mean, and that they tend to have heavier tails. As a result, they provide larger estimated prices for deep out-of-the-money options, for example, which are heavily dependent on an accurate model for volatility.

7.3.1 ARCH(1)

The basic model investigated by Engle was the simplest case in which the process has zero conditional mean (it is reasonable to expect that the market has removed most or all of this) and but that the squares are significantly autocorrelated. Much financial data exibits this property to some degree. Engles ARCH(1) model is: $x_t \sim N(0, h_t)$ and

$$h_t \;=\; \alpha_0 + \alpha_1 x_{t-1}^2$$

whereas an ARCH regression model allows the conditional mean of x_t in (7.4) to depend on some observed predictors. The GARCH-IN-MEAN process fit by French et. al. allow the mean of x_t to be a function of its variance so that $x_t \sim N(a + bh_t^{p/2}, h_t)$. This would allow testing the hypotheses of relative risk aversion, for example. However, there is little evidence that b may be nonzero, and even less evidence to determine whether the linear relation should be between mean and standard deviation ($p = 1$) or between mean and variance ($p = 2$).

7.3.2 Estimating Parameters

The conditional log likelihood to be maximized with respect to the parameters $\alpha_i,\ \beta_j$ is:

$$log(L) \;=\; -\frac{1}{2}\sum_t [log\ h_t + \frac{\hat{a}_t^2}{h_t}]$$

Various modifications of the above GARCH model are possible and have been tried, but the spirit of the models as well as most of the methodology remains basically the same. There is also a system of Yule-Walker equations that can be

solved for the coefficients β_i in an ARCH model. If γ_i is the autocovariance function of the *innovations squared* a_i^2, then

$$\gamma_n = \sum_{i=1}^{s} \alpha_i \gamma_{n-i} + \sum_{i=1}^{r} \beta_i \gamma_{n-i}$$

for $n \geq r+1$. These provide the usual PAF for identification of the suitable order r of the autoregressive part.

7.3.3 Akaike's Information Criterion

Clearly, a model which leads to small estimated variances for the innovations is preferred, all else being equal, to one with large residual variation. In other words we are inclinded to minimize the estimated residual variance $\frac{1}{N-k}\sum \hat{a}_i^2$ (or equivalently its logarithm) in the selection of the model, where k is the number of autoregressive+moving average parameters in the model. However, such a criterion would encourage the addition of parameters for even a marginal improvement in residual variance, and so a better criterion penalizes against an increase in the number of parameters.

$$AIC = log[\frac{1}{N-k}\sum \hat{a}_i^2] + \frac{2k}{N}$$

The AIC criterion chooses that model which minimizes this quantity. It should be noted that the AIC put out by S is $-2log(L) + 2\times k$ and this is approximately N times the above value. The advantge in multiplying by N is that differences operate on a more natural scale. When nested models are compared (i.e. one model is a special case of the other), differences between values of the statistic $-2log(L)$ have a distribution which is Chi-squared with degrees of freedom the difference in the number of parameters in the two models under the null hypothesis that the simpler model holds.

7.3.4 Testing for ARCH effects

Most of the tests for the adequacy of a given time series model are inherited from regression, although in some cases the autocorrellation of the series induces a different limiting distribution. For example, if there is an ARCH effect, then there should be a significant regression of $\hat{a}_t^2$ on its predecessors $\hat{a}_{t-1}^2$, $\hat{a}_{t-2}^2$, $\hat{a}_{t-3}^2 \ldots$. Suppose we are able to obtain residuals $\hat{a}_l$, $\hat{a}_{l+1}, \ldots \hat{a}_N$ from an ARMA model for the original series. We test for ARCH effect by regressing the vector $(\hat{a}_{l+s}^2, \ldots \hat{a}_N^2)$ on a constant as well as the s "predictors"

$$(\hat{a}_{l+s-1}^2, \quad , \hat{a}_{N-1}^2),\ (\hat{a}_{l+s-2}^2, \quad , \hat{a}_{N-2}^2).\ \ldots (\hat{a}_l^2, \quad , \hat{a}_{N-s}^2)$$

and obtaining the usual coefficient of determination or *squared multiple correlation coefficient* R^2. Standardized, $(N-l)R^2$ has an approximate chi-squared

distribution with s degrees of freedom under the null hypothesis of homoscedasticity so values above the 95'th chi-squared percentile would lead to rejecting the homoscedasticity null hypothesis and concluding arch-like effects. The following table provides the approximate critical values for the chi-squared at the 5% level.

Chi-squared critical values
Conclude ARCH effects if $(N-l)R^2$ exceeds:

s	Critical Value
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

7.3.5 Example. Deutschmark Exchange

The original exchange rate series is given in Figure 7.1. There are approximately 4700 daily observations of the value of the Deutschmark priced in American dollars covering a period from May 16, 1972 to December 31, 1990. Fitting an

Figure 7.1: Deutschmark exchange rate
figure=dm.ps,height=3in,width=5in

AR model by AIC to the first differenced series of DM exchange rates led to an estimate $\hat{\sigma}^2 = 1.41141 \times 10^{-5}$ and an order AR(21) model with the following coefficients:

LAG	AR 21 coefficients	LAG	AR 21 coefficients
1	0.017195644	11	0.007974478
2	0.023371179	12	-0.021073932
3	0.017750192	13	0.026812630
4	-0.011688776	14	0.008615040
5	-0.002804700	15	0.037807230
6	0.023669761	16	-0.019937096
7	0.012745515	17	-0.022599909
8	0.040368907	18	-0.017522413
9	0.016654816	19	0.023648366
10	0.041129731	20	0.034871727

LAG	COEFF
21	-0.024458840

and the following values of the AIC as output from Splus corresponding to lags 1-37,

```
 14.3662109 14.4160156 13.6240234 14.1992188 15.7988281 17.7529297
 16.7841797 17.6533203 12.2275391 12.6582031  6.5322266  8.2451172
 8.5869141  6.9638672  8.5751953  4.5263672  4.5263672  4.3544922
 5.0292969  4.3935547  0.8105469  0.0000000  1.7773438  3.2275391
 4.7246094  5.5156250  5.4570312  7.4218750  9.3662109  6.0908203
 4.2158203  5.9472656  7.8769531  9.8632812 10.3320312 11.3339844
12.4189453
```

The minimum AIC has been subtracted from all values so it corresponds to the order 21 model, and after subtraction, gives tabulated value 0.0000000. These coefficients can be compared with twice the standard error of $1/N^{1/2}$ where $N = 4696$ and this is around 0.028. Since few of the coefficients exceed this value (except the coefficient for lag 10), there is little support for any non-trivial autoregressive effect. Fortunately, because the estimated coefficients are generally small as well, there will also be little difference in forecasting whether we use white noise or the suggested AR(21) model above, since the white noise model corresponds to putting all of the above (small) coefficients equal to 0.

However, if we save the residuals $\hat{a}_t$ from the above time series, and then study the series $\hat{a}_t^2$ we obtain the following:

```
    LAG  Autocorrelation Function  LAG  Autocorrelation Function

    [1,]  0.158922836               [6,]  0.046017088
    [2,]  0.101335056               [7,]  0.021516455
    [3,]  0.077508301               [8,]  0.009644603
    [4,]  0.008419966               [9,]  0.002976760
    [5,]  0.032856703               [10,]  0.105071306
                                    [11,] -0.030951777

  LAG  PARTIAL Autocorrelation  LAG  PARTIAL Autocorrelation

    [1,]  2.079889e-01          [6,]  5.511491e-02
    [2,]  1.328089e-01          [7,]  3.411203e-02
    [3,]  9.759974e-02          [8,]  2.005137e-02
    [4,]  3.005405e-02          [9,]  1.561776e-02
    [5,]  4.940604e-02          [10,]  1.002484e-01
                                [11,] -3.095178e-02
```

Since the AF seems to decay somehat more quickly than the PAF with 3 values significantly non-zero, we might try an ARCH(3) model to describe the process. The large coefficient at lag 10 gives some concern. Does it indicate some sort of biweekly seasonality that we may wish to remove (if it has any reasonable explanation)? If an explanation can be found, then the seasonality can be dealt with by using differences of the form $(1 - B^{10})$.

Diebold and Nerlove, in ARCH Models of Exchange Rate Fluctuations confirm the ARCH effect on the exchange rate for a number of different currencies. However, they observe substantially longer effects (e.g. at lag ≥ 12) although their data is weekly.

Chapter 8

Appendix A: Some Basic Theory of Probability

8.1 Probability Models.

Basic Definitions.

Probabilities are defined on sets or events, usually denoted with capital letters early in the alphabet such as A, B, C. These sets are subset of a *Sample Space or Probability Space* Ω, which one can think of as a space or set containing all possible outcomes of an experiment. We will say that an event $A \subset \Omega$ occurs if one of the outcomes in A (rather than one of the outcomes in Ω but outside of A) occurs. Not only should we be able to describe the probability of individual events, we should also be able to define probabilities of various combinations of them including

1. Union of sets or events $A \cup B =$ A or B (occurs whenever A occurs or B occurs or both A and B occur.)

2. Intersection of sets $A \cap B = A$ and B (occurs whenever A and B occur).

3. Complement : $A^c =$ not A (occurs when the outcome is not in A).

4. Set differences : $A \setminus B = A \cap B^c$ (occurs when A occurs but B does not)

5. Empty set : $\phi = \Omega^c$ (an impossible event-it never occurs since it contains no outcomes)

Recall *De Morgan's rules* of set theory: $(\cup_i A_i)^c = \cap_i A_i^c$ and $(\cap_i A_i)^c = \cup_i A_i^c$

Events are subsets of Ω . We will call $\mathcal{F}$ the class of all events (including ϕ and Ω).

Axioms of Probability

A probability measure is a set function $P : \mathcal{F} \rightarrow [0,1]$ such that

1. $P(\Omega) = 1$
2. If A_k is a disjoint sequence of events so $A_k \cap A_j = \phi, \quad k \neq j$, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Proposition

$P(\phi) = 0$.

Proposition

If $A_k, k = 1, \ldots N$ is a finite sequence of disjoint events so $A_k \cap A_j = \phi, \quad k \neq j$, then

$$P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i)$$

Proposition

$P(A^c) = 1 - P(A)$

Proposition

Suppose $A \subset B$. Then $P(A) \leq P(B)$.

Proposition

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proposition

(inclusion-exclusion) $P(\cup_k A_k) = \sum_k P(A_k) - \sum \sum_{i<j} P(A_i \cap A_j) + \sum \sum \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots$.

Proposition

$P(\cup_{i=1}^{\infty} A_i) \leq \sum_i P(A_i)$.

Proposition

Suppose $A_1 \subset A_2 \subset \ldots$. Then $P(\cup_{i=1}^{\infty} A_i) = lim_{i \rightarrow \infty} P(A_i)$.

Example.

A coin is tossed twice. List Ω and the class $\mathcal{F}$ of possible events. Define the probability of an event A to be

$$P(A) = \frac{number\ \ of\ \ points\ \ in\ \ A}{number\ \ of\ \ points\ \ in\ \ \Omega}$$

Would this be the correct definition of probability if we defined the sample space using the number of heads observed $\Omega = \{0, 1, 2\}$?

Counting Techniques

Permutations.

The number of ways of arranging n distinct objects in a row is $n! = n(n-1)\ldots 1$ and $0! = 1$. Define $n^{(r)} = n(n-1)\ldots(n-r+1)$ (called "n to r factors") for arbitrary n, and r a non-negative integer. Define $n^{(0)} = 1$.

Example.

How many distinct ways are there of rearranging the 15 letters

$$AAAAABBBBCCCDDE?$$

Example

There are ten students seated at a table of which 5 are Pure Math, and 5 are Impure Math. The organisers are concerned about (intellectual) disputes. How many arrangements are there such that no two pure math students sit together? If the students are seated at random, what is the probability no two pure math students are seated together?

Combinations

Suppose the order of selection is not considered to be important. We wish, for example, to distinguish only different *sets* selected, without regard to the order in which they were selected. Then the number of distinct sets of r objects that can be constructed from n distinct objects is

$$\binom{n}{r} = \frac{n^{(r)}}{r!}$$

Note this is well defined for r a non-negative integer for any real number n .

8.2 Independence and Conditional Probabilities.

Independent Events.

Two events $A,\ B$ are said to be *independent* if

$$P(A \cap B) = P(A)P(B) \tag{2.1}$$

Compare this definition with that of *mutually exclusive or disjoint* events A, B . Events A, B are mutually exclusive if $A \cap B = \phi$.

Independent experiments are often built from *Cartesian Products* of sample spaces. For example if Ω_1 and Ω_2 are two sample spaces, and $A_1 \subset \Omega_1,\ A_2 \subset \Omega_2$ then an experiment consisting of *both of the above* would have sample space the Cartesian product

$$\Omega_1 \times \Omega_2 \ = \ \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

and probabilities of events such as $A_1 \times A_2$ are easily defined, in this case as $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$. Verify in this case that an event entirely determined by the first experiment such as $A = A_1 \times \Omega_2$ is independent of one determined by the second $B = \Omega_1 \times A_2$.

Definition.

A fine or countably infinite set of events $A_1, A_2, \ldots ..$ are said to be mutually independent if

$$P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}) \tag{2.2}$$

for any $k \geq 2$ and $i_1 < i_2 < \ldots i_k$.

Properties.

1. A,B independent implies $A,\ B^c$ independent.
2. Any A_{i_j} can be replaced by $A_{i_j}^c$ in equation (2.2).

Why not simply require that every pair of events is independent?

Example:

Pairwise independence does not imply independence. Two fair coins are tossed. Let A = first coin is heads, B = second coin is heads, C = we obtain exactly one heads. Then A is independent of B and A is independent of C but $A,\ B,\ C$ are **not mutually independent**.

Example

Players A and B decide to play chess until one of them wins. The probability A wins a given game is .3, the probability B wins is .2 and the probability of a draw is .5. What is the probability A wins first?

Lim Sup of events

For a sequence of events $A_n, n = 1, 2, ...$ we define another event $[A_n \ i.o.] = \limsup A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$. Note that this is the set of all points x which lie in infinitely many of the events $A_1, A_2, ...$ The notation i.o. stands for "infinitely often".

Borel Cantelli Lemmas

Clearly if events are individually too small, then there little or no probability that their lim sup will occur, i.e. that they will occur infinitely often.

Lemma 1: For an arbitrary sequence of events A_n , if $\sum_n P(A_n) < \infty$ then $P[A_n \ i.o.] = 0$.

Lemma 2: For a sequence of *independent events* A_n , $\sum_n P(A_n) = \infty$ implies $P[A_n \ i.o] = 1$.

Conditional Probability.

Suppose we are interested in the probability of the event A but we are given some relevant information, namely that another related event B occurred. How do we revise the probabilities assigned to points of Ω in view of this information? If the information does not effect the relative probability of points in B then the new probabilities of points outside of B should be set to 0 and those within B simply rescaled to add to 1.

Definition: Conditional Probability:

For $B \in \mathcal{F}$ with $P(B) > 0$, define a new probability

$$Q(A) \ = \ P(A|B) \ = \ \frac{P(A \cap B)}{P(B)}. \tag{2.3}$$

This is also a probability measure on the same space $(\Omega, \mathcal{F})$, and satisfies the same properties. Note that $P(B|B) = 1, \ \ P(B^c|B) = 0.$

Theorem: Bayes Rule

If $P(\cup_n B_n) = 1$ for a *disjoint* finite or countable sequence of events B_n all with positive probability, then

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)} \tag{2.4}$$

Theorem: Multiplication rule.

If $A_1 \dots A_n$ are arbitrary events,

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 A_1) \dots P(A_n|A_1 A_2 \dots A_{n-1}) \tag{2.5}$$

Example. Diagnostic Testing.

Suppose a blood test for HIV tests positive for 95% of the people who are actually HIV positive and tests negative for 99% of the people who are HIV negative. Suppose the probability that a male is HIV positive is .0001 and the probability that a female is HIV positive is .00005. Assume equal proportions of males and females in the population.

1. Find the probability that a randomly selected person who tested positive on the diagnostic test is indeed HIV positive.
2. Find the probability that a randomly selected male who tested positive on the diagnostic test is indeed HIV positive.

Random Variables and Discrete Distributions

Random Variables

Properties of $\mathcal{F}$.

The class of events $\mathcal{F}$ (called a $\sigma - a\,\mathrm{lg}\,ebra$ or $\sigma - field$) should be such that the operations normally conducted on events, for example countable unions or intersections, or complements, keeps us within that class. In particular it is such that

(a) $Varphi \in \mathcal{F}$

(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(c) If $A_n \in \mathcal{F}$ for all $n = 1, 2, ...,$ then $\cup_{n=1}^{\infty} \in \mathcal{F}$.

It follows from these properties that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is also closed under countable intersections, or countable intersections of unions, etc.

Definition

Let X be a function from a probability space Ω into the real numbers. We say that the function is *measurable* (in which case we call it a random variable) if for $x \in \Re$, the set $\{\omega; X(\omega) \leq x\} \in \mathcal{F}$. Since events in $\mathcal{F}$ are those to which we can attach a probability, this permits us to obtain probabilities for the event that the random variable X is less than or equal to any number x.

Definition: Indicator random variables

For an arbitrary set $A \in \mathcal{F}$ define $I_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. This is called an *indicator random variable.* (sometimes a *characteristic function* in measure theory, but not here).

Definition: Simple Random variables.

Consider events $A_i \in \mathcal{F}$ such that $\cup_i A_i = \Omega$. Define $X(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega)$ where $c_i \in \Re$. Then X is measurable and is consequently a random variable. We normally assume that the sets A_i are disjoint. Because this is a random variable which can take only finitely many different values, then it is called *simple*. Any random variable taking only finitely many possible values can be written in this form.

Example.

A coin is tossed 10 times. X is the number of heads. Describe $(\Omega, \mathcal{F})$ and the function $X(\omega)$.

Notation

We will often denote the event $\{\omega \in \Omega; X(\omega) \leq x\}$ more compactly by $[X \leq x]$.

Theorem.

If X_1, X_2 are random variables, so is

1. $X_1 + X_2$
2. $X_1 X_2$
3. $min(X_1, X_2)$.

Cumulative Distribution Functions.

Definition.

The *cumulative distribution function* (c.d.f.) of a *Random variable* X is defined to be the function $F(x) = P[X \leq x],\ x \in \Re$.

Properties of C. D. F.

1. A c.d.f. $F(x)$ is non-decreasing. i.e. $F(x) \geq F(y)$ whenever $x \geq y$.
2. $F(x) \to 0$, as $x \to -\infty$.
3. $F(x) \to 1,\ x \to \infty$.
4. $F(x)$ is right continuous. i.e. $F(x) = lim\ F(x + h)$ as h decreases to 0.

There are two types of distributions that we consider in this course, discrete distributions and continuous ones. Discrete distributions are those whose cumulative distribution function at any point x can be expressed as a finite or countable sum of values. For example

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

for some probabilities p_i which sum to one. In this case the cumulative distribution is piecewise constant, with jumps at the values that the random variable can assume. The values of those jumps are the individual probabilities. For example $P[X = x]$ is equal to the size of the jump in the graph of the c.d.f. at the point x. We refer to the function $f(x) = P[X = x]$ as the *probability function* of the distribution.

Some Special Discrete Distributions

The Discrete Uniform Distribution

Many of the distributions considered so far are such that each point is equally likely. For example, suppose the random variable X takes each of the points $a, a+1, \ldots b$ with the same probability $\frac{1}{b-a+1}$. Then the c.d.f. is

$$F(x) = \frac{x - a + 1}{b - a + 1}, \quad x = a, a+1, \ldots b$$

and the probability function is $f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, \ldots b$ and 0 otherwise.

