Numerical Methods

Magdi Salah El-Azab

Numerical Methods. –5th ed.

رقم الإيداع بدار الكتب والوثائق القومية 16253/2004

© Magdi Salah El-Azab
All Rights Reserved.

Preface

Many problems important for applications in engineering and science can be neatly formulated as mathematical models but go no further for lake of analytic solutions. In fact most of the mathematical models found in technical disciplines and in the natural sciences are often very complicated. For this reason and due to the advent of computers, microcomputers and programmable calculators, it is typical to employ numerical techniques to find approximate solutions.

Numerical approximation techniques are given in the literature in a variety of ways. On one hand, in the traditional numerical analysis course much attention concerns with both the approximation methods and the mathematical analysis that produces them. On the other hand, considerable attention is focused on the numerical approximation methods for solving a given problem with less detailed deductions, proofs, and stability considerations.

As the title of this book indicates, the author elected to focus his attention on the use of numerical methods as a solution tool for the problem encountered in engineering and physical practice. It was immediately clear that this book is written to provide engineers, scientists, or graduate students working on a research project with compact and precise information about the numerical methods that can be used for the solution of complex problems encountered in engineering applications. The exposition would not be complete without mentioning some algorithmization, programming, and use of computer possibilities.

This book therefore presents primarily computation procedures that have been thoroughly verified by practical problem solution. It is written to provide college students with the basic skills needed for the use of numerical methods as well as for further course-work in numerical analysis. It also provides the reader with experience in the area of problem solving. It is the author's belief that these tasks should be undertaken by the reader in his or her first course in numerical analysis.

The book is divided into seven chapters. The first two chapters are devoted to approximate a function defined either analytically or by a discrete set of numerical data. Chapter three deals with the solution of nonlinear algebraic and transcendental equations. Chapters four and five deal with the solution of linear system of equations by direct and iterative methods, respectively. In chapter six some of the most famous methods for the numerical solution of ordinary differential equations are considered. Finally, the numerical solution of partial differential equations is dealt with in chapter seven.

There are numerous examples with solutions provided in this book. These are used to illustrate topics and procedures being introduced and reinforce definitions and rules for various arithmetic and algebraic operations. In addition, there are numerous, well graded, exercises usually for each chapter and sometimes for some sections of the book.

Mansoura September, 2024 **Magdi Salah El-Azab**

Contents

1. Numerical solution of nonlinear equations

1.1 Location of roots 4

1.1.1 Analytical approach 4

1.1.2 Graphical approach 5

1.2 Methods of solutions 7

1.2.1 Bisection method 8

1.2.2 False position method 16

1.2.3 The secant method 22

Exercise 1.1 26

1.2.4 Simple iteration method 29

1.2.5 Newton-Raphson method 39

Exercise 1.2 49

2. Numerical solution of linear systems of equations (Iterative methods)

2.1 Jacobi's iterative method 56

2.2 Gauss-Seidel iterative method 64

2.3 Successive over relaxation method (**SOR**) 69

Exercise 2.1 79

3. Interpolation by Polynomials

3.1. Lagrange interpolation 83

Exercise 3.1 88

3.2. Difference operators 90

3.3. General Newton' formulas of interpolation 98

3.4. Newton forward-difference formula 103

3.5. Newton backward-difference formula 108

3.6. Central-difference formulas 112

3.6.1. Gauss forward-difference formula 113

3.6.2. Gauss backward-difference formula 114

Exercise 3.2 119

4. Numerical solution of ordinary differential equations

4.1 Picard's method 124

4.2 Euler method 128

4.3 Modified Euler method 135

4.4 Taylor-series method 138

Exercise 4.1 145

4.5 Runge-Kutta methods 145

4.6 Multistep methods 165

Exercise 4.2 174

5. Numerical solution of partial differential equations

5.1 The finite difference method 182

5.2 The finite difference method for elliptic partial differential equations 186

Exercise 5.1 200

5.3 The finite difference method for parabolic partial differential equations 204

5.3.1 Explicit difference scheme 205

5.3.2 Implicit difference scheme (Crank-Nicolson scheme) 214

Exercise 5.2 218

5.4 The finite difference method for hyperbolic partial differential equations 220

Exercise 5.3 227

6. Curve Fitting

6.1 Principle of least squares 231

6.2 The least square method for polynomials of degree m 238

6.3 The least square method for general curves 243

Exercise 6.1 252

1. Numerical solution of nonlinear equations

This chapter will be concerned with the numerical solution of nonlinear equations and system of equations. First we shall restrict ourselves to single equations. One reason is that this case is more common than that of simultaneous nonlinear equations. An equally important reason, however, is that the solution of simultaneous nonlinear equations is a difficult problem. For the single nonlinear equations of interest here, we assume no solution in closed form can be found. Thus we must seek methods which lead to approximate solutions.

The first basic nonlinear equation in algebra is the quadratic equation

$$ax^2 + bx + c = 0,$$

and we all know that the solution of this equation is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

For cubic and quartic equations, formulas exist but are so complex as to be rarely used; for higher-degree equations it is difficult to exhibit the solution in explicit form. So it is typical to employ numerical methods for the solution of such equations. Before going on we introduce the following definitions.

Definition 1.1 (Algebraic equation) An $algebraic\ equation$ is an equation of the polynomial form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0,$$

where $a_n, a_{n-1}, \cdots, a_1$ and a_0 are real numbers.

For example; the following equations are algebraic equations

$$x^4 + 7x^2 - 5x + 10 = 0,$$
$$x^2 - 4 = 0.$$

Definition 1.2 (Transcendental Equation) Any equation, which cannot be represented by a polynomial form is called *transcendental equation.*

For example; the equations

$$\sin x + \cos 2x = 0, \qquad \ln x - 4x = 0,$$

are transcendental equations.

A nonlinear equation is either algebraic or transcendental. For an algebraic equation all roots, both real and complex, are to be determined, but for a transcendental equation the search is usually restricted to real roots. The purpose of this chapter is to present some numerical methods to solve the general nonlinear equations problem.

One of the most frequently occurring problems in scientific work is to solve the equation

$$f(x) = 0, \tag{1}$$

whether this equation is an algebraic or a transcendental equation.

Definition 1.3 (Roots of an equation) The values of x that satisfy equation (1) are called the ***roots*** of this equation. These values are also called the $\boldsymbol{zeros}$ of the function $f(x)$.

We assume in the foregoing that the exact root of equation (1) will be denoted by α and the approximate solution x_{app} will be approximated by a sequence of estimates $x_0,\ x_1, \cdots$, with

$$x_{app} = \lim_{n \to \infty} x_n$$

The numerical analysis provides a mean where a solution may be found, or, at least approximated as closely as desired. Many of these numerical procedures follow a scheme that may be thought of as providing a series of successive approximations, each more precise than the previous one, so that enough repetitions of the procedure eventually give an approximation which differs from the true value by less than some arbitrary error tolerance. It was immediately clear that, all the numerical methods require obtaining starting values before the method can begin. So that the numerical solution of equation (1) is carried out in two steps:

- **Step 1:** ***Location of roots***

In practical problems, a priori knowledge of the nature, number and approximate location of the roots is often required to ensure the convergence of the iterative methods we now discuss. That is to say, in this step the required roots of equation (1) are located in some intervals so that each root is restricted to some interval $(a,\ b),\ a, b \in \mathcal{R}$.

- **Step 2:** ***Using iterative methods of solution***

In this step, we introduce the following elementary iterative methods for finding a solution of equation (1):

1. Bisection method,
2. False position method,
3. Secant method.
4. Simple iteration method,
5. Newton-Raphson method,

Most practical computer routines implementing these methods will quite when:

- either x_{app} is obtained to a pre-determined tolerance or
- $|f(x_{app})|$ is so small that to machine accuracy it registers as zero.
- Some routines also include a maximum number of iterations as a safety measure.

1.1 Location of roots

Note that all the methods we will use require an initial estimate of the root we are computing. It often requires as much thought and effort to get a good starting value as it does to refine it to acceptable accuracy. Sometimes one's knowledge of the physical problem will suggest a starting value. When this is not available, one normally finds starting values by initial trial-and-error computations (analytical approach), or by making a rough graph of the function (graphical approach). We now discuss the usual approaches to guess the initial approximations.

1.1.1 Analytical approach

In this approach the trial-and-error computations are used to locate the intervals that contain the required roots. By the use of the following theorem we can locate the roots of the equation $f(x) = 0$.

Theorem 1.1 For any continuous real function $f(x): R \to R$ in the period $(a,\ b)$, $a, b \in R$, if $f(a)$ has opposite sign with respect to $f(b)$, i.e., $f(a)f(b) < 0$, then there is ***at least one root*** in this interval (In fact there is an odd number of roots).

The above theorem is illustrated by the following figure.

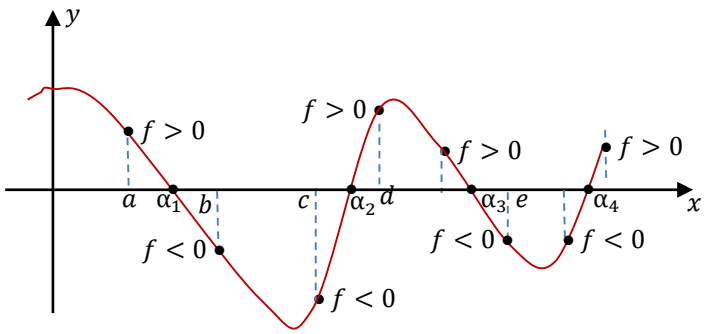


Fig. 1.1. Location of the roots of the equation $f(x) = 0$

Due to the opposite signs of the sketched function through the different intervals, we have

$$f(a)f(b) < 0 \quad \Rightarrow \quad (one\ root)$$

$$f(c)f(d) < 0 \quad \Rightarrow \quad (one\ root)$$

$$f(a)f(e) < 0 \quad \Rightarrow \quad (three\ roots)$$

However, the opposite signs of the sketched function imply odd number of the roots of the function.

1.1.2 Graphical approach

Another approach to locate the zeros of a function $f(x)$ is to draw its graph $y = f(x)$ and observe where it crosses the x axis. (i.e., for which $f(x) = 0$). On occasions it is helpful to obtain a graphical solution of this type in a slightly different manner by writing $f(x)$ in the form $f(x) = g(x) - h(x)$, when the zeros of $f(x)$ will correspond to the equation $g(x) = h(x)$. To see how this may help, suppose that we seek the roots of the transcendental equation

$$\sin x - \cosh x + 1 = 0,$$

where $f(x) = \sin x - \cosh x + 1$. Then by writing

$$f(x) = g(x) - h(x),$$

with $g(x) = \sin x$ and $h(x) = \cosh x - 1$, the zeros of $f(x) = 0$ will correspond to those values of x for which the graphs of $y = g(x)$ and $y = h(x)$ intersect, as shown in Fig. 1.2. Familiarity with the two graphs involved makes it easier to sketch them and to appreciate how many zeros there are likely to be. The two zeros are seen to be $x = 0$ (exact) and $x = 1.3$.

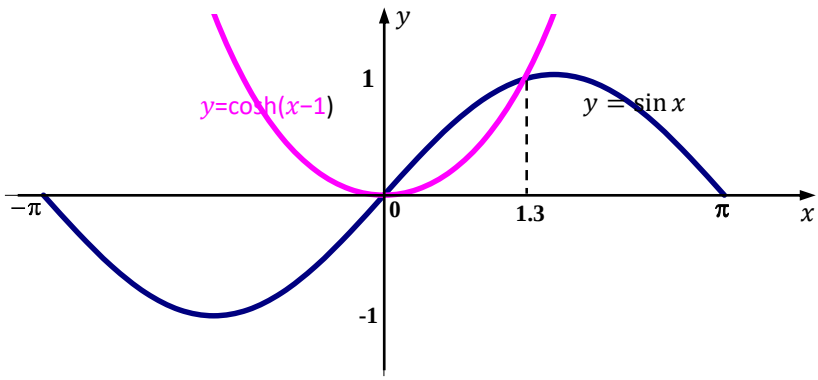


Fig. 1.2. Roots of $\sin x - \cosh x + 1 = 0$

Example 1 Determine approximate values of the roots of the equation

$$\sin x - x + 0.5 = 0$$

Solution We first sketch the graphs of the two curves $y = \sin x$ and $y = x - 0.5$ (see Fig. 1.3). Since the roots of the given equation represent the points of intersections of these two curves, and we know that $|\sin x| \leq 1$, then we have $x - 0.5 \leq 1$. This gives

$$-1 \leq x - 0.5 \leq 1,$$

or

$$-0.5 \leq x - 0.5 \leq 1.5$$

So, we are only interested in this interval $(-0.5, 1.5)$.

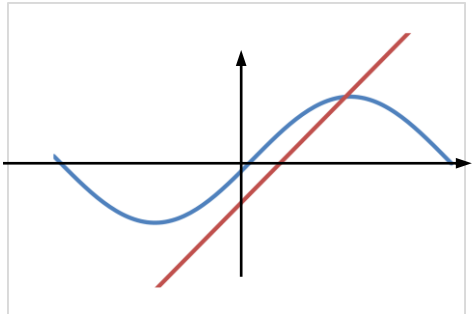


Fig. 1.3 Example 1

We deduce from the graph that the equation has only one root near $x = 1.5$. We then tabulate $f(x)$ near $x = 1.5$ as follows:

x	$f(x)$
1.45	0.0427
1.49	0.0067
1.50	-0.0025

There is at least one root

Since the sign of $f(x)$ is changed from positive for $x = 1.45$ to negative at $x = 1.5$, we conclude that the value of the root of $f(x) = 0$ lies between $x = 1.45$ and $x = 1.5$.. But the value of $f(x)$ at $x = 1.49$ is also positive but less than the value of $f(x)$ at $x = 1.45$. Therefore, we conclude that the root of the equation lies in the interval

$$1.49 < x < 1.5$$

1.2 Methods of solutions

We have seen in the above section how to estimate the number, nature and values of the roots of general nonlinear equations in the form

$f(x) = 0$. In the present section we shall introduce some methods for solving such equations with better accuracy.

1.2.1 Bisection method

The first technique, based on the intermediate value theorem, is called ***the bisection algorithm*** or ***binary − search method***. Suppose a continuous function f, defined on the interval $[a, b]$, is given with $f(a)$ and $f(b)$ of opposite signs. Then by Theorem 3.1, there exists a root α of the equation $f(x) = 0, a < \alpha < b$, for which $f(\alpha) = 0$. Although the procedure will work for the case when $f(a)$ and $f(b)$ have opposite signs and there is more than one root in the interval $[a,\ b]$, it will be assumed for simplicity that the root in this interval is unique.

To solve the equation

$$f(x) = 0, \tag{1}$$

the bisection method calls for a repeated halving of subintervals of $[a,\ b]$ and, at each step, locating the "half" containing the approximate value of the root. To begin, compute the first approximation as the midpoint of $[a, b]$; that is,

$$x_1 = \frac{a+b}{2} \tag{2}$$

At this point, we have one of the following cases:

1. If $f(x_1) = 0$, then $\alpha = x_1$ which is the required root. If not, then $f(x_1)$ has the same sign as either $f(a)$ or $f(b)$.
2. If $f(x_1)$ and $f(a)$ have opposite signs, then $\alpha \in (a,\ x_1)$ and a new approximation is given by

$$x_2 = \frac{a+x_1}{2} \tag{3}$$

3. If $f(x_1)$ and $f(b)$ have opposite signs, then $\alpha \in (x_1,\ b)$ and a new approximation is given by

$$x_2 = \frac{x_1 + b}{2} \tag{4}$$

We can obviously continue this interval-halving to obtain a smaller and smaller subinterval within which a root lies. This produces the bisection algorithm (see Fig. 1.4).

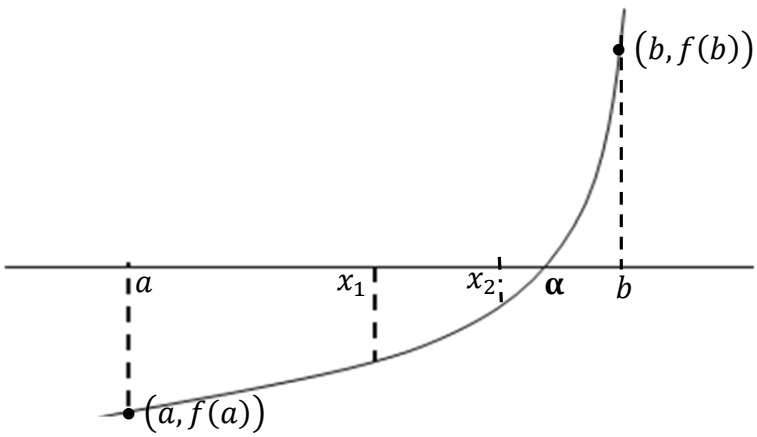


Fig. 1.4. The bisection algorithm

The Bisection method, as conceptually clear, has significant drawbacks and it is suitable for automatic computations. Besides it has the important property that it will always converge to a suitable solution and for that reason is often used as a "starter" for the more efficient methods presented later in this chapter. Another important advantage of the bisection method, beyond its simplicity, is our knowledge of the accuracy of the current approximation to the root.

The accuracy of a computed value is usually expressed either as the ***absolute error***; namely

$$Absolute\ error = |true\ value - approximate\ value|,$$

or as the ***relative error***; namely

$$Relative\ error = \frac{Absolute\ error}{true\ value}$$

The relative error is often the better measure of accuracy for very large or very small values. Sometimes the accuracy is expressed as the number of correct digits after the decimal point.

The error of the Bisection method is bounded by

$$|\alpha - x_n| \leq \frac{b-a}{2^n},$$

where the exact root α lies in the interval (a, b). However, the number of iterations required to determine the root under consideration within the tolerance Tol is given by

$$\frac{b-a}{2^n} \leq Tol,$$

which implies that

$$2^n \geq \frac{b-a}{Tol},$$

or

$$n \geq \frac{1}{\ln 2}\frac{b-a}{Tol}$$

Example 1 Find the smallest positive root of the equation

$$x^3 - 8x + 5 = 0,$$

by the bisection method correct to 5 decimal places.

Solution The solution of this problem lies in two main steps. In the first step we locate the interval containing the root. Put

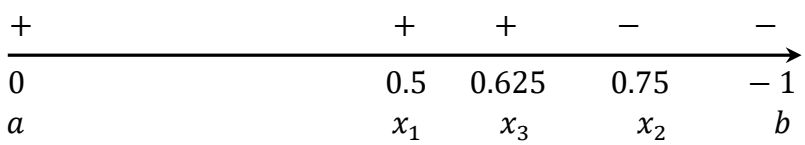
$$f(x) = x^3 - 8x + 5,$$

then

x	0	1
$f(x)$	5	-2

$$\underbrace{\qquad\qquad}_{f(0)f(1)<0}$$

The inequality $f(0)f(1) < 0$ implies that there is at least one root in the interval $(0,1)$.



The bisection method gives the following values.

$$x_1 = \frac{0+1}{2} = 0.5, \qquad f(0.5) = 1.125$$

$$x_2 = \frac{0.5+1}{2} = 0.75, \qquad f(0.75) = -0.578125$$

$$x_3 = \frac{0.5+0.75}{2} = 0.625, \qquad f(0.625) = 0.244141$$

$$x_4 = \frac{0.625+0.75}{2} = 0.6875, \qquad f(0.6875) = -0.175049$$

$$x_5 = \frac{0.625+0.6875}{2} = 0.65625, \qquad f(0.65625) = 0.032623$$

$$\vdots \qquad\qquad \vdots$$

and so on.

The first eighteen iterations are listed in the following table.

n	$a(+)$	$b(-)$	x	$f(x)$
1	0	1	0.5	1.125000
2	0.500000	1.000000	0.750000	-0.578125
3	0.500000	0.750000	0.625000	0.244141
4	0.625000	0.750000	0.687500	-0.175049
5	0.625000	0.687500	0.656250	0.032623
6	0.656250	0.687500	0.671875	-0.071705
7	0.656250	0.671875	0.664063	-0.019666
8	0.656250	0.664063	0.660157	0.006445
9	0.660157	0.664063	0.662110	-0.006618
10	0.660157	0.662110	0.661134	-0.000092
11	0.660157	0.661134	0.660646	0.003173
12	0.660646	0.661134	0.660890	0.001541
13	0.660890	0.661134	0.661012	0.000725
14	0.661012	0.661134	0.661073	0.000316
15	0.661073	0.661134	0.661104	0.000109
16	0.661104	0.661134	0.661119	0.000009
17	0.661119	0.661134	0.661127	-0.000045
18	0.661119	0.661127	0.661123	-0.000018

From this table, the positive root of the given equation is approximated to six decimal places by $x = 0.66112$ after 18 iterations.

Example 2 Find a point of intersection of the two curves

$$y = x^4 \ \ and \ \ y = 12x^2 + 10,$$

by the bisection method

Solution To find a point of intersection of the given two curves we have to solve the equation

$$x^4 = 12x^2 + 10,$$

that is, we solve the equation

$$f(x) = x^4 - 12x^2 - 10 = 0$$

Here, we have

x	0	1	2	3	4
$f(x)$	-10	-21	-42	-37	54

Then there is at least one root in the interval $(3, 4)$. Using the bisection method gives the following values.

n	$a(-)$	$b(+)$	x	$f(x)$
1	3	4	3.500000	-6.937500
2	3.500000	4.000000	3.750000	19.003906
3	3.500000	3.750000	3.625000	4.988525
4	3.500000	3.625000	3.562500	-1.225082
5	3.562500	3.625000	3.593750	1.817765
6	3.562500	3.593750	3.578125	0.280517
7	3.562500	3.578125	3.570313	-0.476170
8	3.570313	3.578125	3.574219	-0.098813
9	3.574219	3.578125	3.576172	0.090605
10	3.574219	3.576172	3.575196	-0.004117
11	3.575196	3.576172	3.575684	0.043228
12	3.575196	3.575684	3.575440	0.019552
13	3.575196	3.575440	3.575318	0.007716
14	3.575196	3.575318	3.575257	0.001799
15	3.575196	3.575257	3.575227	-0.001111
16	3.575227	3.575257	3.575242	0.000344
17	3.575227	3.575242	3.575235	-0.000335
18	3.575235	3.575242	3.575239	0.000053
19	3.575235	3.575239	3.575237	-0.000141
20	3.575237	3.575239	3.575238	-0.000044
21	3.575238	3.575239	3.575239	0.000053
22	3.575238	3.575239	3.575239	0.000053

After 22 iterations, we get $x_{app} = 3.575239$, correct to six decimal places. Therefore, the coordinate of the point of intersection is $(3.575239,\ 163.38806)$.

Advantages and drawbacks of the bisection method

- The bisection method is very easy and simple
- The method is always convergent.
- The error bound decreases by one half with each iteration. That is the error can be controlled.
- The bisection method converges very slowly, so that it is used to get an initial estimate for much faster methods
- The bisection method cannot detect multiple roots

Example 3 Find the positive root of the equation $x = e^{-x}$ correct to 3 decimal places by the Bisection method.

Solution First we use graphs of the curves

$$y = x \quad and \quad y = e^{-x},$$

as shown in Fig. 1.5 and we notice from the table

x	0.5	0.6
$f(x)$	-0.1756	0.0933

that

$$f(x) = xe^{x} - 1$$

Changes its sign in the interval

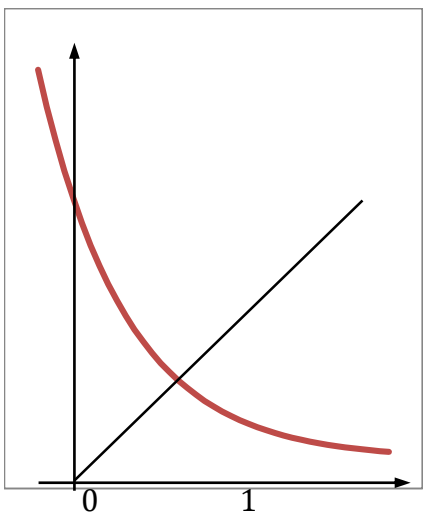


Fig. 1.5

$(0.5,\ 0.6)$. Apply the bisection method as shown in the following table

n	$a(-)$	$b(+)$	x	$f(x)$
1	0.5	0.6	0.550000	-0.046711
2	0.550000	0.600000	0.575000	0.021850
3	0.550000	0.575000	0.562500	-0.012782
4	0.562500	0.575000	0.568750	0.004446
5	0.562500	0.568750	0.565625	-0.004190
6	0.565625	0.568750	0.567188	0.000124
7	0.565625	0.567188	0.566407	-0.002033
8	0.566407	0.567188	0.566798	-0.000954
9	0.566798	0.567188	0.566993	-0.000415
10	0.566993	0.567188	0.567091	-0.000144
11	0.567091	0.567188	0.567140	-0.000009

After 11 iterations, the difference between the last two values of $f(x)$ is

$$0.567140 - 0.567091 \cong 0.000049 = 4.9 \times 10^{-4} < 5 \times 10^{-3},$$

then the last value of x is the required value of the root i.e.,

$$x_{app} \cong 0.567$$

correct to 3-decimal places.

Example 4 Solve the equation

$$4e^{-x} \sin x - 1 = 0,$$

in the interval $(0,\ 0.5)$.

Solution The bisection method yields.

n	$a(-)$	$b(+)$	x	$f(x)$
1	0	0.5	0.25	-0.229286411
2	0.25	0.5	0.375	0.006940729
3	0.25	0.375	0.3125	-0.100292711
4	0.3125	0.375	0.34375	-0.044067942
5	0.34375	0.375	0.359375	-0.017925398
6	0.359375	0.375	0.3671875	-0.005334495
7	0.3671875	0.375	0.37109375	0.000842363
8	0.3671875	0.37109375	0.369140625	-0.002236228
9	0.369140625	0.37109375	0.370117188	-0.000694476
10	0.370117188	0.37109375	0.370605469	7.45574E-05
11	0.370117188	0.370605469	0.370361328	-0.000309806
12	0.370361328	0.370605469	0.370483398	-0.000117586
13	0.370483398	0.370605469	0.370544434	-2.15046E-05
14	0.370544434	0.370605469	0.370574951	2.65288E-05
15	0.370544434	0.370574951	0.370559692	2.51269E-06
16	0.370544434	0.370559692	0.370552063	-9.49581E-06
17	0.370552063	0.370559692	0.370555878	-3.49152E-06
18	0.370555878	0.370559692	0.370557785	-4.89403E-07
19	0.370557785	0.370559692	0.370558739	1.01165E-06
20	0.370557785	0.370558739	0.370558262	2.61123E-07
21	0.370557785	0.370558262	0.370558023	-1.14139E-07
22	0.370558023	0.370558262	0.370558143	7.3492E-08
23	0.370558023	0.370558143	0.370558083	-2.03237E-08
24	0.370558083	0.370558143	0.370558113	2.65841E-08
25	0.370558083	0.370558113	0.370558098	3.13021E-09

From this table, we get

$$x_{app} = 0.370558,$$

correct to six decimal places.

1.2.2 False position method

This is an alternative method for solving the general nonlinear continuous equation of form

$$f(x) = 0, \tag{1}$$

when it is known that the required root is located between two values a and b given that $f(a)$ and $f(b)$ have opposite signs. The false position method can be summarized in the following steps.

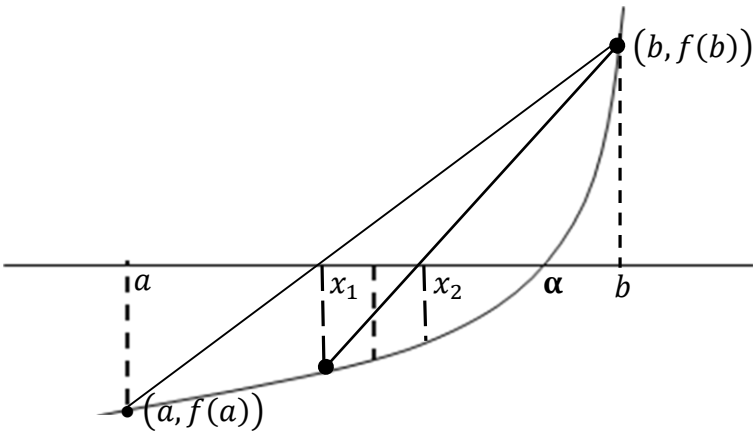


Fig. 1.6. The false position method

***Step* 1**. Approximate the function $f(x)$ in the interval (a, b) to a straight line joining the two points $(a, f(a))$ and $(b, f(b))$ as shown in Fig. 3.8. The equation of this line is

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a)$$

***Step* 2**. The first approximate solution can be considered as the value x_1 which is the intersection of the approximated straight line with the x −axis. That is to say when $y = 0$, $x = x_1$. Substitute in the equation of the straight line we get

$$0 - f(a) = \frac{f(b) - f(a)}{b - a}(x_1 - a),$$

from which we get

$$x_1 = a - \frac{b-a}{f(b)-f(a)} f(a), \tag{2}$$

or

$$\boxed{x_1 = \frac{af(b) - bf(a)}{f(b)-f(a)}} \tag{3}$$

This equation may be written in the form

$$\boxed{x_1 = \frac{\begin{vmatrix} a & f(a) \\ b & f(b) \end{vmatrix}}{f(b)-f(a)}} \tag{4}$$

***Step* 3**. In the second approximation step, we consider a smaller range for the chosen period, that is by considering a new straight line joining between two of the three points $(a, f(a)), (b, f(b)), (x_1, f(x_1))$ at which the values of $f(x)$ have opposite signs. For example: if

$$f(a) \equiv -ve$$
$$f(b) \equiv +ve$$
$$f(x_1) \equiv -ve$$

then we consider the two points $(x_1, f(x_1))$, $(b, f(b))$ and substitute in the equation

$$x_1 = \frac{\begin{vmatrix} x_1 & f(x_1) \\ b & f(b) \end{vmatrix}}{f(b)-f(x_1)}$$

***Step* 3** is repeated until the difference between two successive approximations is in the range of assumed allowable tolerance.

Advantages and drawbacks of the false position method

The false position method is sometimes called the method of linear interpolation. It is a two-point method, whereas Newton-Raphson is a one-point method. It has almost assured convergence, and it may converge to a root faster. However it may happen that most or all of the calculated x values are in the same side of the root in which case convergence may be slow.

Example 1 Find the smallest positive root of the equation

$$x^3 - 8x + 5 = 0,$$

by the false position method.

Solution Here we have $f(x) = x^3 - 8x + 5$. The zeros of this polynomial, we are searching for lie on the interval $(0,1)$. Thus

$$a = 0 \qquad f(a) = 5$$
$$b = 1 \qquad f(b) = -2$$

The first approximation is

$$x_1 = \frac{\begin{vmatrix} a & f(a) \\ b & f(b) \end{vmatrix}}{f(b) - f(a)} = \frac{(0)(-2) - (1)(5)}{-2 - 5} = 0.714286,$$

$$f(x_1) = -0.349854$$

The sign of $f(x_1)$ implies that the required root lies in the interval $(0,\ 0.714286)$. Therefore, the second approximation is given by

$$a = 0 \qquad f(a) = 5$$
$$x_1 = 0.714286 \qquad f(b) = -0.349854$$

Therefore, the second approximation is given by

$$x_2 = \frac{(0)(-0.349854) - (0.714286)(5)}{-0.349854 - 5} = 0.667575,$$

$$f(x_2) = -0.04309,$$

$$\vdots$$

Repeating the process, we get the following table

n	x	$f(x)$
	0	5.000000
	1	-2.000000
1	0.714286	-0.349856
2	0.667575	-0.043091
3	0.661871	-0.005020
4	0.661207	-0.000580
5	0.661130	-0.000065
6	0.661121	-0.000005
7	0.661120	0.000002
8	0.661120	0.000002

After 8 iterations, we see that $x_1 = 0.66112$ approximates the required root correct to 5 decimal places by the false position method while in the Bisection method we need to 18 iterations.

Example 2 Using the false position method, find the three roots of the equation

$$x^3 - 4x^2 - 5x + 10 = 0$$

Solution First, to locate the roots of this equation we construct the table for the function $f(x) = x^3 - 4x^2 - 5x + 10$,

x	0	1	2	3	4	5
$f(x)$	10	2	-8	-14	-10	10

$\underbrace{\qquad\qquad}_{f(1)f(2)<0}$ $\underbrace{\qquad\qquad}_{f(4)f(5)<0}$

It is evident that the given equation has two roots in the intervals $(1,2)$ and $(4,5)$, respectively. On finding these roots by the use of the false position method we obtain

n	x	$f(x)$
	1	2
	2	-8
1	1.200000	-0.032000
2	1.196850	0.000378
3	1.196887	-0.000002
4	1.196887	-0.000002

n	x	$f(x)$
	4	-10
	5	10
1	4.500000	-2.375000
2	4.595960	-0.391427
3	4.611179	-0.060413
4	4.613514	-0.009224
5	4.613870	-0.001410
6	4.613924	-0.000225
7	4.613933	-0.000027
8	4.613934	-0.000005
9	4.613934	-0.000005

From the above tables we conclude that

$$x_1 = 1.196887 \quad \& \quad x_2 = 4.613934,$$

are two roots of the given equation correct to six decimal places. To get the last root we use the relation roots and coefficients of algebraic equations

$$1.196887 + 4.613934 + x_3 = 4,$$

from which the third root is given by $x_3 = -1.810821$.

Example 3 Find the positive root, between 5 and 8, of the equation $x^{2.2} = 69$ correct to 3 decimal places.

Solution Using the false position method with the starting values

$$(5, -34.506758) \quad \text{and} \quad (8, 28.00586)$$

To get the table

n	x	$f(x)$
	5	-34.506758
	8	28.005860
1	6.655990	-4.275627
2	6.834002	-0.406143
3	6.850670	-0.037546
4	6.852209	-0.003458
5	6.852351	-0.000313
6	6.852364	-0.000025
7	6.852365	-0.000003
8	6.852365	-0.000003

so that, the root after two iterations is 6.852365.

1.2.3 The secant method

The secant method is a modification of the false position method. Instead of requiring that the function have opposite signs at the two values used for interpolation, we can choose the two values nearest the root (as indicated by the magnitude of the function at the various points) and interpolate from these. Usually the nearest values to the root will be the last two values calculated. This makes the interval

under consideration shorter and hence improves the assumption that the function can be represented by the line through the two points.

To solve the equation

$$f(x) = 0, \tag{1}$$

the secant method, as illustrated in Fig. 1.7, uses the secant to the curve $y = f(x)$ employing two previous approximations x_{n-1} and x_n. That is

$$\frac{0 - f(x_n)}{x_{n+1} - x_n} = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

or

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n), \quad n = 1,2,\cdots \tag{2}$$

or, alternatively

$$\boxed{x_{n+1} = \frac{\begin{vmatrix} x_{n-1} & f(x_{n-1}) \\ x_n & f(x_n) \end{vmatrix}}{f(x_n) - f(x_{n-1})}, \quad n = 1,2,\cdots} \tag{3}$$

After using (2) or (3) to generate x_{n+1}, x_{n-1} takes on the value of x_n and x_n takes on the value of x_{n+1}; the old value of x_{n-1} is discarded.

Convergence condition of the secant method

From a study of Fig. 3.9 it is clear that the secant method converges for any continuous function. However, from equation (1), we have

$$x_n - x_{n+1} = \frac{f(x_n)}{m_n}, \tag{4}$$

where m_n is the slop defined by $m_n = \frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$. Equation (4) represents the difference between two successive iterations, i.e., it can be named "***correction term***". Then for convergence, it must be:

$$\lim_{n\to\infty} \frac{f(x_n)}{m_n} = 0,$$

i.e., the correction term has decreasing values going from one iteration step to the other.

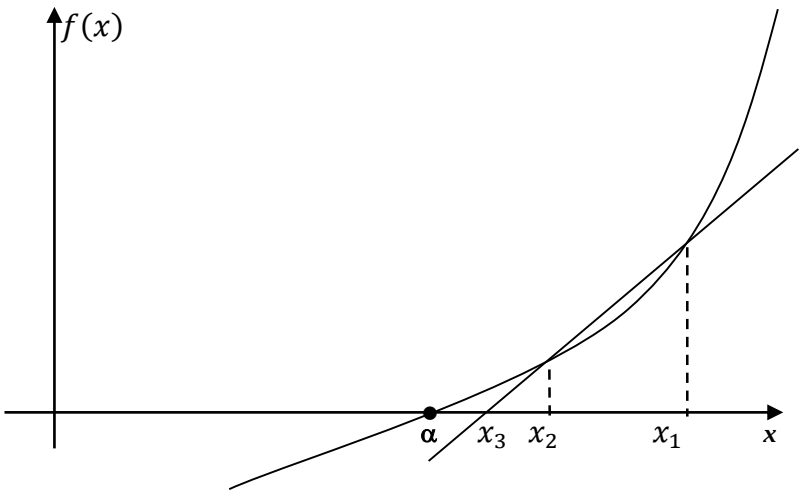


Fig. 1.7. The secant method

Example 1 Find the root of the equation

$$x^3 - 8x + 5 = 0,$$

in the interval $(0,1)$ using the secant method.

Solution Using the secant method to solve the equation

$$f(x) = x^3 - 8x + 5 = 0,$$

with

$$f(0) = 5 \quad \& \quad f(1) = -2$$

so that we get the following results

n	x	$f(x)$
	0	5
	1	-2
1	0.714285714	-0.349854
2	0.653710295	0.049672
3	0.661241475	-0.000810
4	0.661120635	-0.000002
5	0.661120336	0.000000
6	0.661120336	0.000000

From the above table the approximate root correct to 9 decimal places is 0.661120336.

Example 2 Find the positive root of $\sin x - \frac{x}{2} = 0$, using the secant method and the method of false position.

Solution Since the root lies between $\frac{\pi}{2}$ and π (see Fig. 1.8) we use $x_0 = \frac{\pi}{2}$ and $x_0 = \pi$. Using the formulas of the secant and false position methods we obtain the table below

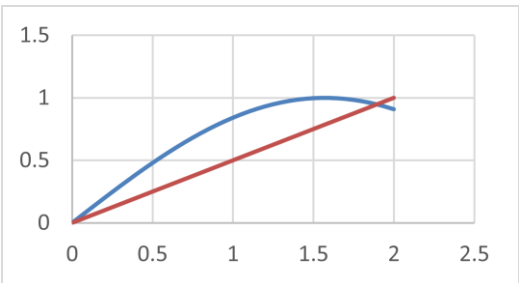


Fig. 1.8. The function $f(x) = \sin x - \frac{x}{2}$

n	False position method		Secant method	
	x	$f(x)$	x	$f(x)$
	1.570796	0.214602	1.570796	0.214602
	3.141593	-1.570796	3.141593	-1.570796
1	1.759603	0.102427	1.759604	0.102427
2	1.844202	0.040756	1.844203	0.040755
3	1.877013	0.014974	1.900109	-0.003790
4	1.888954	0.005336	1.895352	0.000117
5	1.893195	0.001881	1.895494	0.000000
6	1.894688	0.000660	1.895494	0.000000
7	1.895212	0.000231		
8	1.895395	0.000081		
9	1.895460	0.000028		
10	1.895482	0.000010		
11	1.895490	0.000003		
12	1.895493	0.000001		

As we would expect, the secant method converges faster to the true value 1.895494 than the method of false position.

Exercise 1.1

In problems $1-10$, approximate to within 10^{-4} the roots of the following equations by using the bisection method

1. $x^3 - 4x - 8.95 = 0$

[Answer: $x_{10} = 2.7037429$]

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

3. $x^{10} - 10 = 0, \quad (0,\ 1.3)$

[Answer: $x_{10} = 1.258925$]

4. $x - 0.2\sin x - 0.8 = 0, \qquad \left(0, \ \frac{\pi}{2}\right)$

5. $\frac{40}{x}(1 - e^{-0.2x}) - 5 = 0,$

[Answer: $x_{14} \cong 5.1358$]

6. $x - \cos x = 0, \qquad \left(0, \ \frac{\pi}{2}\right)$

7. $4e^{-x}\sin x - 1 = 0,$

[Answer: $x_{10} \cong 0.370, \quad x_{10} \cong 1.364$]
Correct to 3-decimal places

8. $2x^4 - 5x^3 - 5 = 0$

9. $2x^4 - 4.759x^3 + 3x - 14.2 = 0,$

[Answer: $x \cong 2.57051$, after 20 iterations]

10. $x^4 + 2.3x^3 + 5x - 11.5 = 0$

In problems $11 - 15$, Use the false-position and secant methods to:

11. find the smallest positive root of the equation $x^4 - 2x^3 - 9 = 0$ correct to 6 decimal places.

[Answer: $x_7 \cong 2.545598$, 7 iteratios]

12. find the smallest positive root of the equation $1.5x - 10e^{-0.2x} = 0$, correct to four decimal places.

13. find the point of intersection of the two curves $y = e^x - 2$, and $y = \ln(x + 2)$ correct to five correct decimal places.

[Answer: $x \cong 1.14619$, 6 iteratios]

14. solve the problems

(i) $x^{10} - 10 = 0, \quad (0, \ 1.3)$

(ii) $x - 0.2\sin x - 0.8 = 0, \qquad \left(0, \ \frac{\pi}{2}\right)$

(iii) $\frac{650}{x}(1-e^{-0.15x})-40=0, \qquad (14,\ 16)$

(iv) $x^3+3x^2-1=0, \qquad (-4,\ 1)$

(v) $x-\cos x=0, \qquad \left(0,\ \frac{\pi}{2}\right)$

15. find the smallest positive root of the equation $2\sin x+x-2=0$, correct to 5-decimal places.

[Answer: $x \cong 0.70458,\ 5$ iteratios]

16. Derive the false position formula to solve the equation $f(x)=0$. Hence use this method to find the smallest positive root of the equation

$$x^5-3x^4+5x-25=0,$$

correct to two decimal places.

17. Use the false position method to find the root of the equation

$$x^2+4=4x+e^{x/4},$$

which lies in the interval $(25,\ 26)$ correct to five decimal places.

[Answer: $x \cong 25.12854,\ 5$ iteratios]

18 Solve the equation

$$x^4-12x^2-10=0,$$

correct to four decimal places by using the false position method.

19. Use the false position formula to find the smallest positive root of the equation

$$x^2+\ln x=4x-1,$$

correct to four decimal places. Compare your results with those obtained by the bisection method.

[Answer: $x \cong 3.33947,\ 5$ iteratios, 18 iterations]

20. Solve the transcendental equation $\frac{650}{x}(1-e^{-0.15x})-40=0$, in the interval $(14,\ 16)$, correct to three decimal places.

1.2.4 Simple iteration method

We now discuss another method that is of general applicability. We begin with the equation

$$f(x) = 0, \tag{1}$$

and rearrange it into an equivalent expression of the form

$$f(x) = x - \varphi(x) = 0, \tag{2}$$

so that we have

$$x = \varphi(x) \tag{3}$$

Any solution α of (1) satisfies also equation (3), that is

$$f(\alpha) = 0 \quad \Rightarrow \quad \alpha = \varphi(\alpha) \tag{4}$$

Once an iteration function $\varphi(x)$ for solving (1) is chosen, one carries out the following successive iterations

$$x_1 = \varphi(x_0),$$
$$x_2 = \varphi(x_1),$$
$$x_3 = \varphi(x_2),$$
$$\vdots$$

and in general

$$x_{n+1} = \varphi(x_n), \quad n = 0,1,2,\cdots, \tag{5}$$

which enables us to construct a sequence of approximations to some root provided that we have a rough approximation x_0 of the desired root. This process is called ***simple iteration method*** or the ***fixed-point iteration***. If $|x_{n+1} - x_n|$ decreases as n increases, the process converges to a zero of $f(x)$ under suitable conditions.

Practically, the successive iterations in (5) are repeated until one of the following conditions is satisfied.

($\boldsymbol{a}$) Certain number of iterations is achieved.

($\boldsymbol{b}$) If $|x_{n+1} - x_n| \leq Tol$, where Tol is certain defined accuracy.

($\boldsymbol{c}$) If $|f(x_n)| \leq E$, i.e., if we substitute with the obtained root in the function $f(x)$, its modulus must be less than or equal to E.

There are several ways for constructing the iteration function $\varphi(x)$ from equation the given equation. If, for example, we have

$$f(x) = x^3 - 5x + 2 = 0,$$

then among possible choices for $\varphi(x)$ are the following:

(i) $\varphi_1(x) = \sqrt[3]{5x-2}$,

(ii) $\varphi_2(x) = \frac{x^3+2}{5}$,

(iii) $\varphi_3(x) = \frac{5}{x} - \frac{2}{x^2}$,

(iv) $\varphi_4(x) = x^3 - 4x + 2$,

(v) $\varphi_5(x) = \sqrt{x^3 + x^2 - 5x + 2}$

$\vdots$

Each such $\varphi(x)$ is called an ***iteration function*** for solving (1). However, The convergence of the iteration process depends on the choice of $\varphi(x)$ as well as the starting point x_0. The following Theorem shows how to choose the function $\varphi(x)$ and the starting point x_0 which guarantee the convergence of the method.

Theorem 1.2 Let $x = \alpha$ be a solution of the equation $x = \varphi(x)$ and suppose that φ has a continuous derivative in some interval J containing α. Then if

$$|\varphi'(x)| < 1, \qquad \forall x \in J,$$

the iteration process

$$x_{n+1} = \varphi(x_n), \quad n = 0,1,2,\cdots$$

converges to the root $x = \alpha$ for any $x_0 \in J$.

Proof. By the mean value theorem of differential calculus there is an η between x and α such that

$$\begin{aligned} x_{n+1} - \alpha &= \varphi(x_n) - \varphi(\alpha) \\ &= \varphi'(\eta)(x_n - \alpha) \\ &= \varphi'(\eta)\big(\varphi(x_{n-1}) - \varphi(\alpha)\big) \\ &= \big(\varphi'(\eta)\big)^2 (x_{n-1} - \alpha) \\ &\vdots \\ &= \big(\varphi'(\eta)\big)^{n+1} (x_0 - \alpha) \end{aligned}$$

For convergence of the iteration process it must be that

$$\lim_{n\to\infty} |x_{n+1} - \alpha| = 0,$$

i.e.,

$$\lim_{n\to\infty} \big(\varphi'(\eta)\big)^{n+1} (x_0 - \alpha) = 0$$

Since $(x_0 - \alpha)$ is a definite small value, then the convergence condition must be

$$\lim_{n\to\infty} \big(\varphi'(\eta)\big)^{n+1} = 0,$$

which is correct only if the following ***convergence condition*** holds

$$\boxed{|\varphi'(x)| < 1, \quad x \in J} \tag{6}$$

This completes the proof.

Remark 1 The rate of convergence depends on the factor $|\varphi'(x)|^{n+1}$. The smaller the value of $|\varphi'(x)|$, the faster the convergence, which may be very slow if $|\varphi'(x)|$ is

close to 1.

Remark 2 Condition (6) is not necessary for the convergence of the ***simple iteration method***.

The procedure is detailed in the simple iteration Algorithm and described in Fig. 1.9.

Example 1 Use the simple iteration method to obtain the smallest positive root of the equation

$$x^3 - 8x + 5 = 0,$$

correct to five decimal places.

