

# **Topics in Computational Mathematics**

**Notes for**

**Computational Mathematics (MA1611)  
Information Technology (AS1054)**

**Dr G Bowtell**

# Contents

<b>1</b>	<b>Curve Sketching</b>	<b>1</b>
1.1	Curve Sketching . . . . .	1
1.2	Increasing and Decreasing Function . . . . .	1
1.3	Stationary Points . . . . .	2
1.4	Classification of Stationary Points . . . . .	3
1.5	Point of Inflection - Definition and Comment . . . . .	4
1.6	Asymptotes . . . . .	5
<b>2</b>	<b>Root Finding</b>	<b>7</b>
2.1	Introduction . . . . .	7
2.2	Existence of solution of $f(x) = 0$ . . . . .	8
2.3	Iterative method to solve $f(x) = 0$ by rearrangement . . . . .	10
2.4	Iteration using Excel - Method 1 . . . . .	11
2.5	Newton's Method to solve $f(x) = 0$ . . . . .	12
2.6	Iteration using Excel - Method 2 . . . . .	14
2.7	Simultaneous Equations - linear and non-linear . . . . .	15
2.7.1	Linear simultaneous equations . . . . .	15
2.7.2	Matrix product and inverse using Excel . . . . .	18
2.7.3	Non-linear simultaneous equations . . . . .	20
<b>3</b>	<b>Financial Functions in Excel</b>	<b>27</b>
3.1	Introduction . . . . .	27
3.2	Geometric Progression . . . . .	27
3.3	Basic Compound Interest . . . . .	28
3.4	Basic Investment Problem . . . . .	29
3.5	Basic Financial Worksheet Functions in Excel . . . . .	31
3.6	Further Financial Worksheet Functions in Excel . . . . .	34
<b>4</b>	<b>Curve fitting - Interpolation and Extrapolation</b>	<b>39</b>
4.1	Introduction . . . . .	39
4.2	Linear Spline . . . . .	42
4.3	Cubic Spline - natural . . . . .	45
4.4	Linear Least Squares Fitting . . . . .	49
4.4.1	Linear Least Squares - Regression - Single Variable . . . . .	49
4.4.2	Least Squares Criteria for Straight Line Fitting . . . . .	50

4.4.3	Linear Least Squares - Multiple Regression - Many Variables . . .	52
4.4.4	Excel - Linear Regression . . . . .	54
4.4.5	Goodness of Fit . . . . .	59
4.5	Non-Linear Least Squares Fitting . . . . .	63
4.5.1	Polynomial trendline ( <i>y</i> depending on a single <i>x</i> variable only) . . .	63
4.5.2	Power, Exponential and Logarithmic trendlines . . . . .	64
4.5.3	General Fitting using Excel . . . . .	65
<b>5</b>	<b>Minitab - Descriptive Statistics</b>	<b>69</b>
5.1	Introduction . . . . .	69
5.2	Mean and Standard Deviation. . . . .	69
5.3	Sampling Problem . . . . .	70
5.4	Coefficient of Variation . . . . .	71
5.5	Standard Error of the Mean . . . . .	71
5.6	Quartiles and Median . . . . .	72
5.7	Box Plots . . . . .	73
5.8	Histogram . . . . .	75
5.9	Minitab . . . . .	76
5.9.1	Descriptive statistics . . . . .	77
5.9.2	Box Plots . . . . .	77
5.9.3	Histogram . . . . .	78
<b>6</b>	<b>Powers of Matrices - Markov Chains</b>	<b>80</b>
6.1	Introduction . . . . .	80
6.2	Conditional Probability . . . . .	80
6.3	Markov Chain . . . . .	82
<b>7</b>	<b>Google &amp; PageRank</b>	<b>86</b>
7.1	Introduction . . . . .	86
7.2	PageRank . . . . .	86
7.3	Random Surfer & PageRank . . . . .	87
7.4	Google . . . . .	92
<b>8</b>	<b>Polynomial Approximations</b>	<b>98</b>
8.1	Introduction . . . . .	98
8.2	Taylor's Polynomial Approximation . . . . .	99
8.2.1	Linear - tangent approximation . . . . .	100
8.2.2	Quadratic approximation . . . . .	100
8.2.3	General approximation . . . . .	101
8.2.4	Maclaurin's Expansion - Example . . . . .	102
8.2.5	The Error Term . . . . .	103
8.2.6	Convergence of $p_n(x)$ as $n$ tends to infinity . . . . .	106
8.2.7	D'Alembert's Ratio test . . . . .	108
8.3	Polynomial Interpolation . . . . .	109
8.3.1	Direct Method . . . . .	109
8.3.2	Lagrange's Method . . . . .	109

8.3.3	Divided Differences . . . . .	112
8.3.4	Forward Differences . . . . .	116
8.3.5	Newton-Gregory formula . . . . .	119
<b>9</b>	<b>Numerical Integration</b>	<b>121</b>
9.1	Introduction . . . . .	121
9.2	Trapezoidal method - linear approximation . . . . .	121
9.2.1	Geometrical approach . . . . .	122
9.2.2	Analytical approach: rule + error term . . . . .	122
9.2.3	Composite trapezoidal rule . . . . .	124
9.2.4	Degree of accuracy . . . . .	125
9.3	Simpson's rule - quadratic approximation . . . . .	126
9.3.1	Basic formula . . . . .	127
9.3.2	Error term . . . . .	128
9.3.3	Composite Simpson's rule . . . . .	128
9.3.4	Degree of accuracy . . . . .	130
<b>10</b>	<b>Simultaneous Equations</b>	<b>131</b>
10.1	Introduction . . . . .	131
10.2	Matrix representation - Row reduction . . . . .	131
10.3	Partial pivotal selection . . . . .	135
10.4	Existence and Uniqueness . . . . .	138
10.5	Geometrical representation . . . . .	140

## Preface for 2009–10

This course provides material for the Mathematical Science module Computational Mathematics (MA1611) and part of the Actuarial Science module Information Technology (AS1054). It has been taught until last year by Dr Graham Bowtell, who wrote these copious notes. In 2009, after around 40 years at City University, Dr Bowtell retired.

This year the course will essentially follow that of previous years, although there will be some changes which will become apparent as the year progresses. The worksheets used by Dr Bowtell may or may not be used this year. However, you may find them useful, and they are provided for you on my web page for this course. This can be found at

<http://www.staff.city.ac.uk/o.s.kerr/CompMaths>

where the material for his course can be found.

Dr O.S. Kerr

## Preface

This course looks simultaneously at mathematically based problems and associated pieces of software. In many instances the mathematics of a subject area is first developed and then the software used to solve specific problems. By this means you are able to extend your mathematical knowledge whilst at the same time learn how to use various pieces of software.

In detail the presentation of the course consists of two distinct parts. The theory is contained in the lectures which are for one session per week over twenty weeks. The use of the software and development of the mathematical ideas are dealt with in the computational mathematics computer laboratories. These laboratories are scheduled for two hours per fortnight for each Actuarial Science student and one hour per week for the Mathematics students.

During the the following pieces of software are considered:

### **Derive**

This is a package that carries out both algebraic and numerical procedures. For example amongst other things it is capable of solving equations, differentiating and integrating functions, manipulating matrices, calculating Taylor expansions and plotting graphs in two and three dimensions. Although the course starts with this piece of software and associated mathematical problems it will be used quite extensively over the course, especially for plotting graphs.

### **Excel**

Excel is used in this section of the module as well as in the VBA programming section. Here the course concentrates on the use of worksheet functions as well as looking at their mathematical background where appropriate.

### **Minitab**

This is a package that is specifically designed to handle data and carry out statistical analysis. From simple diagrams, which can also be created in Excel, to the more sophisticated hypothesis testing and regression analysis.

### **Google**

Google is a well known search engine, here we not only look at how to use it for specialist searching but also how Google actually carries out its search with emphasis on how a particular page is ranked above another. This particular aspect of the working of the Google search engine involves a topic known as Linear Algebra and Markov Chains. The course will consider the underlying mathematics in a fairly elementary way with the object of giving the user a better understanding of the search process.

### **Important Note**

Not all the material in these notes will be used on the course and as such the notes should be considered as a reference for the lectures and labs and not a replacement. It

is not likely that much or indeed any of the material from the final four chapters will be covered. Finally note that the course may also present material in a different order to which they appear in the notes.

# Chapter 1

## Curve Sketching

### 1.1 Curve Sketching

To sketch the curve of a given function  $y = f(x)$  the following features need to be identified.

- Intervals on which  $f(x)$  is increasing or decreasing
- Stationary points
- Local maxima and minima
- Points of inflection
- $x$  and  $y$  intercepts
- Asymptotes

### 1.2 Increasing and Decreasing Function

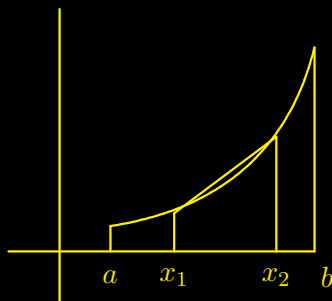


Figure 1.1:

With reference to fig 1.1 we make the following definition and deduction:



- $f(x)$  is an **increasing function** on the interval  $[a, b]$  if for all  $x_1$  and  $x_2 \in [a, b]$

$$x_2 > x_1 \Rightarrow f(x_2) > f(x_1)$$

- Clearly

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} > 0$$

which, since as this tends to  $f'(x_1)$ <sup>1</sup> as  $x_2 \rightarrow x_1$ , implies that  $f'(x_1) \geq 0$ . As this is true for any  $x_1$  in the interval  $[a, b]$  we can make the following statement:

$$f(x) \text{ increasing on } [a, b] \Rightarrow f'(x) \geq 0 \text{ on } [a, b]$$

$$f'(x) > 0 \text{ on } [a, b] \Rightarrow f(x) \text{ increasing on } [a, b]$$

Similarly

$$f(x) \text{ decreasing on } [a, b] \Rightarrow f'(x) \leq 0 \text{ on } [a, b]$$

$$f'(x) < 0 \text{ on } [a, b] \Rightarrow f(x) \text{ decreasing on } [a, b]$$

### 1.3 Stationary Points

We now consider the special points where  $f'(x)$  is zero.

- The point  $(x, f(x))$  is a **stationary point** of  $f(x)$  if  $f'(x) = 0$

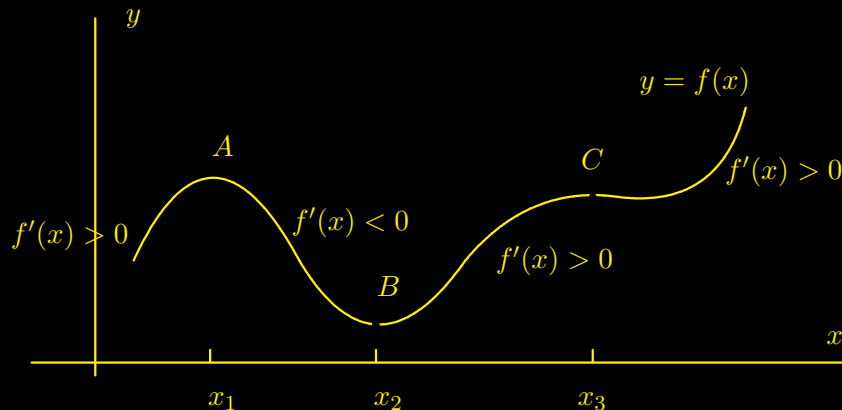


Figure 1.2:

In fig 1.2  $A, B$  and  $C$  are stationary points. We also note that at  $A$  and  $B$  the curve *turns*. This leads to the following definition:

<sup>1</sup>provided  $f$  is differentiable on  $[a, b]$

- A point about which  $f'(x)$  changes sign is called a **turning point**. Clearly at such a point  $f'(x) = 0$ , ie the point is stationary.<sup>2</sup>

It is also clear from fig 1.2 that in the locality of the two points  $A$  and  $B$ ,  $f(x)$  is maximal and minimal respectively. This leads to the following definition:

- $f(x)$  has a **local maximum** at  $x_1$  if for all  $x$  in the neighbourhood of  $x_1$   $f(x) < f(x_1)$
- $f(x)$  has a **local minimum** at  $x_2$  if for all  $x$  in the neighbourhood of  $x_2$   $f(x) > f(x_2)$

## 1.4 Classification of Stationary Points

Having located a stationary point we wish to classify it as, maximum, minimum or, as described below, point of inflection. In the diagram these are the three points  $A$ ,  $B$  and  $C$  respectively. The following criteria allows us to do this by considering either  $f'(x)$  either side of a stationary point or  $f''(x)$  at the stationary point.

- If to the left of the stationary point  $f'(x)$  is positive and to the right of the stationary point  $f'(x)$  is negative, as at the point  $A$  in fig 1.2, then the stationary point is a local maximum.
- An alternative test for a local maximum is to observe that at  $A$  in fig 1.2  $f'(x)$  varies from being positive to the left of  $x_1$  to being negative to the right of  $x_1$ , thus  $f'(x)$  is decreasing and hence its derivative is negative. ie  $f''(x) < 0$ . Thus we can say that<sup>3</sup> if:

$$f'(x_1) = 0 \text{ and } f''(x_1) < 0 \text{ then } f(x) \text{ has a local maximum at } x_1$$

- If immediately to the left of the stationary point  $f'(x)$  is negative and immediately to the right of the stationary point  $f'(x)$  is positive, as at the point  $B$  in fig 1.2, then the stationary point is a local minimum.
- An alternative test for a local minimum is to observe that at  $B$  in fig 1.2  $f'(x)$  varies from being negative to the left of  $x_2$  to being positive to the right of  $x_2$ , thus  $f'(x)$  is increasing and hence its derivative is positive. ie  $f''(x) > 0$ . Thus we can say that<sup>4</sup> if:

$$f'(x_2) = 0 \text{ and } f''(x_2) > 0 \text{ then } f(x) \text{ has a local minimum at } x_2$$

- If  $f'(x)$  has the same sign immediately to the left and right of the stationary point, as at the point  $C$  in fig 1.2, the stationary point is an example of a point of inflection. There is no simple second derivative test, the point of inflection is best identified by considering the sign of  $f'(x)$  either side of the stationary point. See below for the general definition of a point of inflection and further comment.

<sup>2</sup>provided  $f'$  is continuous at the given point

<sup>3</sup>provided  $f''$  is continuous at  $x_1$

<sup>4</sup>provided  $f''$  is continuous at  $x_2$

**Example 1.1**

Find the stationary points of  $f(x) = x^4 - 6x^2 + 8x$  and hence classify as local maxima, local minima or point of inflection.

$$f'(x) = 4x^3 - 12x + 8 = 0 \Rightarrow x = 1 \text{ or } x = -2.$$

$f''(x) = 12x^2 - 12$  thus  $f''(-2) = 36$  which is positive, hence  $x = -2$  is a local minimum

However  $f''(1) = 0$ , thus  $x = 1$  **may be** a point of inflection

Considering  $f'(x)$  either side of  $x = 1$ , for example at  $x = 0$  and  $x = 2$ , we see that  $f'(0) = 8$  and  $f'(2) = 28$ , the sign of  $f'(x)$  does not change at  $x = 1$  thus  $f(x)$  has a point of inflection at  $x = 1$ .

**1.5 Point of Inflection - Definition and Comment**

A point of inflection is any point on the curve (*the point does not necessarily have to be a stationary point*) where its second derivative changes sign. Formally we have<sup>5</sup>:

- If the sign of  $f''(x)$  immediately to the left of  $x = c$  is different to the sign of  $f''(x)$  immediately to the right of  $x = c$  then the curve  $y = f(x)$  has a point of inflection at  $x = c$ . Clearly since  $f''(x)$  is changing sign as  $x$  passes through  $c$  we can deduce that  $f''(c) = 0$
- The following point should be noted: if  $x = c$  is a point of inflection for  $f(x)$  then  $f''(c) = 0$ . However the converse is not true. ie  $f''(c) = 0$  does not imply that  $f(x)$  has a point of inflection at  $x = c$
- A simple second derivative test  $f''(x_3) = 0$  cannot be used to classify the point  $C$  in fig 1.2 as a point of inflection. However we can see that it satisfies the definition of a point of inflection as follows. To the left of  $x = x_3$  the first derivative  $f'(x)$  is decreasing to zero at  $x = x_3$  which means that  $f''(x) < 0$  to the left of  $x = x_3$ . To the right of  $x = x_3$  the first derivative  $f'(x)$  is increasing from zero at  $x = x_3$ , which means that  $f''(x) > 0$  to the right of  $x = x_3$ . Thus  $f''(x)$  changes sign at  $x = x_3$  and hence  $f(x)$  has a point of inflection at  $x = x_3$ . Thus for a stationary point where the second derivative vanishes we can check to see if its a point of inflection by looking at the sign of  $f''(x)$  either side of the point.
- $f''(x) = 0$  is not sufficient to establish a point of inflection:  
In the definition  $f''(x)$  needs to change sign for a point to be a point of inflection, and at such a point  $f''(x) = 0$ . However  $f''(x) = 0$  is not sufficient to guarantee a point of inflection. Consider  $y = x^4$ , clearly this has a minimum at  $x = 0$  since  $x^4 \geq 0$  for all  $x$ . However  $f'(0) = 0$  and  $f''(0) = 0$ . Thus  $(0, 0)$  is a stationary point at which  $f''(x) = 0$  but which is not a point of inflection.

---

<sup>5</sup>provided  $f''$  is continuous at  $x = c$

## 1.6 Asymptotes

A curve  $C$  is an asymptote to  $y = f(x)$  if as the point  $(x, f(x))$  moves away from the origin the graph of  $y = f(x)$  tends to  $C$ . Essentially we have two cases:

- (i)  $x = a$  is an asymptote to  $y = f(x)$  if as  $x \rightarrow a$ ,  $|f(x)| \rightarrow \infty$ .
- (ii) As  $x \rightarrow +\infty$  or as  $x \rightarrow -\infty$   $y$  tends to the curve  $C$   
*The curve  $C$  is quite often just a straight line but in general it can be any suitable curve*

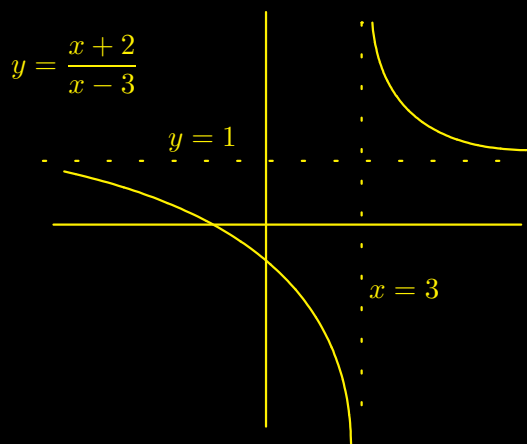


Figure 1.3:

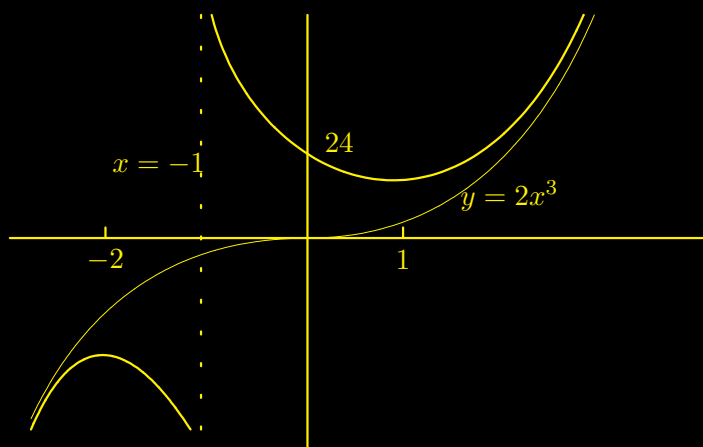


Figure 1.4:

### Example 1.2

**Simple asymptotes**

Consider  $y = \frac{x+2}{x-3} = f(x)$ . Clearly this function is not defined at  $x = 3$  however:

As  $x \rightarrow 3$  from above  $y \rightarrow +\infty$  and as  $x \rightarrow 3$  from below  $y \rightarrow -\infty$

Thus  $x = 3$  is a vertical asymptote. To consider  $y$  as  $x \rightarrow \infty$  we first rearrange  $f(x)$  as

$$y = \frac{1 + \frac{2}{x}}{1 - \frac{3}{x}}$$

Clearly now as  $x \rightarrow \pm\infty$   $y \rightarrow 1$ . Thus  $y = 1$  is an horizontal asymptote. See fig 1.3.

**Example 1.3** - (see Fig 1.4)

Sketch

$$y = \frac{2(x^4 + x^3 + 12)}{x + 1} = f(x).$$

**asymptotes**

Clearly this function is not defined at  $x = -1$  however:

As  $x \rightarrow -1$  from above  $y \rightarrow +\infty$  and as  $x \rightarrow -1$  from below  $y \rightarrow -\infty$ ,

since in both cases the numerator of  $f(x)$  is positive in the neighbourhood of  $x = -1$ . Thus  $x = -1$  is a vertical asymptote. We now ask are there any other asymptotes that are not necessarily straight lines. To see this we rearrange  $f(x)$  as follows:

$$y = 2x^3 + \frac{24}{x+1} \quad \text{clearly now as } x \rightarrow \pm\infty \quad y \rightarrow 2x^3$$

Thus  $y = 2x^3$  is an asymptote for  $y = f(x)$ .

**Stationary point**

Using the rearranged form of  $f(x)$  we see that

$$f'(x) = 6x^2 - \frac{24}{(x+1)^2} \quad \text{and} \quad f''(x) = 12x + \frac{48}{(x+1)^3}$$

Hence  $f'(x) = 0$  implies that  $x = 1$  or  $-2$ . Evaluating  $f''(x)$  at these points gives  $f''(1) > 0$  and  $f''(-2) < 0$ . Hence we deduce that there is a local minimum at  $x = 1$  and a local maximum at  $x = -2$ .

**Intercepts**

Putting  $x = 0$  into  $f(x)$  we see that the  $y$ -intercept is given by  $y = 24$ . The  $x$ -intercepts are given by the solution of  $y = x^4 + x^3 + 12 = 0$ . The solution of this problem is not trivial so will not be considered here.

## Chapter 2

# Root Finding

### 2.1 Introduction

Many problems require the solution of the equation  $f(x) = 0$ . Geometrically this is where the curve  $y = f(x)$  cuts the  $x$ -axis. For example we may wish to find all the solutions of  $x^2 - 30 \sin x = 0$ . Clearly  $x = 0$  is a solution however we need to determine how many other solutions there are and how to find them to a given degree of accuracy.

In general we may have several equations in several unknowns. The simplest example of this is a set of simultaneous linear equations which usually present very few problems, especially if there are only two or three variables. A more difficult problem to solve is when the simultaneous equations are not linear, for example consider solving the equations

$$f_1(x_1, x_2) = 2 - x_1^2 - x_2 = 0 \quad \text{and} \quad f_2(x_1, x_2) = 2x_1 - x_2^2 - 1 = 0 \quad (2.1)$$

By introducing a vector notation can write equations of the above type in a compact form as follows:

Set

$$\underline{x} = (x_1, x_2, \dots, x_N) \quad \underline{f}(\underline{x}) = (f_1(\underline{x}), f_2(\underline{x}) \dots f_N(\underline{x}))$$

Thus Eq(2.1) becomes

$$\underline{f}(\underline{x}) = \underline{0} \quad (2.2)$$

where the zero vector  $\underline{0} = (0, 0 \dots 0)$ .

In this chapter we will consider the general case of  $N$  equations in  $N$  unknowns. In all problems we ideally need to consider the following three questions.

- Does there exist a solution to  $\underline{f}(\underline{x}) = \underline{0}$ . *existence*
- Does there exist more than one solution to  $f(x) = 0$ . *uniqueness*
- Does the method give some idea of the accuracy of its result. ie a bound for the error.

## 2.2 Existence of solution of $f(x) = 0$

We consider the existence of a solution in the case of a single equation in one variable. If we consider the problem  $f(x) = x^2 - 30 \sin x = 0$  and tabulate the value of  $f(x)$  for  $x$  between -4 and +4 we obtain the following:

$x =$	-4	-3	-2	-1	0	1	2	3	4
$f(x) = x^2 - 30 \sin x$	-6.7	13.2	31.3	26.2	0	-24.2	-23.3	4.8	38.7

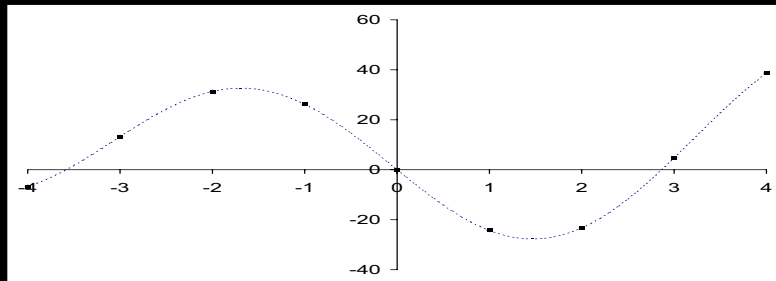


Figure 2.1: Plot of tabulated points and sketch of curve between points.

From the table and Fig 2.1 we see that  $f(-4) < 0$  and  $f(-3) > 0$  which suggest that in between -4 and -3 there exists at least one value of  $x$  at which  $f(x)$  is zero. Similarly we note that  $f(2) < 0$  and  $f(3) > 0$  which suggests there exists at least one solution in the interval  $[2, 3]$ . A more thorough investigation of the function will yield that there is another solution in the interval  $[-6, -5]$ . Finally we note that  $x = 0$  is a solution.

Formally we have the following theorem that provides us with a set of sufficient conditions to guarantee a solution of  $f(x) = 0$  in a given interval.

### Theorem 2.1

If  $f(x)$  is continuous in the interval  $a \leq x \leq b$  and  $f(a) \times f(b) < 0$  then there exists at least one solution to  $f(x) = 0$  in the interval  $[a, b]$ .

The theorem guarantees at least one solution, however as we see in Fig 2.2 there may be more than one solution in the interval  $[a, b]$ . Here the curves cuts the  $x$ -axis three times in the interval, thus indicating that there are three solutions in the interval.

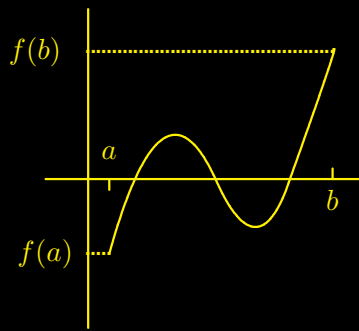


Figure 2.2:

### Notes on the Theorem

- The above requires that  $f(x)$  is continuous on  $[a, b]$ . The following example illustrates why this condition is needed. Let  $f(x) = \frac{x^2}{x-2}$  clearly  $f(1) = -1 < 0$  and  $f(3) = 9 > 0$  but as we see from fig 2.3 there is no solution in  $(1, 3)$ . Thus we see that even though  $f(1) \times f(3) < 0$  there is no solution between  $x = 1$  and  $x = 3$ . The Theorem cannot be used since  $f(x)$  is clearly not continuous at  $x = 2$ , a point in the interval  $(1, 3)$ .

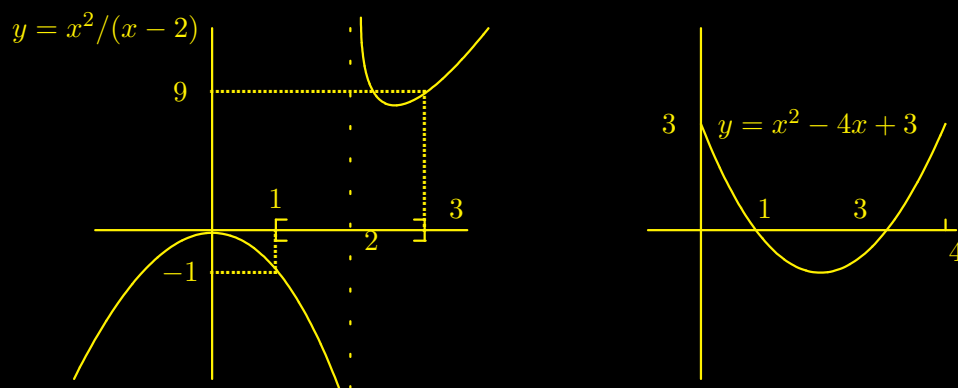


Figure 2.3:

- If  $f(x)$  is continuous on  $[a, b]$  and  $f(a) \times f(b) > 0$  then there may still be a solution in  $[a, b]$ . For example the second graph in fig 2.3 shows  $f(x) = x^2 - 4x + 3$  on the interval  $[0, 4]$ . Clearly  $f(0) \times f(4) > 0$  and  $x = 1$  and  $x = 3$  are solutions of  $f(x) = 0$



## 2.3 Iterative method to solve $f(x) = 0$ by rearrangement

The iterative method for solving  $f(x) = 0$ , or more fully the fixed point iterative method, can be summarised by the following three steps:

- To solve  $f(x) = 0$  we rewrite this equation in the form  $x = g(x)$ . Thus the roots of  $f(x) = 0$  are the same as any value of  $x$  satisfying  $x = g(x)$ .
- To obtain the solution of  $x = g(x)$  we take an initial first guess  $x = x_0$  then calculate  $x_1$  from the equation  $x_1 = g(x_0)$ . The process is then repeated using in general  $x_n = g(x_{n-1})$  to generate the sequence  $x_0, x_1, x_2, x_3, \dots$
- If the sequence 'converges' to  $c$  as  $n$  tends to infinity then  $x_n = g(x_{n-1})$  becomes  $c = g(c)$ <sup>1</sup>. Thus the limit of the sequence satisfies  $c = g(c)$  and thus  $x = c$  is a solution of  $f(x) = 0$ .

The value of  $c$  that satisfies  $c = g(c)$  is called a fixed point of  $g(x)$ . Formally we make the following definition:

### Definition - fixed point

Given a function  $g(x)$  then  $x = c$  is a **fixed point** of  $g(x)$  if  $c = g(c)$ .

### Example 2.1

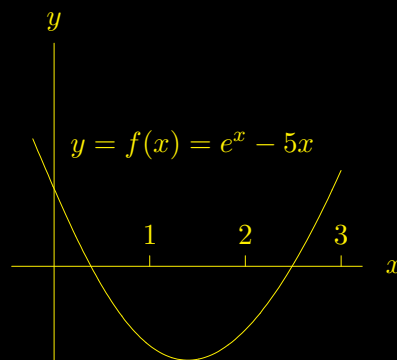


Figure 2.4:

Consider the function  $f(x) = e^x - 5x$ . To solve  $f(x) = 0$  we rewrite  $f(x) = 0$  as  $x = \frac{e^x}{5} = g(x)$

Thus we form the iterative scheme:

$$x_n = \frac{e^{x_{n-1}}}{5}$$

The following table illustrates the iteration with three first guesses for  $x_0$ , namely  $x_0 = 1$ ,  $x_0 = 2$  and  $x_0 = 3$ .

---

<sup>1</sup>assuming  $g(x)$  continuous

$n =$	0	1	2	3	4	5	6	7	8	9
$x_n =$	1	0.544	0.344	0.282	0.265	0.261	0.260	0.259	0.259	0.259
$x_n =$	2	1.478	0.877	0.481	0.323	0.276	0.264	0.260	0.259	0.259
$x_n =$	3	4.017	11.11	13341	$2 \times 10^{5793}$	...	...	...	...	...

We note the following:

- The starting value for  $x_0$  is crucial, with  $x_0 = 1$  or 2 the sequence gives the root at 0.260 whereas  $x_0 = 3$  leads to a sequence that tends to infinity.
- Fig 2.4 shows that there are two solutions to this problem but the iterative scheme has provided only one. As we see below a new arrangement of  $f(x) = 0$  will give us the other root.

We now consider a new rearrangement of  $f(x) = 0$  and see that the new rearrangement produces a sequence that tends to the other fixed point of  $g(x)$  and hence the solution of  $f(x) = 0$ . Consider the following;

$$f(x) = 0 \Rightarrow e^x - 5x = 0 \Rightarrow e^x = 5x \Rightarrow x = \ln(5x) = g(x)$$

Using  $x_n = \ln(5x_{n-1})$  with two starting values gives the following sequences

$n =$	0	1	2	3	4	5	6	7
$x_n =$	0.25	0.223	0.109	-0.602	complex ...	...	...	...
$x_n =$	1.0	1.609	2.085	2.344	2.461	2.510	2.530	2.538

We note that convergence is slow and to attain even four decimal place accuracy we would need to calculate many more terms.

## 2.4 Iteration using Excel - Method 1

The simplest method to implement iteration using Excel is to directly exploit the copy feature of the sheet. Recall that when copying cells that contain formula the references is relative and the cell references change as you copy to different parts of the worksheet. To produce the iterative scheme we need to carry out the following two distinct action:

- **Construct a user function  $g(x)$**   
Excel allows us to construct our own worksheet functions using the VBA module. To construct  $g(x) = \ln(5x)$  for the second part of Example 2.1 above:

Select **Tools** from the top menu and then **Macro, Visual Basic Editor**.

A new display should appear, with a new top menu.

Select **Insert** from the top menu. Select **Module**.

You should now have a blank window with 'Module 1' highlighted along the top. Now type the following piece of code:

```
function g(x) [Enter]
g=WorksheetFunction.In(5*x) [Enter]
end function [Enter]           note: this line appears automatically
```

Return to the Excel worksheet

- **Enter scheme into worksheet**

Enter the following three rows into cells A1:B3

	A	B
1	x	g(x)
2	=1.0	=g(A2)
3	=B2	=g(A3)

Highlight the cells A3:B3 using the mouse and then use the **+** handle in the bottom righthand corner of cell B3 to copy down the cells to row 8. The iterated values of  $x_0, x_1 \dots x_6$  should now appear in column A.

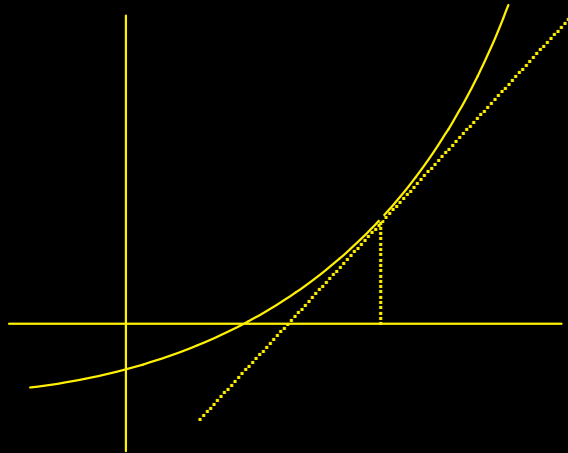
## 2.5 Newton's Method to solve $f(x) = 0$

We saw that one of the problems of the fixed point method using simple rearrangements was that typically  $\{x_n\}$  only tended to one fixed point, and hence only one solution of  $f(x) = x - g(x) = 0$ , no matter how we chose our starting value  $\{x_0\}$ . In fact the terms unstable and stable are introduced to distinguish the two types of fixed point, those that can be reached by iteration and those that can't. We now look at Newton's method that produces a fixed point algorithm such that all the fixed points are stable, and hence theoretically can be reached by a single iteration formula. In practice finding suitable starting values can be a problem; no method is perfect!

Let  $x^*$  be a root of  $f(x) = 0$ , ie  $f(x^*) = 0$ . In the diagram this is the point where the curve  $y = f(x)$  cuts the  $x$ -axis. Let  $x = x_0$  be a first estimate to  $x^*$ , ie a first guess for the solution of  $f(x) = 0$ . Construct the tangent at the point P with  $x$ -coordinate  $x_0$  as in the diagram. We can now think of the tangent as replacing the curve, thus where this tangent cuts the  $x$ -axis can then be used as a new estimate of  $x^*$ . Repeating the process produces a sequence  $\{x_n\}$  which we *hope* will tend to  $x^*$ . Algebraically the sequence  $\{x_n\}$  is produced as follows:

- We first remark that the equation of a straight line with gradient  $m$  through the point  $(x_0, f(x_0))$  is given by  $y - f(x_0) = (x - x_0)m$ .
- From Calculus the gradient of the tangent at  $P$  equals  $f'(x_0)$ .
- Hence the equation of the tangent is given by:

$$y - f(x_0) = (x - x_0)f'(x_0) \tag{2.3}$$



- The tangent cuts the axis at  $y = 0$ ,  $x = x_1$  hence

$$0 - f(x_0) = (x_1 - x_0)f'(x_0) \quad \Rightarrow^2 \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- Thus in general the iterative process is  $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$   $n = 1, 2, \dots$   
and the iterative function  $g(x)$  is given by  $g(x) = x - \frac{f(x)}{f'(x)}$

If the sequence of values tends to  $c$  then

$$c = c - \frac{f(c)}{f'(c)} \quad \Rightarrow \quad f(c) = 0$$

Thus if the sequence converges it tend to a root of  $f(x) = 0$ .

### Example 2.2

Once more considering the problem  $f(x) = e^x - 5x = 0$  which has two roots; we recall that using a simple rearrangement technique one of the roots was stable and the other unstable. Applying Newton's method  $g(x)$  is given by:

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{e^x - 5x}{e^x - 5} = \frac{x - 1}{1 - 5e^{-x}}$$

Thus using the iterative process  $x_n = \frac{x_{n-1} - 1}{1 - 5e^{-x_{n-1}}}$  with the two starting values  $x_0 = 1$  and  $x_0 = 3$  we obtain the following:

---

<sup>2</sup>provided  $f'(x_0) \neq 0$

$n =$	0	1	2	3	4
$x_n =$	1.0	0.0	0.250	0.259	0.259
$x_n =$	3.0	2.663	2.553	2.543	2.543

We see clearly that both the solutions are obtained using the same iterative process with different starting values. Indeed provided that we start close enough to a root Newton's method will converge to it. We can also see that the rate of convergence is quite impressive; another advantage of Newton's method.

## 2.6 Iteration using Excel - Method 2

We now consider the solution of  $f(x) = x^2 - 30 \sin x = 0$  using Newton's method but using the self-referencing facility of Excel rather than entering the scheme directly into the sheet. This method has the advantage of being totally automatic and allows us to set an accuracy for the solution. Applying Newton's method to  $f(x)$  gives:

$$g(x) = x - \frac{x^2 - 30 \sin x}{2x - 30 \cos x} \Rightarrow x_n = x_{n-1} - \frac{x_{n-1}^2 - 30 \sin x_{n-1}}{2x_{n-1} - 30 \cos x_{n-1}}$$

Recall that apart from  $x = 0$  there are three other solutions, one in each of the intervals  $[-6, -5]$ ,  $[-4, -3]$ ,  $[2, 3]$ . Starting with  $x_0$  equal to the mid point of each of these intervals we obtain the following results using Excel:

$n =$	0	1	2	3
$x_n =$	-5.5	-5.22	-5.18	-5.18
$x_n =$	-3.5	-3.58	-3.58	-3.58
$x_n =$	2.5	2.90	2.86	2.86

We see that correct to two decimal places the three roots are -5.18, -3.58 and 2.86

The self reference method we employ to carry out the above calculations is as follows:

- **Select and set up the self reference option:**

Select **Tools, Options** and then select the calculation tab.

Tick the **iterate box**.

Set **Max iterations** equal to 1.

Set **maximum difference** equal to 0.001.

- **Set up the user function:**  $g(x) = x - (x^2 - 30 \sin x)/(2x - 30 \cos x)$

- **Set up the worksheet:**

We first enter the value of  $x_0 = 1$  and the function  $g(x)$  as follows:

	A
1	=1
2	=g(A1)

At this point cell A1 contains the value of  $x_0$  and cell A2 contains  $x_1 = g(A1)$ . Since there is no self reference Excel does not iterate. The next stage is to replace cell A1 with =A2. Thus cell A1 contains  $g(A1)$ , that is to say  $A1 = g(A1)$ . We now have a self reference and Excel starts to iterate  $A1 = g(A1)$ . Since we have set the maximum number of iterations equal to one we can iterate one step at a time by pressing the function key **F9**.

- **General Settings**

In general the scheme will carry out the number of iterations we have set in maximum iterations, unless the maximum difference is attained, in which case it will terminate early. The maximum difference is actually the relative difference  $\left| \frac{x_n - x_{n-1}}{x_{n-1}} \right|$  and the iterations terminate if this becomes less than the specified amount.

## 2.7 Simultaneous Equations - linear and non-linear

We now consider the more general problem mentioned in the introduction, namely the solution of  $N$  equations in  $N$  unknowns. There are many different ways of doing this numerically, however we will consider only two approaches. The first is particular to simultaneous linear equations and involves the use of matrices and their inverses. The second is more general and is applicable to non-linear equations and is an iterative process based on Newton's method.