The Hypergeometric Distribution

Suppose we have a collection (the *population*) of N objects which can be classified into two groups S or F where there are r of the former and $N-r$ of the latter. Suppose we take a random sample of n items without replacement from the population. What is the probability that we obtain exactly x S's?

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \; x = 0, 1, \ldots$$

What is the possible range of values of x ? Note that as long as N, R, n, x are integers, this formula gives 0 unless x is in this range. (Note: while attempting to avoid being too judgemental, S above stands for *success* and F for *Failure*)

The Binomial Distribution

The setup is identical to that in the last paragraph only now we sample *with replacement*. Thus, for each member of the sample, the probability of an S is

$p = r/N$. Then the probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,\ 1, \ldots n$$

With any distribution, the sum of *all* the probabilities should be 1. Check that this is the case for the binomial, i.e. that

$$\sum_{x=0}^{n} f(x) = 1.$$

The Hypergeometric distribution is often approximated by the binomial distribution in the case N large. Problem 6 below justifies this approximation. Note that in the case of the binomial distribution, the two *parameters* (constants that one needs to determine the distribution) n, p are fixed, and usually known. For fixed sample size n we have counted X the number of S's in n *trials* of a simple experiment (e.g. tossing a coin).

The Negative Binomial distribution

The binomial distribution was generated by assuming that we repeated trials a fixed number n of times and then counted the total number of successes X in those n trials. Suppose we decide in advance that we wish a fixed number (k) of successes instead, and sample repeatedly until we obtain exactly this number. Then the number of trials X is random.

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots$$

A special case of most interest is the case $k = 1$ called the *Geometric* distribution. Then

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

The Poisson Distribution.

Suppose that a disease strikes members of a large population (of n individuals) independently, but in each case it strikes with very small probability p . If we count X the number of cases of the disease in the population, then X has the binomial (n, p) distribution. For very large n and small p this distribution can be again approximated as follows:

Theorem. Suppose $f_n(x)$ is the probability function of a binomial distribution with $p = \lambda/n$ for some fixed λ. Then as $n \to \infty$,

$$f_n(x) \to f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for each $x = 0, 1, 2, \ldots$.

The function $f(x)$ above is the probability function of a *Poisson Distribution* named after a French mathematician. This distribution has a single parameter λ, which makes it easier to use than the binomial, since the binomial requires knowledge or estimation of two parameters. For example the size n of the population of individuals who are susceptible to the disease might be unknown but the "average" number of cases in a population of this type λ could be obtained.

Example.

Phone calls arrive at a switchboard at an average rate of one every two minutes. If the operator nips out for a quick drink (5 minutes) what is the probability that there are no calls in this interval? What is the probability that there are more than three calls (in which case the supervisor is alerted).

8.3 Expected Values, Mean, Variances

Expected Value

An indicator random variable I_A takes two values, the value 1 with probability $P(A)$ and the value 0 otherwise. Its expected value, or average over many (independent) trials would therefore be $0(1-P(A))+1P(A)=P(A)$. This is the simplest case of an integral or expectation.

Recall that a simple random variable is one which has only finitely many distinct values c_i on the sets A_i where these sets form a partition of the sample space (i.e. they are disjoint and their union is Ω).

Expectation of simple random Variables.

For a simple random variable $X = \sum_i c_i I_{A_i}$, define $E(X) = \sum_i c_i P(A_i)$. The form is standard:

$$E(X) = \sum (values\ of\ X) \times Probability\ of\ values$$

Thus, for example, if a random variable X has probability function $f(x) = P[X=x]$, then $E(X) = \sum_x x f(x)$.

Properties.

For simple random variables $X,\ Y$,

1. $X(\omega) \le Y(\omega)$ for all ω implies $E(X) \le E(Y)$.
2. For real numbers $\alpha,\ \beta,\ E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proof. Suppose $X = \sum_i c_i I_{A_i} \le \sum_j d_j I_{B_j}$ where A_i forms a disjoint partition of the space Ω (i.e. are disjoint sets with $\cup_i A_i = \Omega$) and B_j also

forms a disjoint partition of the space. Then $c_i \leq d_j$ whenever $A_iB_j \neq \phi$. Therefore

$$E(X) = \sum_i c_i P(A_i) = \sum_i c_i \sum_j P(A_iB_j) \leq \sum_i \sum_j d_j P(A_iB_j) = \sum_j d_j P(B_j) = E(Y)$$

For the second part, note that $\alpha X + \beta Y$ is also a simple random variable that can be written in the form $\sum_i \sum_j (\alpha c_i + \beta d_j) I_{A_iB_j}$ where the sets A_iB_j form a disjoint partition of the sample space Ω. Now take expectation to verify that this equals $\alpha \sum_i c_i P(A_i) + \beta \sum_j d_j P(B_j)$.

Example Find the Expected value of X , a random variable having the Binomial(n, p) distribution.

Expectation of non-negative measurable random variables.

Definition: Suppose X is a non-negative random variable so that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then we define

$$E(X) \;=\; sup\{E(Y);\; Y\; simple,\; Y \leq X\}.$$

Expected value: discrete case.

If a random variable X has probability function $f(x) = P[X = x]$, then the definition of expected value in the case of *finitely many* possible values of x is essentially $E(X) = \sum_x xf(x)$. This formula continues to hold even when X may take a countably infinite number of values provided that the series $\sum_x xf(x)$ is absolutely convergent.

Example.

Find the expected value of a random variable X having the geometric distribution.

Notation.

Note that by $\int_A XdP$ we mean $E(XI_A)$ where I_A is the indicator of the event A.

Properties of Expectation.

Assume $X,\ Y$ are non-negative random variables. Then ;

1. If $X = \sum_i c_i I_{A_i}$ simple, $E(X) = \sum_i c_i P(A_i)$.
2. If $X(\omega) \leq Y(\omega)$ for all ω , $E(X) \leq E(Y)$.

3. If X_n increasing to X, then $E(X_n)$ increases to $E(X)$ (this is usually called the *Monotone Convergence Theorem*).

4. For non-negative numbers $\alpha,\ \beta,\ E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proof of Properties.

(1) If $Z \leq X$ and Z is a simple function, then $E(Z) \leq E(X)$. It follows that since X is a simple function and we take the supremum over all simple functions Z, that this supremum is $E(X)$.

(2.) Suppose Z is a simple function $\leq X$. Then $Z \leq Y$. It follows that the set of Z satisfying $Z \leq X$ is a subset of the set satisfying $Z \leq Y$ and therefore the supremum of $E(Z)$ over the former cannot be greater.

(3.) Since $X_n \leq X$ if follows form property (2) that $E(X_n) \leq E(X)$. Similarly $E(X_n)$ is monotonically non-decreasing and it therefore converges. Thus it converges to a limit satisfying

$$limE(X_n) \leq E(X).$$

We will now show that $lim\ E(X_n) \geq E(X)$ and then conclude equality holds above. Suppose $\epsilon > 0$ is arbitrary and $Y = \sum_i c_i I_{A_i}$ where $Y \leq X$ is a simple random variable. Define $E_n = \{\omega; X_n(\omega) \geq (1-\epsilon)Y(\omega)\}$ Note that as $n \to \infty$, this sequence of sets increases to a set containing $\{\omega; X(\omega) \geq (1-\epsilon/2)Y(\omega)\}$ and since $X \geq Y$ the latter is the whole space Ω. Therefore,

$$E(X_n) \geq \int_{E_n} X_n dP \geq (1-\epsilon)\int_{E_n} Y dP.$$

But

$$\int_{E_n} Y dP = \sum_i c_i P(A_i E_n) \to \sum_i c_i P(A_i)$$

as $n \to \infty$. Therefore

$$lim\ \ E(X_n) \geq (1-\epsilon)E(Y)$$

whenever Y is a simple function satisfying $Y \leq X$. Note that the supremum of the right hand side over all such Y is $(1-\epsilon)E(X)$. We have now shown that for any $\epsilon > 0$, $limE(X_n) \geq (1-\epsilon)E(X)$ and it follows that this is true also as $\epsilon \to 0$.

(4) Take two sequences of simple random variables X_n increasing to X and Y_n increasing to Y. Assume α and β are non-negative. Then by property 2. of 4.1.2,

$$E(\alpha X_n + \beta Y_n) = \alpha E(X_n) + \beta E(Y_n)$$

By monotone convergence, the left side increases to the limit $E(\alpha X + \beta Y)$ while the right side increases to the limit $\alpha E(X) + \beta E(Y)$. We leave the more general case of a proof to later.

General Definition of Expected Value.

For an arbitrary random variable X, define $X^+ = max(X,0)$, $X^- = max(0,-X)$. Note that $X = X^+ - X^-$. Then we define $E(X) = E(X^+) - E(X^-)$. This is well defined even if one of $E(X^+)$ or $E(X^-)$ are equal to ∞ as long as both or not infinite since the form $\infty - \infty$ is meaningless.

Definition.

If both $E(X^+) < \infty$ and $E(X^-) < \infty$ then we say X is *integrable*.

Example:

Define a random variable X such that $P[X = x] = \frac{1}{x(x+1)}, x = 1, 2, \ldots.$.

Is this random variable integrable?

General Properties of Expectation.

In the general case, expectation satisfies 1-4 of 4.1.8 above plus the the additional properties:

1. If $P(A) = 0$, $\int_A X(\omega)dP = 0$
2. If $P[X = c] = 1$ for some constant c, then $E(X) = c$.
3. If $P[X \geq 0] = 1$ then $E(X) \geq 0$. .

Other interpretations of Expected Value

For a discrete distribution, the distribution is often represented graphically with a bar graph or histogram. If the values of the random variable are $x_1 < x_2 < x_3 < \ldots$ then rectangles are constructed around each value, x_i, with *area* equal to the probability $P[X = x_i]$. In the usual case that the x_i are equally spaced, the rectangle around x_i has as base $(\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2})$. In this case, the expected value $E(X)$ is the x-coordinate of the center of gravity of the probability histogram.

We may also think of expected value as a long run average over many independent repetitions of the experiment. Thus, $f(x) = P[X = x]$ is approximately the long run proportion of occasions on which we observed the value $X = x$ so the *long run average* of many independent replications of X is $\sum_x xf(x) = E(X)$.

8.4 Discrete Bivariate and Multivariate Distributions

Definitions.

Example.

Suppose we throw 2 dice and define two random variables $X =$ maximum of the two numbers observed and $Y =$ minimum. We wish to record the probability of all possible combinations of values for both X and Y. We may do so through a formula for these joint probabilities

$$P[X = x, Y = y] = f(x, y) = \begin{cases} 2/36 & x > y \\ 1/36 & x = y \\ 0 & x < y \end{cases}$$

for $x, y = 1, 2, \ldots 6$.

Definitions.

The function $f(x, y) = P[X = x, Y = y]$ giving the probability of all combinations of values of the random variables is called the *joint probability function* of X and Y. The function $F(x, y) = P[X \leq x, Y \leq y]$ is called the *joint cumulative distribution function.* The joint probability function allows us to compute the probability functions of both X and Y. For example

$$P[X = x] = \sum_{all\ y} f(x, y).$$

We call this the *marginal* probability function of X, denoted by $f_X(x) = P[X = x] = \sum_{all\ y} f(x, y)$. Similarly, $f_Y(y)$ is obtained by adding the joint probability function over all values of x. Finally we are often interested in the conditional probabilities of the form

$$P[X = x|Y = y] = f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

This is called the *conditional probability function* of X given Y.

Example Determine all marginal probability functions and conditional probability functions in Example 5.1.1.

Expected Values

For a single (discrete) random variable we determined the expected value of a function of X, say $h(X)$ by

$$E[h(X)] = \sum_{all\ x} (value\ of\ h) \times (Probability\ of\ value) = \sum_{x} h(x) f(x)$$

For two or more random variables we should use a similar approach. However, when we add over all cases, this requires adding over all values of x *and* y. Thus, if h is a function of both X and Y,

$$E[h(X,Y)] = \sum_{all\ x\ and\ y} h(x,y)f(x,y).$$

Definition: Independent Random Variables

Two random variables X, Y are said to be *independent* if the events $[X = x]$ and $[Y = y]$ are independent for all $x,\ y$, i.e. if

$$P[X = x, Y = y] = P[X = x]P[Y = y] \quad all \quad x, y$$

i.e. if

$$f(x,y) = f_X(x)f_Y(y) \quad all \quad x, y.$$

This definition extends in a natural way to more than two random variables. For example we say random variables $X_1, X_2, \ldots X_n$ are (mutually) independent if, for every choice of values $x_1, x_2, \ldots x_n$, the events $[X_1 = x_1], [X_2 = x_2], \ldots [X_n = x_n]$ are independent events. This holds if the joint probability function of all n random variables factors into the product of the n marginal probability functions.

Theorem

If X, Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

Example

Suppose X and Y are two independent random variables with the same distribution (i.e. same probability functions)

$$f_X(x) = (1-p)^x p, \quad x = 0, 1, \ldots$$

and

$$f_Y(y) = (1-p)^y p, \quad y = 0, 1, \ldots$$

where $0 < p < 1$. Find the probability function of $Z = X + Y$ and the conditional probability function $f_{X|Z}(x|z)$.

Definition: Variance

The variance of a random variable measures its variability about its own expected value. Thus if one random variable has larger variance than another, it *tends* to be farther from its own expectation. If we denote the expected value of X by $E(X) = \mu$, then

$$Var(X) = E[(X - \mu)^2].$$

Adding a constant to a random variable does not change its variance, but multiplying it by a constant does; it multiplies the original variance by the constant squared (see 5.1.13, property 2.)

Example

Suppose the random variable X has the binomial (n, p) distribution. Find $E(X)$ and $var(X)$.

Definition:Covariance.

Define the covariance between 2 random variables $X,\ Y$ as

$$cov(X, Y) = E[(X - EX)(Y - EY)]$$

Covariance measures the linear association between two random variables. Note that the covariance between two *independent random variables* is 0. If The covariance is large and positive, there is a tendency for large values of X to be associated with large values of Y . On the other hand, if large values of X are associated with small values of Y , the covariance will tend to be negative. There is an alternate form for covariance, generally easier for hand calculation but more subject to computer overflow problems: $cov(X, Y) = E(XY) - (EX)(EY)$.

Theorem.

For any two random variables $X,\ Y$

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

One special case is of fundamental importance: the case when X, Y are independent random variables and $var(X + Y) = var(X) + var(Y)$ since $cov(X, Y) = 0$.

Example

A population includes a proportion p of unemployed. An interviewer polls members of the population (you may assume with replacement since the population is large) at random until exactly k unemployed have been found and records $X =$ the total number polled. Find the variance of X when $k = 1$ and use this to determine $var(X)$ in general.

Properties of Variance and Covariance

For any random variables X_i and constants a_i

1. $Var(X_1) = cov(X_1, X_1)$.
2. $var(a_1X_1 + a_2) = a_1^2 var(X_1)$.
3. $cov(X_1, X_2) = cov(X_2, X_1)$.
4. $cov(X_1, X_2 + X_3) = cov(X_1, X_2) + cov(X_1, X_3)$.
5. $cov(a_1X_1, a_2X_2) = a_1a_2cov(X_1, X_2)$.

Correlation Coefficient

The covariance has an arbitrary scale factor because of property 5 above. This means that if we change the units in which something is measured, (for example a change from imperial to metric units of weight), the covariance will change. It is desirable to measure covariance in units free of the effect of scale. To this end, define the *standard deviation* of X by $SD(X) = \sqrt{var(X)}$. Then the *correlation coefficient* between X and Y is

$$\rho = \frac{cov(X, Y)}{SD(X)SD(Y)}$$

Theorem

For any pair of random variables X, Y, we have $-1 \leq \rho \leq 1$ with $\rho = \pm 1$ if and only if the points (X, Y) always lie on a line so $Y = aX + b$ for some constants a, b.

The Multinomial Distribution

Suppose an experiment is repeated n times (called "trials") where n is fixed in advance. On each "trial" of the experiment, we obtain an outcome in one of k different categories $A_1, A_2, \ldots A_k$ with the probability of outcome A_i given by p_i. Here $\sum_{i=1}^{k} p_i = 1$. At the end of the n trials of the experiment consider the count of X_i =number of outcomes in category i, $i = 1, 2, \ldots k$. Then the random variables $(X_1, X_2, \ldots X_k)$ have a joint *multinomial* distribution given by the joint probability function

$$P[X_1 = x_1, X_2 = x_2, \ldots X_k = x_k] = \binom{n}{x_1 \quad x_2 \quad \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

whenever $\sum_i x_i = n$ and otherwise this probability is 0. Note that the marginal distribution of each X_i is binomial (n, p_i) and so $E(X_i) = np_i$.

Example

In political poll of 1000 respondents, $X_1 = 400$ indicated that they would vote "Yes" in a referendum, $X_2 = 360$ indicated that they would vote "No" and the remainder were undecided. Write an expression for the probability of exactly this outcome assuming that $p_{no} = p_{yes} = .38$. Find the probability that $X_1 = x$ given that $X_1 + X_2 = y$.

Covariance of a linear transformation.

Suppose $X = (X_1, ..., X_n)'$ is a vector whosse components are possibly dependent random variables. We define the expected value of this random vector by

$$\mu = E(X) = \begin{pmatrix} EX_1 \\ . \\ . \\ . \\ . \\ EX_n \end{pmatrix}$$

and the covariance matrix by a matrix

$$V = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & . & . & cov(X_1, X_n) \\ & & & & \\ & & & & \\ cov(X_n, X_1) & . & . & . & var(X_n) \end{pmatrix}.$$

Then if A is a $q \times n$ matrix of constants, the random vector $Y = AX$ has mean $A\mu$ and covariance matrix AVA'. In particular if $q = 1$, the variance of AX is AVA'.

8.5 Continuous Distributions

Definitions

Suppose a random variable X can take any real number in an interval. Of course the number that we record is often rounded to some appropriate number of decimal places, so we don't actually observe X but $Y = X$ rounded to the nearest $\Delta/2$ units. So, for example, the probability that we record the number $Y = y$ is the probability that X falls in the interval $y - \Delta/2 < X \leq y + \Delta/2$. If $F(x)$ is the cumulative distribution function of X this probability is $P[Y = y] = F(y + \Delta/2) - F(y - \Delta/2)$. Suppose now that Δ is very small and that the cumulative distribution function is piecewise continuously differentiable with a derivative given in an interval by

$$f(x) = F'(x).$$

Then $F(y+\Delta/2)-F(y-\Delta/2) \approx f(y)\Delta$ and so Y is a discrete random variable with probability function given (approximately) by $P[Y = y] \approx \Delta f(y)$. The derivative of the cumulative distribution function of X, provided it exists, is called the *probability density function* of the random variable X. Notice that an interval of small length Δ around the point y has approximate probability given by $length\ of\ interval \times f(y)$. Thus the probability of a (small) interval is approximately proportional to the probability density function in that interval, and this is the motivation behind the term *probability density*.

Example.

Suppose X is a random number chosen in the interval $[0,1]$. Any interval of length $\Delta \subset [0,1]$ is to have the same probability Δ regardless of where it is located. Then the cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

The probability density function is given by the derivative of the c.d.f. $f(x) = 1$ for $0 < x < 1$ and otherwise $f(x) = 0$. Notice that $F(y) = \int_{-\infty}^{y} f(x)dx$ for all y and the probability density function can be used to determine probabilities as follows;

$$P[a < X < b] = P[a \le X \le b] = \int_a^b f(x)dx.$$

In particular, notice that $F(b) = \int_{-\infty}^{b} f(x)dx$ for all b.

Example.

Is it always true that $F(b) = \int_{-\infty}^{b} F'(x)dx$*?* Let $F(x)$ be the binomial $(n, 1/2)$ cumulative distribution function. Notice that the derivative $F'(x)$ exists and is continuous except at finitely many points $x = 0,\ 1,\ 2,\ 3,\ 4$. Is it true that $F(b) = \int_{-\infty}^{b} F'(x)dx$?