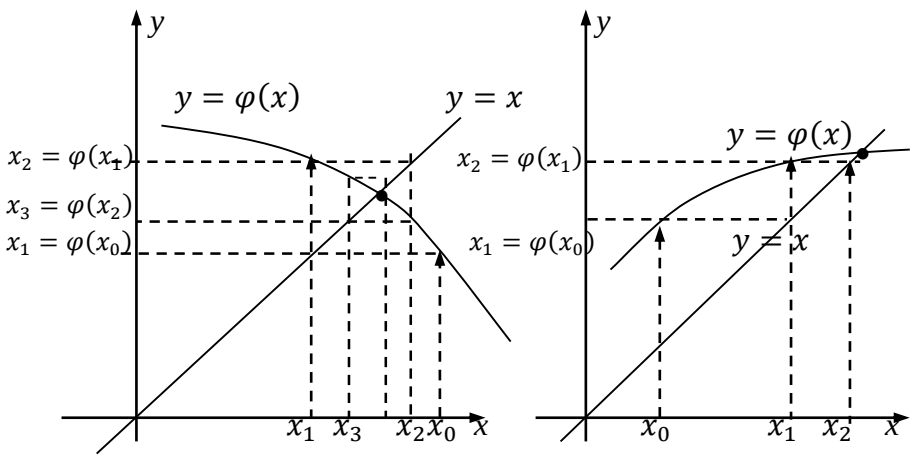


Fig. 1.9

Solution Put $f(x) = x^3 - 8x + 5$, thus, we have

x	0	1
$f(x)$	5	-2

$$\underbrace{\qquad\qquad}_{f(0)f(1)<0}$$

The inequality $f(0)f(1) < 0$ implies that there is at least one root in the interval $(0,1)$. Take as an initial approximation $x_0 = 0$.

The second step is to improve the accuracy of this initial approximation. At this stage we use the simple iteration method to achieve that. Choose the iteration function to be

$$x = \varphi(x) = \frac{x^3 + 5}{8},$$

which satisfies the convergence condition since

$$\varphi'(x) = \frac{3x^2}{8} < 1,$$

throughout the interval $(0,1)$. Therefore, we use the iteration

$$\boxed{x_{n+1} = \frac{x_n^3 + 5}{8}, \qquad n = 0,1,\cdots}$$

On taking $x_0 = 0$, we obtain

n	x
1	0
2	0.625
3	0.655517578
4	0.660209758
5	0.660971275
6	0.661095891
7	0.661116311
8	0.661119658

Since x_7 and x_8 are equal for 5 decimal numbers, then the solution is $x = 0.66111$ correct to five decimal places.

Example 2 Solve the equation

$$x^4 - 12x^2 - 10 = 0,$$

by the use of the simple iteration method correct to five decimal places.

Solution Put

$$f(x) = x^4 - 12x^2 - 10,$$

thus, we have

x	0	1	2	3	4
$f(x)$	-10	-21	-42	-37	54

$\underbrace{\qquad\qquad}_{f(3)f(4)<0}$

From the above table, there is at least one root in the interval $(3,4)$. Choose the iteration function to be

$$x = \varphi(x) = \sqrt[4]{12x^2 + 10},$$

which yields

$$\varphi'(x) = \frac{1}{4}\frac{24x}{(12x^2 + 10)^{3/4}},$$

Now we use the end points of the solution interval, to get

$$|\varphi'(3)| = 0.50276033 < 1,$$

$$|\varphi'(4)| = 0.44791717 < 1$$

Therefore, we use the iteration

$$\boxed{x_{n+1} = \sqrt[4]{12x_n^2 + 10}, \qquad n = 0,1,\cdots}$$

On taking $x_0 = 3$, we obtain

n	x
1	3
2	3.295873
3	3.441959
4	3.512195
5	3.545539
6	3.561274
7	3.568678
8	3.572158
9	3.573792
10	3.574559
11	3.574920
12	3.575089
13	3.575168
14	3.575205
15	3.575223
16	3.575231
17	3.575235
18	3.575237
19	3.575238
20	3.575238

The solution is $x = 3.575238$ correct to six decimal places.

Example 2 Use the simple iteration method to obtain a maximum value correct to four decimal places of the function

$$f(x) = x^4 - 14x^2 + 16x$$

Solution To get the critical points of the given function we put $f'(x) = 0$,

$$f'(x) = 4x^3 - 28x + 16 = 0,$$

or

$$x^3 - 7x + 4 = 0 \tag{7}$$

To solve this equation, we first locate analytically the smallest positive root as follows:

$\boldsymbol{x}$	$\mathbf{0}$	$\mathbf{1}$
$\boldsymbol{f(x)}$	4	-2

$\underbrace{\qquad\qquad}_{f(0)f(1)<0}$

which implies that there is at least one root in the interval $(0,1)$. Using the simple iteration method to solve equation (7) with $x_0 = 0$. Choose the iteration function to be

$$x = \varphi(x) = \frac{x^3 + 4}{7},$$

which satisfies the convergence condition since

$$\varphi'(x) = \frac{3x^2}{7} < 1, \qquad \forall x \in (0,1)$$

Hence, we use the iteration

$$\boxed{x_{n+1} = \frac{x_n^3 + 4}{7}, \qquad n = 0,1,\cdots}$$

with $x_0 = 0$, and this will yield the following table of results

$\boldsymbol{n}$	$\boldsymbol{x}$
1	0
2	0.57142857
3	0.59808413
4	0.60199107
5	0.60259393
6	0.60268765

7	0.60270224
8	0.60270451
9	0.60270487
10	0.60270492
11	0.60270493
12	0.60270493

Since x_7 and x_8 are equal for 4 decimal numbers, then the solution is $x = 0.6027$ correct to four decimal places and it is equal to $x = 0.60270493$ which is correct to 8 decimal places after 12 iterations. Anyway the maximum value of the function at this point is approximated by

$$f_{max} \cong 4.689686534$$

Example 3 Taking $x_0 = 1$, find the root of the equation

$$f(x) = 3x - e^{-x} = 0$$

to accuracy 0.0001, using the method of simple iteration.

Solution We first write

$$3x - e^{-x} = 0,$$

in the form

$$x = \frac{1}{3}e^{-x} = \varphi(x),$$

which satisfies

$$|\varphi'(x_0)| = \frac{1}{3}e^{-1} < 1$$

We have by iteration

$$\boxed{x_{n+1} = \frac{1}{3}e^{-x_n}, \qquad n = 0,1,\cdots}$$

n	x
1	1
2	0.12262648
3	0.29486467
4	0.24821078
5	0.26006516
6	0.25700045
7	0.25778929
8	0.25758601
9	0.25763838
10	0.25762489
11	0.25762836
12	0.25762747
13	0.25762770

$$x_{13} - x_{12} = 2.3 \times 10^{-7} < 5 \times 10^{-6}$$

1.2.5 Newton-Raphson method

Newton-Raphson (or simply Newton's) method is one of the most powerful and well-known numerical methods for solving a root-finding problem

$$f(x) = 0, \tag{1}$$

where f is a differentiable function. There are at least three common ways of introducing Newton's method. The most common approach is by considering the technique graphically (see Fig. 1.10).

The equation of the tangent line to the curve $y = f(x)$ at the point $\big(x_0, f(x_0)\big)$ is given by

$$\frac{y - f(x_0)}{x - x_0} = f'(x_0), \tag{2}$$

The intersection between this tangent and the $x - axis$ gives the first approximated solution x_1. The value of x_1 is computed by substituting $x = x_1$ when $y = 0$ in equation (2), so we have

$$\frac{0 - f(x_0)}{x_1 - x_0} = f'(x_0), \tag{3}$$

or,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{4}$$

Using successive tangents we find that the $(n+1)th$ approximation is given in the form

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0,1,2,\cdots} \tag{5}$$

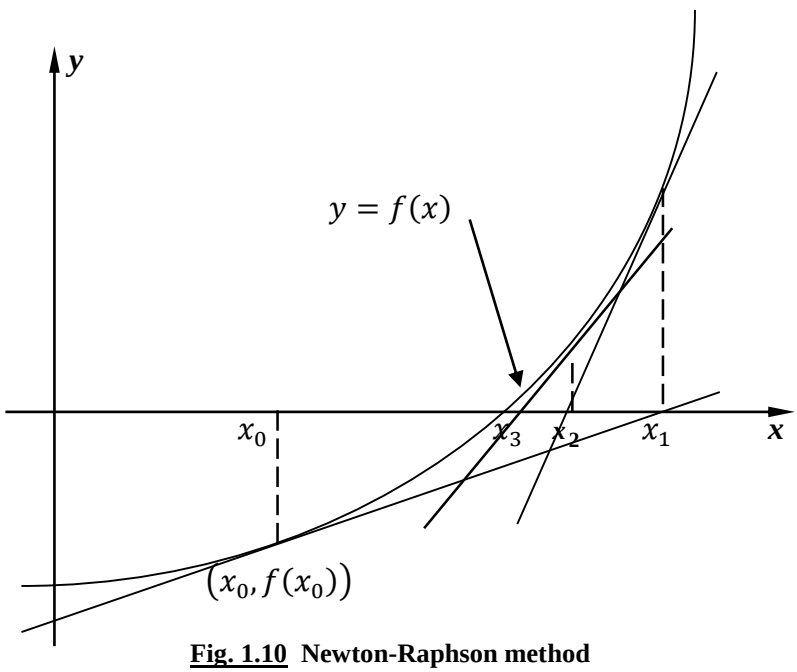


Fig. 1.10 **Newton-Raphson method**

It is clear from this equation that Newton's method cannot be continued if $f'(x_n) = 0$ for some n. We will note that the method is most effective when f' is bounded away from zero near the exact root α.

Newton's algorithm is widely used because at least in the near neighborhood of a root, it is more rapidly convergent than any of the methods so far discussed. However, offsetting this is the need for two function evaluations at each step, $f(x_n)$ and $f'(x_n)$.

♦ Convergence of Newton's method

We now use the result of section 3.2.2 to show a criterion for convergence of Newton's method. The algorithm

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0,1,2,\cdots,$$

is of the form $x_{n+1} = \varphi(x_n)$. Successive iterations converge if the iteration function $\varphi(x)$ satisfies the convergence condition $|\varphi'(x)| < 1$. Since

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

then

$$\varphi'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{\left(f'(x)\right)^2} = \frac{f(x)f''(x)}{\left(f'(x)\right)^2}$$

Hence if

$$\left|\frac{f(x)f''(x)}{\left(f'(x)\right)^2} < 1,\right|$$

on an interval about the root α , the method will converge for any initial value x_0 in the interval. The condition is sufficient only, and requires the usual continuity and existence of $f(x)$ and its derivatives.

Example 1 Solve the equation

$$x^3 - 8x + 5 = 0,$$

using Newton-Raphson method.

Solution Here we have

$$f(x) = x^3 - 8x + 5,$$
$$f'(x) = 3x^2 - 8,$$

and the root we are searching for lies on the interval $(0,1)$ since $f(0)f(1) < 0$. Thus, Newton's method has the iteration form

$$x_{n+1} = x_n - \frac{x_n^3 - 8x_n + 5}{3x_n^2 - 8}, \qquad n = 0,1,2,\cdots$$

which can be simplified as

$$\boxed{x_{n+1} = \frac{2x_n^3 - 5}{3x_n^2 - 8}, \qquad n = 0,1,2,\cdots}$$

where x_0 is yet to be selected. For some problems it suffices to let x_0 be arbitrary, while in others it is important to select a good initial approximation. For the problem under consideration, take $x_0 = 0$ as an initial approximation, the approximations in the following table are generated. An excellent approximation is obtained with $n = 4$.

n	x	$f(x)$	$f'(x)$
1	0	5.000000	-8.000000
2	0.625000	0.244141	-6.828125
3	0.660755	0.002444	-6.690208
4	0.661120	0.000002	-6.688761
5	0.661120	0.000002	-6.688761

The approximate root is $x = 0.66112$, which is the same result of the Bisection method but by carrying out 5 iterations only instead of 18 iterations in the Bisection method.

Advantages and drawbacks of Newton-Raphson method

Newton-Raphson method is an extremely powerful technique, but it has three major difficulties:

- The convergence of the method is not assured as in the case of Bisection method. Thus, when using computer for calculations it is preferred to use the Bisection method or any other method rather than the Newton-Raphson method
- The method fails in the case of multiple roots since in this case $f'(x) = 0$.
- The need to know the value of the derivative of $f(x)$ at each approximation is a major difficulty. Frequently $f'(x)$ is far more difficult and needs more arithmetic operations to calculate than $f(x)$. As a simple example, let

$$f(x) = x^2 3^x \sin 2x,$$

then

$$f'(x) = 2x3^x \sin 2x + (x^2 3^x \sin 2x) \ln 3 + 2x^2 3^x \cos 2x,$$

which is extremely tiresome to evaluate, particularly if the method is being performed on a nonprogrammable calculator.

However, the advantage of the method is

- its convergence rate is one of the fastest when it does converge.

To circumvent the problem of the derivative evaluation in Newton's method, a slight variation is introduced in the false position method.

Example 2 Solve the equation

$$x^6 - x - 1 = 0, \qquad x \in (1, 2)$$

Solution Denoting $f(x) = x^6 - x - 1$, we get $f'(x) = 6x^5 - 1$, and Newton's method yields the iteration formula

$$x_{n+1} = x_n - \frac{x_n^6 - x_n - 1}{6x_n^5 - 1}, \qquad n = 0,1,2,\cdots$$

Taking $x_0 = 1$ as an initial approximation the following table will result

n	x	$f(x)$	$f'(x)$
1	1	-1.000000	5.000000
2	1.200000	0.785984	13.929920
3	1.143576	0.093034	10.734819
4	1.134909	0.001903	10.296826
5	1.134724	-0.000001	10.287622
6	1.134724	-0.000001	10.287622

The approximate root is $x = 1.134724$.

Example 3 Using Newton-Raphson method, find the positive root of the equation

$$\sin x = x^2,$$

correct to 6-decimal places.

Solution On defining $f(x) = \sin x - x^2$, we first locate the root from the following table

x	0	0.25	0.5	0.75	1
$f(x)$	0	0.1849	0.2294	0.1191	-0.1585

We notice from the above table that $f(x) = \sin x - x^2$ changes its sign when x varies from $x = 0.75$ to $x = 1$. Thus the root lies in the interval $0.75 < x < 1$. Then by Newton-Raphson method an approximate value of the root is given by $x_0 = 0.875$ and Newton–Raphson method has the form

$$x_{n+1} = x_n - \frac{\sin x_n - x_n^2}{\cos x_n - 2x_n}, \qquad n = 0,1,2,\cdots$$

from which we construct the following table.

n	x	$f(x)$	$f'(x)$
1	0.75	0.119139	-0.768311
2	0.905066	-0.032679	-1.192498
3	0.877662	-0.001043	-1.116373
4	0.876728	-0.000002	-1.113786
5	0.876726	0.000000	-1.113781
6	0.876726	0.000000	-1.113781

We conclude that the root is 0.876726 correct to 6- decimal places.

Example 4 By applying the Newton-Raphson method find $\sqrt{11}$ correct to 5 decimal places.

Solution To find $\sqrt{11}$ it is equivalent to solve the equation

$$x = \sqrt{11},$$

or

$$x^2 = 11,$$

hence

$$f(x) = x^2 - 11 = 0$$

Then, by Newton-Raphson iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 11}{2x_n}, \qquad n = 0,1,2,\cdots$$

Since an approximate value of $\sqrt{11}$ is 3, we start with $x_0 = 3$, then

n	x	$f(x) = x^2 - 11$	$f'(x) = 2x$
1	3	-2.000000	6.000000
2	3.333333	0.111109	6.666666
3	3.316667	0.000280	6.633334
4	3.316625	0.000001	6.633250
5	3.316625	0.000001	6.633250
6	3.316625	0.000001	6.633250

From the above table, we have $\sqrt{11} = 3.316625$, correct to 6 decimal places.

Example 1 Find the smallest positive root, correct to 6 decimal places, of the following equation

$$x^6 - x - 1 = 0,$$

by the use of:

(i) Bisection method, **(ii)** Simple iteration method,

(iii) Newton-Raphson method, **(iv)** False position method,

(v) Secant method,

Solution The solution of this problem lies in two main steps. In the first step we locate the interval containing the root. Put

$$f(x) = x^6 - x - 1,$$

x	0	1	2
$f(x)$	-1	-1	61

$\underbrace{\qquad\qquad}_{f(1)f(2)<0}$

The inequality $f(1)f(2) < 0$ implies that there is at least one root in the interval $(1,2)$.

(i) Bisection method

n	$a(-)$	$b(+)$	$x = (a+b)/2$	
1	1	2	1.500000	8.890625
2	1	1.500000	1.250000	1.564697
3	1	1.250000	1.125000	-0.097713
4	1.125000	1.250000	1.187500	0.616653
5	1.125000	1.187500	1.156250	0.233269
6	1.125000	1.156250	1.140625	0.061578
7	1.125000	1.140625	1.132813	-0.019570
8	1.132813	1.140625	1.136719	0.020622
9	1.132813	1.136719	1.134766	0.000431
10	1.132813	1.134766	1.133790	-0.009588
11	1.133790	1.134766	1.134278	-0.004585
12	1.134278	1.134766	1.134522	-0.002079
13	1.134522	1.134766	1.134644	-0.000824
14	1.134644	1.134766	1.134705	-0.000197
15	1.134705	1.134766	1.134736	0.000122
16	1.134705	1.134736	1.134721	-0.000032
17	1.134721	1.134736	1.134729	0.000050
18	1.134721	1.134729	1.134725	0.000009
19	1.134721	1.134725	1.134723	-0.000012
20	1.134723	1.134725	1.134724	-0.000001

21	1.134724	1.134725	1.134725	0.000009
22	1.134724	1.134725	1.134725	0.000009
23	1.134724	1.134725	1.134725	0.000009
24	1.134724	1.134725	1.134725	0.000009
25	1.134724	1.134725	1.134725	0.000009

(ii) Simple iteration method

$$\varphi(x) = \sqrt[6]{x+1}, \quad \varphi'(x) = \frac{1}{6}(x+1)^{-5/6},$$

$$\varphi'(1) = 0.1 < 1, \qquad \varphi'(2) = 0.0667 < 1$$

$$x_{n+1} = \sqrt[6]{x_n + 1}, \qquad n = 0,1,\cdots$$

n	x
1	1
2	1.122462048
3	1.133635199
4	1.134627646
5	1.13471559
6	1.134723381
7	1.134724071
8	1.134724132
9	1.134724138
10	1.134724138

(iii) Newton-Raphson

$$f(x) = x^6 - x - 1,$$

$$f'(x) = 6x^5 - 1,$$

$$x_{n+1} = x_n - \frac{x_n^6 - x_n - 1}{6x_n^5 - 1} = \frac{5x_n^6 + 1}{6x_n^5 - 1}, \qquad n = 0,1,\cdots$$

n	x	$f(x)$	$f'(x)$
1	1	-1	5

2	1.2	0.785984	13.92992
3	1.143575843	0.093031957	10.73481137
4	1.134909462	0.001907397	10.29684929
5	1.134724221	8.53719E-07	10.2876329
6	1.134724138	1.70308E-13	10.28762877
7	1.134724138	-8.88178E-16	10.28762877
8	1.134724138	-8.88178E-16	10.28762877
9	1.134724138	-8.88178E-16	10.28762877
10	1.134724138	-8.88178E-16	10.28762877

(iv) False position method

n	x	$f(x)$
	1	-21
	2	-42
1	1.01612903	-0.9153677
2	1.03067475	-0.8319214
3	1.0437166	-0.7510232
4	1.05534703	-0.6737807
5	1.06566728	-0.601031
6	1.07478341	-0.5333466
7	1.08280282	-0.4710581
8	1.08983138	-0.4142873
9	1.09597118	-0.3629862
10	1.10131887	-0.3169753
11	1.10596456	-0.2759807
12	1.1099912	-0.2396656
13	1.11347431	-0.2076572
14	1.116482	-0.1795679
15	1.11907521	-0.1550111
16	1.12130811	-0.1336129
17	1.12322857	-0.1150198
18	1.12487867	-0.0989036
19	1.12629527	-0.0849636
20	1.12751052	-0.0729279
21	1.12855237	-0.0625525
22	1.12944508	-0.0536203
23	1.13020964	-0.0459393
24	1.13086419	-0.0393409

25	1.13142436	-0.0336771
26	1.13190362	-0.0288192
27	1.13231356	-0.024655
28	1.13266412	-0.0210874
29	1.13296385	-0.0180323
30	1.13322008	-0.015417
31	1.13343909	-0.0131791
32	1.13362627	-0.0112645
33	1.13378623	-0.009627
34	1.13392292	-0.0082267
35	1.1340397	-0.0070296

(v) Secant method

n	x	$f(x)$
	1	-1.000000
	2	61.000000
1	1.016129	-0.915368
2	1.030675	-0.831920
3	1.175689	0.465228
4	1.123679	-0.110634
5	1.133671	-0.010807
6	1.134753	0.000297
7	1.134724	-0.000001
8	1.134724	-0.000001

Exercise 1.2

1. Use the simple iteration method to find the root of the equation

$$x^2 + 4 = 2x + e^{x/4},$$

that lies in the interval $(25,\ 26)$ correct to four decimal places.

[Answer: $x \cong 25.1285,\ \ 8\ iterations$]

2. Solve numerically for x the equation

$$e^x - 3x^2 = 0,$$

correct to five decimal places.

3. Use the simple iteration method to find the smallest positive root of the equation

$$x^4 - 12x^2 - 10 = 0,$$

correct to four decimal places. Compare the results with those obtained by the bisection method.

[Answer: $x_8 \cong 1.1138$, $x_{13} = 3.575$]

4. Use the simple iteration method to obtain a maximum value correct to four decimal places of the function

$$f(x) = \frac{x^5}{5} - \frac{4.759}{4}x^4 + \frac{3}{2}x^2 - 14.277x + 12.5$$

5. Use Newton-Raphson method to find an approximation to a root of $f(x) = x^2 + 2xe^x + e^{2x} = 0$ starting with $x_0 = 0$ and performing 10 iterations. If you are using a computer, redo the calculations using extended precision.

[Answer: $x_{10} \cong -0.565741$, $x_{21} \cong -0.567143$]

6. Use the simple iteration method to find the smallest positive root of the equation

$$3x^4 + x^3 - 6 = 0,$$

correct to three decimal places. Compare your results with those obtained by the bisection method (Accuracy, running cost).

7. Use Newton-Raphson formula to find the smallest positive root of the equation

$$x^4 - 6.255x^3 + 10.78x^2 - 6.255x + 9.78 = 0,$$

correct to five decimal places. 3.282690

[Answer: $x_4 = 3.092056$]

8. Use Newton-Raphson formula to find the smallest positive root of the equation $x^4 - 3x^3 - 10 = 0$ correct to 6 decimal places. Compare your results with those obtained by the bisection method.

9. By Newton-Raphson formula, find the point of intersection of the two curves $y = e^x - 2$, and $y = \ln(x + 2)$ correct to four correct decimal places.

[Answer: $x_6 \cong 1.14619322$]

10. Find a minimum value of the function

$$f(x) = \frac{1}{6}x^6 - \frac{2.3}{5}x^5 + \frac{5}{2}x^2 - \frac{23}{2}x + 42,$$

correct to two decimal places using

(i) Newton-Raphson formula

(ii) Bisection method

(iii) Simple iteration method

11. Derive the false position formula to solve the equation $f(x) = 0$. Hence use this method to find the smallest positive root of the equation

$$x^2 + \ln x = 4x - 1,$$

correct to five decimal places. Compare your results with those obtained by Newton-Raphson method.

[Answer: $x_6 \cong 3.33947$, $x_5 \cong 3.33947$]

12. Solve $4\cos x = e^x$ with accuracy 10^{-4}, by using

(**i**) The Bisection method in $\left(\frac{\pi}{4}, \frac{\pi}{2}\right)$,

(**ii**) Newton's method with $x_0 = \frac{\pi}{4}$

(**iii**) The false position and secant methods in $\left(\frac{\pi}{4}, \frac{\pi}{2}\right)$.

13. Compare the results obtained when the bisection method and the method of Newton-Raphson are used to solve the equation

$$3\sin x = x + \frac{1}{x}$$

[Answer: $x_{21} \cong 2.105853$, $x_{13} \cong 2.105852$]

14. Use the simple iteration method to find a root, accurate within 10^{-5}, of $f(x) = \left(\sin x - \frac{x}{2}\right)^2 = 0$, starting with $x_0 = \frac{\pi}{2}$.

15. Find the root of the equation $\tan x + \tanh x = 0$ that lies in the interval $(2,3)$.

[Answer: $x_6 \cong 2.365020$]

16. Find the root of the equation $x^3 - 5x - 7 = 0$, which lies between 2 and 3 correct to four decimal places using

($\boldsymbol{i}$) The Bisection method

($\boldsymbol{ii}$) The method of simple iteration

($\boldsymbol{iii}$) The Newton-Raphson method

($\boldsymbol{iv}$) The method of false position

($\boldsymbol{v}$) The secant method

17. Solve the equation $\cos x = \ln x$ to five correct decimals.

[Answer: $x_5 \cong 1.302964$]

18. The problem "solve $x^2 - x - 2 = 0$" can be reformulated in several ways for fixed point iteration, including

$$x = x^2 - 2,\ \ x = \sqrt{x+2},\ \ x = 1 + \frac{2}{x},\ \ x = x - \frac{x^2 - x - 2}{4}$$

Determine which of these formulations gives a convergent fixed point iteration for the zero $x = 2$. Repeat this for the zero $x = -1$. Verify

the conclusions by making a few iterations on each formulation for the two zeros, starting at $x = 1.7$ and $x = -0.7$.

19. Use the secant method to find solve numerically for x the transcendental equation $650(1 - e^{-0.15x}) - 40x = 0$, which les in the interval $(14,\ 15)$, correct to six decimal places.

[Answer: $x_9 \cong 13.984808$]

20. Derive the convergence condition of the simple iteration method (fixed point method) to solve the equation $f(x) = 0$. Hence use this method to find the smallest positive root of the equation

$$x^4 - 4x^3 - 3x + 32 = 0,$$

correct to five decimal places. Compare your results with those obtained by the Newton-Raphson method.

2. Numerical solution of linear systems of equations (Iterative methods)

An iterative (indirect) method to solve the $n \times n$ linear systems

$$A\mathbf{x} = \mathbf{b}, \tag{1}$$

starts with an initial approximation $\mathbf{x}^{(0)}$ to the solution $\mathbf{x}$, and generates a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converges to $\mathbf{x}$. Most of these techniques involve a process that converts the system (1) into an equivalent system of the form

$$\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{C}, \tag{2}$$

for some $n \times n$ matrix $\mathbf{T}$ and vector $\mathbf{C}$. After the initial vector $\mathbf{x}^{(0)}$ is selected, the sequence of approximate solution vectors is generated by computing

$$\mathbf{x}^{(1)} = \mathbf{T}\mathbf{x}^{(0)} + \mathbf{C},$$
$$\mathbf{x}^{(2)} = \mathbf{T}\mathbf{x}^{(1)} + \mathbf{C},$$
$$\vdots$$

In general

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)} + \mathbf{C}, \tag{3}$$

for each $k = 0,1,2,\cdots$. Thus if we assume that the resulting sequence $\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots \mathbf{x}^{(k)}, \cdots\}$ converges to $\mathbf{x}$ as $k \to \infty$, the iterations in (3) give

$$\underbrace{\mathbf{x}^{(k+1)}}_{\mathbf{x}} = \underbrace{\mathbf{T}\mathbf{x}^{(k)} + \mathbf{C}}_{\mathbf{x}},$$

i.e., $\mathbf{x}$ is the solution of (2) and hence is the solution of (1).

Iterative methods are seldom used for solving linear systems of small dimension since the time required for sufficiently accuracy exceeds that required for direct methods such as the Gaussian

elimination method. For large systems with a high percentage of zero entries, however, these techniques are efficient in terms of computer storage and time requirements. Systems of this type arise frequently in the numerical solution of boundary value problems and partial differential equations.

We need the following definition during the discussion of the convergence of iterative methods.

Definition 2.1 A matrix $A = (a_{ij})$ is said to be ***diagonally dominant*** if

$$|a_{ij}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \tag{4}$$

The next theorem gives without proof the condition for the convergence of iterative methods for solving linear systems.

Theorem 2.1 The iterative method

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)} + \mathbf{C}, \qquad k = 0,1,2,\cdots \tag{5}$$

derived for solving the linear system $A\mathbf{x} = \mathbf{b}$, converges for any initial vector $\mathbf{x}^{(0)}$ and any vector $\mathbf{b}$ if and only if the matrix A is diagonally dominant.

2.1 Jacobi's iterative method

We shall present one of the possible transformations of a general system of the form

$$A\mathbf{x} = \mathbf{b}, \tag{6}$$

where $A = (a_{ij}), i, j = 1,2,\cdots,n,$ into the form

$$\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{C}, \tag{7}$$

Splitting A into its diagonal and off-diagonal parts. To see this, let D be the diagonal matrix whose diagonal is the same as A, and $-L$ be the strictly lower-triangular part of A, and $-U$ be the strictly upper-triangular part of A. With this notation, A is split into

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

$$- \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -a_{n1} & -a_{n1} & \cdots & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \cdots & -a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$= D - L - U \tag{8}$$

The equation

$$A\mathbf{x} = \mathbf{b}, \quad \text{or} \quad (D - L - U)\mathbf{x} = \mathbf{b},$$

is then transformed into

$$D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b}, \tag{9}$$

and finally, provided $a_{ii} \neq 0$ for all i,

$$\mathbf{x} = D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}, \tag{10}$$

This result in the matrix form is called the Jacobi's iterative technique:

$$\boxed{\mathbf{x}^{(k+1)} = D^{-1}(L + U)\mathbf{x}^{(k)} + D^{-1}\mathbf{b}, \quad k = 0,1,2\cdots} \tag{11}$$

This technique consists of solving the ith equation in $A\mathbf{x} = \mathbf{b}$ for x_i, $i = 0,1,2\cdots, n$ to obtain

$$\begin{array}{ccccccccc} a_{11}x_1 & = & 0 & -a_{12}x_2 & \cdots & -a_{1n}x_n & + & b_1 \\ a_{22}x_2 & = & -a_{22}x_1 & 0 & \cdots & -a_{2n}x_n & + & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{nn}x_n & = & -a_{n1}x_1 & -a_{n2}x_2 & \cdots & 0 & + & b_n \end{array}, \tag{12}$$

or

$$a_{ii}x_i = \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} - a_{ij}x_j + b_i \right), \quad i = 1,2,\cdots,n, \tag{13}$$

and generating each $x_i^{(k+1)}$ from components of $x_i^{(k)}$ for $k \geq 0$ by

$$\begin{array}{ccccccccc}
x_1^{(k+1)} & = & 0 & -\dfrac{a_{12}}{a_{11}}x_2^{(k)} & \cdots & -\dfrac{a_{1n}}{a_{11}}x_n^{(k)} & + & \dfrac{b_1}{a_{11}} \\
x_2^{(k+1)} & = & -\dfrac{a_{21}}{a_{22}}x_1^{(k)} & 0 & \cdots & -\dfrac{a_{2n}}{a_{22}}x_n^{(k)} & + & \dfrac{b_2}{a_{22}}, \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_n^{(k+1)} & = & -\dfrac{a_{n1}}{a_{nn}}x_1^{(k)} & -\dfrac{a_{n2}}{a_{nn}}x_2^{(k)} & \cdots & 0 & + & \dfrac{b_n}{a_{nn}}
\end{array} \tag{14}$$

or, in general, we have

$$\boxed{x_i^{(k+1)} = \frac{1}{a_{ii}} \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} - a_{ij}x_j^{(k)} + b_i \right), \qquad i = 1,2,\cdots,n, \quad k = 0,1,\cdots} \tag{15}$$

In practice, equation (15) is used in computation, with equation (11) being reserved for theoretical purposes. The iteration (14) can also be described as follows: To solve the system (6), we choose an initial vector $\mathbf{x}^{(\mathbf{0})}$ with components $x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}$. We substitute the numbers $x_2^{(0)}, \cdots, x_n^{(0)}$ for $x_2, x_3, \cdots, x_n$ into the first equation of (14) and, calculating x_1, we obtain a new $x_1^{(1)}$. We substitute the numbers $x_1^{(0)}, x_3^{(0)}, \cdots, x_n^{(0)}$ for $x_1, x_3, \cdots, x_n$ into the second equation and

calculate $x_2^{(1)}$. We proceed in this way until we calculate $x_n^{(1)}$ from the last equation. The whole procedure is then repeated, that is, we substitute $x_i^{(1)}$ and calculate $x_i^{(2)}$, and so forth.

Example 1 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)}\right) = (0,\ 0,\ 0),$$

solve the following system of equations correct to three decimal places

$$\begin{array}{rcrcrcr} 5x_1 & - & x_2 & + & 3x_3 & = & -2 \\ x_1 & + & 5x_2 & - & 2x_3 & = & 10 \\ 2x_1 & - & 4x_2 & + & 10x_3 & = & 6 \end{array} \qquad (16)$$

Solution

Step 1 Write the given system in the from

$$x_1 = \frac{1}{5}(\qquad x_2 - 3x_3 - 2)$$

$$x_2 = \frac{1}{5}(-x_1 \qquad + 2x_3 + 10)$$

$$x_2 = \frac{1}{10}(-2x_1 + 4x_2 \qquad + 6)$$

or,

$$x_1 = (\qquad 0.2x_2 - 0.6x_3 - 0.4)$$

$$x_2 = (-0.2x_1 \qquad + 0.4x_3 + 2.0)$$

$$x_2 = (-0.2x_1 + 0.4x_2 \qquad + 0.6)$$

Therefore,

$$
\begin{aligned}
x_1^{(k+1)} &= \left(\quad 0.2x_2^{(k)} - 0.6x_3^{(k)} - 0.4 \right) \\
x_2^{(k+1)} &= \left(-0.2x_1^{(k)} + 0.4x_3^{(k)} + 2.0 \right) \\
x_3^{(k+1)} &= \left(-0.2x_1^{(k)} + 0.4x_2^{(k)} + 0.6 \right)
\end{aligned} \tag{17}
$$

Put $k = 0,1,2,\cdots$ in (17) to get the approximate solution of the given system of equations.

Step 2 Starting with the initial approximation

$$
\begin{aligned}
x_1^{(0)} &= 0 \\
x_1^{(0)} &= 0 \\
x_1^{(0)} &= 0
\end{aligned} \tag{18}
$$

Step 3 The first approximation is obtained by substituting $k = 0$ into (17) and using (18) to get

$$
\begin{aligned}
x_1^{(1)} &= \left(\quad 0.2x_2^{(0)} - 0.6x_3^{(0)} - 0.4 \right) = -0.4 \\
x_2^{(1)} &= \left(-0.2x_1^{(0)} + 0.4x_3^{(0)} + 2.0 \right) = 2.0 \\
x_3^{(1)} &= \left(-0.2x_1^{(0)} + 0.4x_2^{(0)} + 0.6 \right) = 0.6
\end{aligned}
$$

For the next approximation $k = 1$, we obtain

$$
\begin{aligned}
x_1^{(2)} &= \left(\quad 0.2x_2^{(1)} - 0.6x_3^{(1)} - 0.4 \right) = -0.36 \\
x_2^{(2)} &= \left(-0.2x_1^{(1)} + 0.4x_3^{(1)} + 2.0 \right) = 2.032 \\
x_3^{(2)} &= \left(-0.2x_1^{(1)} + 0.4x_2^{(1)} + 0.6 \right) = 1.48
\end{aligned}
$$

The third iteration is

$$x_1^{(3)} = \left(\quad 0.2x_2^{(2)} - 0.6x_3^{(2)} - 0.4\right) = -0.824$$
$$x_2^{(3)} = \left(- 0.2x_1^{(2)} + 0.4x_3^{(2)} + 2.0\right) = 2.664$$
$$x_3^{(3)} = \left(- 0.2x_1^{(2)} + 0.4x_2^{(2)} + 0.6\right) = 1.600$$

and the fourth iteration gives

$$x_1^{(4)} = \left(\quad 0.2x_2^{(3)} - 0.6x_3^{(3)} - 0.4\right) = -0.8272$$
$$x_2^{(4)} = \left(- 0.2x_1^{(3)} + 0.4x_3^{(3)} + 2.0\right) = 2.8048$$
$$x_3^{(4)} = \left(- 0.2x_1^{(3)} + 0.4x_2^{(3)} + 0.6\right) = 1.8304$$

the fifth iteration is

$$x_1^{(5)} = \left(\quad 0.2x_2^{(4)} - 0.6x_3^{(4)} - 0.4\right) = -0.9372$$
$$x_2^{(5)} = \left(- 0.2x_1^{(4)} + 0.4x_3^{(4)} + 2.0\right) = 2.8976$$
$$x_3^{(5)} = \left(- 0.2x_1^{(4)} + 0.4x_2^{(4)} + 0.6\right) = 1.88736$$

$$\vdots$$

Proceeding up to the last two iterations

$$x_1^{(25)} = -1$$
$$x_1^{(25)} = 2.999999,$$
$$x_1^{(25)} = 1.999999$$

and

$$x_1^{(26)} = -1$$
$$x_1^{(26)} = 3$$
$$x_1^{(26)} = 2$$

After we reach at the exact solution, which represents the fixed point of the given system, then we get the results obtained Table 2.1.

	x_1	x_2	x_3	1
$x_1 =$	0	0.2	-0.6	-0.4
$x_2 =$	-0.2	0	0.4	2
$x_3 =$	-0.2	0.4	0	0.6
$k = 0$	0	0	0	
$k = 1$	-0.4000	2.0000	0.6000	
$k = 2$	-0.3600	2.3200	1.4800	
$k = 3$	-0.8240	2.6640	1.6000	
$k = 4$	-0.8272	2.8048	1.8304	
$k = 5$	-0.9373	2.8976	1.8874	
$k = 6$	-0.9529	2.9424	1.9465	
$k = 7$	-0.9794	2.9692	1.9675	
$k = 8$	-0.9867	2.9829	1.9836	
$k = 9$	-0.9936	2.9908	1.9905	
$k = 10$	-0.9961	2.9949	1.9950	
$k = 11$	-0.9980	2.9972	1.9972	
$k = 12$	-0.9989	2.9985	1.9985	
$k = 13$	-0.9994	2.9992	1.9992	
$k = 14$	-0.9997	2.9996	1.9996	
$k = 15$	-0.9998	2.9998	1.9998	
$k = 16$	-0.9999	2.9999	1.9999	
$k = 17$	-1.0000	2.9999	1.9999	
$k = 18$	-1.0000	3.0000	2.0000	
$k = 19$	-1.0000	3.0000	2.0000	

Table 2.1. Example 1

Example 2 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)}, x_4^{(0)}\right) = (1,\ 1,\ 3,\ 5),$$

use Gauss-Jacobi iteration method to solve the following system of equations correct to five decimal places. (All computations have to be executed to six decimal places)

$$\begin{array}{ccccccccc} 5x_1 & + & x_2 & - & 2x_3 & + & x_4 & = & 5 \\ x_1 & + & 10x_2 & - & x_3 & + & 2x_4 & = & 15 \\ x_1 & - & 2x_2 & + & 8x_3 & - & x_4 & = & 20 \\ x_1 & - & 2x_2 & + & x_3 & + & 4x_4 & = & 25 \end{array}$$

Solution

	x_1	x_2	x_3	x_4	1
$x_1 =$	0	-0.2	0.4	-0.2	1
$x_2 =$	-0.1	0	0.1	-0.2	1.5
$x_3 =$	-0.125	0.25	0	0.125	2.5
$x_4 =$	-0.25	0.5	-0.25	0	6.25
$k = 0$	1	1	3	5	
$k = 1$	1.00000	0.70000	3.25000	5.75000	
$k = 2$	1.01000	0.57500	3.26875	5.53750	
$k = 3$	1.08500	0.61838	3.20969	5.46781	
$k = 4$	1.06664	0.61891	3.20245	5.48552	
$k = 5$	1.06009	0.61648	3.20709	5.49218	
$k = 6$	1.06110	0.61626	3.20813	5.49145	
$k = 7$	1.06171	0.61641	3.20786	5.49082	
$k = 8$	1.06170	0.61645	3.20774	5.49081	
$k = 9$	1.06164	0.61644	3.20775	5.49087	
$k = 10$	1.06164	0.61644	3.20776	5.49087	
$k = 11$	1.06164	0.61644	3.20776	5.49087	

Table 2.2. Example 2

From the table, the solution of the given system is

$$(1.06164, 0.61644, 3.20776, 5.49087)^t$$

2.2 Gauss-Seidel iterative method

The difference between the Jacobi and Gauss-Seidel methods is that in the latter, as each components of x_i is computed, we use it immediately in the iteration. Therefore, in solving the system

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

we calculate $x_1^{(1)}$ from the first equation as in the Jacobi method. We then substitute this new $x_1^{(1)}$ and not the "old" $x_1^{(0)}$ into the second equation and calculate $x_2^{(1)}$. Into the third equation we already substitute new $x_1^{(1)}$, $x_2^{(1)}$, and old $x_4^{(0)}, x_5^{(0)}, \cdots, x_n^{(0)}$ and calculate $x_3^{(1)}$ etc. That is,

$$\begin{array}{ccccccccc}
x_1^{(k+1)} & = & 0 & -\dfrac{a_{12}}{a_{11}}x_2^{(k)} & \cdots & -\dfrac{a_{1n}}{a_{11}}x_n^{(k)} & + & \dfrac{b_1}{a_{11}} \\
x_2^{(k+1)} & = & -\dfrac{a_{21}}{a_{22}}x_1^{(k+1)} & 0 & \cdots & -\dfrac{a_{2n}}{a_{22}}x_n^{(k)} & + & \dfrac{b_2}{a_{22}}, \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_n^{(k+1)} & = & -\dfrac{a_{n1}}{a_{nn}}x_1^{(k+1)} & -\dfrac{a_{n2}}{a_{nn}}x_2^{(k+1)} & \cdots & 0 & + & \dfrac{b_n}{a_{nn}}
\end{array}$$

or,

$$\boxed{x_i^{(k+1)} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right), \quad i = 1,2,\cdots,n, \quad k = 0,1,\cdots} \tag{1}$$

To write this method in matrix form, multiply both sides of (1) by a_{ii} and collect all kth iterate terms to give

$$a_{ii}x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i \tag{2}$$

for each $i = 1,2,\cdots,n$. Writing all n equations gives

$$\begin{array}{llllll}
a_{11}x_1^{(k)} & 0 & & & = & -a_{12}x_2^{(k-1)} - a_{12}x_3^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1, \\
a_{21}x_1^{(k)} & +a_{22}x_2^{(k)} & & & = & -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2, \\
\vdots & \vdots & & & & \vdots \\
a_{n1}x_1^{(k)} & +a_{n2}x_2^{(k)} & \cdots & +a_{nn}x_n^{(k)} & = & +b_n,
\end{array}$$

and it follows that, in matrix form, Gauss-Seidel method can be represented by

$$(D - L)\mathbf{x}^{(k)} = U\,\mathbf{x}^{(k)} + \mathbf{b},$$

or,

$$\boxed{\mathbf{x}^{(k+1)} = (D - L)^{-1}U\,\mathbf{x}^{(k)} + (D - L)^{-1}\mathbf{b}, \quad k = 0,1,\cdots} \qquad (3)$$

For the lower-triangular matrix $(D - L)$ to be non-singular, it is necessary and sufficient that $a_{ii} \neq 0$ for each $i = 1,2,\cdots,n$.

Example 3 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)}\right) = (0,\ 0,\ 0),$$

use Gauss-Seidel method to solve the following system of equations

$$\begin{array}{rcrcrcr}
5x_1 & - & x_2 & + & 3x_3 & = & -2 \\
x_1 & + & 5x_2 & - & 2x_3 & = & 10, \\
2x_1 & - & 4x_2 & + & 10x_3 & = & 6
\end{array}$$

correct to three decimal places

Solution Step 1 Write the given system in the from

$$x_1 = \frac{1}{5}(\quad x_2 - 3x_3 - 2)$$
$$x_2 = \frac{1}{5}(-x_1 + 2x_3 + 10)$$
$$x_2 = \frac{1}{10}(-2x_1 + 4x_2 + 6)$$

Therefore,

$$\boxed{\begin{aligned} x_1^{(k+1)} &= (\quad 0.2x_2^{(k)} - 0.6x_3^{(k)} - 0.4) \\ x_2^{(k+1)} &= (-0.2x_1^{(k+1)} + 0.4x_3^{(k)} + 2.0) \\ x_3^{(k+1)} &= (-0.2x_1^{(k+1)} + 0.4x_2^{(k+1)} + 0.6) \end{aligned}} \quad (4)$$

Put $k = 0,1,\cdots$ in this equation in order to obtain the approximate solution of the given system of equations.

Step 2 Starting with the initial approximation

$$\begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Step 3 The first approximation is obtained by substituting $k = 0$ into (4) to get

$$x_1^{(1)} = (\quad 0.2x_2^{(0)} - 0.6x_3^{(0)} - 0.4) = -0.4$$
$$x_2^{(1)} = (-0.2x_1^{(1)} + 0.4x_3^{(0)} + 2.0) = 2.08$$
$$x_3^{(1)} = (-0.2x_1^{(1)} + 0.4x_2^{(1)} + 0.6) = 1.512$$

For the next approximation $k = 1$, we obtain

$$x_1^{(2)} = (\quad 0.2x_2^{(1)} - 0.6x_3^{(1)} - 0.4) = -0.8912$$
$$x_2^{(2)} = (-0.2x_1^{(2)} + 0.4x_3^{(1)} + 2.0) = 2.7830$$
$$x_3^{(2)} = (-0.2x_1^{(2)} + 0.4x_2^{(2)} + 0.6) = 1.8914$$

The third and fourth iterations are

$$x_1^{(3)} = \left(\quad 0.2x_2^{(2)} - 0.6x_3^{(2)} - 0.4\right) = -0.9782$$

$$x_2^{(3)} = \left(-0.2x_1^{(3)} + 0.4x_3^{(2)} + 2.0\right) = 2.9552$$

$$x_3^{(3)} = \left(-0.2x_1^{(3)} + 0.4x_2^{(3)} + 0.6\right) = 1.9765$$

For the next approximation $k = 1$, we obtain

$$x_1^{(4)} = \left(\quad 0.2x_2^{(3)} - 0.6x_3^{(3)} - 0.4\right) = -0.9955$$

$$x_2^{(4)} = \left(-0.2x_1^{(4)} + 0.4x_3^{(3)} + 2.0\right) = 2.9897$$

$$x_3^{(4)} = \left(-0.2x_1^{(4)} + 0.4x_2^{(4)} + 0.6\right) = 1.9950$$

$$\vdots$$

Proceeding in this manner, we obtain Table 2.2.

	x_1	x_2	x_3	1
$x_1 =$	0	0.2	-0.6	-0.6
$x_2 =$	-0.2	0	0.4	2.0
$x_3 =$	-0.2	0.4	0	0.6
$k = 0$	0	0	0	
$k = 1$	-0.4000	2.0800	1.5120	
$k = 2$	-0.8912	2.7830	1.8914	
$k = 3$	-0.9782	2.9522	1.9765	
$k = 4$	-0.9955	2.9897	1.9950	
$k = 5$	-0.9991	2.9978	1.9989	
$k = 6$	-0.9998	2.9995	1.9998	
$k = 7$	-1.0000	2.9999	2.0000	
$k = 8$	-1.0000	3.0000	2.0000	
$k = 9$	-1.0000	3.0000	2.0000	

Table 2.3. Example 3

Note that the exact solution of the given system is obtained after 19 iterations in Jacobi's iterative method whereas it is obtained after 9

iterations only in Gauss-Seidel method. Anyway, an approximate solution correct to three decimal places is obtained after 6 iterations in Gauss-Seidel method and after 14 iterations if we use Jacobi's iteration method.