### 2.7.1 Linear simultaneous equations

We first look at the familiar problem of linear simultaneous equations and indeed concentrate our attention just to the case when such a set has a unique solution. Consider the problem:

$$\begin{aligned} 2x + y &= 3 \\ x + y &= 2 \end{aligned} \tag{2.4}$$

We rewrite these equations in matrix form as follows:

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

This is now written in an even shorter form as

$$A\underline{x} = \underline{b}$$

where

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \quad \underline{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \underline{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

#### Definitions

$A$  is called a **square matrix** or a two by two matrix (denoted  $2 \times 2$ ) since it has two rows

and two columns. In general a matrix with  $m$  rows and  $n$  columns is called an  $m$  by  $n$  matrix (denoted  $m \times n$ )

$\underline{x}$  and  $\underline{b}$  are referred to as two by one matrices (denoted  $2 \times 1$ ) or **column matrices** or **column vectors**.

In the above example it is clear how we have identified the terms in the equations with the matrices. We now show how we can extend the idea of simple algebra to that of matrices. Consider the simple problem of solving  $ax = b$  where  $a$ ,  $b$  and  $x$  are real numbers. Formally

$$ax = b \Rightarrow a^{-1}ax = a^{-1}b \Rightarrow x = a^{-1}b = \frac{b}{a}$$

We now ask the question, is it possible to carry out a similar process with matrices? That is to say is it possible to write:

$$A\underline{x} = \underline{b} \Rightarrow A^{-1}A\underline{x} = A^{-1}\underline{b} \Rightarrow \underline{x} = A^{-1}\underline{b} \quad (2.5)$$

In the case of the single variable  $a^{-1} = 1/a$ , the question now is, what is  $A^{-1}$ ? Clearly it is not possible to divide by a matrix in the same way as we divide by a real number. However the key property of  $a^{-1}$  in solving the problem is that  $a^{-1}a = 1$ , as we will see below we will find a matrix denoted  $A^{-1}$  that has a similar property. Since  $A^{-1}A$  will be a product of two matrices we need to say first exactly what we mean by the product of two matrices.

When writing our equations in matrix form we have implicitly defined what we mean by the product of a  $2 \times 2$  matrix and a column vector. Thus for example:

$$\begin{pmatrix} 3 & -4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 3 \times 5 - 4 \times 6 \\ 2 \times 5 + 1 \times 6 \end{pmatrix} = \begin{pmatrix} -9 \\ 16 \end{pmatrix}$$

Thus given two matrices  $A$  and  $B$  the product matrix  $AB$  is defined as the matrix obtained by multiplying  $A$  by each of the columns of  $B$  in turn. Thus for example if  $A$  and  $B$  are given by:

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

Then applying  $A$  to the first column of  $B$  gives as above:

$$\begin{pmatrix} 3 & -4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} -9 \\ 16 \end{pmatrix}$$

and applying  $A$  to the second column of  $B$  gives

$$\begin{pmatrix} 3 & -4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -10 \\ -3 \end{pmatrix}$$

Thus  $AB$  is given by

$$AB = \begin{pmatrix} 3 & -4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & -2 \\ 6 & 1 \end{pmatrix} = \begin{pmatrix} -9 & -10 \\ 16 & -3 \end{pmatrix}$$

**We now return to our example**

From above  $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ .

The matrix corresponding to  $A^{-1}$  is given by  $A^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ . (see notes below)

We can verify this by considering the following product:

$$A^{-1}A = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

The matrix we have denoted  $I$  corresponds to the number 1 when doing ordinary arithmetic; that is to say it has no effect when multiplying other matrices, this can be seen from:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

From Eq(2.5) we have that the solution of our problem is given by  $\underline{x} = A^{-1}\underline{b}$ , thus:

$$\underline{x} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Thus the solution to our problem is  $x = 1, y = 1$ .

**Notes on product and inverse**

The following two points need to be considered when dealing with matrices.

- Bearing in mind that matrices can have  $m$  rows and  $n$  columns there is a restriction on when it is possible to calculate the product. For the product of  $A$  and  $B$  to be defined the number of columns of  $A$  must be the same as the number of rows of  $B$ . This can be seen from the following example where we consider writing out a set of equations in matrix form.

Given

$$\begin{aligned} 2x + 3y + 4z + 5w &= 1 \\ 3x - 4y + 5z - 6w &= 3 \end{aligned}$$

(2.6)

we can write them in matrix form as

$$\begin{pmatrix} 2 & 3 & 4 & 5 \\ 3 & -4 & 5 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

Here we see how the  $2 \times 4$  matrix multiplies the column with 4 rows. This means that the first matrix can multiply any matrix with four rows by simply repeating the process.



- For a general  $2 \times 2$  matrix we can calculate the inverse as follows:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow A^{-1} = \frac{1}{(ad - bc)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \text{provided } (ad - bc) \neq 0$$

- We can not calculate the inverse of matrix that isn't square. Furthermore not all square matrices have an inverse, as we can see above for the  $2 \times 2$  case no inverse exists when  $(ad - bc) = 0$ . This last fact is analogous to not being able to calculate  $a^{-1}$  when  $a = 0$ .

## 2.7.2 Matrix product and inverse using Excel

Excel introduces two array functions one to carry out the product of two suitable matrices and one to calculate the inverse of a square matrix. The following example calculates the product of two matrices:

### Example 2.3      =MMULT(range,range)

Given

$$A = \begin{pmatrix} 1 & 3 & 5 \\ -1 & 4 & -2 \\ 4 & -2 & 8 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} -2 & 4 & -1 \\ 1 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix}$$

Enter the elements of  $A$  into the top left hand corner of a new worksheet. That is to say enter the elements into the rectangular block of cells from A1 to C3. Excel refers to this as a **range** or an array and is denoted by **A1:C3**

Leaving a blank row enter the elements of  $B$  into the rows under  $A$ . That is to say enter  $B$  into cells A5:C7

To calculate the product  $AB$  we need to enter =MMULT(A1:C3,A5:C7). Excel refers to this as an **array function** which needs to be entered in a rather special way in order to see the complete answer. Enter as follows:

- Select cells A9:C11 using the mouse (highlight the cells). This is the range in which the answer will appear.
- type in =MMULT(A1:C3,A5:C7) - do not press any keys yet!  
(If you wish you can point to the ranges with the mouse rather than typing in explicitly)
- To display the product you need to simultaneously press **Ctrl** **Shift** and **Enter**

This should give you:

$$AB = \begin{pmatrix} 6 & 13 & 25 \\ 4 & 8 & 27 \\ -2 & 10 & -10 \end{pmatrix}$$

displayed in the range A9:C11

**Example 2.4**    **=MINVERSE(range)**

- Enter the matrix:

$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -2 \\ 4 & -2 & 2 \end{pmatrix}$$

in the range A13:C15.

- Select the range A17:C19 ready to receive the answer.
- Type **=MINVERSE(A13:C15)** and press **Ctrl Shift** and **Enter**.  $A^{-1}$  should now be displayed in the range A17:C19 and should be:

$$A^{-1} = \begin{pmatrix} -1 & 0 & 0.5 \\ -3 & -1 & 0.5 \\ -1 & -1 & 0 \end{pmatrix}$$

- We can check that we do indeed have the inverse of  $A$  by using **MMULT** with  $A$  and  $A^{-1}$  to verify that the product is  $I$ , the  $3 \times 3$  matrix with ones down the leading diagonal and zeros elsewhere.

**Example 2.5**    **Solution of equations**

Given the equations:

$$\begin{aligned} x + y + z &= 3 \\ 2x + y - 2z &= 1 \\ 3x - y - z &= 1 \end{aligned}$$

the solution is given by:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -2 \\ 3 & -1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$$

- Enter the matrix  $A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -2 \\ 3 & -1 & -1 \end{pmatrix}$  in cells A1:C3 and the column matrix  $\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$  in cells E1:E3
- Select A5:C7 to accept the inverse of  $A$  and enter **=MINVERSE(A1:C3)** followed by **Ctrl Shift** and **Enter**
- To calculate the solution select E5:E7 and enter **=MMULT(A5:C7,E1:E3)** followed by **Ctrl Shift** and **Enter**. The cell E5 should be the  $x$  value, cell E6 the  $y$  value and cell E7 the  $z$  value. (*they all equal 1*)

### 2.7.3 Non-linear simultaneous equations

This section looks at the solution of the simultaneous equations

$$z = f(x, y) = 0 \quad \text{and} \quad z = g(x, y) = 0 \quad (2.7)$$

where  $f(x, y)$  and  $g(x, y)$  are something more than expressions of the type  $ax + by + c$ . That is to say they are not both linear; as this case has been dealt with in the previous section. Although we develop the method for two equations in two unknowns it becomes apparent how the method may be extended to  $n$  equations in  $n$  unknowns.

Geometrically if we plot  $z = f(x, y)$  and  $z = g(x, y)$  we obtain two surfaces. Since we require  $z = 0$  we look to see where these two surfaces intersect the  $x$ - $y$  plane. In general and with reference to Fig 2.6 the surfaces will intersect the  $x$ - $y$  plane in two curves,  $C_1$  and  $C_2$ . Therefore the points we are looking for are the intersections of these two curves. In the figure the curves are shown to intersect at the single point  $P$ .

Following the ideas used in Newton's method for one variable we now start with a point  $(x_0, y_0)$  and construct the tangent planes to the two surfaces at the two points with  $x$ - $y$  coordinates  $(x_0, y_0)$ . We then look to see where the two tangent planes cut  $z = 0$ , that is to say cut the  $x$ - $y$  plane. This in general will give two lines in the  $x$ - $y$  plane,  $L_1$  (see Fig 2.7) and  $L_2$ . The next approximation is then given by the intersection of  $L_1$  and  $L_2$  and is denoted as  $(x_1, y_1)$ . The process can be repeated starting with  $(x_1, y_1)$ .

In Fig 2.7 a tangent plane has been drawn at the point  $Q$  on the surface of  $z = f(x, y)$

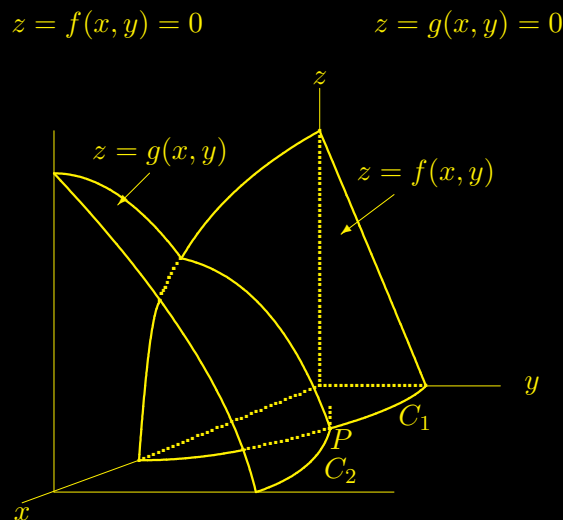


Figure 2.6: The surface  $z = f(x, y)$  meets the plane  $z = 0$  in the curve  $C_1$ . The surface  $z = g(x, y)$  meets the plane  $z = 0$  in the curve  $C_2$ . The two curves meet at  $P$ , the coordinates of which give the solution of  $f(x, y) = g(x, y) = 0$

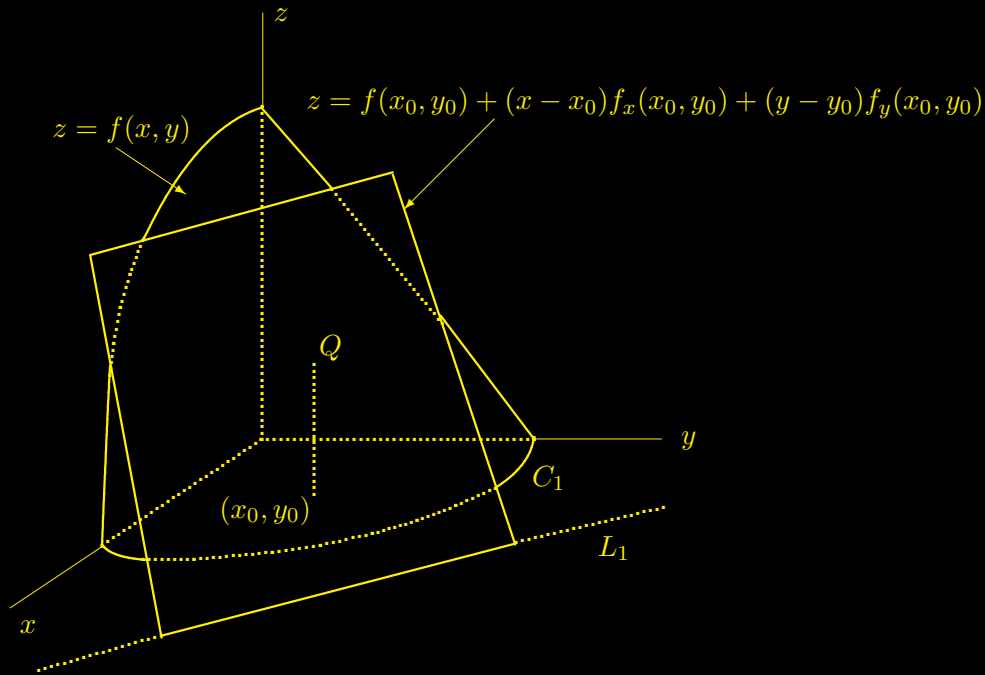


Figure 2.7: The surface  $z = f(x, y)$  meets the plane  $z = 0$  in the curve  $C_1$ . The tangent plane to the surface at  $Q$ , the point with coordinates  $(x_0, y_0)$ , meets the plane  $z = 0$  in the line  $L_1$ . See below and Eq(2.9) for equation of tangent plane.

with  $x$ - $y$  coordinates  $(x_0, y_0)$ . This tangent plane cuts the  $x$ - $y$  plane in the line  $L_1$ . In this method for approximating the roots of Eq(2.7)  $L_1$  replaces  $C_1$ . Carrying out a similar construction for the surface  $z = g(x, y)$ , at the point with  $x$ - $y$  coordinates  $(x_0, y_0)$ , produces a line  $L_2$  to replace the curve  $C_2$ . We now find the point of intersection of  $L_1$  and  $L_2$  as an approximation to the point of intersection of  $C_1$  and  $C_2$ . (recall the point of intersection of  $C_1$  and  $C_2$  gives the solution to Eq(2.7))

Having described the method the only thing left to do is to find the equations of the tangent planes, set  $z = 0$  to give the equations of the two lines  $L_1$  and  $L_2$  and then solve to find their point of intersection.

We recall that for the problem of a function of one variable, that is to say  $f(x) = 0$ , we calculated the equation of the tangent to the curve at  $x = x_0$ , this was given by Eq(2.3) as

$$y = f(x_0) + (x - x_0)f'(x_0) \quad (2.8)$$

Since the function  $f$  now depends on two variable it is no longer clear what we mean by  $f'(x, y)$ . Indeed we introduce a new type of derivative, called the **partial derivative** (*not to be confused with implicit differentiation*), this either differentiates the function as if  $x$  is the only variable and  $y$  treated as a constant, or the other way round. Geometrically the idea of a partial derivative is the gradient of the tangent to the curve on the surface

$z = f(x, y)$  made from the intersection of this surface and the plane  $x = \text{constant}$  or  $y = \text{constant}$ . Thus as we see in fig 2.8, at a given point on the surface we have two such tangents and thus two partial derivatives, one with respect to  $x$  and one with respect to  $y$ . How to obtain the partial derivative and the ensuing notation can perhaps best be

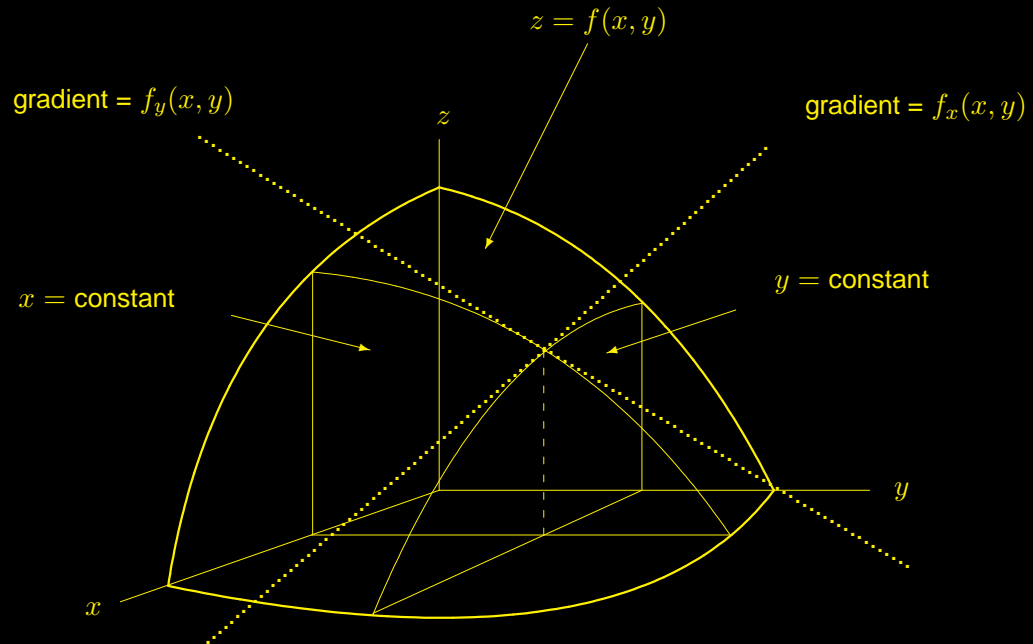


Figure 2.8: The dotted lines are the tangents to the curves formed from the intersection of the surface and the constant planes  $x = \text{const}$  and  $y = \text{const}$ . They have, respectively, gradients  $f_x(x, y)$  and  $f_y(x, y)$ .

seen from the following example:

### Example 2.6

Given

$$f(x, y) = 2 - x^2y + 4x + 5y$$

The partial derivative of  $f(x, y)$  with respect to  $x$  is denoted<sup>3</sup> by  $f_x(x, y)$  and obtained by differentiating with respect to  $x$  and keeping  $y$  constant.

Thus

$$f_x(x, y) = -2xy + 4$$

The 2 and the  $5y$  are treated as constants and thus vanish on differentiation; the  $y$  in the  $-x^2y$  term is treated as a constant multiple, thus we just differentiate the  $x^2$ ; the  $4x$  is differentiated to give the 4

---

<sup>3</sup>Alternative notation  $f_x(x, y) = \frac{\partial f(x, y)}{\partial x}$   $f_y(x, y) = \frac{\partial f(x, y)}{\partial y}$

Similarly differentiating with respect to  $y$  and keeping  $x$  constant is denoted<sup>3</sup> by  $f_y(x, y)$ :

$$f_y(x, y) = -x^2 + 5$$

*The 2 and the 4x are treated as constants and thus vanish on differentiation; the  $x^2$  in the  $-x^2y$  term is treated as a constant multiple, thus we just differentiate the  $y$ ; the  $5y$  is differentiated to give the 5*

We are now in a position to write down the equation of the tangent plane to  $z = f(x, y)$  at the point  $(x_0, y_0)$ . The extension of the tangent equation Eq(2.8) for the function of two variables is given by:

$$z = f(x_0, y_0) + (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0) \quad (2.9)$$

Although this is not proved here I think it is a reasonable result for students to accept. The tangent approximations are also referred to as **linear approximations** to the function at the given point. It is now clear from Eq(2.9) what the linear approximation would be for a function of any number of variables.

In general for a function  $z = f(x_1, x_2, \dots, x_n)$  the linear approximation at  $\underline{a} = (a_1, a_2, \dots, a_n)$  is given by:

$$z = f(\underline{a}) + (x_1 - a_1)f_{x_1}(\underline{a}) + (x_2 - a_2)f_{x_2}(\underline{a}) + \dots + (x_n - a_n)f_{x_n}(\underline{a}) \quad (2.10)$$

This equation can be written in a more compact form as:

$$z = f(\underline{a}) + \sum_{i=1}^{i=n} (x_i - a_i)f_{x_i}(\underline{a}) \quad (2.11)$$

Clearly we can see that for  $n = 1$  this gives the equation of the tangent to a simple curve; for  $n = 2$  it gives the tangent plane to the surface. Past this point we lose the intuitive geometrical description and simply refer to it as the linear approximation to the function at the given point.

### Newton's method in two variables

To solve

$$z = f(x, y) = 0 \quad \text{and} \quad z = g(x, y) = 0$$

we proceed as follows:

- Select a point  $(x_0, y_0)$  close to a solution.
- Construct the tangent planes to  $f(x, y)$  and  $g(x, y)$  at  $(x_0, y_0)$ . These are given by

$$z = f(x_0, y_0) + (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0)$$

$$z = g(x_0, y_0) + (x - x_0)g_x(x_0, y_0) + (y - y_0)g_y(x_0, y_0)$$

- Setting  $z = 0$  to give the equations of the lines  $L_1$  (see Fig 2.7) and  $L_2$ :

$$0 = f(x_0, y_0) + (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0)$$

$$0 = g(x_0, y_0) + (x - x_0)g_x(x_0, y_0) + (y - y_0)g_y(x_0, y_0)$$

- Obtain the point of intersection of these lines,  $(x_1, y_1)$ , by first writing them in matrix notation:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}_{(x_0, y_0)} + \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_0, y_0)} \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix}$$

Where for convenience we have placed  $(x_0, y_0)$  outside the brackets to indicate that the functions inside the brackets are evaluated at  $(x_0, y_0)$ .

$$\Rightarrow - \begin{pmatrix} f \\ g \end{pmatrix}_{(x_0, y_0)} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_0, y_0)} \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix}$$

Thus using the inverse of the matrix:

$$- \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_0, y_0)}^{-1} \begin{pmatrix} f \\ g \end{pmatrix}_{(x_0, y_0)} = \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix}$$

Thus we obtain Newton's formula for two functions of two variables:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_0, y_0)}^{-1} \begin{pmatrix} f \\ g \end{pmatrix}_{(x_0, y_0)}$$

### Remarks

- In general to solve

$$f(x, y) = g(x, y) = 0$$

use the iterative scheme:

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} x_{n-1} \\ y_{n-1} \end{pmatrix} - \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_{n-1}, y_{n-1})}^{-1} \begin{pmatrix} f \\ g \end{pmatrix}_{(x_{n-1}, y_{n-1})} \quad n = 1, 2, \dots \quad (2.12)$$

starting with some point  $(x_0, y_0)$  close to a solution.

- We can see that this formula is the same as Newton's method for a function of one variable if  $g$  is no longer a part of the problem and  $f$  only depends on  $x$ . In this case the matrix becomes  $(f'(x_{n-1}))^{-1}$  and Eq(2.12) gives:

$$x_n = x_{n-1} - (f'(x_{n-1}))^{-1} f(x_{n-1}) = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \Rightarrow$$

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad \text{Newton's method for one variable}$$

- The extension to  $N$  equations in  $N$  unknowns is now very easy to see. The matrix of partial derivatives will be  $N \times N$  and the column vectors will have  $N$  entries. In

detail for  $N = 3$ , to solve  $f(x, y, z) = g(x, y, z) = h(x, y, z) = 0$  for the solution near the given point  $(x_0, y_0, z_0)$  the scheme becomes:

$$\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} = \begin{pmatrix} x_{n-1} \\ y_{n-1} \\ z_{n-1} \end{pmatrix} - \begin{pmatrix} f_x & f_y & f_z \\ g_x & g_y & g_z \\ h_x & h_y & h_z \end{pmatrix}^{-1} \begin{pmatrix} f \\ g \\ h \end{pmatrix} \quad n = 1, 2, \dots$$

where the functions and derivatives on the righthand side of the equation are evaluated at  $(x_{n-1}, y_{n-1}, z_{n-1})$ .

### Example 2.7

Obtain the solution to

$$f(x, y) = 2 - x^2 - y = 0 \quad g(x, y) = 2x - y^2 - 1 = 0$$

close to the point  $x = 0.5$   $y = 0.5$

For convenience we will set up the iterative scheme for  $n = 1$ . To calculate the matrix we have to obtain four partial derivatives:

- $f(x, y) = 2 - x^2 - y \Rightarrow f_x(x, y) = -2x$  and  $f_y(x, y) = -1$
- $g(x, y) = 2x - y^2 - 1 \Rightarrow g_x(x, y) = 2$  and  $g_y(x, y) = -2y$

Thus the first step of the scheme is given by:

$$\begin{aligned} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}_{(x_0, y_0)}^{-1} \begin{pmatrix} f \\ g \end{pmatrix}_{(x_0, y_0)} \\ \Rightarrow \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} -2x_0 & -1 \\ 2 & -2y_0 \end{pmatrix}^{-1} \begin{pmatrix} 2 - x_0^2 - y_0 \\ 2x_0 - y_0^2 - 1 \end{pmatrix} \end{aligned}$$

Explicitly the first step gives, with  $x_0 = 0.5$  and  $y_0 = 0.5$ :

$$\begin{aligned} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} - \begin{pmatrix} -1 & -1 \\ 2 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 1.25 \\ -0.25 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} - \begin{pmatrix} -1/3 & 1/3 \\ -2/3 & -1/3 \end{pmatrix} \begin{pmatrix} 1.25 \\ -0.25 \end{pmatrix} \\ \Rightarrow \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} - \begin{pmatrix} -1/2 \\ -3/4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.25 \end{pmatrix} \end{aligned}$$

Continuing this scheme to iterate gives the following results:



$n =$	0	1	2	3	4
$x_n =$	0.5	1.00	0.99	1.00	1.00
$y_n =$	0.5	1.25	1.00	1.00	1.00

We see that the scheme converges rapidly to the root  $x = 1, y = 1$ .

### Newton's Method using Excel

Since the problem is expressed in terms of matrices we can use the MINVERSE and MMULT together with the automatic iteration provided by the circular or self reference facility. We assume that we have already set up the two functions  $f(x, y)$  and  $g(x, y)$ . We initially enter the following into columns A and B of the sheet.

	A	B	comments
1	x0	=0.5	initial x value
2	y0	=0.5	initial y value
3	The matrix		
4	=-2*B1	= -1	enters $f_x$ and $f_y$
5	=2	= -2 * B2	enters $g_x$ and $g_y$
6	Calc Inverse		
7	=MINVERSE(A4:B5)		enter this array function with <b>Ctrl Shift Enter</b> <sup>4</sup>
8			
9			
10	f	=f(B1,B2)	enters $f(x_0, y_0)$
11	g	=g(B1,B2)	enters $g(x_0, y_0)$
12			
13	x1	=B1:B2-MMULT(A7:B8,B10:B11)	enter iterative formula
14	y1		enter the array function
15			with <b>Ctrl Shift Enter</b> <sup>4</sup>

At this point there is no self reference and therefore no iteration will take place. Select the option for Excel to carry out circular reference as in Sec(2.6), again setting the maximum iteration option equal to one so that we can observe the iteration one step at a time by repeatedly using **F9**. To gain even further control on when the calculations take part tick the **manual** option in the calculations window. This means that no calculation will take place anywhere on the worksheet until the **F9** key is pressed.

To complete the above change the entry in cell B1 to **=B13** and the entry in cell B2 to **=B14**. Since the manual option has been selected no iteration will take place until the **F9** key is pressed. (Note: if the manual option had not been selected a recalculation would have taken place as soon as the cell B1 was changed to **=B13**; this can sometimes be confusing). To continue to iterate repeatedly press **F9**.

<sup>4</sup>recall you must select the range to accept the output from a range/array function before typing in the function. The function is finally entered using **Ctrl-Shift-Enter**. MMULT and MINVERSE are both range/array functions

## Chapter 3

# Financial Functions in Excel

### 3.1 Introduction

This chapter develops the background for the financial functions in Excel that deal with loan repayments and other compound interest problems. For example if I borrow \$5000 over three years at a fixed rate of interest how much do I pay back per month. Although the mathematical ideas behind such problems are quite straightforward some care needs to be employed when trying to implement the functions provided by Excel. Thus a thorough understanding of the mathematical formulas being used by Excel is essential in order to obtain and interpret correctly the answers given. Nearly all the work in compound interest is based on a single formula which in turn is derived from the sum of a geometric progression. We thus start with a few preliminary ideas.

### 3.2 Geometric Progression

A quantity is said to vary geometrically over a given set of time intervals if one value is some constant multiple of the previous value. Symbolically we have:

#### Definition

A geometric progression or series of  $n$  values is an ordered set of the form:

$$a, ar, ar^2, ar^3 \dots ar^{n-1}$$

where  $a$  is referred to as the first term and  $r$  the common ratio.

For example:

$$5, \frac{5}{2}, \frac{5}{4}, \frac{5}{8} \dots 5 \left(\frac{1}{2}\right)^{n-1} \quad a = 5 \quad r = \frac{1}{2}$$
$$3, -\frac{3}{2}, \frac{3}{4}, -\frac{3}{8} \dots 3 \left(-\frac{1}{2}\right)^{n-1} \quad a = 3 \quad r = -\frac{1}{2}$$

We observe that in the second example the common ratio  $r$  is negative.

Usually we are interested in the sum of a geometric progression, a formula for which

can be obtained as follows:

Let

$$S_n = a + ar + ar^2 + ar^3 \dots ar^{n-1}$$

then

$$S_n - rS_n = (a + ar + ar^2 + \dots ar^{n-1}) - (ar + ar^2 + ar^3 + \dots ar^n) = a - ar^n$$

Thus we have

$$S_n(1 - r) = a(1 - r^n) \Rightarrow S_n = a \left( \frac{1 - r^n}{1 - r} \right) \quad r \neq 1 \quad (3.1)$$

We note the following two points:

- If  $r = 1$  the formula is not valid, however in this case  $S_n = a + a + \dots + a$ ,  $n$  time and thus  $S_n = na$ .
- In interest problems  $|r|$  is usually less than 1 and quite often  $n$  is large; we thus note the following result:

If  $|r| < 1$  then:

$$\text{as } n \rightarrow \infty \quad r^n \rightarrow 0 \quad \text{Thus Eq(3.1)} \Rightarrow S_n \rightarrow \frac{a}{1 - r}$$

Indeed  $\frac{a}{1 - r}$  is quite a common approximation for  $S_n$  when  $n$  is large and  $r$  is small.

### 3.3 Basic Compound Interest

The basic compound interest problem is:

If  $\$A$  is invested for  $n$  periods at an interest rate of  $r$  per period, how much is the investment worth at the end of the  $n$  periods?

For example if I invest  $\$1000$  for 12 months compounded monthly at a rate of 1% at the end of each month how much do I have at the end of the year? In this example  $A = 1000$ ,  $r = 0.01$  and  $n = 12$ . Note that  $r$  is expressed as a fraction in the mathematics even though it is expressed as a percentage in the question. Fig 3.1 illustrates the growth of the initial investment  $A$  over the  $n$  periods as the interest is added at the end of each period. At the end of the first month you have your initial investment  $A$  plus the interest on the investment,  $rA$ , giving a total of  $A(1 + r)$ <sup>1</sup>. This new amount is then used as the starting amount for the second period; thus at the end of the second interval we

<sup>1</sup>Note that it is true in general that after compounding over one period the amount at the end of the period is the amount at the start of the period multiplied by  $(1 + r)$

have  $A(1 + r)$  plus the interest on this amount,  $rA(1 + r)$ . As we see this simplifies to give  $A(1 + r)^2$ . Continuing in this manner gives a total amount of  $A(1 + r)^n$  at the end of the  $n^{th}$  period.

In our simple example of \$1000 being invested at 1% per month for a year the amount

period	0	1	2	...	$n$
Amount	$A$	$A + rA$ $A(1 + r)$	$A(1 + r) + rA(1 + r)$ $A(1 + r)^2$	...	$A(1 + r)^n$

Figure 3.1: Shows the detail of compounding after periods 1 and 2.

at the end of the 12 months =  $1000(1 + 0.01)^{12} = \$1126.83$ .

We note that the final amount is \$26.83 more than if the compounding had been carried out only once at the end of the year at a rate of 12%. In all our problems it will be clear what the rate is and when we are applying it. We will not discuss the idea of effective interest rates or indeed the various ways in which banks express the interest rates in their advertising. However knowing mathematically what information is required to carry out the correct calculation will enable us to ask meaningful questions in any given situation.

Thus we summarise:

An amount  $A$  invested at a rate  $r$  for  $n$  periods amounts to  $A(1 + r)^n$ .

### 3.4 Basic Investment Problem

If an amount  $A$  is invested initially at a rate  $r$  for  $n$  periods and in addition we invest an amount  $p$  at the end of each period what will the final amount be?

This simple problem represents someone who initially deposits an amount  $A$  and then continues to invest, say at the end of each month, an amount  $p$ , with the interest being compounded at the rate  $r$  at the end of each month.

As we will see it will be possible to adjust this basic problem to answer a variety of different questions.

With reference to fig 3.2 we see that at the end of the first period the initial investment has grown to  $A(1 + r)^1$ . Adding to this the regular investment  $p$  gives a total of

$$S_1 = \underline{A(1 + r) + p}$$

period	0	1	2	...	$n$
Amount	$A$	$A(1+r) + p$	$\{A(1+r) + p\}(1+r) + p$	...	$S_n$
	$S_0$	$S_1$	$S_2 = A(1+r)^2 + p(1+r) + p$	...	

Figure 3.2: Shows the detail of compounding after periods 1 and 2 with an initial investment  $A$  and subsequent investments  $p$  at the end of each period.

at the start of the second period.

At the end of the second period this amount will have grown to its original value multiplied by  $(1+r)$ . Adding in our next regular payment  $p$  means that we start the third period with:

$$S_2 = \{A(1+r) + p\}(1+r) + p = \underline{A(1+r)^2 + p(1+r) + p}$$

Carrying out this process once more gives:

$$S_3 = \{A(1+r)^2 + p(1+r) + p\}(1+r) + p = \underline{A(1+r)^3 + p(1+r)^2 + p(1+r) + p}$$

From this we can deduce the  $n^{th}$  case:

$$S_n = A(1+r)^n + p(1+r)^{n-1} + p(1+r)^{n-2} \dots + p$$

After the first term we note that we have the sum of a geometric progression which with reference to Eq(3.1),  $a = p$  and the common ratio =  $(1+r)$  gives:

$$S_n = A(1+r)^n + p \left\{ \frac{1 - (1+r)^n}{1 - (1+r)} \right\} = A(1+r)^n + p \frac{(1+r)^n - 1}{r}$$

Thus we summarise:

An initial investment  $A$  and regular investment  $p$  at the end of each period, compounded at the rate  $r$  per period, becomes after  $n$  periods:

$$A(1+r)^n + p \left\{ \frac{(1+r)^n - 1}{r} \right\} \tag{3.2}$$

**Example 3.1**

I initially invest \$1000 in a saving scheme and then at the end of each month I invest an extra \$50. If the interest rate is 0.5% per month and I continue this process for two year, how much will my saving be worth.

Substituting directly into Eq(3.2) for  $A = 1000$ ,  $p = 50$ ,  $r = 0.005$  and  $n = 24$  we obtain the final amount:

$$\text{final amount} = 1000(1 + 0.005)^{24} + 50 \left\{ \frac{(1 + 0.005)^{24} - 1}{0.005} \right\} = \underline{\underline{\pounds 2398.76}}$$

### 3.5 Basic Financial Worksheet Functions in Excel

In Example 3.1 it is clear how we should enter the parameters  $A$ ,  $p$ ,  $r$ , and  $n$ , however to use the worksheet functions in Excel we have to adopt a sign convention.

#### Excel's sign convention

Money that flows away from you is given a negative sign and money that flows towards you is given a positive sign. A more concrete way of looking at this is, if you have to put your hand in your pocket and pay money out then it carries a minus sign, whereas on the other hand if money is put into your pocket then it carries a positive sign.

In Example 3.1 we make an initial investment, thus in the Excel functions we would enter  $-1000$  and not  $+1000$ . Similarly we make a monthly payment thus we would enter  $-50$  into the Excel functions and not  $+50$ . The final amount is paid back to you thus Excel will yield a positive answer.

Thus denoting the final or future value by FV, monthly payments by PMT and the initial or present value of our investment by PV the sign convention would require:

- $A = -PV$  since the initial amount is paid out to the Bank
- $p = -PMT$  since the payments are paid out to the Bank
- $S_n = FV$  since the future value is paid back to you.

With this in mind and substituting into Eq(3.2) Excel is in fact using the formula:

$$FV = -PV(1 + r)^n - PMT \left\{ \frac{(1 + r)^n - 1}{r} \right\} \quad (3.3)$$

#### Excel's Future Value Function

**=FV(r, n, PMT, PV, type)**

The parameters are as in Eq(3.3) with the additional parameter "type". If type is set equal to zero then the formula is as in Eq(3.3) with the payments being made at the end of each period. If type is set equal to 1 then it is assumed that the payments are made

at the start of each period and thus the formula is amended accordingly.

To solve Example 3.1 using Excel we would enter into the worksheet:

$$= \mathbf{FV}(0.005, 24, -50, -1000, 0)$$

or alternatively the rate of 0.005 could be entered as 0.5%. This gives \$2398.76 as before.

### Excel's Present Value Function

$$=\mathbf{PV}(r, n, \text{PMT}, \text{FV}, \text{type})$$

This is simply Eq(3.3) rearranged to make PV the subject of the formula giving:

$$\text{PV} = - \left\{ \text{FV} + \text{PMT} \left\{ \frac{(1+r)^n - 1}{r} \right\} \right\} (1+r)^{-n}$$

### **Example 3.2**

If I wish to accumulate \$5000 in four years time by depositing \$75 per month in a fixed rate account with interest rate of 0.4% per month, what initial investment must I also make.

Here the accumulated amount of \$5000 is the FV which is paid to us, thus it is entered as a positive quantity. We are making payments of \$75 thus we enter PMT as -75. Four years is 48 periods thus  $n = 48$  and the interest rate is entered as 0.004 or 0.4%. Thus we enter:

$$= \mathbf{PV}(0.004, 48, -75, 5000, 0)$$

This gives -858.55, the minus sign indicating that we must pay out, that is to say deposit \$858.55, at the start of the investment in order to reach \$5000 by the end.

### Excel's Payment Function

$$=\mathbf{PMT}(r, n, \text{PV}, \text{FV}, \text{type})$$

Again this is a simply rearrangement of Eq(3.3), making PMT the subject of the formula gives:

$$\text{PMT} = - \left\{ \text{FV} + \text{PV}(1+r)^n \right\} \times \left\{ \frac{r}{(1+r)^n - 1} \right\}$$

The sort of question that you may wish to answer is typically the following concerning the repayment of a loan.

### **Example 3.3**

How much will the monthly repayments be if I borrow \$100,000 over 20 years with an effective monthly interest rate is 0.5%.

In this problem the final value, FV, is zero, since we must pay back all the loan and interest by the end of the period. The loan is paid to us at the start, so the present value, PV is +100,000. The number of payments is 240 and we assume that we pay at the end of each month thus type=0.

Thus we need to enter:

$$= \text{PMT}(0.5\%, 240, 100000, 0, 0)$$

This gives  $-716.43$ . As expected the result is negative since we are paying out money each month.

### Excel's Number of Periods Function

=**NPER**(r, PMT, PV, FV, type)

This function calculates the number of periods in a financial problem, in the case of repaying a loan this is the number of repayments, as in Example 3.3 above. Again a simple rearrangement of Eq(3.3) making  $n$  the subject of the formula gives:

$$n = \frac{1}{\ln(1+r)} \ln \left\{ \frac{\frac{\text{PMT}}{r} - \text{FV}}{\frac{\text{PMT}}{r} + \text{PV}} \right\}$$

A simple example to illustrate this would be:

#### **Example 3.4**

How long would it take me to pay off a loan of \$10,000 at a rate of 0.5% per month if I can afford to pay \$100 per month. The loan is paid to us at the start so PV= +10000, at the end of the loan the final value is zero so we set FV=0, since we pay out the payments PMT=  $-100$  and as usual assuming that we are making the payments at the end of the month type=0. Thus we enter:

$$= \text{NPER}(0.5\%, -100, 10000, 0, 0)$$

This gives 138.98, which means we need to make 138 payments of \$100 plus a final payment of less than normal or alternatively we make 138 payments with the 138<sup>th</sup> payment being larger than normal. We can investigate the two options by using the FV function:

- If we make 139 payments of \$100 then we will overpay, that is to say the future value will be positive instead of zero. We need to find the future value of the loan after 139 payment of \$100 and then adjust the final payment accordingly. Using the FV function:

$$= \text{FV}(0.5\%, 139, -100, 10000, 0)$$

gives 2.42, that is to say we will overpay by \$2.42. Thus the final payment (the 139<sup>th</sup>) should be \$97.58



- If we only make 138 payments of \$100 then how much of the loan will still be outstanding? Changing the number of payments in the FV function gives:

$$= \mathbf{FV}(0.5\%, 138, -100, 10000, 0)$$

which gives  $-97.09$ . Thus we have an under payment of \$97.09 which means our final (the 138<sup>th</sup>) payment must be  $\$100 + \$97.09 = \$197.09$ .