Definition (cumulative distribution function)

Suppose the cumulative distribution function of a random variable $F(x)$ is such that its derivative $f(x) = F'(x)$ exists except at finitely many points. Suppose also that

$$F(b) = \int_{-\infty}^{b} f(x)dx \tag{6.1}$$

for all $b \in \Re$. Then the distribution is called *(absolutely) continuous* and the function $f(x)$ is called the *probability density function*.

Example.

Is it really necessary to impose the additional requirement (6.1) or this just a consequence of the fundamental theorem of calculus? Consider the case $F(x) = 0, x < 0,$ and $F(x) = 1,\ x \geq 0.$ This cumulative distribution function is piecewise differentiable (the only point where the derivative fails to exist is the point $x = 0$). But is the function the integral of its derivative?

For a continuous distribution, probabilities are determined by integrating the probability density function. Thus

$$P[a < X < b] = \int_a^b f(x)dx \tag{6.2}$$

A probability density function is not unique. For example we may change $f(x)$ at finitely many points and it will still satisfy (2) above and all probabilities, determined by integrating the function, remain unchanged. Whenever possible we will choose a continuous version of a probability density function, but at a finite number of discontinuity points, it does not matter how we define the function.

Properties of a Probability Density Function

1. $f(x) \geq 0$ for all $x \in \Re$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

The Uniform Distribution.

Consider a random variable X that takes values with a continuous uniform distribution on the interval $[a, b]$. Then the cumulative distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

and so the probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ and elsewhere the probability density function is 0. Again, notice that the definition of f at the points a and b does not matter.

Example

Let U have a continuous uniform distribution on the interval $[0, 1]$. Define the random variable $X = ln(1/U)$. Find the cumulative distribution function of X and determine whether it is absolutely continuous. If so, find its probability density function.

Expected Values for Continuous Distributions.

Suppose we were to approximate a continous random variable X having probability density function $f(x)$ by a a discrete random variable Y obtained by rounding X to the nearest Δ units. Then the probability function of Y is

$$P[Y = y] = P[y - \Delta/2 \leq X \leq y + \Delta/2] \approx \Delta f(y)$$

and its expected value is

$$E(Y) = \sum_y yP[y - \Delta/2 < X \leq y + \Delta/2] \approx \sum_y y\Delta f(y).$$

Note that as the interval length Δ approaches 0, this sum approaches the integral

$$\int xf(x)dx$$

and thus we define, for *continuous random variables*

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and for any function on the real numbers $h(x)$,

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx.$$

Example

Find the expected value and the variance of a random variable having probability density function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

The Exponential Distribution.

Consider a random variable X having probability density function

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x > 0$$

The cumulative distribution function is given by

$$F(x) = 1 - e^{-x/\mu}$$

and the moments are

$$E(X) = \mu, \quad var(X) = \mu^2$$

Such a random variable is called the *exponential distribution* and it is commonly used to model lifetimes of simple components such as fuses, transistors, etc that are not subject to wear and tear.

Example

If the lifetime of a transistor is exponentially distributed, find the probability that it will survive at least t more months given that it has already survived for x months. Compare this with the probability that a new transistor lives for at least t months.

Note: this property is called the memoryless property of the exponential distribution. A component with this distribution of liffetimes does not exhibit any evidence of aging.

Example

Show that for a uniform[a,b] random variable X, we have $E(X) = \frac{a+b}{2}$ and $var(X) = \frac{(b-a)^2}{12}$.

Generating Random variables with an Exponential Distribution.

Suppose that a computer has a built-in generator for the uniform[0,1] distribution (as is the case for nearly every higher-level computer language). How could I use a Uniform random variable U to generate an exponential random variable? Let $X = -\mu \ln(1-U)$. Find the cumulative distribution function of X.

Two methods for Computer Generation of Random Variables.

By far the simplest and most common method for generating non-uniform variates is based on the inverse cumulative distribution function. For arbitrary c.d.f. $F(x)$, 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 c.d.f. is continuous and strictly increasing. However, in the general case of a possibly discontinuous non-decreasing c.d.f. the function continues to enjoy some of the properties of an inverse. In particular, in the general case, *If* F *is an arbitrary c.d.f. and* U *is uniform*$[0,1]$ *then* $X = F^{-1}(U)$ *has c.d.f.* $F(x)$.

Example: Generating a discrete random variable

Consider generating, using U a uniform$[0,1]$ random variable, a random variable X having the following probability function:

x	1	2	3	4	5	6
P[X=x]	0.1	0.3	0.2	0.1	0.1	0.2

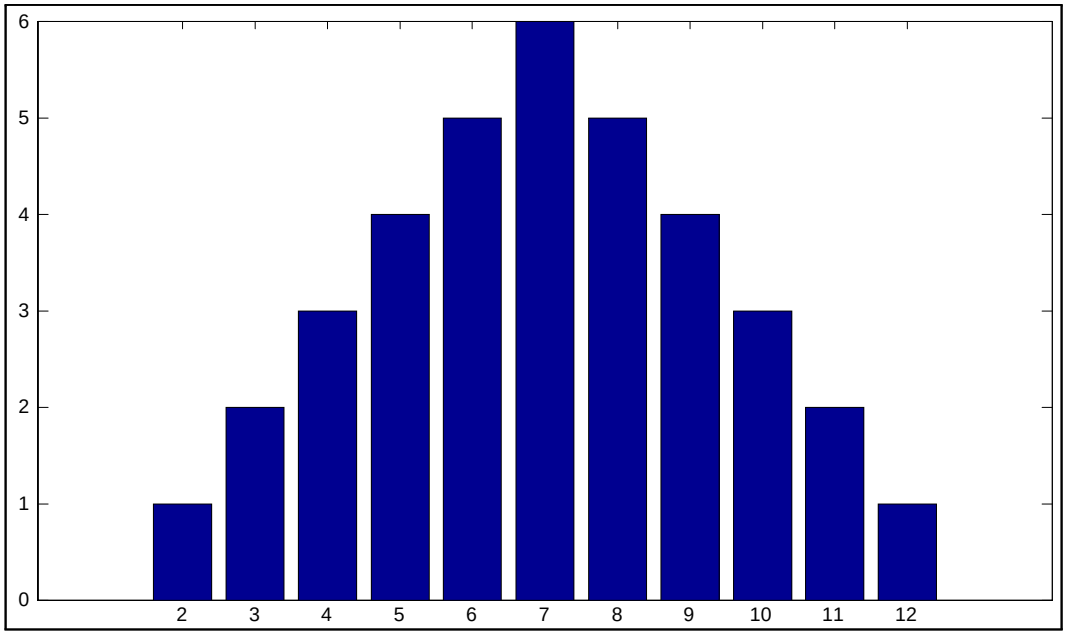


Figure 8.1:

Example: Generating a Geometric (p) random variable.

In this case, the c.d.f. is given by $F(x) = 1-(1-p)^{[x]}$, $x \geq 0$ where $[x]$ denotes the integer part of x. Then

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

where E is exponential(1) generates a geometric random variable. Compare the efficiency of this generator with one defined by $X = \min\{N; U_N < p\}$ where $U_1, U_2, ...$ are independent uniform[0,1] random variables.

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 a point satisfies the condition has a geometric distribution with parameter $p = 1/c$ and mean c so when c is large, the rejection method is not very effective.

The Normal distribution

Normal Approximation to the Poisson distribution

Consider a random variable X which has the Poisson distribution with parameter μ . Recall that $E(X) = \mu$ and $var(X) = \mu$ so $SD(X) = \sqrt{\mu}$. We

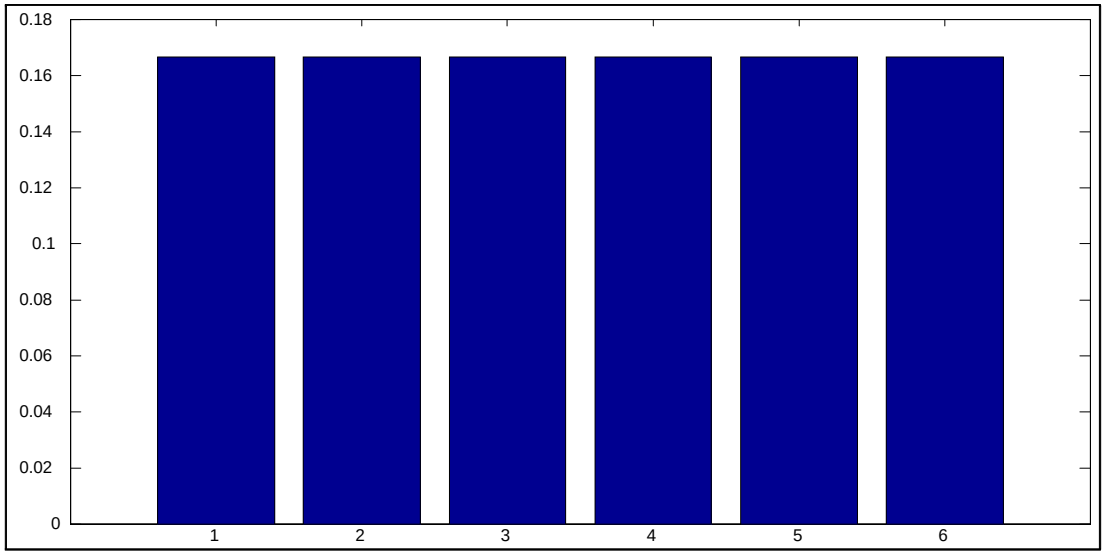


Figure 8.2:

wish to approximate the distribution of this random variable for large values of μ. In order to prevent the distribution from disappearing off to $+\infty$, consider the standardized random variable

$$Z = \frac{X - \mu}{\sqrt{\mu}}.$$

Then $P[Z = z] = P[X = \mu + z\sqrt{\mu}] = \frac{\mu^x}{x!}e^{-\mu}$ where $x = \mu + z\sqrt{\mu}$ is an integer. Using Stirling's approximation $x! \sim \sqrt{2\pi x}x^x e^{-x}$ and taking the limit of this as $\mu \to \infty$, we obtain

$$\frac{\mu^x}{x!}e^{-\mu} \sim \frac{1}{\sqrt{2\pi\mu}}e^{-z^2/2}$$

where the symbol $\sim$ is taken to mean that the ratio of the left to the right hand side approaches 1.

The standard normal distribution

Consider a continuous random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad -\infty < x < \infty$$

Such a distribution we call the *standard normal distribution* or the $N(0,1)$ distribution. The cumulative distribution function

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$$

is not obtainable in simple closed form, and requires either numerical approximation or a table of values. The probability density function $f(x)$ is symmetric about 0 and appears roughly as follows:

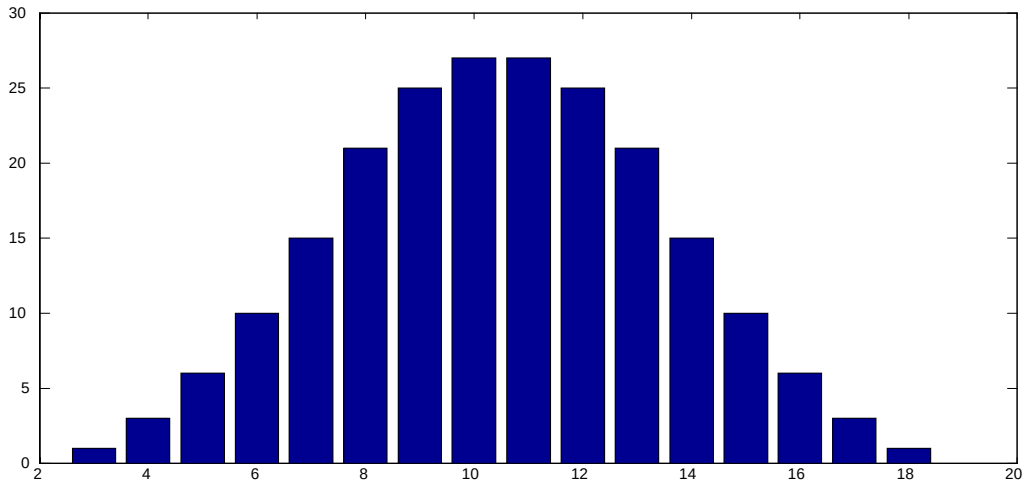


Figure 8.3: Standard Normal Probability Density Function

Example.

Prove that the integral of the standard normal probability density function is 1. The normal cumulative distribution function is as given below:

Note, for example that $F(-x) = 1 - F(x)$ for all x and if Z has a standard normal distribution

$$P[-1 < Z < 1] \approx .68 \quad \textit{and} \quad P[-2 < Z < 2] \approx .95.$$

Example.

If $Z \sim N(0,1)$ find $P[Z^2 \leq 3.84]$.

The General Normal Distribution.

If we introduce a shift in the location in the graph of the normal density as well as a change in scale, then the resulting random variable is of the form

$$X = \mu + \sigma Z, \quad Z \sim N(0,1)$$

for some constants $-\infty < \mu < \infty, \quad \sigma > 0.$

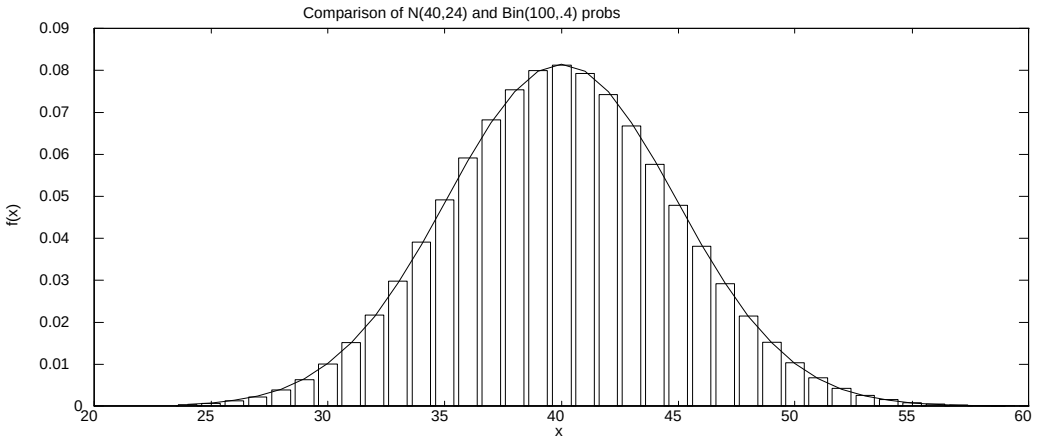


Figure 8.4:

Example.

Show that the probability density function of X is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X has the above normal distribution, we will denote this by $X \sim N(\mu, \sigma^2)$.

Moments

Show that the function $f(x;\mu,\sigma)$ integrates to 1 and is therefore a probability density function. Find the expected value and variance of a random variable having the probability density function $f(x;\mu,\sigma)$.

Linear Combinations.

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent random variables. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$.

Example.

Suppose $X_i \sim N(\mu, \sigma^2)$ are independent random variables. What is the distribution of the sample mean

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}?$$

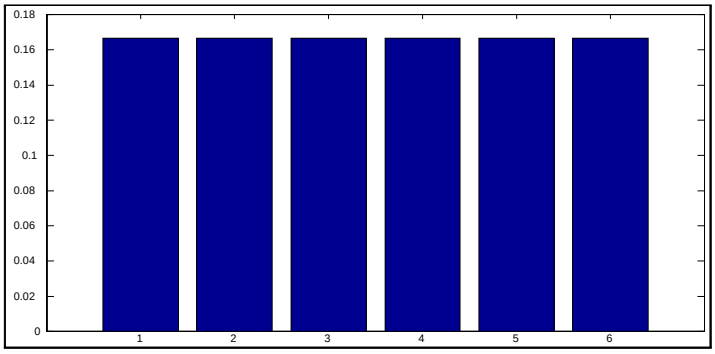


Figure 8.5:

Assume $\sigma = 1$ and find the probability $P[|\bar{X}_n - \mu| > 0.1]$ for various values of n. What happens to this probability as $n \to \infty$?

The Central Limit Theorem

The major reason that the normal distribution is the single most commonly used distibution is the fact that it tends to approximate the distribution of sums of random variables. For example, if we throw n dice and S_n is the sum of the outcomes, what is the distribution of S_n?. The tables below provide the number of ways in which a given value can be obtained. The corresponding probabability is obtained by dividing by 6^n.

$n = 1,$

1	2	3	4	5	6	
1	1	1	1	1	1	

$n = 2,$

2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	5	4	3	2	1

$n = 3$

3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

$n = 4$

4	5	6	7	8	.	.	.
1	4	10	20	35	.	.	.

The distributions show a simple pattern. For $n = 1$, the probability function is a constant (polynomial degree 0). For $n = 2$, two linear functions spliced together. For $n = 3$ a spline consisting of three quadratic pieces (polynomials of degree $n-1$). In general the histogram for S_n consists of n piecewise polynomials of degree $n-1$ which approach very rapidly the shape of the normal probability density function.

Example

Let $X_i = 0$ or 1 when the i'th toss of a biased coin is Tails or Heads respectively. What is the distribution of $S_n = \sum_{i=1}^{n} X_i$? Consider the standardized

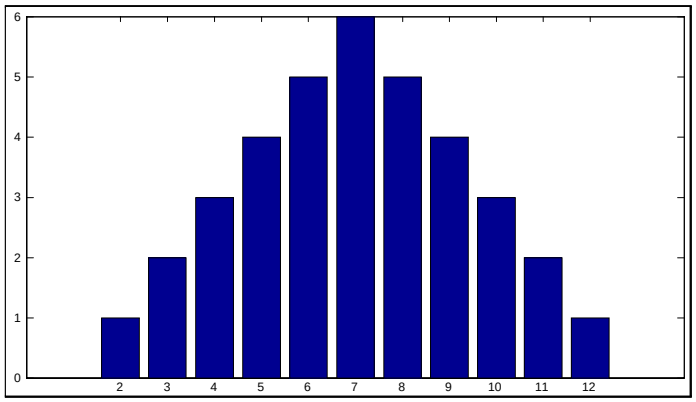


Figure 8.6:

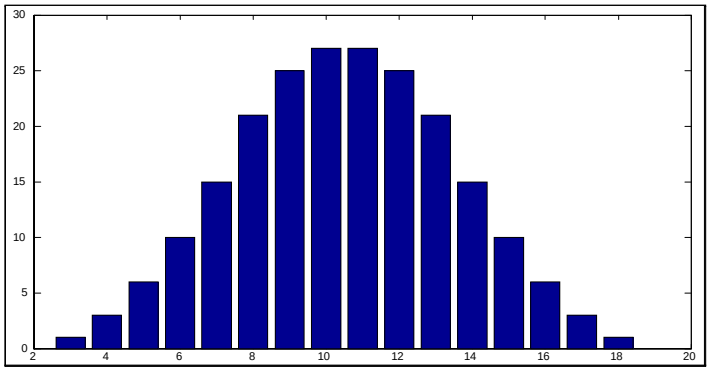


Figure 8.7:

random variable

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$$

Approximate the distribution of S_n^* for large values of n.