Example 4 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right) = (1,\ 4,\ 2,\ 0),$$

solve the following system of equations correct to three decimal places

$$\begin{array}{rcrcrcrcr} 4x_1 & + & x_2 & - & x_3 & + & x_4 & = & 4 \\ x_1 & + & 8x_2 & - & x_3 & + & x_4 & = & 32 \\ 2x_1 & - & 2x_2 & - & 5x_3 & - & 3x_4 & = & 10 \\ 2x_1 & - & x_2 & + & x_3 & - & 5x_4 & = & 30 \end{array}$$

Solution All computations are executed to six decimal places.

	x_1	x_2	x_3	x_4	1
$x_1 =$	0	-0.25	0.25	-0.25	1
$x_2 =$	-0.125	0	0.125	-0.125	4
$x_3 =$	0.4	-0.4	0	-0.6	-2
$x_4 =$	0.4	-0.2	0.2	0	-6
$k = 0$	1	4	2	0	
$k = 1$	0.500000	4.187500	1.000000	-6.437500	
$k = 2$	1.812500	4.703125	2.609375	-5.693750	
$k = 3$	1.900000	4.800391	2.423438	-5.715391	
$k = 4$	1.834610	4.788027	2.428848	-5.737992	
$k = 5$	1.844703	4.790267	2.434498	-5.733273	
$k = 6$	1.844376	4.790424	2.433318	-5.733671	
$k = 7$	1.844141	4.790356	2.433418	-5.733731	

Table 2.4. Example 4

From the table, the solution of the given system after 7 iterations is

$$(1.844,\ \ 4.790,\ \ 2.433, -5.733)^t$$

2.3 Successive over relaxation method (SOR)

To solve the system

$$A\mathbf{x} = \mathbf{b}, \tag{1}$$

we may use instead of Gauss-Seidel iteration a more general iterative method depending on one real parameter ω, called the ***relaxation parameter***. Recall, however, that in the Gauss-Seidel method, $x_i^{(k+1)}$ is chosen to be

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right), \tag{2}$$
$$i = 1,2,\cdots,n, \quad k = 0,1,\cdots$$

or,

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)} + b_i\right),$$
$$i = 1,2,\cdots,n \tag{3}$$

Equation (3) describes $x_i^{(k+1)}$ as being obtained by adding a correction term to the iteration $x_i^{(k)}$. Modifying the Gauss-Seidel procedure as given by equation (3) by adding a multiple of the correction term to $x_i^{(k)}$ and obtain $x_i^{(k+1)}$ from the equation

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)} + b_i\right),$$
$$i = 1,2,\cdots,n \tag{4}$$

For choices of $0 < \omega < 1$, the procedures (4) are called ***under-relaxation methods*** and can be used to obtain convergence of some systems that are not convergent by the Gauss-Seidel method. For

choices of $\omega > 1$, the procedures are called ***over-relaxation methods*** which can be used to accelerate the convergence for systems that are convergent by the Gauss-Seidel method. These methods are often abbreviated **SOR** for ***Successively-Over-Relaxation***, and are particularly useful for solving the linear systems that occur in the numerical solution of certain partial differential equations.

Before illustrating the advantages of the SOR method, we note that equation (4) can be reformulated for calculation purposes to read

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right) \qquad i = 1,2,\cdots,n \qquad (5)$$

To determine the matrix form of the SOR method we rewrite (5) as

$$a_{ii}x_i^{(k+1)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = (1-\omega)a_{ii}x_i^{(k)} - \omega\sum_{j=\,i+1}^{n} a_{ij}x_j^{(k)} + \omega b_i, \qquad i = 1,2,\cdots,n$$

so

$$(D-\omega L)\mathbf{x}^{(k+1)} = [(1-\omega)D + \omega U]\mathbf{x}^{(k)} + \omega\mathbf{b},$$

or

$$\mathbf{x}^{(k+1)} = (D-\omega L)^{-1}[(1-\omega)D + \omega U]\mathbf{x}^{(k)} + \omega(D-\omega L)^{-1}\mathbf{b} \qquad (6)$$

Remark As we note from equation (5), to construct the tabular formula for the SOR method, we use the formula of Gauss-Seidel with the following differences:

1. The entries of the table headlines are multiplied by ω,

2. The zeros of the main diagonal are replaced by $(1-\omega)$,

and after that the iterations are proceeded in the same way as in Gauss-Seidel iterations.

Example 5 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)}\right) = (0,\ 0,\ 0),$$

use SOR method with $\omega = 1.05$ to solve the following system of equations

$$\begin{array}{rcrcrcr} 5x_1 & - & x_2 & + & 3x_3 & = & -2 \\ x_1 & + & 5x_2 & - & 2x_3 & = & 10, \\ 2x_1 & - & 4x_2 & + & 10x_3 & = & 6 \end{array} \qquad (7)$$

correct to three decimal places.

Solution With the aid of the equations

$$x_1 = \frac{1}{5}(\qquad x_2 - 3x_3 - 2)$$

$$x_2 = \frac{1}{5}(-x_1 \qquad + 2x_3 + 10)$$

$$x_3 = \frac{1}{10}(-2x_1 + 4x_2 \qquad + 6)$$

and according to the above remark the equations for the SOR method with $\omega = 1.05$ are:

$$x_1^{(k+1)} = \left(-0.05x_1^{(k)} - 0.21x_2^{(k)} - 0.63x_3^{(k)} - 0.42\right)$$
$$x_2^{(k+1)} = \left(-0.21x_1^{(k+1)} - 0.05x_2^{(k)} + 0.42x_3^{(k)} + 2.1\right)$$
$$x_3^{(k+1)} = \left(-0.21x_1^{(k+1)} + 0.42x_2^{(k+1)} + 0.05x_3^{(k)} + 0.63\right) \tag{8}$$

Starting with the initial approximation

$$\begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The first approximation is obtained by substituting $k = 0$ into (8) to get

$$x_1^{(1)} = \left(-0.05x_1^{(0)} - 0.21x_2^{(0)} - 0.63x_3^{(0)} - 0.42\right) = -0.42$$
$$x_2^{(1)} = \left(-0.21x_1^{(1)} - 0.05x_2^{(0)} + 0.42x_3^{(0)} + 2.1\right) = 2.1882$$
$$x_3^{(1)} = \left(-0.21x_1^{(1)} + 0.42x_2^{(1)} + 0.05x_3^{(0)} + 0.63\right) = 1.637244$$

For the next approximation $k = 1$, we obtain

$$x_1^{(2)} = \left(-0.05x_1^{(1)} - 0.21x_2^{(1)} - 0.63x_3^{(1)} - 0.42\right) = -0.970942$$
$$x_2^{(2)} = \left(-0.21x_1^{(2)} - 0.05x_2^{(1)} + 0.42x_3^{(1)} + 2.1\right) = 2.88213$$
$$x_3^{(2)} = \left(-0.21x_1^{(2)} + 0.42x_2^{(2)} + 0.05x_3^{(1)} + 0.63\right) = 1.96253$$

The third approximation is

$$x_1^{(3)} = \left(-0.05x_1^{(2)} - 0.21x_2^{(2)} - 0.63x_3^{(2)} - 0.42\right) = -1.0026$$
$$x_2^{(3)} = \left(-0.21x_1^{(3)} - 0.05x_2^{(2)} + 0.42x_3^{(2)} + 2.1\right) = 2.990702$$
$$x_3^{(3)} = \left(-0.21x_1^{(3)} + 0.42x_2^{(3)} + 0.05x_3^{(2)} + 0.63\right) = 1.998514$$
$$\vdots$$

so that we construct the tabular formula

x_1	x_2	x_3	1

$x_1 =$	-0.05	0.21	-0.63	-0.42
$x_2 =$	-0.21	-0.05	0.42	2.1
$x_3 =$	-0.21	0.42	-0.05	0.63
$k = 0$	0	0	0	
$k = 1$	-0.420000	2.188200	1.637244	
$k = 2$	-0.970942	2.882130	1.962530	
$k = 3$	-1.002600	2.990702	1.998514	
$k = 4$	-1.000886	3.000027	2.000272	
$k = 5$	-1.000121	3.000138	2.000070	
$k = 6$	-1.000009	3.000024	2.000008	
$k = 7$	-1.000000	3.000002	2.000000	
$k = 8$	-1.000000	3.000000	2.000000	

Table 2.5. Example 5

Note that the approximate solution correct to three decimal places is obtained after 5 iterations in SOR method, after 6 iterations in Gauss-Seidel method and after 14 iterations if we use Jacobi's iteration method.

Example 6 Starting with the initial values

$$\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right)^t = (4,\ 3,\ 2,\ 1)^t,$$

solve the following system of equations correct to four decimal places.

$$\begin{aligned} 10x_1 - 2x_2 + 2x_3 + x_4 &= 60 \\ 4x_1 + 10x_2 - 2x_3 + 2x_4 &= 60 \\ 2x_1 - 2x_2 + 10x_3 - 3x_4 &= 50 \\ 2x_1 + 4x_2 + 3x_3 - 10x_4 &= 30 \end{aligned}$$

by the use of:

(**i**) Gauss-Jacobi method

(**ii**) Gauss-Seidel method

(**iii**) Successive over relation method with $\omega = 1.04$.

Solution All computations are executed to six decimal places.

(i) Gauss-Jacobi method

	x_1	x_2	x_3	x_4	1
$x_1 =$	0	0.2	-0.2	-0.1	6
$x_2 =$	-0.4	0	0.2	-0.2	6
$x_3 =$	-0.2	0.2	0	0.3	5
$x_4 =$	0.2	0.4	0.3	0	-3
$k = 0$	4	3	2	1	
$k = 1$	6.100000	4.600000	5.100000	-0.400000	
$k = 2$	5.940000	4.660000	4.580000	1.590000	
$k = 3$	5.857000	4.222000	5.221000	1.426000	
$k = 4$	5.657600	4.416200	5.100800	1.426500	
$k = 5$	5.720430	4.471820	5.179670	1.428240	
$k = 6$	5.715606	4.462114	5.178750	1.486715	
$k = 7$	5.708001	4.452165	5.195316	1.481592	
$k = 8$	5.703211	4.459544	5.193310	1.481061	
$k = 9$	5.705141	4.461165	5.195585	1.482453	
$k = 10$	5.704871	4.460570	5.195941	1.484170	
$k = 11$	5.704509	4.460406	5.196391	1.483985	
$k = 12$	5.704405	4.460678	5.196375	1.483982	
$k = 13$	5.704462	4.460717	5.196449	1.484065	
$k = 14$	5.704447	4.460692	5.196471	1.484114	
$k = 15$	5.704433	4.460693	5.196483	1.484108	
$k = 16$	5.704431	4.460702	5.196484	1.484109	
$k = 17$	5.704433	4.460703	5.196487	1.484112	
$k = 18$	5.704432	4.460702	5.196488	1.484114	

(ii) Gauss-Seidel method

	x_1	x_2	x_3	x_4	1

$x_1 =$	0	0.2	-0.2	-0.1	6
$x_2 =$	-0.4	0	0.2	-0.2	6
$x_3 =$	-0.2	0.2	0	0.3	5
$x_4 =$	0.2	0.4	0.3	0	-3
$k = 0$	4	3	2	1	
$k = 1$	6.100000	3.760000	4.832000	1.173600	
$k = 2$	5.668240	4.464384	5.111309	1.452794	
$k = 3$	5.725336	4.441569	5.179085	1.475420	
$k = 4$	5.704955	4.458751	5.193385	1.482507	
$k = 5$	5.704823	4.460246	5.195837	1.483814	
$k = 6$	5.704500	4.460605	5.196365	1.484052	
$k = 7$	5.704443	4.460685	5.196464	1.484102	
$k = 8$	5.704434	4.460699	5.196484	1.484112	
$k = 9$	5.704432	4.460702	5.196488	1.484114	
$k = 10$	5.704431	4.460702	5.196488	1.484113	

(iii) SOR method

	x_1	x_2	x_3	x_4	1
$x_1 =$	-0.04	0.208	-0.208	-0.104	6.24
$x_2 =$	-0.416	-0.04	0.208	-0.208	6.24
$x_3 =$	-0.208	0.208	-0.04	0.312	5.2
$x_4 =$	0.208	0.416	0.312	-0.04	-3.12
$k = 0$	4	3	2	1	
$k = 1$	6.184000	3.755456	4.926863	1.225723	
$k = 2$	5.621512	4.521070	5.156459	1.489826	
$k = 3$	5.728037	4.438953	5.190438	1.477860	
$k = 4$	5.700872	4.463095	5.196017	1.484472	
$k = 5$	5.705132	4.460143	5.196357	1.483971	
$k = 6$	5.704329	4.460770	5.196484	1.484125	
$k = 7$	5.704449	4.460689	5.196486	1.484111	
$k = 8$	5.704429	4.460704	5.196488	1.484114	
$k = 9$	5.704432	4.460702	5.196488	1.484114	
$k = 10$	5.704431	4.460702	5.196488	1.484113	

From the table, the solution of the given system correct to four decimal places is

$$(5.70443,\ \ 4.46070,\ \ 5.19648, 1.48411)^t,$$

which is obtained after 18, 10, and 10 iterations on using Gauss-Jacobi method, Gauss-Seidel method and SOR method; respectively.

Example 7 Starting with the initial values

$$\left(x_1^{(0)},\ \ x_2^{(0)},\ \ x_3^{(0)}\right)^t = (2,\ \ 2, 2)^t,$$

solve the following system of equations correct to four decimal places.

$$\begin{aligned} 5x_1 + x_2 + x_3 &= 15 \\ 2x_1 + 8x_2 - 3x_3 &= 4 \\ x_1 - 2x_2 + 8x_3 &= 10 \end{aligned}$$

by the use of:

$(\boldsymbol{i})$ Gauss-Jacobi method

$(\boldsymbol{ii})$ Gauss-Seidel method

$(\boldsymbol{iii})$ Successive over relation method with $\omega = 1.02$.

$(\boldsymbol{iv})$ Successive over relation method with $\omega = 1.2$.

Solution All computations are executed to six decimal places. To construct the tabular formula, we write the system as

$$\begin{aligned} x_1 &= (\qquad\quad - 0.2x_2 - 0.2x_3 + 3) \\ x_2 &= (-0.25x_1 \qquad\quad + 0.375x_3 + 0.5) \\ x_3 &= (-0.125x_1 + 0.25x_2 \qquad\quad + 1.25) \end{aligned}$$

for Jacobi' iterative method and Gauss-Seidel iterative method, whereas for SOR method we use the system form

$$x_1 = ((1-\omega)x_1 - 0.2\omega x_2 - 0.2\omega x_3 + 3\omega)$$

$$x_2 = (-0.25\omega x_1 \quad + \quad (1-\omega)x_2 \quad + \quad 0.375\omega x_3 \quad + \quad 0.5\omega)$$

$$x_3 = (-0.125\omega x_1 \quad + \quad 0.25\omega x_2 \quad + \quad (1-\omega)x_3 \quad + \quad 1.25\omega)$$

So that we apply the different iterative methods as follows.

(i) Jacobi' iterative method

	x_1	x_2	x_3	1
$x_1 =$	0	-0.2	-0.2	3
$x_2 =$	-0.25	0	0.375	0.5
$x_3 =$	-0.125	0.25	0	1.25
$k = 0$	**2**	**2**	**2**	
$k = 1$	2.200000	0.750000	1.500000	
$k = 2$	2.550000	0.512500	1.162500	
$k = 3$	2.665000	0.298438	1.059375	
$k = 4$	2.728437	0.231016	0.991485	
$k = 5$	2.755500	0.189698	0.966699	
$k = 6$	2.768721	0.173637	0.952987	
$k = 7$	2.774675	0.165190	0.947319	
$k = 8$	2.777498	0.161576	0.944463	
$k = 9$	2.778792	0.159799	0.943207	
$k = 10$	2.779399	0.159005	0.942601	
$k = 11$	2.779679	0.158626	0.942326	
$k = 12$	2.779810	0.158453	0.942197	
$k = 13$	2.779870	0.158371	0.942137	
$k = 14$	2.779898	0.158334	0.942109	
$k = 15$	2.779911	0.158316	0.942096	
$k = 16$	2.779918	0.158308	0.942090	
$k = 17$	2.779920	0.158304	0.942087	
$k = 18$	2.779922	0.158303	0.942086	
$k = 19$	2.779922	0.158302	0.942086	

(ii) Gauss-Seidel iterative method

	x_1	x_2	x_3	1
$x_1 =$	0	-0.2	-0.2	3
$x_2 =$	-0.25	0	0.375	0.5
$x_3 =$	-0.4	0.4	0	-4
$k = 0$	2	2	2	
$k = 1$	2.200000	0.700000	1.150000	
$k = 2$	2.630000	0.273750	0.989688	
$k = 3$	2.747312	0.184305	0.952662	
$k = 4$	2.772607	0.164097	0.944448	
$k = 5$	2.778291	0.159595	0.942612	
$k = 6$	2.779559	0.158590	0.942203	
$k = 7$	2.779841	0.158366	0.942111	
$k = 8$	2.779905	0.158315	0.942091	
$k = 9$	2.779919	0.158304	0.942086	
$k = 10$	2.779922	0.158302	0.942085	
$k = 11$	2.779923	0.158301	0.942085	
$k = 12$	2.779923	0.158301	0.942085	

(iii) SOR method with $\omega = 1.02$

	x_1	x_2	x_3	1
$x_1 =$	-0.02	-0.204	-0.204	3.06
$x_2 =$	-0.255	-0.02	0.3825	0.51
$x_3 =$	-0.1275	0.255	-0.02	1.275
$k = 0$	2	2	2	
$k = 1$	2.204000	0.672980	1.125600	
$k = 2$	2.649010	0.251585	0.978893	
$k = 3$	2.756002	0.176614	0.949068	
$k = 4$	2.775241	0.161800	0.943434	
$k = 5$	2.779027	0.158976	0.942344	
$k = 6$	2.779750	0.158431	0.942135	
$k = 7$	2.779890	0.158326	0.942094	
$k = 8$	2.779917	0.158306	0.942087	
$k = 9$	2.779921	0.158302	0.942085	
$k = 10$	2.779923	0.158301	0.942085	
$k = 11$	2.779923	0.158301	0.942085	

(iii) SOR method with $\omega = 1.2$

	x_1	x_2	x_3	1
$x_1 =$	-0.2	-0.24	-0.24	3.6
$x_2 =$	-0.3	-0.2	0.45	0.6
$x_3 =$	-0.15	0.3	-0.2	1.5
$k = 0$	2	2	2	
$k = 1$	2.240000	0.428000	0.892400	
$k = 2$	2.835104	0.065449	0.915889	
$k = 3$	2.797458	0.159823	0.945150	
$k = 4$	2.775315	0.160758	0.942900	
$k = 5$	2.780059	0.158136	0.941852	
$k = 6$	2.779991	0.158209	0.942094	
$k = 7$	2.779929	0.158322	0.942088	
$k = 8$	2.779916	0.158300	0.942085	
$k = 9$	2.779924	0.158301	0.942085	
$k = 10$	2.779923	0.158301	0.942085	
$k = 11$	2.779923	0.158301	0.942085	

Note that an approximate solution correct to four decimal places is obtained after 14 iterations in Jacobi's iterative method while Gauss-Seidel method gives this solution after 9 iterations. In the SOR method the approximate solution is obtained after 7 iterations.

Exercise 2.1

In problems **1-5**, solve the linear systems of equations by the use of

(**i**) Gauss-Jacobi method,

(**ii**) Gauss-Seidel method,

(**iii**) Successive over relation method with $\omega = 1.02$.

1. $\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)}\right)^t = (1,1,1)^t$

$$\begin{array}{rcrcrcr} 4x_1 & - & 2x_2 & + & x_3 & = & 12 \\ 3x_1 & + & 10x_2 & - & 3x_3 & = & 4 \\ x_1 & - & 2x_2 & + & 5x_3 & = & 4 \end{array}$$

[Answer: $(2.775510, -0.408163, 081633)^t$]

2. $\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right) = (2,\ 2,\ 5,\ 5)$

$$\begin{array}{rcrcrcrcr} 10x_1 & + & x_2 & - & 2x_3 & + & x_4 & = & 10 \\ x_1 & - & 10x_2 & - & x_3 & + & x_4 & = & -20 \\ 2x_1 & - & 2x_2 & + & 10x_3 & - & 4x_4 & = & 50 \\ x_1 & - & x_2 & + & x_3 & - & 10x_4 & = & -40 \end{array}$$

3. $\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right) = (5,\ 5,\ 1,\ 1)$

$$\begin{array}{rcrcrcrcl} 10x_1 & - & 2x_2 & + & 2x_3 & + & x_4 & = & 70 \\ 4x_1 & + & 10x_2 & - & 2x_3 & + & 2x_4 & = & 40 \\ 2x_1 & - & 2x_2 & + & 5x_3 & - & x_4 & = & 10 \\ 2x_1 & - & x_2 & + & 3x_3 & - & 10x_4 & = & 30 \end{array}$$

[Answer: $(7.60889, 1.13994, -0.96385, -1.88137)^t$]

4. $\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right) = (2,\ 1,\ 2, -1)$

$$\begin{array}{rcrcrcrcr} 10x_1 & + & x_2 & - & 2x_3 & + & x_4 & = & 50 \\ x_1 & - & 8x_2 & - & x_3 & + & x_4 & = & 32 \\ x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 30 \\ 2x_1 & - & x_2 & + & x_3 & - & 5x_4 & = & 16 \end{array}$$

5. Starting with $\left(x_1^{(0)},\ x_2^{(0)},\ x_3^{(0)},\ x_4^{(0)}\right) = (1,\ 2,\ 1,\ 2)$,

solve the following system of equations correct to three decimal places

$$\begin{array}{rcrcrcrcr} 4x_1 & + & x_2 & - & x_3 & + & x_4 & = & 4 \\ x_1 & + & 8x_2 & - & x_3 & + & x_4 & = & 32 \\ 2x_1 & - & 2x_2 & - & 10x_3 & - & 3x_4 & = & 10 \\ 2x_1 & - & x_2 & + & x_3 & - & 8x_4 & = & 36 \end{array}$$

[Answer: $(1.03684, 4.44436, -0.23383, -4.82556)^t$]

3. Interpolation by Polynomials

Interpolation is the process of **"*reading between lines*"** of a table or of fitting a smooth curve to a limited set of data. We take it up here for a number of reasons that are commonly important for engineers and scientists. One of which is that interpolation is frequently used to estimate the values at intermediate points from tabulated values. A more important application is to numerical differentiation and integration of the underlying function.

In this chapter we examine the question: "Given values of an unknown function corresponding to certain values of *x*, what is the behavior of the function?" Our purpose is to determine the behavior of the function as evidenced by the sample of data given by the table:

x	x_0	x_1	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n
$y = f(x)$	y_0	y_1	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

We wish to approximate other values of the function at values of *x* not tabulated and to estimate the integral of $f(x)$ and its derivative. To do that we construct a polynomial $P_n(x)$ of degree n that ***interpolates*** $f(x)$ at the points $x_r,\ r = 0,1,\cdots,n$ that is, whose values at the points coincide with the values of the given function. This is expressed mathematically as

$$P_n(x_r) = f(x_r), \quad r = 0,1,\cdots,n$$

Since $f(x)$ and $P_n(x)$ have the same values at tabulated values of x, it perhaps reasonable to assume they will be nearly the same at intermediate x −values. This assumption is the basic for the use of

$P_n(x)$ as an interpolating polynomial. We emphasize that $f(x)$ and $P_n(x)$ will, in general, not be the same function. Hence there is some error to be expected in the estimate from such interpolation.

This chapter is devoted to construct such interpolation polynomial. When the polynomial is of the first degree, this leads to the familiar ***linear interpolation*** and if it is of the second degree we obtain the ***quadratic interpolation***. We will be interested in polynomials of degree higher than the first and second degrees using the $(n+1)$ points of the given function. These points are called ***mesh points*** or ***knots***.

Interpolating in tables of data was once a very important topic. Today, with computers and calculators that calculate values for most functions at electronic speeds, we look up values from published tables less frequently. A major reason for discussing this subject, apart from its historical value as mentioned before, is to lay the groundwork for numerical integration and differentiation in addition to determine the intermediate values.

Using the notation $f_r = y_r$ to indicate the value of f at the x −value x_r, we start with the following theorem.

Weierstrass Approximation Theorem. Suppose that $\{f(x_r) = y_r\}_{r=0}^{n}$ is a sequence of function values at distinct x −values, then there exists a unique polynomial $P_n(x)$ of degree at most n that interpolates $f(x)$ at $x_0, x_1, \cdots, x_n$.

For example, suppose we want to find the polynomial

$$P_2(x) = ax^2 + bx + c$$

to interpolate the following readings:

x	1	2	3
$f(x)$	6	10	25

We can write three equations involving the unknown coefficients $a, b,$ and c:

When $x = 1$: $a(1)^2 + b(1) + c = 6$

If $x = 2$: $a(2)^2 + b(2) + c = 10$

If $x = 3$: $a(3)^2 + b(3) + c = 25$

Solving these three simultaneous equations gives the required polynomial:

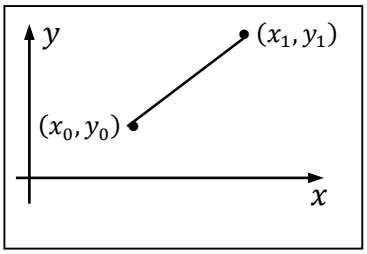
We can then estimate the values of the function at some values of x by substituting for this value in the polynomial.

We seek a better and simpler way of finding such interpolating polynomials. The above procedure is awkward, especially if we have a table of large number of data. So our problem is to find the simplest means of writing the nth degree polynomial that passes through $(n+1)$ pairs of points, $(x_r, f_r), r = 0, 1, \cdots, n$. We begin with the simplest type –***Lagrange interpolation***.

3.1. Lagrange interpolation

To explain the basic idea of Lagrange's interpolating polynomials, let us start with the linear case. The linear polynomial that passes through the distinct points (x_0, y_0) and (x_1, y_1) can be written in the form

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0},$$



or

$$y = y_0 + (x - x_0)\frac{(y_1 - y_0)}{(x_1 - x_0)},$$

or

$$y = y_0 + \frac{(x - x_0)}{(x_1 - x_0)}(y_1) - \frac{(x - x_0)}{(x_1 - x_0)}(y_0)$$

This equation of the line can be written in the form

$$\boxed{P_2(x) = y(x) = \frac{(x - x_1)}{(x_0 - x_1)}(y_0) + \frac{(x - x_0)}{(x_1 - x_0)}(y_1)}$$

When $x = x_0$:

$$P_2(x_0) = 1.(y_0) + 0.(y_1) = y_0,$$

and when $x = x_1$:

$$P_2(x_1) = 0.(y_0) + 1.(y_1) = y_1$$

so $P_2(x)$ interpolates the given data points.

To generalize the concept of linear interpolation to higher polynomial, consider the construction of a polynomial of degree at most *n* that passes through the $(n + 1)$ points:

x	x_0	x_1	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n
$y = f(x)$	y_0	y_1	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

The general form of the nth degree the ***Lagrange polynomial*** take the form

$$P_n(x) = y(x) = \frac{(x-x_1)(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)}(y_0)$$
$$+\frac{(x-x_0)(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}(y_1)$$
$$\vdots$$
$$+\frac{(x-x_0)(x-x_1)\cdots(x-x_{r-1})(x-x_{r+1})\cdots(x-x_n)}{(x_r-x_0)(x_r-x_1)\cdots(x_r-x_{r-1})(x_r-x_{r+1})\cdots(x_r-x_n)}(y_r)$$
$$\vdots$$
$$+\frac{(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}(y_n) \qquad (1)$$

In a compact form, we have

$$P_n(x) = \sum_{r=0}^{n} l_r(x).y_r$$
$$= l_0(x).y_0 + l_1(x).y_1 + \cdots + l_{n-1}(x).y_{n-1} + l_n(x).y_n, \quad (2)$$

with

$$l_r(x) = \prod_{\substack{k=0\\k\neq r}}^{n} \frac{(x-x_k)}{(x_r-x_k)}, \quad r = 0,1,\cdots,n \qquad (3)$$

or equivalently,

$$l_r(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{r-1})(x-x_{r+1})\cdots(x-x_n)}{(x_r-x_0)(x_r-x_1)\cdots(x_r-x_{r-1})(x_r-x_{r+1})\cdots(x_r-x_n)}$$

The definition of Lagrange's polynomials implies

$$l_r(x_r) = 1, \qquad r = 0,1,\cdots,n \qquad (4)$$
$$l_r(x_k) = 0, \qquad \forall\, r \neq k \qquad (5)$$

From equations $(2)-(5)$, we see that

$$P_n(x_r) = f(x_r), \quad r = 0,1,\cdots,n$$

Example 1 By Lagrange's interpolation formula find $y = f(x)$ from the table of readings:

x	3	5	6
$y = f(x)$	6	10	15

Solution Here we have three points, thus there is a unique polynomial unique polynomial $P_2(x)$ of degree at most 2 that interpolates $f(x)$.

To substitute in equation (2), we first compute Lagrange's coefficients:

$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-5)(x-6)}{(3-5)(3-6)} = \frac{1}{6}(x^2-11x+30),$$

$$l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-3)(x-6)}{(5-3)(5-6)} = \frac{-1}{2}(x^2-9x+18)$$

and

$$l_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-3)(x-5)}{(6-3)(6-5)} = \frac{1}{3}(x^2-8x+15)$$

So we have

$$f(x) \cong P_2(x) = \sum_{r=0}^{2} l_r(x).y_r = l_0(x).y_0 + l_1(x).y_1 + l_2(x).y_2$$

$$= \frac{1}{6}(x^2-11x+30)(6) - \frac{1}{2}(x^2-9x+18)(10) + \frac{1}{3}(x^2-8x+15)(15)$$

$$= (x^2-11x+30) - 5(x^2-9x+18) + 5(x^2-8x+15)$$

$$= x^2-6x+15$$

Example 2 Use Lagrange's interpolation formula to find $f(2)$ given that

x	0	1	3	6
$y = f(x)$	2	4	10	18

Solution To get the polynomial that interpolates the given data we substitute into equation (1) to get the result

$$f(x) \cong P_3(x) = \sum_{r=0}^{3} l_r(x).y_r$$

$$= \frac{(x-1)(x-3)(x-6)}{(0-1)(0-3)(0-6)}(2) + \frac{(x-0)(x-3)(x-6)}{(1-0)(1-3)(1-6)}(4)$$

$$+ \frac{(x-0)(x-1)(x-6)}{(3-0)(3-1)(3-6)}(10) + \frac{(x-0)(x-1)(x-3)}{(6-0)(6-1)(6-3)}(18)$$

Then we have the approximation

$$f(2) \cong \frac{(1)(-1)(-4)}{(-1)(-3)(-6)}(2) + \frac{(2)(-1)(-4)}{(1)(-2)(-5)}(4) + \frac{(2)(1)(-4)}{(3)(2)(-3)}(10)$$

$$+ \frac{(2)(1)(-1)}{(6)(5)(3)}(18) = 6.8$$

Exercise 3.1

In problems $1-4$, find the polynomial of the least degree, which attains the present values at the given points

1.

x	1.4	1.3	1.1

$y = f(x)$	3.14	4.15	5.65

[Answer: $f(x) \cong P_2(x) = -8.667x^2 + 13.3x + 1.5$]

2.

x	-2	-1	0	1	2
$y = f(x)$	6	0	2	0	6

3.

x	-2	1	2	4
$y = f(x)$	25	-8	-15	-23

[Answer: $f(x) \cong x^2 - 10x + 1$]

4.

x	11	13	15	17	19	21
y	1342	2210	2758	5850	6878	9282

5. Given the table of readings:

x	1	2	3	4	5
y	0.10	0.11	0.13	0.14	0.16

Approximate $y(1.5)$ and $y(2.6)$, respectively.

[Answer: $y(1.5) = 0.103359$, $y(2.6) = 0.103882$]

6. Given a table of the values of the function

x	1.5	1.54	1.56	1.60	1.63	1.70
y	3.873	3.924	3.950	4.000	4.037	4.123

Using Lagrange' formula, find the values of the function at the indicated points:

$(\boldsymbol{a})$ $x = 1.52$ $(\boldsymbol{b})$ $x = 1.55$

$(\boldsymbol{c})$ $x = 1.58$ $(\boldsymbol{d})$ $x = 1.61$

$(\boldsymbol{e})$ $x = 1.67$

7. Find $y(0.7)$ and $y(1.4)$ from the table:

x	0.5	0.8	1.5	2.5	4
$y = f(x)$	1.15	1.77	3.09	4.67	6.56

[Answer: $y(0.7) = 1.566833$, $y(1.4) = 2.912521$]

8. Knowing the values of $\sin x$ for

$$x = 0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2},$$

find $\sin x$ for $x = \frac{\pi}{12}$ and estimate the error.

9. Find $y(1.5)$ and $y(3.5)$ from the table:

x	1	2	3	4	5
$y = f(x)$	0.25	0.17	0.11	0.1	0.08

[Answer: $y(1.5) = 0.21289$, $y(3.5) = 0.10039$]

10. Knowing the values of $\cos x$ for

$$x = 0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2},$$

find $\cos x$ for $x = \frac{\pi}{12}$ and estimate the error.

3.2. Difference operators

Engineers work frequently with tables. These can be of a mathematical nature (e.g. tables of logarithms or sine) or can be tables

obtained by experiments for design purposes. From these tables, engineers often have to estimate values of quantity for which we have only a certain number of values tabulated. It is important to obtain estimates, which are as accurate as the data permits.

Given a tabulated data, in this section we shall focus our attention on three extracts from the table of data; ***the forward difference***, ***the backward difference*** and ***the divided difference*** of the function represented by the table.

Before proceeding to introduce other interpolation formulas, let us give the following definitions.

Definition 3.1 (Equal spaced mesh points). Mesh points are called equally spaced if

$$x_r = x_0 + rh, \quad r = 1,2,\cdots,n$$

where h is a constant called the ***step size*** of the interval of tabulation.

In this case we have

$$x_r - x_{r-1} = h, \quad r = 1,2,\cdots,n$$

If the step size of the interval of tabulation h is not equal the mesh points are called ***unequal spaced***.

Suppose that we are given a function that represented by the tabulated values

x	x_0	x_1	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n

$y = f(x)$	y_0	y_1	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

We use the following symbols that use the entries of this table to form what are known the ***difference tables***.

Definition 3.2 (Forward difference operator). The forward finite difference of the function of the first order is defined by

$$\Delta y_r = y_{r+1} - y_r, \qquad r = 0,1,\cdots,n-1 \tag{1}$$

and the finite difference of the second order is defined by

$$\Delta^2 y_r = \Delta y_{r+1} - \Delta y_r, \qquad r = 0,1,\cdots,n-2 \tag{2}$$

In general, the finite difference of the *k*th order is defined by

$$\Delta^k y_r = \Delta^{k-1} y_{r+1} - \Delta^{k-1} y_r, \qquad r = 0,1,\cdots,n-k \tag{3}$$

Using definition 2.2, we have

$$\Delta y_0 = y_1 - y_0, \qquad \Delta y_1 = y_2 - y_1, \tag{4}$$

$$\begin{aligned}\Delta^2 y_0 &= \Delta(\Delta y_0) = \Delta y_1 - \Delta y_0 \\ &= (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0,\end{aligned} \tag{5}$$

and

$$\begin{aligned}\Delta^3 y_0 &= \Delta(\Delta^2 y_0) \\ &= \Delta(y_2 - 2y_1 + y_0) = (\Delta y_2 - 2\Delta y_1 + \Delta y_0) \\ &= y_3 - 3y_2 + 3y_1 - y_0,\end{aligned} \tag{6}$$

and so on.

Definition 3.3 (Backward Difference Operator). The backward finite

difference of the function of the first order is defined by

$$\nabla y_r = y_r - y_{r-1}, \qquad r = 1,2,\cdots,n \tag{7}$$

and the finite difference of the second order is defined by

$$\nabla^2 y_r = \nabla y_r - \nabla y_{r-1}, \qquad r = 2,3,\cdots,n \tag{8}$$

In general, the finite difference of the *k*th order is defined by

$$\nabla^k y_r = \nabla^{k-1} y_r - \nabla^{k-1} y_{r-1}, \qquad r = k, k+1, \cdots, n \tag{9}$$

Definition 3.3 gives

$$\nabla y_1 = y_1 - y_0, \qquad \nabla y_2 = y_2 - y_1, \tag{10}$$

$$\begin{aligned} \nabla^2 y_2 &= \nabla(\nabla y_2) = \nabla y_2 - \nabla y_1 \\ &= (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0, \end{aligned} \tag{11}$$

$$\begin{aligned} \nabla^2 y_3 &= \nabla(\nabla y_3) = \nabla y_3 - \nabla y_2 \\ &= (y_3 - y_2) - (y_2 - y_1) = y_3 - 2y_2 + y_1, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \nabla^3 y_3 &= \nabla(\nabla^2 y_3) \\ &= \nabla(y_3 - 2y_2 + y_1) = (\nabla y_2 - 2\nabla y_1 + \nabla y_0) \\ &= y_3 - 3y_2 + 3y_1 - y_0, \end{aligned} \tag{13}$$

$$\begin{aligned} \nabla^4 y_3 &= \nabla(\nabla^3 y_3) \\ &= \nabla^3 y_3 - \nabla^3 y_2, \end{aligned} \tag{14}$$

and so on.

We note that the second- and higher- order differences are generally obtained by differencing the previous differences, but

Equations (6), (7) and (10) show how any difference can be calculated directly from the functional values.

On constructing the forward difference and the backward difference tables we use the above definitions of the difference operators. This is shown in Table 3.1 and Table 3.2, respectively.

x	y	Δ	Δ^2	$\cdots$
x_0	y_0			
		$\Delta y_0 = y_1 - y_0$		
x_1	y_1		$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$	
		$\Delta y_1 = y_2 - y_1$		
x_2	y_2		$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$	
		$\Delta y_2 = y_3 - y_2$		
x_3	y_3			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
x_{n-2}	y_{n-2}		$\Delta^2 y_{n-3} = \Delta y_{n-2} - \Delta y_{n-3}$	
		$\Delta y_{n-2} = y_{n-1} - y_{n-2}$		
x_{n-1}	y_{n-1}		$\Delta^2 y_{n-2} = \Delta y_{n-1} - \Delta y_{n-2}$	
		$\Delta y_{n-1} = y_n - y_{n-1}$		
x_n	y_n			

Table 3.1 Forward difference table

x	y	∇	∇^2	$\cdots$
x_0	y_0			
		$\nabla y_1 = y_1 - y_0$		
x_1	y_1		$\nabla^2 y_2 = \nabla y_2 - \nabla y_1$	
		$\nabla y_2 = y_2 - y_1$		
x_2	y_2		$\nabla^2 y_3 = \nabla y_3 - \nabla y_2$	
		$\nabla y_3 = y_3 - y_2$		
x_3	y_3			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

x_{n-2}	y_{n-2}		$\nabla^2 y_{n-1} = \nabla y_{n-1} - \nabla y_{n-2}$	
x_{n-1}	y_{n-1}	$\nabla y_{n-1} = y_{n-1} - y_{n-2}$	$\nabla^2 y_n = \nabla y_n - \nabla y_{n-1}$	
x_n	y_n	$\nabla y_n = y_n - y_{n-1}$		

Table 3.2 Backward difference table

Example 1 Construct the forward difference table from the readings

x	1	3	5	7	9
y	2	5	15	40	75

Solution The forward difference table is constructed as follows:

x	y	Δ	Δ^2	Δ^3	Δ^4
1	2				
		3			
3	5		7		
		10		8	
5	15		15		-3
		25		5	
7	40		10		
		35			
9	75				

Table 3.3 Difference table of Example 1

We note from the above example that in hand computation of a difference table, great care should be exercised to avoid arithmetic errors in the subtractions –the fact that we subtract the upper entry from the lower adds a real source of confusion. One of the best ways to check for mistakes is to add the sum of the numbers in each column to the top entry in the column to its left. This sum should equal the bottom entry in the column to the left.

For the unequal spaced mesh points we have the following definition of the difference operator.

Definition 3.4 (Divided Difference Operator). The divided difference of the function $y(x)$ of the first order is defined by

$$[y_{r-1}, y_r] = \frac{y_r - y_{r-1}}{x_r - x_{r-1}}, \quad r = 1,2,\cdots,n$$

and the divided difference of the second order is defined by

$$[y_{r-2}, y_{r-1}, y_r] = \frac{[y_{r-1}, y_r] - [y_{r-2}, y_{r-1}]}{x_r - x_{r-2}}, \quad r = 2,3,\cdots,n$$

The process is continued for $r = k+1, k+2, \cdots, n$ using the formula

$$[y_{r-k}, \cdots, y_{r-1}, y_r] = \frac{[y_{r-k+1}, \cdots, y_r] - [y_{r-k}, \cdots, y_{r-1}]}{x_r - x_{r-2}},$$

until the nth divided difference $[y_1, \cdots, y_n]$ has been obtained.

Thus we have

$$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0}, \qquad [y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1},$$

$$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0},$$

and

$$[y_0, y_1, y_2, y_3] = \frac{[y_1, y_2, y_3] - [y_0, y_1, y_2]}{x_3 - x_0},$$

and so on.

The determination of the divided differences from tabulated data points is outlined in Table 3.4.

x	y	$First\ divided\ difference$	$Second\ divided\ difference$
x_0	y_0		
		$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0}$	
x_1	y_1		$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0}$
		$[y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1}$	
x_2	y_2		$[y_1, y_2, y_3] = \frac{[y_2, y_3] - [y_1, y_2]}{x_3 - x_1}$
		$[y_2, y_3] = \frac{y_3 - y_2}{x_3 - x_2}$	
x_3	y_3		$[y_2, y_3, y_4] = \frac{[y_3, y_4] - [y_2, y_3]}{x_4 - x_2}$
		$[y_3, y_4] = \frac{y_4 - y_3}{x_4 - x_3}$	
x_4	y_4		$[y_3, y_4, y_5] = \frac{[y_4, y_5] - [y_3, y_4]}{x_5 - x_3}$
		$[y_4, y_5] = \frac{y_5 - y_4}{x_5 - x_4}$	
x_5	y_5		
$\vdots$	$\vdots$		

Table 3.4 Divided difference table

Example 2 Construct the difference table from the readings

x	1	3	7	10
y	2	5	12	30

Solution The divided difference table is constructed as follows:

x	y	$[.,.]$	$[.,.,.]$	$[.,.,.,.]$
1	2			
		1.5		
3	5		0.041667	
		1.75		0.062831
7	12		0.6071143	
		6		
10	30			

Table 3.5 Divided Difference table of Example 2

3.3. General Newton' formulas of interpolation

The Lagrange interpolation technique is useful for determining the values of successively higher-degree interpolating polynomials at a particular point. Each of the entries in the interpolating table, however, depends on the point being evaluated, so the table cannot be used to give an explicit representation for the interpolating polynomial.

Methods for determining the explicit representation of an interpolating polynomial from tabulated data are known as ***divided-***

difference methods. These methods were more widely used for computational purposes before digital computing equipment became readily available. However, the methods can be used to derive techniques for approximating the derivatives and integrals of functions, as well as for approximating the solutions to differential equations.

Suppose that $P(x)$ is the polynomial of degree at most n that agrees with the function $y = f(x)$ at the distinct numbers $x_r,\ r = 0,1,\cdots,n$ which is given from the table of data:

x	x_0	x_1	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n
$y = f(x)$	y_0	y_1	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

The divided difference formulas of f with respect to $x_0, x_1, \cdots, x_n$ can be derived by showing that $P(x)$ has the representation

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots$$
$$+a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}), \qquad (1)$$

for appropriate constants $a_0, a_1, \cdots, a_n$.

To determine the first of these constants, a_0, note that if $P(x)$ is written in the form of equation (1), then evaluating P at x_0 leaves only the constant term a_0; that is,

$$P(x_0) = a_0 + 0 + 0 + \cdots = y_0,$$

Thus

$$a_0 = y_0 \qquad (2)$$

Similarly, when P is evaluated at x_1, the only nonzero terms in the evaluation of $P(x_1)$ are the constant and linear terms,

$$P(x_1) = y_0 + a_1(x_1 - x_0) + 0 + \cdots = y_1,$$

So

$$a_1 = \frac{(y_1 - y_0)}{(x_1 - x_0)} = [y_0, y_1] \tag{3}$$

Substituting from (2) and (3), the interpolating polynomial in equation (1) becomes:

$$P(x) = y_0 + [y_0, y_1](x - x_0) + a_2(x - x_0)(x - x_1) + \cdots$$
$$+a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}), \tag{4}$$

The constants $a_2, a_3, \cdots, a_n$ in P can be consecutively obtained in a manner similar to the evaluation of a_0 and a_1, but algebraic manipulation become tedious. Therefore, as might be expressed from the evaluation of a_0 and a_1, the required constants are

$$a_r = [y_0, y_1, y_2, \cdots, y_r], \qquad r = 0,1,\cdots,n \tag{5}$$

so P can be rewritten as

$$\boxed{\begin{aligned} P(x) = y_0 + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1) + \cdots \\ +[y_0, y_1, y_2, \cdots, y_n](x - x_0)(x - x_1)\cdots(x - x_{n-1}), \end{aligned}} \tag{6}$$

or as

$$\boxed{P(x) = y_0 + \sum_{r=1}^{n} [y_0, y_1, y_2, \cdots, y_r](x - x_0)(x - x_1)\cdots(x - x_{r-1})} \tag{7}$$

Equation (6) or (7) is known as ***Newton forward divided-difference formula***, or ***General Newton' formula*** and it is often used for readings at unequal intervals. This formula does not depend on the order of the reading points, thus if the interpolating nodes are recorded as $x_n, x_{n-1}, \cdots, x_0$ a formula similar to equation (6) results:

$$P(x) = y_n + [y_{n-1}, y_n](x - x_n)$$
$$+[y_{n-2}, y_{n-1}, y_n](x - x_n)(x - x_{n-1}) + \cdots$$
$$+[y_0, y_1, y_2, \cdots, y_n](x - x_n)(x - x_{n-1}) \cdots (x - x_1), \quad (8)$$

Equation (8) is known as ***Newton backward divided-difference formula***.