### Excel's rate Function

**=RATE**(n, PMT, PV, FV, type, guess)

This function calculates the value of  $r$  from Eq(3.3), however it is not possible to make  $r$  the subject of the formula and hence we are unable to obtain an explicit formula for  $r$  in terms of the other parameters. Excel solves Eq(3.3) using an iterative scheme with a starting value equal to "guess", which is the final parameter of the RATE function. Although not certain, it is most likely that Excel uses Newton's method or one of its variations to implement the iterative scheme. In detail Excel uses an iterative scheme with first value "guess" to solve:

$$f(r) = FV + PV(1+r)^n + PMT \left\{ \frac{(1+r)^n - 1}{r} \right\} = 0$$

In most cases setting "guess" equal to zero works well, indeed omitting the "guess" parameter completely will still work and Excel will assume by default that it is zero.

#### **Example 3.5**

I borrow \$1000 over 1 year making payments of \$100 per month at the end of each month. What is the monthly interest rate? Since we borrow \$1000 the present value  $PV=1000$ ; the payments have a negative sign since they are paid out thus  $PMT = -100$ ; the number of payments  $=12$ ; type is set equal to zero since the payments are at the end of each month;  $guess=0$  as recommended. Thus we enter:

$$= \mathbf{RATE}(12, -100, 1000, 0, 0, 0)$$

which gives 2.92%.

## **3.6 Further Financial Worksheet Functions in Excel**

Although Excel does have many worksheet functions we now consider just four more that are to do with the amount of interest paid in any period of a given loan. One may wish to know this for tax purpose.

### Interest and Principal

When paying back a loan each payment can be thought of as consisting of two parts:

- Payment of Interest
- Payment towards paying off the original loan

Excel provides us with functions to calculate these values for a single payment at any point in the repayment process. It also provides a function to calculate the total amount of interest paid over a given period of the loan. Similarly there is a function that calculates the total amount of principal paid back over a given period.

To see precisely what these values are we construct a repayment diagram similar to the investment diagram in Fig 3.2 (replace  $p$  with  $-p$ ) and add to the diagram the interest due at the end of each month. From Fig 3.3 we see that after the first period, at  $i = 1$ , the

period $i =$	0	1	2	3	$\dots$	$n$
Amount owed	$A$	$A(1+r) - p$	$A(1+r)^2 - p(1+r) - p$	$\dots$		
Interest due	0	$Ar$	$\{A(1+r) - p\}r$	$\{A(1+r)^2 - p(1+r) - p\}r$		

Figure 3.3: Shows the amount outstanding after each payment and the amount of interest due at each stage.

amount of interest is  $Ar$  and the amount owing is the original loan  $A$  plus the interest  $Ar$  less the payment  $p$ . It is this amount,  $A(1+r) - p$ , on which interest is charged over the next period. Thus at  $i = 2$  the interest due is this amount times  $r$ , namely  $\{A(1+r) - p\}r$ . The diagram shows this process for one more step. In general we deduce that at the end of the  $i^{th}$  period the interest that is due is given by:

$$\text{Interest} = \{A(1+r)^{i-1} - p(1+r)^{i-2} - p(1+r)^{i-2} + \dots - p\}r$$

After the first term which involves  $A$  we have a geometric progression with  $(i - 1)$  terms, first term  $-p$  and common ratio  $(1 + r)$ , thus summing this progression gives:

$$\text{Interest} = Ar(1+r)^{i-1} - pr \left\{ \frac{1 - (1+r)^{i-1}}{1 - (1+r)} \right\} = \frac{(1+r)^{i-1} \{Ar - p\} + p}{1+r} \quad (3.4)$$

Since this represents the amount of interest paid off within the single payment  $p$  after the  $i^{th}$  period the amount of principal paid off is simply  $p$  minus this value. Hence

$$\text{Principal paid} = p - \text{Interest} = \frac{(1+r)^{i-1} \{p - Ar\}}{1+r}$$

**Excel's Interest and Principal Payment Functions**

The function that calculates the amount of interest in any given payment is:

$$= \text{IPMT}(r, i, n, PV, FV, type)$$

and the function that calculates the amount of principal paid off in any given payment is:

$$= \text{PPMT}(r, i, n, PV, FV, type)$$

To convert the mathematical formula in Eq(3.4) to the Excel formula we need to make reference to Excel's sign convention. Since in Eq(3.4) the amount  $A$  is a loan it will be input in Excel as +PV, however since the repayments are moving away from us the  $p$  will be replaced with  $-PMT$ , interest with  $-IPMT$  and the principal with  $-PPMT$ .

Thus in Excel:

$$\underline{\text{IPMT}} = - \left\{ (1+r)^{i-1} \{PVr + PMT\} - PMT \right\} \quad \text{and} \quad \underline{\text{PPMT}} = PMT - \text{IPMT}$$

We note that the two functions IPMT and PPMT contain  $n$  and FV these are used internally by Excel to calculate PMT using the function  $=\text{PMT}(r,n,PV,FV,type)$ . The type parameter is as usual set equal to zero for payments at the ends of the periods and one for payments at the start. (*formula not covered here*)

In the case of paying off the whole of a loan we set  $FV=0$ , however the functions do allow us the flexibility to specify a different final value.

### Example 3.6

If I borrow \$20000 making monthly repayments at the end of each month at a monthly rate of 2/3% over 3 years then

- (a) How much interest and principal do I pay in the final payment?
- (b) How much interest and principal do I pay in the first payment?

(a) For the interest in the final payment we set  $i = 36$  and enter

$$=\text{IPMT}((2/3)\%, 36, 36, 20000, 0, 0)$$

which gives  $-\$4.15$ . The minus sign indicating, as expected, that we pay \$4.15 of interest in our final payment.

For the principal in the final payment we again set  $i = 36$  and enter

$$=\text{PPMT}((2/3)\%, 36, 36, 20000, 0, 0)$$

which gives  $-\$622.58$ , again the minus sign indicating that we are paying out this amount, in this case toward the principal.



(b) Check that the total amount of principal is indeed paid back using the CUMPRINC function.

(a) Enter:

**=CUMIPMT**((2/3)%, 36, 20000, 25, 36, 0)

Note that the last year is from month 25 to 36 inclusive. This gives –316.01 indicating that \$316.01 of interest is paid during the final year.

(b) Use CUMPRINC to check that the scheme does indeed pay of the whole loan:

Entering

**=CUMPRINC**((2/3)%, 36, 20000, 1, 36, 0)

with "start" equal to 1 and "end" equal to 36 gives –20000 indicating that the complete loan of \$20000 has been paid.

## Chapter 4

# Curve fitting - Interpolation and Extrapolation

### 4.1 Introduction

The basic problem covered by the title of this chapter is:

- Given a set of data points  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$  how can we best construct a function that in some way approximates this information. This function is then used to approximate  $y$  between the data points, **interpolation**, or to approximate  $y$  outside the range of the data points, **extrapolation**.
- Extend this problem to the case where  $y$  depends on more than one variable  $x$

One can imagine that the data points are just a sample of points generated by some unknown function  $f(x)$  and it our task to find an approximation to this **underlying function**. Such an approximation can then be used to calculate values of  $f(x)$  at points in between the given data points. For example, given just two data points the obvious construction is to draw a straight line through the two points, this can then be used to calculate the value of  $f(x)$  at some value of  $x$  not at a data point.

In Fig 4.1 we have taken the two data points  $\{(1, 2), (3, 8)\}$  and drawn the straight line  $y - 2 = 3(x - 1)$  through them. We can then use this to approximate  $f(x)$  for any value of  $x$ . e.g.  $f(2) \approx 3(2 - 1) + 2 = 5$

Continuing this idea, if we are given three data points that don't lie in a straight line then we can, as seen in Fig 4.2, draw a quadratic through the points. In this diagram we are given the data points  $\{(-1, -1), (1, 1), (2, 5)\}$  and have constructed (*by magic*) the quadratic  $y = x^2 + x - 1$  through them. This construction is unique and can be used to calculate approximate value of  $f(x)$  at given values of  $x$ . e.g.  $f(1.5) \approx (1.5)^2 + (1.5) - 1 = 2.75$ .

Indeed continuing this process, through four points we can construct a cubic, through five points a quartic and so on. In general through  $(n + 1)$  data points we will be able

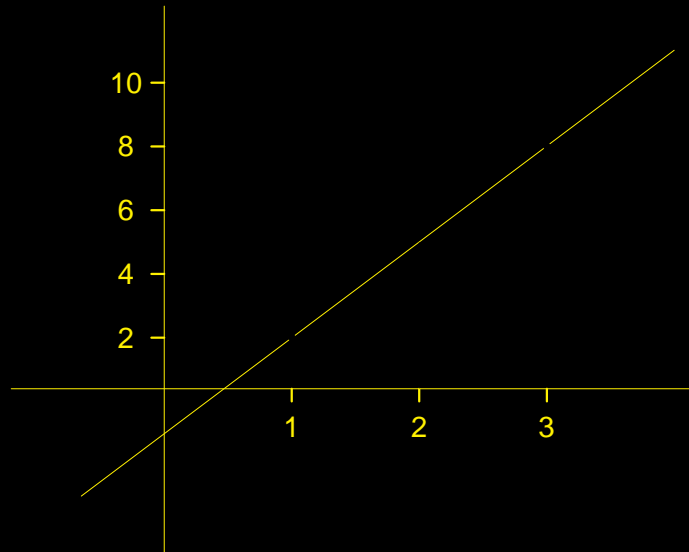


Figure 4.1: Approximation at  $x = 2$  given by  $y = 5$

to uniquely construct<sup>1</sup> a polynomial of degree at most  $n$ . There are many methods for constructing such polynomials (see Chapter 8), however when the number of data points get over about ten the polynomial constructed, although passing through all the points, often gives surprising and unrepresentative results between the data points. This is due to the fact that a polynomial of degree  $n$  may have as many as  $(n - 1)$  turning points and as seen in Fig 4.3 this may give a poor approximation to the underlying function. The diagram also shows a second possible curve which, although it doesn't pass through all the points, gives intuitively a better overall approximation. It is this type of curve that will be considered later.

### Example 4.1

This example illustrates a straightforward method for obtaining a polynomial through a set of points.

Given the points:

$$\{(-2, 3), (-1, 5), (0, 4), (1, 6), (3, 7), (4, 8)\}$$

construct a polynomial that passes through all the points.

Since there are six points we try and construct a polynomial of degree five setting

$$p(x) = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

Using the fact that the polynomial must pass through the six points generates six simultaneous linear equations which we solve for  $a_0, a_1, \dots, a_5$  as follows:

$$p(-2) = 3 = a_0 - 2a_1 + 4a_2 - 8a_3 + 16a_4 - 32a_5$$

<sup>1</sup>A polynomial of degree  $n$  is a function of the form  $p_n(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_0$  with  $a_n \neq 0$

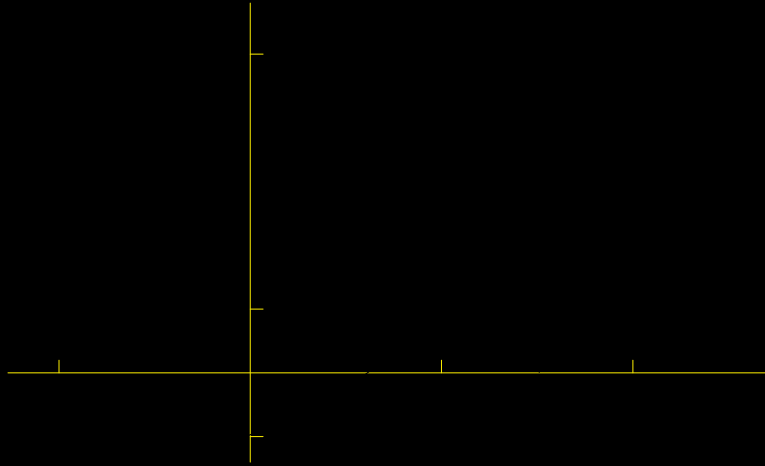


Figure 4.2: Approximation at  $x = 1.5$  given by  $y = 2.75$

$$\begin{aligned}
 p(-1) = 5 &= a_0 - a_1 + a_2 - a_3 + a_4 - a_5 \\
 p(0) = 4 &= a_0 \\
 p(1) = 6 &= a_0 + a_1 + a_2 + a_3 + a_4 + a_5 \\
 p(3) = 7 &= a_0 + 3a_1 + 9a_2 + 27a_3 + 81a_4 + 243a_5 \\
 p(4) = 8 &= a_0 + 4a_1 + 16a_2 + 64a_3 + 256a_4 + 1024a_5
 \end{aligned}$$

Writing this in the standard matrix form  $A\underline{a} = \underline{b}$  the solution is given by  $\underline{a} = A^{-1}\underline{b}$ , which in detail is as follows:

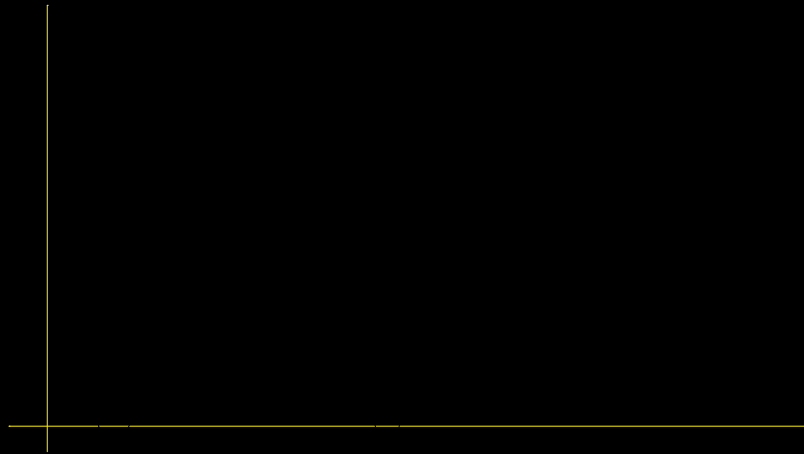
$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 4 & -8 & 16 & -32 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 27 & 81 & 243 \\ 1 & 4 & 16 & 64 & 256 & 1024 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 5 \\ 4 \\ 6 \\ 7 \\ 8 \end{pmatrix} \Rightarrow$$

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0.067 & -0.600 & -0.083 & 0.667 & -0.067 & 0.017 \\ -0.039 & 0.650 & -1.208 & 0.611 & -0.017 & 0.003 \\ -0.061 & 0.075 & 0.125 & -0.194 & 0.075 & -0.019 \\ 0.039 & -0.150 & 0.208 & -0.111 & 0.017 & -0.003 \\ -0.006 & 0.025 & -0.042 & 0.028 & -0.008 & 0.003 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ 4 \\ 6 \\ 7 \\ 8 \end{pmatrix} = \begin{pmatrix} 4 \\ 0.533 \\ 1.872 \\ -0.106 \\ -0.372 \\ 0.072 \end{pmatrix}$$

Thus expressing the coefficients correct to three decimal places the polynomial through the six points is given by:

$$p(x) = 4 + 0.533x + 1.872x^2 - 0.106x^3 - 0.372x^4 + 0.072x^5$$





This polynomial is plotted in Fig 4.4 along with the original data points. The fit looks quite reasonable for the first three points however the final three look more as if they are in a straight line. This prompts us to question just how good is this method of constructing a suitable underlying function even for six points.

The problem with fitting when we have many data points, as remarked above and illustrated in Fig 4.3, together with the final remark at the end of Eg(4.1) leads us to look for other methods of determining an underlying function  $f(x)$ .

## 4.2 Linear Spline

The simplest way to construct an underlying function is to just join adjacent data points with straight lines. This method is fine but crude. As each line segment is constructed from only two data points the overall influence of the distribution of the data points is lost when calculating an approximate value for  $y$  at a given value of  $x$ .

Given the data points  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$  we join adjacent points with a straight line, as in Fig 4.5. The straight line between the points  $(x_k, y_k)$  and  $(x_{k+1}, y_{k+1})$  is denoted by  $y = S_k(x)$  and the complete set of all line segments by  $S(x)$ . It is this complete set of all line segments that is referred to as the **linear spline** through the points.

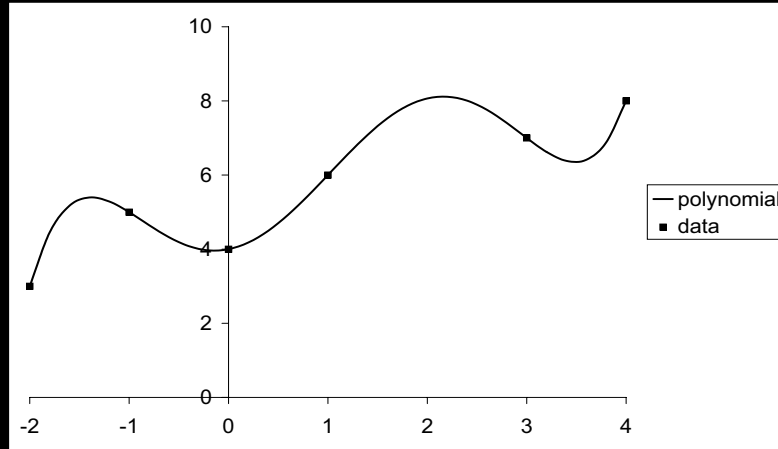


Figure 4.4: Polynomial fit to six data points.

Notationally we write:

$$S(x) = \begin{cases} S_0(x) & x_0 \leq x \leq x_1 \\ S_1(x) & x_1 \leq x \leq x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ S_{n-1} & x_{n-1} \leq x \leq x_n \end{cases}$$

Since the components  $S_k(x)$  of  $S(x)$  are straight lines their equations are given by:

$$\begin{aligned} S_0(x) &= y_0 + \left( \frac{y_1 - y_0}{x_1 - x_0} \right) (x - x_0) \\ S_1(x) &= y_1 + \left( \frac{y_2 - y_1}{x_2 - x_1} \right) (x - x_1) \\ &\vdots \\ S_k(x) &= y_k + \left( \frac{y_{k+1} - y_k}{x_{k+1} - x_k} \right) (x - x_k) \\ &\vdots \\ S_{n-1}(x) &= y_{n-1} + \left( \frac{y_n - y_{n-1}}{x_n - x_{n-1}} \right) (x - x_{n-1}) \end{aligned} \tag{4.1}$$

### Example 4.2

Using the data from table 4.1 and the results of Eq(4.1) construct the linear spline through

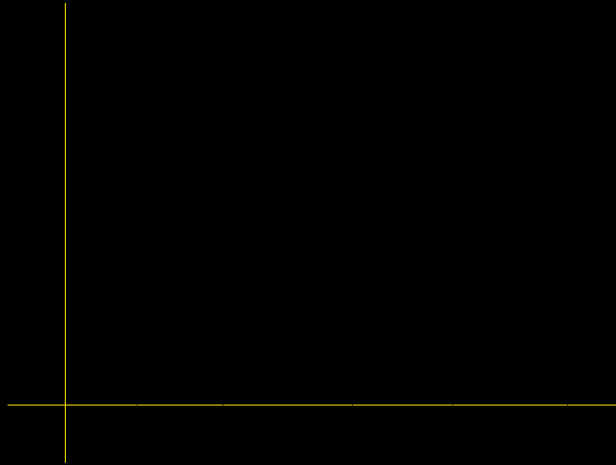


Figure 4.5: Linear spline  $S(x)$  through the data points  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$

the given points. Hence obtain an approximation to the function at  $x = 0.3$

$x_k$	0.1	0.2	0.4
$f(x) = \frac{1}{x \ln x}$	-4.343	-3.107	-2.728

Table 4.1: tabulated values of  $\frac{1}{x \ln x}$

For  $x \in [0.1, 0.2]$

$$S_0(x) = -4.343 + \frac{(-3.107 + 4.343)}{(0.2 - 0.1)}(x - 0.1) = \underline{12.36x - 5.579}$$

For  $x \in [0.2, 0.4]$

$$S_1(x) = -3.107 + \frac{(-2.728 + 3.107)}{(0.4 - 0.2)}(x - 0.2) = \underline{1.895x - 3.486}$$

Thus we can write:

$$S(x) = \begin{cases} 12.36x - 5.579 & 0.1 \leq x \leq 0.2 \\ 1.895x - 3.486 & 0.2 \leq x \leq 0.4 \end{cases}$$

Since  $x = 0.3$  is between 0.2 and 0.4 we have:

$$S(0.3) = S_1(0.3) = 1.895(0.3) - 3.486 = -2.918$$

Comparing this with the exact value of  $f(0.3)$  we have:

$$f(0.3) = \frac{1}{(0.3) \ln(0.3)} = -2.79 \quad \Rightarrow \quad |\text{error}| = 2.918 - 2.769 = 0.149$$

which represents around a 5% error.

### 4.3 Cubic Spline - natural

We now consider a more subtle way of joining the adjacent points. Rather than just joining the points with straight lines we join them with cubic polynomials. Clearly knowing two points does not provide us with sufficient information to construct a cubic. (We would need four points to construct a unique cubic). However by saying how one cubic must join to the next we can generate enough conditions. Put simply we require that at a join of two cubics the two cubics must have the same first and second derivatives. Additionally at the end points where the cubics have "free ends", we impose the condition that the second derivative must be zero. The end conditions are what classify the spline as natural, other end conditions lead to other types of cubic spline.

Denoting now by  $S_k(x)$  the cubic joining the point  $(x_k, y_k)$  to the point  $(x_{k+1}, y_{k+1})$ , point

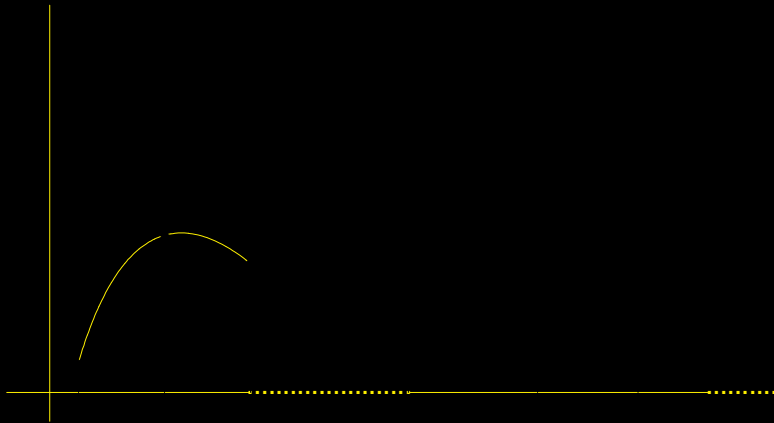


Figure 4.6:

P to Q in fig 4.6, the conditions on  $S_k(x)$  are as follows:

**Spline conditions**

(i)  $S_k(x)$  must pass through P and Q:

$$\Rightarrow S_k(x_k) = y_k \quad \text{and} \quad S_k(x_{k+1}) = y_{k+1} \quad k = 0 \dots (n-1)$$

(ii) At P  $S'_k(x_k) = S'_{k-1}(x_k) \quad k = 1 \dots (n-1)$

(iii) At P  $S''_k(x_k) = S''_{k-1}(x_k) \quad k = 1 \dots (n-1)$

(iv) To produce a natural cubic spline we set the second derivatives equal to zero at the ends of the interval. Thus  $S''_0(x_0) = S''_{n-1}(x_n) = 0$  There are no conditions on the first derivatives at these point.

Condition (i) merely ensures that the completed spline passes through all the data points.

Condition (ii) ensures that at a point where two components meet, say at P, they do so smoothly in the sense that the two have the same tangent at the point of contact.

Condition (iii) is similar to (ii) but now the second derivatives rather than the first are made equal at the point of contact. This further condition ensures that the curvatures of the two components are the same at the meeting point.

Condition (iv) has the effect of setting the curvature of the spline equal to zero at the end points. In many problems this is a reasonable thing to do.

At first sight the construction of the components  $S_k(x)$  appears quite formidable, however a great deal of ingenuity has been put into a method for making the problem quite straightforward. The first trick is to assume that the cubic is written in a certain way; instead of simply writing the cubic in the standard form  $a + bx + cx^2 + dx^3$  we express each  $S_k(x)$  as:

$$S_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \quad k = 0 \dots (n-1)$$

In total there are  $4n$  unknown coefficients, these will be calculated from the  $4n$  equations obtained by imposing the above four conditions.

- At P condition (i) implies

$$S_k(x_k) = y_k = a_k \quad k = 0 \dots (n-1)$$

Thus all the  $a$  coefficients are obtained immediately.

At Q condition (i) implies

$$S_k(x_{k+1}) = y_{k+1} = a_k + b_k(x_{k+1} - x_k) + c_k(x_{k+1} - x_k)^2 + d_k(x_{k+1} - x_k)^3 \quad k = 1 \dots n$$

Note that this gives  $2n$  equations.

- Condition (ii), after differentiating  $S_k(x)$  and  $S_{k-1}(x)$  gives:

$$S'_k(x_k) = S'_{k-1}(x_k) \Rightarrow b_k = b_{k-1} + 2c_{k-1}(x_k - x_{k-1}) + 3d_{k-1}(x_k - x_{k-1})^2 \quad k = 1 \dots (n-1)$$

Note that this gives  $(n - 1)$  equations.

- Condition (iii), after differentiating  $S_k(x)$  and  $S_{k-1}(x)$  twice gives:

$$S''_k(x_k) = S''_{k-1}(x_k) \Rightarrow 2c_k = 2c_{k-1} + 6d_{k-1}(x_k - x_{k-1}) \quad k = 1 \dots (n - 1)$$

Note that this gives  $(n - 1)$  equations.

- Finally imposing condition (iv):

$$S''_0(x_0) = 0 \Rightarrow 2c_0 = 0 \quad \text{and} \quad S''_{n-1}(x_n) = 0 \Rightarrow 2c_{n-1} + 6d_{n-1}(x_n - x_{n-1}) = 0$$

We note that our conditions have generated  $4n$  equations, precisely the same as the number of coefficients. It should therefore be possible for us to solve these equations and hence construct  $S(x)$ . The following example outlines the general method for solving the equations. Since the equations always take the same form a systematic approach is possible for all spline problems; we do not have to solve the completely general simultaneous equation problem of  $4n$  equations in  $4n$  unknowns.

### Example 4.3

Construct the natural cubic spline through the three points  $\{(1, 1), (2, -1), (4, 3)\}$ . The spline  $S(x)$  is given by:

$$S(x) = \begin{cases} S_0(x) = a_0 + b_0(x - 1) + c_0(x - 1)^2 + d_0(x - 1)^3 & 1 \leq x \leq 2 \\ S_1(x) = a_1 + b_1(x - 2) + c_1(x - 2)^2 + d_1(x - 2)^3 & 2 \leq x \leq 4 \end{cases}$$

Applying condition (i):

- $S_0(1) = 1 \Rightarrow a_0 = 1$
- $S_0(2) = -1 \Rightarrow -1 = a_0 + b_0 + c_0 + d_0$
- $S_1(2) = -1 \Rightarrow a_1 = -1$
- $S_1(4) = 3 \Rightarrow 3 = a_1 + 2b_1 + 4c_1 + 8d_1$

Applying condition (ii):

- $S'_1(2) = S'_0(2) \Rightarrow b_1 = b_0 + 2c_0 + 3d_0$

Applying condition (iii):

- $S''_1(2) = S''_0(2) \Rightarrow 2c_1 = 2c_0 + 6d_0$

Applying condition (iv)

- $S_0''(1) = 0 \Rightarrow 2c_0 = 0$
- $S_1''(4) = 0 \Rightarrow 2c_1 + 12d_1 = 0$

We now consider the solution of these eight equations in the following systematic fashion:

- In all case the  $a_k$  coefficients are given straightaway by  $a_k = y_k$  and  $c_0$  is always zero for the natural cubic spline. Thus in general we immediately have  $(n + 1)$  of the unknowns. In this example:

$$c_0 = 0 \quad a_0 = 1 \quad a_1 = -1$$

- From the conditions on  $S_k''(x)$  we can always write the  $d$  coefficients in terms of the  $c$  coefficients. In this example:

$$d_0 = \frac{2c_1}{6} = \frac{c_1}{3} \quad d_1 = -\frac{2c_1}{12} = -\frac{c_1}{6}$$

- From the results of condition (i) and substituting in the known  $a$  values and  $c_0 = 0$ , we can always write the  $b$  coefficients in terms of the  $c$ 's and  $d$ 's. If we now substitute for the  $d$ 's in terms of the  $c$ 's, from above, we obtain the  $b$ 's in terms of the  $c$ 's. In this example:

$$b_0 = -2 - c_0 - d_0 = -2 - \frac{c_1}{3} \quad b_1 = 2 - 2c_1 - 4d_1 = 2 - 2c_1 + \frac{2c_1}{3} = 2 - \frac{4c_1}{3}$$

- At this point we have all the  $b$ 's and  $d$ 's in terms of the  $c$ 's. We now substitute for the  $b$ 's and  $d$ 's into the equations formed from condition (ii) (ie the constraints on  $S_k'(x)$ .) These equations will in general be solvable for the  $c$ 's. By substituting back we can then calculate the  $b$ 's and  $d$ 's. In this example:

$$b_1 = b_0 + 2c_0 + 3d_0 \Rightarrow 2 - \frac{4c_1}{3} = -2 - \frac{c_1}{3} + 3\frac{c_1}{3} \Rightarrow c_1 = 2$$

- Finally substituting back with  $c_1 = 2$  gives

$$b_1 = -\frac{2}{3} \quad b_0 = -2 - \frac{2}{3} = -\frac{8}{3} \quad d_1 = -\frac{1}{3} \quad d_0 = \frac{2}{3}$$

The spline  $S(x)$  is drawn in fig 4.7 and given by:

$$S(x) = \begin{cases} S_0(x) &= 1 - \frac{8}{3}(x-1) + \frac{2}{3}(x-1)^3 & 1 \leq x \leq 2 \\ S_1(x) &= -1 - \frac{2}{3}(x-2) + 2(x-2)^2 - \frac{1}{3}(x-2)^3 & 2 \leq x \leq 4 \end{cases}$$

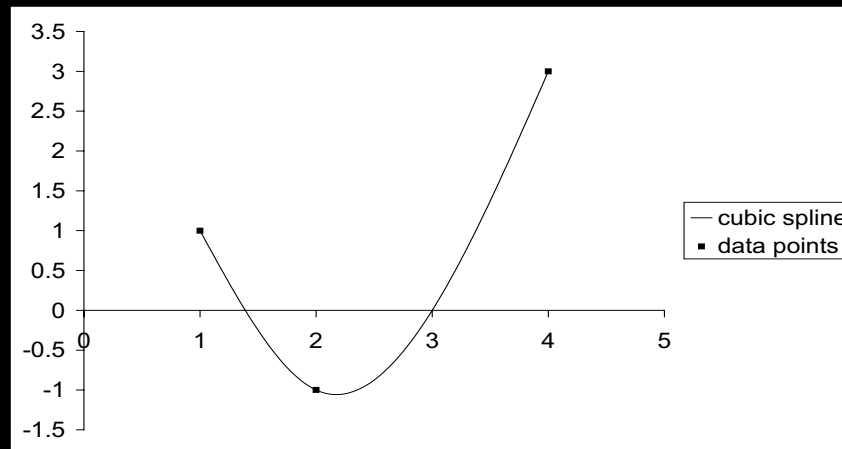


Figure 4.7: Cubic Spline for Example (4.3)

## 4.4 Linear Least Squares Fitting

### 4.4.1 Linear Least Squares - Regression - Single Variable

As we shall see in a later section the method of fitting a curve using the least squares criteria is a very general concept. In this section we consider fitting a straight line to a set of data; this is sometimes referred to as regression. This idea may at first sight seem daft, as clearly it is not likely that the set of given points will lie in a straight line. However if we now drop the condition that the fitted curve should pass through all, or indeed any of the data points, we can imagine trying to construct a straight line that in some sense gets close to as many points as possible whilst at the same time indicating the general trend of the data. The problem therefore is to specify mathematically, in a unique fashion, a way of constructing such a line from the given data. Fig 4.8 shows such a construction; although it is not a good fit as far as the last data point is concerned, however this is more that compensated for by its closeness to the other points.



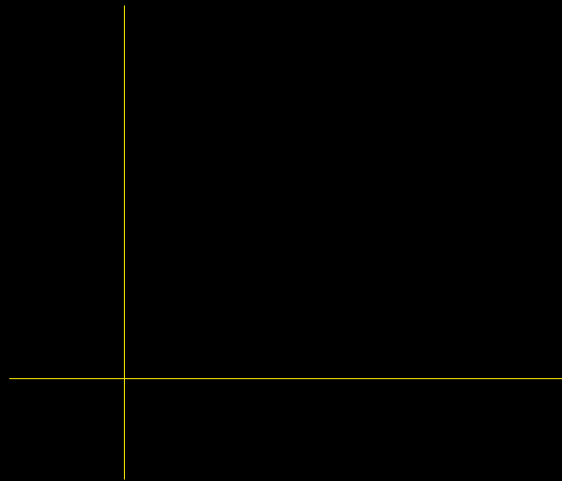


Figure 4.8: *Straight line fitted to a set of data points*

#### 4.4.2 Least Squares Criteria for Straight Line Fitting

In words the least square criteria for fitting a straight line to a set of data is given as:

The line is drawn such that the sum of the squares of the vertical distances of each data point from the line is a minimum.

Mathematically if the data points are  $\{(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)\}$  and the equation of the line is given by  $y = mx + c$  then we require  $m$  and  $c$  such that:

$$S = \sum_{i=0}^N (mx_i + c - y_i)^2 \quad \text{is a minimum} \quad (4.2)$$

In Fig 4.9, where we have only four data points, this sum is the sum of the squares of the dotted lengths. In particular if the point  $Q$  has coordinates  $(x_i, y_i)$  then the point vertically above it,  $P$ , lying on the line  $y = mx + c$  has coordinates  $(x_i, mx_i + c)$ . Thus  $PQ^2 = (mx_i + c - y_i)^2$ . These quantities are then added together for each data point to give in general the expression in Eq(4.2).

To find the values of  $m$  and  $c$  that minimise  $S$  can be achieved either using an algebraic approach, since  $S$  only depends quadratically on  $m$  or  $c$ , or we can extend the ideas of maximum and minimum from the calculus of one variable to that of two. Although it doesn't require any new material the algebraic method is long and tedious, the shortest approach is to use calculus for two variables. For this purpose the function  $S$  is considered to depend on the two variables  $m$  and  $c$ , thus extending the notation the sum in Eq(4.2) is denoted as  $S(m, c)$ . Since  $S(m, c)$  is the sum of squares it will be positive and since it only depends quadratically on its two variables it will be in the shape of an open bowl. This is shown in Fig 4.10 where  $z = S(m, c)$  has been plotted.

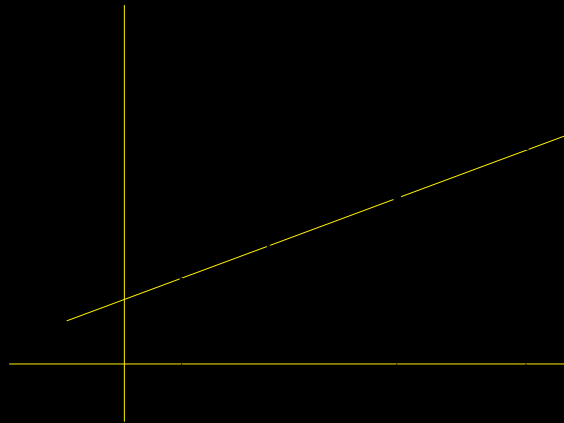
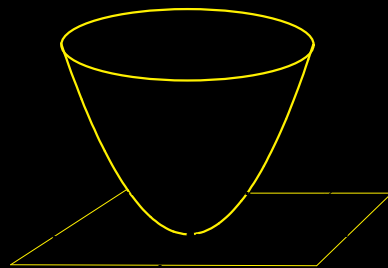


Figure 4.9: Dotted lines show the difference between the data points and the fitted line

(This is an assumption but it will mean that we will not have to prove that the point we locate as a possible extremum is a minimum rather than a maximum, or indeed the equivalent of a stationary point of inflection in two dimensions.)

In Fig 4.10 we can see that at a minimum all the tangents to the surface will be horizontal and in particular the gradients of the tangents in the  $m$  and  $c$  directions will be zero.

Thus using the work on partial derivatives from Chapter 2 this implies that at a mini-



mum  $S_m(m, c)$  and  $S_c(m, c)$  are both zero.<sup>2</sup> This is exactly the same scenario as in one dimension where turning points of  $f(x)$  are located by solving  $f'(x) = 0$ . Thus from Eq(4.2) we have:

$$S_m(m, c) = \sum_{i=0}^N 2(mx_i + c - y_i)x_i = 0 \Rightarrow m \sum_{i=0}^N x_i^2 + c \sum_{i=0}^N x_i - \sum_{i=0}^N x_i y_i = 0 \quad (4.3)$$

$$S_c(m, c) = \sum_{i=0}^N 2(mx_i + c - y_i) = 0 \Rightarrow m \sum_{i=0}^N x_i + \sum_{i=0}^N c - \sum_{i=0}^N y_i = 0 \quad (4.4)$$

Dividing Eq(4.3) and Eq(4.4) by  $(N + 1)$ , introducing the mean<sup>3</sup> of the  $x$  values as  $\bar{x}$  and the mean of the  $y$  values as  $\bar{y}$  and noting that in Eq(4.4),  $\sum_{i=0}^N c = c(N + 1)$ , we obtain:

$$m \sum_{i=0}^N \frac{x_i^2}{(N + 1)} + c\bar{x} - \sum_{i=0}^N \frac{x_i y_i}{(N + 1)} = 0 \quad (4.5)$$

$$m\bar{x} + c - \bar{y} = 0 \quad (4.6)$$

Eq(4.6) is of particular interest as it tells us that the straight line fit,  $y = mx + c$ , passes through the points with coordinates given by the means of the  $x$  and  $y$  data. That is to say through the centroid of the data points. This is very encouraging since it is reasonable to expect that in some way the straight line will correspond with the mean of the data.

As Eq(4.5) and Eq(4.6) are simply two linear simultaneous equations in  $m$  and  $c$  we can solve to give:

$$m = \frac{\sum_{i=0}^N \frac{x_i y_i}{(N + 1)} - \bar{x} \bar{y}}{\sum_{i=0}^N \frac{x_i^2}{(N + 1)} - \bar{x}^2} \quad (4.7)$$

$$c = \bar{y} - m\bar{x} \quad (4.8)$$

#### 4.4.3 Linear Least Squares - Multiple Regression - Many Variables

As we are intending to use Worksheet Functions in Excel to calculate the coefficients  $m$  and  $c$  in the above it is worth first remarking on the natural extension of the linear problem to several variables. Here it is assumed that the dependent variable  $y$  depends on several variables  $\{x_1, x_2, \dots, x_k\}$  and that we are given a set of data points each of

<sup>2</sup>In general  $f(x_1, x_2, \dots, x_k)$  is stationary at  $\underline{x} = \underline{a}$  if all  $k$  partial derivatives of  $f(\underline{x})$  vanish at  $\underline{a}$ .