First let an integer $x \sim np + z\sqrt{np(1-p)}$ for fixed z. Then as $n \to \infty$, $x/n \to p$, $0 < p < 1$, Stirling's approximation implies that

$$\binom{n}{x} \sim \frac{\sqrt{2\pi}n^{n+1/2}e^{-n}}{2\pi x^{x+1/2}(n-x)^{n-x+1/2}} \sim \frac{1}{\sqrt{2\pi np(1-p)}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}.$$

Also using the series expansion $\ln(1+x) = x - \frac{1}{2}x^2 + O\left(x^3\right)$, putting $\sigma = \sqrt{\frac{p(1-p)}{n}}$, and noting $\sigma \to 0$ as $n \to \infty$,

$$\begin{aligned}
\ln\{\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}\} &= x\ln(\frac{p}{p+z\sigma}) + (n-x)\ln(\frac{1-p}{1-p-z\sigma}) \\
&= -x\ln(1+\frac{z\sigma}{p}) - (n-x)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\ln(1+\frac{z\sigma}{p}) - n(1-p-z\sigma)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\{(\frac{z\sigma}{p}) - \frac{1}{2}(\frac{z\sigma}{p})^2 + O(\frac{z\sigma}{p})^3\} \\
&\quad -n(1-p-z\sigma)\{-(\frac{z\sigma}{1-p}) - \frac{1}{2}(\frac{z\sigma}{1-p})^2 + O(\frac{z\sigma}{1-p})^3\} \\
&= -n\{z\sigma + \frac{z^2\sigma^2}{p} - \frac{1}{2}\frac{z^2\sigma^2}{p} - z\sigma + \frac{z^2\sigma^2}{1-p} - \frac{1}{2}\frac{z^2\sigma^2}{1-p} + O(\sigma^3)\} \\
&= -\frac{1}{2}z^2\sigma^2(\frac{n}{p} + \frac{n}{1-p}) + O(n^{-1/2}) \\
&= -\frac{z^2}{2} + O(n^{-1/2})
\end{aligned}$$

Therefore,

$$\begin{aligned}
P[S_n &= x] = P[S_n^* = z] = \binom{n}{x}p^x(1-p)^{n-x} \\
&\sim \binom{n}{x}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \\
&\sim \frac{1}{\sqrt{np(1-p)}}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}.
\end{aligned}$$

This is the standard normal probability density function multiplied by the distance between consecutive values of S_n^*. In other words, this result says that

the area under the probabability histogram for S_n^* for the bar around the point z can be approximated by the area under the normal curve between the same two points $(z \pm \frac{1}{2\sqrt{np(1-p)}})$.

Theorem.

Let $X_i, \quad i = 1, \ldots n$ be independent random variables all with the same distribution, and with mean μ and variance σ^2 . Then the cumulative distribution function of

$$S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges to the cumulative distribution function of a standard normal random variable.

The proof of this result we will defer after the discussion of moment generating functions.

Consider, for example, the case where the X_i are indpendent each with a Bernoulli (p) distribution. Then the sum $\sum_{i=1}^n X_i$ has a binomial distribution with parameters n, p and the above theorem asserts that if we subtract the mean and we divide by the standard deviation of a binomial random variable, then the result is approximately standard normal. In other words, for large values of n a binomial random variable is approximately normal $(np, np(1-p))$. To veriify this fact, we plot both the binomial$(100, 0.4)$ histogram as well as the normal probability density function below.

Example.

Use the central limit theorem and the normal approximation to a probability histrogram to estimate the probability that the sum of the numbers on 6 dice is 20. What is the exact probability?

The Distribution of a Function of a Random Variable.

We have seen that if X has a normal distribution, then a linear function of X, say $aX + b$ also has a normal distribution. The parameters are easily determined since $E(aX + b) = aE(X) + b$ and $var(aX + b) = a^2 var(X)$. Is this true of arbitrary functions and general distributions? For example is X^2 normally distributed? The answer in general is NO. For example, the distribution of X^2 must be concentrated entirely on the positive values of x, whereas the normal distributions are all supported on the whole real line (i.e. the probability density functiopn $f(x) > 0, \quad$ all $x \in \mathcal{R}$. In general, the safest method for finding the distribution of the function of a random variable in the continuous case is to first find the cumulative distribution of the function and then differentiate to obtain the probability density function.

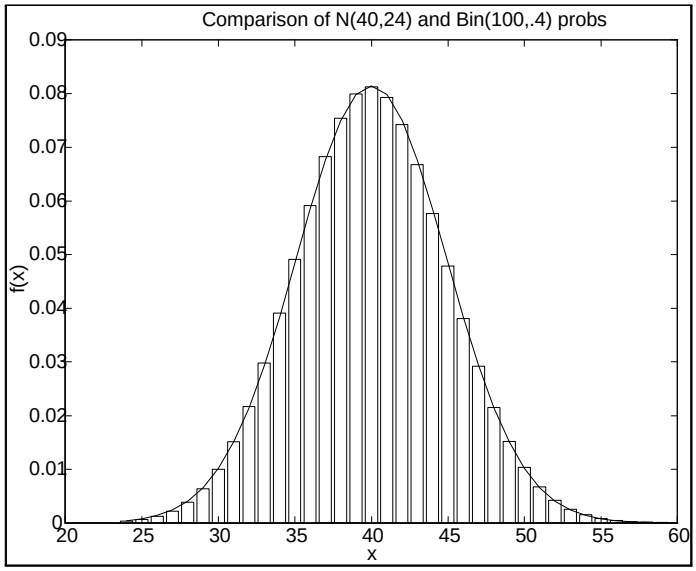


Figure 8.8:

Example.

Find the probability density function of $X = Z^2$ where Z has a standard normal distribution.

Theorem

Suppose a continuous random variable X has probability density function $f_X(x)$. Show that the probability density function of $Y = h(X)$ where $h(.)$ is a continuous monotone increasing function with inverse function $h^{-1}(y)$ is

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y)$$

Moment Generating Functions

Consider a random variable X. We have seen several ways of describing its distribution, using either a cumulative distribution function, a probability density function (continuous case) or probabililty function or a probability histogram or table (discrete case). We may also use some transform of the probability density or probability function. For example, consider the function defined by

$$M_X(t) = Ee^{tX}$$

defined for all values of t such that this expectations exists and is finite. This function is called the moment generating function of the (distribution of the)

random variable X. It is a powerful tool for detemining the distribution of sums of independent random variables and for proving the central limit theorem. In the discrete case we can write $M_X(t) = \sum_x e^{xt} P[X = x]$ and in the continuous case $M_X(t) = \int_{-\infty}^{\infty} e^{xt} f(x)dx$.

Properties of the Moment Generating Function

For these properties we assume that the moment generating function exists at least in some neighbourhood of the value $t = 0$, say for $-Varepsilon < t < Varepsilon$ for some $Varepsilon > 0$. We also assume that $\frac{d}{dt}E[X^n e^{tX}\} = E[\frac{d}{dt}X^n e^{tX}]$ for each value of $n = 0, 1, 2,$ for $-Varepsilon < t < Varepsilon$. The ability to differentiate under an integral or infinite sum is justified under general conditions involving the rate at which the integral or series converges.

1. $M'(0) = E(X)$
2. $M^{(n)}(0) = E(X^n), n = 1, 2, ...$
3. A moment generating function uniquely determines a distribution. In other words if $M_X(t) = M_Y(t)$ for all $-Varepsilon < t < Varepsilon$, then X and Y have the same distribution.
4. $M_{aX+b}(t) = e^{bt} M_X(at)$ for constants a, b.
5. If X and Y are independent random variables, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Example

Let X have a Binomial (n, p) distribution. Then the moment generating function of X is

$$M_X(t) = (pe^t + 1 - p)^n.$$

Example

Let X have a Poisson(λ) distribution. Then the moment generating function of X is

$$M_X(t) = \exp\{\lambda(e^t - 1)\}.$$

Example

Let X have an exponential distribution with mean μ. Then the moment generating function of X is

$$M_X(t) = \frac{1}{1 - \mu t} \quad \text{for } t < 1/\mu.$$

Example

Let X have a Normal (μ, σ^2) distribution. Then the moment generating function of X is

$$M_X(t) = \exp\{\mu t + \sigma^2 t^2/2\}.$$

Use this to show that the sum of independent normal random variables is also normally distributed.

Moment generating functions are useful for showing that a sequence of cumulative distribution functions converge because of the following result, stated without proof. The result implies that convergence of the moment generating functions can be used to show convergence of the cumulative distribution functions (i.e. convergence of the distributions).

Theorem

Suppose Z_n s a sequence of random variables with moment generating functions $M_n(t)$. Let Z be a random variable Z having moment generating function $M(t)$. If $M_n(t) \to M(t)$ for all t in a neighbourhood of 0, then

$$P[Z_n \leq z] \to P[Z \leq z]$$

as $n \to \infty$ for all values of z at which the function $F_Z(z)$ is continuous.

Proof of the Central Limit Theorem

We now use the properties of the moment generating function to prove the central limit theorem; i.e. that the cumulative distribution function of $S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$ converges to the c.d.f. of the standard normal distribution as $n \to \infty$. Note that $S_n^* = \frac{1}{\sqrt{n}} X_i^*$ where $X_i^* = (X_i - \mu)/\sigma$ and so it is sufficient to prove this result for standardized random variables with mean 0 and variance 1. In this case, by the above theorem, it is sufficient to show that the moment generating function of S_n^* converges to the moment generating function of the standard normal, i..e to $m(t) = e^{t^2/2}$. Now let $L_n(t)$ be the logarithm of the moment generating function

$$L_n(t) = \ln[M_n(t)] = \ln[Ee^{S_n^* t}]$$

and

$$L(t) = \ln[M(t)] = \ln[Ee^{X^* t}]$$

Note that

$$L_n(t) = nL(t/\sqrt{n})$$

and that

1. $L(0) = 0$
2. $L'(0) = 0$
3. $L''(0) = 1$

Then

$$\begin{aligned}
\lim_{n\to\infty} L_n(t) &= \lim_{n\to\infty} L(t/\sqrt{n})/(n^{-1}) \\
&= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \text{ by L'Hospital's rule} \\
&= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-L^{''}(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \text{ by L'Hospital's rule} \\
&= \lim_{n\to\infty} L^{''}(t/\sqrt{n})\frac{t^2}{2} \\
&= \frac{t^2}{2}
\end{aligned}$$

It follows on exponentiating that $M_n(t)$ converges to $e^{t^2/2}$ which is the $N(0,1)$ moment generating function and therefore the cumulative distribution function of S_n^* converges to the normal cumulative distribution function pointwise (since the latter c.d.f. is continuous everywhere).

8.6 Stochastic Processes

A Stochastic process is an indexed family of random variables X_t for t ranging over some index set T such as the integers or an interval of the real line. For example a sequence of independent random variables is a stochastic process, as is a Markov chain. For an example of a continuous time stochastic process, define X_t to be the price of a stock at time t (assuming trading occurs continously over time).

Markov Chains

Consider a sequence of (discrete) random variables $X_1, X_2, \ldots$ each of which takes integer values $1, 2, \ldots N$ (called *states*). We assume that for a certain matrix P (called the *transition probability matrix*), the conditional probabilities are given by corresponding elements of the matrix; i.e.

$$P[X_{n+1} = j | X_n = i] = P_{ij}, \quad i = 1, \ldots N, \quad j = 1, \ldots N$$

and furthermore that the chain only cares about the last state occupied in determining its future; i.e. that

$$P[X_{n+1} = j|X_n = i,\ X_{n-1} = i_1, X_{n-2} = i_2 ... X_{n-l} = i_l] = P[X_{n+1} = j|X_n = i] = P_{ij}$$

for all $j, i, i_1, i_2, \ldots$. Then the sequence of random variables X_n is called a *Markov Chain*. Markov Chain models are the most common simple models for dependent variables, including weather (precipitation, temperature), movements of security prices etc.

Properties of the Transition Matrix P

Note that $P_{ij} \geq 0$ for all i, j and $\sum_j P_{ij} = 1$ for all i. This last property implies that the $N \times N$ matrix $P - I$ (where I is the identity matrix) has rank at most $N - 1$ because the sum of the N columns of $P - I$ is identically 0.

Example. Rain-No rain

Suppose that the probability that tomorrow is rainy given that today is not is α and the probability that tomorrow is dry given that today is rainy is β.

Example. Gambler's Ruin

A gambler at each play of a game either wins \$1 or loses \$1 ith probabilities p, 1-p respectively. The gambler quits playing when his fortune reaches either 0 or M. Then the total fortune of the gambler at time t follows a Markov chain. What is the transition probability matrix?

The distribution of X_t

Show that if the chain is started by randomly choosing a state for X_0 with distribution $P[X_0 = i] = q_i,\ i = 1, 2, \ldots N$, then the distribution of X_1 is the vector $q'P$ where $\underline{q}$ is the column vector of values q_i. Similarly the distribution of X_t is the vector $\underline{q}'P^t$ where P^t is the product of the matrix P with itself t times. Under very general conditions, it can be shown that these probabilities converge and in many such cases, the limit does not depend on the initial distribution q.

Definition

A *limiting distribution* of a Markov chain is a vector ($\underline{\pi}$ say) of long run probabilities of the individual states so

$$\pi_i = lim_{t\to\infty} P[X_t = i].$$

Definition

A *stationary distribution* of a Markov chain is the column vector ($\underline{\pi}$ say) of probabilities of the individual states such that

$\underline{\pi}'P = \underline{\pi}'$.

Theorem

Any limiting distribution of a Markov Chain must be a stationary distribution.

Proof.

Note that $\pi' = \lim_{n\to\infty} q'P^n = \lim_{n\to\infty}(q'P^n)P = (\lim_{n\to\infty} q'P^n)P = \pi'P$.

Example

Consider a Markov chain with transition probability matrix

$$P = \begin{pmatrix} .9 & .1 \\ .2 & .8 \end{pmatrix}$$

Find $lim_{t\to\infty}P^t$ and the limiting distribution of the Markov chain. Show that in general for a 2×2 transition matrix, the stationary distribution is proportional to (P_{21}, P_{12}).

Example: Binary information:

Suppose that $X_1, X_2, \ldots$ is a sequence of binary information (Bernoulli random variables) taking values either 0 or 1. Suppose that the probability that a 0 is followed by a 1 is p and the probability that a 1 is followed by a 0 is given by q where $0 < p,\ q < 1$. Find the transition matrix for the Markov chain and the long run proportion of zeros in the sequence.

When is the limiting distribution of a Markov chain unique and independent of the initial state of the chain?

Definition: irreducible, aperiodic

We say that a Markov chain is *irreducible* if every state can be reached from every other state. In other words for every pair i, j there is some m such that $P^m_{i,j} > 0$. We say that the chain is *aperiodic* if $gcd\{N; P^{(N)}_{ii} > 0\} = 1$. For a *periodic chain* (i.e. one with period>1) returns to a state can occur only at multiples of the period $gcd\{N; P^{(N)}_{ii} > 0\}$.

Theorem

If a Markov chain is irreducible and aperiodic, then there exists a *unique* limiting distribution $\underline{\pi}$. In this case $P^n \to \pi'\underline{1}$ the matrix whose rows are all identically π' as $n \to \infty$.

Generating Functions.

Definition: Generating function

Let $a_0, a_1, a_2, \ldots$ be a finite or infinite sequence of real numbers. Suppose the power series

$$\mathcal{A}(t) = \sum_{i=0}^{\infty} a_i t^i$$

converges for all $-\epsilon < t < \epsilon$ for some value of $\epsilon > 0$. Then we say that the sequence has a *generating function* $\mathcal{A}(t)$.

Note. Every bounded seqence has a generating function since the series $\sum_{i=0}^{\infty} t^i$ converges whenever $|t| < 1$. Thus, discrete probability functions have generating functions. The generating function of a random variable X or its associated probability function $f_X(x) = P[X = x]$ is given by

$$\mathcal{F}_X(t) = \sum_x f_X(x) t^x = E(t^X).$$

Note that if the random variable has finite expected value, then this converges on the interval $t \in [-1, 1]$.

The *joy of generating functions* is that they provide a transform of the original distribution to a space where many operations are made much easier. We will give examples of this later. The most important single property is that they are in one-one correspondence with distributions such that the series converges; for each distribution there is a unique generating function and for each generating function there is a unique distribution.

As a consequence of this representation and the following theorem we can use generating functions to determine distributions that would otherwise be difficult to identify.

Theorem

Suppose a random variable X has generating function $\mathcal{F}_X(t)$ and Y has generating function $\mathcal{F}_Y(t)$. Suppose that X and Y are independent. Then the generating function of the random variable $W = X+Y$ is $\mathcal{F}_W(t) = \mathcal{F}_X(t)\mathcal{F}_Y(t)$.

Example

Find the distributions that corresponds to the following generating functions:

(a) $\mathcal{F}(t) = \frac{t}{3-2t}$

(b) $\mathcal{F}(t) = e^{\lambda(t-1)}$

Example

Find the generating function of the Binomial (n, p) distribution. Suppose X_1 and X_2 are independent random variables, both with this binomial distribution.

Find the distribution of $W = X_1 + X_2$. Notice that whenever a Moment generating function exists, we can recover the generating function from it by replacing e^t by t.

Example.

One of six defferent varieties of coupons is placed in each box of cereal. Find the distribution of the number of cereal boxes you need to buy to obtain all six coupons. (Answer: the (probability) generating function of the number is

$$\frac{5!t^6}{(6-t)(6-2t)(6-3t)(6-4t)(6-5t)}$$

$= \frac{5}{324}t^6 + \frac{25}{648}t^7 + \frac{175}{2916}t^8 + \frac{875}{11\,664}t^9 + \frac{11\,585}{139\,968}t^{10} + \frac{875}{10\,368}t^{11} + O\left(t^{12}\right)$
and this expansion as a power series provides the probabilities)

The Poisson Process.

One of the simplest continuous time stochastic processes is the *Poisson Process*. Suppose N_t denotes the total number of arrivals into a system (such as the number of customers arriving at a queue) untill time t. Note that the number of arrivals in time interval $(a, b]$ is then $N_b - N_a$. Assume the following properties;

(a) The probability of exactly one arrival in a small interval of length Δt is $\lambda\Delta t + o(\Delta t)$. (Note that the probability does not depend on where the interval is, only on its length).

(b) The probability of two or more arrivals in an interval of length Δt is $o(\Delta t)$ where by definition of the o notation, $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

(c) For disjoint intervals $I_i = (a_i, b_i]$ (so $I_i \cap I_j = \phi, \quad i \neq j$), the number of arrivals in these intervals $Y_i = N_{b_i} - N_{a_i}$ are mutually independent random variables.

Theorem.

Under the above conditions, (a)-(c), the distribution of the process $N_t, \quad t \in T$ is that of a *Poisson process*. This means that the number of arrivals $N_b - N_a$ in an interval $(a, b]$ has a Poisson distribution with parameter $\lambda(b-a) = \lambda\times$*the length of the interval*, and the number of arrivals in disjoint time intervals are independent random variables. The parameter λ specifies the *rate* of the Poisson process.

Example.

1. (a) (a) Show that if $N(t)$ is a Poisson process and $T_1, T_2, ...$ are the times of the first event, and the time between the first and second events, etc. then T_1, $T_2, ...$ are independent random variables, each with an exponential distribution with expected value $1/\lambda$.

(b) Show that if $T_1, T_2, \ldots T_n$ are independent random variables each with an exponential (1) distribution, then the sum $\sum_{i=1}^{n} T_i$ has a (gamma) probability density function

$$f(x) = \frac{1}{(n-1)!} x^{n-1} e^{-x}, x > 0.$$

Example.

Suppose emergency calls to 911 follow a Poisson process with an average of 10 calls per hour. What is the probability that there are no calls in a five minute period? What is the probability that there are more than 100 calls in an 8-hour shift? Given that there are 5 calls in the first hour, what is the probability that the first call occurred in the first x minutes?

Poisson Process in space.

In an analogous way we may define a Poisson process in space as a distribution governing the occurrence of random points with the properties indicated above; The number of points in a given set S has a Poisson distribution with parameter $\lambda \times |S|$ where $|S|$ is the area or volume of the set, and if $Y_1, Y_2, \ldots$ are the number of points occurring in disjoint sets $S_1, S_2, \ldots,$ they are mutually independent random variables.

Example

Bacteria are immersed in contaminated water at a rate of λ per nanolitre (note: 1 nanolitre $= 10^{-9}$ litres). What is the probability that there are no bacteria in a sample of 1 ml. if $\lambda = 10^{-3}$? What is the probability of more than 1200 bacteria in 1 ml if $\lambda = 10^{-3}$?