Example 1 Find $P(x)$ that interpolates the following readings:

x	0	1	3	6
y	0	1	27	36

Solution. Constructing the difference table as follows:

x	y	[.,.]	[.,.,.]	[.,.,.,.]
0	$\underline{0}$			
		$\underline{1}$		
1	1		$\underline{4}$	
		13		$\underline{-1}$
3	27		$\underline{-2}$	
		$\underline{3}$		
6	$\underline{36}$			

Table 3.6 Divided Difference table of Example 1

The coefficients of the Newton's forward divided difference formula of the interpolatory polynomial result from a diagonally downward path in Table 2.6 (the solid line). Thus the interpolating polynomial is

$$\begin{aligned} P(x) &= y_0 + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1) \\ &\quad + [y_0, y_1, y_2, y_3](x - x_0)(x - x_1)(x - x_2) \\ &= (0) + (1)(x - 0) + (4)(x - 0)(x - 1) \\ &\quad + (-1)(x - 0)(x - 1)(x - 3) \\ &= -x^3 + 8x^2 - 6x \end{aligned}$$

Note that we can use the backward formula of Newton to obtain the polynomial. To substitute in this formula, we use the coefficients from a diagonally upward path in Table 2.6 (the dashed line). Thus the interpolating polynomial is

$$\begin{aligned} P(x) &= y_3 + [y_2, y_3](x - x_3) + [y_1, y_2, y_3](x - x_3)(x - x_2) \\ &\quad + [y_0, y_1, y_2, y_3](x - x_3)(x - x_2)(x - x_1) \\ &= (36) + (3)(x - 6) + (-2)(x - 6)(x - 3) \\ &\quad + (-1)(x - 6)(x - 3)(x - 1) \\ &= -x^3 + 8x^2 - 6x \end{aligned}$$

Example 1 Approximate $y(1.25)$ and $y(5)$ from the following readings

x	0	1.5	3	4	7
y	3	12	20	30	65

Solution. Constructing the difference table as follows:

x	y	$[.,.]$	$[.,.,.]$	$[.,.,.,.]$	$[.,.,.,.,.]$
0	3				
		6			
1.5	12		-0.22222		
		5.333333		0.522222	
3	20		1.866667		-0.11227
		10		-0.26364	
4	30		0.416667		
		11.66667			
7	65				

Table 3.7 Divided difference table of Example 2

Using Newton's forward divided difference formula

$$P(x) = 3 + 6(x-0) - 0.22222(x-0)(x-1.5)$$
$$+0.522222(x-0)(x-1.5)(x-3)$$
$$-0.11227(x-0)(x-1.5)(x-3)(x-4),$$

from which we substitute approximate $x = 1.25$, to get

$$y(1.25) \cong 3 + 7.5 + 0.069444 + 0.28559 + 0.168837 = 11.02387$$

To get $y(5)$, the backward formula of Newton formula in the form

$$P(x) = 65 + 11.66667(x-7) + 0.416667(x-7)(x-4)$$
$$-\ 0.26364(x-7)(x-4)(x-3)$$
$$-0.11227(x-7)(x-4)(x-3)(x-1.5)$$

So that we have

$$y(5) \cong 65 - 23.333 - 0.8333 + 1.05454 + 1.5717 = 43.4596$$

3.4. Newton forward-difference formula

When $x_0, x_1, \cdots, x_n$ are arranged consecutively with equal spacing, equation (6) in the previous section can be expressed in a simplified form. Assuming equal spaced mesh points, i.e.,

$$x_r - x_{r-1} = h, \;\; r = 1,2,\cdots,n,$$

and recalling the notations of Δ and the divided difference symbol we get

$$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y_0}{h},$$

$$[y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y_1}{h}$$

$$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0}$$

$$= \frac{\frac{\Delta y_1}{h} - \frac{\Delta y_0}{h}}{2h} = \frac{\Delta^2 y_0}{2h^2},$$

and in general

$$[y_0, y_1, \cdots, y_r] = \frac{\Delta^r y_0}{r\,!\,h^r}, \quad r = 1,2,\cdots,n$$

Consequently, equation (6) becomes

$$\boxed{\begin{aligned} P(x) = y_0 &+ \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2h^2}(x - x_0)(x - x_1) + \cdots \\ &+ \frac{\Delta^n y_0}{n\,!\,h^n}(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \qquad (1) \end{aligned}}$$

This form is called the ***Newton forward–difference formula***. It is used to interpolate the values of y nearer to the beginning value of the given table. Introducing the transformation

$$x = x_0 + sh \quad \Rightarrow \quad s = \frac{x - x_0}{h},$$

so,

$$\frac{x - x_1}{h} = \frac{(x - x_0) - (x_1 - x_0)}{h} = \frac{(x - x_0) - h}{h} = s - 1,$$

$$\frac{x - x_2}{h} = \frac{(x - x_0) - (x_2 - x_0)}{h} = \frac{(x - x_0) - 2h}{h} = s - 2,$$

$$\vdots$$

so equation (1) can be rewritten in the form

$$P(x) = y_0 + \Delta y_0 \frac{(x - x_0)}{h} + \frac{\Delta^2 y_0}{2\,!} \frac{(x - x_0)}{h} \frac{(x - x_1)}{h} + \cdots,$$

or,

$$P(x_0 + sh) = y_0 + s\Delta y_0 + \frac{s(s-1)}{2\,!}\Delta^2 y_0 + \cdots$$

$$+ \frac{s(s-1)(s-2)\cdots(s-n+1)}{n\,!}\Delta^n y_0 \qquad (2)$$

Using the binomial-coefficient notation

$$\binom{s}{r} = \frac{s(s-1)(s-2)\cdots(s-r+1)}{r\,!},$$

we can express $P(x)$ compactly as

$$\boxed{\begin{aligned} P(x) &= y_0 + \binom{s}{1}\Delta y_0 + \binom{s}{2}\Delta^2 y_0 + \cdots + \binom{s}{n}\Delta^n y_0 \\ &= \sum_{r=0}^{n} \binom{s}{r} \Delta^r y_0 \qquad (3) \end{aligned}}$$

Note that this formula uses the entries of the upper diagonal of the difference table as shown in Table 2.8.

x	y	Δ	Δ^2	Δ^3	$\cdots$
x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		
		Δy_2		$\Delta^3 y_1$	
x_3	y_3				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

Table 3.8 Forward difference table

Remark 1 Formula (3) is used for interpolation at points x which are close to the beginning of the table (x_0) in order to make maximum use of the data points and so to increase the degree of the interpolating polynomial.

Remark 2 For $n = 1$ and $n = 2$ we get particular cases:

♦ Linear Interpolation

$$P(x) = y_0 + s\Delta y_0$$

♦ Quadratic Interpolation

$$P(x) = y_0 + s\Delta y_0 + \frac{s(s-1)}{2\,!}\Delta^2 y_0$$

Example 1 Use the readings

x	0	0.2	0.4	0.6	0.8	1
y	-0.002	0.058	0.149	0.283	0.470	0.717

to find y when

(i) $x = 0.15$ **(ii)** $x = 0.3$ **(iii)** $x = 0.7$ **(iv)** $x = 0.9$

Solution We first form a difference table which is shown as Table 2.9. In this example, it is obvious that $h = 0.2$.

(i) Here we take $x_0 = 0$, then $s = \frac{0.15-0}{0.2} = 0.75$ and the use of Newton forward-difference formula gives

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5
0	-0.002					
		0.06				
0.2	0.058		0.031			
		0.091		0.012		
0.4	0.149		0.043		-0.002	
		0.134		0.01		-0.001
06	0.283		0.053		-0.003	

		0.187		0.07		
0.8	0.470		0.06	⋮	0.283	
		0.247				
1	0.717					

Table 3.9 Forward difference table of Example 1

$$y(0.15) \cong -0.002 + (0.06)(0.75) + (0.031)\frac{(0.75)(-0.25)}{2!}$$
$$+ (0.012)\frac{(0.75)(-0.25)(-1.25)}{3!}$$
$$- 0.002\frac{(0.75)(-0.25)(-1.25)(-2.25)}{4!}$$
$$- 0.001\frac{(0.75)(-0.25)(-1.25)(-2.25)(-3.25)}{5!}$$
$$= -0.002 + 0.045 - 0.00291 + 0.00469 - 0.000044 - 0.000014$$
$$= 0.040592$$

(ii) 0.3 lies halfway between 0.2 and 0.4, so we use $x_0 = 0.2$ and $s = \frac{0.3-0.2}{0.2} = 0.5$. This yield

$$y(0.3) \cong 0.058 + 0.091(0.5) + 0.043\frac{(0.5)(-0.5)}{2!}$$
$$+ (0.01)\frac{(0.5)(-0.5)(-1.5)}{3!} - 0.003\frac{(0.5)(-0.5)(-1.5)(-2.5)}{4!}$$
$$= 0.058 + 0.0455 - 0.00538 + 0.000625 + 0.000117 = 0.098867$$

(iii) This time take $x_0 = 0.6$ with $s = \frac{0.7-0.6}{0.2} = 0.5$. Here, we can only pick up the first and second differences and

$$y(0.7) \cong 0.283 + 0.187(0.5) + 0.06\frac{(0.5)(-0.5)}{2!}$$
$$= 0.283 + 0.0935 - 0.0075 = 0.369$$

(iv) $x_0 = 0.8$, $s = \frac{0.9-0.8}{0.2} = 0.5$. We are restricting to linear interpolation:

$$y(0.7) \cong 0.47 + 0.247(0.5) = 0.47 + 0.1235 = 0.5935$$

Hence the formula ceases to be of use at the end of a table and we wonder if a similar formula exists which allows us to go back into a table to pick up values.

3.5. Newton backward-difference formula

Newton backward divided difference formula is used to derive a more commonly applied formula known as the ***Newton backward-difference formula***. Using the notation of the backward finite difference operator, we have

$$[y_{n-1}, y_n] = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} = \frac{\nabla y_n}{h},$$

$$[y_{n-2}, y_{n-1}] = \frac{y_{n-1} - y_{n-2}}{x_{n-1} - x_{n-2}} = \frac{\nabla y_{n-1}}{h},$$

$$[y_{n-2}, y_{n-1}, y_n] = \frac{[y_{n-1}, y_n] - [y_{n-2}, y_{n-1}]}{x_n - x_{n-2}}$$

$$= \frac{\frac{\nabla y_n}{h} - \frac{\nabla y_{n-1}}{h}}{2h} = \frac{\nabla^2 y_n}{2h^2},$$

and in general

$$[y_{n-r}, \cdots, y_{n-1}, y_n] = \frac{\nabla^r y_0}{r\,!\,h^r}, \quad r = 1,2,\cdots,n-1$$

Consequently, Newton backward divided difference formula becomes

$$\boxed{\begin{aligned} P(x) = y_n &+ \frac{\nabla y_n}{h}(x - x_n) + \frac{\nabla^2 y_n}{2h^2}(x - x_n)(x - x_{n-1}) + \cdots \\ &+ \frac{\nabla^n y_0}{n\,!\,h^n}(x - x_n)(x - x_{n-1}) \cdots (x - x_1) \qquad (1) \end{aligned}}$$

Let

$$x = x_n + sh \quad \Rightarrow \quad s = \frac{x - x_n}{h},$$

so,

$$\frac{x - x_{n-1}}{h} = \frac{(x - x_n) + (x_n - x_{n-1})}{h} = \frac{(x - x_0) + h}{h} = s + 1,$$

$$\frac{x - x_{n-2}}{h} = \frac{(x - x_n) + (x_n - x_{n-2})}{h} = \frac{(x - x_0) + 2h}{h} = s + 2,$$

$$\vdots$$

so equation (1) can be rewritten in the form

$$P(x) = P(x_n + sh) = y_n + s\nabla y_n + \frac{s(s+1)}{2\,!}\nabla^2 y_n + \cdots + \frac{s(s+1)(s+2)\cdots(s+n-1)}{n\,!}\nabla^n y_n \tag{2}$$

or, compactly, we get

$$\boxed{\begin{aligned} P(x) &= y_n + \binom{s}{1}\nabla y_n + \binom{s+1}{2}\nabla^2 y_n + \cdots \binom{s+n-1}{n}\nabla^n y_n \\ &= \sum_{r=0}^{n} \binom{s+r-1}{r}\nabla^r y_n \end{aligned} \tag{3}}$$

Note that this formula uses the entries of the diagonal indicated by the arrow shown in the difference Table 2.10.

x	y	∇	∇^2	∇^3	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

x_{n-2}	y_{n-2}		$\nabla^2 y_{n-1}$		
		∇y_{n-1}		$\nabla^3 y_n$	
x_{n-1}	y_{n-1}		$\nabla^2 y_n$		
		∇y_n			
x_n	y_n				

Table 3.10 Backward difference table

Remark Formula (3) is used for interpolation at points x which are close to the end of the table, i.e. to x_n.

Example 1 Use Newton backward difference formula (3) to case **(iv)** of the example in the previous section where the forward interpolation formula reduced to linear interpolation.

Solution To obtain y at $x = 0.9$ by the use of Newton backward difference formula (see Table 2.11); We have

$$h = 0.2, \quad x_n = 0.8, \quad s = \frac{0.9 - 1}{0.2} = 0.5$$

x	y	∇	∇^2	∇^3	∇^4	∇^5
0	-0.002					
		0.06				

0.2	0.058		0.031			
		0.091		0.012		
0.4	0.149		0.043		−0.002	
		0.134		0.01		−0.001
06	0.283		0.053		−0.003	
		0.187		0.07		
0.8	0.470		0.06			
		0.247				
1	0.717					

Table 3.11 Backward difference table of Example 1

Therefore

$$y(0.9) \cong 0.47 + (0.187)(0.5) + (0.053)\frac{(0.5)(1.5)}{2!}$$
$$+ (0.01)\frac{(0.5)(1.5)(2.5)}{3!} - 0.002\frac{(0.5)(1.5)(2.5)(3.5)}{4!}$$
$$= 0.47 + 0.0935 + 0.019875 + 0.003125 - 0.00054687$$
$$= 0.585953$$

Alternatively, we may take $x_n = 1$, hence $s = \frac{0.9-1}{0.2} = -0.5$ to obtain

$$y(0.9) \cong 0.717 + (0.247)(-0.5) + (0.06)\frac{(-0.5)(0.5)}{2!}$$
$$+ (0.07)\frac{(-0.5)(0.5)(1.5)}{3!} - 0.003\frac{(-0.5)(0.5)(1.5)(2.5)}{4!}$$
$$- 0.001\frac{(-0.5)(0.5)(1.5)(2.5)(3.5)}{5!}$$
$$= 0.717 - 0.1235 - 0.0075 - 0.00044 + 0.000117 + 0.000027$$
$$= 0.585707$$

3.6. Central-difference formulas

Given the difference table

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6	$\cdots$
$\vdots$								

$x_{-3} = x_0 - 3h$	y_{-3}							
		Δy_{-3}						
$x_{-2} = x_0 - 2h$	y_{-2}		$\Delta^2 y_{-3}$					
		Δy_{-2}		$\Delta^3 y_{-3}$				
$x_{-1} = x_0 - h$	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$			
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$		
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$	
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$		
$x_1 = x_0 + h$	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		$\Delta^6 y_{-2}$	
		Δy_1		$\Delta^3 y_0$		$\Delta^5 y_{-1}$		
$x_2 = x_0 + 2h$	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$		$\Delta^6 y_{-1}$	
		Δy_2		$\Delta^3 y_1$				
$x_3 = x_0 + 3h$	y_3							
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3.12 Forward difference table

The Newton formulas are not appropriate for approximating a value x, that lies near the centre of the table, since employing either the forward or backward method in such a way that the highest-order difference is involved will not allow x_0 to be close to x. A number of difference formulas are available in this instant, each of which has situations when it can be used to maximum advantage. These methods which are known as ***centred-difference formulas***, will be discussed in this section

3.6.1. Gauss forward-difference formula

Gauss's first interpolation formula (for forward interpolation) uses the differences of f with respect to $x_0, x_1, x_{-1}, x_2, x_{-2}, \cdots$

Denoting by $s = \frac{x-x_0}{h}$ and substitute in Newton forward-difference formula from Table 2.12 as indicated by the solid arrow $(\rightarrow)$ we get

$$P(x) = y_0 + s\Delta y_0 + \frac{s(s-1)}{2\,!}\Delta^2 y_{-1} + \frac{s(s-1)(s+1)}{3\,!}\Delta^2 y_{-1}$$

$$+\frac{s(s-1)(s+1)(s-2)}{4\,!}\Delta^3 y_{-2} + \frac{s(s-1)(s+1)(s-2)(s+2)}{5\,!}\Delta^4 y_{-2}$$

$$+\cdots \tag{1}$$

or, in compact form

$$\boxed{\begin{aligned} P(x) = y_0 &+ \binom{s}{1}\Delta y_0 + \binom{s}{2}\Delta^2 y_{-1} + \binom{s+1}{3}\Delta^3 y_{-1} \\ &+ \binom{s+1}{4}\Delta^4 y_{-2} + \binom{s+2}{5}\Delta^5 y_{-2} \\ &+ \binom{s+2}{6}\Delta^6 y_{-3} + \binom{s+3}{7}\Delta^7 y_{-3} + \cdots \end{aligned} \qquad (2)}$$

Here the differences

$$y_0,\ \Delta y_0,\ \Delta^2 y_{-1},\ \Delta^3 y_{-1},\ \Delta^4 y_{-2},\ \Delta^5 y_{-2}, \cdots, \tag{3}$$

used from Table 2.12 are derived with respect to

$$x_0,\ x_1,\ x_{-1},\ x_2,\ x_{-2},\ \cdots,$$

and are indicated by the solid arrow $(\rightarrow)$. In Gauss' forward formulas these differences are multiplied by

$$1,\ s,\ \frac{s(s-1)}{2!},\ \frac{s(s-1)(s+1)}{3!},\ \frac{s(s-1)(s+1)(s-2)}{4!},$$

$$\frac{s(s-1)(s+1)(s-2)(s+2)}{5!}, \cdots \tag{4}$$

respectively, as given by formula (1) and (2).

3.6.2. Gauss backward-difference formula

Gauss's second interpolation formula (for backward interpolation) uses the divided differences of f with respect to

$$x_0,\ x_{-1},\ x_1,\ x_{-2},\ x_2,\cdots,$$

and substitute in Newton forward-difference formula from the differences in Table 2.10;

$$P(x) = y_0 + s\Delta y_{-1} + \frac{s(s+1)}{2\,!}\Delta^2 y_{-1} + \frac{s(s+1)(s-1)}{3\,!}\Delta^2 y_{-2}$$
$$+\frac{s(s+1)(s-1)(s+2)}{4\,!}\Delta^3 y_{-2}$$
$$+\frac{s(s+1)(s-1)(s+2)(s-2)}{5\,!}\Delta^4 y_{-3} + \cdots \qquad (5)$$

or, in compact form

$$\boxed{\begin{aligned} P(x) = y_0 &+ \binom{s}{1}\Delta y_{-1} + \binom{s+1}{2}\Delta^2 y_{-1} \\ &+ \binom{s+1}{3}\Delta^3 y_{-2} + \binom{s+2}{4}\Delta^4 y_{-2} \\ &+ \binom{s+2}{5}\Delta^5 y_{-3} + \binom{s+3}{6}\Delta^6 y_{-3} \\ &+ \binom{s+3}{7}\Delta^7 y_{-4} + \cdots \qquad (6) \end{aligned}}$$

Here $s = \frac{x-x_0}{h}$ and the differences used from the table are indicated by the dotted arrow($\cdots\rightarrow$), namely

$$y_0,\ \Delta y_{-1},\ \Delta^2 y_{-1},\ \Delta^3 y_{-2},\ \Delta^4 y_{-2},\ \Delta^5 y_{-3},\cdots, \qquad (7)$$

are multiplied respectively, by

$$1,\ s,\ \frac{s(s+1)}{2!},\ \frac{s(s+1)(s-1)}{3!},\ \frac{s(s+1)(s-1)(s+2)}{4!},$$

$$\frac{s(s+1)(s-1)(s+2)(s-2)}{5!}, \cdots \tag{8}$$

Remark The Gaussian formulas are used for interpolation in the middle of the table (close to x_0). The forward formula (1) is applied for $x > x_0$, and the backward (4) for $x < x_0$. Both of the two formulas goes through the table in a zigzag path, the first step being a forward one in (1) and a backward one in (5) as shown in Table 2.12.

Example 1 Using the table of the values of the function $y = e^x$ (Table 2.13), find $e^{1.13}$ and $e^{1.16}$.

Solution For computing $e^{1.13}$, we use of Gauss backward difference formula (6) with

$$h = 0.05, \quad x_0 = 1.15, \qquad s = \frac{1.13 - 1.15}{0.05} = -0.4,$$

$$\begin{aligned} P(x) = y_0 &+ \binom{s}{1}\Delta y_{-1} + \binom{s+1}{2}\Delta^2 y_{-1} \\ &+ \binom{s+1}{3}\Delta^3 y_{-2} + \binom{s+2}{4}\Delta^4 y_{-2} \\ &+ \binom{s+2}{5}\Delta^5 y_{-3} + \binom{s+3}{6}\Delta^6 y_{-3} \end{aligned}$$

x	y	Δ	Δ^2	Δ^3	Δ^4

1.00	2.7183				
		0.1394			
1.05	2.8577		0.0071		
		0.1465		0.0004	
1.10	3.0042		0.0075		0
		0.1540		0.0004	
1.15	3.1582		0.0079		0
		0.1619		0.0004	
1.20	3.3201		0.0083		0
		0.1702		0.0004	
1.25	3.4903		0.0088		
		0.1790			
1.30	3.6693				

Table 3.13 Difference table of Example 1

Thus,

$$e^{1.13} = 3.1582 + 0.1540\binom{-0.4}{1} + 0.0079\binom{0.6}{2} + 0.0004\binom{0.6}{3}$$

$$= 3.1582 + 0.1540(-0.4) + 0.0079\left(\frac{(0.6)(-0.4)}{2!}\right)$$

$$+0.0004\frac{(0.6)(-0.4)(-1.4)}{3!}$$

$$= 3.1582 - 0.0616 - 0.000948 + 0.0000224$$

$$= 3.0956744$$

To obtain $e^{1.16}$ by the use of Gauss forward difference formula (2) with

$$h = 0.05, \;\; x_0 = 1.16, \qquad s = \frac{1.16 - 1.15}{0.05} = 0.2,$$

$$P(x) = y_0 + \binom{s}{1}\Delta y_0 + \binom{s}{2}\Delta^2 y_{-1} + \binom{s+1}{3}\Delta^3 y_{-1}$$

$$+\binom{s+1}{4}\Delta^4 y_{-2}+\binom{s+2}{5}\Delta^5 y_{-2}$$

$$+\binom{s+2}{6}\Delta^6 y_{-3}+\binom{s+3}{7}\Delta^7 y_{-3}+\cdots,$$

then

$$e^{1.16}=3.1582+0.1619\binom{0.2}{1}+0.0079\binom{0.2}{2}+0.0004\binom{1.2}{3}$$

$$=3.1582+0.1540(0.2)+0.0079\left(\frac{(0.2)(-0.8)}{2!}\right)$$

$$+0.0004\frac{(1.2)(0.2)(-0.8)}{3!}$$

$$=3.1582-0.0616-0.000948+0.0000224=3.0956744$$

Example 2 Given the readings:

x	1.0	1.5	2.0	2.5	3
y	0.12	0.52	1.47	3.24	6.21

By the use of two different interpolation formulas, find $y(1.25)$ and $y(2.25)$ from this table.

Solution Here, we have equal spaced mesh points with step $h=0.5$, so that the forward formula of Newton and Gauss may be used to evaluate $y(1.25)$ and $y(2.25)$, respectively. Hence, on constructing the difference table and computing s for each case, we get

x	y	Δ	Δ^2	Δ^3	Δ^4
1.00	0.12				
		0.4			
1.50	0.52		0.55		
		0.95		0.27	
2.0	1.47		0.82		0.11
		1.77		0.38	
2.5	3.24		1.2		0
		2.97			
3	6.21				0

Table 3.14 Difference table of Example 2

$$y(1.25) = 0.12 + 0.4(0.5) + 0.55\left(\frac{(0.5)(-0.5)}{2}\right)$$
$$+0.27\left(\frac{(0.5)(-0.5)(-1.5)}{6}\right) + 0.11\left(\frac{(0.5)(-0.5)(-1.5)(-2.5)}{24}\right)$$
$$= 0.12 + 0.2 - 0.06875 + 0.016875 - 0.0043 = 0.263828,$$

$$y(2.25) = 1.47 + 1.77(0.5) + 0.82\left(\frac{(0.5)(-0.5)}{2}\right)$$
$$+0.38\left(\frac{(0.5)(-0.5)(1.5)}{6}\right) + 0.11\left(\frac{(0.5)(-0.5)(1.5)(-1.5)}{24}\right)$$
$$= 1.47 + 0.885 - 0.1025 - 0.02375 + 0.002578 = 2.231328$$

The use of Gauss' backward formula to approximate $y(2.25)$ yields

$$y(2.25) = 1.47 + 0.95(0.5) + 0.82\left(\frac{(0.5)(1.5)}{2}\right)$$
$$+0.27\left(\frac{(0.5)(1.5)(-0.5)}{6}\right) + 0.11\left(\frac{(0.5)(1.5)(-0.5)(2.5)}{24}\right)$$
$$= 1.47 + 0.475 + 0.3075 - 0.01688 - 0.0043 = 2.231328$$

Exercise 3.2

In problems 1 to 3 make use of the interpolation formulas to determine the values of the function at the indicated points:

1. Find $y(3)$ and $y(10)$

x	1	4	9	16	25
y	0	2.89	5.75	8.65	11.53

$[Ans.\ y(3) = 2.066377,\ \ y(10) = 6.184821]$

2. Find $y(12), y(14),\ y(16)$, and $y(20)$.

x	11	13	15	17	19	21
y	1342	2210	2758	5850	6878	9282

3. Find $y(2.5)$ and $y(3.5)$

x	1	2	3	4	5
y	7.85	3.02	1.24	0.58	0.34

$[Ans.\ y(2.5) = 1.898203,\ \ y(3.5) = 0.842578]$

4. Find. $y(13),\ y(25)$, and $y(30)$.

x	14	17	31	35
y	68.7	64.0	44.0	39.1

5. By the use of the backward interpolation formulas of Newton and Gauss, find $y(8)$ and $y(6)$, respectively.

x	1	3	5	7	9
y	2.7	1.2	0.5	0.2	0.1

$[Ans.\ y(8) = 0.1296875,\ \ y(6) = 0.3140625]$

6. Find $y(94), y(99)$ and $y(102)$.

x	93.0	96.2	100.0	104.2	108.7
y	11.38	12.80	14.70	17.07	19.1

7. Approximate $y(6.5)$ and $y(9.5)$ using two different interpolation formulas from the table of readings:

x	2	4	6	8	10
y	22.3	24.9	26.9	28.6	30.1

$[Ans.\ y(6.5)(Gauss\ Forward) = 27.347754,\ \ y(9.5)(Newton) = 29.745801]$

8. Use Newton' formulas to obtain $y(1.5), y(3)$ and $y(6)$ from the following readings

x	1	2	4	7
y	12	18	36	16

9. Given the table of readings:

x	1	1.5	2	2.5	3
y	3.00	2.36	1.84	1.43	1.11

By the use of different interpolation formulas, find $y(1.25)$, $y(1.75)$, $y(2.25)$ and $y(2.75)$.

$[Ans.\ y(1.25) = 2.664766,\ \ y(2.25) = 1.622266]$

10. Use different interpolation formulas to find y when $x = 3.25$ by using the readings

x	2.5	3	3.5	4	4.5
y	2.07	2.75	3.37	3.94	4.49

11. Use Gauss forward and Gauss backward formulas to obtain $f(5)$ by using the following readings

x	0	2	4	6	8
y	3	5	9	16	30

12. Use Gauss forward and Gauss backward, to find $y(0.7)$ from the readings

x	0.4	0.6	0.8	1.0
y	0.3799	0.5370	0.6640	0.7616

13. Use a suitable interpolation formula to find $y(0.5)$, $y(2.5)$, and $y(3.5)$ from the readings:

x	0	1	2	3	4
y	3	4	19	84	259

14. When the steady-state heat-flow equation is solved numerically, temperature $T(x)$ is calculated at the nodes of a gridwork constructed in the domain of interest. When a certain problem was solved, the values given in the table below were obtained. This procedure does not give the temperature at points other than the nodes of the grid; if they are desired, one can interpolate to find them. Use the data to estimate the temperature at the points $x = 0.7, 1.6\ and\ 2.8$.

x	0	0.5	1.0	1.5	2.0	2.5	3.0
y	5	3.56	3	5.56	15	36.56	77

15. In studies of radiation-induced polymerization, a source of gamma rays was employed to give measured doses of radiation. However, the dosage varied with position in the apparatus, with these figures being recorded:

Position (in) from base pt.	0	0.5	1.0	1.5	2.0	3	3.5	4
Dosage rad/hr	19	2.39	2.71	2.98	3.20	3.20	2.98	2.74

For some reason, the reading at 2.5 inches was not reported, but the value of reading there is needed. Fit interpolating polynomials of various degrees to the data to supply the missing information. What do you think is the best estimate for the dosage level at 2.5 inches?

4. Numerical Solution of Ordinary Differential Equations

In this chapter we discuss some of the numerical methods, which are used in solving ordinary differential equation. There are at present a large number of numerical methods for solving differential equations. Since our treatment is admittedly introductory and brief, we touch upon only a few of the elementary methods.

In our subsequent discussion, we assume that the differential equation in question is of the form

$$\frac{dy}{dx} = y' = f(x, y),$$

where $\frac{dy}{dx}$ means the change in y with respect to x and f is any function of the independent variable x and the unknown variable y. We note that all differential equations can be written in this form or as a system of differential equations of this form. Hence, from one point of view there is no loss of generality in making this assumption. Furthermore, to avoid cumbersome treatment of integration constants, we concentrate on initial value problems. Thus we investigate the initial value problems

$$y' = f(x, y), \tag{1}$$

$$y(x_0) = y_0 \tag{2}$$

To obtain a feeling for the goal of numerical methods, let us imagine that the curve in Fig. 4.1 represents the solution to the initial value problem (1)–(2). Suppose that we want to know the coordinates of the

point $(x_1,\ y_1)$. In the case that y is given explicitly by a formula involving x, it is simply a matter of knowing x_1.

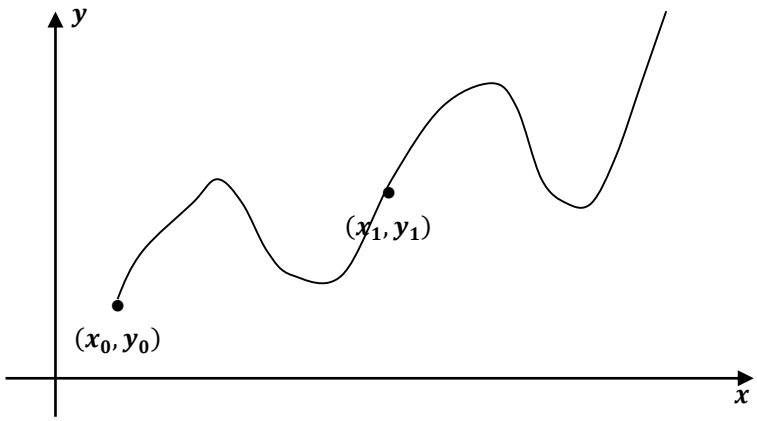


Fig. 4.1

If y is not known in terms of x and the solution curve of Fig. 4.1 exists only in theory, is it still possible to determine y_1 to knowing x_1? The answer to this question is the basic goal of any numerical method.

We will say that we have solved a differential equation numerically on the interval $[a,\ b]$ when we have obtained a set of points $(x_0,\ y_0)$, $(x_1,\ y_1), \cdots, (x_N,\ y_N)$ with $x_0 = a$, $x_N = b$. Indeed y_n on the solution curve is the approximation of the solution when x has the value x_n.

4.1 Picard's method

Consider the initial value problems

$$y' = f(x, y) \tag{1}$$

$$y(x_0) = y_0 \tag{2}$$

From the differential equation we have

$$dy = f(x, y)\, dx,$$

which implies by integration

$$\int_{y_0}^{y} dy = \int_{x_0}^{x} f(x,y)\,dx$$

The integration of the left hand side gives

$$y - y_0 = \int_{x_0}^{x} f(x,y)\,dx$$

or

$$y = y_0 + \int_{x_0}^{x} f(x,y)\,dx \qquad (3)$$

If $x - x_0$ is small enough, y will not be much different from y_0 in that integral. We can substitute $y = y_0$ as an approximation in the integral formula (3) to get the first approximation y_1:

$$y_1 = y_0 + \int_{x_0}^{x} f(x,y_0)\,dx,$$

which is a better approximation. Repeat this process to obtain the second, the third, the fourth, ⋯, approximations until the required accuracy is obtained.

In general we can write the $(n+1)th$ approximation in the from

$$\boxed{y_{n+1} = y_0 + \int_{x_0}^{x} f(x,y_n)\,dx\,, \quad n = 0,1,\cdots} \qquad (4)$$

Example 1 Solve by Picard's method the differential equation

$$y' = x^2 + y,$$

up to the third approximation, given that $y = 0$ at $x = 0$.

Solution Here we have

$$x_0 = 0, \quad y_0 = 0, \qquad f(x,y) = x^2 + y\,.$$

Substituting into equation (4) we find:

$$y_{n+1} = 0 + \int_0^x (x^2 + y_n)\,dx\,, \quad n = 0,1,\cdots \tag{5}$$

For $n = 0$, we get

$$y_1 = \int_0^x (x^2 + y_0)\,dx = \int_0^x (x^2 + 0)\,dx = \frac{x^3}{3}$$

Put $n = 1$, in (5) to obtain

$$y_2 = \int_0^x (x^2 + y_1)\,dx = \int_0^x \left(x^2 + \frac{x^3}{3}\right) dx = \frac{x^3}{3} + \frac{x^4}{12}$$

To find the third approximation we put $n = 2$ in (5) to get

$$y_3 = \int_0^x (x^2 + y_2)\,dx = \int_0^x \left(x^2 + \frac{x^3}{3} + \frac{x^4}{12}\right) dx$$

$$= \frac{x^3}{3} + \frac{x^4}{12} + \frac{x^5}{60}$$

Thus the approximate solution

$$y(x) = \frac{x^3}{3} + \frac{x^4}{12} + \frac{x^5}{60}$$

Now we can use this formula to evaluate the solution on the interval $(0,1)$ with step size $h = 0.2$ which is shown in table 3.1

x	exact	approx
0	0	0
0.2	0.002806	0.002805
0.4	0.023649	0.023637
0.6	0.084238	0.084096
0.8	0.211082	0.210261
1.0	0.436564	0.433333

Table 4.1. Values of the exact and approximate solution at the mesh points

Example 2 Solve $y' = x - y^2$ by using Picard's method up to the third approximation, given that $y(1) = -1$.

Solution Since $x_0 = 1$ and $y_0 = -1$, then we use the axis transformation:

$$\left.\begin{aligned} X &= x - x_0 = x - 1 \\ Y &= y - y_0 = y + 1 \end{aligned}\right\}, \tag{6}$$

to simplify the calculations. Thus

$$\frac{dy}{dx} = \frac{dy}{dY} \cdot \frac{dY}{dX} \cdot \frac{dX}{dx} = \frac{dY}{dX} \tag{7}$$

On substituting with (6) and (7) in the given differential equation we get the initial value problem

$$y' = x - y^2 \;\Rightarrow\; \frac{dY}{dX} = (X + 1) - (Y - 1)^2,$$

or

$$\frac{dY}{dX} = X + 2Y - Y^2, \;\; X_0 = 0, \;\; Y_0 = 0$$

From formula (4) we get

$$\boxed{Y_{n+1} = \int_0^X (X + 2Y_n - Y_n^2)\, dX\,, \quad n = 0,1,\cdots} \tag{8}$$

From which we obtain

$$Y_1 = \int_0^X (X + 0)\, dX = \frac{X^2}{2},$$

$$Y_2 = \int_0^X (X + 2Y_1 - Y_1^2)\, dX$$

$$= \int_0^X \left[X + 2\left(\frac{X^2}{2}\right) - \left(\frac{X^2}{2}\right)^2\right] dX = \frac{X^2}{2} + \frac{X^3}{3} - \frac{X^5}{20},$$

$$Y_3 = \int_0^X (X + 2Y_2 - Y_2^2)\, dX$$

$$= \int_0^X \left[X + 2\left(\frac{X^2}{2} + \frac{X^3}{3} - \frac{X^5}{20}\right) - \left(\frac{X^2}{2} + \frac{X^3}{3} - \frac{X^5}{20}\right)^2\right] dX$$

$$= \int_0^X \left[X + X^2 + \frac{2X^3}{3} - \frac{X^5}{10} - \left(\frac{X^4}{4} + \frac{X^6}{9} + \frac{X^{10}}{400} + \frac{X^5}{3} - \frac{X^7}{20} - \frac{X^8}{30}\right)\right] dX$$

$$= \int_0^X \left[X + X^2 + \frac{2X^3}{3} - \frac{X^4}{4} - \frac{13X^5}{30} - \frac{X^6}{9} + \frac{X^7}{20} + \frac{X^8}{30} - \frac{X^{10}}{400}\right] dX$$

$$= \frac{X^2}{2} + \frac{X^3}{3} + \frac{X^4}{6} - \frac{X^5}{20} - \frac{13X^6}{180} - \frac{X^7}{63} + \frac{X^8}{160} + \frac{X^9}{270} - \frac{X^{11}}{4400}$$

Therefore, the third approximation of the given initial value problem is

$$y + 1 = \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} + \frac{(x-1)^4}{6} - \frac{(x-1)^5}{20}$$
$$- \frac{13(x-1)^6}{180} - \frac{(x-1)^7}{63} + \frac{(x-1)^8}{160} + \frac{(x-1)^9}{270} - \frac{(x-1)^{11}}{4400}$$

4.2 Euler method

Euler method is the simplest numerical method for solving the initial value problem

$$y' = f(x, y), \quad a \le x \le b, \tag{1}$$

$$y(x_0) = y(a) = y_0 \tag{2}$$

Referring to Fig. 5.2, the slope of the solution curve (assuming it is known) at the initial point $(x_0,\ y_0)$ is given and we may approximate it by

$$y_0' = f(x_0, y_0) \cong \frac{y_1 - y_0}{x_1 - x_0},$$

or

$$y_1 = y_0 + (x_1 - x_0)f(x_0, y_0)$$

Defining $(x_1 - x_0) = h$, which is a positive real number called the step size, then we obtain the formula:

$$\begin{aligned} x_1 &= x_0 + h \\ y_1 &= y_0 + hf_0 \end{aligned} \tag{3}$$

Of course, if this approximation is to be reasonably accurate, then h should be quite small (see Fig. 3.2).

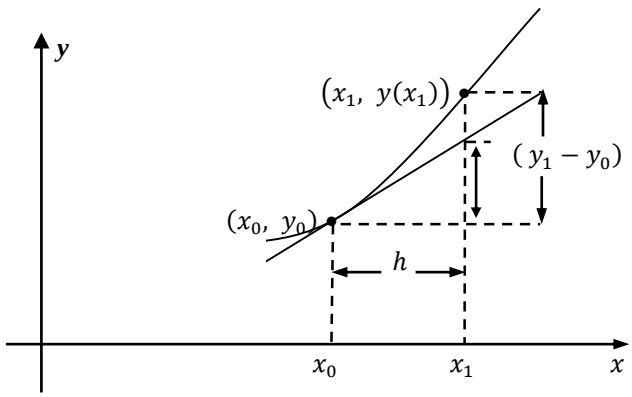


Fig. 4.2

This process can then be repeated to obtain additional points. Usually, this process is carried out in a specified interval with a prescribed spacing between the x coordinates. The spacing is normally taken to be uniform and is denoted by h. Thus Euler method is used for solving the initial value problem $(1) - (2)$ at discrete set of nodes For simplicity, this sequence of points are taken to be equal spaced and the approximate solution will be estimated at these points:

$$\boxed{\begin{aligned} x_{n+1} &= x_n + h, \qquad h = \frac{b-a}{N} \\ y_{n+1} &= y_n + hf_n, \qquad n = 0,1,\cdots,N \end{aligned}} \tag{4}$$

where we denote by $f_n \equiv f(x_n, y_n)$. Euler method is symbolically represented by Fig. 4.3.

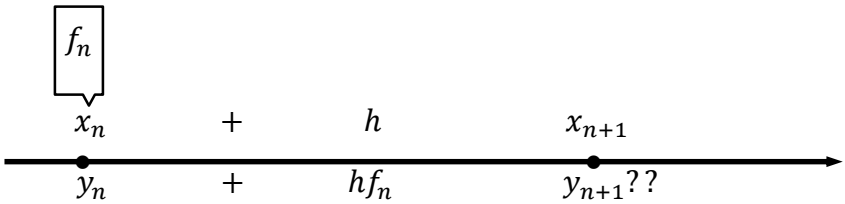


Fig. 4.3

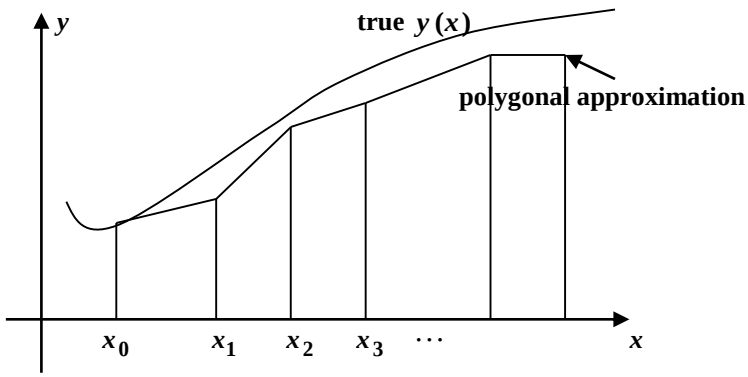


Fig. 4.4

The solution to the initial value problem is then the sequence $\{(x_n, y_n)\}_{n=0}^{N}$. This solution is usually displayed in tabular form, although it could be displayed geometrically by plotting each of the points (See Fig. 4.4).

Example 1 Given that

$$y' = x^2 + y, \qquad y(0) = 1,$$

obtain by Euler method the value of y at $0.02, 0.04, 0.06, 0.08$ and 0.1.

Solution Here we have

$$f(x,y) = x^2 + y, \qquad x_0 = 0, \qquad y_0 = 1, \qquad h = 0.02, \qquad (7)$$

thus formulas (6) gives

$$\boxed{\begin{aligned} &x_{n+1} = x_n + 0.02, \\ &y_{n+1} = y_n + 0.02\left(x_n^2 + y_n\right), \qquad n = 0, 1, 2, \cdots, \end{aligned}}$$

from which we have

$$\begin{aligned} x_1 &= x_0 + 0.02 = 0.02, \\ y_1 &= y(0.02) = y_0 + 0.02(x_0^2 + y_0) \\ &= 1 + 0.02(0 + 1) = 1.02, \end{aligned}$$

$$\begin{aligned} x_2 &= x_1 + 0.02 = 0.04, \\ y_2 &= y(0.04) = y_1 + 0.02(x_1^2 + y_1) \\ &= 1.02 + 0.02[(0.02)^2 + 1.02] = 1.040408, \end{aligned}$$

$$x_3 = x_2 + 0.02 = 0.06,$$

$$\begin{aligned} y_3 &= y(0.06) = y_2 + 0.02(x_2^2 + y_2) \\ &= 1.040408 + 0.02[(0.04)^2 + 1.040408] = 1.06124816, \end{aligned}$$

$$x_4 = x_3 + 0.02 = 0.08,$$

$$\begin{aligned} y_4 &= y(0.08) = y_3 + 0.02(x_3^2 + y_3) \\ &= 1.06124816 + 0.02[(0.06)^2 + 1.06124816] \\ &= 1.082545123, \end{aligned}$$

and

$$x_5 = x_4 + 0.02 = 0.1,$$

$$\begin{aligned} y_5 &= y(0.1) = y_4 + 0.02(x_4^2 + y_4) \\ &= 1.082545123 + 0.02[(0.08)^2 + 1.082545123] \\ &= 1.104324026. \end{aligned}$$

Example 2 Use Euler method to obtain a numerical solution of the initial value problem

$$y' = x^2 + y^2, \qquad y(0) = 0, \tag{8}$$

on the interval $0 \le x \le 1$.

Solution We were not told how many points to obtain; therefore, the choice is at our disposal. The more points we ask for, the smaller the value of h will be and so the more accurate our results will be. The price to be paid, of course, is that the smaller h is, the more calculations we require. We choose $N = 5$. Thus

$$h = \frac{1-0}{5} = 0.2.$$

Therefore, we have the results:

$$f(x,y) = x^2 + y^2,\ x_0 = 0,\ y_0 = 0,\ h = 0.2.$$

So that

$$\boxed{\begin{aligned} &x_{n+1} = x_n + 0.2, \\ &y_{n+1} = y_n + 0.2(x_n^2 + y_n^2), \qquad n = 0,1,2,3,4 \end{aligned}}$$

Now we put the counter $n = 0,1,2,3,4$, to get

$$\begin{aligned} x_1 &= x_0 + 0.2 = 0.2, \\ y_1 &= y(0.2) = y_0 + 0.2(x_0^2 + y_0^2) \\ &= 0 + 0.2(0+0) = 0, \\ x_2 &= x_1 + 0.2 = 0.4, \\ y_2 &= y(0.4) = y_1 + +0.2(x_1^2 + y_1^2) \\ &= 0 + 0.2[(0.2)^2 + (0)^2] = 0.008, \end{aligned}$$

$$x_3 = x_2 + 0.2 = 0.6,$$
$$y_3 = y(0.6) = y_2 + +0.2(x_2^2 + y_2^2)$$
$$= 0.008 + 0.2[(0.4)^2 + (0.008)^2] = 0.0400128,$$
$$x_4 = x_3 + 0.2 = 0.8,$$
$$y_4 = y(0.8) = y_3 + +0.2(x_3^2 + y_3^2)$$
$$= 0.0400128 + 0.2[(0.6)^2 + (0.0400128)^2]$$
$$= 0.112333005,$$

and

$$x_5 = x_4 + 0.2 = 1,$$
$$y_5 = y(1) = y_4 + +0.2(x_4^2 + y_4^2)$$
$$= 0.112333005 + 0.2[(0.8)^2 + (0.112333005)^2]$$
$$= 0.242856746,$$

Example 3 Solve by Euler method the initial boundary value problem

$$y' = x - y + 1, \quad y(0) = 1, \ 0 \le x \le 1 \tag{9}$$

Estimate the accuracy of your results (**Hint:** the exact solution is $y(x) = x + e^{-x}$).

Solution Suppose $N = 5, 10, 20$ so that $h = 0.2, 0.1, 0.05$. Using

$$f(x, y) = x - y + 1, \ x_0 = 0, \ y_0 = 1,$$

and

$$\boxed{\begin{aligned} &x_{n+1} = x_n + h, \\ &y_{n+1} = y_n + h(x_n - y_n + 1), \quad n = 0, 1, \cdots, N-1 \end{aligned}}$$

On substituting in these formulas, we get

x	$Exact\ sol.$	$Euler\ Approx.\ sol.$		
		$h = 0.2$	$h = 0.1$	$h = 0.05$
0	1	1	1	1
0.2	1.018730753	1	1.01	1.01450625
0.4	1.070320046	1.04	1.0561	1.06342043
0.6	1.148811636	1.112	1.131441	1.14036009
0.8	1.249328964	1.2096	1.23046721	1.24012667
1.0	1.367879441	1.32768	1.34867844	1.35848592

The following table shows the comparison between the approximate value and the actual values.

x	$Absolute\ error = \lvert Exact\ sol. - approx.\ sol. \rvert$		
	$h = 0.2$	$h = 0.1$	$h = 0.05$
0	0	0	0
0.2	0.0000	0.0000	0.0000
0.4	0.0187	0.0087	0.0042
0.6	0.0303	0.0142	0.0069
0.8	0.0368	0.0174	0.0085
1.0	0.0397	0.0189	0.0092

Note that the error grows slightly as the value of x_n increases. The error become larger with each successive point. Accuracy can be improved by using small h.