<sup>3</sup>since the summations run from 0 to  $N$  we have  $(N + 1)$  data points, thus  $\bar{x} = \sum_{i=0}^N \frac{x_i}{(N+1)}$  etc

the form  $(x_1, x_2, \dots, x_k, y)$ . For example if  $y$  depends on two variables,  $x_1$  and  $x_2$ , the data points are of the form  $(x_1, x_2, y)$  and geometrically are represented by points in three dimensional space. In general we find the linear function  $y = m_1x_1 + m_2x_2 + \dots + m_kx_k + c$  that best fits the data in the sense of least squares. In the case of  $y$  depending on two variables this amounts to minimising the sum of the squares of the vertical distances of the data points from the plane  $y = m_1x_1 + m_2x_2 + c$ . (ie the sum of the squares of the dotted lines in Fig 4.11)

In Fig 4.11, where  $y$  depends on two variables, it is necessary to label each data point

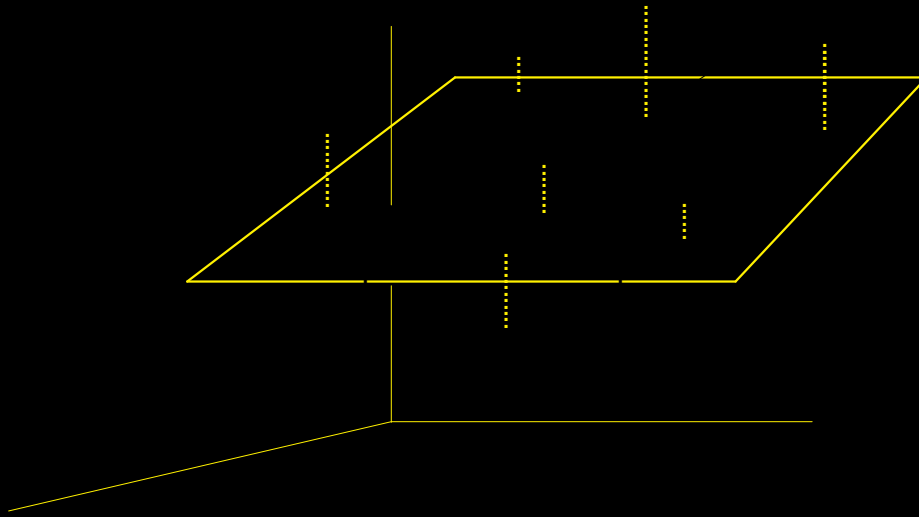


Figure 4.11:

as well as each independent variable, thus we introduce the index  $i$ . Each data point is therefore of the form  $({}^ix_1, {}^ix_2, {}^iy)$ , with the index  $i$  running from 0 to  $N$ . The least squares criteria now requires the determination of  $m_1, m_2$  and  $c$  such that:

$$S = \sum_{i=0}^N [(m_1 {}^ix_1 + m_2 {}^ix_2 + c) - {}^iy]^2 \quad \text{is a minimum.}$$

It is clear that the notation is getting rather involved, as in general we have to both label the variables  $\{x_1, x_2, \dots, x_k, y\}$  and the data points. Thus the general  $i$ th data point, in the case where  $y$  depends on  $k$  variables, is denoted as  $({}^ix_1, {}^ix_2, \dots, {}^ix_k, {}^iy)$  and the least square criteria for fitting  $y = m_1x_1 + m_2x_2 + \dots + m_kx_k + c$  means finding  $\{m_1, m_2, \dots, m_k, c\}$  such that:

$$S = \sum_{i=0}^N [(m_1 {}^ix_1 + m_2 {}^ix_2 + \dots + m_k {}^ix_k + c) - {}^iy]^2 \quad \text{is a minimum} \quad (4.9)$$

The determination of the parameters  $\{m_1, m_2 \dots m_k, c\}$  follows the same mathematical steps for general  $k$  as it does for the case  $k = 1$  carried out in Sec(4.4.2). We formulate  $(k + 1)$  linear simultaneous equations by equating to zero the partial derivatives of  $S$  with respect to the  $c$  parameter and each of the  $m$  parameters. These equations are then solved for the  $(k + 1)$  parameters  $\{m_1, m_2 \dots m_k, c\}$ . Although this only establishes the values of the parameters at which  $S$  is stationary the fact that  $S$  is simply the sum of squared terms implies that this stationary point is a minimum<sup>4</sup>.

#### 4.4.4 Excel - Linear Regression

As we would expect we will not be required to carry out any of the above calculations by hand. Excel provides us with several Worksheet Functions that allow us in the most general case to carry out both linear and nonlinear multiple variable fitting.

##### 1. =LINEST(*known y values, known x values, c constant, statistics*)

In general LINEST assumes that the data points are all of the form  $({}^i x_1, {}^i x_2, \dots, {}^i x_k, {}^i y)$  and finds the values of  $\{m_1, m_2 \dots m_k, c\}$  such that

$$y = m_1 x_1 + m_2 x_2 \dots + m_k x_k + c$$

is a best fit using the least squares criteria.

The parameter of LINEST are as follows:

- *known y values* A range of cells identifying a single column of  $y$  values.
- *known x values* A range of cells identifying  $k$  columns, one column for each of the  $x$  variables.
- *c constant* Optional parameter set to be **true** or **false**. If **true** or omitted then  $c$  is calculated normally as above. If **false** then  $c$  is set to zero and the  $m$  parameters are found such that  $y = m_1 x_1 + \dots m_k x_k$  is the best least squares fit. (*we will seldom use this option*)
- *statistics* Optional parameter set to be **true** or **false**. If **false** or omitted then LINEST returns only  $c$  and the  $m$  values. If **true** then LINEST also returns a set of regression statistics that enable us to assess the goodness of the approximation. (see Sec(4.4.5))

LINEST is an array function and will therefore return the values  $m_1, \dots m_k, c$  together with certain statistics which can be used to judge how good a fit has been achieved. This last facility is discussed later in Sec(4.4.5).

With the optional parameters omitted LINEST places in a preselected row and in the following order the  $(k + 1)$  values  $\{m_k, m_{k-1} \dots m_1, c\}$ .

---

<sup>4</sup>It is possible to analyse the nature of a stationary point for a function of several variables using a multiple variable version of the second derivative test; this is not carried out here.

$m_k$	$m_{k-1}$	$\dots$	$\dots$	$m_1$	$c$
-------	-----------	---------	---------	-------	-----

As with any array function to implement LINEST you need to first highlight the required space, in this case  $(k + 1)$  cells in any vacant row, and finally, after entering the  $y$  and  $x$  data ranges into LINEST, use Ctrl-Shift-Enter to obtain the results.

The following example omits both the optional parameters and therefore only finds the  $c$  and  $m$  values. These are used in  $y = m_1x_1 + c$  to create a set of values to plot on the same graph as the original data.

#### Example 4.4

Find the least squares linear fit to the following data points:

$x =$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$y =$	0	0.100	0.199	0.296	0.389	0.479	0.565	0.644	0.717	0.783

- Enter the  $x$  values in cells A1:A10 and the  $y$  values in cells B1:B10.
- Highlight cells A13:B13
- enter **=LINEST(B1:B10,A1:A10)** followed by **Ctrl-Shift-Enter**

Cell A13 should now contain  $m_1$  and cell B13 contain  $c$ . This should give the straight line fit as:

$$y = (0.878061)x + (0.022073) \quad (4.10)$$

To illustrate the approximation plot the graph of the data points and the graph generated by the straight line in Eq(4.10) as follows:

- Set up a column of  $y$  values using  $y = m_1x + c$ :
  - Enter **=\$A\$13\*A1+\$B\$13** in cell C1
  - Copy this down to cell C10.
  - Cells C1 to C10 should now contain the predicted  $y$  values.
- Plot the original data:
  - Highlight cells A1 to B10.
  - Select the Chart Wizard.
  - Select Scatter
  - Select the sub type that produces just points (first in the list of subtypes)
  - Select Finish
- Plot the straight line fit on the same graph:

- Highlight and Copy the  $y$  values C1 to C10
- Click on the chart and Paste the values just copied. The points will be plotted with the same  $x$  values as the last set of data. (*this is the simplest way of adding data to a chart and is appropriate when the new data has exactly the same set of  $x$  values as the existing set.*)
- To change the points to a straight line display **right-click** on one of the new data points (*make sure this is flagged as Series 2*); Select Chart Type; Select the subtype that give only straight lines. (*the last of the options.*) Select OK.

You should now have the graph in Fig(4.12).

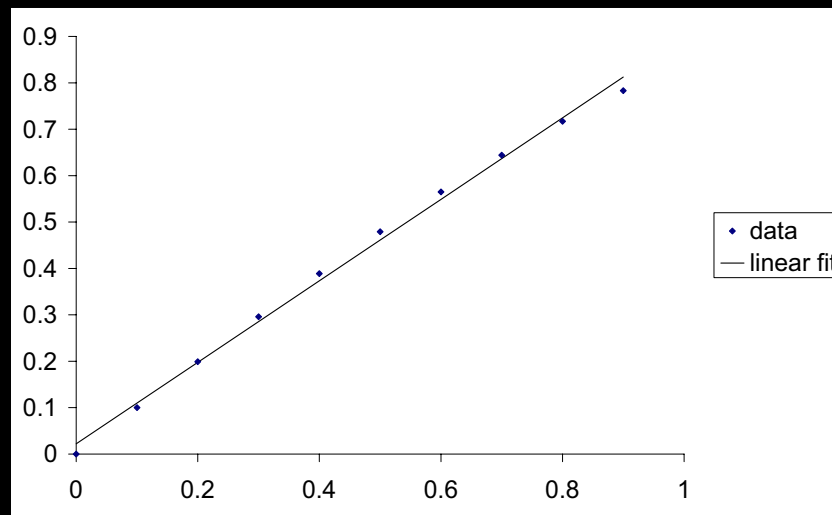


Figure 4.12: *Least squares linear fit to the set of ten data points in Eg(4.4)*

**Example 4.5**

Given the following data obtain the best least squares fit of the form:

$$y = m_1x_1 + m_2x_2 + m_3x_3 + c$$

Hence obtain an approximation to  $y$  at  $x_1 = x_2 = x_3 = 4$ .

$x_1 =$	1	1.7	2.3	-0.5	2.3	3.7	-3.0	2.5	2.0	1.6	3.0
$x_2 =$	2.1	1.0	1.7	0.3	0.5	0.6	-1.2	-0.2	2.0	3.1	-1.0
$x_3 =$	1.0	2.1	3.0	3.1	3.6	1.4	5.0	1.2	6.0	7.1	-3.0
$y =$	2.17	2.272	2.795	1.515	2.519	2.419	0.611	1.911	3.169	3.356	0.950

- Enter the  $x_1$  values in cells A1:A11,  $x_2$  values in B1:B11, the  $x_3$  values in C1:C11 and the  $y$  values in cells D1:D11.
- Highlight cells A13:D13
- Enter `=LINEST(D1:D11,A1:C11)` followed by **Ctrl-Shift-Enter**
- A13 contains  $m_3$ , B13 contains  $m_2$ , C13 contains  $m_1$  and D13 contains  $c$ .

Numerically the above process gives:

$$y = (0.155)x_3 + (0.309)x_2 + (0.273)x_1 + 1.062$$

Thus the value of  $y$  at  $x_1 = x_2 = x_3 = 4$  is given by

$$y = (0.155)(4) + (0.309)(4) + (0.273)(4) + 1.062 = \underline{3.948}$$

In an example of this type, where there are more than two independent  $x$  variables, it is clearly not possible to draw a graph of  $y$  against the  $x$  values; thus there is not a realistic geometric interpretation of the results as in Fig 4.9 and Fig 4.11.

**2. =TREND(known  $y$  values, known  $x$  values, new  $x$  values,  $c$  constant)**

In Ex 4.4 it was necessary to calculate the values of  $y$  at given values of the  $x$  variables using the values of  $m_1$  etc given by LINEST. The **TREND** function will fit the function  $y = m_1x_1 + \dots + m_kx_k + c$  in exactly the same manner as LINEST but **not** reveal the  $c$  and  $m$  values. Instead it will store the  $c$  and  $m$  values and then use them to calculate  $y$  at a new given set of  $x$  values.

In detail:

- *known  $y$  values* A range of cells identifying a single column of  $y$  values.
- *known  $x$  values* A range of cells identifying  $k$  columns, one column for each of the  $x$  variables.
- *new  $x$  values* A range of cells identifying  $k$  columns, one column for each of the  $x$  variables at which you want the  $y$  values to be evaluated.



- *c constant* Optional parameter set to be **true** or **false**. If **true** or omitted then  $c$  is calculated normally as above. If **false** then  $c$  is set to zero and the  $m$  parameters are found such that  $y = m_1x_1 + \dots + m_kx_k$  is the best least squares fit. (*we will seldom use this option*)

The following example illustrates this function for the data in Ex 4.4.

### Example 4.6

Given the following data points:

$x =$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$y =$	0	0.100	0.199	0.296	0.389	0.479	0.565	0.644	0.717	0.783

complete the table:

$X =$	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7
$Y =$															

where  $Y = m_1X + c$  is the least squares fit to the original data.

- Enter the  $x$  values in cells A1:A10 and the  $y$  values in cells B1:B10.
- Enter the  $X$  values in cells C1:C15
- Highlight cells D1: D15
- enter **=TREND(B1:B10,A1:A10,C1:C15)** followed by **Ctrl-Shift-Enter**

Cells C1:D15 should now contain the required values and can be used to plot the least squares approximation given by  $Y = m_1X + c$ . The values of  $m_1$  and  $c$  are not available to the user when using the TREND function.

### 3. Trendline and Chart

If we have already plotted a chart of the data then Excel will allow us to add a straight line least squares fit directly onto the chart. By selecting the correct options the equation of the fit is also given.

To add a trendline:

- Plot the given  $(x, y)$  data on a chart: use XY(scatter) and points only for the best effect.
- Activate the chart by left clicking just inside the boundary of the chart.
- Select **Chart** from the top menu. Select **Add Trendline**.
- Select **linear** OK (*there are other options which we discuss below*)

### 4. =FORECAST( $x$ , known $y$ values, known $x$ values)

If  $y$  depends on a single  $x$  variable then the function FORECAST will give the value of  $y = m_1x + c$  for a given value of  $x$ . The function does not reveal the value of  $c$  or  $m_1$  or indeed any other values of  $y$ . In detail:

- $x$ : the value of  $x$  at which you wish to evaluate  $y$  using the linear least squares fit to the data.
- *known  $y$  values* the given data value for  $y$
- *known  $x$  values* the given data value for  $x$

With reference to Ex 4.4 we found in Eq(4.10) that  $y = (0.878061)x + (0.022073)$ , thus at  $x = 1$ , for example,  $y = 0.900134$ .

Entering =FORECAST(1,B1:B10,A1:A10) into the worksheet gives, as expected, 0.900134.

#### 4.4.5 Goodness of Fit

There are many ways of assessing how good a fit has been achieved. Setting the options to be true LINEST returns a lot of statistical data that can be used for this purpose. Here two of the more straightforward indicators given by LINEST are considered, namely the coefficient of determination and the standard error of the  $y$ -estimate. The coefficient of determination is related to the correlation coefficient and the standard error to the standard deviation of the difference between the  $y$ -data and the  $y$ -estimate. Although some discussion of the meaning of these statistical concepts is carried out here it is far from complete and the student is advised to consult any standard statistics text for further reference.

##### Standard Error

Consider here the problem of fitting the regression line  $Y = mx + c$  to the data  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$  by the method of least squares. The **standard error**<sup>5</sup> of the estimate  $Y$  is defined as

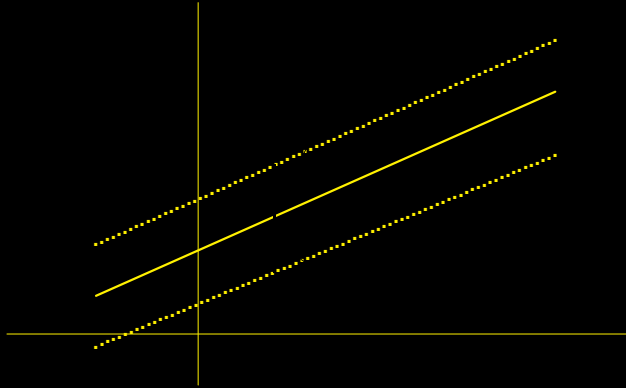
$$SE_y = \sqrt{\frac{\sum_{i=0}^n (y_i - Y_i)^2}{(n - 1)}}$$

Recall that the straight line through the data was calculated such that  $\sum_{i=0}^n (y_i - Y_i)^2$  was a minimum. Hence of all the straight lines passing amongst the data the regression line is the line that minimises  $SE_y$ . We are therefore interested in fits that give as small a value as possible for  $SE_y$ .

Statistically  $SE_y$  corresponds to an estimate of the standard deviation of the error variable  $E_i = (y_i - Y_i)$ . Indeed under certain assumptions concerning the population from

<sup>5</sup>In a more general discussion this definition of standard error could be used for any type of fit, rather than just linear. However the comments that follow this definition are only relevant to linear fitting

which  $E_i$  is selected we can state that with increasing  $n$  we can expect about 65% of the data points to lie within the dotted band in Fig 4.13. This is deliberately a rather vague statement as the details of this result are not covered in this text. Although  $SE_y$  and



the idea of 65% of the data points lying within  $SE_y$  of the regression line is useful the actual value of  $SE_y$  is less informative. Clearly calculating  $SE_y$  for a data set consisting of large values of  $y$  will yield a large value of  $SE_y$ , whereas if we calculate it for a data set consisting of small values of  $y$  it will yield a small value of  $SE_y$ . It may however be the case that we have fitted a line to the data set of large values much better than to the data set of small values. Thus it is not sensible to try and use  $SE_y$  as an index of goodness of fit for all data sets. To this end the Coefficient of Determination and its relation to the Correlation Coefficient are discussed.

#### Coefficient of Determination ( $r^2$ ) - Correlation Coefficient ( $r$ ).

A measure of dispersion of the data about the curve of best fit that doesn't suffer from the  $y$ -data scaling discussed above is the **coefficient of determination**<sup>6</sup>. When fitting  $Y = mx + c$  to the data set  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$  the coefficient of determination is defined as:

$$r^2 = \frac{\sum_{i=0}^n (\bar{y} - Y_i)^2}{\sum_{i=0}^n (\bar{y} - y_i)^2} \quad (4.11)$$

The coefficient is denoted by  $r^2$  to stress the fact that it is always positive or zero, this follows from the fact that it consists of the quotient of the sum of squared terms. Additionally  $\bar{y}$  denotes the mean of the  $y$ -data.

<sup>6</sup>As with  $SE_y$  this definition can be applied to any type of fitting, however we only discuss linear fitting in this text.

The coefficient compares the spread of the fitted  $Y_i$  values about  $\bar{y}$ , the mean of the data, with the spread of the  $y$ -data values  $y_i$  about the data mean  $\bar{y}$ .

Although **not proved here** the significance of  $r^2$  is summarised by the following results:

- If the data lies on the line then  $y_i=Y_i$  and thus from Eq(4.11)  $r^2 = 1$ . This is in fact the largest value of  $r^2$  and corresponds to perfect agreement between the fit and the data.
- As the fit gets worse then  $r^2$  decreases, with the worst case being indicated by  $r^2 = 0$
- Under the assumption that we have carried out a linear fit according to the least squares criteria  $r^2$  can be rewritten using just the data points as:

$$r^2 = \frac{\left( \sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=0}^n (x_i - \bar{x})^2 \sum_{i=0}^n (y_i - \bar{y})^2}$$

- Statisticians use the Pearson product-moment correlation coefficient to determine how good a linear relation exists between two sets of data. This is **defined** to be:

$$r = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2 \sum_{i=0}^n (y_i - \bar{y})^2}}$$

As can be seen this is just the square root of the coefficient of determination, however it can take either a positive or negative sign. Since from above  $0 \leq r^2 \leq 1$  the range of values of the correlation coefficient is given by  $-1 \leq r \leq 1$ .

It can be shown that the sign of  $r$  and the sign of the gradient of the best line of fit are the same (*not the values, just the signs*). Thus a value of  $r$  close to  $-1$  represents a good linear relationship between the  $x$ - $y$  data values such that as  $x$  increases the  $y$  values tend overall to decrease. Alternatively a value of  $r$  close to  $+1$  represents a good linear relationship between the  $x$ - $y$  data values such that as  $x$  increases the tendency in the  $y$  data is to also increase.

### Excel's Statistics

As already mentioned in Sec(4.4.4) when using LINEST in order to display the  $c$  and  $m$  values together with all the fitting statistics one needs to enter

=LINEST(*y-values*, *x-values*, TRUE, TRUE).

For the case of fitting  $y = m_1x_1 + m_2x_2 + \dots + m_kx_k + c$  to  $(n + 1)$  data points, each of the form  $(x_1, x_2, \dots, x_k, y)$ , LINEST will return an array with 5 rows and  $k + 1$  columns containing the following:

$m_k$	$m_{k-1}$	...	$m_1$	$c$
$se_k$	$se_{k-1}$	...	$se_1$	$se_c$
$r^2$	$SE_y$			
$F$	df			
$S_{reg}$	$S$			

Thus before entering the array function LINEST it will be necessary to highlight an array of 5-rows and  $k + 1$ -columns. The output is as follows with the items covered in this text in bold type:

- **The first row contains the  $c$  and  $m$  fitting parameters which have been seen before and can be displayed by entering LINEST without the two extra parameters.**
- The second row consists of standard errors for each of the parameters; these can be ignored as they are not discussed in this text.
- **The third row contains the two important parameters discussed above. Namely the coefficient of determination  $r^2$  and the standard error of the  $y$  estimate  $SE_y$ .**
- For the purpose of hypothesis testing the fourth row contains the  $F$  statistic and df, the number of degrees of freedom for the system.  
(neither are covered here but it is worth noting that the number of degrees of freedom for the system =  $(n - k)$ )
- Row five contains some useful sums of squares.

$S_{reg}$  is the sum of the squares of the differences between the fitted  $y$ -values and the mean of the  $y$ -data. This appears in the numerator of the definition of  $r^2$  in Eq(4.11).

$S$  is the sum of the squares of the differences between the fitted  $y$ -values and the data  $y$ -values as given by Eq(4.9).

A result that is sometimes of use is that if we denote the denominator of the expression in the definition of  $r^2$  by  $SS$ , that is to say let:

$$SS = \sum_{i=0}^n (\bar{y} - y_i)^2$$

then it can be shown that

$$SS = S + S_{reg}$$

## 4.5 Non-Linear Least Squares Fitting

In the above section the least squares fitting was restricted to a linear function of the independent variables. We found the values of the parameters  $\{m_1, m_2, \dots, m_k, c\}$  such that  $y = m_1x_1 + \dots + m_kx_k + c$  best fits the data according to the criteria of least squares. It is possible to generalise this idea in several different ways. For simplicity the following only uses the function LINEST, even though Excel offers other worksheet functions that carry out procedures similar to TREND. With this approach it is possible to produce all the information about the fitting process and hence use it in a variable and informative manner.

### 4.5.1 Polynomial trendline (*y depending on a single x variable only*)

The following problem is considered:

Given the data set  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$

Fit the polynomial  $y = c + m_1x + m_2x^2 + \dots + m_kx^k$

The method involves relabelling the powers of  $x$  as our  $x$ -data and then applying LINEST. The original data points are of the form  $(x, y)$  however if we introduce  $X_i = x^i$  then we can apply LINEST to the data points  $(X_1, X_2, \dots, X_k, y)$ . Thus LINEST will find the  $c$  and  $m$  values such that

$$y = m_1X_1 + m_2X_2 + \dots + m_kX_k + c$$

is a best fit which with  $X_i = x^i$  is the same as saying

$$y = m_1x + m_2x^2 + \dots + m_kx^k + c$$

is a best fit. For example if it is decided to fit a cubic to 10 values, that is to say  $k = 3$  and  $n = 9$ , the process would be to create four columns of values, one for  $X_1 = x$ , one for  $X_2 = x^2$ , one for  $X_3 = x^3$  and one for  $y$ . LINEST is then applied to these data as follows:

- In A1:A10 enter the given  $x$  values,  $\{x_0, x_1 \dots x_9\}$   
(column A contains the  $X_1 = x$  values for LINEST)
- In B1 enter =A1^2, and copy down<sup>7</sup> to B10.  
(column B now contains the  $X_2$  values =  $x^2$ )
- In C1 enter =A1^3, and copy down to C10  
(column C now contains the  $X_3$  values =  $x^3$ )
- In D1:D10 enter the given  $y$  values,  $\{y_0, y_1 \dots y_9\}$ .
- Highlight E1:H1, Enter =LINEST(D1:D10,A1:C10), use **Ctrl-Shift-Enter** to give the values of  $m_3, m_2, m_1$  and  $c$  in E1 to H1 respectively.  
(this applies LINEST to the data set of the form  $(X_1, X_2, X_3, y)$ )

<sup>7</sup>copy down by grabbing the + sign in the righthand corner of B1, drag and drop it to B10

**Example 4.7**

Consider the ten data points from Ex 4.4.

$x =$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$y =$	0	0.100	0.199	0.296	0.389	0.479	0.565	0.644	0.717	0.783

- Create the four columns as above in cells A1:D10.
- Highlight E1:H1, enter **=LINEST(D1:D10,A1:C10)**, use **Ctrl-Shift-Enter**.
- E1 contains  $m_3 = -0.1449$ , F1 contains  $m_2 = -0.0188$   
G1 contains  $m_1 = 1.0042$  and H1 contains  $c = 0.0000$
- the polynomial fit is given by  $y = 1.0042x - 0.0188x^2 - 0.1449x^3$

Excel provides us with a shortcut that saves us creating the columns containing the extra powers of  $x$ . If the data is as above, with the  $x$  values in A1:A10 and the  $y$  values in D1:D10, then

$$\mathbf{=LINEST(D1:D10,A1:A10\wedge\{1,2,3\})}$$

gives the required  $c$  and  $m$  values.

**4.5.2 Power, Exponential and Logarithmic trendlines**

In this section the following three problems are considered:

Given the data set  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$

Fit one of:

<b>power curve:</b>	$y = bx^m$
<b>exponential curve:</b>	$y = be^{mx}$
<b>logarithmic curve:</b>	$y = m \ln x + c$

In the first two cases a simple log transformation will indicate immediately the variable to use in LINEST:

**Power curve<sup>8</sup>**

$$y = bx^m \Rightarrow \ln y = m \ln x + \ln b \Rightarrow Y = mX + c$$

where  $Y = \ln y$ ,  $X = \ln x$  and  $c = \ln b$ . Thus applying LINEST to two new columns of data given by  $X = \ln x$  and  $Y = \ln y$  will give values for  $m$  and  $c$  which can then be used to calculate  $m$  and  $b$ .

Excel provides a similar shortcut as found with the polynomial fitting. If the  $x$  data is in A1:A10 and the  $y$  data is in B1:B10 then we simply use:

<sup>8</sup>The method of least squares is used to fit the transformed curve  $Y = m_1X + c$  to the transformed data. However it should be noted that using the value of  $m_1$  and  $b = \ln c$  so found does not give a least squares fit of the original function to the original data.

**= LINEST(LN(B1:B10),LN(A1:A10))**

rather than creating two new data columns.

(recall that *LINEST* is an array function and in this case must be entered into two adjacent cells to give  $m$  and  $c$ )

### Exponential<sup>8</sup>

Using an argument similar to that of the power curve;

$$y = be^{mx} \Rightarrow \ln y = mx + \ln b \Rightarrow Y = mx + c$$

where  $Y = \ln y$  and  $c = \ln b$ . Thus applying *LINEST* to the original  $x$  data and the  $Y = \ln y$  data set will give values for  $m$  and  $c$  which can then be used to calculate  $m$  and  $b$ .

Again Excel provides a shortcut. Assuming the original data is as above the shortcut is given by:

**= LINEST(LN(B1:B10),A1:A10)**

### Logarithmic

The logarithmic fit needs no transformation. We either apply *LINEST* directly to the original  $y$  data and the log of the  $x$  data or use the shortcut method. In this case the shortcut method would be:

**= LINEST(B1:B10,LN(A1:A10))**

with the  $m$  and  $c$  values being given without further calculation.

### Trendlines and charts

As we saw in Sec(4.4.4) on Linear Regression the straight line trend could be added directly to a chart of the given data. Activating the chart by left clicking just inside the boundary of the chart and then selecting Chart - Add trendline from the top menu gives a pop up menu with several options. These include the option to add a polynomial fit up to degree six, a power trendline, an exponential trendline or a logarithmic trendline. Additionally there is the option of adding a moving average trend; not covered here.

Visually this facility is very useful, however it does not give us any quantitative information about the trendline.

### 4.5.3 General Fitting using Excel

In fitting a polynomial we touched on a more general problem that might be solved using *LINEST*. Given a set of data where  $y$  depends on  $k$  variables  $\{x_1, x_2, \dots, x_k\}$  we can use *LINEST* to fit a function of the form:

$$y = m_1 f_1(x_1 \dots x_k) + m_2 f_2(x_1 \dots x_k) + \dots + m_n f_n(x_1 \dots x_k) + c$$



Although the  $f$  functions may not be linear we are still fitting a linear function of the  $c$  and  $m$  parameters. An even more general problem can be solved using Excel, namely that of fitting a function to a set of data where the dependence on the parameters is not linear. For example fitting a function of the form  $y = axe^{bx} \cos(cx+d)$  to a given set of  $x$ - $y$  data can be carried out using Excel's Solver programme. The following two examples illustrate both of the above two methods. In each case we fit a given function to the data:

$x =$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$y =$	0	0.0354	0.0846	0.169	0.317	0.495	0.544	0.464	0.383	0.342	0.338

Table 4.2: Data for Ex 4.8 and Ex 4.9

**Example 4.8** Find the values of  $m_1, m_2, m_3, m_4$  and  $c$  such that

$$y = m_4 \sin 2x + m_3 \cos 2x + m_2 \sin x + m_1 \cos x + c \quad (4.12)$$

is a least squares fit to the data in Tab(4.2).

- Set up the following column headers in A1:N1 :

x	y		Squares		cos x	sin x	cos 2x	sin 2x	m4	m3	m2	m1	c
---	---	--	---------	--	-------	-------	--------	--------	----	----	----	----	---

- Enter the  $x$ -values in A3:A13 and the  $y$ -values in B3:B13
- Enter = cos(A3) in F3, copy down to F13. (*this creates a column of cos  $x$  values*)
- Enter = sin(A3) in G3, copy down to G13. (*this creates a column of sin  $x$  values*)
- Enter = cos(2\*A3) in H3, copy down to H13. (*this creates a column of cos  $2x$  values*)
- Enter = sin(2\*A3) in I3, copy down to I13. (*this creates a column of sin  $2x$  values*)
- Highlight J2 to N2 ready to accept the values of  $m_4, m_3, m_2, m_1$ , and  $c$ .
- Enter =LINEST(B3:B13,F3:I13) **Ctrl-Shift-Enter**

This gives the fit as:

$$F(x) = 14.765 \sin 2x + 7.555 \cos 2x - 30.830 \sin x - 51.168 \cos x + 43.637$$

This is plotted in Fig 4.14 where it is referred to as the "trig-fit". (*for the purpose of plotting the values of  $F(x)$  were evaluated for  $x = 0$  to 1 in steps of 0.025*)

**Example 4.9** In the previous example the fitting parameters  $m$  and  $c$  formed the coefficients of given functions of  $x$ . In short the trial function was a linear function of its parameters. This example fits a function whose parameters are not simply the coefficients of functions of  $x$ . In such a case it is not possible to use LINEST for the fitting process, however Excel provides a goal seeking program called Solver. Solver uses a numerical

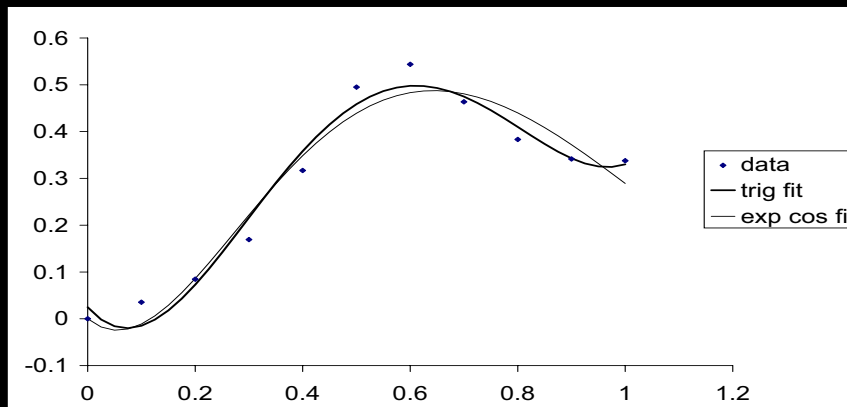


Figure 4.14: Data points; Trig-fit Eq(4.12); Exponential-Cos fit

method to minimise the contents of a given cell (*target cell*) subject to varying the contents of other cells (*the parameters*).

In this problem the values of the parameters  $a$ ,  $b$ ,  $c$  and  $d$  are determined such that  $y = F(x) = axe^{bx} \cos(cx + d)$  forms a least squares best fit to the given data. (Table 4.2.) Thus in detail Solver must find the values of  $a$ ,  $b$ ,  $c$  and  $d$  such that

$$S = \sum_{i=0}^N \left( y_i - ax_i e^{bx_i} \cos(cx_i + d) \right)^2 \quad (4.13)$$

is minimised. (*where here  $N = 10$ , the number of data points less one.*) Using the same Worksheet as in Ex(4.8) construct a user function for the trial function  $F(x)$ , four cells containing  $a$ ,  $b$ ,  $c, d$  and a cell containing  $S$  before invoking Solver.

- In a VBA module construct the user function  $F(a,b,c,d,x)$
- Enter headers  $a, b, c$  and  $d$  into cells J4:M4 and 1 as starter values for  $a$ ,  $b$ ,  $c$  and  $d$  into cells J5:M5
- In D3 enter  $(B3-F(\$J\$5, \$K\$5, \$L\$5, \$M\$5, A3))^2$ . Copy D3 down to D13. (*this gives the square of the differences between the  $y$  values for each data point and the trial function*)
- In D15 enter  $=\text{Sum}(D3:D13)$ , D15 now contains  $S$  and is the target cell for Solver.
- Select Tools, Solver. (*If Solver is not available then Select the Add-Ins and tick the Solver Add-In check box*).
- In the Solver dialogue box: Make the target cell D15; Select Equal to min; Enter J5:M5 in changing cells; Select Solve to run the program. The best values of the parameters will now be in J5:M5 and the minimum value of  $S$  in D15.

The above calculation gives

$$y = F(x) = 3.288xe^{-2.226x} \cos(2.480x - 1.861)$$

and is shown plotted in Fig 4.14 where it is referred to as "exp cos fit".

It is clear from the last two examples that the form of the trial function plays an important roll in this type of approximation. The trigonometric fit of Ex(4.8) appears to give a better fit and suggest that perhaps we increase the number of parameters by adding to the trial function  $m_5 \sin 3x + m_6 \cos 3x$ .

#### **A Final General Comment**

Expanding a given function in terms of a series of functions is already a familiar idea; we have seen in calculus that functions are expanded in terms of powers of  $x$  to give a Taylor Series. A variation of this idea is to expand a function in terms of a trigonometric series of sine and cosine terms similar to the above, producing what it known as a Fourier series. In Ex(4.8) above we are not given the function  $f(x)$ , only a sample of its values, however we are able to construct a trigonometric approximation to  $f(x)$  using the method of least squares.

## Chapter 5

# Minitab - Descriptive Statistics

### 5.1 Introduction

Minitab is a purpose built piece of software for displaying and analysing statistical data. Although many of the operations can be done in Excel this package offers direct access to such processes as hypothesis testing, analysis of variance, regression, data display using histograms and box plots and many other features. In this chapter we look at some of the basic ideas used in characterising data and how to graphically represent them. The hypothesis testing and the analysis of variance are not covered and although Minitab deals with linear regression this is discussed in Chapter 4 and will not be considered here.

### 5.2 Mean and Standard Deviation.

Two of the fundamental problems in elementary Statistics are:

- Given a set of data can we use a single value to typify the whole set. This is usually taken to be the average (*mean*) of the data though in some cases the middle value or median is used (*see below for definition*).
- Identify the amount of spread of the data about a central value in order to distinguish one set of data from another. One natural way to do this is to relate the central value to the difference between the highest and lowest values in the data set. Although this simple approach may at first seem attractive it does not indicate how well the data is clustered about the centre and tends to give a false impression if there are one or two very large values in the data set. This later shortcoming can be overcome in a graphical representation of the data using a box plot which introducing the idea of quartiles and outliers (*see below*).

The usual way to describe spread is to use the standard deviation which is a measurement based on the square of the distances of the data points from the mean.

Given a set of data  $\{x_1, x_2 \dots x_N\}$  the following definitions are made:

**Mean**

The mean of the data set, denoted  $\bar{x}$  is given by:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

**Standard Deviation**

The standard deviation of the data set, denote  $s$ , is given by:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$

Here the average of the sum of the squares of the distances of the data points from the mean are calculated and clearly the smaller this value the more clustered the points are about the mean. The square root is taken primarily to make  $s$  have the same units as the data.

The following result gives us an alternative expression for the standard deviation.

**Result**

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N x_i^2 - N\bar{x}^2$$

**Proof**

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N x_i^2 + \sum_{i=1}^N \bar{x}^2 - 2\bar{x} \sum_{i=1}^N x_i = \sum_{i=1}^N x_i^2 + N\bar{x}^2 - 2N\bar{x}^2 = \sum_{i=1}^N x_i^2 - N\bar{x}^2$$

**Alternative expression for  $s$** 

$$\text{Standard Deviation} = s = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \bar{x}^2}$$

**5.3 Sampling Problem**

The above definitions are directly applicable to given sets of data, however most of statistics is concerned with obtaining information about a large set of data (*the population*) by considering only a few of the items from the population (*sample*). For example we could estimate the average height of everyone in the UK over the age of twenty by calculating the average height of a group of, say, a thousand randomly chosen people over the age of twenty. Similarly we could estimate the standard deviation of the population by considering the standard deviation of our sample.

Consider the following general scenario where the population has  $M$  data elements  $X_i$ , with mean  $\mu$  and standard deviation  $\sigma$ , both of which we would like to estimate. A sample is taken consisting of  $N$  elements denoted  $x_i$  whose mean  $\bar{x}$  and standard deviation  $s$  we can calculate.

Population	Sample
$X_1, X_2 \dots X_M$	$x_1, x_2 \dots x_N$
mean = $\mu$ std = $\sigma$	mean = $\bar{x}$ std = $s$

The mean of the sample  $\bar{x}$  is usually used to estimate the mean of the population  $\mu$ , though for small values of  $N$  this may be a poor estimate.

Similarly the standard deviation  $s$  of the sample may be used to estimate the standard deviation  $\sigma$  of the population. However in this case it is found that if  $N$  is small this will be an under estimate of the standard deviation of the population. There is also a more technical problem using  $s$  concerning whether or not the estimate is 'biased'. (*not covered here.*)

Thus when estimating the standard deviation of a population from a sample the following adjusted formula is used:

$$\sigma \approx \hat{s} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$

Clearly replacing  $N$  with  $N - 1$  makes very little difference if  $N$  is large. By default **Minitab always calculates**  $\hat{s}$  for the standard deviation.

## 5.4 Coefficient of Variation

In the above definitions the size of the standard deviation will depend on the size, or indeed units, of the individual data items. Thus carrying out the height measurements in the sampling problem of Sec(5.3) and measuring in metres will give a different value for the standard deviation to that obtained if the measurements are taken in inches. To overcome this dependence on the units used and to still keep the idea of the amount of spread the following coefficient is introduced.

The **coefficient of variation**  $v$  is defined as:

$$v = \frac{\sigma}{\mu} \approx \frac{\hat{s}}{\bar{x}}$$

That is to say the amount of spread or deviation about the mean is measured relative to the size of the mean.

## 5.5 Standard Error of the Mean

If, in the above, the sampling process is continued then for each sample we can calculate a mean. Denote these means by  $\{\bar{x}_1, \bar{x}_2 \dots\}$ . The set of all such means then forms a

new data set which itself has a mean and standard deviation. The standard deviation of these means is called the standard error of the mean and can be shown to be given by:

$$\text{Standard Error of Means} = \frac{\sigma}{\sqrt{N}}$$

As expected if we only take one sample then we estimate  $\sigma$  with  $\hat{\sigma}$ .

By default **Minitab** calculates:

$$\text{Standard Error of Mean} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N(N-1)}}$$

Knowing this value can then give us information on how good an estimate of the mean is provided by a single sample. Under certain assumptions there will be about a 65% chance of the population mean  $\mu$  being within a standard error of the estimated value of the mean derived from a single sample.

## 5.6 Quartiles and Median

Another measure of central tendency is the median. If the data is arranged into ascending order then the data item halfway along the list (*for a list containing an odd number of items*), or the average of the middle two (*for a list containing an even number of items*), is called the **median** of the data. Clearly the median divides the set of data into two halves in the sense that there are an equal number of data items on either side of it. In a similar way the ordered data set may be split into four quarters in order to give an idea of the spread of the values about the median. Taken in order the points that do this are called, the first, second and third quartiles and denoted  $Q_1$ ,  $Q_2$  and  $Q_3$  respectively. The second quartile is always taken to be the median however the positions of the first and third quartiles are not so clearly defined.

For the data set  $\{x_1, x_2, \dots, x_N\}$  **Minitab** defines the **quartiles** as:

$$Q_1 = x_{\frac{N+1}{4}} \quad Q_2 = x_{\frac{N+1}{2}} \quad \text{and} \quad Q_3 = x_{\frac{3(N+1)}{4}}$$

and uses linear interpolation to calculate these values when the indices are not whole numbers. (*this process does in fact give  $Q_2$  as the median*)

### Example 5.1

Given the data set  $\{1, 3, 5, 9, 11, 11, 12, 15\}$  calculate  $Q_1$ ,  $Q_2$  and  $Q_3$ .