8.7 Conditional Expectation and Martingales

8.7.1 Conditional Expectation.

Theorem.

Let $\mathcal{G} \subset \mathcal{F}$ be sigma-algebras and X a random variable on $(\Omega, \mathcal{F}, P)$. Assume $E(X^2) < \infty$. Then there exists an almost surely unique $\mathcal{G}$-measurable Y such that

$$E[(X-Y)^2] \;=\; inf_Z E(X-Z)^2 \quad (6.1)$$

where the infimum is over all $\mathcal{G}$-measurable random variables. *Note.* We denote the minimizing Y by $E(X|\mathcal{G})$.

For two such minimizing Y_1, Y_2, i.e. random variables Y which satisfy (6.1), we have $P[Y_1 = Y_2] = 1$. This implies that conditional expectation is almost surely unique.

Example.

Suppose $\mathcal{G} = \{\varphi, \Omega\}$. What is $E(X|\mathcal{G})$?

Example.

Suppose $\mathcal{G} = \{\varphi, A, A^c, \Omega\}$ for some event A. What is $E(X|\mathcal{G})$? Consider the special case: $X = I_B$.

Example.

Suppose $\Omega = (0, 1]$ and the function $X(\omega)$ is Borel measurable. Assume that $\mathcal{G}$ is generated by the intervals $(\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, 2, ..., n$. What is $E(X|\mathcal{G})$?

Properties of Conditional Expectation.

(a) If a random variable X is $\mathcal{G}$-measurable, $E(X|\mathcal{G}) = X$.

(b) If a random variable X independent of a sigma-algebra $\mathcal{G}$, then $E(X|\mathcal{G}) = E(X)$.

(c) For any square integrable $\mathcal{G}$-measurable Z, $E(ZX) = E[ZE(X|\mathcal{G})]$.

(d) (special case of (c)): $\int_A XdP = \int_A E(X|\mathcal{G})dP$ for all $A \in \mathcal{G}$.

(e) $E(X) = E[E(X|\mathcal{G})]$.

(f) If a $\mathcal{G}$-measurable random variable Z satisfies $E[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y , then $Z = E(X|\mathcal{G})$.

(g) If Y_1, Y_2 are distinct $\mathcal{G}-$measurable random variables both minimizing $E(X - Y)^2$, then $P(Y_1 = Y_2) = 1$.

(h) *Additive* $E(X + Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.
Linearity $E(cX + d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i) If Z is $\mathcal{G}-$measurable, $E(ZX|\mathcal{G}) = ZE(X|\mathcal{G})$ a.s.

(j) If $\mathcal{H} \subset \mathcal{G}$ are sigma-algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$.

(k) If $X \leq Y$, $E(X|\mathcal{G}) \leq E(Y|\mathcal{G})$ a.s.

(l) *Conditional Lebesgue Dominated Convergence.* If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some integrable random variable Y , then $E(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ *in distribution*

Notes. In general, we define $E(X|Z) = E(X|\sigma(Z))$ and conditional variance $var(X|\mathcal{G}) = E\{(X - E(X|\mathcal{G}))^2|\mathcal{G}\}$. For results connected with property (l) above providing conditions under which the conditional expectations converge, see Convergence in distribution of conditional expectations, (1994) E.M. Goggin, *Ann. Prob* 22, 2. 1097-1114.

Conditional Expectation for integrable random variables.

For non-negative integrable X choose simple random variables $X_n \uparrow X$. Then $E(X_n|\mathcal{G}) \uparrow$ and so it converges. Define $E(X|\mathcal{G})$ to be the limit. In general, for random variables taking positive and negative values, we define $E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G})$.

8.7.2 Martingales.

Intuitively, a martingale is the total fortune of an individual participating in a "fair game". In order to be fair, the expected value of one's future fortune given the history of the process up to and including the present should be equal to one's present wealth. Suppose the fortune at time s is denoted X_s. The current process and any other related processes up to time s generate a sigma-algebra $\mathcal{F}_s$. Then the assertion that the game is fair implies $E(X_t|\mathcal{F}_s) = X_s$ for $t > s$.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *martingale* if

(a) $\mathcal{F}_t$ is increasing (in t) family of sigma-algebras

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_t|\mathcal{F}_s) = X_s$ a.s.

Example.

Suppose Z_t are independent random variables with expectation 0. Define $\mathcal{F}_t = \sigma(Z_1, Z_2, \ldots Z_t)$ and $S_t = \sum_{i=1}^{t} Z_i$. Then $\{(S_t, \mathcal{F}_t),\ t = 1, 2, ...\}$ is a martingale.

Example.

Let X be any integrable random variable, and $\mathcal{F}_t$ an increasing family of sigma-algebras. Put $X_t = E(X|\mathcal{F}_t)$. Then $(X_t, \mathcal{F}_t)$ is a martingale.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *reverse martingale* if

(a) $\mathcal{F}_t$ is decreasing (in t) family of sigma-algebras.

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_s|\mathcal{F}_t) = X_t$ a.s.

Example.

Let X be any integrable random variable, $\mathcal{F}_t$ be any decreasing family of sigma-algebras. Put $X_t = E(X|\mathcal{F}_t)$. Then $(X_t, \mathcal{F}_t)$ is a reverse martingale.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *sub (super) martingale* if

(a) $\mathcal{F}_t$ is increasing (in t) family of sigma-algebras.

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_t|\mathcal{F}_s) \geq (\leq)\ X_s$ a.s.

Example.

Let Y_i be independent identically distributed, $\mathcal{F}_n = \sigma(Y_{(1)}, ..., Y_{(n)}, Y_{n+1}, Y_{n+2}, ...)$, where $(Y_{(1)}, ..., Y_{(n)})$ denote the order statistics. Then $\mathcal{F}_n$ is a decreasing family of sigma fields and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = E(Y_1|\mathcal{F}_n)$ is a reverse martingale.

Definition.

A random variable τ is a (optional) *stopping time* for a martingale $(X_t, \mathcal{F}_t)$ if for each t, $[\tau \leq t] \in \mathcal{F}_t$.

Definition.

For an optional stopping time τ define

$$\mathcal{F}_\tau = \{A \in \mathcal{F}; A \cap [\tau \leq t] \in \mathcal{F}_t, \quad for\ all\ t\}.$$

Then this is a sigma-algebra.

Theorem.

If $(X_t, \mathcal{F}_t)\ t = 1, 2, \ldots n$ is a (sub) martingale and α, β are stopping times with values in $\{1, .., n\}$, such that $\alpha \leq \beta$, then

$$E(X_\beta|\mathcal{F}_\alpha)\ (\geq) = X_\alpha$$

(Sub)martingale Convergence Theorem.

Let $(X_n, \mathcal{F}_n);\ n = 1, 2, \ldots$ be a submartingale such that $lim\ sup_{n\to\infty} E|X_n| < \infty$. Then there is a (finite) random variable X such that $X_n \to X$ a.s.

Reverse martingale convergence Theorem.

If $(X_n, \mathcal{F}_n);\ n = 1, 2, \ldots$ is a reverse martingale,

$$X_n \to E(X_1 | \cap_{n=1}^{\infty} \mathcal{F}_n) \quad a.s.$$

8.7.3 Martingales and Finance

Let $S(t)$ denote the price of a security at the beginning of period $t = 0, 1, 2, ...T$. We assume that the security pays no dividends. Define the *(cumulative) returns process* associated with this security by R_S where

$$\Delta R_S(t) = R_S(t) - R_S(t-1) = \frac{\Delta S(t)}{S(t-1)} = \frac{S(t) - S(t-1)}{S(t-1)}, \quad R_S(0) = 0.$$

Then $100\Delta R_S(t)\%$ is the percentage return in an investment in the stock in the $t-1'$st period. The returns process is a more natural characterisation of stock prices than the original stock price process since it is invariant under artificial scale changes such as stock splits etc. Note that we can write the stock price in terms of the returns process;

$$S(t) = S(0) \prod_{i=1}^{t} (1 + \Delta R_S(i)).$$

Now consider another security, a *riskless discount bond* which pays no coupons. Assume that the price of this bond at time t is $B(t), \quad B(0) = 1$ and $R_B(t)$ is the return process associated with this bond. Then $\Delta R_B(t) = r(t)$ is the interest rate paid over the $t-1$'st period. It is usual that the interest paid over the $t-1$st period should be declared in advance, i.e. at time $t-1$ so that if $S(t)$ is adapted to a filtration $\mathcal{F}_t$, then $r(t)$ is *predictable*, i.e. is $\mathcal{F}_{t-1}-$measurable. The *discounted stock price process* is the process given by

$$S^*(t) = S(t)/B(t).$$

Consider a *trading strategy* of the form $(\beta(t), \alpha(t))$ representing the total number of shares of bonds and stocks respectively held at the beginning of the period $(t-1, t)$. Since our investment strategy must be determined by using only the present and the past values of this and related processes, both $\beta(t)$ and $\alpha(t)$ are predictable processes. Then the value of our investment at time $t-1$ is $V_{t-1} = \beta(t)B(t-1) + \alpha(t)S(t-1)$ and at the end of this period, this changes to $\beta(t)B(t) + \alpha(t)S(t)$ with the difference $\beta(t)\Delta B(t) + \alpha(t)\Delta S(t)$ representing the *gain* over this period. An investment strategy is *self-financing* if the value after rebalancing the portfolio is the value before- i.e. if all investments are paid for by the above gains. In other words if $V_t = \beta(t)B(t) + \alpha(t)S(t)$ for all t. An *arbitrage opportunity* is a trading strategy that makes money with no initial investment; i.e. one such that $V_0 = 0, \quad V_t \geq 0$ for all $t = 1, \ldots T$ and $E(V_T) > 0$. The basic theorem of no-arbitrage pricing is the following:

Theorem

There are no arbitrage opportunities in the above economy if and only if there is a measure Q equivalent to the underlying measure P i.e. $P << Q$ and $Q << P$ such that under Q the discounted process is a martingale; i.e. $E_Q(S^*(t)|\mathcal{F}_{t-1}] = S^*(t-1)$ a.s. for all $t \leq T$.

Proof; See Pliska (3.19)) page 94.

Note: The measure Q is called the equivalent martingale measure and is used to price derivative securities. For any attainable contingent claim X; (a for any random variable X which can be written as a linear function of the available investments), the arbitrage-free price at time t is given by the conditional expected value under Q of the discounted return X given $\mathcal{F}_\tau$.

Chapter 9

Appendix B: Stochastic Integration and Continuous Time Models

The single most important continuous time process in the construction of financial models is the Brownian motion process. A Brownian motion is the oldest continuous time model used in finance and goes back to Bachelier around the turn of the last century. It is also the most common building block for more sophisticated continuous time models called diffusion processes.

The Brownian motion process is a random continuous time process $W(t)$ defined for $t \geq 0$ such that $W(0)$ takes some predetermined value, usually 0, and for each $0 \leq s < t$, $W(t) - W(s)$ has a normal distribution with mean $\mu(t-s)$ and variance $\sigma^2(t-s)$. The parameters μ and σ are the drift and the diffusion parameters of the Brownian motion and the special case $\mu = 0, \sigma = 1$, $W(t)$ is often referred to as a standard Brownian motion or a Wiener process. Further properties of the Brownian motion process that are important are:

A Brownian motion process exists such that the sample paths are each continuous functions (with probability one)

The joint distribution of any finite number of increments $W(t_2) - W(t_1), W(t_4) - W(t_3), ...W(t_k) - W(t_{k-1})$ are independent normal random variables for $0 \leq t_1 \leq t_2 ... \leq t_k$.

Further properties can be derived from these. For example suppose we consider a sum of squared increments of the form $\sum_{i=1}^{k-1}(W(t_{i+1}) - W(t_i))^2$ for for $0 = t_1 \leq t_2 ... \leq t_k = t$ If we allow the number of increments k to go to infinity and the mesh size $\max(t_{i+1} - t_i)$ to go to zero, then it is easy to show that the limit of this random sum of squares is a totally non-random value t,

this limit taken in probability. A sum of squares process defined in this way is

The Stochastic Exponential and logarithm.

It is natural to define the exponential of a process

$$\mathcal{E}(X) = \prod(1 + dX) = \exp\{X - \frac{1}{2} < X^c >\} \prod(1 + \Delta X)e^{-\Delta X}$$

Then if $Y = \mathcal{E}(X)$, we have

$$dY = Y_- dX$$

and therefore we define the stochastic logarithm

$$\mathcal{L}(Y) = \int \frac{1}{Y_-} dY$$

9.1 Ordinary Differential Equations

Consider a stochastic differential equation of the form **??** in the special case that the drift term is linear $a(X_t) = \alpha + \beta X_t$. Most of the standard models for interest rates, for example, take this form, including the CIR, Vasicek, Geometric Brownian motion.

Suppose we wish to determine the expected value of the process. We have seen that if we denote the expected value by $m(s) = E(X_s|X_0)$ then it satisfies the ordinary differential equation $m'(t) = \alpha + \beta m(t)$. This is an example of the simplest form of ordinary differential equation, one in which the derivatives are of order at most 1 and the coefficients are constant. Let us consider the general first order differential equation

$$\frac{dy}{dt} + p(t)y = g(t).$$

These are solved by introducing an integrating factor $\mu(t)$ satisfying

$$\mu(t)[\frac{dy}{dt} + p(t)y] = [\mu(t)y]'.$$

In order that μ satisfy this, we require $\mu p = \mu'$ or

$$\mu(t) = exp\{\int^t p(s)ds\}.$$

Then $(\mu y)' = \mu g$ resulting in the general solution

$$y(t) = \frac{1}{\mu(t)}[\int^t \mu(s)g(s)ds + c].$$

MAPLE can be used to solve differential equations such as this one. For example we would first define the differential equation (here we substituted a, b for α, β) and then request a solution; (:> is the MAPLE prompt)

```
> deq := diff(y(x),x) - a - b*y(x) = 0;
> dsolve(deq, y(x));
```

We might have wished to specify an initial condition in the above equation. Suppose we wish to specify $y(0) = c$

```
> dsolve(deq,y(0)=c,y(x));
```

and we have the solution

$$y(x) \ = \ -a/b \ + \ exp(bx)[\ \frac{a}{b} \ + \ c] \ . \tag{1.7}$$

Consider as an example the Cox-Ingersoll-Ross model, here written with a slightly different specification of parameters. If r_t denotes the spot interest rate at time t ,

$$dr_t \ = \ (\ \alpha + \ \beta r_t)\ dt \ + \ \sigma r_t^{1/2}\ dW_t \tag{1.8}$$

If we let $m(t)$ denote the mean, taking expectations on both sides gives

$$m'(t) = (\alpha + \beta m(t))$$

and therefore the solution is given by (1.7) with $a,\ b$ replaced by α, β.

Non-linear ordinary differential equations are usually somewhat more difficult to solve. These are equations of the form

$$\frac{dy}{dx} = f(x, y)$$

where $f(x, y)$ is not a linear function of y. For example if the function $f(x, y)$ is a quadratic function of y, the equation is called *Ricatti's equation.* Suppose this equation is homogeneous, and can be written in the form

$$M(x, y)dx + N(x, y)dy = 0. \tag{9.1}$$

There is sometimes a function $\mu(x, y)$ called an *integrating factor* satisfying

$$\frac{\partial}{\partial y}(\mu M) = \frac{\partial}{\partial x}(\mu M)$$

and in this case the homogeneous equation can be solved explicitly. Note that after multiplication by μ, the equation

$$\mu M dx + \mu N dy = 0$$

is exact in the sense that it is obtained as the differential of an equation of the form $\psi(x, y) = c$. When no explicit solution to a differential equation can be obtained, we may either approximate the solution numerically or obtain a power series expansion of the solution, to as many terms as are wished.

For example, consider the differential equation

$$\frac{d^2}{dt^2}y(t) + 5\frac{d}{dt}y(t) + 6y(t) = 0$$

This is solved in MAPLE with

```
> de1 := diff(y(t),t$2) + 5*diff(y(t),t) + 6*y(t) = 0;
> dsolve(de1, y(t));
```

yielding:

$$y(t) = C1exp(-3t) + C2exp(-2t)$$

and although in this case an analytic solution is available, we might for a more difficult differential equation wish a series expansion, as obtained by

```
> dsolve(de1, y(t), series);
```

Suppose the initial condition is $y'(0) = c$.

This yields the series expansion of the solution

$$y(t) = y(0) + ct + \left(-3\,y(0) - \frac{5\,c}{2}\right)\,t^2 + \left(5\,y(0) + \frac{19\,c}{6}\right)t^3$$

$$+ \left(-\frac{19\,y(0)}{4} - \frac{65\,c}{24}\right)t^4 + \left(\frac{13\,y(0)}{4} + \frac{211\,c}{120}\right)t^5 + O\left(t^6\right)$$

Problem:

Solve the ordinary differential equation

$$3t^2y''(t) - 2ty'(t) + 2y(t) = 0$$

and sketch the possible solutions.

Problem: (Finding the stationary distribution of an Ito process).

Consider an Ito process of the form:

$$dX_t = a(X_t)dt + \sigma(X_t)dW_t. \tag{9.2}$$

Suppose there is a stationary density $\pi(x)$ satisfying

$$\int p(s,z,t,x)\pi(z)dz = \pi(x) \text{ for all } s < t, \text{ and all } x.$$

Multiply Kolmogorov's forward differential equation (1.3) by $\pi(z)$ and integrate over z. Thus show that the stationary distribution $\pi(x)$ must satisfy a differential equation of the form

$$\frac{d^2}{dx^2}(\sigma^2(x)\pi(x)) = 2\frac{d}{dx}(a(x)\pi(x))$$

Solve this differential equation to obtain the form of the stationary distribution assuming that the diffusion models a process (like interest rate, or asset price) that is positive, so that $\pi(0) = \pi'(0) = 0$.

9.2 Systems of Ordinary Differential Equations.

Consider now a system of ordinary differential equations of the form

$$\frac{d}{dt}\mathbf{y}(t) = -\mathbf{p}(t)\mathbf{y}(t) + \mathbf{g}(t) \tag{1.9}$$

where $\mathbf{y}(\mathbf{t})$ is an n-dimensional column vector of functions $y_i(t)$and $\mathbf{p}(t)$ is a $n \times n$ matrix of functions and $\mathbf{g}$ is an n-dimensional column vector of functions. Then the solution is exactly analogous to the one-dimensional case. First, for a square matrix A we define the exponential

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{1.10}$$

Then the integrating factor μ is defined by integrating componentwise: μ(t) = e $^{\int_{t_0}^{t} \mathbf{p}(s)\ ds}$.

Note that this is a matrix. The solution is then given by $y(t) = [\,\mu(t)]^{-1}\,[\,\int_{t_0}^{t} \mu(s)\, g(s)\ ds\,]$. Maple also permits the solution of systems of differential equations. For example consider the second order differential equation $y''(x) = y(x)$. This is equivalent to the system $y'(x) = z(x), \quad z'(x) = y(x)$. Solve as follows:

```
> sys := diff(y(x),x)=z(x), diff(z(x),x)=y(x)
> fcns := y(x), z(x)
> dsolve(sys,y(0)=0,z(0)=1,fcns);y(x)=1/2 exp(x -1/2 exp(-x);z(x)
=1/2 exp(x) + 1/2 exp(- x)
```

Problem:

Find the general solution to the following system of first order differential equations:

$$\begin{aligned} g'(x) &= -\sqrt{f(x)}, \\ h''(x) &= \frac{g(x)}{f(x)} \\ f'(x) &= \exp\{-f(x)\} \end{aligned}$$

9.3 Partial Differential Equations

Many of the pricing formulae encountered in finance can be derived as solutions to one or more partial differential equations, including the most important, the Black-Scholes formula. In general, this is because the most common models for the underlying asset are diffusion models. Maple provides some facility for the solution of simple partial differential equations. For example:

:>PDE := x*diff(f(x,y),y)-y*diff(f(x,y),x) = 0;

:>pdsolve(PDE);

provides the solution $f(x, y) = F1(x + y)$ for arbitrary function $F1$.