4.3 Modified Euler method

Given the initial value problem

$$y' = f(x, y), \quad a \le x \le b, \tag{1}$$

$$y(x_0) = y(a) = y_0, \tag{2}$$

and assuming that y_n is given, we want to find y_{n+1} at $x_{n+1} = x_n + h$ (See Fig. 5.6). In the modified Euler method, instead using of the slope of the tangent line at the point (x_n, y_n), the average of slopes at the points (x_n, y_n) and (x_{n+1}, y_{n+1}). In this formula y_{n+1} is first predicted by the use of Euler method so we have the modified formula The following steps illustrate the modification of Euler method for all $n = 0,1,\cdots,N$:

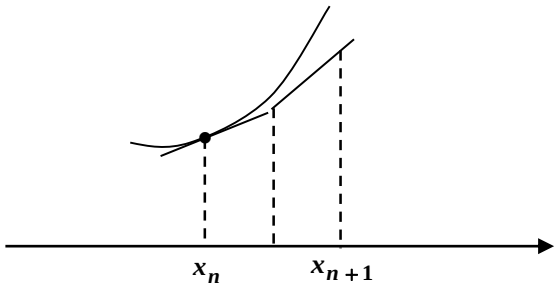
1. In each step of this method the first approximation (predicted value) is

$$\boxed{\begin{aligned} x_{n+1} &= x_n + h, \\ y^*_{n+1} &= y_n + hf_n \end{aligned}} \tag{3}$$

2. In the second step we use the predicted value to get new value (corrected value)

$$\boxed{y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})]} \tag{4}$$

This method has a simple geometric interpretation as shown in Fig. 3.5. In the interval from x_n to $x_n + h/2$ we approximate the solution y by the straight line through (x_n, y_n) with slope $f(x_n, y_n)$ and then we continue along the straight line with slope $f(x_{n+1}, y^*_{n+1})$ until x reaches x_{n+1}.



Denoting by $f_{n+1}^* \equiv f(x_{n+1}, y_{n+1}^*)$, the modified Euler method is symbolically represented by Fig. 3.6.

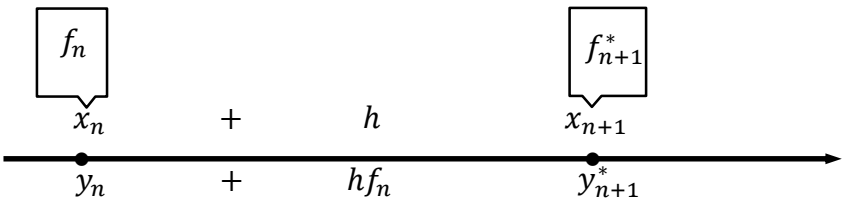


Fig. 4.6

$$\boxed{\begin{aligned} x_{n+1} &= x_n + h, \\ y_{n+1}^* &= y_n + hf_n, \\ y_{n+1} &= y_n + \frac{h}{2}(f_n + f_{n+1}^*), n = 0,1,\cdots,N \end{aligned}} \tag{5}$$

Example 1 Use modified Euler's method to obtain the value of y at $x = 0.1$ for the problem

$$y' = x^2 + y, \qquad y(0) = 1, \tag{6}$$

Solution Here $f(x, y) = x^2 + y$, $h = 0.1$ and we see that

$f_0 = f(0,1) = 0 + 1 = 1,$

$$x_1 = x_0 + h = 0 + 0.1 = 0.1$$

$$y_1^* = y_0 + hf_0 = 1 + 0.1 \times 1 = 1.1 \quad \Rightarrow \quad f_1^* = f(0.1,1.1) = 1.11$$

$$y_1 = y_0 + \frac{h}{2}(f_0 + f_1^*)$$

$$= 1 + \frac{0.1}{2}[1 + 1.11] = 1.1055$$

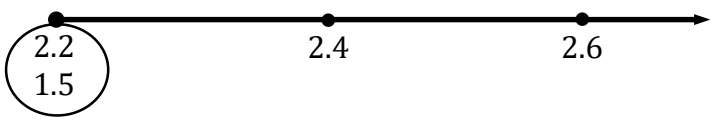
Example 2 Find the value of y at $x = 2.6$ by using modified Euler's method with $h = 0.2$, if

$$y' = \sqrt{x^2 + 4y^2}, \qquad y(2.2) = 1.5 \tag{7}$$

Solution Here we use the modified Euler method two times with

$$f(x, y) = \sqrt{x^2 + 4y^2}, \;\; x_0 = 2.2, \;\; y_0 = 1.5,$$

$$\Rightarrow \qquad f_0 = f(2.2,1.5) = 3.72021505$$



<u>At $x_1 = 2.4$</u>

$$x_1 = x_0 + h = 2.2 + 0.2 = 2.4,$$

$$y_1^* = y_0 + hf_0 = 1.5 + 0.2(3.72021505) = 2.24404301,$$

$$\Rightarrow \qquad f_1^* = f(2.4, 2.24404301) = 5.08949075$$

$$y_1 = y_0 + \frac{h}{2}(f_0 + f_1^*)$$

$$= 1.5 + \frac{0.2}{2}(3.72021505 + 5.08949075) = 2.38097058$$

To get the value of y at $x_2 = 2.6$, we use the information of the previous point $y(2.4) = 2.38097058$, and

$$f_1 = f(2.4, 2.38097058\ ) = 5.33254945,$$

thus

$$x_2 = x_1 + h = 2.4 + 0.2 = 2.6,$$

$$y_2^* = y_1 + hf_1$$
$$= 2.38097058 + 0.2(5.33254945) = 3.447480469$$

$$\Rightarrow \quad f_2^* = f(2.6, 3.447480469) = 7.36888637$$

$$y_2 = y_1 + \frac{h}{2}(f_1 + f_2^*)$$

$$= 2.38097058 + \frac{0.2}{2}(5.33254945 + 7.36888637) = 3.651114162$$

4.3 Taylor-series method

Taylor series can be used to generate series representations of solutions of an initial value problem at specific values of x. Given the initial value problem

$$y' = f(x, y), \quad a \le x \le b, \tag{1}$$

$$y(x_0) = y(a) = y_0 \tag{2}$$

We again wish to obtain a sequence of points (x_n, y_n), $n = 1,2,\cdots,N$ that lie close to the solution of the curve on the interval $[a, b]$. Recall that the Taylor series representation of the solution about $x_0 = a$ is

$$y(x) = \sum_{r=0}^{\infty} \frac{y^{(r)}(x_0)}{r\,!}(x - x_0)^r,$$

where, $y^{(r)}(x_0) = \left.\frac{d^r y}{dx^r}\right|_{x=x_0}$.

Setting $x = x_1 = x_0 + h$, we have

$$y(x_1) = y(x_0 + h) = \sum_{r=0}^{\infty} \frac{y^{(r)}(x_0)}{r\,!} h^r$$

$$= y(x_0) + \frac{y'(x_0)}{1\,!}h + \frac{y''(x_0)}{2\,!}h^2 + \frac{y'''(x_0)}{3\,!}h^3 + \cdots \tag{3}$$

If we assume that all partial derivatives of $f(x, y)$ of any order exist at $(x_0,\ y_0)$, then the differential equation (1) with the boundary conditions (2) can be used to determine each $y^{(r)}(x_0)$, $r = 1,2,\cdots$, and consequently $y(x)$. In general Taylor series (3) is approximated by

$$\boxed{\begin{aligned} x_{n+1} &= x_n + h, \\ y_{n+1} &= y_n + hf(x_n, y_n) + \cdots + \frac{h^k}{k\,!} f^{(k-1)}(x_n, y_n) \end{aligned}} \qquad (4)$$

where $h = x_{n+1} - x_n$, is called the step size and we say that the approximation (4) is of order k.

Knowing $(x_1,\ y_1)$ and x_2, we can then calculate y_2 in a similar manner. Thus

$$y_2 = y(x_2) = y(x_1 + h) = \sum_{r=0}^{\infty} \frac{y^{(r)}(x_1)}{r\,!} h^r$$

Additional points are obtained in the same way.

In any application of the Taylor method, one must specify the order k of the approximation. The larger k is and smaller h is, the more accurate we anticipate the results to be.

To obtain the values of the derivatives in (4) we have two methods:

1. Either differentiating the differential equation (1) successively with respect to x and we compute the values of these derivatives at point $(x_n,\ y_n)$, or

2. Differentiating the differential equation (1) m -times by Leibentz formula to get a recurrence relation for the derivatives, then substitute the point $(x_n,\ y_n)$ to calculate the values of these derivatives.

Example 1 Use the Taylor method of order 4 to obtain $y(0.2)$ if

$$y' = \sin y + x, \qquad y(0) = 0 \tag{5}$$

Solution Since $y(0) = 0$, then

$y' = \sin y + x$	$y'(0) = 0$
$y'' = y' \cos y + 1$	$y''(0) = 1$
$y''' = y'' \cos y - (y')^2 \sin y$	$y'''(0) = 1$
$y^{(4)} = y''' \cos y - 3y'y'' \sin y - (y')^3 \cos y$	$y^{(4)}(0) = 1$

Hence, we have the fourth order Taylor series in the form

$$y(x) = y(0) + \frac{y'(0)}{1\,!}h + \frac{y''(0)}{2\,!}h^2 + \frac{y'''(0)}{3\,!}h^3 + \frac{y^{(4)}(0)}{4\,!}h^4 + \cdots$$

$$= 0 + 0 + \frac{1}{2\,!}h^2 + \frac{1}{3\,!}h^3 + \frac{1}{4\,!}h^4 + \cdots \tag{6}$$

Taking $h = 0.2$ in the expansion (6), we get

$$y(0.2) = \frac{(0.2)^2}{2\,!} + \frac{(0.2)^3}{3\,!} + \frac{(0.2)^4}{4\,!} + \cdots = 0.021466667$$

Example 2 Use the Taylor approximation of order 2 to obtain a numerical solution of the initial value problem

$$y' = x^2 + y^2, \qquad y(0) = 0, \tag{7}$$

on the interval $0 \le x \le 1$.

Solution We choose $h = 0.2$. From the given differential equation, we obtain

$$y' = x^2 + y^2$$

$$y'' = 2x + 2yy'$$

$$= 2x + 2y(x^2 + y^2) = 2x + 2x^2y + 2y^3$$

In general, we have

$$y_{n+1} = y(x_{n+1}) = y(x_n) + \frac{h}{1!}y'(x_n) + \frac{h^2}{2!}y''(x_n) + \cdots$$

$$= y_n + \frac{h}{1!}(x_n^2 + y_n^2) + h^2(x_n + x_n^2 y_n + y_n^3) + \cdots \qquad (8)$$

Thus, our numerical solution of the initial value problem (7) on the interval $0 \le x \le 1$ is calculated from (8) as:

n	x	approx
0	0	0
1	0.2	0
2	0.4	0.016
3	0.6	0.064153764
4	0.8	0.161911281
5	1.0	0.331469044

Remark Taylor method can be used to solve differential equations of kth order if we are given the values of $y, y', y'', \cdots, y^{(k-1)}$ at the point $(x_0,\ y_0)$. This is illustrated by the following example.

Example 3 Use the Taylor approximation of order 4 to obtain $y(0.1)$ if

$$y'' + xy' + 2y = 0, \quad y(0) = 3,\ \ y'(0) = -2, \qquad (9)$$

Solution Since $x_0 = 0,\ y(0) = 3,\ \ y'(0) = -2,\ h = 0.1,$ then

$\boldsymbol{y'' = -xy' - 2y}$	$\boldsymbol{y''(0) = 0 - 2y(0) = -6}$
$y''' = -xy'' - 3y'$	$y'''(0) = 0 - 3y'(0) = 6$
$y^{(4)} = -xy''' - 4y''$	$y^{(4)}(0) = -4y''(0) = 24$

We have the fourth order Taylor series in the form

$$y(x_1) \cong y(0) + \frac{y'(0)}{1\,!}h + \frac{y''(0)}{2\,!}h^2 + \frac{y'''(0)}{3\,!}h^3 + \frac{y^{(4)}(0)}{4\,!}h^4$$

$$= 3 + \frac{(-2)}{1\,!}h + \frac{(-6)}{2\,!}h^2 + \frac{(6)}{3\,!}h^3 + \frac{(24)}{4\,!}h^4 \qquad (10)$$

Thus

$$y(0.1) \cong 3 - 2(0.1) - 3(0.1)^2 + (0.1)^3 + (0.1)^4 = 2.7711$$

Example 4 Solve by Taylor's method the differential equation

$$y'' + 3xy' - 6y = 0, \qquad (11)$$

given that

$$y(0) = 1,\ \ y'(0) = 0.1 \qquad (12)$$

Solution Taylor series of the approximate solution is

$$y(x) = 1 + \frac{(x-0)}{1\,!}y'(0) + \frac{(x-0)^2}{2\,!}y''(0) + \frac{(x-0)^3}{3\,!}y'''(0)$$

$$+ \frac{(x-0)^4}{4\,!}y^{(4)}(0) + \cdots$$

$$= 1 + xy'(0) + \frac{x^2}{2\,!}y''(0) + \frac{x^3}{3\,!}y'''(0) + \frac{x^4}{4\,!}y^{(4)}(0) + \cdots \qquad (13)$$

All derivatives are taken at the point $(0, 1)$, thus

$$y_0' = y'(0) = 0.1,\ \ y_0'' = y''(0) = 6y_0 = 6$$

Equation (11) can rewritten in the form

$$y'' = 6y - 3xy' \tag{14}$$

Differentiating this equation k −times by using Leibentz theorem, we get

$$y^{(k+2)} = 6y^{(k)} - 3\left[(x)\left(y^{(k+1)}\right) + k(1)\left(\left(y^{(k)}\right)\right)\right]$$

$$= 6y^{(k)} - 3xy^{(k+1)} - 3ky^{(k)} = (6-3k)y^{(k)} - 3xy^{(k+1)}$$

<u>At $x = 0$,</u> we get

$$y^{(k+2)}(0) = (6-3k)y^{(k)}(0) \tag{15}$$

from which we have for $k = 2$,

$$y^{(4)}(0) = (6 - 3 \times 2)y^{(k)}(0) = 0,$$

and so

$$y^{(6)}(0) = y^{(8)}(0) = \cdots = 0,$$

and for:

$\boldsymbol{k = 1}$	$\boldsymbol{y'''(0) = 3y'(0) = 0.3}$
$k = 3$	$y^{(5)}(0) = -3y'''(0) = -0.9$
$k = 5$	$y^{(7)}(0) = -9y^{(5)}(0) = 8.1$

Then we get from the substitution in (13)

$$y(x) = 1 + 0.1x + \frac{0.3}{3\,!}x^3 - \frac{0.9}{5\,!}x^5 + \frac{0.81}{7\,!}x^7 + \cdots$$

$$= 1 + 0.1x + \frac{1}{20}x^3 - \frac{3}{400}x^5 + \frac{9}{5600}x^7 + \cdots$$

Example 5 Given that

$$y' = 1 - \frac{y}{x}, \qquad y(2) = 2,$$

find to four correct decimal places the value of y when $x = 2.05$ by using Taylor series.

Solution Since $x_0 = 2,\ y_0 = 2$, then

$\boldsymbol{y' = 1 - \frac{y}{x}}$	$\boldsymbol{y'_0 = 1 - 1 = 0}$
$y'' = \frac{y}{x^2} - \frac{y'}{x}$	$y''_0 = \frac{2}{4} - 0 = \frac{1}{2}$
$y''' = \frac{2y'}{x^2} - \frac{2y}{x^3} - \frac{y''}{x}$	$y'''_0 = 0 - \frac{2(2)}{(2)^3} - \frac{(1/2)}{(2)} = -\frac{3}{4}$

Substituting in Taylor series with $h = 0.05$, we get

$$y_1 = y_0 + \frac{y'_0}{1\,!}h + \frac{y''_0}{2\,!}h^2 + \frac{y'''_0}{3\,!}h^3 + \cdots$$

$$= 2 + 0 + \frac{(1/2)}{2\,!}(0.05)^2 + \frac{(-3/4)}{3\,!}(0.05)^3$$

$$\cong 2.000609375$$

Exercise 4.1

1. Find $y(0.9)$ by the use of Taylor method if

$$2y' = \ln(x + y), \qquad y(0.7) = 1.575864$$

[$\boldsymbol{Answer}$: $y(0.9) = 1.65809987$]

2. Solve by Picard's method the initial value problem:

$$y' = 2x + y, \quad y(0) = 0$$

up to the third approximation. Hence evaluate $y(0.1)$.

3. Evaluate $y(0)$ by the use of Taylor method if

$$5y' = 4(\sqrt{x} + \sqrt{y}), \quad y(0.25) = 1,$$

[$\boldsymbol{Answer}$: $y(0) = 0.703583$]

4. Solve the following initial value problems

$$y' = \left(\frac{y}{x}\right)^2 + \left(\frac{y}{x}\right), \ 1 \le x \le 1.2 \quad y(1) = 1, \ h = 0.1$$

by the use of, Euler's method, the Modified Euler's method, Taylor method.

5. Given the initial value problem

$$4y' = 4x^2 + y, \qquad y(1) = 2.4$$

(i) Find $y(1.2)$ by the use of Taylor method

(ii) Find $y(1.4)$ by the use of Euler method. $y(1.2)$

(iii) Find $y(1.3)$ and $y(1.6)$ by the use of modified Euler method.

[$\boldsymbol{Answer}$: $y = 2.77146667, \ 3.19804, \ 2.999919417, \ 3.82374$]

4.5 Runge-Kutta methods

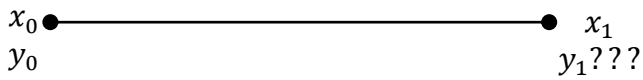
As mentioned previously, Euler's method is not very useful in practical problems because it requires a very small step size for reasonable accuracy. Taylor's algorithm of higher order is unacceptable as a general-purpose procedure because of the need to obtain higher total derivative of $y(x)$. The Runge-Kutta methods attempt to obtain greater accuracy, and at the same time avoid the need for higher derivatives, by evaluating the function $f(x, y)$ at selected points on each subinterval. In fact the procedure of Runge-Kutta is an alternative approach to obtaining high accuracy using points that lie between x_n and x_{n+1}; the values of the solution at these points have no importance as they are only temporary.

Again, given the initial value problem

$$y' = f(x, y), \quad a \le x \le b, \tag{1}$$

$$y(x_0) = y(a) = y_0 \tag{2}$$

and we want to find y_1 at $x_1 = x_0 + h$. In the description of the Taylor approximation of order 2, we wrote



$$y_1 = y(x_0 + h) = y(x_0) + hf(x_0, y_0) + \frac{h^2}{2!} f'(x_0, y_0)$$

Since

$$f'(x_0, y_0) = f_x(x_0, y_0) + f_y(x_0, y_0) \cdot y_0',$$

and

$$y_0' = f(x_0, y_0)$$

This implies

$$y(x_0 + h) = y(x_0) + hf(x_0, y_0) + \frac{h^2}{2!}\left(f_x + f_y f\right)_{(x_0, y_0)}$$

$$= y(x_0) + h\left[f(x_0, y_0) + \frac{h}{2}\left(f_x + f_y f\right)_{(x_0, y_0)}\right] \tag{3}$$

The first step in deriving Runge-Kutta method is to evaluate values for $a, \alpha, and\ \beta$ with the property that

$$af(x_0 + \alpha, y_0 + \beta) = \left[f + \frac{h}{2}\left(f_x + f f_y\right)\right]_{(x_0, y_0)} \tag{4}$$

From Taylor's theorem for the expansion of functions of two variables, we write the left hand side of (4) as

$$af(x_0 + \alpha, y_0 + \beta) = a\left[f + \alpha f_x + \beta f_y\right]_{(x_0, y_0)} \tag{5}$$

Substituting from (5) into (4) and matching the coefficients of f and its derivatives give the three equations

$$f(x_0, y_0): \qquad a = 1$$

$$f_x(x_0, y_0): \qquad a\alpha = \frac{h}{2}$$

$$f_y(x_0, y_0): \qquad a\beta = \frac{h}{2} f(x_0, y_0)$$

The parameters $a, \alpha, and\ \beta$ are uniquely determined to be

$$a = 1, \qquad \alpha = \frac{h}{2}, \qquad \beta = \frac{h}{2} f(x_0, y_0),$$

so, equation (3) is rewritten as

$$y(x_0 + h) = y(x_0) + hf\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} f(x_0, y_0)\right) \tag{6}$$

In general,

$$\boxed{y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{k}{2}\right)} \tag{7}$$

where

$$\boxed{k = hf(x_n, y_n)} \tag{8}$$

Formulas (7) and (8) are called the ***Runge-Kutta approximation of order two***.

Since only three parameters are present in $af(x_0 + \alpha, y_0 + \beta)$ and all are needed in the match of $(x_n, y_n) + \frac{h}{2} f'(x_n, y_n)$, we need a more complicated form to satisfy the conditions required for any of the higher order Taylor methods. As mentioned above, the basic idea of the Runge-Kutta methods is to preserve the order of a Taylor approximation while eliminating the necessity of calculating the various partial derivatives of f that are involved. The alternative

proposed by these methods involves evaluating the function f at certain judicious points rather than evaluating the specific partial derivatives.

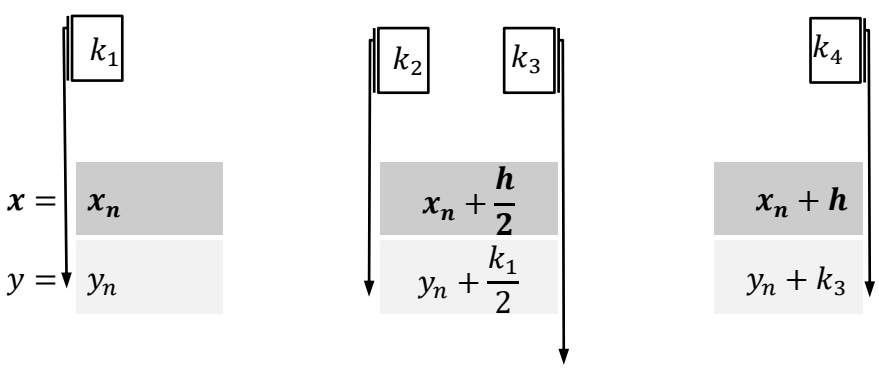
We present here the most common Runge-Kutta method in use. This is the Runge-Kutta approximation of order four (RK4) and is defined by the following formulas for each $n = 0,1,\cdots,N$:

$$\boxed{y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),} \tag{9}$$

where

$$\boxed{\begin{aligned} k_1 &= hf(x_n, y_n), \\ k_2 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right), \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned}} \tag{10}$$

This method has local truncation error of order 4, provided the solution $y(x)$ has five continuous derivatives In each subinterval we apply Runge-Kutta approximation of order 4 with the aid of the following scheme (Fig. 3.7), so that we obtain:



$$y_n + \frac{k_2}{2}$$

Fig. 3.7

Example 1 Use the Runge-Kutta approximation of order 2 with $h = 0.2$ to solve the initial value problem

$$y' = x^2 + y^2, \qquad y(0) = 0, \tag{11}$$

on $0 \le x \le 1$.

Solution Using $(x, y) = x^2 + y^2,\ x_0 = 0,\ y_0 = 0,\ h = 0.2,$ and

$$\boxed{\begin{aligned} &y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{k}{2}\right), \quad n = 0,1,\cdots,4 \\ &k = hf\left(x_n, y_n\right) \end{aligned}} \tag{12}$$

Hence we have

$$k = hf(x_0, y_0) = 0.2hf(0,0) = 0,$$

$$x_1 = x_0 + h = 0.2,$$

$$y_1 = y(0.2) = y_0 + hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k}{2}\right)$$

$$= 0 + 0.2(0.1 + 0) = 0.002$$

$$k = hf(x_1, y_1) = 0.2hf(0.2, 0.002) = 0.008001,$$

$$x_2 = x_1 + h = 0.4,$$

$$y_2 = y(0.4) = y_1 + hf\left(x_1 + \frac{h}{2}, y_1 + \frac{k}{2}\right)$$

$$= 0.002 + 0.2f(0.3, 0.006) = 0.020007,$$

$$k = hf(x_2, y_2) = 0.2hf(0.4, 0.020007) = 0.03208,$$

$$x_3 = x_2 + h = 0.6,$$

$$y_3 = y(0.6) = y_2 + hf\left(x_2 + \frac{h}{2}, y_2 + \frac{k}{2}\right)$$

$$= 0.020007 + 0.2f(0.5, 0.036047) = 0.070267,$$

$$k = hf(x_3, y_3) = 0.2hf(0.6, 0.070267) = 0.072987,$$

$$x_4 = x_3 + h = 0.8,$$

$$y_4 = y(0.8) = y_3 + hf\left(x_3 + \frac{h}{2}, y_3 + \frac{k}{2}\right)$$

$$= 0.070267 + 0.2f(0.7, 0.106761) = 0.170547,$$

$$k = hf(x_4, y_4) = 0.2hf(0.8, 0.170547) = 0.133817,$$

$$y_5 = y(1.0) = y_4 + hf\left(x_4 + \frac{h}{2}, y_4 + \frac{k}{2}\right)$$

$$= 0.170547 + 0.2f(0.9, 0.237455) = 0.343824,$$

Our numerical solution using the Runge-Kutta approximation of order 2 with $h = 0.2$ of the initial value problem (11) on $0 \le x \le 1$. is given by the table

startpt		***k***	**midpt**	
x	y		x	y
0	0	0	0.1	0
0.2	0.002	0.008001	0.3	0.006
0.4	0.020007	0.03208	0.5	0.036047
0.6	0.070267	0.072987	0.7	0.106761
0.8	0.170547	0.133817	0.9	0.237455
1	0.343824			

Example 2 Use the fourth order Runge-Kutta method to approximate y at $x = 0.1$ given that

$$y' = 3x + y^2, \qquad y(0) = 1, \tag{13}$$

Solution Here we have

$$f(x, y) = 3x + y^2, \;\; x_0 = 0, \;\; y_0 = 1, \;\; h = 0.1 - 0 = 0.1,$$

and the formula

$$\begin{aligned}
k_1 &= hf(x_0, y_0) = 0.1f(0,1) = 0.1,\\
k_2 &= hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = hf(0.05,1.05) = 0.12525,\\
k_3 &= hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = hf(0.05,1.062625)\\
&= 0.127917189,\\
k_4 &= hf(x_0 + h, y_0 + k_3) = hf(0.1,1.127917189)\\
&= 0.157219719
\end{aligned}$$

From which we obtain

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4)$$

$$= 1.0 + \frac{1}{6}[0.1 + 2(0.12525) + 2(0.127917189) + 0.157219719],$$

which implies that

$$\boxed{y_1 = y(0.1) \cong 1.127259}$$

Example 3 Obtain the value of y at $x = 0.8$ by the fourth order Runge-Kutta method if $2y' = \sqrt{x+y}$ and $x_0 = 0.4,\ y_0 = 0.41$ (Take $h = 0.4$). Also calculate $y(1.2)$.

Solution Here we have

$$f(x,y) = 0.5\sqrt{x+y}, \qquad x_0 = 0.4, \qquad y_0 = 0.41, \qquad h = 0.4,$$

and the formulas for each subinterval

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad n = 0,1.$$

where

$$\begin{aligned} k_1 &= hf(x_n, y_n), \\ k_2 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right), \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned}$$

First we calculate y at $x = 0.8$ so we have the diagram:

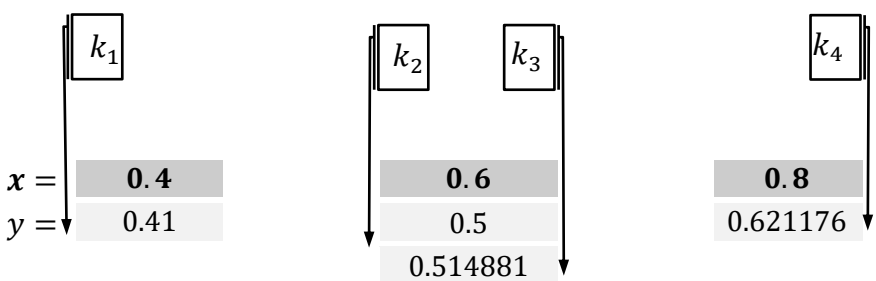


Fig. 4.8

$$\begin{aligned} k_1 &= hf(x_0, y_0) = 0.4f(0.4, 0.41) \\ &= 0.4\left(0.5\sqrt{0.4 + 0.41}\right) = 0.18, \end{aligned}$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.4f(0.6, 0.5)$$
$$= 0.4\left(0.5\sqrt{0.6 + 0.5}\right) = 0.209762,$$
$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.4f(0.6, 0.514881)$$
$$= 0.4\left(0.5\sqrt{0.6 + 0.514881}\right) = 0.211176,$$
$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.4f(0.8, 0.621176)$$
$$= 0.4\left(0.5\sqrt{0.8 + 0.621176}\right) = 0.238426$$

The substitution of these constants in the fourth order Runge-Kutta method gives

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4)$$
$$= 0.41 + \frac{1}{6}[0.18 + 2(0.209762) + 2(0.211176) + 0.238426],$$

which implies that

$$\boxed{y_1 = y(0.8) \cong 0.62005023}$$

If we want to calculate $y(1.2)$ then it is better to start from the value of y at $x = 0.8$. Thus we get the diagram

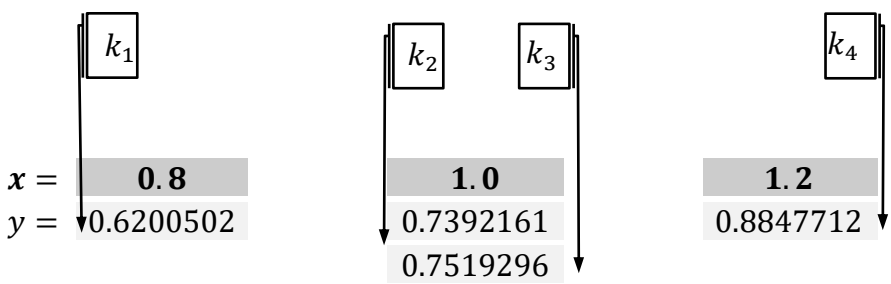


Fig. 4.9

$$k_1 = hf(x_0, y_0) = 0.4f(0.4, 0.41) = 0.2383317,$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.4f(1.0, 0.7392161) = 0.2637587,$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.4f(1.0, 0.7519296) = 0.264721,$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.4f(1.2, 0.8847712) = 0.2887747$$

Therefore,

$$\boxed{y_2 = y(1.2) \cong 0.884061185.}$$

Example 4 Use the Runge-Kutta approximation of order 4 with $h = 0.2$ to solve the initial value problem:

$$y' = x^2 + y^2, \qquad y(0) = 0, \tag{13}$$

on $0 \le x \le 1$.

Solution

$$f(x, y) = x^2 + y^2, \qquad x_0 = 0, \qquad y_0 = 0, \qquad h = 0.2,$$

and the formulas for each subinterval

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \;\; n = 0,1,\cdots,4$$

where

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right),$$

$$k_4 = hf(x_n + h, y_n + k_3)$$

At the point $x_1 = x_0 + h = 0.2$, we have

	k_1	k_2	k_3	k_4
$\boldsymbol{x} =$	$\mathbf{0}$	$\mathbf{0.1}$		$\mathbf{0.2}$
$y =$	0	0	0.001	0.0020002

$$k_1 = hf(x_0, y_0) = 0.2f(0,0) = 0,$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.2f(0.1, 0) = 0.002,$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.2f(0.1, 0.001) = 0.002,$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.2f(0.2, 0.0020002) = 0.0080008$$

Therefore,

$$\boxed{y_1 = y(0.2) \cong 0.002666867.}$$

<u>At the point $x_2 = x_1 + h = 0.4$</u>

	k_1	k_2	k_3	k_4
$\boldsymbol{x} =$	$\mathbf{0.2}$	$\mathbf{0.3}$		$\mathbf{0.4}$
$y =$	0.002666867	0.006667578	0.011671312	0.020694111

$$k_1 = hf(x_0, y_0) = 0.2f(0.2, 0.002666867) = 0.0080014,$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.2f(0.3, 0.006667578) = 0.0180089,$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.2f(0.3, 0.011671312) = 0.0180272,$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.2f(0.4, 0.020694111) = 0.032085649,$$

$$\boxed{y_2 = y(0.4) \cong 0.02136009.}$$

<u>At the point $x_3 = x_2 + h = 0.6$</u>

	k_1	k_2	k_3	k_4
$\boldsymbol{x} =$	$\mathbf{0.4}$	$\mathbf{0.5}$		$\mathbf{0.6}$
$y =$	0.02136009	0.037405716	0.046500009	0.071792541

$$k_1 = 0.2f(0.4, 0.02136009) = 0.0320913,$$
$$k_2 = 0.2f(0.5, 0.037405716) = 0.0502798,$$
$$k_3 = 0.2f(0.5, 0.046500009) = 0.0504325,$$
$$k_4 = 0.2f(0.6, 0.071792541) = 0.073030834$$

Therefore,

$$\boxed{y_3 = y(0.6) \cong 0.0724512.}$$

<u>At the point $x_4 = x_3 + h = 0.8$</u>

	k_1	k_2	k_3	k_4
$\boldsymbol{x} =$	$\mathbf{0.6}$	$\mathbf{0.7}$		$\mathbf{0.8}$
$y =$	0.0724512	0.108976118	0.12263878	0.173459254

$$k_1 = 0.2f(0.6, 0.0724512) = 0.0730498,$$
$$k_2 = 0.2f(0.7, 0.108976118) = 0.1003752,$$
$$k_3 = 0.2f(0.7, 0.12263878) = 0.1010081,$$
$$k_4 = 0.2f(0.8, 0.173459254) = 0.134017623,$$

$$\boxed{y_4 = y(0.8) \cong 0.174090181.}$$

<u>At the point $x_5 = x_4 + h = 0.6$</u>

	k_1	k_2	k_3	k_4
$\boldsymbol{x} =$	$\mathbf{0.8}$	$\mathbf{0.9}$		$\mathbf{1.0}$
$y =$	0.174090181	0.24112092	0.260904111	0.349704372

$$k_1 = 0.2f(0.8, 0.174090181) = 0.1340615,$$
$$k_2 = 0.2f(0.9, 0.24112092) = 0.1736279,$$
$$k_3 = 0.2f(0.9, 0.260904111) = 0.1756142,$$
$$k_4 = 0.2f(1.0, 0.349704372) = 0.22445863$$

Therefore,

$$\boxed{y_5 = y(1.0) \cong 0.350257549.}$$

Our numerical solution using the Runge-Kutta approximation of order 4 with $h = 0.2$ of the initial value problem (13) on $0 \le x \le 1$ is given by the set

$$\{(0, 0), (0.2, 0.002666867), (0.4, 0.02136009), (0.6, 0.0724512),$$

$$(0.8, 0.174090181), (1.0, 0.350257549)\}$$

The Runge-Kutta method is highly accurate (small truncation error) even when n is relatively small (small round-off error). For these reasons it is a widely used method. It can be shown that the error involved in the Runge-Kutta approximation of order k is proportional to h^k. Thus the Runge-Kutta method is more accurate than the Euler method and although it has the same order of accuracy as the Taylor approximation of order k, it has the feature that the calculation depend on knowledge of the function f only and not on derivatives of f as in the case of the Taylor method.

The main computational effort in applying the Runge-Kutta methods is the evaluation of f. In the second order method the local truncation error is of order h^2, but the cost is two functional evaluations per step. The Runge-Kutta of order four requires four evaluations per step, but the local truncation error is of order 4. However, the Runge-Kutta of order four gives more accurate answers than Euler's method with one-quarter of mesh size. Similarly the Runge-Kutta method of order four gives more accuracy with step size h than a second order method with step size $\frac{1}{2}h$, because the fourth order method requires twice as many evaluations per step.

Example 5 Consider the initial value problem

$$y' = 1 - y, \qquad y(0) = 0,$$

which has the exact solution

$$y(x) = 1 - e^{-x}$$

Euler's method with $h = 0.025$, the modified Euler's method with $h = 0.05$, and the Runge-Kutta fourth-order method with $h = 0.1$ will be compared at the mesh points $h = 0.1, 0.2, 0.3, 0.4,$ and 0.5. The results are given in the following table. For this example the fourth order method is clearly superior.

x	Euler's method	Modified Euler's method	Fourth-order Runge-Kutta method	Exact solution
	$h = 0.025$	$h = 0.05$	$h = 0.1$	
0.1	0.096312	0.095123	0.09516250	0.095162582
0.2	0.183348	0.181198	0.18126910	0.181269247
0.3	0.262001	0.259085	0.25918158	0.259181779
0.4	0.333079	0.329563	0.32967971	0.329679954
0.5	0.397312	0.393337	0.39346906	0.393469340

Example 6 If $y' = \cos xy$, $y(0) = 2$, find $y(0.1)$ by using:

(i) Taylor's method up three decimals.

(ii) Modified Euler method up to four decimal places.

(iii) The second-order Runge-Kutta method.

(iv) The fourth-order Runge-Kutta method.

Solution

(i) Taylor's method. Since $f(x, y) = \cos xy$, $x_0 = 0$, $y_0 = 2$, then

$y' = \cos xy,$ | $y_0' = 1$

$y'' = -(y + xy')\sin xy$ | $y_0'' = 0$

$$y''' = -(y + xy')^2 \cos xy - (2y' + xy'') \sin xy \qquad \Big| \; y_0''' = -4$$

Substituting in Taylor series with $h = 0.1$, we get

$$y_1 = y_0 + \frac{y_0'}{1!}h + \frac{y_0''}{2!}h^2 + \frac{y_0'''}{3!}h^3 + \cdots$$

$$= 2 + 0.1(1.0) + \frac{(0.1)^2}{2\,!}(0) + \frac{(0.1)^3}{3\,!}(-4)^3 \cong 2.0993333$$

(ii) Modified Euler's method.

$$y_1^* = y_0 + hf_0 = 2 + 0.1f(0,2) = 2.1$$

$$\Rightarrow \qquad f_1^* = f(0.1,2.1) = 0.978030915$$

$$y_1 = y_0 + \frac{h}{2}(f_0 + f_1^*)$$

$$= 2 + \frac{0.1}{2}[1 + 0.978030915] = 2.098901546$$

(iii) The second-order Runge-Kutta method. Using

$$f(x,y) = \cos xy\,, \qquad x_0 = 0, \qquad y_0 = 2, \;\; h = 0.1,$$

so

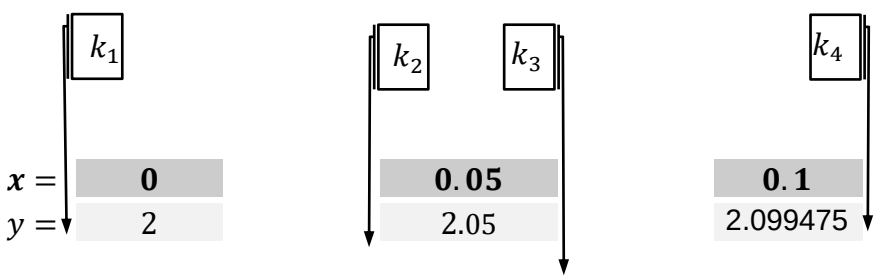
$$k = hf(x_0, y_0) = 0.1hf(0,2) = 0.1,$$

$$y_1 = y(0.1) = y_0 + hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k}{2}\right)$$

$$= 2 + 0.2(0.05 + 2.05) = 2.099475147$$

(iv) The fourth order Runge-Kutta method

$$f(x,y) = \cos xy\,, \qquad x_0 = 0, \qquad y_0 = 2, \;\; h = 0.1,$$



2.049738

Fig. 4.10

$$k_1 = hf(x_0, y_0) = 0.1f(0, 2) = 0.1$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.1f(0.05, 2.05) = 0.099475147$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.1f(0.05, 2.049738) = 0.099475282$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.1f(0.1, 2.099475) = 0.097804185$$

The substitution of these constants in the fourth order Runge-Kutta method gives

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 2.099284174$$

Example 7 Given the initial value problem

$$2y' = \ln(x + y), \qquad y(0.4) = 1.5$$

(i) Evaluate $y(0.7)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.2)$ by the use of Taylor method of order 3,

(iii) Use Euler method to evaluate $y(0.8)$.

Solution

$$y' = f(x, y) = 0.5\ln(x + y), \qquad x_0 = 0.4, \qquad y_0 = 1.5$$

Taylor method	Initial pt.	Rk4	Euler method
■	■	■	■
0.2	0.4	0.7	0.8
	1.5		

(i) Runge-Kutta approximation of order 4, $y(0.7)???$

$$h = 0.7 - 0.4 = 0.3$$

0.4
2

0.55
2.06565
2.0721

0.7
2.1446

$$k_1 = hf(x_0, y_0) = 0.0963$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.1112$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.1117$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.1257$$

$$\Rightarrow \quad y_1 = y(0.7) = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 1.6113$$

(ii) Find $y(0.2)$ by the use of Taylor method of order 3,

$$h = 0.2 - 0.4 = -0.2$$

$y' = 0.5\ln(x+y)$	$y_0' = 0.3209$
$y'' = 0.5 * \dfrac{(1+y')}{(x+y)}$	$y_0'' = 0.3476$
$y''' = 0.5 * \dfrac{(x+y)(y'') - (1+y')^2}{(x+y)^2}$	$y_0''' = -0.08811$

$$y(0.2) = 1.5 + \frac{(-0.2)}{1!}(0.3209) + \frac{(-0.2)^2}{2!}(0.3476) + \frac{(-0.2)^3}{3!}(-0.08811) + \cdots \cong 1.4429$$

(iii) Use Euler method to evaluate $y(0.8)$. Take the reference pt. $(0.7,\ 1.6113)$ so that $h = 0.8 - 0.7 = 0.1$,

$$y(0.8) = 1.6113 + 0.1 * f(0.7, 1.6113) = 1.6532$$

4.6 Multistep methods

The methods discussed previously in this chapter are called one-step methods since the approximations for the mesh point x_{n+1} involves information from only one of the previous mesh points x_n. Although these methods generally use functional evaluation information at points between x_n and x_{n+1}, they do not retain that information for direct use in future approximations. All the information used by these methods is consequently obtained within the interval over which the solution is being approximated.

Since the approximate solution is available at each of the mesh points $x_0, x_1, \cdots, x_n$ before the approximation at x_{n+1} is obtained, and because the error $|y_n - y(x_n)|$ tends to increase with n, it seems reasonable to develop methods that use this more accurate previous data when approximating the solution at x_{n+1}.

Methods using the approximation at more than one previous mesh point to determine the approximation at the next point are called ***multistep methods***.

Given the differential equation

$$y' = f(x, y), \tag{1}$$

and let us assume that we have already obtained approximations to y' and y at a number (here we take this number equals 4) of equally spaced points, say $x_{n-3}, x_{n-2}, x_{n-1}, x_n$ and we wish to find $y(x_{n+1})$. One class of multistep method is based on the principle of numerical integration. If we integrate both sides of equation (1) from x_n to x_{n+1}, we will have

$$\int_{x_n}^{x_{n+1}} y'\,dx = \int_{x_n}^{x_{n+1}} f(x,y)\,dx, \tag{2}$$

or

$$y_{n+1} - y_n = \int_{x_n}^{x_{n+1}} f(x,y)\,dx, \tag{3}$$

To carry out the integration in (3) we now approximate $f\big(x, y(x)\big)$ by a polynomial which interpolates $f\big(x, y(x)\big)$ at the four points $x_{n-3}, x_{n-2}, x_{n-1}, x_n$. If we use the notation

$$f_k = f\big(x_k, y(x_k)\big),$$

we can use the Newton backward formula (See §3.5) of degree 3 for this purpose: Putting $s = \frac{x - x_n}{h}$ and using the difference table

$\boldsymbol{x}$	$\boldsymbol{f}$	$\boldsymbol{\nabla}$	$\boldsymbol{\nabla^2}$	$\boldsymbol{\nabla^3}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{n-3}	f_{n-3}		$\nabla^2 f_{n-2}$		
		∇f_{n-2}			
x_{n-2}	f_{n-2}		$\nabla^2 f_{n-1}$		
		∇f_{n-1}		$\nabla^3 f_n$	
x_{n-1}	f_{n-1}		$\nabla^2 f_n$		
		∇f_n			
x_n	f_n				

Table 4.2 Backward difference table of *f*

Hence

$$P_3(s) = f_n + s\nabla f_n + \frac{s(s+1)}{2\,!}\nabla^2 f_n + \frac{s(s+1)(s+2)}{3\,!}\nabla^3 f_n \tag{4}$$

Inserting this into (3) and noting that $dx = hds$, we obtain

$$y_{n+1} = y_n + h\int_0^1 \left[f_n + s\nabla f_n + \frac{s(s+1)}{2\,!}\nabla^2 f_n + \frac{s(s+1)(s+2)}{3\,!}\nabla^3 f_n\right] ds$$

$$= y_n + h\int_0^1 \left[f_n + s\nabla f_n + \frac{(s^2+s)}{2\,!}\nabla^2 f_n + \frac{(s^3+3s^2+2s)}{3\,!}\nabla^3 f_n\right] ds$$

$$= y_n + h\left[f_n + \frac{1}{2}\nabla f_n + \frac{\left(\frac{1}{3}+\frac{1}{2}\right)}{2}\nabla^2 f_n + \frac{\left(\frac{1}{4}+1+1\right)}{6}\nabla^3 f_n\right],$$

or

$$y_{n+1} = y_n + h\left(f_n + \frac{1}{2}\nabla f_n + \frac{5}{12}\nabla^2 f_n + \frac{9}{24}\nabla^3 f_n\right) \qquad (5)$$

In practice, it is more convenient computationally to work with ordinates instead of differences. From the definition of the backward difference operator ∇ we find that

$$\nabla f_n = f_n - f_{n-1},$$

$$\nabla^2 f_n = \nabla f_n - \nabla f_{n-1} = f_n - 2f_{n-1} + f_{n-2},$$

$$\nabla^3 f_n = \nabla^2 f_n - \nabla^2 f_{n-1} = f_n - 2f_{n-1} + f_{n-2} - f_{n-1} + 2f_{n-2} - f_{n-3}$$

$$= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$$

Substituting in (5) and regrouping, we obtain

$$\boxed{y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \qquad (6)}$$

Formula (6) is known as the ***Adams-Bashforth*** method and in symbols we express it as shown in Fig. 4.12.