In this example  $N = 8$  thus:

$$Q_1 = x_{\frac{8+1}{4}} = x_{2.25} \quad Q_2 = \text{median} = x_{\frac{8+1}{2}} = x_{4.5} \quad \text{and} \quad Q_3 = x_{\frac{3(8+1)}{4}} = x_{6.75}$$

Thus

- $Q_1 = x_{2.25}$  is a quarter of the way between  $x_2 = 3$  and  $x_3 = 5$  that is to say  $\underline{Q_1 = 3.5}$
- $Q_2 = x_{4.5}$  is half way between  $x_4 = 9$  and  $x_5 = 11$  that is to say  $\underline{Q_2 = 10}$
- $Q_3 = x_{6.75}$  is three quarters of the way between  $x_6 = 11$  and  $x_7 = 12$  that is to say  $\underline{Q_3 = 11.75}$

**Note**

With  $Q_2$  defined as the median of the data set an alternative way of defining  $Q_1$  and  $Q_3$  is as follows. Make  $Q_1$  the median of all the data points less than  $Q_2$  and  $Q_3$  the median of all the data points greater than  $Q_2$ . In the above example this would give  $Q_1 = 4$  and  $Q_3 = 11.5$ . Other definitions do exist.

**Remark**

Minitab's definition follow logically from the case when each quarter is separated by an actual data item. In such a case where we have  $k$  items in each quarter we would require  $N = 4k + 3$ , with the +3 part of this being the data items given over to the values of the three quartiles. See fig 5.1.

That is to say:

$$Q_1 = x_{k+1} \quad Q_2 = x_{2k+2} \quad \text{and} \quad Q_3 = x_{3k+3}.$$

Substituting for  $k$  into these expressions from  $N = 4k + 3$  gives:

$$Q_1 = x_{\frac{N+1}{4}} \quad Q_2 = x_{\frac{N+1}{2}} \quad \text{and} \quad Q_3 = x_{\frac{3(N+1)}{4}}.$$

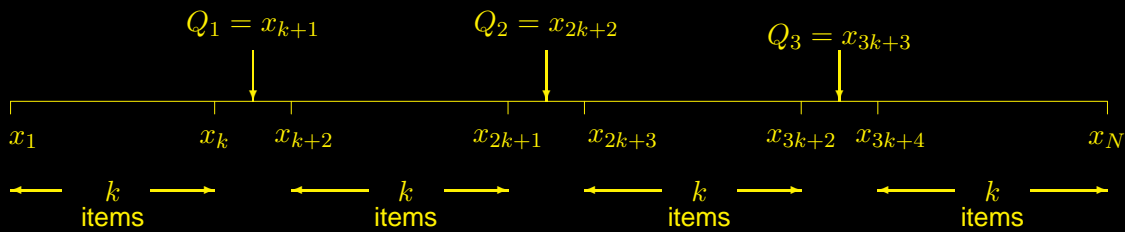


Figure 5.1:  $N = 4k + 3$  items split into four quarter of  $k$  items in each. Each quarter is separated by a data item assigned to one of the quartiles

## 5.7 Box Plots

To visually display and compare the central tendency and spread of one data set with another a diagram consisting of box plots is constructed. A box plot consist of a diagram



that illustrates the range of the data and the position of the three quartiles. Certain adjustments are made to the diagram when either the distance of the maximum or minimum value to its nearest quartile is large compared to the distance between the first and third quartile.

The general form of a box plot is shown in fig 5.2, the ends of the box are the two quartiles  $Q_1$  and  $Q_3$  and the line across the box is the median. The arms or whiskers extend to  $R$  and  $L$  and are given by:

$$R = \text{Min} \begin{cases} \text{Max data item} \\ Q_3 + 1.5(Q_3 - Q_1) \end{cases} \quad L = \text{Max} \begin{cases} \text{Min data item} \\ Q_1 - 1.5(Q_3 - Q_1) \end{cases}$$

Any data items outside these ranges are shown as single items and referred to as **outliers**.

The reason the whiskers do not in general run from the smallest data item to the largest is that the diagram would be a misrepresentation of a data set containing one very large positive or large negative data item. In such a case the values of  $R$  and or  $L$  are restricted to being within the range from one and a half the interquartile range below the first quartile to one and a half the interquartile range above the third quartile. All other values, if any, outside this range are plotted separately.

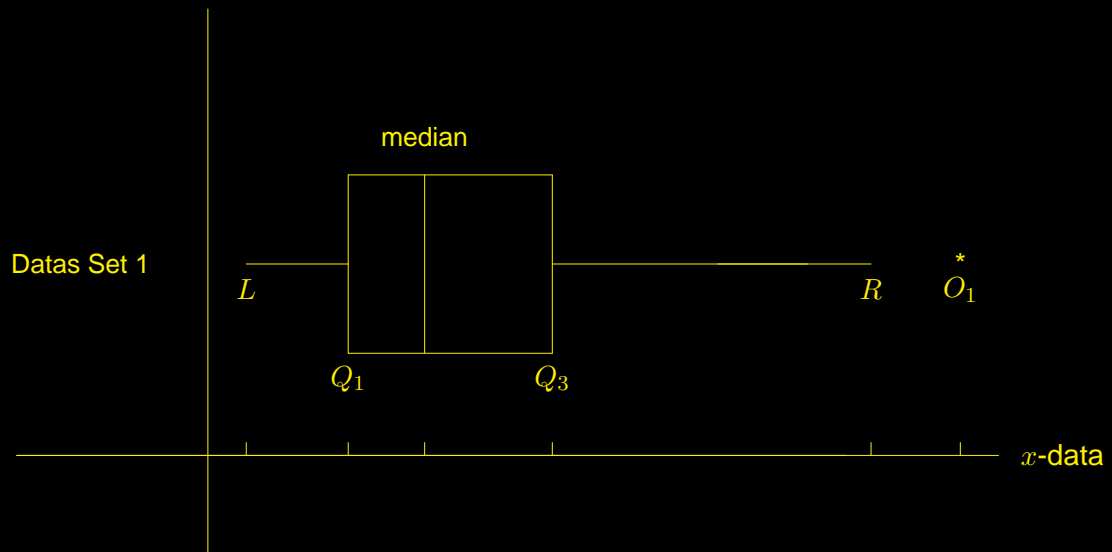


Figure 5.2: *Box Plot: The diagram shows one outlier denoted  $O_1$ , thus  $R$  is given by  $Q_3 + 1.5(Q_3 - Q_1)$ .  $L$  is at the smallest data value.*

## 5.8 Histogram

Given a set of data a Histogram is a type of bar chart that indicate the frequency of occurrence of the data items across its range. If the data are contained in the interval  $[a, b]$  and a partition of this interval is formed, that is to say the interval is divided into non-overlapping subintervals, then on each subinterval a rectangle is constructed such that the area of the rectangle is proportional to the number of data points in the subinterval. The graph so formed is called a histogram.

In the literature the requirement that the area of the rectangle is proportional to the frequency is sometimes dropped and it is the height of the rectangle that becomes proportional to the frequency. This distinction is only relevant in the case that the widths of the subintervals are not all equal.

Minitab provides us with three types of histogram; a **frequency histogram** where the height of the rectangle is proportional to the frequency; a **percentage histogram** where again the height of the rectangle is proportional to the frequency but is now expressed as a percentage; a **density histogram** where the area of the rectangle is proportional to the frequency and the sum of the areas of all the rectangles equals unity. It is really only the last type that strictly satisfies the definition of histogram where the area of each rectangle is proportional to frequency, however since we are usually taking the widths of the intervals to be the same as each other this distinction is not generally relevant. The following example summarises the different types of histograms that are available in Minitab.

### Example 5.2

Consider the following set of  $N = 20$  data points:

$x_i =$	1	2	2	3	6	9	10	20	21	22	23	27	30	35	44	45	47	50	52	60
---------	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Table 5.1:

Minitab's default gives the following split of the extended range when constructing a histogram.

interval <sup>1</sup> width $h_i = 10$	$[-5, 5)$	$[5, 15)$	$[15, 25)$	$[25, 35)$	$[35, 45)$	$[45, 55)$	$[55, 65]$
frequency $f_i =$	4	3	4	2	2	4	1
percentage $100 \times \frac{f_i}{N} =$	20%	15%	20%	10%	10%	20%	5%
density $\frac{f_i}{Nh_i}$	0.02	0.015	0.02	0.01	0.01	0.02	0.005

Since in each case the widths of the intervals are all equal the distinction between whether or not the height or the area of each rectangle is proportional to the frequency is not relevant. As we see in fig 5.3 all three histograms are identical in shape.

<sup>1</sup>notation:  $[-5, 5)$  for example indicates all  $x$  such that  $-5 \leq x < 5$  etc.

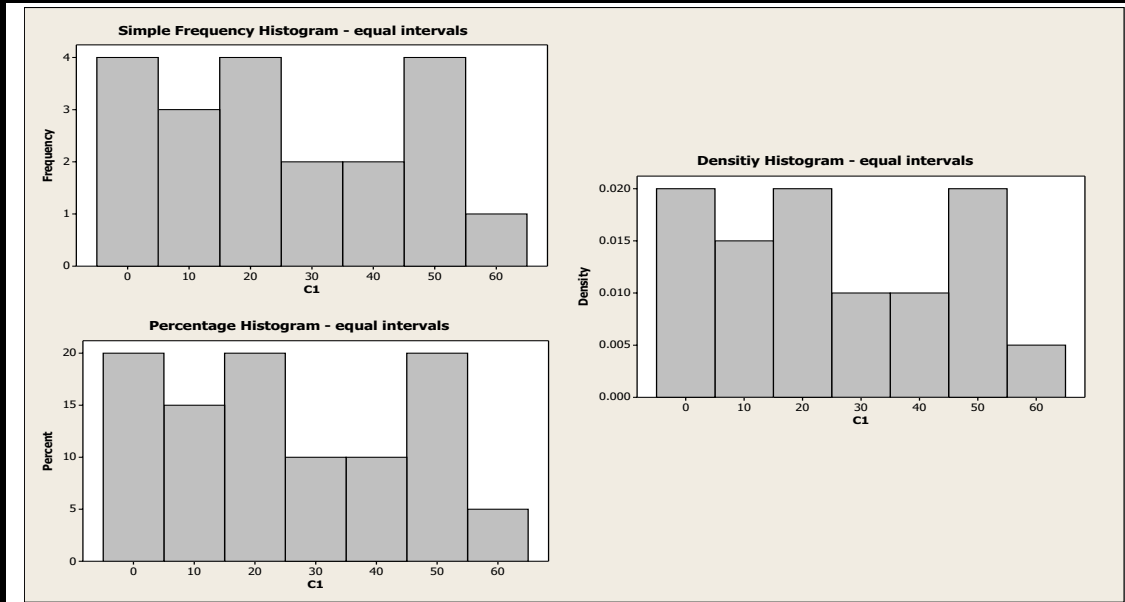


Figure 5.3:

If we now set the subinterval not to be equal to each other then we obtain the following results: In this case since the intervals are not of equal length it is now more obvious

intervals	[0, 5)	[5, 10)	[10, 20)	[20, 25)	[25, 35)	[35, 50)	[50, 60]
Widths $h_i =$	5	5	10	5	10	15	10
frequency $f_i =$	4	2	1	4	2	4	3
percentage $100 \times \frac{f_i}{N} =$	20%	20%	5%	20%	10%	20%	15%
density $\frac{f_i}{Nh_i}$	0.04	0.02	0.005	0.04	0.01	0.013	0.015

that for the frequency and percentage histograms it is the height that is proportional to the frequency whereas for the density histogram it is the area. This can be seen in detail (fig 5.4) by looking at the intervals [35, 50) and [20, 25). These intervals are of different widths but contain four data items each. On the frequency and percentage histograms the rectangles for both these intervals are the same height, thus indicating that the frequency is represented by height and not area. On the the density histogram the heights of the rectangle are adjusted in order that the total areas of the rectangles are equal, thus the rectangle on the narrower interval [20, 25) is higher than the one on the wider interval [35, 50).

## 5.9 Minitab

This section looks at the Minitab implementation of the above ideas, namely printing out descriptive statistics, plotting Box Plots and constructing Histograms and displaying them

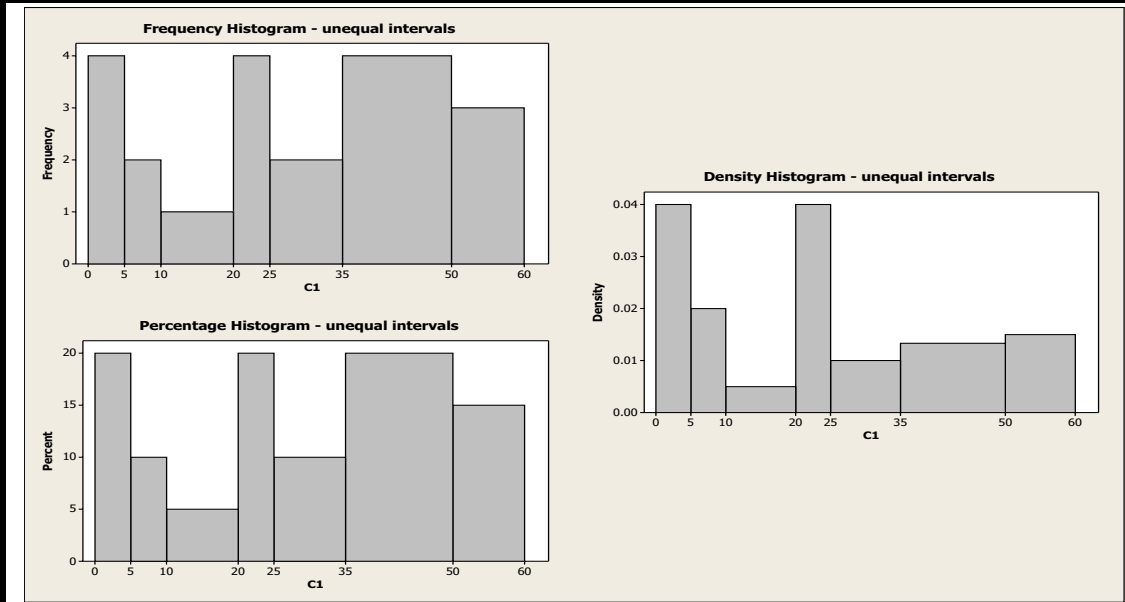


Figure 5.4:

in a Layout format as seen in fig 5.3 and fig 5.4.

### 5.9.1 Descriptive statistics

- Enter the data from table 5.1 into column 1 of the worksheet. The worksheet behaves in a similar but **restricted way** to the Excel worksheet.
- Select **Stat** from the top menu and **Basic Statistics** from the sub menu.
- Set **Variable set** equal to C1
- Select **Statistics** and tick: mean; SE of mean; standard deviation; variance; coefficient of variation; Min; Max; First quartile; Median; Third quartile; N. Select OK OK to close all sub windows.
- The basic statistics should now be visible in the **Session Window** and given by:

Data	N	Mean	SE Mean	StDev	Variance	CoefVar	Min	$Q_1$	Median	$Q_3$	Max
C1	20	25.45	4.28	19.14	366.47	75.22	1.00	6.75	22.50	44.75	60.00

### 5.9.2 Box Plots

- With the data from table 5.1 in column 1 select **Graph** from the top menu followed by **Box Plot**.
- Select **Simple OK**
- Enter C1 as **Graph variables**.

- By default Minitab creates vertical box plots, here we produce horizontal box plots by:

Selecting **Scale** and ticking **transpose**. OK. OK

This gives a basic box plot for the data in column 1. Pointing to the box will activate a note giving the data used to construct the plot. Double clicking the title of the plot makes it possible to format and change the title.

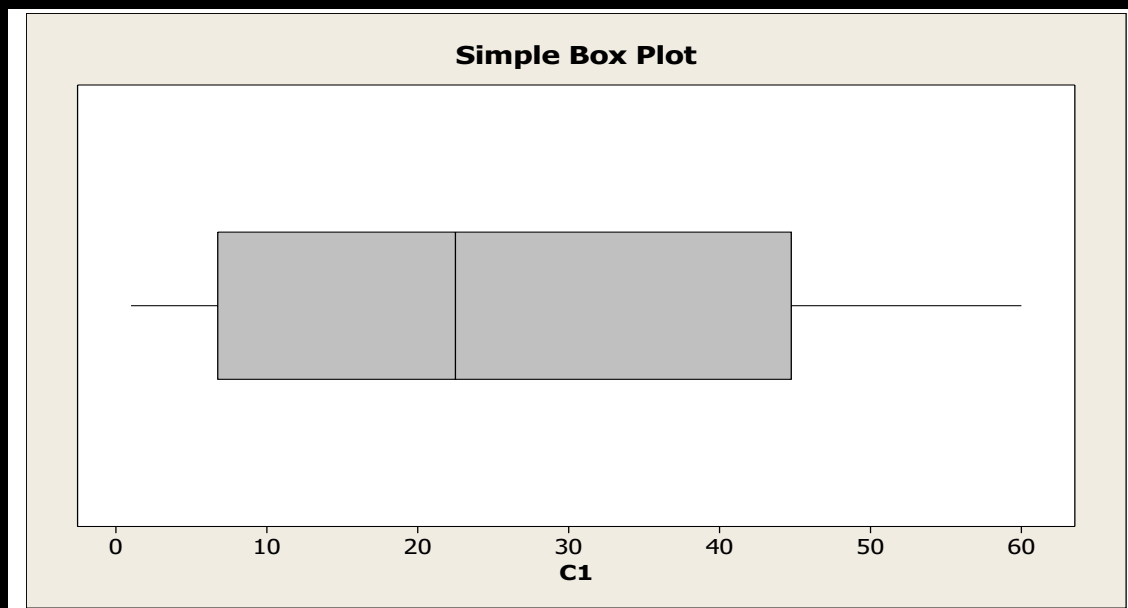


Figure 5.5: *Box Plot for the data in table 5.1*

### 5.9.3 Histogram

- With the data from table 5.1 in column 1 select **Graph** from the top menu followed by **Histogram**.
- Select Simple type. OK
- Set **Graph variables** to C1. Select **Scale**
- Select the **y-scale type** tag and select one of the three types, namely **Frequency**, **Percent** or **Density**. OK. OK

This gives a default Histogram of the selected type with equal subintervals over a slightly extended range.

To produce a layout of histograms as in Fig 5.3

- Create two more histograms of the other two types

- With one of the histograms active (*click inside one of the histograms windows if not active*) Select **Editor** from the top menu followed by **Layout tools**.
- The layout tool is fairly easy to follow. Position your your three histograms. Select **Finish**

You should now have a new window containing all three histograms.

To produce a histogram with unequal sub intervals you have to fist create a histogram using equal intervals as above and then edit the  $x$ -axis as follows:

- With an histogram active select **Editor** from the top menu followed by **Select Item** and **x-Scale** (*alternatively you can point to the x-axis and left click*)
- With the  $x$ -axis now active select **Editor** followed by **Edit x-scale** (*alternatively you can right click the active x-axis*)
- Select the **Binning** tag. Check the **Cut point** and **Midpoint/cut point positions** boxes.
- In the Midpoint/cut point position box enter the cut points: 0 5 10 20 25 35 50 60 (*leave spaces between the values*) OK

The old histogram will be replaced with a new one with intervals as specified.

## Chapter 6

# Powers of Matrices - Markov Chains

### 6.1 Introduction

This chapter looks at a couple of applications for the power of a matrix. In particular we look at the case where it is the probability of events that vary with each application of the matrix. For example we may know that if it is sunny today there is a given probability of it being sunny tomorrow whereas if it's not sunny today there is a given probability of it not being sunny tomorrow. The question that we might then pose is, if it is sunny today and we apply our probabilistic model, what will the probability of it being sunny be in 10 days time.

An identical problem is considered below where instead of talking about sunny and non-sunny days we talk about the probability of a stock rising or falling.

A similar problem can be constructed in the area of manpower planning. For example the army may each year decide to move troops between three overseas postings, A, B and C in order to broaden their experience. Each year it decides to transfer: 10% of A to B and 15% to C; 5% of B to A and 10% to C; 10% of C equally divided between A and B. The question is what will happen if this policy is adopted for several years.

Each of these applications require us to find the  $n^{\text{th}}$  power of a special type of matrix. In order to develop the model first consider the idea of conditional probability.

### 6.2 Conditional Probability

The following simple example concerning conditional probability gives us the necessary ideas and notation required to develop a simple stochastic model.

**Example 6.1** Consider the box in fig 6.1 containing three red balls and two white. Select one ball at random. Without replacing this ball select a second ball at random from the box. The problem asks you to calculate the probability that the second ball is red.

This is a fairly elementary problem in probability that involves the idea of conditional

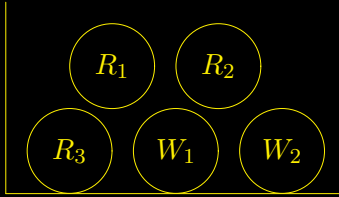


Figure 6.1: Three red balls and two white balls in a box.

probability. In the figure the balls have been labelled  $R_1, R_2$  etc, thus we can choose the first ball in five ways and the second in four way giving  $5 \times 4 = 20$  possible pairs of the form  $R_1W_1, R_1W_2$ , etc. To calculate the number of pairs where the second choice is a red ball we could list all 20 possible pairs and then count how many ended in a red choice.

A more general approach would be to say that we have a success if we choose either a red followed by a red (RR) or a white followed by a red (WR).

- For WR we can choose the first ball in 2 ways from 5 and the second in 3 ways from 4, giving a total of  $2 \times 3 = 6$  ways from the 20 possible selections.
- Thus the probability of selecting WR,

$$P(WR) = \frac{2 \times 3}{20} = \frac{2}{5} \times \frac{3}{4} = P(W)P(R|W) = \frac{6}{20}$$

Where  $P(W)$  is the probability of selecting a white ball on the first draw and  $P(R|W)$  is the probability of selecting a red ball on the second draw given that a white ball was chosen on the first draw.  $P(R|W)$  is referred to as the **conditional probability**<sup>1</sup> of selecting a red ball given that a white has already been chosen.

- For RR we can choose the first ball in 3 ways from 5 and the second in 2 ways from 4, giving a total of  $3 \times 2 = 6$  ways from 20. As before we then have

$$P(RR) = \frac{3}{5} \times \frac{2}{4} = P(R)P(R|R) = \frac{6}{20}$$

- Since to succeed only one or other of these can happen the total chance is obtained by adding the two, thus:

$$P(\text{Red Second}) = P(W)P(R|W) + P(R)P(R|R) = \frac{6}{20} + \frac{6}{20} = \frac{12}{20} = \frac{3}{5}$$

The above example can be represented by the following transition diagram which will then be developed into a Markov Chain<sup>2</sup> in the next section.

<sup>1</sup>In general  $P(A|B)$  denotes the probability of event  $A$  occurring given that event  $B$  has already occurred

<sup>2</sup>Markov 1856-1922 Russian Mathematician



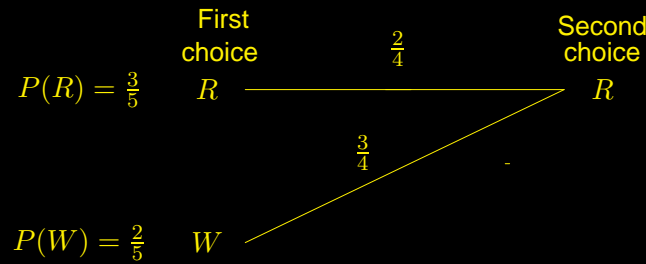


Figure 6.2:

### 6.3 Markov Chain

The following example illustrates the idea of a Markov Chain and the associated Markov Matrix. It is an example of a discrete stochastic system.

**Example 6.2** It is noted that if a given stock has risen at the end of the days trading there is a 1 in 3 chance that it will rise at the end of the following day's trading. Similarly if the stock has not risen at the end of trading there is a 1 in 2 chance that it will not rise at the end of the following day. The question is: what is the probability of a rise on day  $n$  given the chance of a rise on day 0. The following diagram drawn along the lines of fig 6.2 illustrates the essential information of this problem: Denoting the probability

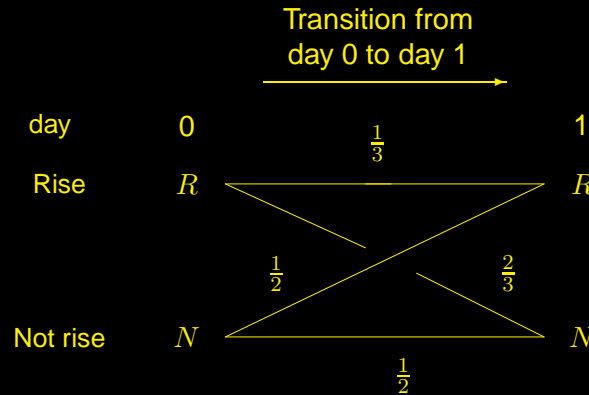


Figure 6.3:

of a rise on day  $k$  by  $P_k(R)$  and not a rise by  $P_k(N)$  we deduce from the diagram that:

$$P_1(R) = \frac{1}{3}P_0(R) + \frac{1}{2}P_0(N) \quad \text{and} \quad P_1(N) = \frac{2}{3}P_0(R) + \frac{1}{2}P_0(N)$$

Writing this in matrix form gives:

$$\begin{pmatrix} P_1(R) \\ P_1(N) \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} P_0(R) \\ P_0(N) \end{pmatrix} \quad \text{which is denoted as} \quad \underline{P}_1 = M\underline{P}_0$$

Thus moving from day to day we have:

$$\underline{P}_1 = M\underline{P}_0 \rightarrow \underline{P}_2 = M\underline{P}_1 = M^2\underline{P}_0 \dots \text{continuing} \dots \underline{P}_n = M^n\underline{P}_0$$

Calculating  $M^n$  is clearly a difficult problem if  $M$  has many rows and columns. The most realistic method of evaluating  $\underline{P}_n$  is to use an iterative process based on:

$$\underline{P}_{k+1} = M\underline{P}_k \text{ given } \underline{P}_0$$

If we are given that the stock did in fact rise on day 0 that is to say  $\underline{P}_0$  is known and given by  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  we are able to calculate  $\underline{P}_n$  for any  $n$ . Using Excel with a similar worksheet used for solving the matrix iteration carried out when using Newton's methods we have:

	A	B	C	
1	PR =	1		<i>replace with =B7 to iterate</i>
2	PN =	0		<i>replace with =B8 to iterate</i>
3				
4	M=	= 1/3	= 1/2	
5		= 2/3	= 1/2	
6				
7		=MMULT(B4:C5,B1:B2)		<i>highlight B7:B8</i>
8				<i>use Ctrl-Shift-Enter</i>

For this to carry out the iteration correctly set the options and proceed as follows:

- Under **Tools, Options** select the **Calculation** tab and set:
  - iteration, 1 step**, and also of importance set **manual calculation**
- Enter the worksheet with B1=1 and B2=0.
- Change the contents of B1 and B2 as indicated
- To start the iterative process press **F9**.

On iterating it is found that the contents of B1 and B2 converge to 0.43 and 0.57 respectively. Other valid starting values will also converge to these two values. It is no coincidence that we attain convergence to a single set of values, this is a consequence of  $M$  being a positive Markcov matrix and the column sum of the state vector being unity. In terms of the problem these values will represent the probabilities of the stock rising or not rising if we arbitrarily observe the stock at some point in its lifetime.

This section concludes with a couple of useful definitions and one of the fundamental properties of the Markov matrix.

**Definition**

An  $n \times n$  non-negative matrix  $M$  such that its columns each sum to 1 is called a Markov Matrix. That is to say if  $M$  has  $i$ - $j$ th term  $M_{ij}$  then:

$$M_{ij} \geq 0 \quad \text{and} \quad \sum_{i=1}^n M_{ij} = 1 \quad j = 1 \dots n$$

**Definition**

The vector  $\underline{P}_k$  used above to describe the system at any particular point is called the **state** of the system. The sequence of states generated from applying the Markov Matrix  $M$  is called a **Markov chain**.

The following diagram represents a Markov chain:

$$\underline{P}_0 \xrightarrow{M} \underline{P}_1 \xrightarrow{M} \underline{P}_2 \xrightarrow{M} \dots$$

**Result**

The column sum of a state vector is preserved under the application of a Markov matrix. The proof of this statement is quite straightforward but not necessary for this course, instead consider the application of  $M$  from Ex(6.2) above.

$$\text{Let } \underline{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \text{ such that } p_1 + p_2 = 1$$

Then

$$M\underline{p} = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3}p_1 + \frac{1}{2}p_2 \\ \frac{2}{3}p_1 + \frac{1}{2}p_2 \end{pmatrix}$$

Thus the column sum of  $M\underline{p}$  is given as:

$$\frac{1}{3}p_1 + \frac{1}{2}p_2 + \frac{2}{3}p_1 + \frac{1}{2}p_2 = p_1 + p_2 = 1$$

Thus the column sum of  $\underline{p}$  and  $M\underline{p}$  are equal, which in this problem is unity since the state vector  $\underline{p}$  represents a probability distribution.

**Example 6.3** Consider now the example in the introduction concerning the movement of troops.

*Each year it is decided to transfer: 10% of A to B and 15% to C; 5% of B to A and 10% to C; 10% of C equally divided between A and B. The question is what will happen if this policy is adopted for several years.*

A transition network for this problem is in Figure 6.4.

As can be seen:

$$\begin{aligned} \text{Numbers at A after transfer} &= 0.75A + 0.05B + 0.05C \\ \text{Numbers at B after transfer} &= 0.1A + 0.85B + 0.05C \\ \text{Numbers at C after transfer} &= 0.15A + 0.1B + 0.9C \end{aligned}$$

Thus in matrix form:

$$\begin{pmatrix} 0.75 & 0.05 & 0.05 \\ 0.1 & 0.85 & 0.05 \\ 0.15 & 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} A_0 \\ B_0 \\ C_0 \end{pmatrix} = \begin{pmatrix} A_1 \\ B_1 \\ C_1 \end{pmatrix} \quad \text{rewrite as } M\underline{P}_0 = \underline{P}_1$$

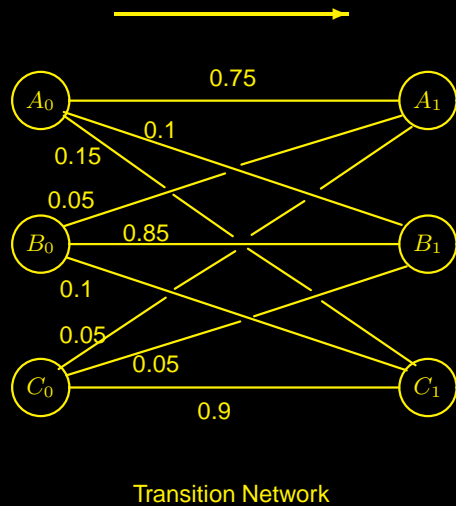


Figure 6.4: The troop levels at the three bases  $A$ ,  $B$  and  $C$  are denoted by  $A_i$ ,  $B_i$  and  $C_i$ , where  $i$  stands for the number of years from the start of the transfer scheme.

The matrix  $M$  is clearly a Markov matrix since its entries are all non-negative and its columns sum to unity. The conservation of the column sum of  $\underline{P}_0$  under the application of  $M$  implies that in this model the number of troops remains constant.

In general to calculate the distribution of troupes after  $n$  year it is necessary to calculate  $M^n \underline{P}_0$ , where  $\underline{P}_0$  is the initial distribution of troops amongst the three bases. This calculation is carried out using Excel. In doing this it turns out that no matter what the initial distribution of troops is, after enough years the distribution will tend to a steady state.

That is to say  $M^n \underline{P}_0 \rightarrow$  some vector  $\underline{P}$  as  $n \rightarrow \infty$

It can be shown that the vector  $\underline{P}$  is unique up to a non-negative multiple. This multiple is dependent on the column sum of  $\underline{P}_0$ . In this example:

$$\text{with } \underline{P}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{Excel gives } \underline{P} = \begin{pmatrix} 0.5 \\ .875 \\ 1.625 \end{pmatrix}$$

If  $\underline{P}_0$  is changed from a vector with column sum 3 to one with column sum 6 then all the components of  $\underline{P}$  would be doubled.

Indeed:

$$\text{with, say, } \underline{P}_0 = \begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix} \quad \text{Excel gives } \underline{P} = \begin{pmatrix} 1 \\ 1.75 \\ 3.25 \end{pmatrix} = 2 \times \begin{pmatrix} 0.5 \\ .875 \\ 1.625 \end{pmatrix}$$

## Chapter 7

# Google & PageRank

### 7.1 Introduction

Google is an extraordinary search engine of high speed and perhaps more importantly high quality. Almost immediately after entering your search into Google it delivers many possible web sites or pages that might be of interest. The addresses of the pages plus other information are ordered according to how *good* they are. It is this measure of *goodness*, apart from the sheer size and speed of Google that has made it so popular.

It is not difficult to imagine how Google collects its data; it employs a piece of software known as a *crawler* which continually visits web sites collecting relevant information, using the links on one web site to connect to another and so on. To some extent exactly what data Google collects is a secret, however it clearly collects *key words* and other web site addresses. It is also known that it downloads pages and compresses them into a huge archive. From the data it creates an index which is consulted every time you input a query. In addition to this Google orders its index according to something called *PageRank*. It is the ability of Google to look at PageRank and its index simultaneously that contributes to its quality and speed.

### 7.2 PageRank

A page's rank is based on the number of pages that point to it (link to it) and the PageRank of these pages.
--

The above statement may seem recursive in the sense that a page's rank depends on the PageRank of the pages that point to it. Thus one may ask how do we define the PageRank of the pointing pages. However mathematically this is nothing more than solving a problem that has the unknown variable on both sides of the equation. That is to say the unknown variable in the equation is given implicitly by the equation.

The idea of ranking is applicable to many different problems, for example consider the following:

In the academic world a researcher may be ranked according to how many people cite his/her work in their academic papers. However a simple count is not really such a good measure as we should also take into account the ranking of the people that are doing the citing. Thus if some eminent professor refers to my work it is definitely worth more in terms of ranking than if my mum recommends me.

In terms of the internet many pages point to Google, thus if Google is the only page linking to you this clearly makes you more important than if just your mum's web site points to you.

These ideas can be captured by the "random surfer". The random surfer is condemned for ever to move about the internet by:

- (a) Clicking arbitrarily from one page to another using existing page links.
- (b) If no link exists from a page he just arbitrarily moves to another page.
- (c) If at any time he gets bored using page links he again just arbitrarily moves to another page.

In (a) it is assumed that no preference is given to any one particular link on the current page, the choice of link is completely arbitrary. In (b) and (c) the initial model assumes that he can arbitrarily select a web site from a complete list of sites, this certainly simplifies the model though of course it is not practically possible for a real surfer to do this. However it is in (b) and (c) that we can if we so wish introduce bias into the model, for example he may just arbitrarily select UK web sites. This would have the effect of increasing the PageRank of all UK sites.

As the random surfer clicks on into infinity each site develops a probability of being visited. Thus we define the Page Rank of a site by:

The **PageRank** of a site is the probability that it will be visited by the random surfer.

### 7.3 Random Surfer & PageRank

It has already been stated that the PageRank of a page is the probability that it will be visited by a random surfer. To illustrate how this is achieved the following example is given. It only consists of four sites, each of which links to at least one other site. That is to say the problem of moving to a site without a link is not covered, however random hopping to another site due to the surfer getting bored is covered. Let the four sites be linked as follows:

- site 1 is linked to sites 2 and 3
- site 2 is linked to sites 1 and 4
- site 3 is linked to site 2 only
- site 4 is linked to all the other sites

The following diagram illustrates the above linked structure:

Figure 7.1: Sites on the left are linked, via the line segments, to sites on the right.

To construct the model consider initially the surfer at node 1.

- He either moves to site 2 or 3 by clicking
- If he gets bored he selects a site arbitrarily from the complete list (*including selecting his current site*)

If we assume that he decides to click to another site with a probability of  $\alpha$  then there is a probability of  $(1 - \alpha)$  that he selects a site because of boredom. Thus for site 1 we construct the diagram in fig 7.2: We have three similar diagrams for the other three sites shown in fig 7.3

With reference to the fig 7.2 and fig 7.3 the probability of the surfer selecting site 1 is calculated as follows:

- If the surfer is at site 1 then he only selects site 1 if he selects it because he is bored, thus:  $P(1|1) = (1 - \alpha) \left(\frac{1}{4}\right)$
- If the surfer is at site 2 then he can select site 1 by either selecting because he is bored or by clicking on a link, thus:  $P(1|2) = (1 - \alpha) \left(\frac{1}{4}\right) + \alpha \left(\frac{1}{2}\right)$

Figure 7.2: Possible routes out of site 1

Figure 7.3: Routes out of sites 2, 3 and 4

- If the surfer is at site 3 then he can only select site 1 by virtue of being bored, thus:  $P(1|3) = (1 - \alpha) \left(\frac{1}{4}\right)$ .
- If the surfer is at site 4 then he can select site 1 by virtue of being bored or by clicking on the link, thus  $P(1|4) = (1 - \alpha) \left(\frac{1}{4}\right) + \alpha \left(\frac{1}{3}\right)$

Denoting the probability that the surfer is currently at site  $n$  by  $P_0(n)$  the probability of selecting site 1 is given by:

$$P_1(1) = P_0(1) \frac{(1 - \alpha)}{4} + P_0(2) \frac{(1 - \alpha)}{4} + P_0(3) \frac{(1 - \alpha)}{4} + P_0(4) \frac{(1 - \alpha)}{4} + \alpha P_0(2) \frac{1}{2} + \alpha P_0(4) \frac{1}{3}$$

Using the fact that the sum of the probabilities at a given stage equals 1, that is  $P_0(1) + P_0(2) + P_0(3) + P_0(4) = 1$ , gives;

$$P_1(1) = \frac{(1 - \alpha)}{4} + \alpha \left\{ \frac{1}{2} P_0(2) + \frac{1}{3} P_0(4) \right\} \quad (7.1)$$

Repeating the above calculations for the other three sites gives:

$$P_1(2) = \frac{(1 - \alpha)}{4} + \alpha \left\{ \frac{1}{2} P_0(1) + P_0(3) + \frac{1}{3} P_0(4) \right\} \quad (7.2)$$



$$P_1(3) = \frac{(1-\alpha)}{4} + \alpha \left\{ \frac{1}{2}P_0(1) + \frac{1}{3}P_0(4) \right\} \quad (7.3)$$

$$P_1(4) = \frac{(1-\alpha)}{4} + \alpha \left\{ \frac{1}{2}P_0(2) \right\} \quad (7.4)$$

Writing Eqs 7.1 - 7.4 in matrix and column vector form gives:

$$\begin{pmatrix} P_1(1) \\ P_1(2) \\ P_1(3) \\ P_1(4) \end{pmatrix} = \frac{1-\alpha}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix} \quad (7.5)$$

At this stage it is not obvious that this actually constitutes a Markov process, however with a little rewriting this fact becomes evident. The fact that  $P_0(1) + P_0(2) + P_0(3) + P_0(4) = 1$  can be expressed in matrix form as:

$$1 = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix}$$

Working this into Equ(7.5) gives:

$$\begin{pmatrix} P_1(1) \\ P_1(2) \\ P_1(3) \\ P_1(4) \end{pmatrix} = \frac{1-\alpha}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix} + \alpha \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix}$$

Which gives:

$$\begin{pmatrix} P_1(1) \\ P_1(2) \\ P_1(3) \\ P_1(4) \end{pmatrix} = \left\{ \frac{1-\alpha}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \alpha \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \right\} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix}$$

This is of the form  $\underline{P}_1 = M\underline{P}_0$ . The matrix  $M$  clearly has positive values since  $0 < \alpha < 1$  and a simple check shows that each of its columns sums to unity. Hence  $M$  is a Markov matrix and the random surfing defines a Markov chain. If it is given that the surfer starts

at site 1 then  $\underline{P}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$  and the Markov chain corresponding to the activity of the

surfer, namely

$$\underline{P}_0 \xrightarrow{M} \underline{P}_1 \xrightarrow{M} \underline{P}_2 \xrightarrow{M} \dots$$

will tend to some steady state vector. This vector represents the probabilities of a site being visited by the surfer and hence its components are the PageRanks of each site. Using Excel with  $\alpha = 0.85$  the steady state vector to which the process converges is

$$\text{given by } P_{\infty} = \begin{pmatrix} 0.247 \\ 0.364 \\ 0.197 \\ 0.192 \end{pmatrix}$$

Thus the PageRanks are: site 1 = 0.247, site 2 = 0.364, site 3 = 0.197 and site 4 = 0.192

As expected site 2 has the largest page rank since it has the most number of links pointing to it. Sites 1 and 3 both have two links pointing to them however since the most important site, ie site 2, points to site 1 we see that site 1 has a greater PageRank than site 3. These facts reassure us that the basic idea of ranking the pages in this manner is effective.

## 7.4 Google

This section looks at a few of the facilities offered by Google. From Explorer or any other browser that you may be using enter the site [www.google.co.uk](http://www.google.co.uk) this will give the google home search page for the UK. Using this rather than [www.google.com](http://www.google.com) gives us a button to just search UK sites if we wish.