Before discussing methods of solution in general, we develop the Black-Scholes equation in a general context. Suppose that a security price satisfies

$$dS_t \;=\; a(S_t,\; t\;)\; dt \;+\; \sigma(S_t,\; t)\; dW_t \tag{1.11}$$

Our assumed market allows investment in the stock as well as in discount bonds, whose price at time t is β_t. There are various other assumptions as well; for example partial shares may be purchased, there are no dividends paid and no commissions, and no possibility of default for the bonds. Since bonds are assumed risk-free, they satisfy an equation

$$d\beta_t = r_t \beta_t dt$$

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.

A quick reminder of one of the most important single results of the twentieth century in finance and in science. This single mathematical result underlies the research leading to 1997 Nobel Prize to Merton and Black for their work on hedging in financial models.

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*

$$dV = [a(S_t,\ t\ )\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t,\ t)}{2}\ \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]dt + \ \sigma(S_t,\ t)\frac{\partial V}{\partial S}dW_t. \qquad (1.12)$$

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

$$V(S_{t+h},t+h) \approx V(S_t,t) + \frac{\partial V}{\partial S}(S_{t+h}-S_t) + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}(S_{t+h}-S_t)^2 + \frac{\partial V}{\partial t}h \qquad (1.13)$$

where we have ignored remainder terms that are $o(h)$. Note that substituting from (1.11) into (1.13), 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 **??**.

By Ito's lemma, provided V is smooth, it also satisfies a diffusion equation of the form **??**. 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 at the exercise price. Fortunately, such exceptional points can be worked around in the argument. 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 $V(S_t,t)$? Suppose such a linear combination is $u_tS_t + w_t\beta_t$ where the predictable functions $u_t,\ w_t$ represent the number of shares of stock and bonds respectively owned at time t. The portfolio is assumed to be self-financing and this requires that all returns obtain from the changes in the value of the securities and bonds held, i.e. it is assumed that $dV = u_tdS_t + w_td_t.\ Substituting from (1.11),$

$$dV = u_tdS_t + w_td\beta_t = [u_ta(S_t,t) + w_tr_t\beta_t]dt + u_t\sigma(S_t,t)dW_t \qquad (1.14)$$

It follows on comparing the coefficients of dt and dW_t in **??** and (1.14), that $u_t = \frac{\partial V}{\partial S}$ called the *delta* corresponding to *delta hedging*. Consequently,

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

and solving for w_t we obtain:

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

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 Delta 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 (said to exist only in a Japanese garden), is possible. The equation we obtained by equating both coefficients in **??** and (1.14) is of the form;

$$-r_t V + r_t S_t \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} + \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{1.15}$$

The negative of the first two terms $r_t(V - S_t \frac{\partial V}{\partial S})$ represents the amount made by the portion of our portfolio devoted to risk-free bonds. The last two terms represents the return on a hedged portfolio long one option and short delta stocks. 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 the remarkable feature of the arbitrage-free theory. Essentially, no matter what the drift term for the particular security is, in order to avoid arbitrage, all securities are priced as if they had drift the spot interest rate. 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

$$-rV + \frac{\partial V}{\partial t} + (r - D_0) S \frac{\partial V}{\partial S} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{1.16}$$

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.

We should also note that 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

$$\frac{\partial}{\partial t}u = k\frac{\partial^2}{\partial x^2}u \tag{1.17}$$

A solution of **??** 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 **??** of the form $u(x,t) = g(x/\sqrt{t})$. By differentiating and substituting above, we obtain an ordinary differential equation of the form

$$g''(\omega) + \frac{1}{2k}\omega g'(\omega) = 0, \omega = x/\sqrt{t} \tag{1.18}$$

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

$$\frac{\partial}{\partial w}g(\omega) = C_2 \exp\{-w^2/4k\} = C_2 \exp\{-x^2/4kt\} \tag{1.19}$$

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

$$u(x,t) = C_1 + C_2 \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^{x} \exp\{-z^2/4kt\}dz \tag{1.20}$$

is a solution of this, the heat equation for $t > 0$. The role of the two constants is simple. Clearly if a solution to **??** 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 **??** 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

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

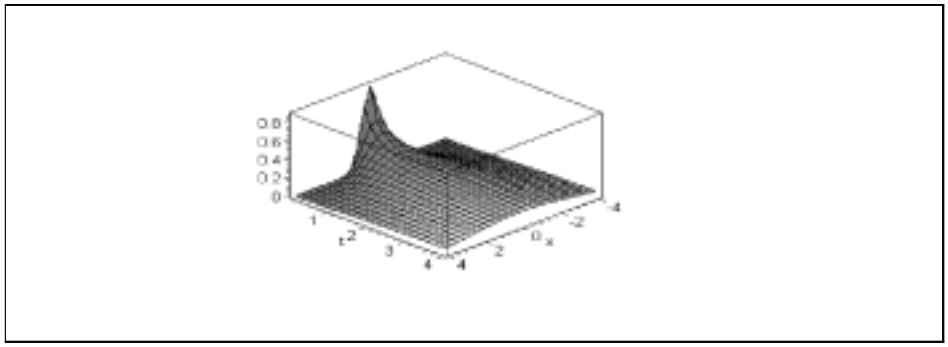


Figure 9.1:

Let us consider a basic solution to **??**:

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-x^2/4kt\} \tag{1.21}$$

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:

```
> 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 1.1: u(x,t)

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 (1.20) 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 **??**. 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 **??** subject to the required boundary condition.

Problem

Use separation of variables to solve the heat equation **??** on $0 < x < 1, t > 0$ subject to initial condition $u(x,0) = u_0(x)$ and $u(0,t) = u(1,t) = 0, t > 0$.

We may also solve the heat equation with given initial/boundary conditions using the *Laplace Transforms.* The problem is to solve **??** with $k = 1$ subject to

$$u(x,0) = u_0(x), u(x,t) \text{ bounded.}$$

Define the Laplace transform with respect to the variable t to be $U(x,s) = \int_0^{\infty} u(x,t) \exp\{-st) dt$. We suppose that we may differentiate twice under the integral sign so that **??** implies

$$sU(x,s) - u_0(x) = \frac{\partial^2}{\partial x^2} U(x,s)$$

which can be solved as an ordinary differential equation for fixed s. The solution is

$$U(x,s) = \frac{1}{2\sqrt{s}} \int_{-\infty}^{\infty} \exp\{-\sqrt{s}|x-y|\} u_0(y) dy$$

The solution (1.22) to the heat equation can no be found by inverting this Laplace transform

Chapter 10

Appendix: Numerical Solutions of DE's and PDE's

10.0.1 Difference and Differential Operators and solving ODE's.

Normally, differential and partial differential equations are solved numerically by replacing the derivatives by differences. Recall that a Taylor series approximation to a function f takes the form

$$f(x+h) = f(x) + ah + \frac{b}{2}h^2 + O(h^3)$$

where $a = f'(x)$ and $b = f''(x)$ are the first and second derivative of the function at x respectively.

If we wish to estimate the first derivative a we might use the forward (first) difference approximation $\frac{f(x+h)-f(x)}{h} = a+\frac{b}{2}h+O(h^2)$ or the backward analogy $\frac{f(x)-f(x-h)}{h} = a - \frac{b}{2}h + O(h^2)$ but note that the approximation based on symmetric first differences appears better;

$$\frac{f(x+h)-f(x-h)}{2h} = a + O(h^2).$$

Similarly, in approximating the second derivative b we can use the second difference

$$\frac{f(x+h)-2f(x)+f(x-h)}{h^2} = b + O(h)$$

These approximations, generally based on central differences rather than forward or backward differences form the basis of the more precise numerical solutions to differential and partial differential equations.

Many of these approximations are easily obtained using difference and differential operator notation, designed to permit easy access to various formulae for approximating derivatives. To begin with, suppose we are interested in approximating the derivative of a function $f(x)$ of a single variable. Denote by Df the derivative function f', so

$$D^0 f(x) = f(x),\ (Df)(x) = f'(x), \quad (D^2 f)(x) = f'' ,$$

etc. Then for h small, by a Maclaurin's series expansion,

$$f(x+h) = \sum_{i=0}^{\infty} \frac{h^i D^i}{i!} f(x) = e^{hD} f(x)$$

where the second equality is really the definition of the operator e^{hD}. Thus the forward difference $\Delta f(x) = f(x+h) - f(x)$ can be written in operator notation $\Delta f(x) = e^{hD} f(x) - f(x)$ or symbolically $\Delta = e^{hD} - 1$ and

$$D = \frac{1}{h} log(1+\Delta) = \frac{1}{h}(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \dots)$$

Similarly

$$D^2 = \frac{1}{h^2}(\Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \dots).$$

The first term in these expansions provides a simple estimator of the first and second derivative respectively. Thus, for example, the estimator $\frac{\Delta^2}{h^2}$ provides a simple approximation to the second derivative. The next term in the series provides an indication of the order of the error. We have seen that central differences tend to provide more accurate estimates of derivatives. Introducing the notation for central differences $\delta f(x) = f(x+h/2) - f(x-h/2)$, note that

$$\delta = (1+\Delta)^{1/2} - (1+\Delta)^{-1/2} \ = (e^{hD/2} - e^{-hD/2}) = 2sinh(hD)$$

Expanding the inverse of the hyperbolic sine in a series,

$$D = \frac{1}{h} sinh^{-1}(\delta/2) = \frac{1}{h}(\delta - \frac{1}{24}\delta^3 + \dots)$$

and similarly

$$D^2 = \frac{1}{h^2}(\delta^2 - \frac{1}{12}\delta^4 + \dots).$$

The more rapid convergence of the estimators using central differences rather than forward or backward differences is apparent. Thus, the estimator of first derivative $\frac{\delta}{h}$ has error $O(h^2)$ and the estimator of second derivative

$$\frac{\delta^2}{h^2} f = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

has error $O(h^2)$.

Let us now see how these difference approximations to derivatives can be used to numerically solve an ordinary differential equation. The first and simplest method is the Euler method, but there is a hugh number of more sophisticated alternatives including the trapezoidal rule, or the improved Euler method, the modified Euler method, and the Runge-Kutta method. Consider the simplest, the forward Euler method and the simple differential equation of second order

$$\frac{d^2}{dt^2}y(t) = 100e^{-10t} + 100e^{10t} \tag{2.1}$$

with initial conditions $y(0) = 2, y'(0) = 0$. Note that if we replace the second derivative by a second difference, the equation becomes

$$\frac{1}{h^2}\Delta^2 y(t) = 100e^{-10t} + 100e^{10t}, t = 2h, 3h, ... \tag{2.2}$$

allowing us to approximate values of the function on the lattice of points of the form $kh, k = 2, 3, ...$ using the initial values. The values corresponding to $k = 0, 1$ result from the initial conditions.

Problem.

Solve the equation (2.1) with initial conditions $y(0) = 2, y'(0) = 0$ numerically using (2.2) recursively and $h = 0.1$.

Use the simplest Euler approximation for the initial conditions by replacing $y'(0) = 0$ by the condition $\Delta y(0) = 0$. Compare with the exact solution.

Problem.

Solve the equation (2.1) with initial conditions $y(0) = 2, y'(0) = 0$ numerically using the recursion

$$\frac{1}{h^2}\delta^2 y(t) = 100e^{-10t} + 100e^{10t}, t = 2h, 3h, ... \tag{2.3}$$

and $h = 0.1$. Compare with the exact solution.

The improved Euler method is a simple modification of the Euler method for solving an equation of the form $y'(t) = f(t, y)$. The problem with Euler's method is it uses as slope the derivative at one end only of the interval and an improved approximation of the derivative using the average slope of at the two endpoints is usually more accurate. In this case, we make a preliminary estimate of $y((k+1)h)$ denoted by $y^* = hf(kh, y(kh)) + y(kh)$ and then solve for $y((k+1)h)$ the equations

$$y((k+1)h) = \frac{f(kh, y(kh)) + f((k+1)h, y^*)}{2}h + f(kh, y(kh)), k = 1, 2, ...$$

The Runge-Kutta method differs from the above two methods only in the way in which the slope of the line is estimated. For example we may take an average of the slope $f(t,y)$ at points $t = kh, (k+1/2)h, (k+1)h$ and *various* corresponding values of y in the interval. In fact it is possible to select 4 combinations of values (t,y) so that the approximation is perfect when $y(t)$ is a polynomial of degree 4. In general. we may select k points so that the approximation is perfect for a polynomial of degree k. This is a general description of the Runge-Kutta method.

$y((k+1)h) =$ (weighted average of values of $f(t,y)$ in the interval)$h + f(kh, y(kh)), k = 1, 2, ...$

Problem:

Use Euler's method and the improved Euler's method to solve the equation $y'(t) = ty(t)$ using step size $h = 0.1$ and 0.05 and initial condition $y(0) = 1$. Compare the solution with the exact solution on the interval $0 < t < 1$. How does the error change as we (a) increase h, (b) let t get farther away from the initial condition $t = 0$.

Example: Numerical Methods for ODE's in MAPLE:

Consider the second order differential equation

$$\frac{d^2}{dt^2}y(t) = ty(t)$$

with initial conditions $y(0) = 0, y'(0) = 1$.. We wish a numerical solution. There are a number of methods available in MAPLE including classical, lsode, mgear, rfk45, taylorseries,dverk78. Each has a number of options. The simplest method is classical[foreuler] for forward euler.

> deq := diff(y(t), t$2) = y(t)*t;

> ans:=dsolve(deq,y(t),numeric,method=classical[foreuler],
initial=array([0,1]), start=0);

We may then plot the result for $0 < t < 2$ as follows;

> with(plots);

> odeplot(ans,[t,y(t)],0..2);

Compare with an alternate more accurate method;

> ans2 := dsolve(deq3, y(t), numeric, method=mgear[msteppart], initial=array([2,0]), start=0);

> odeplot(ans2,[t,y(t)],0..2, axes=boxed);

> ans(2);

[t = 2, y(t) = 3.588338680829067, y'(t) = 4.643239714056109]

> ans2(2);

[t = 2, y(t) = 3.611076600977132, y'(t) = 4.676276660728110]

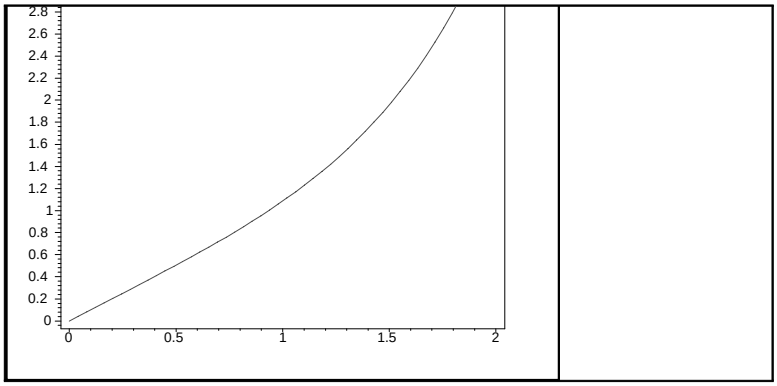


Figure 10.1:

Stability:

Instability is the possible accumulation of relatively small errors over time, and results in errors that get larger as $t \to \infty$. A stable numerical method is forgiving of small errors over time, while an unstable method may exaggerate these errors until they are the dominant part of the function. For an example of instability, consider the ordinary differential equation

$$\frac{dy}{dt} = y - t$$

with solution given by $y = (y_0 - 1)e^t + (t + 1)$ where the initial value is $y(0) = y_0$. Suppose that true value of the y_0 is 1 but it has been approximated numerically by $1 + \epsilon$ with ϵ small. The error might, for example, be the error due to expressing the initial condition with floating point arithmetic. For partial differential equations, often the initial condition is also approximated using a lattice of values, and this results in initial error. Note that even if the computer were able to obtain the exact analytic solution to the differential equation with the erroneous initial value, the difference between the true solution $t + 1$ and the approximation is ϵe^t (assuming no subsequent errors). This grows without bound as $t \to \infty$.

10.0.2 Numerical Methods for P.D.E.'s. Explicit Finite Difference Method.

We now return to the heat conduction or diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{2.4}$$

subject to the initial condition $u(x, 0) = u_0(x)$. A simple approach to solving this equation numerically is to approximate the first derivative on the left and

the second derivative on the right using finite differences and solve the resulting system of equations for the approximate value of the function on grid points. Suppose, for example we replace the derivative on the left side by a forward difference and the second derivative on the right side by a central second difference, This allows us to solve for the values of $u(x,t+\Delta t)$ in terms of the values of $u(x,t)$for all $x = n\Delta x$ and $t = m\Delta t,\ m = 0,1,2,\ldots.$ Denoting $U_{i,j} = u(i\Delta x, j\Delta t)$the equation becomes

$$U_{i,j+1} - U_{i,j} = \frac{\Delta t}{(\Delta x)^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})$$

or

$$U_{i,j+1} = rU_{i+1,j} + (1-2r)U_{i,j} + rU_{i-1,j} \tag{2.5}$$

where $r = \frac{\Delta t}{(\Delta x)^2}$. It turns out that this strategy of solving for $U_{i,j+1}$ in terms of its predecessors is not stable, i.e. the cumulative error does not converge, unless $r < 1/2$. This imposes a lower limit on the possible size of Δx, viz. $(\Delta x)^2 > 2\Delta t$. This method is called the *explicit finite difference method*, not because it is uncensored (although this is also true) but because we are able to solve explicitly for the values at time step $j+1$ using the values at time step j.

Consider the stability of an equation of the form (2.5). Let $U^*_{i,j}$ be the solution to (2.5) beginning with the exact initial conditions and $U_{i,j}$ the values obtained solving (2.5) if we begin with the (slightly) erroneous values $U_{i,0}$. Then the error due to the error in the initial condition is $E_{i,j} = U^*_{i,j} - U_{i,j}$. It is easy to see that this, since it is a linear combination of terms satisfying (2.5), also satisfies the equation (2.5). Suppose, for example, $E_{i,j} = \epsilon\lambda^j sin(i\omega)$ for each $i,\ j$ and for some real λ and frequency ω and for (small) ϵ. In general it is possible to show that the general solution of (2.5) is a linear combination of such terms with different values of λ, ω. Substituting in (2.5) and solving,

$$\lambda = 1 + r\frac{sin((i+1)\omega) - 2sin(i\omega) + sin((i-1)\omega)}{sin(i\omega)}.$$

Since $sin(A+B) = sin(A)cos(B) + cos(A)sin(B)$ this results in

$$\lambda = 1 + 2r[cos\omega - 1] = 1 - 4rsin^2(\omega/2)$$

Note that if $r > 1/2$, it is possible that for some frequencies ω, the corresponding value of $\lambda < -1$. This results in λ^j blowing up in magnitude as $j \to \infty$, i.e. the absolute value of at least some of the errors will go to infinity as $j \to \infty$. We have a similar instability if $r < 0$. However, if $0 < r < 1/2$ the errors will all converge to 0 as $j \to \infty$ and the solution to (2.5) is stable. For the explicit method to be stable, we need $\Delta t < \frac{1}{2}(\Delta x)^2$.

Problem:

What is the general form for the solution of the difference equation $y_{k+2} - 3y_{k+1} + 2y_k = (1/2)^k, k = 1,2,...$?

(Hint: solve as you would a differential equation: first, find the roots of the characteristic equation and hence all solutions to the homogeneous difference equation. Then find a particular solution to the homogeneous equation.)

Problem:

Consider solving the partial difference equation (2.5) by separation of variables. i.e. assume a solution exists of the form $U_{i,j} = d_i\phi_j$ for sequences d_i, ϕ_j. Show that the general solution is of the form $\phi_j = \epsilon\lambda^j$ for some λ and $d_i = C_1 \sin(i\omega) + C_2 \cos(i\omega)$ for some C_1, C_2, ω.

The next problem shows that if we were to replace the left side of (2.4) by a central difference that should be more accurate, we nevertheless obtain a method which is unstable for all values of r.

Problem:

We have already seen that *central differences* are generally more precise estimates of derivatives than are forward or backward differences. Suppose we replace (2.2) using a central difference estimator of $\frac{\partial u}{\partial t}$. Show that we obtain the partial difference equation

$$U_{i,j+1} - U_{i,j-1} = \frac{2\Delta t}{\Delta x^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) \tag{2.6}$$

The *Fourier* method approaches the stability of this difference equation by mapping the changes in the two directions into the complex plane. For example, suppose we consider a solution which is a constant multiple of $U_{i,j} = z^i q^j$ for some complex $z = e^{\beta\Delta x\sqrt{-1}}$ and real $q = e^{\alpha\Delta t}$, where α, β are arbitrary real numbers. General solutions could be obtained by taking linear combinations of such terms for different values of α and β and then taking the real or imaginary part of the linear combination. Substituting in (2.6), obtain the equation

$$q - \frac{1}{q} = 4r[cos(\beta\Delta x) - 1] = -8rsin^2(\beta\Delta x/2).$$

Show that even for r close to zero but positive, there is a solution for q with $|q| > 1$ leading to an exploding solution to the difference equation as $j \to \infty$. This indicates instability in the difference equation (2.6) making it undesirable for any value of $r = \frac{\Delta t}{\Delta x^2}$.

For the explicit finite-difference method it can be proven that if we Δt and Δx tend to zero in such a way that the ratio $r = \frac{\Delta t}{(\Delta x)^2}$ remains between 0 and $\frac{1}{2}$ then the finite difference approximation converges to the actual solution. To see this we can consider the difference between the exact solution and the finite-difference approximation:

$$D_{n,m} = U_{n,m} - V_{n,m},$$

where $V_{n.m}$ is the solution of the explicit finite-difference method. From the Taylor's theorem applied to the forward and the central difference approximation we can find that

$$D_{n.m} = (1-2r)D_{n,m} + r(D_{n+1,m} + D_{n-1,m}) + \Delta t(R_1\Delta t + R_2(\Delta x)^2),$$

where R_1 and R_2 are two bounded in absolute value functions. For $\hat{D}^m = \max_n |D_{n,m}|$ and $\hat{D}^{m+1} = \max_n |D_{n,m+1}|$, the largest errors at time-step m and $m+1$ respectively, we get

$$\hat{D}^{m+1} \leq (|1-2r| + 2|r|)\hat{D}^m + \Delta t(R_1\Delta t + R_2(\Delta x)^2).$$

Provided that $0 \leq r \leq \frac{1}{2}$, we have $|1-2r| + 2|r| = 1$, and hence

$$\hat{D}^{m+1} \leq \hat{D}^m + \Delta t(R_1\Delta t + R_2(\Delta x)^2).$$

By induction it follows

$$\hat{D}^{m+1} \leq \hat{D}^0 + (m+1)\Delta t(R_1\Delta t + R_2(\Delta x)^2),$$

so if we assume zero error at time-step $m = 0$, which we can do since $V_{n,0} = U_{n,0}$ from the initial condition, we see that

$$\hat{D}^{m+1} \leq (m+1)\Delta t(R_1\Delta t + R_2(\Delta x)^2) \to 0 \quad \text{as} \quad \Delta t \to 0,$$

which proves that the method gives an approximation which converges to the actual solution. A modification to this argument shows that if $r > \frac{1}{2}$ the error actually grows without bound as we let $\Delta t \to 0$. Therefore, for explicit finite-difference method for the diffusion equation the stability and the convergence problems are equivalent.

The moral of this story is that one should be careful with discrete approximations to continuous differential equations. The stability of the method and the convergence of the numerical solution are among the many concerns. Various methods have been proposed that solve some of the stability problems encountered above, and these are implemented in many packages. We discuss these methods further in the following sections.

10.0.3 Implicit Finite Difference Solutions to the Diffusion Equation.

The implicit finite difference scheme for solving (2.4) is similar to the explicit method in that we use first differences for the derivative on the left hand side of (2.4) but we use the backward difference instead. This results in a system of linear equations in $u(x,t)$ in terms of the values of $u(x, t-\Delta t)$for all $x = n\Delta x$ and $t = m\Delta t$, $m = 0, 1, 2, \ldots$, and these equations fortunately take a simple form.

$$U_{i,j} - U_{i,j-1} = r(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) \tag{2.7}$$

or

$$-rU_{i+1,j} + (1+2r)U_{i,j} - rU_{i-1,j} = U_{i,j-1}$$

where $r = \Delta t/(\Delta x)^2$. Boundary conditions determine the value of $U_{i,j}$on the upper and lower boundary of a rectangle, e.g. for $i = \pm N$, and all j, while initial conditions prescribe the value of the function at time 0.

The implicit method requires solving this system of equations at time j in terms of the solution at time $j-1$. There are three ways in which solutions to systems such as this can be approached. The first is matrix inversion. For large matrices this is very difficult. The second method uses an LU decomposition of the matrix into two components, L , a lower triangular matrix, and U an upper triangular matrix. The advantage of triangular matrices is that systems of equations of the form $Lx = y$, for example, can be easily solved by simple substitution. A third method, called successive over relaxation (SOR) is a method for solving the equation

$$MU_{*,j} = b_j$$

for a matrix M and vector b_j. Here $U_{*,j}$ denotes the column vector $(U_{-N,j},\ U_{-N+1,j}, \ldots ,\ U_{N,j})$, b depends on $U_{*,j-1}$ and the boundary conditions and

$$M = \begin{pmatrix} 1+2r & -r & 0 & . & . & 0 \\ -r & 1+2r & -r & 0 & . & 0 \\ 0 & -r & 1+2r & -r & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -r & 1+2r \end{pmatrix}.$$

This is then rewritten

$$U_{i,j} = \frac{1}{1+2r}[b_{i,j} + r(U_{i-1,j} + U_{i+1,j})] \tag{2.8}$$

and then starting with an initial estimate of the solution (always for fixed j), substitute in the left to obtain an updated estimate on the right. This continues until there is little change in the components of $U_{*,j}$. There is one simple modification of this that is normally applied in practice, motivated by the fact that sequences often converge in such a way that the errors decrease like a geometric series. When this is the case, convergence may be accelerated by over-relaxation (extrapolating the solution). This is most easily described with an example. Consider the sequence $x_n = 0.1 + 0.9x_{n-1} = 0.1000, 0.1900, 0.2710, 0.3439, 0.4095, 0.4686, 0.5217, \ldots$ if we begin with $x_1 = 0.1$. Clearly this sequence converges rather slowly to 1. Consider the sequence defined by $y_n = 0.1 + 0.9x_{n-1}$, and $x_n = x_{n-1} + \omega(y_n - x_{n-1})$ for some over-relaxation parameter $\omega > 1$. How much faster does the sequence converge if $\omega = 5$? The first few terms are$0.1, 0.595, 0.7975, 0.8987$ which is clearly a substantial improvement. Of course $\omega \approx 9$ provides even faster convergence.

Problem (SOR):

Consider the sequence

$$x_n = \frac{x_{n-1}}{2} + \frac{1}{x_{n-1}}$$

$x_1 = 10$. Investigate the speed of convergence of this sequence and one obtained by successive over-relaxation. For what value of the parameter ω does the speed of convergence seem to be greatest?

The SOR method consists of first solving

$$Y_{i,j}^{k+1} = \frac{1}{1+2r}[b_{i,j} + r(U_{i-1,j}^{k+1} + U_{i+1,j}^{k})]$$

and then putting

$$U_{i,j}^{k+1} = U_{i,j}^{k} + \omega(Y_{i,j}^{k+1} - U_{i,j}^{k}), k = 1, 2, 3, ...$$

until convergence.

10.0.4 The Crank-Nicolson Method

This method differs from both the forward and backward approach of the explicit and fully implicit methods primarily in that the first forward or backward difference is now replaced by a symmetric first difference, improving the accuracy to $O(\Delta t)^2$. Note that

$$\frac{\partial u}{\partial t}(x, t + \Delta t/2) \ = \frac{u(x, t+\Delta t) - u(x,t)}{\Delta t} + O(\Delta t)^2.$$

Similarly, we can obtain an approximation to the right hand side accurate to the same order by averaging the symmetric second difference at points t and $t + \Delta t$, i.e. using

$$\frac{u(x+\Delta x, t) - 2u(x,t) + u(x-\Delta x, t)}{2(\Delta x)^2}$$

$$+\frac{u(x+\Delta x, t+\Delta t) - 2u(x,t+\Delta t) + u(x-\Delta x, t+\Delta t)}{2(\Delta x)^2}$$

Setting these equal results in a system of linear equations in the values of $u(x, t+\Delta t)$ solvable in terms of the values of $u(x,t)$ for all x. These equations take the form;

$$U_{i,j+1} - U_{i,j} = \frac{r}{2}\{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1} + U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\}$$

The system of equations that result from the Crank-Nicolson approach is similar to the form shown of the implicit method

$$MU_{*,j+1} = BU_{*,j} + b_{*,j}$$

where

$$M = \begin{pmatrix} 1+r & -r/2 & 0 & . & . & 0 \\ -r/2 & 1+r & -r/2 & 0 & . & 0 \\ 0 & -r/2 & 1+r & -r/2 & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -r/2 & 1+r \end{pmatrix},$$

$$B = \begin{pmatrix} 1-r & r/2 & 0 & . & . & 0 \\ r/2 & 1-r & r/2 & 0 & . & 0 \\ 0 & r/2 & 1-r & r/2 & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & r/2 & 1-r \end{pmatrix},$$

and $b_{*,j}$ is a vector of zeros in the center, with non-zero elements at the boundary which depend on the boundary conditions.

As we will show in the next section, the Crank-Nicolson method has convergence and stability properties similar to that of the implicit method, the method is stable if $r > 0$.

10.0.5 Stability and Consistency of the Crank-Nicolson.

We now consider the stability of the Crank-Nicolson method using a matrix approach. Consider once again the numerical solution to the heat equation

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

on $0 < x < 1$ subject to the initial condition $u(x,0) = f(x)$ and some boundary conditions $u(1,t) = g_1(t),\ \ u(0,t) = g_2(t)$ for all t . We have seen that methods such as the Crank-Nicolson involve discretizing in both the time and space directions, and then writing a recursive formula of the form

$$MU_{*,j+1} = BU_{*,j} + b_{*,j}$$

where the matrices M and B are defined in the last section . Note that

$$U_{*,j+1} = PU_{*,j} + M^{-1}b_{*,j} = \ldots$$

$$= P^{j+1}U_{*,0} + P^jM^{-1}b_{*,0} \ + \ P^{j-1}M^{-1}b_{*,1} + \ldots + M^{-1}b_{*,j} \qquad (2.9)$$

where P is the matrix $M^{-1}B$. Let the eigenvalues of the matrix P be $\lambda_1, \lambda_2, \ldots$ and the corresponding normalized eigenvectors $v_1, v_2, \ldots$. Recall from linear algebra that P^j has the same eigenvectors with eigenvalues λ_n^j. Then provided that these eigenvectors span the space, we can write $U_{*,0} = \sum_i \alpha_i v_i$ for some constant coefficients α_i, and then

$$P^j U_{*,0} = \sum_i \alpha_i \lambda_i^j v_i. \tag{2.10}$$

Similarly, each of the other terms on the right hand side of (2.9) can also be written in a similar form with coefficients involving powers of the eigenvalues λ_i^j. For stability with respect to small errors in the initial conditions, we would like to be assured that as j increases, the vector on the right hand side of (2.10) remains bounded. This is clearly the case if all of the eigenvalues λ_i are less than or equal to 1. A similar argument implies stability with respect to small errors in the initial conditions. Thus the condition for matrix stability is more generally, for a recursion of the form $MU_{*,j+1} = BU_{*,j} + b_{*,j}$, that the maximum eigenvalue of P , i.e. the maximum root of the equation $det(M - \lambda B) = 0$ is less than one in absolute value. In the case of the Crank Nicolson method, this results in the same condition as does the Fourier method, i.e. that $0 < r$.

Besides stability, which ensures that errors do not tend to grow without bound, we generally require some indication that the discrete approximation to a partial DE is close to the continuous solution. Let the difference operator F correspond to the discrete Crank-Nicolson method.

$$FU_{i,j} = \frac{U_{i,j+1} - U_{i,j}}{\Delta t} - \frac{\{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1} + U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\}}{2(\Delta x)^2}$$

Then the difference between the discrete Crank-Nicolson and the corresponding differential operator can be expanded, as was done in section 2.1, to obtain

$$FU_{i,j} - (\frac{\partial}{\partial t}u - \frac{\partial^2}{\partial x^2}u) = \frac{1}{2}(\Delta t)\frac{\partial}{\partial t}[\frac{\partial}{\partial t}u - \frac{\partial^2}{\partial x^2}u] - \frac{1}{12}(\Delta x)^2\frac{\partial^4 u}{\partial x^4} \tag{2.11}$$