Another multistep method can be obtained by using a polynomial of degree 3 which interpolates $f\big(x, y(x)\big)$ at $x_{n-2}, x_{n-1}, x_n, x_{n+1}$. by

the use of Newton backward formula at these points in terms of $s = \frac{x-x_{n+1}}{h}$. In this case the interpolating polynomial will be

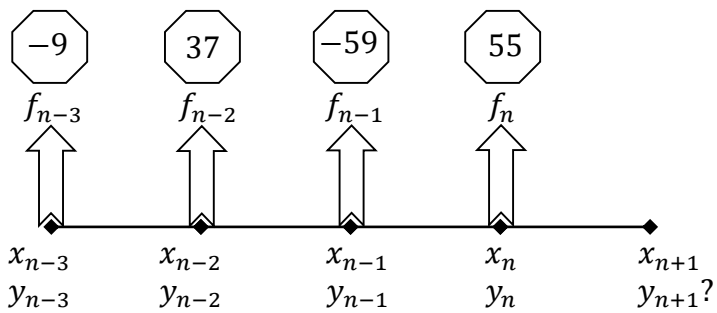


Fig. 4.12 Adams-Bashforth method

$$P_3(s) = f_{n+1} + s\nabla f_{n+1} + \frac{s(s+1)}{2\,!}\nabla^2 f_{n+1} + \frac{s(s+1)(s+2)}{3\,!}\nabla^3 f_{n+1} \tag{7}$$

The substitution of (7) into (3) gives

$$y_{n+1} = y_n + h\int_{-1}^{0}\left[f_{n+1} + s\nabla f_{n+1} + \frac{s(s+1)}{2\,!}\nabla^2 f_{n+1} + \frac{s(s+1)(s+2)}{3\,!}\nabla^3 f_{n+1}\right]ds$$

$$= y_n + h\int_{-1}^{0}\left[f_{n+1} + s\nabla f_{n+1} + \frac{(s^2+s)}{2\,!}\nabla^2 f_{n+1} + \frac{(s^3+3s^2+2s)}{3\,!}\nabla^3 f_{n+1}\right]ds$$

$$= y_n + h\left[f_{n+1} - \frac{1}{2}\nabla f_{n+1} + \frac{\left(\frac{1}{3}-\frac{1}{2}\right)}{2}\nabla^2 f_{n+1} + \frac{\left(-\frac{1}{4}+1-1\right)}{6}\nabla^3 f_{n+1}\right]$$

$$= y_n + h\left[f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1} - \frac{1}{24}\nabla^3 f_{n+1}\right] \tag{8}$$

If the differences in (8) are expressed in terms of ordinates, we obtain

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \qquad (9)$$

Formula (9) defines an implicit multistep method known as the ***Adams-Moulton*** method and in symbols we express it as

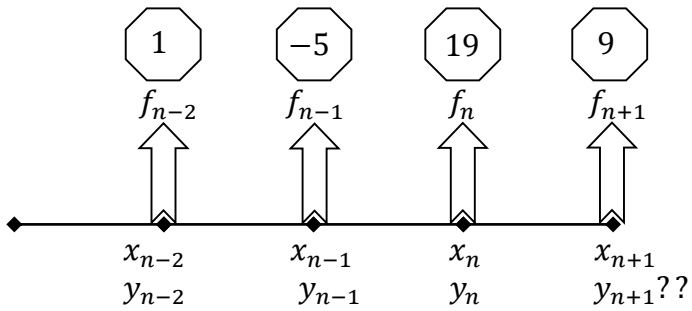


Fig. 4.13. Adams-Moulton method

The starting values in either (6) or (9) must be specified, generally, by assuming $y(x_0) = y_0$ and generating the remainder values by either a Runge-Kutta method or some other one-step technique.

To apply an implicit method such as (9) directly, we must solve the implicit equation for y_{n+1}. It is not clear that this can be done in general, or that a unique solution for y_{n+1} will always be obtained. So that equation (9) must therefore be solved by iteration. A convenient predictor to use with this corrector is the Adams-Bashforth formula (6). The combination of an explicit (Adam-Bashforth) and implicit (Adams-Moulton) technique is called a predictor-corrector method.

Given the initial value problem

$$y' = f(x, y), \quad y(x_n) = y_n, \ n = 0,1,2,3 \qquad (10)$$

The following steps give specifications for a predictor-corrector method to solve problem (10).

Step 1 Compute y_{n+1}^*, using the predictor formula (Adams-Bashforth formula (6))

$$y_{n+1}^* = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \qquad (11)$$

Step 2 Compute $f_{n+1}^* = f(x_{n+1}, y_{n+1}^*)$.

Step 3 Compute the corrected value y_{n+1}, using the corrector formula (Adams-Moulton formula (9))

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1}^* + 19f_n - 5f_{n-1} + f_{n-2}) \qquad (12)$$

In using predictor-corrector methods, it is necessary to have a method for obtaining the necessary starting values initially. Special forms can be worked out for this case. However, a fairly ideal combination is to use the fourth order Runge-Kutta method together with the predictor-corrector formula. The Runge-Kutta can then be used for starting the solution initially while the predictor-corrector formula can be used after that. This is illustrated by the following example.

Example 1 For the initial value problem

$$y' = x - y + 1, \qquad y(0) = 1 \qquad (13)$$

(a) Use Runge-Kutta method to compute $y(0.1), y(0.2), and\ y(0.3)$.

(b) Use the predictor-corrector formula (11)–(12) to find $y(0.4)$.

Solution Here we have

$$f(x, y) = x - y + 1, \;\; x_0 = 0, \;\; y_0 = 1, \;\; h = 0.1$$

(a) At $x_1 = 0.1$

$$k_1 = hf(x_0, y_0) = hf(0,1) = 0,$$
$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = hf(0.05,1) = 0.005,$$
$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = hf(0.05,1.0025) = 0.00475,$$
$$k_4 = hf(x_0 + h, y_0 + k_3) = hf(0.1,1.00475) = 0.009525$$

From which we obtain

$$y_1 = y(0.1) = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 1.0048375$$

At $x_2 = 0.2$

$$k_1 = hf(x_1, y_1) = hf(0.1,1.0048375) = 0.009516,$$
$$k_2 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}\right) = hf(0.15,1.009595625) = 0.01404,$$
$$k_3 = hf\left(x_1 + \frac{h}{2}, y_1 + \frac{k_2}{2}\right) = hf(0.15,1.011857719) = 0.013814,$$
$$k_4 = hf(x_1 + h, y_1 + k_3) = hf(0.2,1.018651728) = 0.018135,$$
$$y_2 = y(0.2) = y_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 1.018730901$$

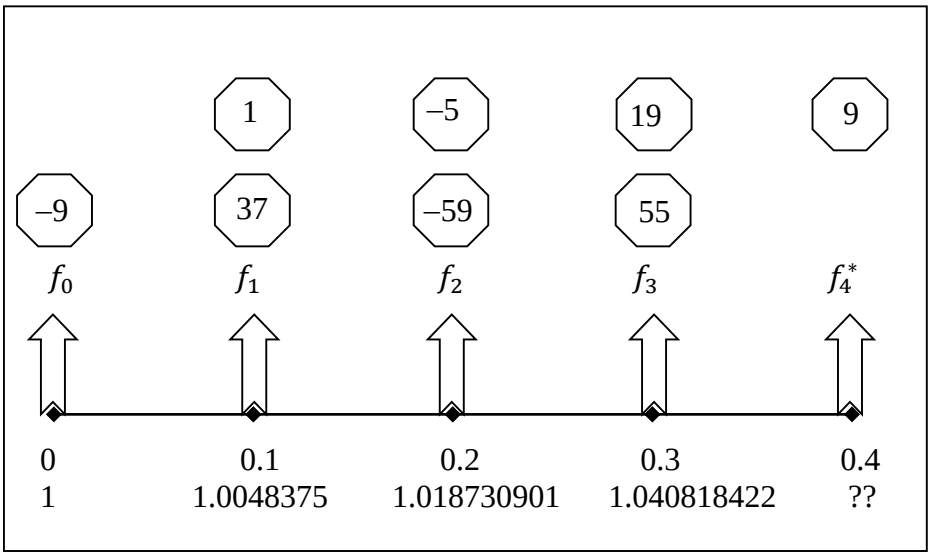
At $x_3 = 0.3$

$$k_1 = hf(x_2, y_2) = hf(0.2,1.018730901) = 0.018127,$$
$$k_2 = hf\left(x_2 + \frac{h}{2}, y_2 + \frac{k_1}{2}\right) = hf(0.25,1.027794356) = 0.022221,$$
$$k_3 = hf\left(x_2 + \frac{h}{2}, y_2 + \frac{k_2}{2}\right) = hf(0.25,1.029841184) = 0.022016,$$
$$k_4 = hf(x_2 + h, y_2 + k_3) = hf(0.3,1.040746783) = 0.025925,$$
$$y_3 = y(0.3) = y_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 1.040818422$$

(b) To use the predictor-corrector form we first compute the function value at the known points:

x	y	$f(x,y)$
0	1	0
0.1	1.0048375	0.0951625
0.2	1.018730901	0.1812691
0.3	1.040818422	0.25918158

Table 4.3 Function evaluation



Therefore, Adams-Bashforth formula implies the predicted value

$$y_4^* = y_3 + \frac{h}{24}(55f_3 - 59f_2 + 37f_1 - 9f_0) = 1.070323099,$$

and Adams-Moulton formula gives

$$y_4 = y_3 + \frac{h}{24}(9f_4^* + 19f_3 - 5f_2 + f_1) = 1.070319918$$

Example 2] Given the initial value problem:

$$y' = \sqrt{x} + \sqrt{y}, \quad y(0.16) = 1.44,$$

(i) Evaluate $y(0.2)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.12)$ by the use of Taylor method,

(iii) Find $y(0.24)$ by the use of Euler method.

(iv) Use a predictor-corrector method to evaluate $y(0.28)$.

Solution Here we have

$$f(x,y) = \sqrt{x} + \sqrt{y}, \quad x_0 = 0.16, \quad y_0 = 1.44, \quad h = 0.04$$

(i) <u>At $x_1 = 0.2$</u> by <u>$RK4$</u>

$$k_1 = hf(x_0, y_0) = hf(0.16,1.44) = 0.064,$$
$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = hf(0.18,1.472) = 0.06550097,$$
$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = hf(0.18,1.47275048) = 0.06551334,$$
$$k_4 = hf(x_0 + h, y_0 + k_3) = hf(0.2,1.505513) = 0.066968$$

From which we obtain

$$y_1 = y(0.2) = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = 1.50549948$$

(ii) <u>At $x_1 = 0.12$</u> by the use of Taylor method

$\boldsymbol{y' = \sqrt{x} + \sqrt{y}}$	$\boldsymbol{y'_0 = 1.6}$
$y'' = \frac{1}{2\sqrt{x}} + \frac{y'}{2\sqrt{y}}$	$y''_0 = 1.916667$
$y''' = -\frac{1}{4x\sqrt{x}} + \frac{yy'' - (y')^2}{4y\sqrt{y}}$	y'''_0 =-0.32331

Substituting in Taylor series, we get

$$y(0.12) = y_0 + \frac{y_0'}{1\,!}h + \frac{y_0''}{2\,!}h^2 + \frac{y_0'''}{3\,!}h^3 + \cdots$$

$$= 1.44 + \frac{(1.6)}{1\,!}(0.04) + + \frac{(1.916667)}{2\,!}(0.04)^2 +$$

$$+ \frac{(0.32331)}{3\,!}(0.05)^3 \cong 1.37753678$$

(iii) At $x_1 = 0.24$ by the use of Euler method. Take

$$f(x,y) = \sqrt{x} + \sqrt{y}, \quad x_0 = 0.2, \quad y_0 = 1.50549948, \quad h = 0.04,$$

thus formulas

$$x_1 = 0.2 + 0.04,$$

$$y_1 = y(0.24) = y_0 + 0.04f(0.2, 1.50549948) = 1.57246754$$

(iv) To use the predictor-corrector form we first compute the function value at the known points:

n	x	y	$f(x,y)$
0	0.12	1.37753678	1.52009529
1	0.16	1.44	1.6
2	0.2	1.50549948	1.674201567
3	0.24	1.57246754	1.743878628

Table 4.4 Function evaluation

Therefore, Adams-Bashforth formula implies the predicted value

$$y_4^* = y_3 + \frac{h}{24}(55f_3 - 59f_2 + 37f_1 - 9f_0) = 1.643558502,$$

and Adams-Moulton formula gives

$$y_4 = y_3 + \frac{h}{24}(9f_4^* + 19f_3 - 5f_2 + f_1) = 1.64357281$$

Exercise 4.2

1. Given the initial value problem:

$$y' = \sqrt{10x^2 + y}, \qquad x_0 = 0.2, \qquad y_0 = 0.41$$

Evaluate $y(0.4)$ by Runge-Kutta approximation of order 4,

[$\boldsymbol{Answer}$: $y(0.4) = 0.648545$]

2. Find $y(1.1)$ by the use of fourth order Runge-Kutta method if

$$y' = \left(\frac{y}{x}\right)^2 + \left(\frac{y}{x}\right), \qquad y(1) = 1$$

3. Given the initial value problem

$$y' = xy + \sqrt{y}, \qquad y(0) = 1$$

(i) Evaluate $y(0.2)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.4)$ by the use of Euler method

(iii) Find $y(-0.2)$ by the use of Taylor method.

[$\boldsymbol{Answer}$: $y(0.2) = 1.233696, y(0.4) = 1.505188, y(-0.2) = 0.826667$]

4. Given the initial value problem

$$y' = \sin x + e^{-x}, \qquad y(0) = 0$$

Evaluate $y(0.5)$ by the use of, Euler's method, the modified Euler's method, Taylor method, and fourth order Runge-Kutta method.

5. Given the initial value problem

$$2y' = \sqrt{x + y}, \qquad y(0.4) = 0.41$$

(i) Evaluate $y(0.8)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0)$ and $y(1.2)$ by the use of Euler method

(iii) Use a predictor-corrector method to evaluate $y(1.6)$

[$\boldsymbol{Answer}$: $y(0.8) = 0.62005023$, $y(0) = 0.23$,

$y(1.2) = 0.858381951$, $y(1.6) = 1.168251982$]

6. Solve the following initial value problems **(i)–(v)** by the use of, Euler's method, the Modified Euler's method, Taylor method, fourth

order Runge-Kutta method and the Adam-Bashforth predictor-corrector formula.

(i). $y' = \frac{1}{x}\left(y^2 + y\right), \quad 1 \le x \le 3, \quad y(1) = -2$ with $h = 0.5$.

(ii). $y' = -xy + \frac{4x}{y}, \quad 0 \le x \le 1, \quad y(0) = 1$ with $h = 0.25$.

(iii). $y' = 4x - 2y, \quad 0 \le x \le 0.6, \quad y(0) = 2$ with $h = 0.2$.

(iv). $y' = y - y^2, \quad 0 \le x \le 0.6, \quad y(0) = 0.5$ with $h = 0.2$.

(v). $y' = x + y^2, \quad 0 \le x \le 0.8, \quad y(0) = 0$ with $h = 0.2$.

7. Given the initial value problem:

$$5y' = 4(\sqrt{x} + \sqrt{y}), \quad y(0.25) = 1,$$

(i) Evaluate $y(0.5)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0)$ by the use of Taylor method,

(iii) Find $y(0.75)$ by the use of Euler method.

(iv) Use a predictor-corrector method to evaluate $y(1)$.

[***Answer***: $y(0.5) = 1.33734286$, $y(0) = 0.70358333$, $y(0.75) = 1.7100513$, $y(1) = 2.175057105$]

8. Given the initial value problem

$$2y' = 4\sqrt{x} + 5\sqrt{y}, \quad y(0.36) = 1.21,$$

find y at $x = 0.4, 0.44, 0.48$ by using

(i) Euler's method,

(ii) The Modified Euler's method,

(iii) The Runge-Kutta method and

(iv) The Adams predictor-corrector formula.

9. Given the initial value problem:

$$y' + \frac{2x + y^2}{xy} = 0, \quad y(1.2) = 1,$$

(i) Evaluate $y(1.4)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(1)$ by the use of Taylor method,

(iii) Find $y(1.6)$ by the use of Euler method.

(iv) Use a predictor-corrector method to evaluate $y(1.8)$.

[***Answer***: $y(1.4) = 0.216661$, $y(1) = 1.584155$, $y(1.6) = 0.546398$, $y(1.8) = 0.112873306$]

10. Given the initial value problem

$$y' = \frac{1}{x^2} - \frac{y}{x} - y^2, \quad 1 \le x \le 2, \quad y(1) = -1$$

with exact solution $y = \frac{1}{x}$:

(a) Use Euler's method with $h = 0.05$ to approximate the solution and compare it with the actual values of y.

(b) Use the answers generated in part (a) and linear interpolation to approximate the following values of y and compare them with the actual values.

(i) $y(1.052)$ **(ii)** $y(1.555)$ **(iii)** $y(1.978)$

11. Given the initial value problem

$$4y' = 4x^2 + y, \qquad y(1) = 2.4$$

(i) Evaluate $y(1.2)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.8)$ by the use of Taylor method

(iii) Find $y(1.4)$ by the use of Euler method.

[$\boldsymbol{Answer}$: $y(1.2) = 2.77151042, y(0.8) = 2.124533, y(1.4) = 3.198086$]

12. Given the initial value problem

$$y' = \frac{2}{x}y + x^2e^x, \quad 1 \le x \le 2, \quad y(1) = 0$$

with exact solution $y = x^2\left(e^x - e\right)$:

(a) Use Taylor's method of order two with $h = 0.1$ to approximate the solution and compare it with the actual values of y.

(b) Use the answers generated in part (a) and linear interpolation to approximate the following values of y and compare them to the actual values of y.

(i) $y(1.04)$ **(ii)** $y(1.55)$ **(iii)** $y(1.97)$

(c) Use Taylor's method of order four with $h = 0.1$ to approximate the solution and compare it with the actual values of y.

(d) Use the answers generated in part (c) and linear interpolation to approximate the following values of y and compare them to the actual values of y.

(i) $y(1.04)$ **(ii)** $y(1.55)$ **(iii)** $y(1.97)$

(e) Use the Modified Euler method with $h = 0.1$ to approximate the solution and compare it with the actual values of y.

(f) Use the Runge-Kutta method of order four with $h = 0.1$ to approximate the solution and compare it with the actual values of y.

13. Given the initial value problem

$$2y' = \ln(x + y), \qquad y(0.4) = 2$$

(i) Evaluate $y(0.7)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.2)$ by the use of Taylor method of order 3,

(iii) Use Euler method to evaluate $y(0.8)$.

[***Answer***: $y(0.7) = 2.1443, y(0.2) = 1.9186, y(0.8) = 2.1966$]

14. In problems **(i)–(iii)**, use the Adams–Bashforth/Adams–Moulton method to approximate $y(1)$, where $y(x)$ is the solution of the given initial value problem. Use $h = 0.2$ and $h = 0.1$ and the Runge–Kutta method of order 4 to compute y_1, y_2, and y_3.

(i) $y' = 1 + y^2, \quad y(0) = 0$

(ii) $y' = y + \cos x, \quad y(0) = 1$

(iii) $y' = xy + \sqrt{y}, \quad y(0) = 1$

15. Given the initial value problem:

$$y' = \sqrt{x^2 + 4y^2}, \quad y(2.4) = 2,$$

(i) Evaluate $y(2.6)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(2.2)$ by the use of Taylor method,

(iii) Find $y(2.8)$ by the use of Euler method.

(iv) Use the Adams-Bashforth/Adams-Moulton method to evaluate $y(3)$.

[***Answer***: $y(2.6) = 3.12957539$, $y(2.2) = 1.21342518$, $y(2.8) = 4.48511172$, $y(3) = 6.752358895$]

In problems **16–18**, use the Adams–Bashforth/Adams–Moulton method to approximate the solution of the given problems and compare the results with the exact values.

16. $y' = \dfrac{y}{x} - \left(\dfrac{y}{x}\right)^2, \quad 1 \le x \le 2, \quad y(1) = 1, \quad h = 0.1,$

exact solution $y(x) = x(1 + \ln x)^{-1}$,

17. $y' = 1 + \dfrac{y}{x} - \left(\dfrac{y}{x}\right)^2, \quad 1 \le x \le 3, \quad y(1) = 0, \quad h = 0.2$,

exact solution $y(x) = x\tan(\ln x)$,

18. $y' = -(y+1)(y+3), \quad 1 \le x \le 2, \quad y(0) = -2, \quad h = 0.1$,

exact solution $y(x) = -3 + 2(1 + e^{-2x})^{-1}$,

19. Given the initial value problem:

$$y' = \frac{y}{x} + \ln x, \quad y(1) = 4$$

(i) Find $y(1.1)$ by Runge-Kutta approximation of order 4,

(ii) Evaluate $y(1.2)$ by the use of Taylor method of order 3,

(iii) Find $y(1.3)$ using Euler method,

(iv) Use a predictor-corrector method to evaluate $y(1.4)$.

[***Answer***: $y(1.1) = 4.404996$, $y(1.2) = 4.819947$, $y(1.3) = 5.239841$, $y(1.4) = 5.673895169$]

20. Given the initial value problem

$$y' = xy + 2\sqrt{x}, \quad y(0) = 1$$

(i) Evaluate $y(0.3)$ by Runge-Kutta approximation of order 4,

(ii) Find $y(0.4)$ by the use of Picard method

(iii) Use Euler method to evaluate $y(0.6)$

5. Numerical Solution of Partial Differential Equations

Partial differential equations (PDEs) occur in almost all areas of science and engineering. In fact, many ordinary differential equations (ODEs) arise either as approximation to or as part of the solution of partial differential equations.

This chapter is devoted mainly to the study of the finite difference method for solving partial differential equations of second order in two independent variables. This choice has been made because second order equations occur far more after in applications than other kind and because these equations have been the focus of a great deal of analytical work and are therefore the best understood type of PDEs. However, the finite difference methods can be applied to equations of first, third, and fourth order and to equations with a larger number of independent variables.

The most general second order PDE in two independent variables that is linear in all of the highest derivatives is

$$\boxed{Au_{xx} + Bu_{xy} + Cu_{yy} = f(x, y)} \qquad (1)$$

where $A, B, C,$ and f are functions of x and y.

The classification of equation (1) depends on the sign of $B^2 - 4AC$. If $B^2 - 4AC < 0$, the equation is called ***elliptic partial differential equation*** (EPDE). If $B^2 - 4AC = 0$, the equation is ***parabolic partial***

differential equation (PPDE). Finally, if $B^2 - 4AC > 0$, the equation is ***hyperbolic partial differential equation*** (HPDE). All three cases associated with some constraints on the domain of solution occur commonly in physics and engineering problems. All three cases associated with some constraints on the boundary of the domain of solution occur commonly in physics and engineering problems.

5.1 The finite difference method

Methods involving finite differences for solving boundary value problems consist of replacing each of the partial derivatives in the differential equation by an appropriate difference relation are called ***finite difference methods***. These methods are ones of the most widely spread at present methods of numerical solution of equations with partial derivatives.

Let us suppose that the problem is to solve the PDE

$$Au_{xx} + Bu_{xy} + Cu_{yy} + \cdots = f(x, y) \quad in\, \Omega \tag{1}$$

with the boundary conditions

$$u|_{\partial\Omega} = g(x, y) \tag{2}$$

where Ω is a smooth domain with boundary Lipschitz boundary $\partial\Omega$. The functions f and g are continuous functions in the variables x and y.

The above problem is to be solved by the finite difference method in three mainly steps:

- Mesh generation.
- Applying the boundary conditions to the boundary $\partial\Omega$.
- Replacing the derivatives by difference relations

♦ Step 1: Mesh generation

We first create a grid points in space on which the solution shall be computed, that is, we shall discretize the equations with respect to the independent variables.

Let the domain Ω with its boundary $\partial\Omega$ that lie in the $xy-$ plane (Fig. 5.1) be the rectangle

$$\Omega \equiv \{(x,y)|a < x < b,\ \ c < y < d\}$$

Choose two integers m and n and define step sizes h and k by

$$h = \frac{b-a}{m},\ \ k = \frac{d-c}{n}$$

Partitioning the interval $[a,b]$ into m equal parts of width h and the interval $[c,d]$ into n equal parts of width k (see Fig.5.1) provides a mean of placing a grid on the rectangle Ω by drawing vertical and horizontal lines through the points with coordinates (x_i, y_j), where

$$x_i = a + ih, \qquad i = 0,1,\cdots,m \tag{4}$$

$$y_j = c + jk, \qquad j = 0,1,\cdots,n \tag{5}$$

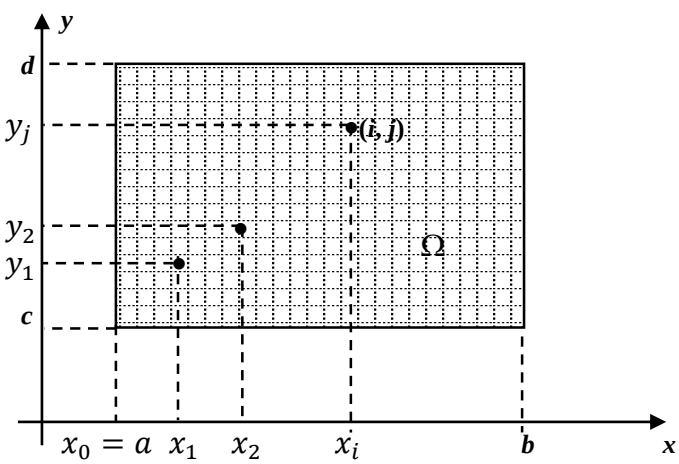


Fig. 5.1

The lines $x = x_i$ and $y = y_j$ are called ***grid lines*** and their intersections are called the ***mesh points*** or ***nodes*** of the grid. Two mesh points are termed neighboring ones if the distance between them in the direction of x- and y-axes is equal to the grid spacing h or k, respectively.

Let us denote the values of the required function $u = u(x, y)$ at the mesh points by $u_{i,j} = u(x_i, y_j)$. A recommended labeling of the mesh points is to let

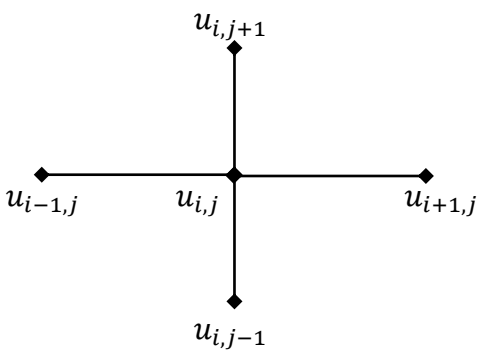


Fig. 5.2

This in effect labels the mash points consecutively from left to right and from top to bottom. For example, with $m = 5$ and $n = 4$, the labeling would result in a grid whose points are shown in Fig. 5.3.

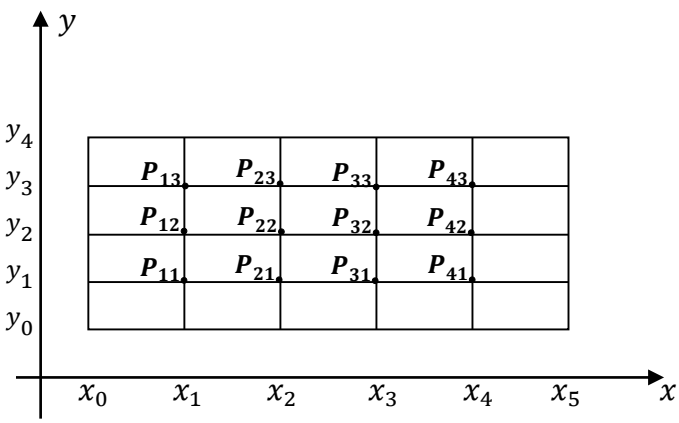


Fig. 5.3

♦ Step 2: Applying the boundary conditions

The second step after the mesh generation is to apply the boundary conditions (3) and/or (4) to the boundary mesh points of the grid.

♦ Step 3: Formulation of difference equations

Finally, in the third step we use the Taylor series for each mesh point $(x_i, y_j), i = 1,2,\cdots,m-1,\ j = 1,2,\cdots,n-1$ in the interior of the grid to generate the central difference formulas

$$\left(\frac{\partial u}{\partial x}\right)_{i,j} \cong \frac{u_{i+1,j} - u_{i-1,j}}{2h}, \tag{6}$$

$$\left(\frac{\partial u}{\partial y}\right)_{i,j} \cong \frac{u_{i,j+1} - u_{i,j-1}}{2k}, \tag{7}$$

For the boundary points we are forced to use the difference formulas

$$\left(\frac{\partial u}{\partial x}\right)_{i,j} \cong \frac{u_{i+1,j} - u_{i,j}}{h}, \tag{8}$$

$$\left(\frac{\partial u}{\partial y}\right)_{i,j} \cong \frac{u_{i,j+1} - u_{i,j}}{k}, \tag{9}$$

Partial derivatives of the second order are replaced analogously;

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \cong \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2}, \tag{10}$$

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} \cong \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{k^2}, \tag{11}$$

The indicated substitutions of derivatives at each mesh point enable us to reduce the solution of partial differential equations to that of a system of difference equations.

5.2 The finite difference method for elliptic partial differential equations

The elliptic partial differential equation we consider is the Poisson equation,

$$\nabla^2 u(x,y) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y) \qquad in\ \Omega \tag{1}$$

and the boundary conditions

$$u|_{\partial\Omega} = g(x,y) \tag{2}$$

where $\Omega \equiv \{(x,y)|a < x < b,\ c < y < d\}$. After carrying out the step of the mesh generation, we replace the partial derivatives in equation (1) by the finite difference formulas (10)–(11) of §6.1 to obtain

$$\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)_{ij} = \\ = \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2} + \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{k^2} = f_{ij}, \tag{3}$$

for each $, i = 1,2,\cdots,m-1,\ j = 1,2,\cdots,n-1,$ and the boundary conditions as

$$\begin{aligned} u(x_0,y_j) &= g(x_0,y_j), & j &= 0,1,\cdots,n \\ u(x_m,y_j) &= g(x_m,y_j), & j &= 0,1,\cdots,n \\ u(x_i,y_0) &= g(x_i,y_0), & i &= 0,1,\cdots,m \\ u(x_i,y_n) &= g(x_i,y_n), & i &= 0,1,\cdots,m \end{aligned} \tag{4}$$

Equations (3) and (4) form a system of linear algebraic equations with respect to the values of the function $u = u(x,y)$ at the mesh points

(x_i, y_j). For a rectangular domain with $h = k$, this system has the form

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{i,j} = h^2 f_{i,j}, \qquad (5)$$

For $f(x, y) = 0$, equation (1) is called the Laplace equation, and the corresponding finite difference equations have the form

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{i,j} = 0, \qquad (6)$$

The typical equation (5) or (6) involves approximations to $u(x, y)$ at the points

$$(x_{i+1}, y_j), (x_i, y_{j+1}), (x_{i-1}, y_j), (x_{i-1}, y_{j-1})$$

Reproducing the portion of the grid where these points are located (Fig. 5.4) shows that each equation involves approximations in a cross-shaped region about (x_i, y_j).

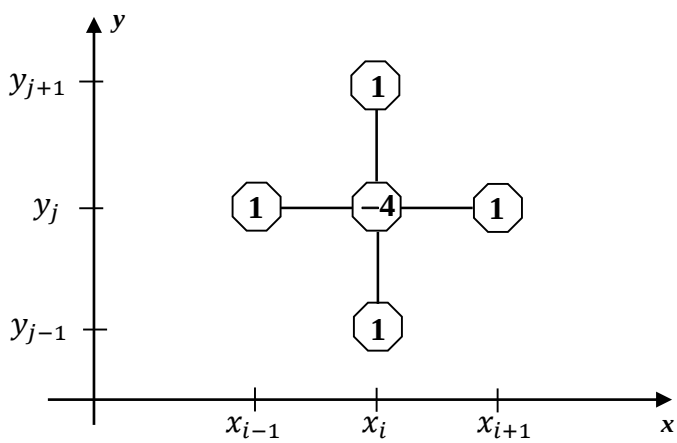


Fig. 5.4

Example 1 Solve the boundary value problem

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0, \quad 0<x<3,\ 0<y<3 \tag{7}$$

$$u(0,y)=50(3-y), \qquad 0\le y\le 3 \tag{8}$$

$$u(3,y)=50(3-y), \qquad 0\le y\le 3 \tag{9}$$

$$u(x,3)=0, \qquad 0\le x\le 3 \tag{10}$$

$$u(x,0)=\begin{cases}50(x+3), & 0\le x\le \frac{3}{2}\\ 50(6-x), & \frac{3}{2}\le x\le 3\end{cases} \tag{11}$$

Solution To apply the finite difference method let us start with a mesh size $h=k=1$.

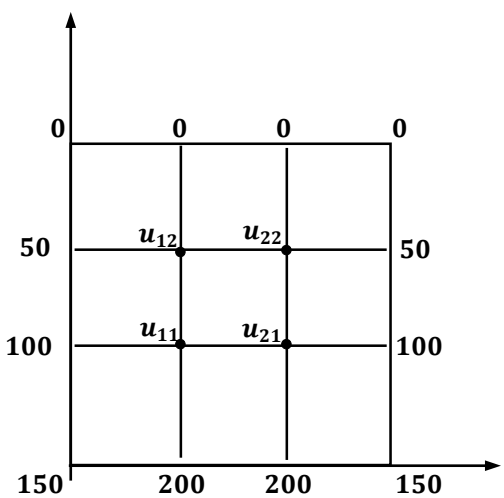


Fig. 5.5

As we see in Fig. 5.5, that choice yields four interior points and twelve boundary points. The numbers listed next to the boundary points are

the exact values of u obtained from the specified conditions along that boundary. Thus the boundary conditions (8) and (9) imply

$$u_{0,1} = u(0,1) = u_{3,1} = u(3,1) = 50(3-1) = 100$$

$$u_{0,2} = u(0,2) = u_{3,2} = u(3,2) = 50(3-2) = 50$$

$$u_{0,3} = u(0,3) = u_{3,3} = u(3,3) = 50(3-3) = 0,$$

and from the boundary condition (10) we get

$$u_{03} = u_{13} = u_{23} = u_{33} = 0,$$

and finally the boundary condition (10) gives

$$u_{00} = u(0,0) = 50(0+3) = 150,$$

$$u_{10} = u(1,0) = 50(1+3) = 200,$$

$$u_{20} = u(2,0) = 50(6-2) = 200,$$

$$u_{30} = u(3,0) = 50(6-3) = 150,$$

We now apply the difference equation (6) at each interior point to get the equations

$$\begin{array}{ccccccccccc} 100 & + & 200 & + & u_{21} & + & u_{12} & - & 4u_{11} & = & 0 \\ 50 & + & u_{11} & + & u_{22} & + & 0 & + & -4u_{12} & = & 0 \\ u_{11} & + & 200 & + & 100 & + & u_{22} & + & -4u_{21} & = & 0' \\ u_{12} & + & u_{21} & + & 50 & + & 0 & + & -4u_{22} & = & 0 \end{array}$$

or

$$\begin{array}{ccccccc} u_{21} & + & u_{12} & - & 4u_{11} & = & -300 \\ u_{11} & + & u_{22} & + & -4u_{12} & = & -50 \\ u_{11} & + & u_{22} & + & -4u_{21} & = & -300 \\ u_{12} & + & u_{21} & + & -4u_{22} & = & -50 \end{array} \qquad (12)$$

From the symmetry of the problem, we conclude that

$$u_{21} = u_{11}, \qquad u_{12} = u_{22},$$

thus the system of equations (12) reduces to the simultaneous two equations

$$\begin{array}{ccccc} -3u_{11} & + & u_{22} & = & -300 \\ u_{11} & - & 3u_{22} & = & -50 \end{array}$$

Solving these equations, we find the approximate solution at the four interior points to be

$$u_{21} = u_{11} = 118.75,$$
$$u_{12} = u_{22} = 56.25.$$

Example 2 Obtain the solution of the steady state boundary value problem $\nabla^2 u = 0$ inside the square bounded by the lines $x = 0$, $x = 1$, $y = 0$ and $y = 1$, given that $u = 640x(1-x)$ along $y = 1$ and $u = 0$ elsewhere on the boundary $\left(take\ h = k = \frac{1}{4}\right)$.

Solution From symmetry; $u_{11} = u_{31}$, $u_{12} = u_{32}$, $u_{13} = u_{33}$, thus

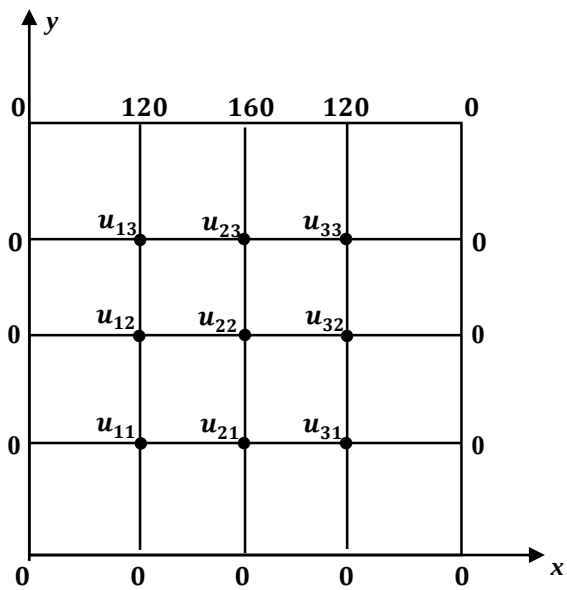


Fig. 5.6

Hence we consider only the interval

$$0 \le x \le \frac{1}{2}$$

The boundary conditions are applied to the mesh points as shown in Fig. 5.6, and the difference equations at the interior points are:

$$\begin{array}{ccccccccccl}
0 & + & 0 & + & u_{21} & + & u_{12} & - & 4u_{11} = 0 & & \text{at the point 11} \\
0 & + & u_{11} & + & u_{22} & + & u_{13} & - & 4u_{12} = 0 & & \text{at the point 12} \\
0 & + & u_{12} & + & u_{23} & + & 120 & - & 4u_{13} = 0 & & \text{at the point 13} \\
u_{11} & + & 0 & + & u_{31} & + & u_{22} & - & 4u_{21} = 0 & & \text{at the point 21} \\
u_{12} & + & u_{21} & + & u_{32} & + & u_{23} & - & 4u_{22} = 0 & & \text{at the point 22} \\
u_{13} & + & u_{22} & + & u_{33} & + & 160 & - & 4u_{23} = 0 & & \text{at the point 23}
\end{array}$$

or

$$\begin{array}{ccccccccl}
 & & u_{21} & + & u_{12} & - & 4u_{11} = 0 & & \text{at the point 11} \\
u_{11} & + & u_{22} & + & u_{13} & - & 4u_{12} = 0 & & \text{at the point 12} \\
u_{12} & + & u_{23} & + & 120 & - & 4u_{13} = 0 & & \text{at the point 13} \\
2u_{11} & & & + & u_{22} & - & 4u_{21} = 0 & & \text{at the point 21} \\
2u_{12} & + & u_{21} & + & u_{23} & - & 4u_{22} = 0 & & \text{at the point 22} \\
2u_{13} & + & u_{22} & + & 160 & - & 4u_{23} = 0 & & \text{at the point 23}
\end{array}$$

Solving these six equations in six unknowns by the method of Relaxation or any other method, we get;

$$u_{11} \cong 9, \quad u_{21} \cong 13, \quad u_{12} \cong 24,$$

$$u_{22} \cong 34, \quad u_{13} \cong 55, \quad u_{23} \cong 76$$

Example 3 Determine the steady temperature distribution in two dimensions, in a square of uniform conductivity when the temperature along the sides of the square is maintained at known values. Suppose the square is bounded by $x = 0$, $x = 4$, $y = 0$ and $y = 4$, and that the boundary values of the temperature u are given to be $(x^2 + y^3)°C$. Take $h = k = 1$.

Solution The values of u at the boundary points are calculated from the equation $u = (x^2 + y^3)°C$ and recorded in Fig. 5.7, so we have,

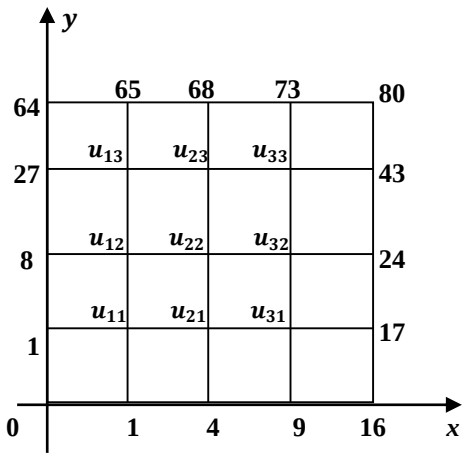


Fig. 5.7

$$u_{00} = 0^2 + 0^3 = 0, \qquad u_{40} = 4^2 + 0^3 = 16,$$
$$u_{01} = 0^2 + 1^3 = 1, \qquad u_{41} = 4^2 + 1^3 = 17,$$
$$u_{02} = 0^2 + 2^3 = 8, \qquad u_{42} = 4^2 + 2^3 = 24,$$
$$u_{03} = 0^2 + 3^3 = 27, \qquad u_{43} = 4^2 + 3^3 = 43,$$
$$u_{04} = 0^2 + 4^3 = 64, \qquad u_{44} = 4^2 + 4^3 = 80,$$
$$u_{10} = 1^2 + 0^3 = 1, \qquad u_{14} = 1^2 + 4^3 = 65,$$
$$u_{20} = 2^2 + 0^3 = 4, \qquad u_{24} = 2^2 + 4^3 = 68,$$
$$u_{30} = 3^2 + 0^3 = 9, \qquad u_{34} = 3^2 + 4^3 = 73$$

The difference equations are:

$$
\begin{array}{lccccccccc}
\text{at node 11:} & 1 & + & 1 & + & u_{21} & + & u_{21} & -4u_{11} = 0 \\
\text{at node 12:} & 8 & + & u_{11} & + & u_{22} & + & u_{13} & -4u_{12} = 0 \\
\text{at node 13:} & 27 & + & u_{12} & + & u_{23} & + & 65 & -4u_{13} = 0 \\
\text{at node 21:} & u_{11} & + & 4 & + & u_{31} & + & u_{22} & -4u_{21} = 0 \\
\text{at node 22:} & u_{12} & + & u_{21} & + & u_{32} & + & u_{23} & -4u_{22} = 0 \\
\text{at node 23:} & u_{13} & + & u_{22} & + & u_{33} & + & 68 & -4u_{23} = 0 \\
\text{at node 31:} & u_{21} & + & 9 & + & 17 & + & u_{32} & -4u_{31} = 0 \\
\text{at node 32:} & u_{22} & + & u_{31} & + & 24 & + & u_{33} & -4u_{32} = 0 \\
\text{at node 33:} & u_{23} & + & u_{32} & + & 43 & + & 73 & -4u_{33} = 0
\end{array}
$$

Solving these nine simultaneous equations by the Relaxation or any other method, we get values of at the internal points.

In the Relaxation method, notice that for the model equation

$$u_1 + u_2 + u_3 + u_4 - 4u_0 = 0$$

Assuming approximate values for the $u's$, the residue becomes

$$R_0 = u_1 + u_2 + u_3 + u_4 - 4u_0,$$

Hence if we change u_1, u_2, u_3, u_4, u_0 by $\Delta u_1, \Delta u_2, \Delta u_3, \Delta u_4$ and Δu_0, respectively. Then the corresponding changes in R_0 are $\Delta u_1, \Delta u_2, \Delta u_3, \Delta u_4$ and $-4\Delta u_0$, respectively. This gives what is called the Relaxation pattern. If we change one u by unity the corresponding changes in $R's$ are as shown in the diagram shown in Fig. 5.8:

The process of Relaxation is aiming at reducing the residuals to zero. In the above example at the nine internal nodes initially guessed values of u are:

$$
\begin{array}{lll}
u_{11}^{(0)} = 10, & u_{21}^{(0)} = 20, & u_{31}^{(0)} = 10, \\
u_{12}^{(0)} = 20, & u_{22}^{(0)} = 30 & u_{32}^{(0)} = 40, \\
u_{13}^{(0)} = 30, & u_{23}^{(0)} = 40, & u_{33}^{(0)} = 50
\end{array}
$$

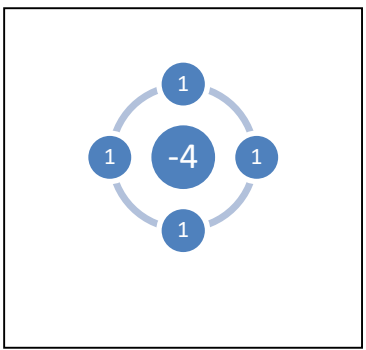


Fig. 5. 8

Substituting these initial values in the above equations, the initial residuals are

$$R_{11} = 2, \qquad R_{21} = -6, \qquad R_{31} = -34,$$
$$R_{12} = -2, \qquad R_{22} = 0, \qquad R_{32} = -26,$$
$$R_{13} = 32, \qquad R_{23} = 18, \qquad R_{33} = -4,$$

Picking out the largest of the initial residuals (-34) and with the use of point relaxation alone, apply an increment Δu to bring-that residual to zero. Take for example $\Delta u_0 = -9$. The effect of this increment is to change the residual from -34 to 2 and the surrounding residuals (one above and. one to the left), from $-26, -6$ to $-35, -15$, respectively. This is the first step.

Now the residual (-35) is the largest, take

$$\Delta u_0 = -9,$$

and the new residual at this point becomes 1 and the residuals below and to the right change from $-2, 18$ to $6, 26$ respectively. This is the second step.

Now the residue at node u_{13} is the largest one $(R_{13} = 32)$, so we take

$$\Delta u_{13} = 8,$$

and the new residual at this point becomes 0 and the residuals above and to the left and below change from $-4, 0, 2$ to $-13, -9, -7$ respectively. This is the third step.

Continue this process by decreasing the largest residual at this moment, which is 26 at node 23, by taking $\Delta u_0 = 7$ and so on.

Finally we get the approximate solution:

$$u_{11} = 9, \qquad u_{21} = 14, \qquad u_{31} = 17,$$
$$u_{12} = 20, \qquad u_{22} = 26, \qquad u_{32} = 28,$$
$$u_{13} = 39, \qquad u_{23} = 42, \qquad u_{33} = 44,$$

All the above steps are recorded in the following table.

64		65		68		73	80
27	u_{13}	R_{13}	u_{23}	R_{23}	u_{33}	R_{33}	43
	30	32	**40**	18	**50**	-4	
	8	0	**5**	0	**-2**	-9	
	2	32	**-1**	18	**-3**	-13	
	-1	-32	**-1**	8	**0**	5	
	0	0	**-1**	18	**-1**	-8	
		5		-20		-2	
		5		-2		-10	
		2		-2		-1	
		39		**42**		**44**	
8	u_{12}	R_{12}	u_{22}	R_{22}	u_{32}	R_{32}	24
	20	-2	**30**	0	**40**	-26	
	0	**1**	**0**	-1	**-9**	-9	
	2	-1	**-2**	-1	**-2**	-35	
	-1	0	**-1**	0	**-1**	0	
	0	-1	**-1**	-1	**0**	-35	
	-1	0	**0**	0	**0**	36	
		-1		-1		1	
		20		**26**		**28**	
1	u_{11}	R_{11}	u_{21}	R_{31}	u_{31}	R_{31}	17
	10	2	**20**	-6	**30**	-34	
	1	-4	**-1**	**1**	**-9**	-1	
	-1	-2	**-3**	-5	**-3**	-35	
	0	-1	**-1**	4	**-1**	36	
	-1	-3	**-1**	-1	**0**	1	
	0	0	**0**	-9	**0**	-9	
		-3		-10		-8	
		4		0		-3	
		1		-10		-11	
0		**9**		**14**		**17**	16
		1		4		9	

Fig. 5.9

Example 4 Solve

$$\nabla^2 u = -10x, \quad 0 \le x \le 3, \qquad 0 \le y \le 3$$

and $u = 20x + 30y$, on the boundaries (take $h = k = 1$).