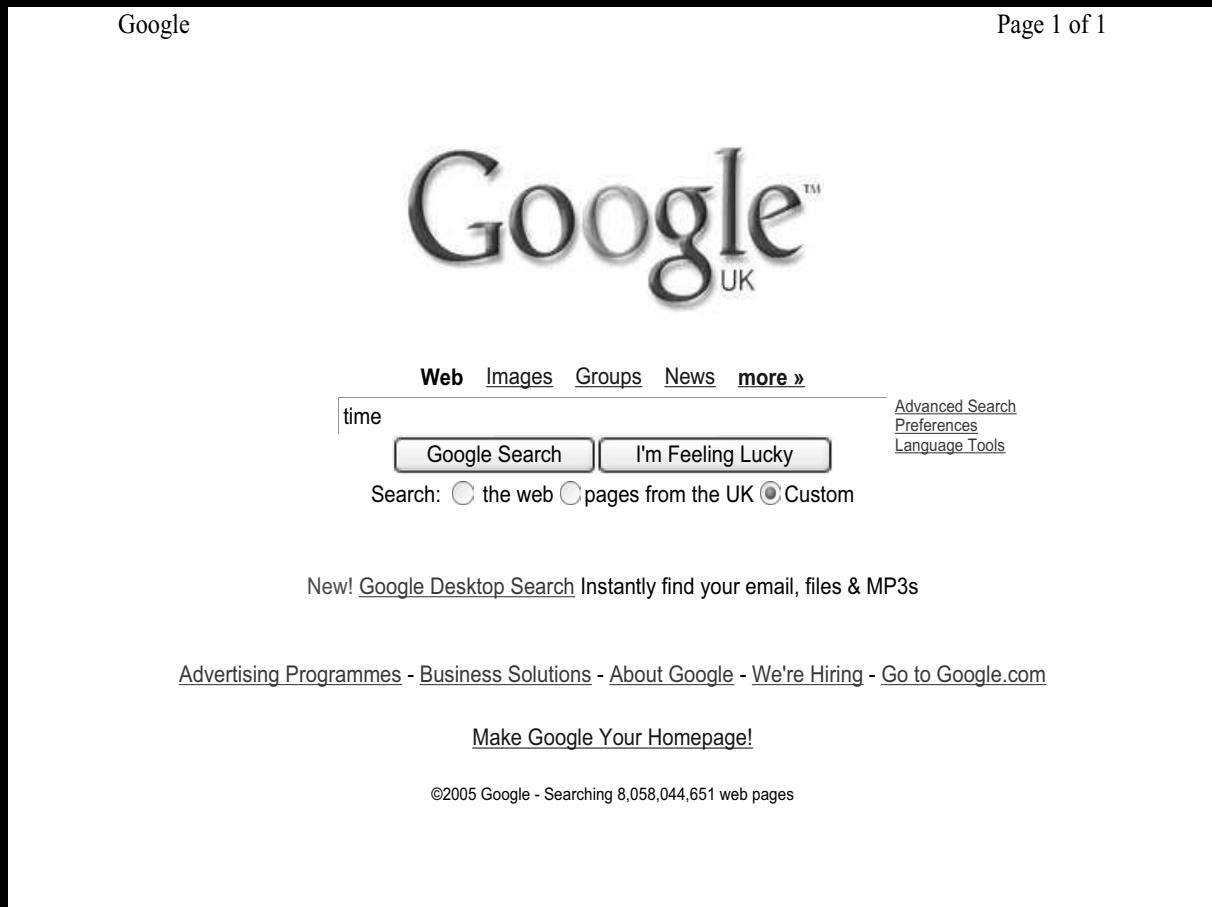


Figure 7.4: Home search page

### Simple search

With reference to fig 7.4:

- Enter the word **time** *(Google is not case sensitive)*
- Check **Pages from the UK**.
- Select **Google search**.

This gives a selection of possible web sites with snippets of information. Notice that the word **time** appears emboldened in the results. For further searching there


is no need to return to the first page as Google provides a search window at the top of the current page.

*If you had clicked **I'm Feeling Lucky** then instead of a list of sites you would have been taken to the site at the top of the list.*

Alternatively you can select to search the whole of the Web or you can make a custom search that you have created. (see later)

Preferences Page 1 of 1

---


**Preferences**
[Preferences Help](#) | [About Google](#)

Save your preferences when finished and **return to search**.

**Global Preferences** (changes apply to all Google services)

**Interface Language** Display Google tips and messages in:  
 English  
 If you do not find your native language in the pulldown above, you can help Google create it through our [Google in Your Language program](#).

**Search Language**  Search for pages written in any language ([Recommended](#)).  
 Search only for pages written in these language(s):

<input type="checkbox"/> Arabic	<input checked="" type="checkbox"/> English	<input type="checkbox"/> Indonesian	<input type="checkbox"/>
<input type="checkbox"/> Bulgarian	<input type="checkbox"/> Estonian	<input type="checkbox"/> Italian	<input type="checkbox"/>
<input type="checkbox"/> Catalan	<input type="checkbox"/> Finnish	<input type="checkbox"/> Japanese	<input type="checkbox"/>
<input type="checkbox"/> Chinese (Simplified)	<input checked="" type="checkbox"/> French	<input type="checkbox"/> Korean	<input type="checkbox"/>
<input type="checkbox"/> Chinese (Traditional)	<input type="checkbox"/> German	<input type="checkbox"/> Latvian	<input type="checkbox"/>
<input type="checkbox"/> Croatian	<input type="checkbox"/> Greek	<input type="checkbox"/> Lithuanian	<input type="checkbox"/>
<input type="checkbox"/> Czech	<input type="checkbox"/> Hebrew	<input type="checkbox"/> Norwegian	<input type="checkbox"/>
<input type="checkbox"/> Danish	<input type="checkbox"/> Hungarian	<input type="checkbox"/> Polish	<input type="checkbox"/>
<input type="checkbox"/> Dutch	<input type="checkbox"/> Icelandic	<input type="checkbox"/> Portuguese	<input type="checkbox"/>

**SafeSearch Filtering** Google's SafeSearch blocks web pages containing explicit sexual content from appearing in search results.  
 Use strict filtering (Filter both explicit text and explicit images)  
 Use moderate filtering (Filter explicit images only - default behavior)  
 Do not filter my search results.

**Number of Results** Google's default (10 results) provides the fastest results.  
 Display  results per page.

**Results Window**  Open search results in a new browser window.

Save your preferences when finished and **return to search**.

(Note: Setting preferences will not work if you have disabled cookies in your browser.)

©2005 Google

Figure 7.5: Preferences Page

## Preferences

With reference to fig 7.4 we note there is a preference link. Click on this to set one or two simple preference to determine how google operates. Clicking the preference link gives fig 7.5

- The **interface Language** is the language that Google operates in. If you change it to French then all the buttons and instructions will be in French. Google will still search all pages of all languages. Change this to suit.
- If you wish to only search pages in a specific language then under **Search Language** check the languages of the pages you wish to search. This will cause the **Custom** button to be visible on the search page of fig 7.4, and indeed in other places. Checking English and French will search all pages written in these languages and no others.
- The safe search filter is recommended as you can get a lot of unwanted rubbish.
- Set the **Number of Results** per page, 10 is good as if you have lots then you get very little information per result.
- Checking the **Results Window** option will mean that every time you click on a page it will open a new window, this can be useful for looking back.
- Don't forget to **Save Preferences**

### Advanced Search

It is possible to carry out quite advanced searches by either typing in words and symbols into the search window or alternatively use the advanced search page. As you get used to using Google you may well find more often than not that you are typing in quite complicated expressions into the search window. The advanced search link is always to be found next to the search window wherever it appears. Clicking on this option gives fig 7.6

- Find the results with **all** the following words. You enter here as many words as you wish, however lots of words are not necessarily a good thing. Try and keep your searches general but short. For example entering: *a really nice place to spend the night in london next wednesday* will most likely not give a good result whereas *london hotels* would probably give what you wanted.

Logically Google **AND's** the words and returns pages that contain all the words you have typed, less possible small words like "it", "a", "the" etc. It does not look for the complete sentence as you have typed it.

- If we wish to include in our search a complete phase, like *chocolate bar* then we can either enter it in the **exact phrase** box or we can type it in the search line including it in quotes. That is to say we would type in "**chocolate bar**" in the search line.

Consider the search for: *a healthy chocolate bar* , using the advanced search page we would enter *a healthy* in the first box and *chocolate bar* in the second.

Google Advanced Search Page 1 of 1

**Google** **Advanced Search** [Advanced Search Tips](#) | [About Google](#)

**Find results** with **all** of the words 10 results  
 with the **exact phrase**  
 with **at least one** of the words  
**without** the words

**Language** Return pages written in

**File Format** Only  return results of the file format

**Date** Return web pages updated in the

**Occurrences** Return results where my terms occur

**Domain** Only  return results from the site or domain  [More info](#)

**SafeSearch**  No filtering  Filter using [SafeSearch](#)

**Froogle Product Search (BETA)**

**Products** Find products for sale    
 To browse for products, start at the [Froogle home page](#)

**Page-Specific Search**

**Similar** Find pages similar to the page

**Links** Find pages that link to the page

Figure 7.6: *Advanced Search*

Alternatively if we are using the search window we just enter; **a healthy “chocolate bar”**

- Up to this point all the words will appear on the retrieved pages however we can refine the search even further:

If we wish to search for healthy chocolate bars that come from Belgium or Britain or both but are not organic then we would now enter *belgium britain* in the **at least one** window and *organic* in the **without** the word window.

Alternatively this could be entered in the search window as:  
**a healthy “chocolate bar” (belgium or britain) -organic**

*Note the use of the minus sign - to create not organic*

Finally we note that Google ignores the word 'a' at the start of the sentence, if we wish for some reason to force google to include a word we prefix it with a + sign. Thus on the search line we would enter **+a** rather than just a. Note that Google remarks on the fact that it didn't include 'a' when not preceded with a +.

- Just for the current search you can select to search just pages in a specific **language**.
- You may wish to restrict your search to *pdf* files or *jpeg* file or indeed exclude such files from your search. To do this use the **File Format** option.

Alternatively you can also do this directly in the search window as **filetype:pdf** to restrict to just *pdf* or **-filetype:pdf** to exclude *pdf*

- **Date** and **Occurrences** can be useful though not that often
- **Domain**: Restricting or excluding sites can be useful. If you select **only** and enter **.ac.uk** in the domain box then you will only search uk university sites. If you enter **city.ac.uk** you will only search the City University site. If you select **don't** then the sites are avoided.

Alternatively you can enter **site:.ac.uk** in the search window to just search UK university sites or **-site:.ac.uk** to avoid them.

### Search within results

Having retrieved many pages of possible sites Google will allow you to carry out a further search just on these pages. At the foot of the screen containing your retrieved sites there is a window containing your search criteria plus a link option to **search within results**. Selecting this gives a second window for entering further criteria. This in fact does little more than add your extra criteria to the existing search string. Since logically the default concatenation is AND this simply gives the existing pages that also contain the new criteria.

### More

Above the search window on the home page, fig 7.4, or indeed above the copy of this on the retrieved sites page there is a **more** link. Clicking this gives fig 7.7

- The **Alerts** is quite useful if used sensibly as it can keep you up to date with any topic; most useful for new. On selecting this option you can choose a topic with which you wish Google to keep you up to date. For example you might type in "arsenal football club". Hopefully this would keep you up to date with their results etc. As the new pages are detected by Google they are emailed directly to you. You can stop this at any time; there is an option on the emails to do this.
- The Google Toolbar is useful however I have not been able to download it to my profile in the lab; City seems to block it.

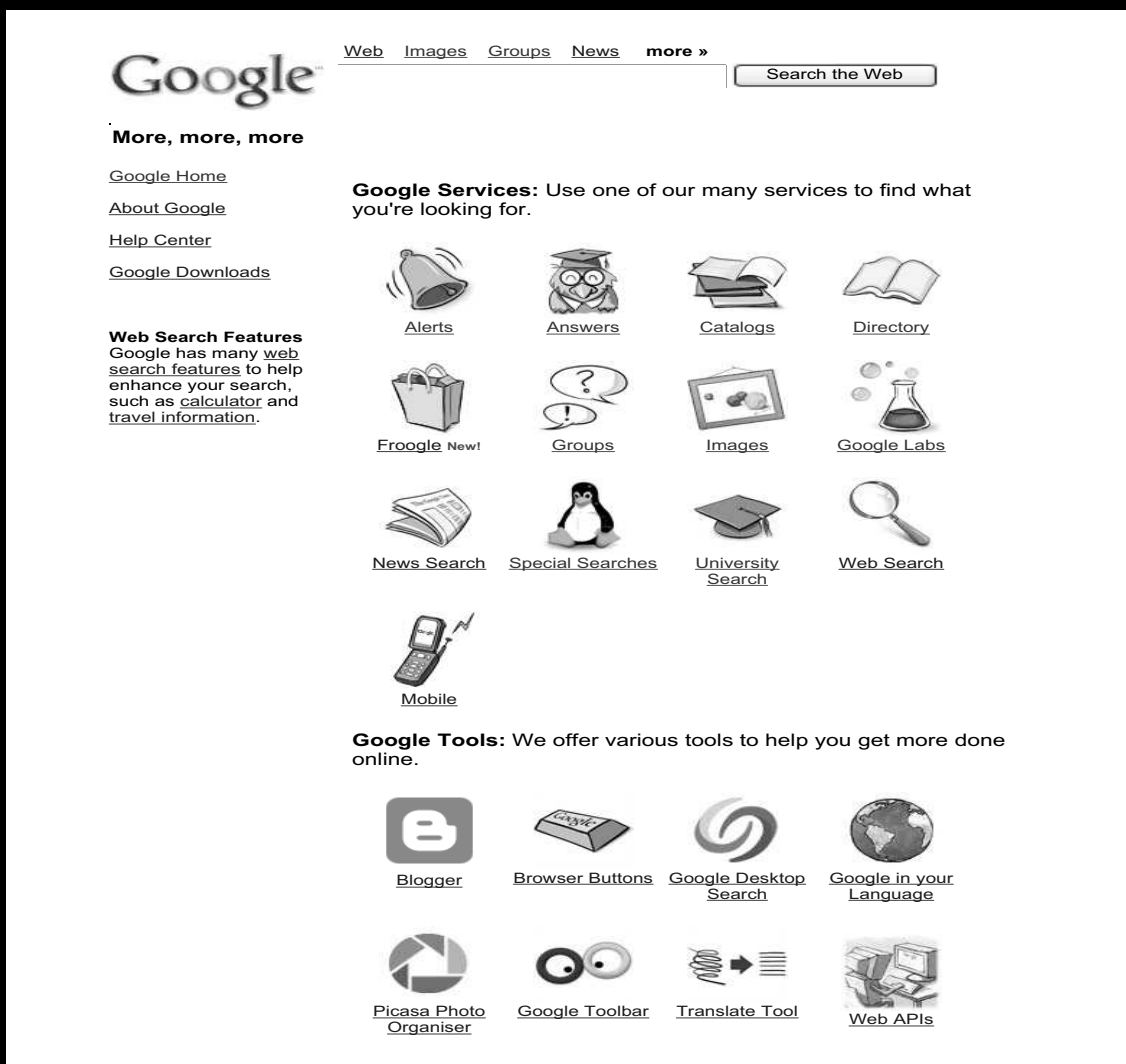


Figure 7.7: More, more, more - Tools and Services



## Chapter 8

# Polynomial Approximations

### 8.1 Introduction

This chapter looks at the construction and use of polynomials in approximating a given function or set of data. For clarity we make the following definition:

**Definition**

A polynomial in  $x$  of **degree**  $n$  is a function of the form:

$$p_n(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n$$

where the coefficients  $a_i$  are independent of  $x$ .

If  $n = 1$  then  $p_1(x) = a_0 + a_1x$  which is a linear polynomial or straight line

If  $n = 2$  then  $p_2(x) = a_0 + a_1x + a_2x^2$  a quadratic function.

We consider two types of problem:

- (i) If at a point  $x = a$  we are given, or can obtain <sup>1</sup>,  $f(a), f'(a), f''(a) \dots f^{(n)}(a)$  then we will construct a polynomial of degree at most  $n$  which will approximate  $f(x)$  near  $x = a$ .
- (ii) Given a set of data points  $\{(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$  we will construct a polynomial of degree at most  $n$  which passes through all these points.

The polynomial can be used to approximate, in (i),  $f(x)$  near to  $x = a$  and in (ii), the underlying function defining the data. The following examples illustrates how a polynomial may be used to simplify a problem.

**Example 8.1**

Assume we wish to evaluate  $\int_0^{0.5} \frac{x^2}{\sqrt{1-x^2}} dx$ . This integral can, with some ingenuity, be evaluated exactly but in many problems this is not possible. We will show here how a

---

<sup>1</sup>notation:  $f^{(n)}(x)$  denotes the  $n^{th}$  derivative of  $f(x)$

polynomial approximation to the integrand can be used to obtain a good approximation to the integral. Assuming the binomial expansion works for rational powers, we can write:

$$f(x) = \frac{x^2}{\sqrt{1-x^2}} = x^2(1-x^2)^{-1/2} = x^2\left(1 + \frac{x^2}{2} + \dots\right)$$

Thus provided that  $x$  is small, that is to say  $x$  is close to 0, then we can approximate  $f(x)$  with  $p_4(x) = x^2 + \frac{x^4}{2}$

Thus:

$$\int_0^{0.5} f(x) dx \approx \int_0^{0.5} \left(x^2 + \frac{x^4}{2}\right) dx = \left[\frac{x^3}{3} + \frac{x^5}{5}\right]_0^{0.5} = 0.0447916$$

In this example it is possible to evaluate the integral exactly to give  $\frac{\pi}{12} - \frac{\sqrt{3}}{8} = 0.0452930$ . Comparing the two values we have only incurred a 1% error using the polynomial, however the integration has been rendered almost trivial using  $p_4(x)$  instead of  $f(x)$ .

### Example 8.2

We now consider the same problem but this time we construct a polynomial through three data points generated from  $f(x)$  as follows:

$$x_0 = 0 \Rightarrow y_0 = f(0) = 0; \quad x_1 = 0.25 \Rightarrow y_1 = f(0.25) = 0.06455; \quad x_2 = 0.5 \Rightarrow y_2 = f(0.5) = 0.2887;$$

The quadratic through these three points can be shown to be given by:

$$p_2(x) = 1.2768x^2 - 0.061x$$

Thus as in the last example we can use  $p_2(x)$  to approximate  $f(x)$ .

$$\int_0^{0.5} f(x) dx \approx \int_0^{0.5} (1.2768x^2 - 0.061x) dx = \left[1.2768\frac{x^3}{3} - 0.061\frac{x^2}{2}\right]_0^{0.5} = 0.0445575$$

Thus using a quadratic through three points in the range of integration gives an even better answer than in the first case where we constructed an approximation that was valid for small  $x$ .

## 8.2 Taylor's Polynomial Approximation

In this approach we use the information regarding  $f(x)$  and its derivatives at a given point  $x = a$ . The approximation is then good for values of  $x$  near  $x = a$ .

### 8.2.1 Linear - tangent approximation

Consider the function  $f(x) = \sin(x)$  near the point  $x = \frac{\pi}{4}$ . In fig.8.1 the coordinates of  $P$  are  $(\frac{\pi}{4}, \frac{1}{\sqrt{2}})$  and the gradient of the tangent to the curve at  $P$  is given by  $f'(\frac{\pi}{4}) = \cos(\frac{\pi}{4}) = \frac{1}{\sqrt{2}}$ . The tangent at  $P$  is then used as an approximation to the curve near  $\pi/4$  and its equation is given by:

$$y - \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left( x - \frac{\pi}{4} \right) \Rightarrow y = p_1(x) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \left( x - \frac{\pi}{4} \right)$$

The general approach to this problem is to construct a polynomial  $p_1(x)$  which has the same value and same first derivative as  $f(x)$  at the point  $x = a$ . The smartest way to write out the polynomial is in the form  $p_1(x) = a_0 + a_1(x - a)$ , as in the above example, and then find  $a_0$  and  $a_1$  as follows:

- i)  $p_1(a) = f(a) \Rightarrow a_0 = f(a)$
- ii)  $p_1'(a) = f'(a) \Rightarrow a_1 = f'(a)$

Thus the general linear approximation to  $f(x)$  about  $x = a$  is given by:

$$p_1(x) = f(a) + f'(a)(x - a)$$

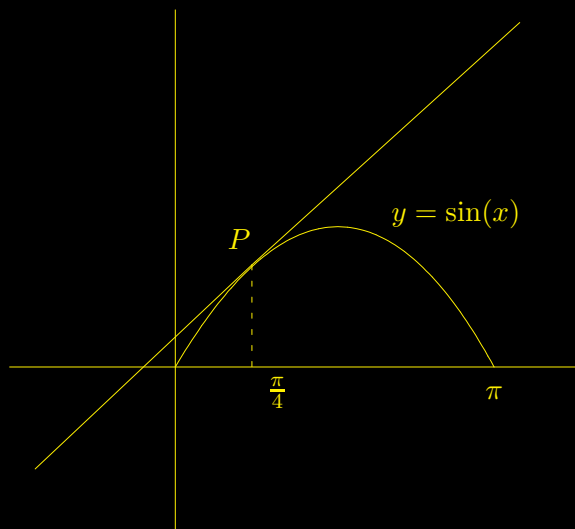


Figure 8.1:

### 8.2.2 Quadratic approximation

We construct the quadratic polynomial approximation to  $f(x)$  about  $x = a$  by requiring that the value of the polynomial and its first and second derivatives agree with those of

$f(x)$  at  $x = a$ . Thus again being smart and writing

$$p_2(x) = a_0 + a_1(x - a) + a_2(x - a)^2$$

we obtain the coefficients as follows:

i)  $p_2(a) = a_0$  and since we require  $p_2(a) = f(a)$  we have  $a_0 = f(a)$

ii)  $p_2'(x) = a_1 + 2a_2(x - a)$  and since we require  $p_2'(a) = f'(a)$  we have  $a_1 = f'(a)$ .

iii)  $p_2''(x) = 2a_2$  and since we require  $p_2''(a) = f''(a)$  we have

$$2a_2 = f''(a) \quad \Rightarrow \quad a_2 = \frac{f''(a)}{2}$$

Thus the general quadratic approximation to  $f(x)$  about  $x = a$  is given by

$$p_2(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2$$

### 8.2.3 General approximation

To approximate  $f(x)$  about  $x = a$  with an  $n^{\text{th}}$  degree polynomial we start the above process by taking :

$$p_n(x) = a_0 + a_1(x - a) + a_2(x - a)^2 + a_3(x - a)^3 + a_4(x - a)^4 + \dots + a_n(x - a)^n$$

By applying the condition that  $p_n(x)$  and all its derivatives up to the  $n^{\text{th}}$  must agree with those of  $f(x)$  at  $x = a$  we obtain, as above, the following expression for  $a_i$ :

$$a_i = \frac{f^{(i)}(a)}{i!}$$

(The  $i!$  is built up from repeatedly differentiating  $(x - a)^i$ )

Thus we have

$$p_n(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + \frac{f^{(iv)}(a)}{4!}(x - a)^4 + \dots + \frac{f^{(n)}(a)}{n!}(x - a)^n$$

or using the summation convention

$$p_n(x) = \sum_{i=0}^{i=n} \frac{f^{(i)}(a)}{i!} (x - a)^i$$

This is referred to as the  $n^{\text{th}}$  order<sup>2</sup> Taylor Polynomial of  $f(x)$  about  $x = a$ .

### Example 8.3

<sup>2</sup>The word **order** is used to indicate we have used the  $n^{\text{th}}$  derivative of  $f(x)$ , if this is non zero at  $x = a$  then  $p_n(x)$  will have degree  $n$ , otherwise the degree of  $p_n(x)$  will be less than  $n$

We now calculate, up to the fifth order, the Taylor polynomials of  $f(x) = \sin(x)$  about  $x = \pi/4$ .

$n =$	$f^{(n)}(x)$	$f^{(n)}(\frac{\pi}{4})$	$p_n(x)$
0	$\sin(x)$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$
1	$\cos(x)$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}\{1 + (x - \frac{\pi}{4})\}$
2	$-\sin(x)$	$-\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}\{1 + (x - \frac{\pi}{4}) - \frac{1}{2}(x - \frac{\pi}{4})^2\}$
3	$-\cos(x)$	$-\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}\{1 + (x - \frac{\pi}{4}) - \frac{1}{2}(x - \frac{\pi}{4})^2 - \frac{1}{3!}(x - \frac{\pi}{4})^3\}$
4	$\sin(x)$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}\{1 + (x - \frac{\pi}{4}) - \frac{1}{2}(x - \frac{\pi}{4})^2 - \frac{1}{3!}(x - \frac{\pi}{4})^3 + \frac{1}{4!}(x - \frac{\pi}{4})^4\}$
5	$\cos(x)$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}\{1 + (x - \frac{\pi}{4}) - \frac{1}{2}(x - \frac{\pi}{4})^2 - \frac{1}{3!}(x - \frac{\pi}{4})^3 + \frac{1}{4!}(x - \frac{\pi}{4})^4 + \frac{1}{5!}(x - \frac{\pi}{4})^5\}$

Fig.8.2 and Fig.8.3 show how the agreement between the polynomial approximations to  $f(x) = \sin x$  improves as the order of the approximation is increased.

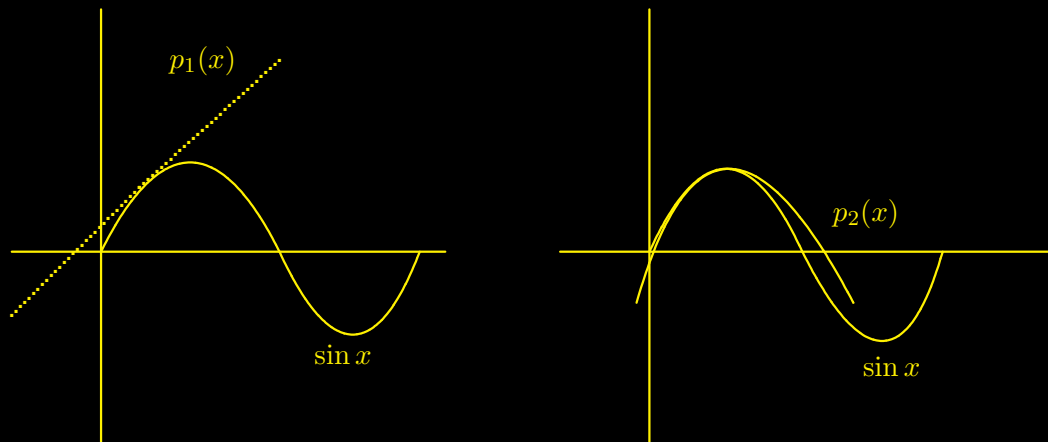
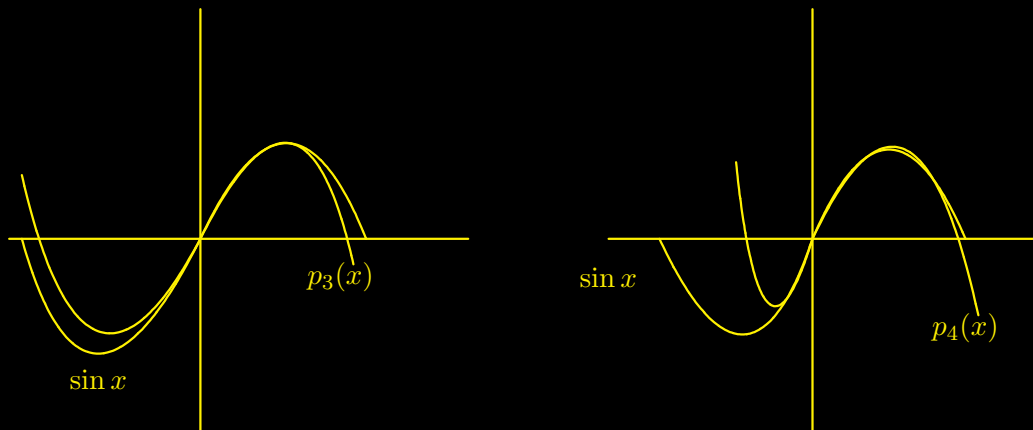


Figure 8.2: linear and quadratic approximations to  $\sin x$  about  $x = \frac{\pi}{4}$

### 8.2.4 Maclaurin's Expansion - Example

If we choose  $a = 0$  in the above Taylor polynomial we obtain

$$p_n(x) = \sum_{i=0}^{i=n} \frac{f^{(i)}(0)}{i!} x^i = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(0)\frac{x^3}{3!} + \dots + f^{(n)}(0)\frac{x^n}{n!}$$

Figure 8.3: cubic and quartic approximations to  $\sin x$  about  $x = \frac{\pi}{4}$ 

This expansion is often referred to as Maclaurin's polynomial. Consider the case  $f(x) = (1+x)^{\frac{1}{2}}$  then we can obtain the expansion as follows:

$$f(0) = 1; \quad f'(x) = \left(\frac{1}{2}\right) (1+x)^{-\frac{1}{2}} \Rightarrow f'(0) = \frac{1}{2}; \quad f''(x) = \left(\frac{1}{2}\right) \left(-\frac{1}{2}\right) (1+x)^{-\frac{3}{2}} \Rightarrow f''(0) = -\frac{1}{2^2} = -\frac{1}{4}$$

Repeated application of this process will give us a polynomial expansion of any order. Carrying the process on up to the 4<sup>th</sup> order gives:

$$p_4(x) = 1 + \frac{x}{2} + \left(\frac{1}{2}\right) \left(-\frac{1}{2}\right) \frac{x^2}{2} + \left(\frac{1}{2}\right) \left(-\frac{1}{2}\right) \left(-\frac{3}{2}\right) \frac{x^3}{3!} + \left(\frac{1}{2}\right) \left(-\frac{1}{2}\right) \left(-\frac{3}{2}\right) \left(-\frac{5}{2}\right) \frac{x^4}{4!}$$

This is clearly just the binomial expansion

$$(1+x)^\alpha = 1 + \alpha x + \alpha(\alpha-1) \frac{x^2}{2!} + \alpha(\alpha-1)(\alpha-2) \frac{x^3}{3!} + \alpha(\alpha-1)(\alpha-2)(\alpha-3) \frac{x^4}{4!} \dots \quad \text{with } \alpha = \frac{1}{2}$$

. Carrying out the arithmetics gives us:

$$p_4(x) = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5}{128}x^4$$

### 8.2.5 The Error Term

In all the above examples approximating  $f(x)$  with a polynomial leads to an error. Formally we can write:

$$f(x) - p_n(x) = R_n(x)$$

where  $R_n(x)$  is the error term<sup>3</sup>. We now consider a method for calculating a bound for the error term. ie for a given problem we find a value  $B(x)$  such that  $|R_n(x)| < B(x)$ .

<sup>3</sup>also referred to as the remainder term

**Taylor's Theorem**

Suppose that  $f(x)$  is an  $n$ -fold continuously differentiable function defined on an interval  $[\alpha, \beta]$  and that the  $(n + 1)^{th}$  derivative of  $f(x)$ , namely  $f^{(n+1)}(x)$  exists on  $[\alpha, \beta]$  and that  $a \in [\alpha, \beta]$ .

For every  $x \in [\alpha, \beta]$  there exists  $t \in (a, x)^4$  such that:

$$f(x) - p_n(x) = R_n(x)$$

where  $p_n(x)$  is the Taylor polynomial:

$$\sum_{i=0}^n f^{(i)}(a) \frac{(x-a)^i}{i!} \quad (8.1)$$

and the remainder <sup>5</sup>is given by:

$$R_n(x) = f^{(n+1)}(t) \frac{(x-a)^{n+1}}{(n+1)!} \quad (8.2)$$

It would appear that we can always calculate the error when making an approximation, this is not the case, as we are **not** able to find the value of  $t$ , we only know that its some value in the interval  $(a, x)$ . In practice we find the maximum value,  $M$ , of  $|f^{(n+1)}(t)|$  for  $t \in [a, x]$  which enables us to write:

$$|R_n(x)| \leq M \left| \frac{(x-a)^{n+1}}{(n+1)!} \right|$$

The righthand side of this equation then forms a bound for our error.

**Example 8.4**

Consider  $f(x) = \ln(x)$  about  $x = 1$ . By repeated differentiation of  $f(x)$  it is easy to show that:

$$f^{(i)}(x) = (-1)^{i-1} \frac{(i-1)!}{x^i} \quad \text{and} \quad f^{(i)}(1) = (-1)^{i-1} (i-1)! \quad (8.3)$$

Hence we obtain the following polynomials of degree 0 to  $n$ :

<sup>4</sup>If  $x < a$  then read as  $(x, a)$ , etc

<sup>5</sup>Known as Lagrange's form of the remainder

$i =$	$f^{(i)}(x)$	$f^{(i)}(1)$	$p_i(x)$
0	$\ln(x)$	0	0
1	$\frac{1}{x}$	1	$(x - 1)$
2	$-\frac{1}{x^2}$	-1	$(x - 1) - \frac{(x - 1)^2}{2}$
3	$\frac{2}{x^3}$	2	$(x - 1) - \frac{(x - 1)^2}{2} + \frac{(x - 1)^3}{3}$
4	$-\frac{2 \times 3}{x^4}$	$-2 \times 3$	$(x - 1) - \frac{(x - 1)^2}{2} + \frac{(x - 1)^3}{3} - \frac{(x - 1)^4}{4}$
$n > 0$	$(-1)^{n-1} \frac{(n-1)!}{x^n}$	$(-1)^{n-1} (n - 1)!$	$(x - 1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} \dots (-1)^{n-1} \frac{(x-1)^n}{n}$

Consider now  $p_3(x)$  and the error incurred when approximating  $f(1.5)$ . From the above :

$$p_3(x) = (x - 1) - \frac{(x - 1)^2}{2} + \frac{(x - 1)^3}{3} \Rightarrow p_3(1.5) = 0.41667$$

For  $x \geq 1$  the absolute value of the remainder term (ie its size) is given by:

$$|R_3(x)| = |f^{(iv)}(t)| \times \left| \frac{(x - 1)^4}{4!} \right| = \frac{3!}{t^4} \times \frac{(x - 1)^4}{4!} \quad 1 < t < x$$

Since for  $t \in [1, x]$ ,  $\frac{1}{t^4}$  is maximum at  $t = 1$  (ie always less than or equal to 1) we can say:

$$|R_3(x)| \leq \frac{(x - 1)^4}{4}$$

At the point  $x = 1.5$  this gives:

$$|R_3(1.5)| \leq \frac{(0.5)^4}{4} = 0.015625$$

Thus

$$f(1.5) = \ln(1.5) = 0.41667 \pm 0.015625$$

A simple check using our calculator to evaluate  $\ln(1.5)$  shows that the actual error does indeed lie within the calculated limits. In detail:

$$|p_3(1.5) - \ln(1.5)| = 0.0112 < 0.015625$$

### Example 8.5

Consider the evaluation of

$$\int_1^2 \ln x \, dx$$



by, as in the above example, writing

$$\ln x = p_3(x) + R_3(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} + R_3(x).$$

Thus we have:

$$\int_1^2 \ln x \, dx \approx \int_1^2 p_3(x) \, dx = \int_1^2 \left\{ (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} \right\} dx = 0.41667$$

The size of the error incurred is given by:<sup>6</sup>

$$\left| \int_1^2 \ln x \, dx - \int_1^2 p_3(x) \, dx \right| \leq \int_1^2 |\ln x - p_3(x)| \, dx = \int_1^2 |R_3(x)| \, dx$$

From the above example we see that

$$|R_3(x)| \leq \frac{(x-1)^4}{4}$$

thus

$$\text{size of error} \leq \int_1^2 \frac{(x-1)^4}{4} dx = 0.05$$

Thus we can say that

$$\int_1^2 \ln x \, dx = 0.41667 \pm 0.05$$

In this example it is possible to evaluate the integral by standard methods, to give correct to three decimal places:

$$\int_1^2 \ln x \, dx = 0.386,$$

which is clearly within 0.05 of our approximation.

### 8.2.6 Convergence of $p_n(x)$ as $n$ tends to infinity

We have seen in our examples so far that by increasing the order of the Taylor polynomial we are able to improve the accuracy of our approximation. This can be seen graphically in fig.8.2 and fig.8.3 and in the above example the approximation to  $\ln(1.5)$  can be shown to improve with an increase in the order of the approximation. ie calculating  $p_4(1.5)$  would give a better approximation to  $\ln(1.5)$  than  $p_3(1.5)$  However this is not always the case, increasing the number of term can make the approximation worse. We can analyse this problem by looking at  $R_n(x)$  as  $n \rightarrow \infty$ .

Since  $f(x) = p_n(x) + R_n(x)$  we see that:

---

<sup>6</sup>using the general result for integrals that  $\left| \int_a^b f(x) \, dx \right| \leq \int_a^b |f(x)| \, dx$

- If  $R_n(x)$  tends to zero as  $n$  tends to infinity then overall the accuracy will improve with increasing  $n$  and in the limit as  $n$  tends to infinity  $p_n(x)$  will tend to  $f(x)$  - the nice case!
- If  $R_n(x)$  grows without bound as  $n$  tends to infinity then the overall accuracy will get worse as  $n$  is increased and the polynomial will diverge as  $n$  tends to infinity. - the nasty case!
- If  $R_n(x)$  tends to some finite value  $\alpha(x)$  as  $n$  tends to infinity then the polynomial will tend to some value that is finite but different to  $f(x)$ . This case is not common and is not considered here - the rare and pathological case ! <sup>7</sup>

As  $n \rightarrow \infty$  the expansion of  $p_n(x)$  is referred to as the **Taylor series** of  $f(x)$  about the point  $x = a$  and is denoted as:

$$\sum_{i=0}^{\infty} f^{(i)} \frac{(x-a)^i}{i!} \quad (8.4)$$

As we have noted above, this infinite series may or may not tend to  $f(x)$  for all values of  $x$ .

### Example 8.6

Returning to our example of  $f(x) = \ln(x)$  expanded about  $x = 1$  we have from Eq.8.3 that

$$f^{(i)}(t) = (-1)^{i-1} \frac{(i-1)!}{t^i}$$

Thus using this with  $i = n + 1$  the size of the remainder term is given by:

$$|R_n(x)| = \left| \frac{(x-1)^{n+1}}{(n+1)!} f^{(n+1)}(t) \right| = \frac{|(x-1)|^{n+1}}{(n+1)!} \frac{n!}{|t^{n+1}|} = \frac{|(x-1)|^{(n+1)}}{(n+1)|t|^{(n+1)}}$$

where  $t$  is some value between 1 and  $x$ . If we consider the case  $x > 1$  then  $1 < t < x$ , thus since  $\frac{1}{t^{(n+1)}} < 1$  in this range we can write:

$$|R_n(x)| \leq \frac{(x-1)^{(n+1)}}{(n+1)} \quad (8.5)$$

We now consider this inequality as  $n \rightarrow \infty$ .

- If  $0 < (x-1) < 1$  then  $(x-1)^{(n+1)} \rightarrow 0$  as  $n \rightarrow \infty$ . Thus for  $1 < x < 2$ , the righthand side of Eq.8.5 tends to zero as  $n$  tends to infinity, hence  $R_n(x)$  tend to zero and the polynomial  $p_n(x)$  tend to  $f(x)$ .
- If  $x = 2$  then the righthand side of Eq.8.5 becomes  $\frac{1}{(n+1)}$  which tend to zero as  $n$  tends to infinity. Thus we can say that  $p_n(2)$  tend to  $f(2)$ .

<sup>7</sup>The inquisitive might like to consider  $f(x)$  defined by:  $f(x) = e^x + e^{-\frac{1}{x^2}}$   $x \neq 0$  and  $f(0) = 1$ , expanded about  $x = 0$ .

- **Summary**

Since  $p_n(1) = f(1)$  for all  $n$  we have shown that:

$$1 \leq x \leq 2 \Rightarrow p_n(x) \rightarrow f(x) \quad \text{as } n \rightarrow \infty$$

That is to say, overall the polynomial approximations get better for value of  $x$  in the region  $1 \leq x \leq 2$ .

- If  $(x - 1) > 1$  then it can be shown that the righthand side of Eq.8.5 tends to infinity as  $n$  tends to infinity. Since Eq.8.5 is only an inequality this tells us nothing about  $R_n(x)$ , since the statement  $|R_n(x)| \leq \infty$ , is, for what its worth, never false.