$$+\frac{1}{6}(\Delta t)^2\left(\frac{\partial^3 u}{\partial t^3} - \frac{3}{2}\frac{\partial^4 u}{\partial x^2 \partial t^2}\right) + \ldots$$

Note that as long as both Δt and Δx approach 0, the error in the approximation on the right side of (2.11) also approaches 0 (assuming sufficient smoothness of the solution). Thus the numerical solution is consistent under these conditions.

10.0.6 The Method of Lines.

In general, the method of lines is a device which reduces a partial differential equation to a coupled system of ordinary differential equations. Consider for an example the non-linear equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u + (\frac{\partial}{\partial x}u)^2$$

Suppose we use symmetric first and second order differences for the derivatives on the right side of this equation, but continue to leave the left side as a derivative. Then at a given point (x,t) the equation takes the form

$$\frac{\partial}{\partial t}u(x,t) = \frac{1}{h^2}[u(x+h,t) - 2u(x,t) + u(x-h,t)] + \frac{1}{4h^2}[u(x+h,t) - u(x-h,t)]^2.$$

If we now denote $U_i(t) = u(ih,t), \quad i = 0,1,2,\ldots$, then this becomes a coupled system of first order differential equations

$$U_i'(t) = \frac{1}{h^2}[U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)] + \frac{1}{4h^2}[U_{i+1}(t) - U_{i-1}(t)]^2, \quad i = 1,\ 2,\ldots$$

which, together with the appropriate initial or boundary conditions, may be solved, for example in MAPLE, to obtain an approximate solution to the PDE.

10.0.7 Finite Elements and the Galerkin Method.

Much of the theory of ordinary and partial differential equations parallels corresponding results in linear algebra. We begin with some elementary results concerning linear operators. Consider a vector space spanned by a complete set of vectors $v_1, v_2, \ldots$. By this we mean that any element of the vector space can be written as a limit of a linear combination of the spanning vectors $lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i v_i$. Suppose the vector space has an inner product denoted by (u,v). A complete inner product space is called a Hilbert Space. In the case of a finite-dimensional vector space, this inner product is usually the dot product between the two coefficient vectors. However, we will deal here with an infinite dimensional vector space in which the vectors are functions. Now let A be a positive linear operator on the vector space. This means that when v s a vector, so is Av and it is linear; $A(v_1+v_2) = Av_1 + Av_2$. Of course we usually represent linear operators in the finite dimensional case by matrices, but when the vectors consist of functions, then objects like derivatives of the function, first differences, etc. are also linear operators. A linear operator is positive definite if it is symmetric ($(Av,u) = (v,Au)$ for all u,v) and $(Av,v) \geq C||v||^2$ for all v and some $C > 0$.

We now consider a particular minimization problem, easiest to interpret and prove in the finite dimensional case, but useful in the infinite dimensional problems in differential equations as well. To motivate the result, pretend for the moment that the vectors f,u,v are finite-dimensional column vectors and A is a positive definite matrix. Suppose we wish to minimize the length of the residuals $Av - f$ over possible vectors v restricted to some linear subspace of Euclidean space. The notion of distance is adapted to the positive definite matrix A, so that the squared distance is given by $(Av-f)^T A^{-1}(Av-f) = (Av,v) - 2(f,v) + (f,f)$. Since we are minimizing over v, it is equivalent to minimize the quantity $(Av,v) - 2(f,v)$ and this minimization is equivalent to minimizing the "length" of the residuals $Av - f$.

Proposition

Suppose A is a linear operator on a given domain D_A in a Hilbert space. Suppose we wish to minimize the quadratic functional

$$(Av, v) - 2(f, v) \tag{2.12}$$

over all vectors v for a given function $f \in D_A$. Then the following are equivalent

(a) v^* is the desired minimum.

(b) v^* is the unique solution to the equation $Av^* = f$.

(c) Provided D_A is spanned by base vectors $v_1, \ldots v_n$ scaled to have length 1, then $v^* = \sum a_i v_i$ where $a_1, a_2, \ldots a_n$ solves the system of equations

$$\sum_i (Av_i, v_j)a_i = (f, v_j), \quad j = 1, 2, \ldots n \tag{2.13}$$

The last equation above can be written in simpler form and applies generally, even in circumstances in which the operator is not positive definite. Suppose we wish to solve $Av^* = f$. Suppose there are finitely many basis vectors $v_1, \ldots v_n$ that we think "almost" span the desired space. Then any proposed solution that is a linear combination of these vectors, say $v_n^* = \sum_{i=1}^n a_i v_i$ must be assessed through the size of the residuals $Av_n^* - f$. As in regression, we have achieved the projection of the solution onto this finite dimensional subspace only if these residuals are perpendicular to each of the basis vectors, i.e. if $(Av_n^* - f, v_j) = 0, \quad j = 1, 2, \ldots n$. This is the system of equations (2.13).

Now consider as a simple example the heat equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u, \quad 0 < x < 1, \quad 0 < t$$

subject to initial and boundary conditions

$$u(x, 0) = 1, \quad 0 < x < 1$$

$$u(0, t) - \frac{\partial}{\partial x}u(0, t) = 0, \quad t > 0$$

$$\frac{\partial}{\partial x}u(1, t) = 0, \quad t > 0$$

We begin with a base of functions of x that satisfy the boundary conditions. For example, the polynomials

$$v_j(x) = (1 + x) - \frac{x^{j+1}}{j + 1}$$

all satisfy the boundary conditions

$$v_j(0) - v_j'(0) = 0, \qquad v_j'(1) = 0.$$

Now consider approximating a solution in the form

$$\bar{u}(x,t) = a_1(t)v_1(x) + a_2(t)v_2(x)$$

In this example, the space of functions should include (possibly infinite) linear combinations of twice differentiable functions of x which satisfy the boundary conditions, with coefficients that are differentiable functions of t. The inner product is $(u,v) = \int_0^1 u(x,t)v(x,t)dx$. The residual at time t is

$$A\bar{u} - f = \frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}, \quad t > 0$$

and

$$1 - \bar{u}(x,0), \quad t = 0.$$

Then the conditions (2.13) lead to the fact that the residuals are orthogonal to the two basis vectors, i.e. that

$$\int_0^1 (1 - \bar{u}(x,0))v_j(x)dx = 0, \quad j = 1,\ 2$$

$$\int_0^1 [\frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}]v_j(x)dx = 0, \quad j = 1,\ 2 \tag{10.1}$$

These equations reduce to

$$\frac{9}{5}a_1(0) + \frac{691}{360}a_2(0) = \frac{4}{3}$$

$$\frac{691}{360}a_1(0) + \frac{1291}{630}a_2(0) = \frac{17}{12}$$

and

$$\frac{9}{5}a_1'(t) + \frac{691}{360}a_2'(t) + \frac{4}{3}a_1(t) + \frac{17}{12}a_2(t) = 0$$

$$\frac{691}{360}a_1'(t) + \frac{1291}{630}a_2'(t) + \frac{17}{12}a_1(t) + \frac{23}{15}a_2(t) = 0$$

and the solution to this system of equations is given by

$$a_1(t) = 0.586e^{-.742t} + 2.448e^{-11.770t}, \quad a_2 = 0.144e^{-.742t} - 2.295e^{-11.770t}$$

and so the approximate solution to the initial-boundary value problem is

$$\bar{u}(x,t) = 0.586e^{-.742t} + 2.448e^{-11.770t}(1 + x - x^2/2)$$

$$+0.144e^{-.742t} - 2.295e^{-11.770t}(1 + x - x^3/3).$$

Of course, because the second exponent is so large and negative, it is clear that the first term in these coefficients dominates for moderate or large values of t.

Example.

The methods we consider here are generally applicable to numerically solving an equation of the form

$$\frac{\partial}{\partial t}u + A(u) = 0$$

where $A(u)$ is usually required to be a positive definite linear differential operator with respect to x. An example of a positive definite operator is one of the form $A(u) = -\frac{\partial}{\partial x}[a(x,t)\frac{\partial}{\partial x}u(x,t)]$ for a non-negative function a. It is easy to show, for example, that the operator $-\frac{\partial^2}{\partial x^2}$ is positive definite on a suitable subspace determined by the boundary conditions. We return to solving equation (2.13) on $0 < x < \pi$, $0 < t < 1$ under the boundary, initial conditions $u(x,0) = 1$, $0 < x < \pi$, and $u(0,t) = u(\pi,t) = 0$. The method of separation of variables leads to solutions of the form $u_n(x,t) = e^{-n^2t}sin(nx)$ and it is easy to see in this case that if we try a linear combination of the form $u(x,t) = \sum_n a_n u_n$ the equation and boundary conditions are satisfied for $a_n = \frac{4}{\pi n}$, $n = 1,3,5,\ldots$, and otherwise $a_n = 0$. Therefore in this case, an explicit solution to the equation is known;

$$u(x,t) = \sum_{n=1,3,5,\ldots} \frac{4}{\pi n}e^{-n^2t}sin(nx).$$

Suppose for the moment we did not know this solution. We could attempt a solution as a linear combination of finitely many of the basis vectors $v_n(x) = \frac{sin(nx)}{n}$, $n = 0,1,2,\ldots 5$ so the attempted solution might take the form $\bar{u}(x,t) = a_0(t) + \sum_{n=1}^{5} a_n(t)v_n(x)$. Proceeding as in the last example, if we require that the residuals are orthogonal to the basis vectors, we obtain the equations

$$a_0(t) = 1$$

$$\int_0^\pi [\frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}]v_j(x)dx = 0, \quad j = 0,1,\ldots 5 \tag{2.14}$$

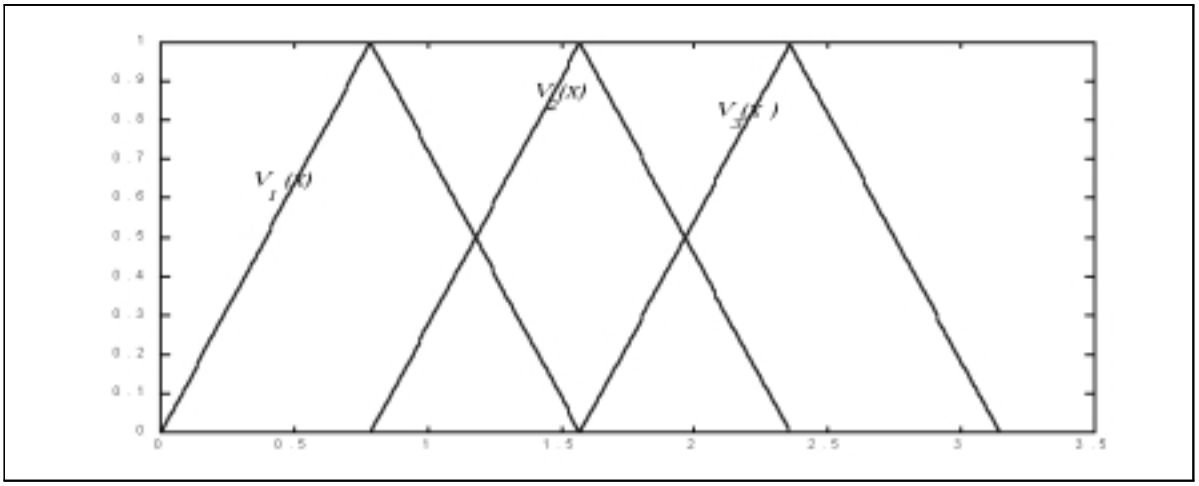


Figure 10.2:

which reduces a coupled system of first order differential equations that may be solved for the coefficients $a_j(t)$. Often the system of equations (2.14) is simplified further by discretizing time, i.e. replacing $\frac{\partial}{\partial t}\bar{u}$ by a first forward difference for small (time) step size h . Then the equations are simply linear equations, with no derivatives involved, and they are solved sequentially in $t = jh,\ j = 1, 2,$.

There is an alternative for the base functions v_n commonly used and often referred to as a finite element method. Suppose, for example, we use the three simple spline functions represented by the triangles below.

Clearly these functions are continous,

and indeed have piecewise continuous first derivatives. For example, $v_j(x)$ can be recovered by integrating its first derivative. Unfortunately, the second derivatives do not have this property and if we express $\bar{u}$ as a linear combination of these base functions, we still not be able to apply the operator A to it, since this requires second derivatives with respect to x. This seemingly harmless failure has led to a mountain of mathematical literature providing a weak interpretation of differential equations. Roughly speaking, functions are defined by the result upon integration after multiplication by well-behaved test functions. For example, the dirac delta function could be defined either as a measure, or as the "function"δ providing $\int \delta(x)\phi(x)dx = \phi(0)$ for all smooth functions ϕ. Under a weak interpretation, in order to solve for the coefficients above, we require integrals of the form $\int[-\frac{\partial^2}{\partial x^2}v_n(x)]v_j(x)dx$ and this can be interpreted using integration by parts. For example if we are integrating over a strip $a < x < b$ and if the basis vectors $v_j(x)$ are zero at $x = a, x = b$ (because

boundary conditions imply that the solution is also zero on this boundary),

$$\int [-\frac{\partial^2}{\partial x^2} v_n(x)] v_j(x) dx = \int [\frac{\partial}{\partial x} v_n(x)][\frac{\partial}{\partial x} v_j(x)] dx \tag{10.2}$$

even when the second derivative fails to exist. Notice that this identity also shows that the linear operator $A = -\frac{\partial^2}{\partial x^2}$ satisfies $(Av.v) \geq 0$ and is therefore positive definite.

Summary

We have seen several alternative methods that can be used to numerically solve a system of the form

$$\frac{\partial}{\partial t} u + A(u) = f$$

for positive definite operator A. We may discretize both time and space, using a method such as Crank-Nicolson. We may discretize space and not time, resulting in a system of first order differential equations in t. Alternatively, we may discretize time. Putting $u_j(x) = u(j\Delta t, x)$ this requires solving a system of the form

$$\frac{u_j(x) - u_{j-1}(x)}{\Delta t} + A u_j(x) = 0$$

with the inherited boundary conditions. This may result in a system of second order differential equations in x which may be solved analytically or numerically (e.g. by finite element methods).

10.0.8 Solution of the Diffusion Equation.

In this section we consider the general solution to the diffusion equation of the form (1.15), rewritten as

$$\frac{\partial V}{\partial t} = r_t V - r_t S_t \frac{\partial V}{\partial S} - \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} \tag{2.15}$$

where S_t is an asset price driven by a diffusion equation

$$dS_t = a(S_t, t)dt + \sigma(S_t, t)dW_t, \tag{2.16}$$

$V(S_t, t)$ 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

$$V(S_T, T) = V_0(S_T). \tag{2.17}$$

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 (1.2) that if a related process X_t satisfies the stochastic differential equation

$$dX_t = r(X_t, t)X_t dt + \sigma(X_t, t)dW_t \tag{2.18}$$

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 (2.15);

$$\frac{\partial p}{\partial t} = -r(s, t)s\frac{\partial p}{\partial s} - \frac{\sigma^2(s, t)}{2}\frac{\partial^2 p}{\partial s^2} \tag{2.19}$$

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

$$f(t, s, T, z) = p(t, s, T, z)q(t, T) \tag{2.20}$$

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 (2.20) with weight function $w(z)$. i.e. try a solution of the form

$$V(s, t) = \int p(t, s, T, z)q(t, T)w(z)dz \tag{2.21}$$

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:

$$V(t, s) = E[w(X_T)q(t, T)|X_t = s] \tag{2.22}$$

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 **??**, we choose $w(x) = V_0(x)$. Now

$$\begin{aligned}
\frac{\partial V}{\partial t} &= \frac{\partial}{\partial t}\int p(t, s, T, z)q(t, T)w(z)dz \\
&= \int [-r(S_t, t)S_t\frac{\partial p}{\partial s} - \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 p}{\partial s^2}]q(t, T)w(z)dz \\
&\quad + r(S_t, t)\int p(t, S_t, T, z)q(t, T)w(z)dz \text{ by (2.19)} \\
&= -r(S_t, t)S_t\frac{\partial V}{\partial S} - \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 V}{\partial S^2} + r(S_t, t)V(S_t, t)
\end{aligned}$$

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

$$V(t,s) = E[V_0(X_T)q(t,T)|X_t = s] \tag{2.23}$$

satisfies both the partial differential equation (2.15) and the terminal conditions **??** 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 (2.18). This result is a special case when the spot interest rates are functions only of time of the following more general theorem.

Theorem(Feynman-Kac)

Suppose the conditions for a unique solution to (2.15,2.17) (see for example Duffie, appendix E) are satisfied. Then the general solution to (2.15) under the terminal condition **??** is given by

$$V(S,t) = E[V_0(X_T)exp\{-\int_t^T r(X_v,v)dv\} \mid X_t = S] \tag{2.24}$$

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 (**??**) 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

$$X_T = \frac{X_t}{q(t,T)} + \int_t^T \frac{\sigma(u)}{q(u,T)} dW_u \tag{2.25}$$

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 $\int_t^T \frac{\sigma^2(u)}{q^2(u,T)}du$.

Problem.

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.

Problem.

Give an example of a function $g(t, W_t)$ such that the random variable $\int_0^1 g(t, W_t)dW_t$ does not have a normal distribution but has larger tails than the normal distribution has.

The special case of (**??**) 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

$$g(x|\eta, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(log\ x - log\ \eta\ - \sigma^2/2)^2/2\sigma^2\}. \tag{2.26}$$

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

$$X_T = X_t exp\{\int_t^T (r(u) - \sigma^2(u)/2)du + \int_t^T \sigma(u)dW_u\}. \tag{2.27}$$

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 (**??**). 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

$$e^{\mu+\sigma^2/2}\Phi(\frac{\mu}{\sigma} + \sigma) - \Phi(\frac{\mu}{\sigma}) \tag{2.28}$$

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

$$V(S,t) = S\Phi(d_1) - Ee^{-r(T-t)}\Phi(d_2) \tag{2.29}$$

where $d_1 < d_2$ are the values $\pm\sigma^2/2$ 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

$$e^{-r(T-t)} \int_0^\infty V_0(x)g(x|S(t)e^{r(T-t)},\ \sigma\sqrt{T-t})dx. \tag{2.30}$$

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

$$dS = a(S)dt + \sigma S dW_t \tag{2.31}$$

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.

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 lack-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 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.

Problem

Verify that for any pair of constants $a \neq 0$ and $b > 0$

$$dX_t = (X_t^{-1} + ab)X_t dt + bX_t dW_t$$

does not have a solution in the form $X_t = f(t, Y_t)$, where $f(t, y)$ is, say, a real function and Y_t is a Gaussian process.

10.0.9 Black-Scholes with Transaction Costs.

We now modify the argument in section 2.8 to accommodate transaction costs. As in Leland (1985), Hoggard et al. (1993), we assume delta hedging and the transaction costs in each time interval is a constant proportion $k/2$ of the value of the trades in that interval. Suppose we have, at time t, u_t units of the security, and bank deposits or bonds, earning constant interest rate r, to a total value of B_t. Then the value of the portfolio $V_t = u_t S_t + B_t$. Therefore the change in value over a small time interval is of the form

$$dV_t = u_t dS_t + rB_t dt - (k/2)S_t|du_t|$$

plus terms of smaller order. The last term represents the transaction costs over this time interval. Using Ito's lemma on $V(S,t)$ we obtain

$$dV = \frac{\partial V}{\partial S}dS + \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}\right]dt$$

Equating terms, $u_t = \frac{\partial V(S_t,t)}{\partial S}$ resulting in delta-hedging and

$$rB_t dt - (k/2)S_t|du_t| = \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}\right]dt \tag{2.32}$$

Applying Ito's lemma to evaluating du_t will show that

$$du_t = \frac{\partial^2 V}{\partial S^2}dS + smaller \;\; order \; terms$$

Therefore

$$(k/2)S_t|du_t| \approx (k/2)S_t|\frac{\partial^2 V}{\partial S^2}||dS|. \tag{2.33}$$

Now we regard this differential equation as an approximation to a discrete process for which dS represents the change in the process over a discrete time interval of length dt. In this case, since approximately, $dS \sim N(a(S_t)dt, \sigma^2 S_t^2 dt)$ it follows that $E|dS| = \sqrt{2/\pi}\sigma S_t\sqrt{dt}$ and since, by the law of large numbers, the sum of many such increments converges to the expected value, the term $|dS|$ may be replaced in (2.33) by $\sqrt{2/\pi}\sigma S_t\sqrt{dt}$. Therefore, substituting in (2.32), we obtain the equation

$$rB_t dt - (k/2)S_t|\frac{\partial^2 V}{\partial S^2}|\sqrt{2/\pi}\sigma S_t\sqrt{dt} = \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}\right]dt.$$

Collecting terms, and substituting $B_t = V_t - \frac{\partial V}{\partial S}S_t$, this reduces to

$$\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}(1 + A\; sgn(\Gamma))S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

where $\Gamma = \frac{\partial^2 V}{\partial S^2}$ and

$$A = \sqrt{\frac{2}{\pi}}\frac{k}{\sigma\sqrt{dt}}$$

is the so-called Leland number. This is exactly the same equation that was solved to give the Black-Scholes formula except that the volatility σ^2 is inflated (deflated) by the factor $(1 + A\; sgn(\Gamma)))$. In the case $0 < A < 1$, this equation has a solution for arbitrary payoff function V_0. In the case $A > 1$, if the payoff function is convex, then again the solution is given by Black-Scholes with inflated volatility. However, if the payoff function is non-convex, then the mathematical problem is ill-posed (see Avellanda and Paras (1994)).

10.0.10 Methods for American Options

The valuation of American options is what is known as a free boundary problem. Typically at each time t there is a value $S_f(t)$ which marks the boundary between two regions: to one side one should hold the option on the other side one should

exercise it. Since we don't know a priori $S_f(t)$, however, we cannot apply boundary conditions in the way we do for European options.

For an American put option, with value $P(S,t)$, the valuation problem can be written as a free boundary problem as follows. For each time t, we must divide the S axis into two distinct regions: the first, $0 \le D < S_f(t)$, is where early exercise is optimal and

$$P = E - S, \quad \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP < 0,$$

the second, $S_f(t) < S < \infty$, where early exercise is not optimal and

$$P > E - S, \quad \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP = 0.$$

The boundary conditions at $S = S_f(t)$ are that P and its slope are continuous:

$$P(S_f(t),t) = \max(E - S_f(t), 0), \quad \frac{\partial P}{\partial S}(S_f(t),t) = -1.$$

Chapter 11

Appendix C: Glossary.

- Annuity A uniform series of payments or receipts over a specified period of time
- Asset A Physical or intangible item of value to a company or individual
- Bankruptcy A legal proceeding of the disposal of assets of a business or individual to satisfy creditors' claims in full or in part and protecting the debtor from further action.
- Beta A regression coefficient indicating the rate at which a given stock changes for a unit changge in the market as a whole.
- Bond A financial instrument representing a form of corporate long-term debt issued to investors.
- Bond rating A published ranking of a bond developed by fianancial organisations to express relative soundness on a defined scale.
- Book value the recorded value of an asset or liability as reflected in financial statements
- Call provision a provission permitting the issuing company to redeem part or all of a bond or preferred stock at a date determined by the company.
- Capital the funds committed to an ienterprise in the form of onership equity and long-term financing.
- Capitalization The sum of all long-term sources of capital of a company. The difference between current liabilities and total assets.
- Cash flow The postive inflow or negative outflow of cash caused by an activity over a specifies period of time.

- Common stock (common shares) Securities representing a derect ownership interest in a corporation and a residual claim on its assets.
- Compounding The process of calculating the growing value of a sum of money over time caused by the reinvestment of earned interest.
- Coupon rate The stated interest rate specifies on the interest coupons attached to bonds and calculated as a percentage of their face value.
- Credit the recognised ability of an individual or corporation to assume indebtetdness with the prospect of servicing such debt.
- Debt (libability) An obligation to pay amounts due under specified terms.
- Default failure to make a payment on a debt obligation when due.
- Discounting the process of calculating the reduced value of a future sum of money in proportion to the opportunity of earning interest over that period.
- Diversification the process of investing in a number of unrelated or partially independent assets or activities to establish a more satisfactory portfolio and reduce the volatility.
- Dividend payout the ratio of the amount of dividends distributed to the aftertax earnings o a corporation.
- Dividend yield The ratio of the total amount of dividends payable on one share over a specified period to the current market price per share.
- Earnings the difference between revenue and costs and expenses for a specifies period.
- Earnings per share; the total aftertax earnings divided by the number of common shares outstanding.
- Equity the recorded ownership claim of all common and preferred shareholders as reflected on the balance sheet.
- Fair market value the price for an asset on which two rational parties with sufficient information would agree in the absence of other factors.
- Financial Model the representation in a mathematical model or computer program relating the effect of various input factors to some measures of performance.
- Fixed-income security any security which provides a constant stream of interest or dividend over its life.
- Foreign exchange exposure The potential loss from changes in the exchange rate with one or more foreign currencies.

- Hedge a strategy to neutralize the risk of an investment by engaging in offsetting contracts whereby potential gains and losses will tend to cancel.
- Junk bond any bond with risk characteristics higher than normal investment grade.
- Leverage Any transaction which magnifies a given effect.
- Liability an obligation to pay a specified amount or perform a specified service at specifed times.
- Liquidity The degree to which a company is readily able to meet its current obligations, or the ease with which a security can be bought or sold.
- Market value The value of and asset as determined in an uncontrained market of many buyers and sellers.
- Net present value The difference between the present values of cash inflows and cash outflows
- Nominal amount Any quantity not adjusted for changes in maket conditions, purchasing powers.
- Option A contractual opportunity to purchase or sell an asset or security at a predetermined price, without obligation to do so.
- Par value the nominal value established by the issuer of a security. For a bond, the issuing company will pay the par value on maturity.
- Perpetuity a series of level periodic payments or receiptss expected to last forever.
- Portfolio a set of diverse investments held by and individual or company.
- Preferred Stock a special class of capital stock that receives a form of preference over common stock in its claim on earnings and assets.
- Present Value the value today of a future sum or a series of sums calculated by discounting the future amounts.
- Principal the original amount of a loan or bond (also called face value) on which interest is based.
- Risk analysis a process of integrating risk into an analysis.
- Risk aversion a subjective unwillingness to accept a given level of risk, unless there is a trade-off for higher average return.
- Risk-free interest rate The assumed yield obtainable on a guaranteed security.

- Risk premium The increased return required for an investment to compensate the holder for the level of risk involved.

-