Solution The values of u at the boundary points are calculated from

$$u(x, y) = 20x + 30y,$$

and recorded in Fig. 5.10,

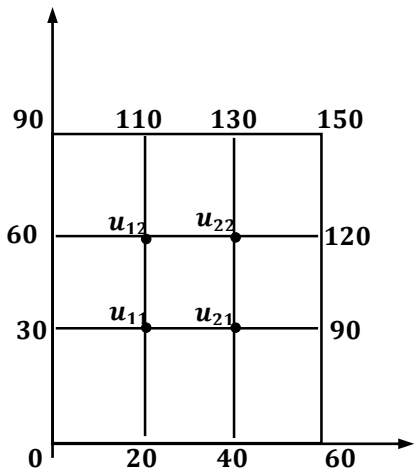


Fig. 5.10

$$u_{00} = 20(0) + 30(0) = 0$$
$$u_{01} = 20(0) + 30(1) = 30,$$
$$u_{02} = 20(0) + 30(2) = 60,$$
$$u_{03} = 20(0) + 30(3) = 90,$$
$$u_{10} = 20(1) + 30(0) = 20,$$
$$u_{20} = 20(2) + 30(0) = 40,$$
$$u_{30} = 20(3) + 30(0) = 60,$$
$$u_{31} = 20(3) + 30(1) = 90,$$
$$u_{32} = 20(3) + 30(2) = 120,$$
$$u_{33} = 20(3) + 30(3) = 150,$$
$$u_{13} = 20(1) + 30(3) = 110,$$
$$u_{23} = 20(2) + 30(3) = 130,$$

The difference equation is;

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-} - 4u_{i,j} = -10x_j \qquad (13)$$

Substituting for the shown nodes we get the following equation;

$$30 + 20 + u_{21} + u_{12} - 4u_{11} = -10(1) = -10$$
$$60 + u_{11} + u_{22} + 110 - 4u_{12} = -10(1) = -10$$
$$u_{11} + 40 + 90 + u_{22} - 4u_{21} = -10(2) = -20$$
$$u_{12} + u_{21} + 120 + 130 - 4u_{22} = -10(2) = -20$$

or

$$\begin{array}{ccccccccc} u_{21} & + & u_{12} & - & 4u_{11} & + & 60 & = 0 \\ u_{11} & + & u_{22} & - & 4u_{12} & + & 180 & = 0 \\ u_{11} & + & u_{22} & - & 4u_{21} & + & 150 & = 0 \\ u_{11} & + & u_{21} & - & 4u_{22} & + & 270 & = 0 \end{array} \qquad (14)$$

These four equations in the system (13) can be solved by any method (such as matrices, Relaxation method).

We will solve them by the Relaxation method. On taking

$$\begin{bmatrix} u_{11}^{(0)} \\ u_{12}^{(0)} \\ u_{21}^{(0)} \\ u_{22}^{(0)} \end{bmatrix} = \begin{bmatrix} 50 \\ 70 \\ 60 \\ 100 \end{bmatrix},$$

the residuals are

$$\begin{array}{ccccccccc} R_{11} = & (60) & + & (70) & - & 4(50) & + & 60 & = -10 \\ R_{12} = & (50) & + & (100) & - & 4(70) & + & 180 & = 50 \\ R_{21} = & (50) & + & (100) & - & 4(60) & + & 150 & = 60 \\ R_{22} = & (70) & + & (60) & - & 4(100) & + & 270 & = 0 \end{array} \qquad (15)$$

We tabulate these values as shown in Fig. 5.11, and proceed in the Relaxation procedure to get the following approximated values of the solution:

$$\begin{bmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 57 \\ 86 \\ 79 \\ 109 \end{bmatrix}$$

u_{12}	R_{12}
70	50
12	– 48
	2
	8
	10
	6
4	16
	– 16
	0
	1
	1
	1
	2
86	

u_{22}	R_{22}
100	20
	12
8	32
	– 32
	0
	4
1	4
	– 4
	0
109	

u_{11}	R_{11}
50	10
	12
6	22
	– 24
	–2
	4
1	2
	– 4
	– 2
57	

u_{21}	R_{21}
80	– 20
	8
	– 12
	6
–1	–6
	4
	– 2
	1
	– 1
	1
	0
79	

Fig. 5.11

Exercise 5.1

In problems **1–5**, approximate the solution to the following elliptic partial differential equations

1. $$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial x^2}=4,\quad 0<x<1, 0<y<2,$$

$$u(x,0)=x^2,\quad u(x,2)=(x-2)^2,\ ,\quad 0\le x\le 1,$$

$$u(0,y)=y^2,\quad u(1,y)=(y-1)^2,\ ,\quad 0\le y\le 2$$

Use $h=k=0.5$ and compare the results with the actual solution $u(x,y)=(x-y)^2$.

2. $$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial x^2}=x,\quad 0<x<1, 0<y<1,$$

$$u(x,0)=u(x,1)=\frac{1}{6}x^3,\qquad 0\le x\le 1,$$

$$u(0,y)=0,\quad u(1,y)=\frac{1}{6}\ ,\quad 0\le y\le 1$$

Use $h=k=\frac{1}{3}$.

3. $$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial x^2}=0,\quad 0<x<1, 0<y<1,$$

$$u(x,0)=0,\quad u(x,1)=x,\ ,\quad 0\le x\le 1,$$

$$u(0,y)=0,\quad u(1,y)=y,\ ,\quad 0\le y\le 1$$

Use $h=k=0.25$ and compare the results with the actual solution $u(x,y)=xy$.

4. $$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial x^2}=(x^2+y^2)e^{xy},\quad 0<x<2, 0<y<1,$$

$$u(x,0)=1,\quad u(x,1)=e^x,\quad 0\le x\le 2,$$

$$u(0,y)=1,\quad u(1,y)=e^{2y},\quad 0\le y\le 1$$

Use $h=k=0.2$ and compare the results with the actual solution $u(x,y)=e^{xy}$.

5. $$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial x^2}=-2,\quad 0<x<1, 0<y<1,$$

$$u(x,0)=\ u(x,1)=x(1-x),\qquad 0\le x\le 1,$$

$$u(0,y)=0,\quad u(1,y)=x\sinh\pi\sin\pi y\,,\quad 0\le y\le 1$$

Use $h=k=0.25$ and compare the results with the actual solution $u(x,y)=\sinh\pi x\sin\pi y$.

6. Find a solution of Laplace equation $\Delta u=0$ in the strip $0\le x\le 4, 0\le y\le 4$, where the solution satisfies the conditions

$$u(x,0)=5x(4-x),\quad u(x,4)=0,\quad 0\le x\le 4,$$

$$u(0,y)=u(4,y)=0\ \ ,\quad 0\le y\le 4$$

Use the finite difference method for $h=k=1$.

7. Apply the Relaxation method to solve the equation $\Delta u=0$ in a square of side unit with $u=100$ along the boundaries. Take $h=k=0.25$.

8. A long square pipe with a square hole carrying a hot fluid. The pipe is half submerged in an ice bath so that the bottom half of the outside of the pipe is at $0°C$. The top surface of the pipe is at $100°C$. Assume that the temperature varies linearly between the top of the ice bath and the top of the pipe. The fluid inside the pipe is at $200°C$. The dimensions of the pipe are show in Fig. 5.12. The temperature in the material of the pipe for the steady state satisfies $u_{xx}+u_{yy}=0$. Solve to find u inside the material of the pipe.

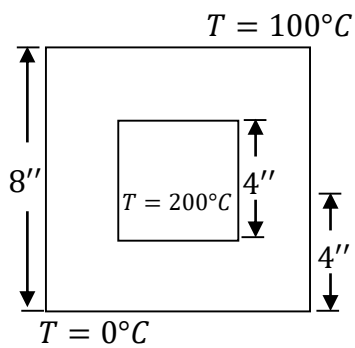


Fig. 5.12

9. Solve Laplace equation $\Delta u = 0$ at the interior points of a unit square. Use the mesh size $h = k = 0.25$ and the boundary conditions

$$u(x,0) = 0, \quad u(x,1) = 100x, \quad 0 \le x \le 1,$$
$$u(0,y) = 0, \quad u(1,y) = 100y, \quad 0 \le y \le 1$$

10. Solve the equation $\Delta u = -100$ over a rectangular region of sides 4 units and 3 units. Use a mesh length of one unit assuming $u = 0$ on the boundaries.

11. Rework problem-(7) if the equation is $\Delta u = -200$.

12. A function u satisfies $\Delta u = 0$ inside the quadrant of the ellipse $4x^2 + 9y^2 = 144$. u vanishes on the ellipse and takes the value 100 on both the semi major and semi minor axes. Use a square network of mesh length one- twelfth the major axis to find the approximate values of the function at the internal nodal points of the mesh.

13. Solve Poisson's equation $\Delta u = 100y$ for the shown square in Fig. 5.13, of side four units ($h = k = 1$ is one unit).

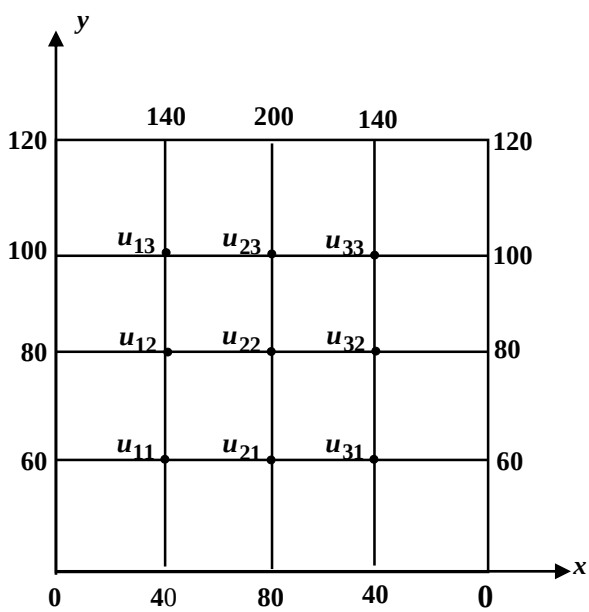


Fig. 5.13

14. Solve Poisson's equation $\Delta u = -2$ at the interior points of the region in Fig. 5.14. The mesh size is $h = k = 0.5$, $u = 1$ at every point along $ABCD$, and $u = 0$ at every point along $DEFGA$. Use symmetry and, if necessary, Gauss-Seidel iteration.

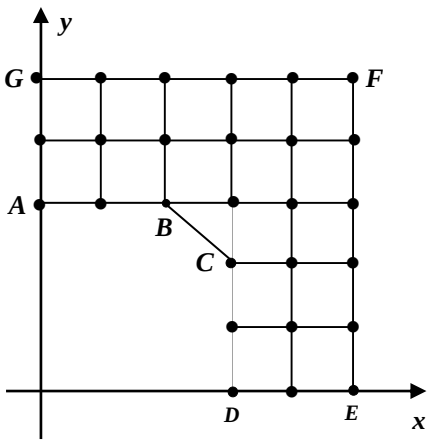


Fig. 5.14

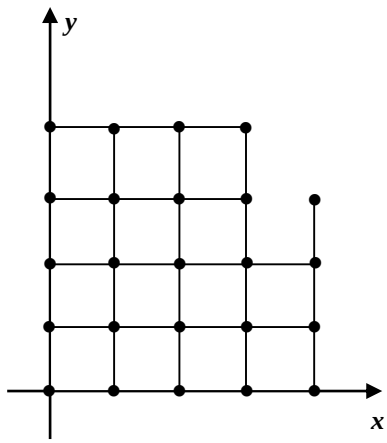


Fig. 5.15

15. Solve Poisson's equation $\Delta u = -64$ at the interior points of the region in Fig. 5.15. The mesh size is $h = k = \frac{1}{8}$, and $u = 0$ at every point on the boundary of the region. If necessary, use Gauss-Seidel iteration.

5.3 The finite difference method for parabolic partial differential equations

The parabolic partial differential equation we will study is the heat or diffusion equation,

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < \ell, \;\; t > 0 \tag{1}$$

subject to the boundary and initial conditions

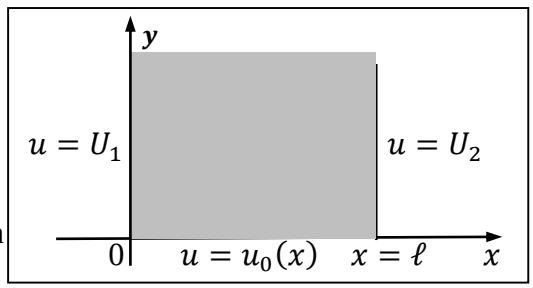
$$u(0,t) = U_1(t), \qquad u(\ell,t) = U_2(t), \qquad t > 0 \tag{2}$$

$$u(x,0) = u_0(x), \qquad 0 \le x \le \ell, \tag{3}$$

The first step in the finite difference method is to create a grid with two-mesh constants h and k. The grid points for this situation are $(x_i,\ t_j)$ where

$x_i = ih, \quad i = 0,1,\cdots,m$

$t_j = jk, \quad j = 0,1,\cdots$



To approximate the solution of (1) we replace the heat equation by a difference equation as we did in the previous section. Using the scheme labeling

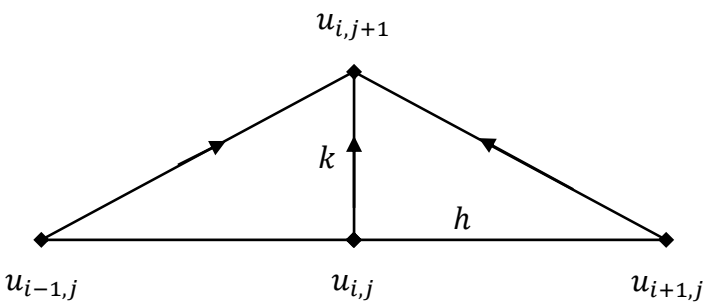


Fig. 5.16

to form the forward difference approximation

$$\frac{\partial u}{\partial t} \cong \frac{u_{i,j+1} - u_{i,j}}{k}, \tag{4}$$

and the difference quotient

$$\frac{\partial^2 u}{\partial x^2} \cong \frac{u_{i-1,j} + u_{i+1,j} - 2u_{i,j}}{h^2}, \tag{5}$$

5.3.1 Explicit difference scheme

The partial differential equation (1) implies that at the interior grid point $(x_i,\ t_j)$ for each $i = 1,2,\cdots,m-1$ and $j = 1,2,\cdots$, we have

$$\underbrace{\frac{\partial u}{\partial t}} = c^2 \underbrace{\frac{\partial^2 u}{\partial x^2}}$$

$$\frac{u_{i,j+1} - u_{i,j}}{k} = c^2 \frac{u_{i-1,j} + u_{i+1,j} - 2u_{i,j}}{h^2}, \tag{6}$$

Collecting the coefficients of $u_{i,j}$ in this equation reduces it to the form

$$u_{i,j+1} = \frac{kc^2}{h^2}\left(u_{i-1,j} + u_{i+1,j}\right) + \left(1 - \frac{2kc^2}{h^2}\right)u_{i,j}, \tag{7}$$

Denoting by

$$\lambda = \frac{kc^2}{h^2},$$

we get the explicit difference formula

$$\boxed{u_{i,j+1} = \lambda\left(u_{i-1,j} + u_{i+1,j}\right) + (1 - 2\lambda)u_{i,j}} \tag{8}$$

for each $i = 1,2,\cdots,m-1$ and $j = 1,2,\cdots$.

Since the initial condition $u(x,0) = u_0(x)$ for each $0 \le x \le \ell$, implies $u_{\ell,0}(x) = u_0(x_\ell)$, for each $i = 0,1,\cdots,m$, these values can be used in equation (8) to find the value of

$$u_{i,1}, \qquad \forall i = 1,2,\cdots,m-1$$

Using the additional boundary conditions $u(0,t) = U_1(t)$ and $(\ell,t) = U_2(t)$, all the entries of the form $u_{i,1}$ can be determined. If the procedure is reapplied once all the approximations $u_{i,1}$, are known, the values of $u_{i,2}, u_{i,3}, \cdots$ can be obtained in a similar manner.

Algorithm (8) is unstable if the error grows too rapidly as the computation proceed. However, it can be proved that the condition of round-off stability of the difference scheme (8) is

$$\boxed{0 \le \lambda \le \frac{1}{2}} \tag{9}$$

Since

$$\lambda = \frac{kc^2}{h^2},$$

the inequality (9) requires that h and k be chosen so that $\lambda \le \frac{1}{2}$.

Remark Practically, the simplest application of the difference scheme (8) is carried out when we chose $\lambda = \frac{1}{2}$. In this case the difference scheme takes the form

$$u_{i,j+1} = \frac{1}{2}\left(u_{i-1,j} + u_{i+1,j}\right) \tag{10}$$

Example1 Find an approximate solution to the equation

$$25\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \;\; 0 \le t \le 2,$$

that satisfies the conditions:

$$u(0,t) = 0, \qquad 0 \le t \le 2$$

$$u(1,t) = 0.5, \qquad 0 \le t \le 2$$

$$u(x,0) = x(1.5 - x), \qquad 0 \le x \le 1$$

Solution Suppose that $h = 0.2$ and we want to use formula (10), then we choose k such that

$$\lambda = \frac{kc^2}{h^2} = \frac{k(1/25)}{(0.2)^2} = \frac{1}{2},$$

which gives $k = 0.5$. Using the two-mesh constants h and k to create the grid shown in Fig. 6.16.

Applying the initial condition

$$u(x,0) = x(1.5 - x), \qquad 0 \le x \le 1,$$

to the boundary points, we get

$$u_{0,0} = u(0,0) = x(1.5 - x) = 0,$$

$$u_{1,0} = u(0.2,0) = (0.2)(1.5 - 0.2) = 0.26,$$

$$u_{2,0} = u(0.4,0) = (0.4)(1.5 - 0.4) = 0.44,$$

$$u_{3,0} = u(0.6,0) = (0.6)(1.5 - 0.6) = 0.54,$$

$$u_{4,0} = u(0.8,0) = (0.8)(1.5 - 0.8) = 0.56,$$

$$u_{5,0} = u(1,0) = (1)(1.5 - 1) = 0.5,$$

List these values on the grid in addition to the boundary conditions

$$u(0,t) = 0, \qquad 0 \le t \le 2$$

$$u(1,t) = 0.5, \qquad 0 \le t \le 2$$

The results of applying the difference scheme (10) to the interior points of the grid are: (see Fig. 5.17).

$$u_{11} = \frac{1}{2}(0 + 0.44) = 0.22,$$

$$u_{21} = \frac{1}{2}(0.26 + 0.54) = 0.4,$$

$$u_{31} = \frac{1}{2}(0.44 + 0.56) = 0.5,$$

$$u_{41} = \frac{1}{2}(0.54 + 0.5) = 0.52,$$

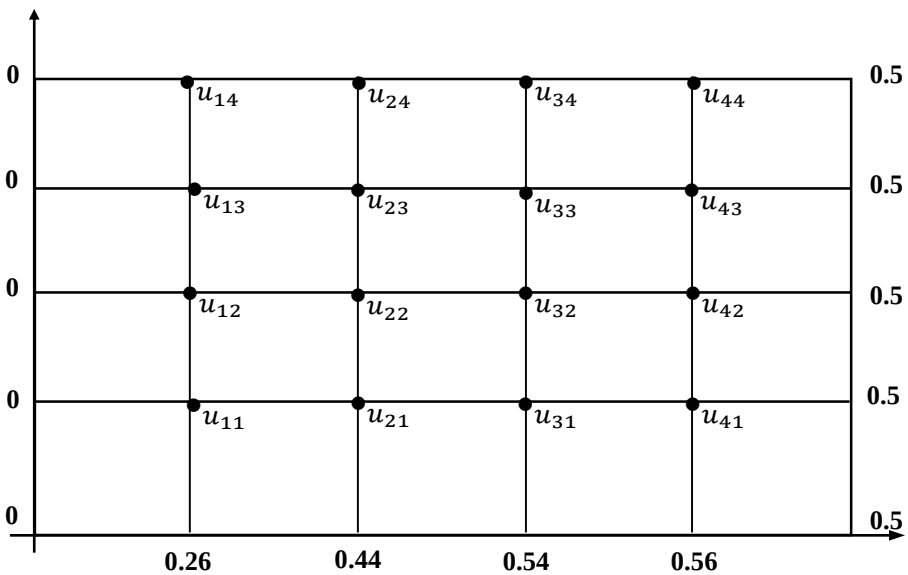


Fig. 5.17

These values are the approximate solutions at the time level $t = 0.5$. The values of this level will be used to obtain the solution at the next level $t = 1$:

$$u_{12} = \frac{1}{2}(0 + 0.4) = 0.2,$$

$$u_{22} = \frac{1}{2}(0.22 + 0.5) = 0.36,$$

$$u_{32} = \frac{1}{2}(0.4 + 0.52) = 0.46,$$

$$u_{42} = \frac{1}{2}(0.5 + 0.5) = 0.5,$$

The solution at the next level $(t = 1.5)$ is

$$u_{13} = \frac{1}{2}(0 + 0.36) = 0.18,$$

$$u_{23} = \frac{1}{2}(0.2 + 0.46) = 0.33,$$

$$u_{33} = \frac{1}{2}(0.36 + 0.5) = 0.43,$$

$$u_{43} = \frac{1}{2}(0.46 + 0.5) = 0.48,$$

and finally, the solution at the time level $t = 2$

$$u_{14} = \frac{1}{2}(0 + 0.33) = 0.165,$$

$$u_{24} = \frac{1}{2}(0.18 + 0.43) = 0.305,$$

$$u_{34} = \frac{1}{2}(0.33 + 0.48) = 0.405,$$

$$u_{44} = \frac{1}{2}(0.43 + 0.5) = 0.465$$

Example 2 Solve the initial boundary-value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \;\; 0 \le t \le 0.5,$$

that satisfies the conditions:

$$u(0,t) = u(1,t) = 0, \qquad 0 \le t \le 0.5$$

$$u(x,0) = \sin \pi x\,, \qquad 0 \le x < 1$$

Solution Suppose that $h = 0.2$ and $k = 0.01$, then

$$\lambda = \frac{kc^2}{h^2} = \frac{(0.01)(1)}{(0.2)^2} = \frac{1}{4},$$

Using the two-mesh constants h and k to create the grid, we see a portion of it in the following figure.

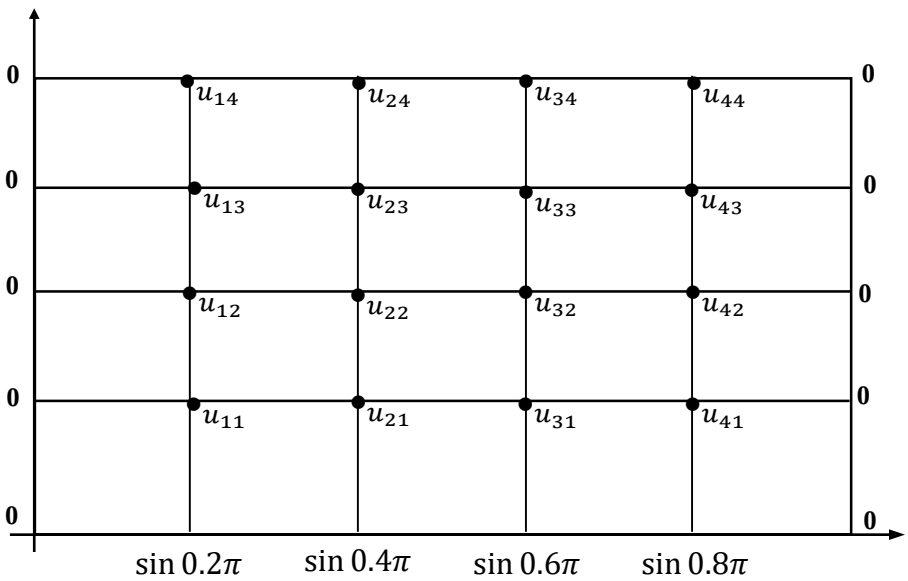


Fig. 5.18

The difference scheme (8) becomes

$$u_{i,j+1} = 0.25\left(u_{i-1,j} + u_{i+1,j}\right) + 0.5u_{i,j}$$

By setting $j = 0$ in this formula we get a formula for the approximation to u on the first time level:

$$\begin{aligned} u_{11} &= 0.25\left(u_{00} + u_{2,0}\right) + 0.5u_{10} \\ &= 0.25(0 + \sin 0.4\pi) + 0.5(\sin 0.2\pi) = 0.5316568, \end{aligned}$$

and in a similar manner, we compute the other values of u at the time level $t = 0.01$:

$$u_{21} = 0.25(\sin 0.2\pi + \sin 0.6\pi) + 0.5(\sin 0.4\pi) = 0.8602387,$$

$$u_{31} = 0.25(\sin 0.4\pi + \sin 0.8\pi) + 0.5(\sin 0.6\pi) = 0.8602387,$$

$$u_{41} = 0.25(\sin 0.6\pi + 0) + 0.5(\sin 0.8\pi) = 0.5316568,$$

These values are used to obtain the solution at the next level $t = 0.02$:

$$u_{12} = 0.25(0 + 0.8602387) + 0.5(0.5316568) = 0.480881,$$

$$\begin{aligned} u_{22} &= 0.25(0.5316568 + 0.8602387) + 0.5(0.8602387) \\ &= 0.7780932, \end{aligned}$$

$$\begin{aligned} u_{32} &= 0.25(0.8602387 + 0.5316568) + 0.5(0.8602387) \\ &= 0.7780932, \end{aligned}$$

$$u_{42} = 0.25(0.8602387 + 0) + 0.5(0.5316568) = 0.4808881,$$

and

$$u_{13} = 0.25(0 + 0.7780932) + 0.5(0.480881) = 0.4349673,$$

$$\begin{aligned} u_{23} &= 0.25(0.480881 + 0.7780932) + 0.5(0.7780932) \\ &= 0.7037919, \end{aligned}$$

$$\begin{aligned} u_{33} &= 0.25(0.7780932 + 0.4808881) + 0.5(0.7780932) \\ &= 0.7037919, \end{aligned}$$

$$u_{43} = 0.25(0.7780932 + 0) + 0.5(0.4808881) = 0.4349673,$$

and finally, we compute the solution at the time level $t = 0.04$:

$$u_{14} = 0.25(0 + 0.7037919) + 0.5(0.4349673) = 0.3934316,$$
$$u_{24} = 0.25(0.4349673 + 0.7037919) + 0.5(0.7037919)$$
$$= 0.6365858,$$
$$u_{34} = 0.25(0.7037919 + 0.4349673) + 0.5(0.7037919)$$
$$= 0.6365858,$$
$$u_{44} = 0.25(0.7037919 + 0) + 0.5(0.4349673) = 0.3934316,$$

and so on we can evaluate the solution at all the interior mesh points of our grid.

Example 3 Given the initial boundary-value problem

$$20\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 0.5, \quad t > 0$$
$$u(0,t) = 200, \qquad t > 0$$
$$u(0.5,t) = 100, \quad t > 0$$
$$u(x,0) = 50(1-2x), \qquad 0 < x < 0.5,$$

Take $h = 0.1$ and $k = 0.04$, to approximate $u(x, 0.08)$, $0 \le x \le 0.5$ correct to three decimal places.

Solution Here we have $h = 0.1$ and $k = 0.04$, then

$$\lambda = \frac{kc^2}{h^2} = \frac{(0.04)(1/20)}{(0.1)^2} = 0.2,$$

The mesh generation and labeling are shown in Fig. 5.19. Applying the initial condition

$$u(x,0) = 50(1-2x), \qquad 0 < x < 0.5,$$

to the boundary points, we get

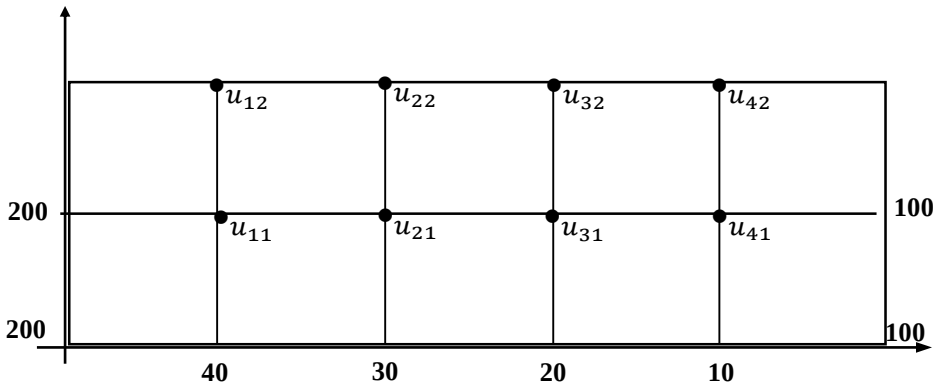


Fig. 5.19

$$u_{1,0} = u(0.1,0) = 50\big(1 - 2(0.1)\big) = 40,$$
$$u_{2,0} = u(0.2,0) = 50\big(1 - 2(0.2)\big) = 30,$$
$$u_{3,0} = u(0.3,0) = 50\big(1 - 2(0.3)\big) = 20,$$
$$u_{4,0} = u(0.4,0) = 50\big(1 - 2(0.4)\big) = 10,$$

Using the difference scheme:

$$u_{i,j+1} = 0.2\big(u_{i-1,j} + u_{i+1,j}\big) + 0.6u_{i,j},$$

to obtain the solution at the time level $t = 0.04$:

$$u_{1,1} = 0.2(200 + 30) + 0.6(40) = 70$$
$$u_{2,1} = 0.2(40 + 20) + 0.6(30) = 30$$
$$u_{3,1} = 0.2(30 + 10) + 0.6(20) = 20$$
$$u_{4,1} = 0.2(20 + 100) + 0.6(10) = 30$$

Therefore, the solution at the time level $t = 0.08$:

$$u_{1,2} = 0.2(200 + 30) + 0.6(70) = 88$$
$$u_{2,2} = 0.2(70 + 20) + 0.6(30) = 36$$
$$u_{3,2} = 0.2(30 + 30) + 0.6(20) = 24$$
$$u_{4,2} = 0.2(20 + 100) + 0.6(30) = 42$$

5.3.2 Implicit difference scheme (Crank-Nicolson scheme)

To obtain a method that is unconditionally stable, we consider an implicit difference scheme that results from replacing the second partial derivatives $\frac{\partial^2 u}{\partial x^2}$ by an average of two central difference quotients, one evaluated at the node $u_{i,j}$ and the other at $u_{i,j+1}$, i.e.,

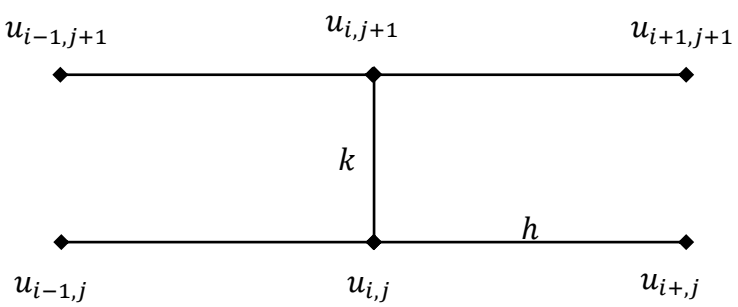


Fig. 5.20

$$\frac{\partial^2 u}{\partial x^2} \cong \frac{1}{2}\left[\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} + \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j+1}\right],$$

which implies that

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \cong \frac{1}{2}\left[\frac{u_{i-1,j} + u_{i+1,j} - 2u_{i,j}}{h^2} + \frac{u_{i-1,j+1} + u_{i+1,j+1} - 2u_{i,j+1}}{h^2}\right]$$

Thus, the heat equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2},$$

is approximated by the difference equation

$$\frac{u_{i,j+1}-u_{i,j}}{k} = \frac{c^2}{2}\left[\frac{u_{i-1,j}+u_{i+1,j}-2u_{i,j}}{h^2}+\frac{u_{i-1,j+1}+u_{i+1,j+1}-2u_{i,j+1}}{h^2}\right], \quad (11)$$

If we again define

$$\lambda = \frac{kc^2}{h^2},$$

then we can write (11) as

$$u_{i,j+1}-u_{i,j} = \\ = \frac{\lambda}{2}\left[u_{i-1,j}+u_{i+1,j}-2u_{i,j}+u_{i-1,j+1}+u_{i+1,j+1}-2u_{i,j+1}\right], \quad (12)$$

or,

$$-\frac{\lambda}{2}u_{i-1,j+1}+(1+\lambda)u_{i,j+1}-\frac{\lambda}{2}u_{i+1,j+1} = \\ = \frac{\lambda}{2}u_{i-1,j}+(1-\lambda)u_{i,j}+\frac{\lambda}{2}u_{i+1,j}, \quad (13)$$

or,

$$-u_{i-1,j+1}+\frac{2(1+\lambda)}{\lambda}u_{i,j+1}-u_{i+1,j+1} = \\ = u_{i-1,j}+\frac{2(1-\lambda)}{\lambda}u_{i,j}+u_{i+1,j}, \quad (14)$$

Equation (14) can be rewritten in the form

$$-u_{i-1,j+1}+\alpha u_{i,j+1}-u_{i+1,j+1} = u_{i-1,j}-\beta u_{i,j}+u_{i+1,j}, \quad (15)$$

where

$$\alpha = 2\left(1+\frac{1}{\lambda}\right), \quad and \quad \beta = 2\left(1-\frac{1}{\lambda}\right),$$

$i = 1,2,\cdots,m-1$ and $j = 0,1,\cdots$.

The scheme (15) is known as ***Crank-Nicolson method*** and is unconditionally stable for all values of λ.

The implicit nature of the difference method expressed in equation (15) implies a system of $(n-1)$ equations in the $(n-1)$ unknowns $u_{1,j+1}, u_{2,j+1}, \cdots, u_{n-1,j+1}$, which can be solved by any approximate method.

Example 4 Solve the initial boundary-value problem

$$\frac{\partial u}{\partial t} = 6.25\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \ \ t > 0$$

$$u(0,t) = u(1,t) = 0, \qquad t > 0$$

$$u(x,0) = \sin \pi x, \qquad 0 < x < 1,$$

Take $h = 0.25$ and $k = 0.01$.

Solution Here we have $c^2 = 6.25$, thus

$$\lambda = \frac{kc^2}{h^2} = \frac{(0.01)(6.25)}{(0.25)^2} = 1,$$

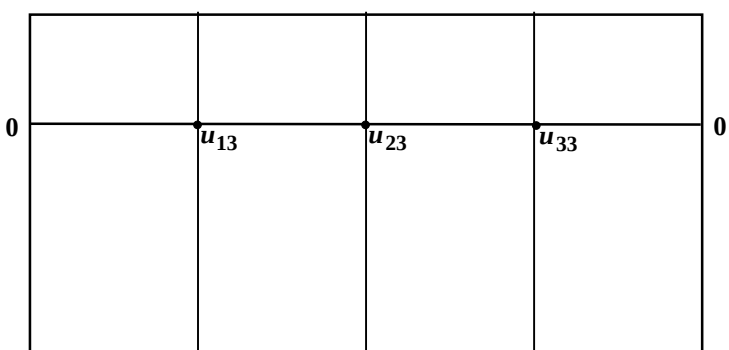
$$\alpha = 2\left(1 + \frac{1}{\lambda}\right) = 4, \qquad \beta = 2\left(1 - \frac{1}{\lambda}\right) = 0$$

Using the mesh steps h and k to create the grid, we see a portion of it in Fig. 5.21.

The difference equation (15) can be rewritten in the form

$$-u_{i-1,j+1} + 4u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j},$$

for $i = 1,2,\cdots,m-1$ and $j = 0,1,\cdots$. By setting $j = 0$ in this formula we get a formula for the approximation to u on the first time level:



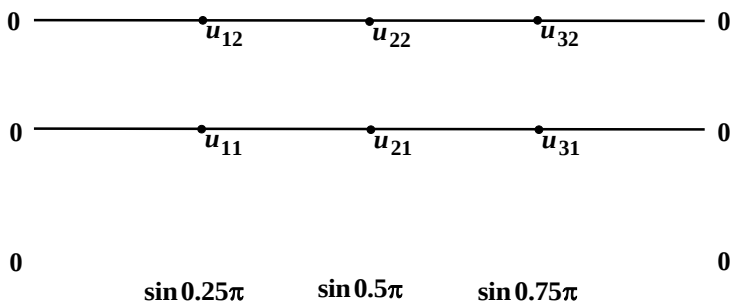


Fig. 5.21

$$0 + 4u_{11} - u_{21} = 0 + \sin(0.5\pi)$$
$$-u_{11} + 4u_{21} - u_{31} = \sin(0.25\pi) + \sin(0.75\pi)$$
$$-u_{21} + 4u_{31} - 0 = \sin(0.5\pi) + 0$$

or,

$$\begin{array}{rcrcrcr} 4u_{11} & - & u_{21} & & & = & 1 \\ -u_{11} & + & 4u_{21} & - & u_{31} & = & 1.414213562 \\ & - & u_{21} & + & 4u_{31} & = & 1 \end{array}$$

On using Gauss elimination method to solve this system of equations we get

$$\begin{bmatrix} u_{11} \\ u_{21} \\ u_{31} \end{bmatrix} = \begin{bmatrix} 0.38672954 \\ 0.54691816 \\ 0.38672954 \end{bmatrix}$$

In a similar way we can compute the solution at the other time levels.

Exercise 5.2

In problems 1-5, approximate the solution to the following parabolic partial differential equations using the explicit and implicit methods.

1. $$\frac{\partial u}{\partial t} = 0.05\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2,\ \ 0 < t < 2$$
$$u(0,t) = u(1,t) = 0, \qquad 0 \le t \le 2$$
$$u(x,0) = 10x(2-x), \qquad 0 < x < 2$$

Take $h = 0.2$.

2. $$\frac{\partial u}{\partial t} = \frac{4}{25}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2,\ \ t > 0$$
$$u(0,t) = 2, \qquad 0 \le t \le 2$$
$$u(2,t) = 12, \qquad 0 \le t \le 2$$
$$u(x,0) = 5x + 2, \qquad 0 \le x \le 2$$

Take $h = 0.4$.

3. $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1,\ \ t > 0$$
$$u(0,t) = u(2,t) = 0, \qquad 0 \le t \le 2$$
$$u(x,0) = \sin\left(\frac{\pi x}{2}\right), \qquad 0 \le x \le 1$$

Take $h = 0.25$.

4. $$\frac{\partial u}{\partial t} = \frac{1}{16}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2,\ \ 0 < t < 0.1$$
$$u(0,t) = u(2,t) = 0, \qquad 0 < t < 0.1$$
$$u(x,0) = 2\sin \pi x, \qquad 0 \le x \le 2$$

Take $h = 0.4,\ and\ \ k = 0.05$.

5. $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1,\ \ 0 < t < 0.5$$
$$u(0,t) = u(2,t) = 0, \qquad 0 < t < 0.5$$
$$u(x,0) = \sin \pi x, \qquad 0 \le x \le 1$$

Take $h = 0.2,\ and\ \ k = 0.01$.

In problems **6-10**, use the forward difference method to approximate the solution to the following parabolic partial differential equations

6. $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2, \;\; t > 0$$
$$u(0,t) = u(2,t) = 0, \;\; t > 0$$
$$u(x,0) = \sin 2\pi x, \qquad 0 \le x \le 2$$

Use $h = 0.2$ and $k = 0.01$, and compare your answers at $t = 0.5$ to the actual solution. Then use $h = 0.1$ and $k = 0.005$ and compare the answers.

7. $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2, \;\; t > 0$$
$$u(0,t) = u(2,t) = 0, \;\; t > 0$$
$$u(x,0) = x(2-x), \qquad 0 \le x \le 2$$

Take $h = 0.2$ and choose an appropriate value for k.

8. $$\frac{\partial u}{\partial t} = \frac{4}{\pi^2}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 4, \;\; t > 0$$
$$u(0,t) = u(4,t) = 0, \;\; t > 0$$
$$u(x,0) = \sin\left(\frac{\pi}{4}x\right)\left(1 + 2\cos\left(\frac{\pi}{4}x\right)\right), \quad 0 \le x \le 4$$

Use $h = 0.2$ and $k = 0.04$. Compare your answers to the actual solution $u(x,t) = e^{-t}\sin\left(\frac{\pi}{4}x\right) + e^{-t/4}\sin\left(\frac{\pi}{4}x\right)$ at $t = 0.4$.

9. $$\frac{\partial u}{\partial t} = \frac{1}{\pi^2}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \;\; t > 0$$
$$u(0,t) = u(1,t) = 0, \;\; t > 0$$
$$u(x,0) = \cos\pi(x - 1/2), \quad 0 \le x \le 1$$

Use $h = 0.1$ and $k = 0.04$. Compare your answers to the actual solution $u(x,t) = e^{-t}\cos\pi(x - 1/2)$ at $t = 0.4$.

10. $$\frac{\partial u}{\partial t} = \frac{1}{\pi^2}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 2, \;\; 0 < t < 1$$
$$u(0,t) = u(2,t) = 0, \;\; 0 \le t \le 1$$

$$u(x,0) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & 1 \le x \ge 2 \end{cases}$$

Take $h = 0.25$ and $k = 0.025$.

5.4 The finite difference method for hyperbolic partial differential equations

In this section we consider the numerical solution to the wave equation, an example of a hyperbolic partial differential equation. The wave equation is given by the differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < \ell, \;\; t > 0, \tag{1}$$

subject to the conditions

$$u(0,t) = u(\ell,t) = 0, \qquad t > 0 \tag{2}$$

$$u(x,0) = u_0(x), \qquad 0 \le x \le \ell, \tag{3}$$

$$\frac{\partial u}{\partial t}(x,0) = g(x), \qquad 0 \le x \le \ell, \tag{4}$$

To set the finite difference method, select the mesh constants h and k, so that we have the mesh points $(x_i,\ t_j)$ which are defined by $x_i = ih, \quad i = 0,1,\cdots,m$ and $t_j = jk, \quad j = 0,1,\cdots$.

Using the shown scheme labeling, to replace the second-order partial derivatives by the central difference quotients

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{ij} \cong \frac{u_{i-1,j} + u_{i+1,j} - 2u_{i,j}}{h^2}, \tag{5}$$

$$\left(\frac{\partial^2 u}{\partial t^2}\right)_{ij} \cong \frac{u_{i,j-1} + u_{i,j+1} - 2u_{i,j}}{k^2}, \tag{6}$$

Thus we replace the differential equation (1) by

$$\frac{u_{i,j-1}+u_{i,j+1}-2u_{i,j}}{k^2}=c^2\frac{u_{i-1,j}+u_{i+1,j}-2u_{i,j}}{h^2} \qquad (7)$$

If λ is used to denote ck/h, we can write the difference equation as

$$u_{i,j-1}+u_{i,j+1}-2u_{i,j}=\lambda^2\left(u_{i-1,j}+u_{i+1,j}-2u_{i,j}\right), \qquad (8)$$

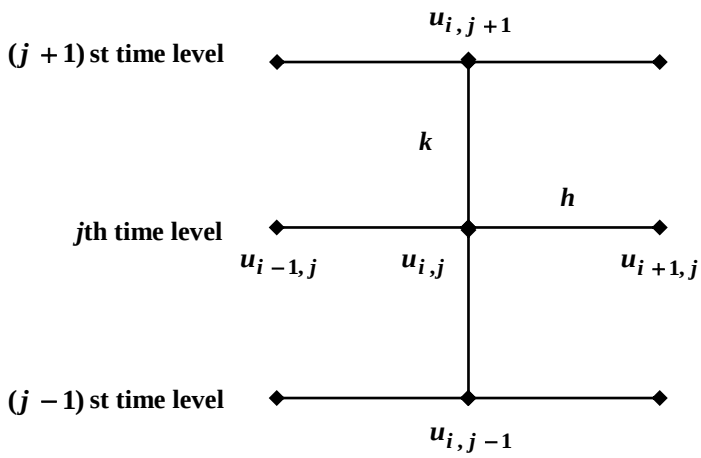


Fig. 5.22

and solve for $u_{i,j+1}$, the most advanced time step approximation, to obtain

$$u_{i,j+1}=\lambda^2\left(u_{i-1,j}+u_{i+1,j}\right)+2\left(1-\lambda^2\right)u_{i,j}-u_{i,j-1}, \qquad (9)$$

This equation holds for each $i=1,\cdots,m-1$ and $j=1,2,\cdots$. The boundary conditions give

$$u_{0,j}=u_{m,j}=0 \qquad (10)$$

for each $j=1,2,\cdots$, and the initial condition implies that

$$u_{i,0}=u_0(x_i), \qquad i=1,2,\cdots,m-1 \qquad (11)$$

Equations (9) and (10) imply that the $(j+1)st$ time level requires values from the jth and $(j-1)st$ time steps as shown in Fig. 5.16. This produces a minor starting problem since values for $j=0$ are given by equation (11), but values for $j=1$, which are needed in equation (9) to compute $u_{i,2}$, must be obtained from the initial-velocity condition (4). It follows that

$$\left(\frac{\partial u}{\partial t}\right)_{i,0} \cong \frac{u_{i,1}-u_{i,-1}}{2k} = g(x_i), \qquad (12)$$

Solving for $u_{i,-1}$ gives

$$u_{i,-1} = u_{i,1} - 2kg(x_i), \qquad (13)$$

By substituting (13) into (9) when $j=0$, we get the special case

$$u_{i,1} = \lambda^2\left(u_{i-1,0}+u_{i+1,0}\right) + 2\left(1-\lambda^2\right)u_{i,0} - u_{i,-1}$$

$$= \lambda^2\left(u_{i-1,0}+u_{i+1,0}\right) + 2\left(1-\lambda^2\right)u_{i,0} - u_{i,1} + 2kg(x_i)$$

As a consequence

$$u_{i,1} = \frac{\lambda^2}{2}\left(u_{i-1,0}+u_{i+1,0}\right) + \left(1-\lambda^2\right)u_{i,0} + kg(x_i) \qquad (14)$$

As in the case of the forward-difference method for the heat equation, the explicit finite difference method for the wave equation has stability problem. In fact, it is necessary that $\lambda \le 1$ for the method to be stable.

Example1 Approximate the solution of the initial-boundary value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \qquad 0<x<1,\ \ 0<t<1,$$

$$u(0,t) = u(1,t) = 0, \qquad 0 < t < 1,$$
$$u(x,0) = \sin 2\pi x\,, \qquad 0 \le x \le 1,$$
$$\frac{\partial u}{\partial t}(x,0) = 2\pi \sin 2\pi x\,, \qquad 0 \le x \le 1$$

Use $h = 0.2$ and $k = 0.1$. Compare your answers to the actual solution $u(x,t) = \sin 2\pi x\,(\cos 2\pi t + \sin 2\pi t)$ at $t = 0.2$.