Since in most cases if the Taylor series converges then it converges to  $f(x)$  it is worth developing tests which give ranges of values of  $x$  for which the series converges. One of the most useful test is D'Alembert's ratio test

### 8.2.7 D'Alembert's Ratio test

Given the infinite series  $u_0 + u_1 + u_2 + u_3 + \dots$  then the ratio test states that if

$$\left| \frac{u_{i+1}}{u_i} \right| \rightarrow L \quad \text{as } n \rightarrow \infty$$

then

- If  $L < 1$  then the series is convergent
- If  $L > 1$  then the series is divergent
- If  $L = 1$  the test gives no information, the series may converge or diverge

If we apply this to the Taylor series of  $f(x)$  about  $x = a$ , see Eq.8.4, then

$$u_i = \frac{f^{(i)}(a)(x - a)^i}{i!}$$

and hence for convergence the ratio test requires:

$$\left| \frac{u_{i+1}}{u_i} \right| = \left| \frac{f^{(i+1)}(a)(x - a)^{(i+1)}}{(i + 1)!} \times \frac{i!}{(x - a)^i f^{(i)}(a)} \right| = \left| \frac{(x - a)}{(i + 1)} \right| \times \left| \frac{f^{(i+1)}(a)}{f^{(i)}(a)} \right| < 1 \quad \text{as } i \rightarrow \infty$$

For the  $\ln(x)$  example where  $a = 1$  and the  $i^{\text{th}}$  derivative of  $f(x)$  at  $x = 1$  is given by (see Eq.8.3)

$$f^{(i)}(1) = (-1)^{i-1}(i - 1)!$$

the ratio test gives:

$$\left| \frac{u_{i+1}}{u_i} \right| = \left| \frac{(x - 1)}{i + 1} \right| \times \frac{i!}{(i - 1)!} = |(x - 1)| \times \frac{i}{i + 1} = |(x - 1)| \times \frac{1}{(1 + \frac{1}{i})} \rightarrow |(x - 1)| \quad \text{as } i \rightarrow \infty$$

Thus provided  $|(x - 1)| < 1$  the series converges. Thus by the ratio test the series expansion of  $\ln(x)$  about  $x = 1$  converges provided  $0 < x < 2$  and diverges if  $x < 0$  or  $x > 2$ . The test gives no information for  $x = 2$  or  $x = 0$ . (we can expect problems with the series at  $x = 0$  since  $\ln(0)$  is not defined; we have already discussed the case  $x = 2$ .)

## 8.3 Polynomial Interpolation

We now consider the problem of constructing a polynomial through a set of points, which may be given or generated from a given function. The polynomial is said to interpolate the points as it gives us a means of calculating values of the data between the given points. We consider in this section three different methods, the first being that due to Lagrange. This is perhaps the most straightforward method but not always the best, it does however lead to other results later in the course. The second method uses Newton's difference table to create the polynomial, this is a more flexible method than Lagrange's as it allows us to easily vary the degree of the polynomial. The third method is that of Splines where we don't construct a single polynomial through all the points but construct either linear or cubic polynomials between adjacent points. Before embarking on our three methods consider first a simple example illustrating a direct approach to the problem.

### 8.3.1 Direct Method

#### Example

Find a degree 2 polynomial passing through the given points  $\{(0, 1), (1, 2), (2, 5)\}$ . Set  $p_2(x) = a_0 + a_1x + a_2x^2$ . Thus

$$p_2(0) = 1 = a_0 \quad p_2(1) = 2 = a_0 + a_1 + a_2 \quad p_2(2) = 5 = a_0 + 2a_1 + 4a_2$$

Solving these gives  $a_0 = 1$ ,  $a_1 = 0$  and  $a_2 = 1$ . Thus the required polynomial is

$$p_2(x) = x^2 + 1.$$

Thus we have found the polynomial of degree at most two through the three points. The term "at most" suggests that there may have been a polynomial of degree less than two. Had this been the case we would have found  $a_2 = 0$ . There are of course an infinite number of polynomials of degree greater than two that pass through the points.

This direct method will always require us to solve a set of simultaneous equations, this may not be trivial for a problem involving a large number of points. We therefore consider other methods of generating the polynomial.

### 8.3.2 Lagrange's Method

Lagrange's method is centred about a set of polynomials referred to as the Lagrange Basis Polynomials which have the property of either vanishing at the data points or being equal to one. We now construct the Lagrange polynomials for a set of three data points.

Consider the three points  $\{(1, -2), (2, -1), (3, 2)\}$ .

Consider the quadratics:

$$l_0(x) = \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{x^2}{2} - \frac{5}{2}x + 3$$

$$l_1(x) = \frac{(x-1)(x-3)}{(2-1)(2-3)} = -x^2 + 4x - 3$$

$$l_2(x) = \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{x^2}{2} - \frac{3}{2}x + 1$$

It is easy to see that  $l_0(1) = 1$ ,  $l_0(2) = 0$  and  $l_0(3) = 0$ . Similarly  $l_1(2) = 1$  and is zero at the other two points and  $l_2(3) = 1$  and is zero at the other two points. Note that these three polynomials depend only on the  $x$ -coordinate of the three points and not the ordinates.

We now use the ordinate of the three points to construct the polynomial of degree at most two passing through the points. Consider

$$p_2(x) = (-2)l_0(x) + (-1)l_1(x) + (2)l_2(x). \quad (8.6)$$

$p_2(x)$  is clearly at most a quadratic since the  $l$ 's are quadratics. Using the properties of the  $l$ 's we see that

$$p_2(1) = (-2)l_0(1) + (-1)l_1(1) + (2)l_2(1) = (-2) \times 1 + (-1) \times 0 + (2) \times 0 = -2$$

$$p_2(2) = (-2)l_0(2) + (-1)l_1(2) + (2)l_2(2) = (-2) \times 0 + (-1) \times 1 + (2) \times 0 = -1$$

$$p_2(3) = (-2)l_0(3) + (-1)l_1(3) + (2)l_2(3) = (-2) \times 0 + (-1) \times 0 + (2) \times 1 = 2$$

Substituting into Eq 8.6 we obtain

$$p_2(x) = x^2 - 2x - 1.$$

### Definition - Lagrange Basis Polynomial

Given the set of values  $\{x_0, x_1, \dots, x_n\}$  then the Lagrange basis polynomials corresponding to these values are the  $n^{\text{th}}$  degree polynomials  $l_k(x)$  given by:

$$l_k(x) = \frac{(x-x_0)(x-x_1)\dots \text{omit } (x-x_k) \dots (x-x_n)}{(x_k-x_0)(x_k-x_1)\dots \text{omit } (x_k-x_k) \dots (x_k-x_n)} \quad k = 0, 1, \dots, n$$

which can be seen to have the following properties:

- $l_k(x_k) = 1 \quad k = 0, \dots, n$
- $l_k(x_i) = 0 \quad i \neq k, \quad i, k = 0, \dots, n$

Using the product notation  $\prod$ , similar to the summation notation  $\sum$  except that we multiply terms rather than add, the basis polynomial can be expressed as

$$l_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^n \frac{(x-x_i)}{(x_k-x_i)} \quad k = 0, 1, \dots, n \quad (8.7)$$

**NB** The Lagrange Basis polynomials depend only on the  $x$ -values and not the  $y$ -values.

Thus the Lagrange polynomial through the points  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$  is given by:

$$p_n(x) = y_0 l_0(x) + y_1 l_1(x) + \dots + y_n l_n(x) = \sum_{k=0}^n y_k l_k(x) \quad (8.8)$$

### Example 8.7

If we are given a function  $y = f(x)$  and a set of  $x$ -values  $\{x_0, x_1, \dots, x_n\}$ , then denoting  $f(x_k)$  by  $y_k$  we can generate the points  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$ . Lagrange's method can then be used to construct the polynomial  $p_n(x)$  through these points, which in turn can be used to approximate  $f(x)$ .

Consider  $f(x) = \ln(1 + x)$  and the set of  $x$ -values  $x_0 = 0$ ,  $x_1 = 1$  and  $x_2 = 2$ . Using a calculator we can obtain correct to six decimal places

$$y_0 = f(0) = \ln(1) = 0, \quad y_1 = f(1) = \ln(2) = 0.693147 \quad \text{and} \quad y_2 = f(2) = \ln(3) = 1.098612$$

From Eq 8.7 and Eq 8.8 we obtain:

$$p_2(x) = 0 \times l_0(x) + y_1 \frac{(x-0)(x-2)}{(1-0)(1-2)} + y_2 \frac{(x-0)(x-1)}{(2-0)(2-1)} = -0.143841x^2 + 0.836988x$$

This polynomial can then be used in any calculation instead of  $\ln(1 + x)$ . For example:

$$\int_{0.5}^{1.5} \frac{\ln(1+x)}{x} dx \approx \int_{0.5}^{1.5} -0.143841x + 0.836988 dx = 0.693147$$

Using a more accurate numerical method a value of 0.699, correct to three decimal places, can be used for comparison.

### Error Term

If the points are generated using a function  $f(x)$ , as in the above example, it is theoretically possible to find an error bound on the polynomial approximation. A similar result was obtained when considering the Taylor polynomial approximation; see Eq 8.2 for the remainder(error) term in that case. The following theorem gives an expression for the remainder (error) term.

### Theorem

Suppose  $\{x_0, x_1 \dots x_n\}$  are distinct points in the interval  $[a, b]$ , and that  $f(x)$  is *nice*<sup>8</sup> on  $[a, b]$  then if  $p_n(x)$  is the Lagrange polynomial through the points  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$  where  $y_k = f(x_k)$  then for each  $x \in [a, b]$  we can find a  $t \in (a, b)$  such that:

$$f(x) - p_n(x) = \frac{(x-x_0)(x-x_1) \dots (x-x_n)}{(n+1)!} f^{(n+1)}(t) \quad (8.9)$$

<sup>8</sup>  $f(x)$  is  $(n+1)$  times continuously differentiable on  $[a, b]$

As with the Taylor expansion we are not able to find the value of  $t$  but by using a bound on the  $(n + 1)^{th}$  derivative of  $f(x)$  in the interval  $(a, b)$  we can write down an error bound as follows:

$$|\text{error}| = |f(x) - p_n(x)| \leq \frac{(x - x_0)(x - x_1) \dots (x - x_n)}{(n + 1)!} \max_{t \in [a, b]} |f^{(n+1)}(t)| \quad (8.10)$$

If we apply this result to the above example where  $f(x) = \ln(1 + x)$ ,  $n = 2$ , the interval is  $[0, 2]$  and  $p_2(x) = -0.143841x^2 + 0.836988x$  then:

$$|\text{error}| = \left| \frac{(x - 0)(x - 1)(x - 2)}{3!} \right| |f^{(3)}(t)| \leq \left| \frac{x(x - 1)(x - 2)}{6} \right| \max_{t \in [0, 2]} \left| \frac{2}{(1 + t)^3} \right|$$

Since the maximum value of  $2/(1 + t)^3$  for  $t \in [0, 2]$  occurs at  $t = 0$ ,

$$|\text{error}| < \left| \frac{x(x - 1)(x - 2)}{6} \right| \times 2$$

If we consider the approximation at  $x = 0.5$  then

$$f(0.5) = \ln(1.5) \approx p_2(0.5) = 0.382533$$

and the bound for the error is given by

$$|\text{error}| < \left| \frac{0.5(0.5 - 1)(0.5 - 2)}{3} \right| = 0.125$$

Thus we can say that

$$\ln(1.5) = 0.382533 \pm 0.125$$

Using the more accurate value of  $\ln(1.5) = 0.4054$  we see that our actual error is approximately 0.02, which is well within the calculated range of  $\pm 0.125$ . It is quite common for our actual error to be well within the error bound calculated from Eq 8.10.

### 8.3.3 Divided Differences

Lagrange's method is good but if you wish to increase the degree of the polynomial  $p(x)$  by introducing more points then it is necessary to calculate another set of basis polynomials  $l_i(x)$ . The method of differences overcomes this problem for the case where the additional points are added after  $x_n$ , as well as giving a lead into other numerical methods.

Given the set of data points  $\{(x_0, y_0), (x_1, y_1) \dots (x_n, y_n)\}$  where the  $y_n$  are either given or calculated from a given function using  $y_n = f(x_n)$  we construct an  $n^{th}$  degree polynomial as follows:

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \dots + a_n(x - x_0)(x - x_1) \dots (x - x_{n-1})$$

and find the coefficients  $a_i$  such that the polynomial passes through the data points. That is to say we find the coefficients such that  $p_n(x_i) = f(x_i)$  or  $p_n(x_i) = y_i$  as appropriate.

$$p_n(x_0) = f(x_0) \Rightarrow f(x_0) = a_0$$

$$p_n(x_1) = f(x_1) \Rightarrow f(x_1) = a_0 + a_1(x_1 - x_0) \Rightarrow a_1 = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}$$

$$p_n(x_2) = f(x_2) \Rightarrow f(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

$$\Rightarrow a_2 = \frac{\frac{f(x_2) - f(x_1)}{(x_2 - x_1)} - \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}}{(x_2 - x_0)}$$

If we continue with this process of evaluating the coefficients then we will soon see that a certain pattern is emerging, this leads us to the following definitions for the divided difference. The evaluation of the divided differences will then be carried out by constructing a difference table.

### Definition - Divided Differences

- The **zero order** divided difference of  $f(x)$  with respect to  $x_i$  is trivially defined to be  $f(x_i)$ . Introducing a square bracket notation we write  $f[x_i] = f(x_i)$
- The first order divided difference with respect to  $x_i$  and  $x_{i+1}$  is defined and denoted by:

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i} \quad (8.11)$$

Thus we define the **first order** divided difference as the difference of two zero order divided differences.

- Repeating the inductive definition we define the **second order** divided difference of  $f(x)$  with respect to  $\{x_i, x_{i+1}, x_{i+2}\}$  as:

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+2}, x_{i+1}] - f[x_{i+1}, x_i]}{x_{i+2} - x_i}$$

- Thus we continue to define higher order divided difference in terms of the lower ones. Note that as we increase the order we need to increase the number of points involved. Starting with just a single point for the zero order, two points for the first order, three points for the second order etc.

The polynomial  $p_n(x)$  can now be written as:

$$p_n(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \dots \quad (8.12)$$

The pattern is clear but how to evaluate the divided differences in a simple manner is not. If we are calculating by hand we construct a divided difference table to obtain the different differences. For programmers the inductive definition is a gift and can be easily programmed. Table 8.1 shows the general divided difference table for five points.

### Example 8.8



Given:  $f(x) = \ln(x)$  and the four points  $f(1) = 0$ ,  $f(1.4) = 0.3365$ ,  $f(1.5) = 0.4055$  and  $f(2) = 0.6931$ , the divided difference table is given by table 8.2. We can now use the first entries in each of the columns (*underlined in table 8.2*) to form the linear polynomial through the first two points, the quadratic through the first three points and the cubic through all four points as follows:

$$p_1(x) = f[x_0] + (x - 1)f[x_0, x_1] = 0 + 0.841(x - 1) = \underline{-0.841 + 0.841x}$$

$$\begin{aligned} p_2(x) &= p_1(x) + (x - 1)(x - 1.4)f[x_0, x_1, x_2] \\ &= -0.841 + 0.841x + (x - 1)(x - 1.4)(-0.3026) \\ &= \underline{-1.2649 + 1.5675x - 0.3026x^2} \end{aligned}$$

$$\begin{aligned} p_3(x) &= p_2(x) + (x - 1)(x - 1.4)(x - 1.5)f[x_0, x_1, x_2, x_3] \\ &= -1.2649 + 1.5675x - 0.3026x^2 + (x - 1)(x - 1.4)(x - 1.5)(0.1113) \\ &= \underline{-1.4987 + 2.1240x - 0.7367x^2 + 0.1113x^3} \end{aligned}$$

If we wish to approximate  $f(1.7)$  that is to say  $\ln(1.7)$ , which from our calculator has the value 0.5306 to four places of decimal, then using each of the above polynomials in turn we obtain the following three values:

- $p_1(1.7) = 0.5889$      |error|  $\approx 0.0583$
- $p_2(1.7) = 0.5254$      |error|  $\approx 0.0052$
- $p_3(1.7) = 0.5300$      |error|  $\approx 0.0006$

$x$	$f(x)$	$1^{st}$	$2^{nd}$	$3^{rd}$
$x_0$	$f[x_0]$			
		$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$		
$x_1$	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1}$		$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$
$x_2$	$f[x_2]$		$f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \frac{f[x_3] - f[x_2]}{x_3 - x_2}$		$f[x_1, x_2, x_3, x_4] = \frac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$
$x_3$	$f[x_3]$		$f[x_2, x_3, x_4] = \frac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \frac{f[x_4] - f[x_3]}{x_4 - x_3}$		$f[x_2, x_3, x_4, x_5] = \frac{f[x_3, x_4, x_5] - f[x_2, x_3, x_4]}{x_5 - x_2}$
$x_4$	$f[x_4]$		$f[x_3, x_4, x_5] = \frac{f[x_4, x_5] - f[x_3, x_4]}{x_5 - x_3}$	
		$f[x_4, x_5] = \frac{f[x_5] - f[x_4]}{x_5 - x_4}$		
$x_5$	$f[x_5]$			

Table 8.1: Divided Difference Table for six general point

$x$	$\ln(x)$	$1^{st}$	$2^{nd}$	$3^{rd}$
1	<u>0</u>			
		$\frac{0.3365 - 0}{1.4 - 1} = \underline{\underline{0.8413}}$		
1.4	0.3365		$\frac{0.6900 - 0.8413}{1.5 - 1} = \underline{\underline{-0.3026}}$	
		$\frac{0.4055 - 0.3365}{1.5 - 1.4} = 0.6900$		$\frac{-0.1913 - (-0.3026)}{2 - 1} = \underline{\underline{0.1113}}$
1.5	0.4055		$\frac{0.5752 - 0.6900}{2 - 1.4} = -0.1913$	
		$\frac{0.6931 - 0.4055}{2 - 1.5} = 0.5752$		
2	0.6931			

Table 8.2: Divided Differences Table -  $f(x) = \ln x$

The error term for  $p_3(x)$  is given by:

$$\frac{f^{iv}(t)}{4!}(x - 1)(x - 1.4)(x - 1.5)(x - 2) \quad t \in (1, 2)$$

Since  $f(t) = \ln(t)$  we can deduce that:

$$|f^{iv}(t)| = \left| \frac{-6}{t^4} \right| < 6 \quad \text{for } t \in (1, 2)$$

Thus at  $x = 1.7$ :

$$|\text{error}| < \frac{6}{4!} |(1.7 - 1)(1.7 - 1.4)(1.7 - 1.5)(1.7 - 2)| = 0.00315$$

We see from above that the error incurred using  $p_3(1.7)$  to approximate  $\ln(1.7)$ , namely 0.0006, is well within this limit of accuracy.

### 8.3.4 Forward Differences

In many problems the data points are equally spaced, something that wasn't a requirement for the above methods. If however this is the case, a simplification to the method of constructing a polynomial through the data points is possible. Thus in general if we are given the data points  $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$  the  $x$ -coordinates are given by  $x_k = x_0 + kh$ , where  $h$  is referred to as the length of the tabular interval. This is illustrated in fig 8.4.

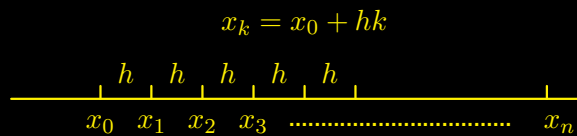


Figure 8.4: Equal tabular intervals

#### Definition - forward difference

The forward difference operator  $\Delta$  is defined by:

$$\Delta f(x_k) = f(x_{k+1}) - f(x_k)$$

or equivalently, if we are only give data points, by:

$$\Delta y_k = y_{k+1} - y_k$$

We say that  $\Delta f(x_k)$  is the **first order forward difference** of  $f(x)$  at  $x = x_k$ . This is related to the first order divided difference of  $f(x)$  at  $x = x_k$  (see Eq 8.11) as follows:

$$f[x_k, x_{k+1}] = \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} = \frac{\Delta f(x_k)}{h}$$

We continue the definition of the forward differences to define the  $n^{\text{th}}$  order forward difference in terms of the  $(n - 1)^{\text{th}}$  order forward difference by:

$$\Delta^n f(x_k) = \Delta^{n-1} f(x_{k+1}) - \Delta^{n-1} f(x_k)$$

Thus in detail, for example, the second order forward difference is given by:

$$\begin{aligned} \Delta^2 f(x_k) &= \Delta f(x_{k+1}) - \Delta f(x_k) \\ &= \{f(x_{k+2}) - f(x_{k+1})\} - \{f(x_{k+1}) - f(x_k)\} \\ &= f(x_{k+2}) - 2f(x_{k+1}) + f(x_k) \end{aligned}$$

Direct evaluation of the second order divided difference leads to the result:

$$f[x_k, x_{k+1}, x_{k+2}] = \frac{\Delta^2 f(x_k)}{2h^2}$$

In general it can be shown that the  $n^{\text{th}}$  order divided difference and the  $n^{\text{th}}$  order forward difference are related by:

$$f[x_k, x_{k+1}, x_{k+2}, \dots, x_{k+n}] = \frac{\Delta^n f(x_k)}{n!h^n} \quad (8.13)$$

The advantage of using the forward difference operator instead of the divided difference formulae is not immediately obvious. However calculating the forward differences proves to be simpler and more accurate than calculating the divided differences. Additionally many of the formulae involving divided differences simplify when we use the result in Eq(8.13). Table(8.4) shows the construction of a general forward difference table (*also simply referred to as a difference table*) and table(8.5) the difference table for the data given in table(8.3).

Using the data from table(8.3) and Eq(8.13) we reinterpret Eq(8.12) in terms of forward

$x$	1	1.5	2	2.5
$\cos x$	0.540	0.071	-0.416	-0.801

Table 8.3: tabulated values of  $\cos x$

differences.

For this problem Eq(8.12) gives:

$$p_3(x) = f(x_0) + (x-1)f[x_0, x_1] + (x-1)(x-1.5)f[x_0, x_1, x_2] + (x-1)(x-1.5)(x-2)f[x_0, x_1, x_2, x_3]$$

In terms of forward differences using Eq(8.13) we have:

$$p_3(x) = f(x_0) + (x-1)\frac{\Delta f(x_0)}{0.05} + (x-1)(x-1.5)\frac{\Delta^2 f(x_0)}{2!(0.5)^2} + (x-1)(x-1.5)(x-2)\frac{\Delta^3 f(x_0)}{3!(0.5)^3}$$

$x$	$f(x)$	$1^{st}$	$2^{nd}$	$3^{rd}$
$x_0$	$f(x_0)$			
		$\Delta f(x_0) = f(x_1) - f(x_0)$		
$x_1$	$f(x_1)$		$\Delta^2 f(x_0) = \Delta f(x_1) - \Delta f(x_0)$	
		$\Delta f(x_1) = f(x_2) - f(x_1)$		$\Delta^3 f(x_0) = \Delta^2 f(x_1) - \Delta^2 f(x_0)$
$x_2$	$f(x_2)$		$\Delta^2 f(x_1) = \Delta f(x_2) - \Delta f(x_1)$	
		$\Delta f(x_2) = f(x_3) - f(x_2)$		$\Delta^3 f(x_1) = \Delta^2 f(x_2) - \Delta^2 f(x_1)$
$x_3$	$f(x_3)$		$\Delta^2 f(x_2) = \Delta f(x_3) - \Delta f(x_2)$	
		$\Delta f(x_3) = f(x_4) - f(x_3)$		$\Delta^3 f(x_2) = \Delta^2 f(x_3) - \Delta^2 f(x_2)$
$x_4$	$f(x_4)$		$\Delta^2 f(x_3) = \Delta f(x_4) - \Delta f(x_3)$	
		$\Delta f(x_4) = f(x_5) - f(x_4)$		
$x_5$	$f(x_5)$			

Table 8.4: Forward Difference Table for six general points up to the third order forward difference

$x$	$\cos(x)$	$1^{st}$	$2^{nd}$	$3^{rd}$
1	<u>0.540</u>			
		$0.071 - 0.540 = \underline{\underline{-0.469}}$		
1.5	0.071		$(-0.487) - (-0.469) = \underline{\underline{-0.018}}$	
		$-0.416 - 0.071 = -0.487$		$0.102 - (-0.018) = \underline{\underline{0.120}}$
2.0	-0.416		$(-0.385) - (-0.487) = 0.102$	
		$-0.801 - (-0.416) = -0.385$		
2.5	-0.801			

Table 8.5: Forward Differences Table -  $f(x) = \cos x$

Substituting for the forward differences from table(8.5) we obtain:

$$\begin{aligned} p_3(x) &= 0.540 + (x - 1)\frac{-0.469}{0.05} + (x - 1)(x - 1.5)\frac{-0.018}{2!(0.5)^2} + (x - 1)(x - 1.5)(x - 2)\frac{0.120}{3!(0.5)^3} \\ &= 0.540 - 0.938(x - 1) - 0.036(x - 1)(x - 1.5) + 0.16(x - 1)(x - 1.5)(x - 2) \end{aligned}$$

Thus the cubic polynomial through the data points is given by:

$$p_3(x) = 0.16x^3 - 0.756x^2 + 0.192x + 0.944$$

### 8.3.5 Newton-Gregory formula

Clearly going through the above steps for each example is not acceptable, thus the following general formula is devised to express  $p_n(x)$  in terms of the forward difference operator. The first step is to move the origin to one of the tabular points, usually the first, namely  $x_0$ , and then introduce the parameter  $s$  as follows: Set

$$x = x_0 + sh \quad (8.14)$$

where  $h$  is the common tabular interval.

(recall the distance between consecutive  $x$  values is constant and denoted by  $h$ .)

Fig 8.5 illustrates the relationship between  $x$  and the new variable  $s$ .

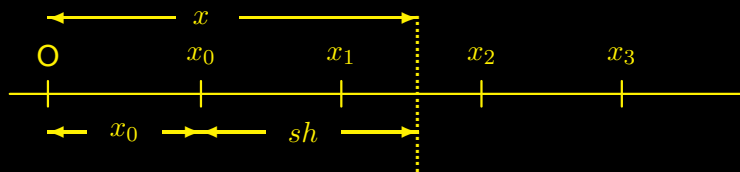


Figure 8.5: Moving the origin from  $O$  to  $x_0$ , new variable  $sh$ .  $x$  is shown at  $s = 1.5$ .  $\{x_0, x_1 \dots\}$  equally spaced distance  $h$  apart.

We see for example that if  $s = 1$  then  $x = x_0 + h = x_1$ , if  $s = 2$  then  $x = x_0 + 2h = x_2$  etc. Clearly from Eq(8.14), for each value of  $x$  there will be a unique value of  $s$  given by

$$s = \frac{x - x_0}{h}$$

From Eq(8.12) we have that:

$$p_n(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \dots \quad (8.15)$$

thus in order to make the change of variable we need to consider  $(x - x_k)$ . From Eq(8.14) we have

$$x = x_0 + sh$$

thus

$$(x - x_0) = sh$$

$$(x - x_1) = x - (x_0 + h) = (x - x_0) - h = h(s - 1)$$

$$(x - x_2) = x - (x_0 + 2h) = (x - x_0) - 2h = h(s - 2)$$

$$(x - x_3) = x - (x_0 + 3h) = (x - x_0) - 3h = h(s - 3)$$

.....

$$(x - x_k) = h(s - k)$$

Substituting these values and the values of the divided differences in terms of the forward differences from Eq (8.13) into the first four terms of Eq(8.15) gives the following:

$$p_n(s) = f(x_0) + (sh) \frac{\Delta f(x_0)}{h} + (sh)(h(s-1)) \frac{\Delta^2 f(x_0)}{2!h^2} + (sh)(h(s-1))(h(s-2)) \frac{\Delta^3 f(x_0)}{3!h^3} + \dots$$

Cancelling the powers of  $h$  in each term gives in general:

$$\begin{aligned} p_n(s) = & f(x_0) + s\Delta f(x_0) + \frac{s(s-1)}{2!} \Delta^2 f(x_0) \\ & + \frac{s(s-1)(s-2)}{3!} \Delta^3 f(x_0) + \dots \text{ up to the } \Delta^n f(x_0) \text{ term} \end{aligned} \quad (8.16)$$

The pattern is clear for all terms up to the  $\Delta^n f(x_0)$  term. It is also clear that if we treat  $\Delta$  simply as an algebraic symbol the expansion is symbolically the same as the binomial expansion of  $(1 + \Delta)^s f(x_0)$  up to the  $n^{\text{th}}$  power.

### Example 8.9 - Newton-Gregory

We will now use the data from table(8.5) and the Newton-Gregory formula in Eq(8.16) to find an approximation to  $\cos(1.25)$  using  $p_3(x)$ . Since in this example  $x_0 = 1$  and  $h = 0.5$ , from  $x = x_0 + sh$  the value of  $s$  is given by  $1.25 = 1 + 0.5 \times s \Rightarrow s = 0.5$

Thus

$$p_3 = f(x_0) + 0.5\Delta f(x_0) + (0.5)(-0.5)\Delta^2 f(x_0)/2! + (0.5)(-0.5)(-1.5)\Delta^3 f(x_0)/3!$$

Hence:

$$p_3 = 0.540 + 0.5(-0.469) - (0.25)(-0.018)/2! + (0.125)(0.120)/6 = \underline{0.310}$$

To three decimal places  $\cos 1.25 = 0.315$  thus we have an error of about 0.005. Thus the method is clearly quite good as the original data was only accurate to three decimal places.

## Chapter 9

# Numerical Integration

### 9.1 Introduction

In practical problems the integration of a given function is usually impossible in terms of simple elementary functions. The aim of this section is to formulate a numerical approximation to the definite integral:

$$\int_{x=a}^{x=b} f(x) dx$$

whilst at the same time estimating a bound on the incurred error. In chapter 3 we found ways of approximating  $f(x)$  with a polynomial  $p(x)$  and wrote:

$$f(x) = p(x) + \text{error}(x)$$

As it is always possible to integrate a polynomial, the obvious way forward is to write:

$$\int_{x=a}^{x=b} f(x) dx = \int_{x=a}^{x=b} p(x) dx + \int_{x=a}^{x=b} \text{error}(x) dx$$

We also saw that it was quite often possible to find a bound on the error. If  $|\text{error}(x)| \leq B(x)$  then using a result <sup>1</sup> from the theory of integration we have:

$$\left| \int_{x=a}^{x=b} \text{error}(x) dx \right| \leq \int_{x=a}^{x=b} |\text{error}(x)| dx \leq \int_{x=a}^{x=b} B(x) dx$$

As we have seen,  $B(x)$  is also a polynomial, thus it will be possible to evaluate the final integral in this expression.

As usual the overall game plan is quite straightforward and the difficulty, or devil, lies in the detail. The first method we consider is the Trapezoidal method where the polynomial  $p(x)$  is linear.

### 9.2 Trapezoidal method - linear approximation

Because this method is very simple we can first consider it geometrically even though this does not deliver an estimate for the error.

---

<sup>1</sup>  $\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$



### 9.2.1 Geometrical approach

The definite integral has an interpretation involving the area under a curve as illustrated in fig 9.1. The area of the trapezium  $A$  is an approximation to the total area under the curve from  $x = a$  to  $x = b$ , thus the **trapezoidal rule** is:

$$\int_{x=a}^{x=b} f(x) dx \approx \frac{h}{2} \{f(a) + f(b)\} \quad (9.1)$$

The error involved in making this approximation is given by the area indicated by  $E$  in

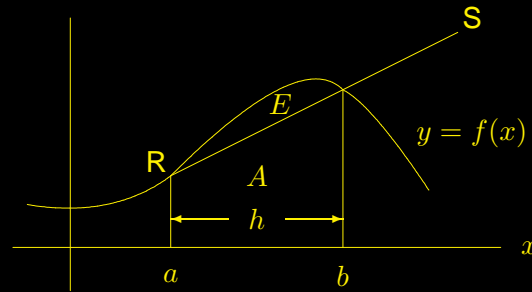


Figure 9.1: Area under the curve approximated by the trapezium  $A$

fig 9.1, however we have no means, using just the diagram, of estimating  $E$ .

### 9.2.2 Analytical approach: rule + error term

Analytically we construct a linear approximation to  $f(x)$  using the two points  $(a, f(a))$  and  $(b, f(b))$ . This gives the straight line approximation (line  $RS$  in fig 9.1 )

$$p(x) = f(a) + (x - a) \left( \frac{f(b) - f(a)}{h} \right) \quad (9.2)$$

Using now the expression for the error from Chapter 3 Eq(8.9) we can write:

$$f(x) = f(a) + (x - a) \left( \frac{f(b) - f(a)}{h} \right) + \frac{(x - a)(x - b)}{2} f''(t) \quad (9.3)$$

where  $t$  is some unknown value in the interval  $(a, b)$ . Considering the approximation using  $p(x)$  from Eq(9.2) we obtain:

$$\begin{aligned} \int_{x=a}^{x=b} f(x) dx &\approx \int_{x=a}^{x=b} p(x) dx = \int_{x=a}^{x=b} \left\{ f(a) + (x - a) \left( \frac{f(b) - f(a)}{h} \right) \right\} dx \\ &= f(a) \int_{x=a}^{x=b} dx + \left( \frac{f(b) - f(a)}{h} \right) \int_{x=a}^{x=b} (x - a) dx = (b - a)f(a) + \left( \frac{f(b) - f(a)}{h} \right) \left[ \frac{(x - a)^2}{2} \right]_a^b \end{aligned}$$

Using the fact that  $h = (b - a)$  we obtain:

$$\int_{x=a}^{x=b} f(x) dx \approx hf(a) + \left( \frac{f(b) - f(a)}{h} \right) \times \frac{h^2}{2} = \frac{h}{2} \{f(a) + f(b)\}$$

Which is the same formula for the trapezoidal rule obtained in Eq(9.1).

We now continue to find an estimate of the error using the remainder term in Eq(9.3). If we are able to find a bound  $M$  on  $|f''(t)|$  in the interval  $[a, b]$  and bearing in mind <sup>2</sup> that  $a \leq x \leq b$  then we can write:

$$\left| \int_a^b \text{error} dx \right| \leq \int_a^b |\text{error}| dx = \int_a^b \left| \frac{(x-a)(x-b)}{2} \right| |f''(t)| dx \leq \int_a^b \frac{(x-a)(b-x)}{2} M dx$$

Integrating by parts and replacing  $b - a$  with  $h$  gives:

$$\begin{aligned} \int_a^b \frac{(x-a)(b-x)}{2} M dx &= \frac{M}{2} \left[ -\frac{(x-a)(b-x)^2}{2} \right]_a^b + \frac{M}{2} \int_a^b (b-x)^2 dx \\ &= 0 + \frac{M}{4} \left[ \frac{(b-x)^3}{-3} \right]_a^b = \frac{M}{12} (b-a)^3 = \frac{Mh^3}{12} \end{aligned}$$

Thus the full statement of the **Trapezoidal rule with error bound** is given by:

$$\boxed{\int_a^b f(x) dx = \frac{h}{2} \{f(b) + f(a)\} \pm \frac{Mh^3}{12} \quad \text{where} \quad |f''(t)| \leq M, \quad t \in [a, b]}$$

### Example 9.1

Use the trapezoidal method to obtain an approximation to  $\int_0^1 e^x dx$  and obtain an error bound as above.

$$\int_0^1 e^x dx \approx \frac{1}{2} (e^0 + e^1) = \underline{1.859}$$

Since  $|f''(t)| = |e^t|$ , which is clearly increasing on the interval  $[0, 1]$ , its maximum value occurs at  $t = 1$  and hence a suitable value of  $M$  is  $e^1$ .

In this example  $h = 1$ , thus the error bound is given by  $\frac{Mh^3}{12} = \frac{e}{12} \approx 0.23$ . Thus we can say that:

$$\int_0^1 e^x dx = 1.859 \pm 0.23$$

A simple calculation using the exact value of the integral shows that correct to two decimal places the error is 0.14, which is well within the predicted error bound.

<sup>2</sup> $a \leq x \leq b$  implies that  $|x - b| = (b - x)$  and  $|x - a| = (x - a)$

### 9.2.3 Composite trapezoidal rule

Clearly in the above example the error was quite large, which suggests we should look at an alternative way of applying the trapezoidal rule. Since the error clearly depends on  $h$  it would seem reasonable to look at the effect of splitting the interval  $[a, b]$  into  $n$  equal divisions or subintervals. Applying the trapezoidal rule to each subinterval means that  $h$  is reduced by a factor of  $n$  and is given by  $h = \frac{(b-a)}{n}$ . To obtain the complete integral the results from estimating the integrals over each subinterval are added. Split

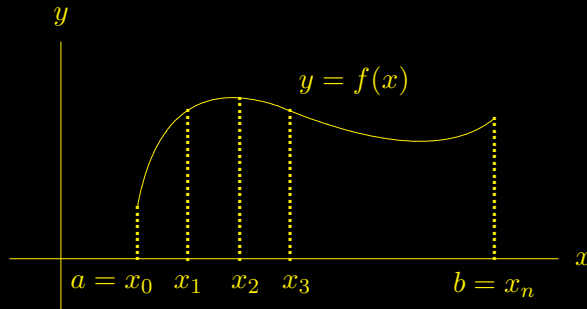


Figure 9.2: Composite trapezoidal rule; equal subintervals width  $h$

the interval  $[a, b]$  into  $n$  equal divisions, as in fig 9.2

Thus:

$$\int_a^b f(x) dx = \int_{x_0}^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx + \int_{x_2}^{x_3} f(x) dx + \dots + \int_{x_{n-1}}^{x_n} f(x) dx$$

Applying the trapezoidal rule to each interval gives:

$$\int_a^b f(x) dx \approx \frac{h}{2} \left\{ \{f(x_0) + f(x_1)\} + \{f(x_1) + f(x_2)\} + \{f(x_2) + f(x_3)\} + \dots + \{f(x_{n-1}) + f(x_n)\} \right\} \quad (9.4)$$

As we see  $f(x_0)$  and  $f(x_n)$  appear only once in Eq(9.4) whereas every other term appears twice, this gives the **composite trapezoidal rule** as:

$$\int_a^b f(x) dx \approx \frac{h}{2} \left\{ f(x_0) + f(x_n) + 2 \sum_{i=1}^{n-1} f(x_i) \right\} \quad (9.5)$$

In the above the trapezoidal rule is applied, in total,  $n$  times in evaluating the complete integral and at each application we incur an error. Although the error bound  $\left(\frac{Mh^3}{12}\right)$  is reduced, due to the use of a smaller  $h$ , it is incurred  $n$  times. Thus an error bound for the composite rule is given by:

$$\text{error bound} = n \times \left(\frac{Mh^3}{12}\right) = \frac{(b-a)}{h} \times \frac{Mh^3}{12} = (b-a) \frac{Mh^2}{12} \quad (9.6)$$

where we have substituted for  $n$  using  $n = \frac{(b-a)}{h}$ .

Thus the full statement for the composite trapezoidal rule is given as:

$$\int_a^b f(x) dx = \frac{h}{2} \left\{ f(x_0) + f(x_n) + 2 \sum_{i=1}^{n-1} f(x_i) \right\} \pm (b-a) \frac{Mh^2}{12} \quad \text{where } |f''(t)| \leq M, \quad t \in [a, b]$$

We see that as we decrease  $h$ , that is to say increase the number of subintervals into which  $[a, b]$  is divided, the error decreased according to the power of  $h^2$ . It is therefore technically possible to obtain the integral to any degree of accuracy provide that  $|f''(t)|$  is bounded on  $[a, b]$ .

### Example 9.2

If we return to the example  $\int_0^1 e^x dx$  and halve the interval then the value of  $h$  is halved and the error bound, which depends on  $h^2$ , will be reduced by a factor of four. Carrying out the calculations;

$$\int_0^1 e^x dx \approx \frac{h}{2} \{f(x_0) + f(x_2) + 2f(x_1)\} = \frac{0.5}{2} \{e^0 + e^1 + 2e^{0.5}\} = 1.7539$$

Using the exact value of the integral the error has now been reduced to approximately 0.035.

From Eq(9.6) the error bound is now given by:

$$\text{error bound} = (1-0) \times \frac{e^1 \times (0.5)^2}{12} = 0.057$$

where  $h = 0.5$  and the same value of  $M$  has been used as in Eg(9.1) Again as expected we see that the actual error is within the error bound.

### 9.2.4 Degree of accuracy

Following the same idea as in Chapter 2 Eg(??) if we are told that  $x$  equals  $a$  correct to  $N$  decimal places then

$$x = a \pm 0.5 \times 10^{-N}$$

Thus if in the above example we wish to ensure that we obtain a value of the integral to within at least 4 decimal places we can use a value of  $h$  given by:

$$(b-a) \frac{Mh^2}{12} \leq 0.5 \times 10^{-4}$$

Substituting in the values of  $M$ ,  $a$  and  $b$  for the above example gives:

$$\frac{e^1 \times h^2}{12} \leq 0.5 \times 10^{-4} \Rightarrow h^2 \leq 0.00022 \Rightarrow h \leq 0.014$$

Since  $nh = (b - a) = 1$  we are forced to take  $h$  such that we obtain an integer value of  $n$ . A suitable value of  $h$  is therefore  $h = 0.0125$  which means that we have  $n = 80$ . That is to say we will apply the trapezoidal rule with 80 subintervals. Doing this gives:

$$\int_0^1 e^x dx \approx 1.71830$$

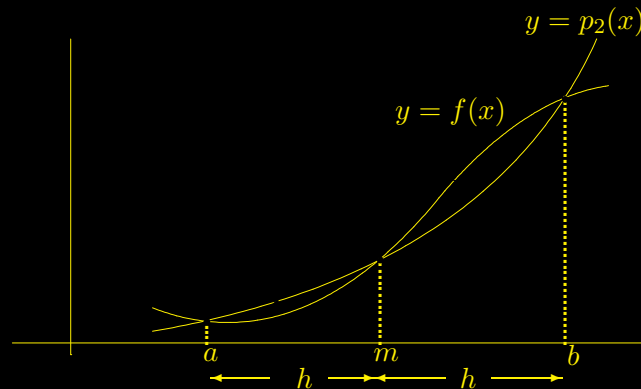
which comparing it with the exact value (correct to 5 decimal places) of 1.71828 gives an error of 0.00002, which as expected is within the tolerance for the accuracy required for 4 decimal places.

### 9.3 Simpson's rule - quadratic approximation

As we saw the trapezoidal rule used two points on the curve to formulate an approximation to the integral. Simpson's rule takes this one step further and uses three points on the curve. Given three points it is possible to construct a quadratic through the points and hence by integrating this quadratic it is possible to approximate the integral. Thus in general:

$$\int_a^b f(x) dx = \int_a^b p_2(x) dx + \int_a^b \text{error}(x) dx \quad (9.7)$$

where, with ref to fig 9.7,  $p_2(x)$  is the quadratic approximation to  $f(x)$  and the error function is given as in Chapter 3, Eq(8.9)



In terms of fig 9.3 we have <sup>3</sup> :

$$\int_a^b f(x) dx \approx \int_a^b p_2(x) dx = A_1 + A_2 \quad \text{and} \quad \int_a^b \text{error}(x) dx = -E_1 + E_2$$

<sup>3</sup>Since  $p_2(x) \geq f(x)$  on the interval  $[a, m]$  it follows that  $\int_a^m \{f(x) - p_2(x)\} dx$  is negative and hence its value is  $-E_1$ , where  $E_1$  is the positive area between the curves.

Replacing  $p_2(x)$  directly using the method of Lagrange or the method of divided differences proves to be quite tedious, however by considering the problem centred about the origin the task of obtaining a basic integration formula can be considerably simplified.

### 9.3.1 Basic formula

Considering the simplified problem, with the midpoint of the range of integration at the origin, we have the following diagram:

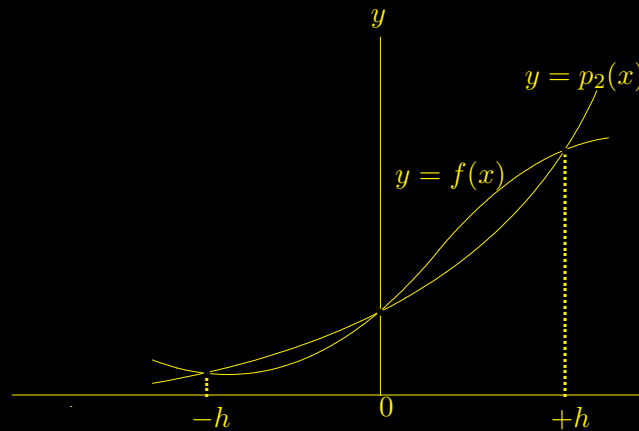


Figure 9.4:

Setting  $p_2(x) = \alpha x^2 + \beta x + \gamma$  gives:

$$\int_{-h}^h f(x) dx \approx \int_{-h}^h p_2(x) dx = \int_{-h}^h (\alpha x^2 + \beta x + \gamma) dx = \frac{h}{3}(2\alpha h^2 + 6\gamma) \quad (9.8)$$

Since the  $p_2(x)$  passes through the points A, B and C (ref fig 9.4) we have:

$$f(0) = p_2(0) = \gamma \quad f(h) = p_2(h) = \alpha h^2 + \beta h + \gamma \quad (i) \quad f(-h) = p_2(-h) = \alpha h^2 - \beta h + \gamma \quad (ii)$$

Using  $\gamma = f(0)$  and adding equations (i) and (ii) gives:

$$2\alpha h^2 = f(h) + f(-h) - 2f(0) \quad (9.9)$$

Substituting from Eq(9.9) into Eq(9.8) for  $2\alpha h^2$  and setting  $\gamma = f(0)$  gives:

$$\int_{-h}^h f(x) dx \approx \frac{h}{3}\{f(h) + f(-h) - 2f(0) + 6f(0)\} = \frac{h}{3}\{f(-h) + 4f(0) + f(h)\}$$

Transforming this result back to the general interval  $[a, b]$  gives

**Simpson's rule:**

$$\boxed{\int_a^b f(x) dx \approx \frac{h}{3}\{f(a) + 4f(a+h) + f(b)\} \quad \text{where} \quad h = \frac{(b-a)}{2}}$$

### 9.3.2 Error term

Using a different approach to the one used to calculate an error bound for the trapezoidal rule an error bound for Simpson's rule can be shown to be:

$$|\text{Simpson's error}| \leq \frac{h^5}{90} M \quad \text{where} \quad |f^{iv}(t)| \leq M, \quad t \in [a, b]$$

#### Example 9.3

Use Simpson's rule to approximate  $\int_0^1 e^x dx$

Splitting the interval into two parts gives  $h = 0.5$ , hence Simpson's rule gives:

$$\int_0^1 e^x dx \approx \frac{0.5}{3} \{f(0) + 4f(0.5) + f(1)\} = \frac{0.5}{3} \{e^0 + 4e^{0.5} + e^1\} = \underline{1.718861}$$

The error bound can now be calculated.

$$|f^{iv}(t)| = e^t \leq e^1 \quad \text{for} \quad t \in [0, 1] \quad \Rightarrow \quad M = e^1$$

$$\text{Thus the error bound} = \frac{h^5}{90} M = \frac{(0.5)^5}{90} \times e^1 = \underline{0.000943}$$

Hence:

$$\int_0^1 e^x dx = \underline{1.718861 \pm 0.000943}$$

As we have seen the exact value to five decimal places is 1.718281, thus the actual error is approximately 0.00058, which as expected is within the accuracy of the error bound.

A fair comparison of this method with the trapezoidal method would be to compare this example with Eg(9.2) which also involved two subintervals and hence the same number of evaluations of  $f(x)$ . In Eg(9.2) the error bound was 0.057 which is much greater than the 0.000943 obtained above. Hence we conclude that Simpson's rule is much more accurate than the trapezoidal rule.

### 9.3.3 Composite Simpson's rule

As with the trapezoidal rule the accuracy of the method can be increased by increasing the number of subintervals into which the interval  $[a, b]$  is divided. It should however be noted at the outset that since a single application of Simpson's rule requires two subintervals any number of applications of Simpson's rule requires an even number of subintervals. Thus if we divide the interval  $[a, b]$  into  $2n$  equal subintervals of width  $h$ , denoted by  $\{x_0, x_1, \dots, x_{2n}\}$  where  $x_0 = a$  and  $x_{2n} = b$  then repeated application of Simpson's rule  $n$  times gives:

$$\begin{aligned} \int_a^b f(x) dx \approx \frac{h}{3} \{ & f(x_0) + 4f(x_1) + f(x_2) \\ & + f(x_2) + 4f(x_3) + f(x_4) \\ & \dots + f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n}) \} \end{aligned} \quad (9.10)$$

From Eq(9.10) it is clear that apart from  $f(x_0)$  and  $f(x_{2n})$  all the even ordinates are multiplied by 2 and all the odd ordinates are multiplied by 4, thus the **composite Simpson's rule** is given by:

$$\int_a^b f(x) dx \approx \frac{h}{3} \{f(x_0) + f(x_{2n}) + 4 \sum_{k=1}^n f(x_{2k-1}) + 2 \sum_{k=1}^{n-1} f(x_{2k})\} \quad (9.11)$$

If the rule is applied  $n$  times then an error is incurred each time; as with the trapezoidal rule the error bound will be multiplied by  $n$ . Thus bearing in mind that  $2nh = (b - a)$  the error bound for the composite rule is given by:

$$\text{error bound} = n \times \left( \frac{h^5}{90} M \right) = \frac{(b-a)}{2h} \times \frac{h^5}{90} M = \frac{(b-a)}{180} h^4 M \quad \text{where } |f^{(iv)}(t)| \leq M \quad t \in [a, b]$$

Thus the full statement for the composite Simpson's rule is:

$$\int_a^b f(x) dx = \frac{h}{3} \{f(x_0) + f(x_{2n}) + 4 \sum_{k=1}^n f(x_{2k-1}) + 2 \sum_{k=1}^{n-1} f(x_{2k})\} \pm \frac{(b-a)}{180} h^4 M$$

Where  $M$  is an upper bound for  $f^{(iv)}(t)$  on the interval  $a \leq t \leq b$

#### Example 9.4

Use two applications of Simpson's rule (take  $n = 2$ ) to approximate  $\int_0^1 e^x dx$

With  $n = 2$  the interval is split into four parts giving  $h = 0.25$ , hence Simpson's rule gives:

$$\begin{aligned} \int_0^1 e^x dx &\approx \frac{0.25}{3} \{f(0) + f(1) + 4f(0.25) + 4f(0.75) + 2f(0.5)\} \\ &= \frac{0.25}{3} \{e^0 + e^1 + 4e^{0.25} + 4e^{0.75} + 2e^{0.5}\} = \underline{1.718318} \end{aligned}$$

The error bound can now be calculated.

$$|f^{(iv)}(t)| = e^t \leq e^1 \quad \text{for } t \in [0, 1] \quad \Rightarrow \quad M = e^1$$

$$\text{Thus the error bound} = (1 - 0) \frac{h^4}{180} M = \frac{(0.25)^4}{180} \times e^1 = \underline{0.000059}$$

Hence:

$$\int_0^1 e^x dx = \underline{1.718318 \pm 0.000059}$$

As we have seen the exact value to five decimal places is 1.718281, thus the actual error is approximately 0.000037, which as expected is within the accuracy of the error bound.



### 9.3.4 Degree of accuracy

As in section(9.2.4) we can calculate how many divisions to take in order to guarantee an accuracy of at least a certain number of decimal places. To achieve an answer accurate to at least four decimal places in Eg(9.4) above, we require:

$$\frac{1}{180}h^4 \times e^1 \leq 0.5 \times 10^{-4} \quad \Rightarrow \quad h \leq 0.24$$

The subinterval length  $h$  must be chosen such that  $n$ , the number of applications of Simpson's rule is an integer. That is to say  $h$  must be such that  $n = \frac{(b-a)}{2h}$  is an integer. Our best choice here would be  $h = 0.125$  which gives eight subdivisions and requires Simpson's rule to be applied four times. One might be tempted to use  $n = 3$  which implies that  $h = 1/6$ . Although this value of  $h$  is less than 0.24 it cannot be expressed exactly as a finite decimal; this may lead to other errors.

# Chapter 10

## Simultaneous Equations

### 10.1 Introduction

This section looks very briefly at the solution of sets of linear simultaneous equations. We look at the standard representation of the equations in terms of matrices and then formulate the process of elimination in terms of elementary row operations for a matrix. The error involved in the standard method of eliminations due to round off error in the computer is addressed using the method of partial pivoting. Additionally the problem of inconsistency, that is to say when the equations do not possess a solution, and the problem of there existing more than one solution are also addressed.

### 10.2 Matrix representation - Row reduction

The theory of matrices is vast, here we simply use the matrix notation to simplify our working. The following problem is considered in two ways, the left hand column uses a traditional representation and the right hand column a matrix representation. This simple example will enable us to see the type of operations we can make on the rows of a matrix in order to obtain the solution to the equations.

Equations	Matrices
$\begin{aligned}x_1 + 2x_2 &= 1 & (i) \\3x_1 + 4x_2 &= 2 & (ii)\end{aligned}$	$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ <p><i>write in augmented form</i></p> $\begin{pmatrix} 1 & 2 & 1 \\ 3 & 4 & 2 \end{pmatrix}$

Equations	Matrices
<p><i>fix (i) : replace (ii) with (ii) - 3 × (i)</i></p> $\begin{array}{rcl} x_1 + 2x_2 & = & 1 \quad (iii) \\ 0 - 2x_2 & = & -1 \quad (iv) \end{array}$ <p><i>from (iv) <math>-2x_2 = -1 \Rightarrow x_2 = \frac{1}{2}</math></i></p>	<p><i>fix row 1: replace row 2 with row 2 - 3 × row 1</i></p> $\begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & -1 \end{pmatrix}$ <p><i>divide row 2 by -2</i></p> $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & \frac{1}{2} \end{pmatrix}$ <p><i>row 2 <math>\Rightarrow x_2 = \frac{1}{2}</math></i></p>
<b>Backward Substitution</b>	
<p><i>substituting back into (iii)</i></p> $x_2 = \frac{1}{2} \Rightarrow x_1 + 2x_2 = 1 \Rightarrow x_1 + 1 = 1 \Rightarrow x_1 = 0$	<p><i>row 1 <math>\Rightarrow x_1 + 2x_2 = 1</math></i></p> <p><i>substitute <math>x_2 = \frac{1}{2} \Rightarrow x_1 = 0</math></i></p>

Using the above example we formulate the following **elementary row operations**. The solution of the equations represented by the augmented matrix is unaltered if we:

(i) **interchange two rows**

*This is the same as writing the equations in a different order*

(ii) **multiply any row by a non-zero number**

*Since this includes multiplication by a fraction the rule includes division by a non-zero number*

(iii) **add to any row any multiple of any other row**

*If the multiple is negative this is the same as saying subtract from any row any multiple of any other row*

**Example 10.1**

Solve using the elementary row operations the equations:

$$\begin{array}{rcl} x_1 + 2x_2 + x_3 & = & 2 \\ 2x_1 - x_2 + 3x_3 & = & -2 \\ 4x_1 - 2x_2 + x_3 & = & 1 \end{array}$$

Writing this as an augmented matrix we get:

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & -1 & 3 & -2 \\ 4 & -2 & 1 & 1 \end{array} \right)$$

The vertical line isn't really necessary and has only been inserted to clearly distinguish the coefficients of the variables from the righthand side of the equations.

- **Step 1:** create a 1 in the (row 1-col 1) position

*This is referred to as the (row 1) pivot, as this is already in place, move to step 2*

- **Step 2:** create zeros below the (row 1) pivot.

*To achieve this we freeze (row 1) and take multiples of (row 1) from (row 2) and (row 3). This is equivalent to eliminating  $x_1$  from the second and third equations.*

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & -1 & 3 & -2 \\ 4 & -2 & 1 & 1 \end{array} \right) \rightarrow \begin{array}{l} \text{(row 1)} \\ \text{(row 2) - 2(row 1)} \\ \text{(row 3) - 4(row 1)} \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & -5 & 1 & -6 \\ 0 & -10 & -3 & -7 \end{array} \right)$$

- **Step 3:** create a 1 in the (row 2-col 2) position, that is to say create the row 2 pivot.

*To do this we freeze (row 1) and (row 3) and multiply (row 2) by  $-\frac{1}{5}$  or equivalently divide (row 2) by -5.*

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & -5 & 1 & -6 \\ 0 & -10 & -3 & -7 \end{array} \right) \rightarrow \begin{array}{l} \text{(row 1)} \\ -\frac{1}{5}(\text{row 2}) \\ \text{(row 3)} \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & -10 & -3 & -7 \end{array} \right)$$

- **Step 4:** create zeros below the (row 2) pivot.

*To achieve this we freeze (row 1) and (row 2) and add 10 times (row 2) to (row 3)*

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & -10 & -3 & -7 \end{array} \right) \rightarrow \begin{array}{l} \text{(row 1)} \\ \text{(row 2)} \\ \text{(row 3) + 10(row 2)} \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & 0 & -5 & 5 \end{array} \right)$$

- **Step 5:** create a 1 in the (row 3-col 3) position, that is to say create the row 3 pivot.

*This is achieved by freezing (row 1) and (row 2) and dividing (row 3) by -5.*

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & 0 & -5 & 5 \end{array} \right) \rightarrow \begin{array}{l} \text{(row 1)} \\ \text{(row 2)} \\ -\frac{1}{5}(\text{row 3}) \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & 0 & 1 & -1 \end{array} \right)$$

The matrix of coefficients, that is the  $3 \times 3$  matrix to the left of the vertical line, is in **upper triangular form** with 1's on the diagonal and zeros below it. In this form we can, using backward substitution, obtain the unique solutions to the equations.

### Backward substitution

- (row 3)  $\Rightarrow x_3 = -1$
- (row 2)  $\Rightarrow x_2 - x_3/5 = 6/5 \Rightarrow x_2 = x_3/5 + 6/5 = (-1)/5 + 6/5 = 1$
- (row 1)  $\Rightarrow x_1 + 2x_2 + x_3 = 2 \Rightarrow x_1 = -2x_2 - x_3 + 2 = -2(1) - (-1) + 2 = 1$

Thus the solution is given by  $x_1 = 1, x_2 = 1, x_3 = -1$ .

An equivalent method to backward substitution would be to continue to use the row operations to create zeros above the pivots as follows.

$$\begin{aligned} \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & -\frac{1}{5} & \frac{6}{5} \\ 0 & 0 & 1 & -1 \end{array} \right) &\xrightarrow{\substack{(row\ 1) - (row\ 3) \\ (row\ 2) + \frac{1}{5}(row\ 3) \\ (row\ 3)}} \left( \begin{array}{ccc|c} 1 & 2 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array} \right) \\ \left( \begin{array}{ccc|c} 1 & 2 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array} \right) &\xrightarrow{\substack{(row\ 1) - 2(row\ 2) \\ (row\ 2) \\ (row\ 3)}} \left( \begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array} \right) \end{aligned}$$

The three rows now give immediately that  $x_1 = 1, x_2 = 1$  and  $x_3 = -1$ .

The next example requires that we interchange rows in order to get the pivots on the leading diagonal, as in the above example.

### Example 10.2

Using elementary row operations solve:

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 2x_1 + 2x_2 - x_3 &= 5 \\ 3x_1 - x_2 + x_3 &= 1 \end{aligned}$$

- **Step 1:** Write in augmented form:

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 2x_1 + 2x_2 - x_3 &= 5 \\ 3x_1 - x_2 + x_3 &= 1 \end{aligned} \rightarrow \text{augmented matrix} \rightarrow \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 2 & 2 & -1 & 5 \\ 3 & -1 & 1 & 1 \end{array} \right)$$

- **Step 2:** As the (row 1) pivot is already in place {(row 1 - col 1) already =1} we create zeros below this pivot.

$$\left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 2 & 2 & -1 & 5 \\ 3 & -1 & 1 & 1 \end{array} \right) \xrightarrow{\substack{(row\ 1) \\ (row\ 2) - 2(row\ 1) \\ (row\ 3) - 3(row\ 1)}} \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & -3 & 3 \\ 0 & -4 & -2 & -2 \end{array} \right)$$

- **Step 3:** Since the term in (row 2 - col 2) is zero it is not possible to use this as a pivot for (row 2). We are not able to make this term equal 1 by multiplying or dividing (row 2) by a factor. The way round this is to swap (row 2) with one of the rows below (row 2) that has a non-zero term in (col 2). In this example there is only (row 3) to consider for a swap.

$$\left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & -3 & 3 \\ 0 & -4 & -2 & -2 \end{array} \right) \xrightarrow{\substack{\text{(row 1)} \\ \text{swap} \\ \text{(row 2) and (row 3)}}} \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & -4 & -2 & -2 \\ 0 & 0 & -3 & 3 \end{array} \right)$$

- **Step 4:** Since the matrix of coefficients is already in upper triangular form we can create the (row 2) and (row 3) pivots in one step.

$$\left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & -4 & -2 & -2 \\ 0 & 0 & -3 & 3 \end{array} \right) \xrightarrow{\substack{\text{(row 1)} \\ -\frac{1}{4}(\text{row 2}) \\ -\frac{1}{3}(\text{row 3})}} \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & -1 \end{array} \right)$$

- **Step 5:** Using backward substitution gives:

$$\begin{aligned} (\text{row 3}) &\Rightarrow x_3 = -1 \\ (\text{row 2}) &\Rightarrow x_2 + \frac{1}{2}x_3 = \frac{1}{2} \Rightarrow x_2 = -\frac{1}{2}x_3 + \frac{1}{2} = -\frac{1}{2}(-1) + \frac{1}{2} = 1 \\ (\text{row 1}) &\Rightarrow x_1 + x_2 + x_3 = 1 \Rightarrow x_1 = -x_2 - x_3 + 1 = -(1) - (-1) + 1 = 1 \end{aligned}$$

Thus the solution to the problem is  $x_1 = 1$ ,  $x_2 = 1$  and  $x_3 = -1$ .

### 10.3 Partial pivotal selection

There are two principal types of error that occur when solving the equations.

- Errors in the data. By which is meant that some of the coefficients may only be quoted to say two decimal places, this immediately gives a possible error  $\pm 0.005$ . This type of error is a problem when the solution to the equations are sensitive to small changes in the coefficients - this will always be a problem and is referred to as **ill conditioning**. Although this problem is not that common we must always be on our guard to identify when it may occur. (*not covered here*)
- Errors due to round off within the computer. The problem is that as we carry out the row reduction process round off errors may grow and affect the final solution so much as to render it useless. This type of error is referred to as **induced error**. The following method of partial pivoting is a very simple way of reducing the growth of such errors. Consider the following example where some computation has taken place leading us to the problem of solving a set of simultaneous equations. However due to the computing process and the effect of round off the entries of column 2 are stored as  $2 + \epsilon$ ,  $-1$  and  $-2 + \delta$  instead of 2, -1 and -2. This is a somewhat artificial example but it will demonstrate how the errors grow as we now attempt to row reduce the problem to solve the equations.

$$\left( \begin{array}{ccc|c} 1 & 2+\epsilon & 1 & 2 \\ 2 & -1 & 3 & -2 \\ 4 & -2+\delta & 1 & 1 \end{array} \right) \rightarrow \begin{array}{l} (row\ 1) \\ (row\ 2) - 2(row\ 1) \\ (row\ 3) - 4(row\ 1) \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & 2+\epsilon & 1 & 2 \\ 0 & -5-2\epsilon & 1 & -6 \\ 0 & -10-4\epsilon+\delta & -3 & -7 \end{array} \right)$$

As we can see after just two simple row operations the  $\epsilon$  error in (row 1) appears doubled in (row 2) and quadrupled in (row 3). The  $\delta$  error remains unchanged. The way to avoid this build up of error is to identify the row whose first term in column 1 has the largest modulus of all the terms in column 1 and then swap this row with (row 1). In the example the term of largest size in column 1 is the 4 in (row 3), thus we swap (row 1) and (row 3).

$$\left( \begin{array}{ccc|c} 1 & 2+\epsilon & 1 & 2 \\ 2 & -1 & 3 & -2 \\ 4 & -2+\delta & 1 & 1 \end{array} \right) \rightarrow \begin{array}{l} \text{swap with (row 3)} \\ (row\ 2) \\ \text{swap with (row 1)} \end{array} \rightarrow \left( \begin{array}{ccc|c} 4 & -2+\delta & 1 & 1 \\ 2 & -1 & 3 & -2 \\ 1 & 2+\epsilon & 1 & 2 \end{array} \right)$$

We now create the (row 1) pivot by multiplying (row 1) by  $\frac{1}{4}$ .

$$\left( \begin{array}{ccc|c} 4 & -2+\delta & 1 & 1 \\ 2 & -1 & 3 & -2 \\ 1 & 2+\epsilon & 1 & 2 \end{array} \right) \rightarrow \begin{array}{l} \frac{1}{4}(row\ 1) \\ (row\ 2) \\ (row\ 3) \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & -\frac{1}{2}+\frac{\delta}{4} & \frac{1}{4} & \frac{1}{4} \\ 2 & -1 & 3 & -2 \\ 1 & 2+\epsilon & 1 & 2 \end{array} \right)$$

Finally creating zeros below the (row 1) pivot gives:

$$\left( \begin{array}{ccc|c} 1 & -\frac{1}{2}+\frac{\delta}{4} & \frac{1}{4} & \frac{1}{4} \\ 2 & -1 & 3 & -2 \\ 1 & 2+\epsilon & 1 & 2 \end{array} \right) \rightarrow \begin{array}{l} (row\ 1) \\ (row\ 2) - 2(row\ 1) \\ (row\ 3) - (row\ 1) \end{array} \rightarrow \left( \begin{array}{ccc|c} 1 & -\frac{1}{2}+\frac{\delta}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & -\frac{\delta}{2} & \frac{5}{2} & -\frac{5}{2} \\ 0 & \frac{5}{2}+\epsilon-\frac{\delta}{4} & \frac{3}{4} & \frac{7}{4} \end{array} \right)$$

We now see that the error of  $\delta$  in the first row second column position appears reduced by a factor of a quarter in (row 1) and (row 3) and a factor of a half in (row 2). In this case the  $\epsilon$  error remains unchanged. The above technique of swapping rows when constructing the pivots is called the method of **partial pivotal selection** or simply **partial pivoting**.

One final note, even if a set of equations is not ill conditioned when we start but we do nothing to minimise the effect of round off error the equations may become ill conditioned at some stage of the row reduction process and therefore lead to a final solution that is useless. This scenario is more likely to occur when we are solving large sets of simultaneous equations, thus even though modern computers calculate values to a great degree of accuracy a piece of software for carrying our row reductions would always use partial pivoting or some other method to reduce any induced error.

### Example 10.3

It is not easy to illustrate the effectiveness of partial pivoting if we carry out our calculations on a small set of equations to the degree of accuracy provided by a modern computer. In this sense the following example is artificial as we only carry out each calculation to three significant figures<sup>1</sup>, however it will demonstrate both the procedure and

<sup>1</sup>eg.  $2.35 \times 3.56 = 8.37$  correct to three significant figures; after each calculation the number is rounded to 3 figures to emulate a computer that can only store three figures for the principal part of a number

effectiveness of partial pivoting.

We consider the solution of the equations:

$$\begin{aligned}x - 2.23y + z &= 3 \\ -2.23x + 5y &= 3 \\ 5y - 2.23z &= 1\end{aligned}\tag{10.1}$$

These equations can be solved exactly using Derive, expressing the answers in terms of rational number as follows

$$x = \frac{38368700}{11210433} \approx 3.42 \quad y = \frac{106900}{50271} \approx 2.13 \quad z = \frac{48422900}{11210433} \approx 4.32$$

Now consider the solution of these equations, first without partial pivoting and then with, remembering that at all times we are working each calculation to three significant figures only.

### Without partial pivoting

$$\left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ -2.23 & 5 & 0 & 3 \\ 0 & 5 & -2.23 & 1 \end{array} \right) \rightarrow (\text{row } 2) + 2.23(\text{row } 1) \rightarrow \left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 0.03 & 2.23 & 9.69 \\ 0 & 5 & -2.23 & 1 \end{array} \right)$$

$$\left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 0.03 & 2.23 & 9.69 \\ 0 & 5 & -2.23 & 1 \end{array} \right) \rightarrow \frac{100}{3}(\text{row } 2) \rightarrow \left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 1 & 74.3 & 323 \\ 0 & 5 & -2.23 & 1 \end{array} \right)$$

$$\left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 1 & 74.3 & 323 \\ 0 & 5 & -2.23 & 1 \end{array} \right) \rightarrow (\text{row } 3) - 5(\text{row } 2) \rightarrow \left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 1 & 74.3 & 323 \\ 0 & 0 & -374 & -1620 \end{array} \right)$$

$$\left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 1 & 74.3 & 323 \\ 0 & 0 & -374 & -1620 \end{array} \right) \rightarrow -(\text{row } 3)/347 \rightarrow \left( \begin{array}{ccc|c} 1 & -2.23 & 1 & 3 \\ 0 & 1 & 74.3 & 323 \\ 0 & 0 & 1 & 4.33 \end{array} \right)$$

Applying backward substitution gives:

- (row 3)  $\Rightarrow z = 4.33$
- (row 2)  $\Rightarrow y + 74.3z = 323 \Rightarrow y = -(74.3)(4.33) + 323 = -322 + 323 = 1$
- (row 1)  $\Rightarrow x - 2.23y + z = 3 \Rightarrow x = (2.23)(1) - 4.33 + 3 = 0.9$

We see that  $x$  and  $y$  are nowhere near the correct answers given above. The process is now repeated but this time using partial pivoting.

### With partial pivoting

Swapping (row 2) and (row 1) to bring the -2.23 term in (row 2) into the (row 1) pivot



position, dividing the new (row 1) by -2.23 and then creating zeros below the (row 1) pivot gives:

$$\begin{pmatrix} 1 & -2.23 & 1 & | & 3 \\ -2.23 & 5 & 0 & | & 3 \\ 0 & 5 & -2.23 & | & 1 \end{pmatrix} \rightarrow \begin{pmatrix} -2.23 & 5 & 0 & | & 3 \\ 1 & -2.23 & 1 & | & 3 \\ 0 & 5 & -2.23 & | & 1 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & -2.24 & 0 & | & -1.35 \\ 1 & -2.23 & 1 & | & 3 \\ 0 & 5 & -2.23 & | & 1 \end{pmatrix} \rightarrow (\text{row } 2) - (\text{row } 1) \rightarrow \begin{pmatrix} 1 & -2.24 & 0 & | & -1.35 \\ 0 & 0.01 & 1 & | & 4.35 \\ 0 & 5 & -2.23 & | & 1 \end{pmatrix}$$

Since there is a term below the 0.01 in the (row 2) pivot position that is of larger magnitude than 0.01, namely the 5 in row 3, then we interchange (row 2) and (row 3). The new (row 2) is then divided by 5 to create the (row 2) pivot.

$$\rightarrow \begin{pmatrix} 1 & -2.24 & 0 & | & -1.35 \\ 0 & 5 & -2.23 & | & 1 \\ 0 & 0.01 & 1 & | & 4.35 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & -2.24 & 0 & | & -1.35 \\ 0 & 1 & -.446 & | & .2 \\ 0 & 0.01 & 1 & | & 4.35 \end{pmatrix}$$

Finally carrying out  $(\text{row } 3) - 0.01 \times (\text{row } 2)$  gives:

$$\rightarrow \begin{pmatrix} 1 & -2.24 & 0 & | & -1.35 \\ 0 & 1 & -.446 & | & .2 \\ 0 & 0 & 1 & | & 4.35 \end{pmatrix}$$

Applying backward substitution:

- (row 3)  $\Rightarrow z = \underline{4.35}$
- (row 2)  $\Rightarrow y = 0.446z + 0.2 = (0.446)(4.35) + 0.2 = 1.94 + 0.2 = \underline{2.14}$
- (row 1)  $\Rightarrow x = 2.24y - 1.35 = (2.24)(2.14) - 1.35 = 4.79 - 1.35 = \underline{3.44}$

As expected there is no significant change in  $z$ , as even without partial pivoting the solution was not unreasonably inaccurate, bearing in mind that the calculations are only carried out to three figures. The values of  $x$  and  $y$  however can now be seen to be much closer to the exact answers. The results are summarised in table 10.1.

## 10.4 Existence and Uniqueness

Given a set of  $n$  equations in  $n$  unknowns then one of the following scenarios is possible:

- (i) There exists a unique solution as in the above example of 3 equations in 3 unknowns.
- (ii) There exists an infinite number of solutions.
- (iii) There does not exist a solution.

	exact	without pivoting	with pivoting
$x$	3.42	0.9	3.44
$y$	2.13	1.00	2.14
$z$	4.32	4.33	4.35

Table 10.1: Solution of Eq(10.1) using three significant figure arithmetic: exact; without partial pivoting; with partial pivoting.

In cases (i) and (ii) the equations are said to be **consistent** and in case (iii) the equations are said to be **inconsistent**.

To illustrate (ii) and (iii) consider the solution of the following set of equations, where  $k$  is some parameter.

$$\begin{aligned}
 x_1 + x_2 + x_3 &= -1 \\
 2x_1 - x_2 + 3x_3 &= 1 \\
 x_1 - 2x_2 + 2x_3 &= k
 \end{aligned}
 \tag{10.2}$$

Expressing in terms of the augmented matrix and row reducing without partial pivoting we obtain:

$$\begin{aligned}
 \left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 2 & -1 & 3 & 1 \\ 1 & -2 & 2 & k \end{array} \right) &\rightarrow \left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 0 & -3 & 1 & 3 \\ 0 & -3 & 1 & k+1 \end{array} \right) \rightarrow \left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 0 & 1 & -\frac{1}{3} & -1 \\ 0 & -3 & 1 & k+1 \end{array} \right) \\
 &\rightarrow \left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 0 & 1 & -\frac{1}{3} & -1 \\ 0 & 0 & 0 & k-2 \end{array} \right)
 \end{aligned}$$

The (row 3) pivot position is zero and since there are no other rows below this row it is not possible to do a row interchange to produce a nonzero value in row3-col3. Writing out the final row in terms of the variables gives:

$$0x_1 + 0x_2 + 0x_3 = (k - 2) \tag{10.3}$$

It is clear from Eq(10.3) that unless  $k = 2$  there are no values of  $x_1$ ,  $x_2$  and  $x_3$  that satisfy this equation. Summarising:

- If  $(k - 2) \neq 0$  then there does **not** exist a solution to Eq(10.2). In this case the equations are said to be inconsistent.
- If  $k = 2$  then (row 3) is satisfied no matter what values the variables take. Thus the solution is given entirely by (row 2) and (row 3).

Setting  $x_3 = t$  we can express the solution in the following parametric form:

$$(\text{row } 2) \Rightarrow x_2 - \frac{1}{3}x_3 = -1 \Rightarrow x_2 = \frac{1}{3}x_3 - 1 = \frac{1}{3}t - 1 = \frac{(t-3)}{3}$$

$$(\text{row } 1) \Rightarrow x_1 + x_2 + x_3 = -1 \Rightarrow x_1 = -x_2 - x_3 - 1 = -\frac{(t-3)}{3} - t - 1 = -\frac{4t}{3}$$

Thus in this case the solution is given by:

$$x_1 = -\frac{4t}{3} \quad x_2 = \frac{(t-3)}{3} \quad x_3 = t$$

for any value of the parameter  $t$ . That is to say we have an infinite number of solutions to the problem.

## 10.5 Geometrical representation

The solution of simultaneous equations, or more precisely linear simultaneous equations, can in general be discussed in terms of a subject called Linear Algebra. For equations in three variables the problem reduces to the study of planes and lines in three dimensional space. The equation of a plane in three dimensions is given by  $Ax_1 + Bx_2 + Cx_3 + D = 0$ , thus the solutions of three such equations are the common points of intersection of the three planes. If two distinct planes intersect then they do so in a line, if this line then pierces the third plane we have a unique solution. However if this line lies in the third plane then we have an infinite number of solutions. Finally it is easy to see that three planes may not have any points in common. This may occur when two of the planes are parallel or when the third plane is parallel to the line of intersection of the first two. In the example given by Eq(10.2) with  $k = 2$  the solution can be written as:

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -\frac{4t}{3} \\ \frac{t}{3} - 1 \\ t \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -\frac{4}{3} \\ \frac{1}{3} \\ 1 \end{pmatrix}$$

Geometrically this is the equation of a straight line in three dimensions passing through the point  $\begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$  and running parallel to the direction of  $\begin{pmatrix} -\frac{4}{3} \\ \frac{1}{3} \\ 1 \end{pmatrix}$

Alternatively without the use of vectors

$$x_1 = -\frac{4t}{3} \Rightarrow \frac{(x_1 - 0)}{(-\frac{4}{3})} = t \quad \text{and} \quad x_2 = \frac{t}{3} - 1 \Rightarrow \frac{(x_2 + 1)}{(\frac{1}{3})} = t \quad \text{and} \quad x_3 = t$$

Which gives the classical equation of a straight line in three dimensions:

$$\frac{(x_1 - 0)}{(-\frac{4}{3})} = \frac{(x_2 + 1)}{(\frac{1}{3})} = \frac{(x_3 - 0)}{1}$$

Finally we note that if  $k \neq 2$  then geometrically we have the case where each plane is parallel to the line of intersection of the other two. This fact is not immediately obvious and is not here developed any further.

# Index

- Asymptote, 1, 5
- Augmented matrix, 131
- Backward substitution, 132, 134
- Classification of Stationary Point, 3
- Coefficient of determination, 60
- Compound interest, 28
- Conditional probability, 80
- Convergence
  - Ratio test, 108
  - Taylor's polynomial, 107
- Correlation coefficient, 60, 61
- Curve fitting, 39
- Decreasing function, 1
- Divided Differences
  - Definition, 113
  - polynomial expansion, 113
  - table, 115
- Elementary row operations, 132
- Error term
  - Lagrange polynomials, 111
  - Taylor's polynomial, 104
- Excel
  - =CUMIPMT(...), 37
  - =CUMPRINC(...), 37
  - =FV(...), 31
  - =IPMT(...), 36
  - =LINEST, 54
  - =MINVERSE(A), 19
  - =MMULT(A,B), 18
  - =NPER(...), 33
  - =PMT(...), 32
  - =PPMT(...), 36
  - =PV(...), 32
  - =RATE(...), 34
  - =TREND, 57
  - financial functions, 31
  - iteration by method I, 11
  - iteration by method II, 14
  - money sign convention, 31
  - Solver, 66
- Existence of solution of  $f(x) = 0$ , 8
- Fixed point
  - definition, 10
  - method, 10
- Forward differences, 116
  - definition of  $\Delta^n$ , 116
  - general table, 118
  - Newton- Gregory formula, 119
- Function
  - decreasing, 1
  - increasing, 1
  - maximum and minimum, 3
  - stationary, 2
- Geometric progression
  - approximation to sum, 28
  - definition, 27
  - sum, 28
- Google, 92
- Ill-conditioning, 135
- Increasing function, 1
- Induced error, 135
- Inflection - point of, 4
- Initial guess, 10
- Interpolation
  - direct method, 109
  - Lagrange basis polynomials, 110
  - Lagrange's Method - example, 109
- Investment problem, 29
- Iterative method by rearrangement, 10
- Lagrange basis polynomials, 110

- Least squares
  - Excel =LINEST, 54
  - Excel =TREND, 57
  - exponential trendline (fit), 64
  - general fit, 65
  - linear fitting, 49
  - linear formula, 52
  - Logarithmic trendline (fit), 64
  - many variable example, 57
  - many variables, 52
  - polynomial trendline (fit), 63
  - power trendline (fit), 64
  - straight line example, 55, 58
  - trendlines and charts, 58, 65
- Linear approximation
  - general, 23
  - tangent line, 12, 21
  - tangent plane, 23
- Linear polynomial, 98
- Maclaurin's polynomial, 102
- Markov
  - solution of chain, 83
  - chain, 82, 83
  - matrix, 83
  - state vector, 84
- Matrix
  - column, 16
  - definition  $m \times n$ , 16
  - product, 16
  - square, 16
- Maximum and Minimum, 3
- Mean of data, 69
- Newton's Method
  - example, 13
  - Excel using 2 variable, 26
  - formula, 12
  - two variables, 23, 24
- Newton-Gregory formula, 119
- Order of Taylor polynomial, 101
- Outliers, 74
- PageRank, 86
- Partial derivative, 21
- Partial pivoting, 136
- Pivot, 133
- Point of Inflection, 4
- Polynomial
  - definition, 98
  - degree, 98
  - linear, 98
  - quadratic, 98
- Polynomial fit, 40
- Quadratic polynomial, 98
- Quartiles, 72
- Random surfer, 87
- Ratio Test, 108
- Rearrangement, 10
- Simpson's rule
  - basic formula, 126
  - composite, 128
  - error term, 128
- Simultaneous Equations
  - augmented matrix, 131
  - backward substitution, 132, 134
  - consistent, 139
  - existence and uniqueness, 139
  - inconsistent, 139
  - infinite number of solutions, 140
  - introduction, 131
  - partial pivoting, 136
  - pivot, 133
  - row operations, 132
- Simultaneous equations
  - linear, 15
- Spline
  - conditions, 46
  - cubic (*natural*), 45
  - linear, 42
- Standard deviation - estimate, 71
- Standard deviation of data, 69
- Standard error, 59
- Standard error of the mean, 71
- Stationary points
  - many variables, 52
  - single variable, 2
- Taylor polynomial
  - $n^{\text{th}}$  order, 101

- convergence, 107
- error term, 104
- example, 102
- linear, 100
- Maclaurin, 102
- order, 101
- quadratic, 100
- Taylor's theorem, 104
- Taylor's theorem, 104
- Trapezoidal method
  - basic formula, 121
  - composite, 124
  - error term, 123