Solution We make the identifications $h = 0.2$, $k = 0.1$ and $c = 1$, so

$$\lambda = ck/h = \frac{0.1 \times 1}{0.2} = 0.5$$

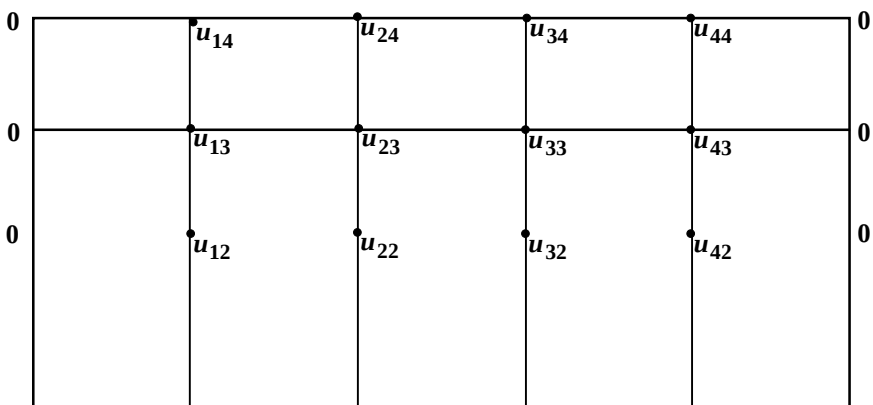
Thus, with $g(x) = 2\pi \sin 2\pi x$, equations (14) and (9) become, respectively,

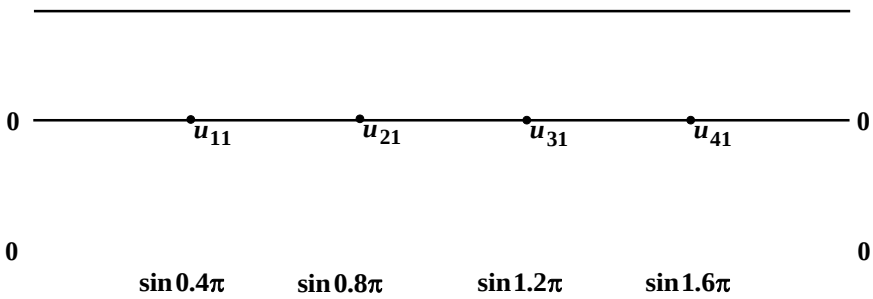
$$u_{i,1} = 0.125\left(u_{i-1,0} + u_{i+1,0}\right) + 0.75u_{i,0} + 0.2\pi \sin 2\pi x_i \quad (15)$$

$$u_{i,j+1} = 0.25\left(u_{i-1,j} + u_{i+1,j}\right) + 1.5u_{i,j} - u_{i,j-1}, \quad (16)$$

For $i = 1,2,3,4$, equation (15) yields the following values for the $u_{i,1}$ on the first time level:

$$\begin{aligned}
u_{1,1} &= 0.125\left(u_{0,0} + u_{1,0}\right) + 0.75u_{1,0} + 0.2\pi \sin 0.4\pi = 1.384332\\
u_{2,1} &= 0.125\left(u_{1,0} + u_{3,0}\right) + 0.75u_{2,0} + 0.2\pi \sin 0.8\pi = 0.855564\\
u_{3,1} &= 0.125\left(u_{2,0} + u_{4,0}\right) + 0.75u_{3,0} + 0.2\pi \sin 1.2\pi = -0.855564\\
u_{4,1} &= 0.125\left(u_{3,0} + u_{5,0}\right) + 0.75u_{4,0} + 0.2\pi \sin 1.6\pi = -1.384332
\end{aligned} \quad (17)$$





Note that the results given in (17) were obtained from the initial condition $u(x,0) = \sin 2\pi x$. For example, $u_{2,0} = \sin 0.4\pi$, and so on. Now $j = 1$ in (16) gives

$$u_{i,2} = 0.25\left(u_{i-1,1} + u_{i+1,1}\right) + 1.5u_{i,1} - u_{i,0},$$

and so for $i = 1,2,3,4$, we get, by the use of the boundary conditions,

$$\begin{aligned}
u_{12} &= 0.25(u_{01} + u_{21}) + 1.5u_{11} - u_{10} \\
&= 0.25(0 + 0.855564) + 1.5(1.384332) - \sin 0.4\pi \\
&= 1.339333, \\
u_{22} &= 0.25(u_{11} + u_{31}) + 1.5u_{21} - u_{20} \\
&= 0.25(1.384332 - 0.855564) + 1.5(0.855564) - \sin 0.8\pi \\
&= 0.827753, \\
u_{32} &= 0.25(u_{21} + u_{41}) + 1.5u_{31} - u_{30} \\
&= 0.25(0.855564 - 1.384332) + 1.5(-0.855564) - \sin 1.2\pi \\
&= -0.827753, \\
u_{42} &= 0.25(u_{31} + u_{51}) + 1.5u_{41} - u_{40} \\
&= 0.25(-0.855564 + 0) + 1.5(-1.384332) - \sin 1.6\pi \\
&= -1.339333
\end{aligned}$$

Computations for higher levels are carried out by the same manner. Anyway, the following table records the approximate and actual solutions at $t = 0.2$.

at $t = 0.2$			
x	$Approximate\ solution$	$Actual\ solution$	$Absolute\ Error$
0	0	0	0
0.2	1.339333	1.328438	0.010895
0.4	–0.827753	0.82102	0.006733
0.6	–0.827753	–0.82102	0.006733
0.8	1.339333	–1.328438	0.010895
0	0	0	0

Example2 Solve numerically the wave equation

$$\frac{\partial^2 u}{\partial t^2} = 4\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1,\ \ 0 < t < 1,$$

with the boundary and initial conditions

$$u(0,t) = u(1,t) = 0, \qquad 0 < t < 1,$$

$$u(x,0) = \sin \pi x,\ \ \frac{\partial u}{\partial t}(x,0) = 0,\ \ 0 \le x \le 1,$$

by using $h = 0.2$ and $k = 0.05$.

Solution Here we have $h = 0.2$, $k = 0.05$ and $c = 2$, so

$$\lambda = ck/h = \frac{2(0.05)}{0.2} = 0.5$$

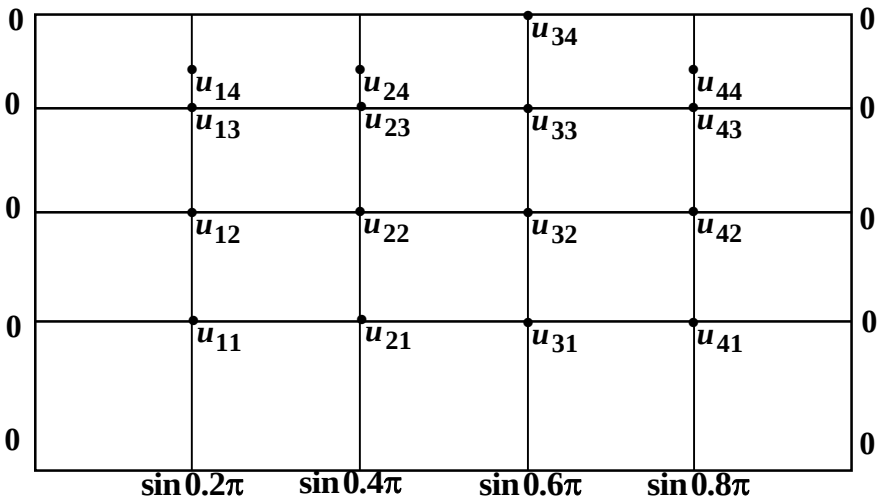
Thus, with $g(x) = 0$, equation (14) and (9) become, respectively,

$$u_{i,1} = 0.125\left(u_{i-1,0} + u_{i+1,0}\right) + 0.75u_{i,0}, \tag{18}$$

$$u_{i,j+1} = 0.25\left(u_{i-1,j} + u_{i+1,j}\right) + 1.5u_{i,j} - u_{i,j-1}, \qquad (19)$$

for $i = 1,2,3,4$, equation (18) yields the following values for the $u_{i,1}$ on the first time level:

$$\begin{aligned} u_{11} &= 0.125(u_{00} + u_{10}) + 0.75u_{10} = 0.559721 \\ u_{21} &= 0.125(u_{10} + u_{30}) + 0.75u_{20} = 0.90564761 \\ u_{31} &= 0.125(u_{20} + u_{40}) + 0.75u_{30} = 0.90564761 \\ u_{41} &= 0.125(u_{30} + u_{50}) + 0.75u_{40} = 0.559721 \end{aligned} \qquad (20)$$



Note that the results given in (20) were obtained from the initial condition $u(x,0) = \sin \pi x$. Now $j = 1$ in (19) gives

$$u_{i,2} = 0.25\left(u_{i-1,1} + u_{i+1,1}\right) + 1.5u_{i,1} - u_{i,0},$$

and so for $i = 1,2,3,4$, we get, by the use of the boundary conditions,

$$\begin{aligned} u_{12} &= 0.25(u_{01} + u_{21}) + 1.5u_{11} - u_{10} \\ &= 0.25(0 + 0.9056476) + 1.5(0.55972) - \sin 0.2\pi \\ &= 4782, \end{aligned}$$

$$\begin{aligned} u_{22} &= 0.25(u_{11} + u_{31}) + 1.5u_{11} - u_{20} \\ &= 0.25(0.9056476 + 0.559721) + 1.5(0.559721) - \sin 0.4\pi \\ &= 7738, \end{aligned}$$

$$u_{32} = 0.25(u_{21} + u_{41}) + 1.5u_{31} - u_{30}$$

$$= 0.25(0.559721 + 0.9056476) + 1.5(0.9056476) - \sin 0.6\pi$$
$$= 7738,$$
$$u_{42} = 0.25(u_{31} + u_{51}) + 1.5u_{41} - u_{40}$$
$$= 0.25(0 + 0.90564761) + 1.5(0.559721) - \sin 0.8\pi$$
$$= 4782,$$

Computations for higher levels are carried out by the same manner.

Exercise 5.3

In problems 1–5, approximate the solution to the following hyperbolic partial differential equations

1.
$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \ \ t > 0,$$
$$u(0,t) = u(1,t) = 0, \qquad t > 0,$$
$$u(x,0) = \sin 2\pi x, \qquad 0 \le x \le 1,$$
$$\frac{\partial u}{\partial t}(x,0) = 0, \qquad 0 \le x \le 1$$

Use $h = 0.1$ and $k = 0.05$, and then $h = 0.05$ $k = 0.05$. Compare your results to the exact solution at $t = 0.5$.

2.
$$\frac{\partial^2 u}{\partial t^2} = \frac{1}{4}\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 0.5, \ \ 0 < t < 0.5,$$
$$u(0,t) = u(0.5,t) = 0, \qquad 0 < t < 0.5,$$
$$u(x,0) = 0, \qquad 0 \le x \le 0.5,$$
$$\frac{\partial u}{\partial t}(x,0) = \sin 4\pi x, \qquad 0 \le x \le 0.5$$

Use $h = 0.25$ and $k = 0.25$.

3.
$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \ \ 0 < t < 0.5,$$
$$u(0,t) = u(1,t) = 0, \qquad 0 < t < 0.5,$$

$$u(x,0) = \sin 2\pi x\,, \qquad 0 \le x \le 1,$$

$$\frac{\partial u}{\partial t}(x,0) = \sin \pi x\,, \qquad 0 \le x \le 1$$

Use $h = 0.25$ and $k = 0.25$.

4.

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \;\; t > 0,$$

$$u(0,t) = u(1,t) = 0, \qquad t > 0,$$

$$u(x,0) = \sin 2\pi x\,, \qquad 0 \le x \le 1,$$

$$\frac{\partial u}{\partial t}(x,0) = 2\pi \sin 2\pi x\,, \qquad 0 \le x \le 1$$

Use $h = 0.1$ and $k = 0.1$. Compare your results to the exact solution $u(x,t) = \sin 2\pi x\,(\cos 2\pi t + \sin 2\pi t)$, at $t = 0.3$.

5.

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \;\; t > 0,$$

$$u(0,t) = u(1,t) = 0, \qquad t > 0,$$

$$u(x,0) = \begin{cases} 1, & 0 \le x \le \dfrac{1}{2} \\ -1, & \dfrac{1}{2} \le x \le 1 \end{cases},$$

$$\frac{\partial u}{\partial t}(x,0) = 0, \qquad 0 \le x \le 1$$

Use $h = 0.1$ and $k = 0.1$. Find u at $t = 0.5$.

6. Curve Fitting

One of the most important problems that appears in engineering applications is to fit functions to given data and find the "best" (in some sense) function that can be used to represent these data. That is to say, it is required to create the best fit model from the relationship between two variables (independent and dependent variables).

Given a table of readings y_r corresponds to x_r, $r = 1,2,\cdots,n$; namely

x	x_1	x_2	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n
y	y_1	y_2	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

The question is how can we find the equation of a curve to ***fit*** these readings? In some cases we know from physical considerations the general shape of a curve, e.g. a straight line, exponential or curve, logarithmic or a parabola, etc. In other cases we plot x_r and y_r and from the shape of the plot we can decide what curve fits the points best. Hence, the function $y(x)$ describes the relationship between two physical quantities x and $y = y(x)$ and the problem of ***data fitting*** is to recover $y(x)$ from the given approximate data y_r, $r = 1,2,\cdots,n$. This means that the function $y(x)$ contains or represents most (if not all) the information contained in the data with little error.

The present chapter is devoted to consider the problem of approximating a function whose values at a sequence of points are generally obtained through measurements or experiments. This approximate function can then, for example, be used to approximate the derivative and integral of the function.

Let $y(x)$ be a function and $x_r,\ r = 1,2,\cdots,n$ be a sequence of data points at which we have observed values of $y(x)$ which generally will be in error. We denote $y(x_r)$, the true value at x_r, and we denote the observed value at x_r by y_r. So we define the deviations at the data points by (the 'error' between data and true values) (see Fig. 1.1)

$$d_1 = y(x_1) - y_1$$
$$d_2 = y(x_2) - y_2$$
$$d_3 = y(x_3) - y_3$$
$$\vdots$$

or, in general

$$d_r = y(x_r) - y_r, \quad r = 1,2,\cdots,n \tag{1}$$

$\boldsymbol{x_r}$	$\boldsymbol{y_r}$	$\boldsymbol{y(x_r)}$	$\boldsymbol{d_r = y(x_r) - y_r}$
$\boldsymbol{x_1}$	y_1	$y(x_1)$	$d_1 = y(x_1) - y_1$
$\boldsymbol{x_2}$	y_2	$y(x_2)$	$d_2 = y(x_2) - y_2$
$\boldsymbol{x_3}$	y_3	$y(x_3)$	$d_3 = y(x_3) - y_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 6.1 Deviations at the data points

Throughout this chapter we shall assume that the errors at different data points are independent. Thus our problem is to minimize the deviations $d_r,\ r = 1,2,\cdots,n$ in some sense. In fact, there are several methods for minimizing the deviation, but we will only discuss and study here the principle of least square.

6.1 Principle of least squares

In this section we concerns with the ***least square technique*** at which we minimize the sum of the squares of deviations d_r. If we have reason to believe that $y(x)$ is a straight line the given data are plotted in Fig. 6.1.

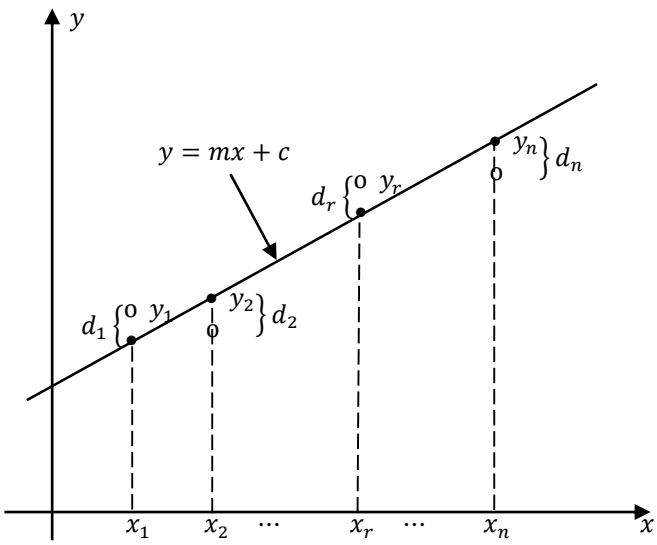


Fig. 6.1. The deviations at the data points

Choosing the equation of the straight line to be

$$y(x) = mx + c, \tag{2}$$

where m and c are two arbitrary constants. The general idea is to choose the constants m and c to minimize the error given by equation (1). It is clear that the minimization of a function of two variables needs to set its partial derivatives with respect to these variables to zero and simultaneously solve the resulting equations.

The least squares approach to this problem involves of minimizing the best approximating line when the error involved is the sum of the

squares of the differences between the values on the approximating line and the given values ($L_2 - norm$);

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2}, \tag{3}$$

or equivalently we minimize

$$e(m,c) = \|d\|^2 = \sum_{r=1}^{n} {d_r}^2 = \sum_{r=1}^{n} (y(x_r) - y_r)^2$$
$$= \sum_{r=1}^{n} (mx_r + c - y_r)^2 \tag{4}$$

If $\|d\|^2$ is considered to be a function of two variables $e(m,c)$, an elementary result from calculus of several variables implies, for a minimum to occur at (m,c), it is necessary for the minimization of $e(m,c)$ that

$$\frac{\partial}{\partial m} e(m,c) = 0, \tag{5}$$

and

$$\frac{\partial}{\partial c} e(m,c) = 0, \tag{6}$$

which simplifies to the linear system of equations

$$\sum_{r=1}^{n} 2(mx_r + c - y_r)(x_r) = 0, \tag{7}$$

$$\sum_{r=1}^{n} 2(mx_r + c - y_r)(1) = 0, \tag{8}$$

Equations (7) and (8) simplify to what is known as the ***normal equations*** for a table of n data points:

$$\boxed{\begin{aligned} m\sum_{r=1}^{n} x_r^2 \;+\; c\sum_{r=1}^{n} x_r &= \sum_{r=1}^{n} x_r y_r, \\ m\sum_{r=1}^{n} x_r \;+\; nc \quad &= \sum_{r=1}^{n} y_r \end{aligned}} \tag{9}$$

The coefficients of the unknowns in these equations are constructed with the aid of the given data as follows:

	x_r	y_r	x_r^2	$x_r y_r$
	x_1	y_1	x_1^2	$x_1 y_1$
	x_2	y_2	x_2^2	$x_2 y_2$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_{n-1}	y_{n-1}	x_{n-1}^2	$x_{n-1} y_{n-1}$
	x_n	y_n	x_n^2	$x_n y_n$
$\sum$	$\sum_{r=1}^{n} x_r,$	$\sum_{r=1}^{n} y_r,$	$\sum_{r=1}^{n} x_r^2,$	$\sum_{r=1}^{n} x_r y_r,$

Table 6.2. Coefficients of the normal equations

The solution to the normal equations is given by the use of Cramer method in the form

$$m = \frac{\begin{vmatrix} \sum_{r=1}^{n} x_r y_r & \sum_{r=1}^{n} x_r \\ \sum_{r=1}^{n} y_r & n \end{vmatrix}}{\begin{vmatrix} \sum_{r=1}^{n} x_r^2 & \sum_{r=1}^{n} x_r \\ \sum_{r=1}^{n} x_r & n \end{vmatrix}},$$

$$c = \frac{\begin{vmatrix} \sum_{r=1}^{n} x_r^2 & \sum_{r=1}^{n} x_r y_r \\ \sum_{r=1}^{n} x_r & \sum_{r=1}^{n} y_r \end{vmatrix}}{\begin{vmatrix} \sum_{r=1}^{n} x_r^2 & \sum_{r=1}^{n} x_r \\ \sum_{r=1}^{n} x_r & n \end{vmatrix}}$$

Remark 1 Recently the programmable calculators are used to calculate the unknowns m and c directly. In fact, these calculators are used for the fitting process itself. Just we insert the input data the values of the unknowns m and c are approximated and the equation of the straight line and its graph are obtained.

Remark 2 As a result of the above analysis, we see that the normal equations corresponding to the equation of the straight line $y(x) = mx + c$ are obtained by multiplying the equation of the line by x and take the summation of both sides and again multiply the equation of the line by 1 and take the summation of both sides we get directly equations (9).

Example 1.1. Fit the following readings to the equation of the straight line $y(x) = mx + c$.

x	1	2	3	4	5
y	2.1	3.9	5.1	6.7	8.3

Solution. The required line is $y = mx + c$, so that the normal equations will be

$$m\sum_{r=1}^{n} x_r^2 + c\sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r y_r,$$
$$m\sum_{r=1}^{n} x_r + nc = \sum_{r=1}^{n} y_r$$

To compute the components of these equations, extend the following table as shown in the third and fourth column, and sum the columns

	x_r	y_r	x_r^2	$x_r y_r$
	1	2.1	1	2.1
	2	3.9	4	7.8
	3	5.1	9	15.3
	4	6.7	16	26.8
	5	8.3	25	41.5
$\sum$	15	26.1	55	93.5

Table 6.3. The coefficients of the normal equations in Example 1.1

Substitute in the norm al equations with $n = 5$, we get

$$55m + 15c = 93.5$$
$$15m + 5c = 26.1$$

The solution of these equations is

$$m = \frac{\begin{vmatrix} 93.5 & 15 \\ 26.1 & 5 \end{vmatrix}}{\begin{vmatrix} 55 & 15 \\ 15 & 5 \end{vmatrix}} = \frac{(5)(93.5) - (15)(26.1)}{(5)(55) - (15)^2} = 1.52,$$

$$c = \frac{\begin{vmatrix} 55 & 93.5 \\ 15 & 26.1 \end{vmatrix}}{\begin{vmatrix} 55 & 15 \\ 15 & 5 \end{vmatrix}} = \frac{(55)(26.1) - (15)(93.5)}{(5)(55) - (15)^2} = 0.66$$

Then, the equation of the approximating line is

$$\boxed{y(x) = 1.52x + 0.66,}$$

and the graph of this line with the data points is given in Fig. 6.2.

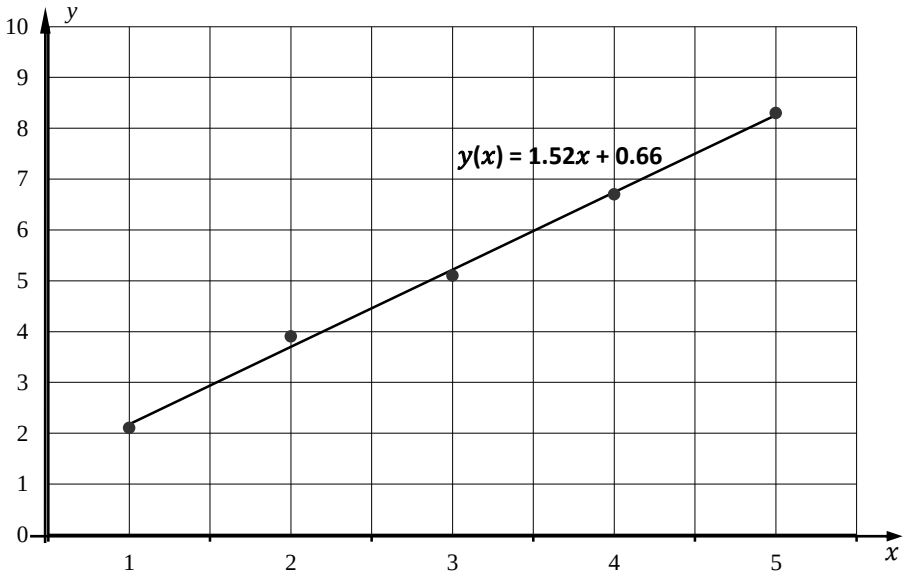


Fig. 6.2. The graph of the fitting curve with the data points of Example 1.1

Moreover, the approximate values given by the least squares technique and the data points are given in table 6.4, from which, the $L_2 - norm$ of the error is

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.064} \cong 0.25$$

x_r	y_r	$y(x_r) = 1.52x_r + 0.66$	$d_r^2 = (y(x_r) - y_r)^2$
1	2.1	2.18	0.0064
2	3.9	3.70	0.04
3	5.1	5.22	0.0144
4	6.7	6.74	0.0016
5	8.3	8.26	0.0016
			0.064

Table 6.4. Deviations of the best approximating line in Example 1.1

Example 1.2. Fit the curve $y = ax + b$ for the following readings:

x	3	3.25	3.5	3.75	4
y	8.3	8.7	9.2	9.6	10.1

Solution. The normal equations corresponding to the line $y = ax + b$ are

$$a\sum_{r=1}^{n} x_r^2 + b\sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r y_r,$$
$$a\sum_{r=1}^{n} x_r + nb = \sum_{r=1}^{n} y_r$$

The given readings imply that

	x_r	y_r	x_r^2	$x_r y_r$
	3	8.3	9	24.9
	3.25	8.7	10.5625	28.275
	3.5	9.2	12.25	32.2
	3.75	9.6	14.0625	36
	4	10.1	16	40.4
$\sum$	17.5	45.9	61.875	161.775

Table 6.5. The coefficients of the normal equations in Example 1.2

Substitute in the norm al equations with $n = 5$, we get

$$61.875\,a + 17.5b = 161.775$$
$$17.5\,a + 5b = 45.9$$

The solution of these equations is

$$a = \frac{\begin{vmatrix} 161.775 & 17.5 \\ 45.9 & 5 \end{vmatrix}}{\begin{vmatrix} 61.875 & 17.5 \\ 17.5 & 5 \end{vmatrix}} = 1.8,$$

$$b = \frac{\begin{vmatrix} 61.875 & 161.775 \\ 17.5 & 45.9 \end{vmatrix}}{\begin{vmatrix} 61.875 & 17.5 \\ 17.5 & 5 \end{vmatrix}} = 2.88$$

Therefore, the equation of the required approximating straight line takes the form

$$\boxed{y(x) = 1.8x + 2.88,}$$

and

x_r	y_r	$y(x_r) = 1.8x_r + 2.82$	$d_r^2 = (y(x_r) - y_r)^2$
3	8.3	8.28	0.0004
3.25	8.7	8.73	0.0009
3.5	9.2	9.18	0.0004
3.75	9.6	9.63	0.0009
4	10.1	10.08	0.0004
			0.003

Table 6.6. Deviations of the best approximating line in Example 1.2

Therefore, the least squares of the errors is

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.003} \cong 0.054772256$$

6.2 The least square method for polynomials of degree *m*

The general problem of approximating a set of data

x	x_1	x_2	$\cdots$	x_r	$\cdots$	x_{n-1}	x_n
y	y_1	y_2	$\cdots$	y_r	$\cdots$	y_{n-1}	y_n

with a polynomial

$$P_m(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0 \tag{10}$$

of degree $m < n$ using the least squares procedure is handled in a similar manner and requires choosing the constants

$$a_m, a_{m-1}, \cdots, a_1, a_0$$

to minimize the least square error

$$e(a_0, a_1, \cdots, a_{m-1}, a_m) = \|d\|^2 = \sum_{r=1}^{n} (P_m(x_r) - y_r)^2$$

$$= \sum_{r=1}^{n} (a_m x_r^m + a_{m-1}x_r^{m-1} + \cdots + a_1 x_r + a_0 - y_r)^2 \tag{11}$$

As in the linear case, for *e* to be minimized it is necessary that

$$\frac{\partial}{\partial a_i} e(a_0, a_1, \cdots, a_{m-1}, a_m) = 0, \qquad i = 0,1,\cdots,m \tag{12}$$

Thus, for each *i*

$$\frac{\partial e}{\partial a_m} = 0 \Rightarrow \sum_{r=1}^{n} 2(a_m x_r^m + a_{m-1}x_r^{m-1} + \cdots + a_1 x_r + a_0 - y_r)(x_r^m) = 0$$

$$\frac{\partial e}{\partial a_{m-1}} = 0 \Rightarrow \sum_{r=1}^{n} 2(a_m x_r^m + a_{m-1}x_r^{m-1} + \cdots + a_1 x_r + a_0 - y_r)(x_r^{m-1}) = 0$$

$$\vdots$$

$$\frac{\partial e}{\partial a_1} = 0 \Rightarrow \sum_{r=1}^{n} 2(a_m x_r^m + a_{m-1}x_r^{m-1} + \cdots + a_1 x_r + a_0 - y_r)(x_r) = 0$$

$$\frac{\partial e}{\partial a_0} = 0 \Rightarrow \sum_{r=1}^{n} 2(a_m x_r^m + a_{m-1}x_r^{m-1} + \cdots + a_1 x_r + a_0 - y_r)(1) = 0$$

It is helpful to write the normal equations as follows

$$a_m \sum_{r=1}^{n} x_r^{2m} + a_{m-1} \sum_{r=1}^{n} x_r^{2m-1} + \cdots + a_1 \sum_{r=1}^{n} x_r^{m+1} + a_0 \sum_{r=1}^{n} x_r^m = \sum_{r=1}^{n} x_r^m y_r, \tag{13}$$

$$a_m \sum_{r=1}^{n} x_r^{2m-1} + a_{m-1} \sum_{r=1}^{n} x_r^{2m-2} + \cdots + a_1 \sum_{r=1}^{n} x_r^m + a_0 \sum_{r=1}^{n} x_r^{m-1} = \sum_{r=1}^{n} x_r^{m-1} y_r, \tag{14}$$

$$\vdots$$

$$a_m \sum_{r=1}^{n} x_r^{m+1} + a_{m-1} \sum_{r=1}^{n} x_r^{m} + \cdots + a_1 \sum_{r=1}^{n} x_r^{2} + a_0 \sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r\, y_r, \quad (15)$$

$$a_m \sum_{r=1}^{n} x_r^{m} + a_{m-1} \sum_{r=1}^{n} x_r^{m-1} + \cdots + a_1 \sum_{r=1}^{n} x_r^{m} + a_0 n = \sum_{r=1}^{n} y_r \quad (16)$$

It can be shown that the normal equations always have a unique solution provided that the x_r for $r = 1, \cdots, n$ are distinct.

Remark 3 The normal equations corresponding to the polynomial

$$P_m(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0$$

are obtained by multiplying the equation of the polynomial by

$$x^m, x^{m-1}, \cdots, x, 1,$$

respectively, and take the summation of both sides in each case to get equations (13)-(16).

Example 1.3. Fit the data in the following table with the least squares polynomial of degree two.

x	1	1.5	2	2.5	3
y	1.41	1.96	2.88	4.16	5.8

Solution. The required polynomial may be written in the form

$$y(x) = P_2(x) = ax^2 + bx + c$$

from which we can write the normal equations as

$$a\sum_{r=1}^{n} x_r^4 + b\sum_{r=1}^{n} x_r^3 + c\sum_{r=1}^{n} x_r^2 = \sum_{r=1}^{n} x_r^2 y_r$$
$$a\sum_{r=1}^{n} x_r^3 + b\sum_{r=1}^{n} x_r^2 + c\sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r y_r$$
$$a\sum_{r=1}^{n} x_r^2 + b\sum_{r=1}^{n} x_r + cn = \sum_{r=1}^{n} y_r$$

From the given data we can construct the table

	x_r	y_r	x_r^2	x_r^3	x_r^4	$x_r y_r$	$x_r^2 y_r$
	1	1.41	1	1	1	1.41	1.41
	1.5	1.96	2.25	3.375	5.0625	2.94	4.41
	2	2.88	4	8	16	5.76	11.52
	2.5	4.16	6.25	15.625	39.0625	10.4	26
	3	5.8	9	27	81	17.4	52.2
$\sum$	10	16.21	22.5	55	142.125	37.91	95.54

Table 1.7. Coefficients of the normal equations in Example 1.3

Substitute in the norm al equations with $n = 5$, we get

$$142.125a + 55b + 22.5c = 95.54$$
$$55a + 22.5b + 10c = 37.91$$
$$22.5a + 10b + 5c = 16.21$$

Solving this system by the use of Crammer method we get

$$a = \frac{\begin{vmatrix} 95.54 & 55 & 22.5 \\ 37.91 & 22.5 & 10 \\ 16.21 & 10 & 5 \end{vmatrix}}{\begin{vmatrix} 142.125 & 55 & 22.5 \\ 55 & 22.5 & 10 \\ 22.5 & 10 & 5 \end{vmatrix}} = \frac{7.9375}{10.9375} = 0.7257,$$

$$b = \frac{\begin{vmatrix} 142.125 & 95.54 & 22.5 \\ 55 & 37.91 & 10 \\ 22.5 & 16.21 & 5 \end{vmatrix}}{\begin{vmatrix} 142.125 & 55 & 22.5 \\ 55 & 22.5 & 10 \\ 22.5 & 10 & 5 \end{vmatrix}} = \frac{-7.73125}{10.9375} = -0.7069,$$

$$c = \frac{\begin{vmatrix} 142.125 & 55 & 95.54 \\ 55 & 22.5 & 37.91 \\ 22.5 & 10 & 16.21 \end{vmatrix}}{\begin{vmatrix} 142.125 & 55 & 22.5 \\ 55 & 22.5 & 10 \\ 22.5 & 10 & 5 \end{vmatrix}} = \frac{15.20312}{10.9375} = 1.39$$

Thus, the least squares polynomial of degree two fitting the preceding data is

$$\boxed{y(x) = P_2(x) = 0.7257x^2 - 0.7069x + 1.39}$$

At the given values of x_r, we have the approximations shown in the following table

x_r	y_r	$y(x_r) = 0.7257x_r^2 - 0.7069x + 1.39$	$d_r^2 = (y(x_r) - y_r)^2$
1	1.41	1.408843	1.33898E-06
1.5	1.96	1.962539	6.44797E-06
2	2.88	2.879086	8.35918E-07
2.5	4.16	4.158482	2.30389E-06
3	5.8	5.800729	5.30816E-07
			1.146E-05

Table 1.8. Deviations of the best approximating curve in Example 1.3

The error

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{1.146\text{E}-05} \cong 0.0034,$$

is the least that can be obtained by using a quadratic polynomial.

6.3 The least square method for general curves

Although the least square method using polynomials is the most extensively used procedure, occasionally it is appropriate to assume that the data is exponentially related. This requires the approximating function to be of the form

$$y = ce^{mx}, \tag{1}$$

for some constants c and m.

In this case, some transformation is considered to obtain an equation of straight line in the new axis. Namely, taking the natural logarithm of the approximating equation we get

$$\ln y = \ln(ce^{mx}) = \ln c + \ln(e^{mx}),$$

or

$$\ln y = \ln c + mx, \tag{2}$$

which may be written in the form

$$\boxed{Y = mX + C} \qquad \begin{array}{ll} X = x, & Y = \ln y \\ C = \ln c & \Rightarrow \quad c = e^{C} \end{array} \tag{3}$$

Here we use the substitutions

$$Y = \ln y, \qquad C = \ln c, \tag{4}$$

to obtain an equation of straight line in the new variables X and Y.

To fit the given data to the straight line (3), the normal equations shall be of the form

$$m\sum_{r=1}^{n} x_r^2 + C\sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r Y_r \tag{5}$$

$$m\sum_{r=1}^{n} x_r \;+\; nC \qquad = \sum_{r=1}^{n} Y_r$$

Substituting from the given readings, the constants m and C are determined from the system of equations (5). To get the constant c we use the transformation (4); namely

$$c = e^{C}$$

A similar treatment is done if it is required to fit some readings to the exponential curve

$$y = bx^{a}, \tag{6}$$

for some constants a and b. Taking the logarithm of both sides,

$$\ln y = \ln b + \ln x^{a} = \ln b + a \ln x, \tag{7}$$

Using the transformations

$$Y = \ln y, \qquad X = \ln x, \qquad B = \ln b, \tag{8}$$

we can write Equation (7) as

$$\boxed{Y = aX + B} \qquad \begin{aligned} &X = \ln x, \;\; Y = \ln y, \\ &B = \ln b \quad \Rightarrow \quad b = e^{B} \end{aligned} \tag{9}$$

So that the normal equations shall be

$$\begin{aligned} a\sum_{r=1}^{n} X_r^2 \;+\; B\sum_{r=1}^{n} X_r &= \sum_{r=1}^{n} X_r Y_r \\ a\sum_{r=1}^{n} X_r \;+\; nB \qquad &= \sum_{r=1}^{n} Y_r \end{aligned} \tag{10}$$

Solving these equations and using the transformation (8) we obtain the constants a and b.

Example 1.4. Fit the following readings for the curve $y(x) = be^{ax}$.

x	1	1.25	1.5	1.75	2	2.25
y	1.12	0.92	0.75	0.61	0.51	0.42

Solution. Since $y = be^{ax}$, then

$$\ln y = ax + \ln b,$$

or

$$\boxed{Y = AX + B,} \qquad \begin{aligned} &X = x, \;\; Y = \ln y \\ &A = a, \;\; B = \ln b \quad \Rightarrow \quad b = e^{B} \end{aligned}$$

from which we can write the normal equations as

$$A\sum_{r=1}^{n} X_r^2 \; + \; B\sum_{r=1}^{n} X_r = \sum_{r=1}^{n} X_r Y_r$$

$$A\sum_{r=1}^{n} X_r \; + \; nB \qquad = \sum_{r=1}^{n} Y_r$$

Construct the table

	x_r	y_r	$X_r = x_r$	$Y_r = \ln y_r$	X_r^2	$X_r Y_r$
	1	1.12	1	0.11333	1	0.11333
	1.25	0.92	1.25	-0.0834	1.5625	-0.1042
	1.5	0.75	1.5	-0.2877	2.25	-0.4315
	1.75	0.61	1.75	-0.4943	3.0625	-0.865
	2	0.51	2	-0.6733	4	-1.3467
	2.25	0.42	2.25	-0.8675	5.0625	-1.9519
$\sum$			9.75	-2.2929	16.9375	-4.5860

Table 6.9. Coefficients of the normal equations in Example 1.4

Substitute in the norm al equations with $n = 6$, we get

$$16.9375A + \; 9.75\text{B} \; = -4.586$$

$$9.75A \qquad + \; 6B \qquad = -2.2929$$

Solving this system by the use of Crammer method we get

$$a = A = -0.7863, \qquad B = 0.8956$$

From which we have

$$b = e^{0.8956} = 2.4488$$

Therefore, the approximation assumes the form

$$\boxed{y(x) = 2.4488e^{-0.7863x},}$$

which gives, at the data points, the values in the following table

x_r	y_r	$y(x_r) = 2.4488e^{-0.7863x}$	$d_r^2 = (y(x_r) - y_r)^2$
1	1.12	1.11549	2.03E-05
1.25	0.92	0.91642	1.279E-05
1.5	0.75	0.75288	8.284E-06
1.75	0.61	0.61852	7.258E-05
2	0.51	0.50814	3.466E-06
2.25	0.42	0.41746	6.474E-06
			0.0001239

Table 6.10. Deviations of the best approximating curve in Example 1.4

Therefore, the least squares of the errors is

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.0001239} \cong 0.01113$$

Example 1.5. Fit the readings

x	1	1.5	2	2.5	3	3.5	4
y	0.12	0.52	1.47	3.24	6.21	10.78	17.35

for the curve $y(x) = mx^c$.

Solution. Taking the logarithm of both sides of the given curve we get the equation

$$\ln y = \ln m + C \ln x,$$

which may be written as

$$\boxed{Y = M + CX,} \quad \begin{aligned} &X = \ln x, \quad Y = \ln y \\ &M = \ln m \quad \Rightarrow \quad m = e^M, \ C = c \end{aligned}$$

from which we can write the normal equations as

$$\begin{aligned} Mn \quad &+ \ C\sum_{r=1}^{n} X_r = \sum_{r=1}^{n} Y_r \\ M\sum_{r=1}^{n} X_r &+ \ C\sum_{r=1}^{n} X_r^2 = \sum_{r=1}^{n} X_r Y_r \end{aligned}$$

Extend the table of data as shown below

x_r	y_r	$X_r = \ln x_r$	$Y_r = \ln y_r$	X_r^2	$X_r Y_r$
1	0.12	0	-2.1203	0	0
1.5	0.52	0.40547	-0.6539	0.1644	-0.2651
2	1.47	0.69315	0.38526	0.48045	0.26704
2.5	3.24	0.91629	1.17557	0.83959	1.07717
3	6.21	1.09861	1.82616	1.20695	2.00624
3.5	10.78	1.25276	2.37769	1.56942	2.97869
4	17.35	1.38629	2.85359	1.92181	3.95592
		5.75257	5.84409	6.18262	10.0199

Table 6.11. Coefficients of the normal equations in Example 1.5

The substitution in the normal equations yields $(n = 7)$,

$$\begin{aligned} 7\mathrm{M} \quad\quad\quad\quad &+ 5.75257\mathrm{C} = 5.84409 \\ 5.75257\mathrm{M} &+ 6.18262\mathrm{C} = 10.0199 \end{aligned}$$

Solving this system we obtain

$$M = -2.1115 \qquad \Rightarrow \qquad m = 0.1211,$$
$$c = C = 3.5853$$

Thus, the approximation curve is

$$\boxed{y(x) = 0.1211x^{3.5853},}$$

which gives, at the data points, the values in the following table

x_r	y_r	$y(x_r) = 0.1211x^{3.5853}$	$d_r^2 = (y(x_r) - y_r)^2$
1	0.12	0.1211	1.2E-06
1.5	0.52	0.51818	3.3E-06
2	1.47	1.45354	0.00027
2.5	3.24	3.23503	2.5E-05
3	6.21	6.21966	9.3E-05
3.5	10.78	10.8091	0.00085
4	17.35	17.4466	0.00932
			0.01057

Table 6.12. Deviations of the best approximating curve in Example 1.5

The last column of this table gives the $L_2 - norm$ of the error as

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.01057} \cong 0.102795$$

Example 1.6. Fit the curve $y(x) = \ln(ax + b)$ to the readings:

x	2	4	5	7	8
y	2.4	2.9	3.2	3.5	3.6

Solution. The approximation curve $y = \ln(ax + b)$ can be written as $e^y = ax + b$, or

$$\boxed{Y = AX + B,} \qquad X = x, \quad Y = e^y \qquad A = a\,, \quad B = b$$

Hence, the normal equations are

$$A\sum_{r=1}^{n} x_r^2 + B\sum_{r=1}^{n} x_r = \sum_{r=1}^{n} x_r Y_r$$

$$A\sum_{r=1}^{n} x_r + nB = \sum_{r=1}^{n} Y_r$$

Construct the table

	x_r	y_r	$X_r = x_r$	$Y_r = e^{y_r}$	X_r^2	$X_r Y_r$
	2	2.4	2	11.0232	4	22.0464
	4	2.9	4	18.1741	16	72.6966
	5	3.2	5	24.5325	25	122.663
	7	3.5	7	33.1155	49	231.808
	8	3.6	8	36.5982	64	292.786
$\sum$	26			123.4435	158	742

Table 6.13. The coefficients of the normal equations in Example 1.6

Substitute in the norm al equations with $n = 5$, we obtain

$$158A + 26B = 742$$
$$26A + 5B = 123.4435$$

Solving this system to get

$$a = A = 4.3901, \qquad b = B = 1.8604$$

That is, the approximating curve is

$$\boxed{y(x) = \ln(4.3901x + 1.8604),}$$

so that we have the following table

x_r	y_r	$y(x_r) = \ln(4.3901x_r + 1.8604)$	$d_r^2 = (y(x_r) - y_r)^2$
2	2.4	2.36468	0.00125
4	2.9	2.96634	0.0044
5	3.2	3.17014	0.00089
7	3.5	3.48404	0.00025
8	3.6	3.61041	0.00011
			0.0069

Table 6.14. Deviations of the best approximating curve in Example 1.6

The last column of this table gives the $L_2 - norm$ of the error as

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.0069} \cong 0.0831$$

Example 1.7. Fit the curve $y(x) = \dfrac{x}{a + bx}$ to the readings

x	2	3	6	7	8	9
y	1	5.95	-1.49	-1.29	-1.14	-1.05

Solution. Writing the approximation curve in the form

$$\frac{1}{y} = \frac{a + bx}{x} = a\left(\frac{1}{x}\right) + b$$

So that we have the linear form

$$\boxed{Y = AX + B,} \qquad X = \frac{1}{x}, \quad Y = \frac{1}{y}$$

$$A = a\,, \quad B = b$$

from which we can write the normal equations as

$$A\sum_{r=1}^{n} X_r^2 + B\sum_{r=1}^{n} X_r = \sum_{r=1}^{n} X_r Y_r$$

$$A\sum_{r=1}^{n} X_r + nB = \sum_{r=1}^{n} Y_r$$

From the given readings, we have

x_r	y_r	$X_r = \frac{1}{x_r}$	$Y_r = \frac{1}{y_r}$	X_r^2	$X_r Y_r$
2	1	0.5	1	0.25	0.5
3	5.95	0.33333	0.16807	0.11111	0.05602
6	-1.49	0.16667	-0.6711	0.02778	-0.1119
7	-1.29	0.14286	-0.7752	0.02041	-0.1107
8	-1.14	0.125	-0.8772	0.01563	-0.1096
9	-1.05	0.11111	-0.9524	0.01235	-0.1058
		1.3789	-2.1078	0.4373	0.1180

Table 6.15. Coefficients of the normal equations in Example 1.7

Substitute in the normal equations we obtain$(n = 6)$,

$$0.4373a + 1.3789\text{b} = 0.1180$$

$$1.3789a + 6b = -2.1078$$

Solving this system

$$a = A = 5.0057, \qquad b = B = -1.5018$$

Hence, the approximating curve is

$$\boxed{y(x) = \frac{x}{5.0057 - 1.5018x},}$$

which gives

x_r	y_r	$y(x_r) = \frac{x}{5.0057 - 1.5018x}$	$d_r^2 = (y(x_r) - y_r)^2$
2	1	0.99895	1.1E-06
3	5.95	5.9964	0.0021532
6	-1.49	-1.4981	6.545E-05
7	-1.29	-1.2711	0.000356
8	-1.14	-1.1414	2.069E-06
9	-1.05	-1.0575	5.651E-05
			0.0026343

The last column of this table gives the $L_2 - norm$ of the error as

$$\|d\| = \left(\sum_{r=1}^{n} {d_r}^2\right)^{1/2} = \sqrt{0.0026343} \cong 0.05$$

Exercise 6.1

1. Fit the equation of an straight line for the readings

x	1	1.5	2	2.5	3	4
y	0.27	0.64	1.02	1.39	1.77	2.5

[Answer: $y(x) = 0.744857x - 0.473, \ \|d\| = 0.008641$]

2. Fit the equation $y = ax + b$ for the readings

x	2	2.25	2.5	2.75	3	3.25	3.5
y	6.76	7.40	8.04	8.67	9.31	9.95	10.59

3. Fit the equation $y = bx^a$ to the following readings

x	2	3	5	6	10
y	2.83	5.20	11.18	14.70	31.62

[Answer: $y(x) = x^{1.4995}$, $\|d\| = 0.0038921$]

4. Fit the equation $y = cx^m$ to the following readings

x	3	4	6	7	9	10
y	1.39	1.19	0.95	0.87	0.76	0.72

5. Use the readings

x	2	2.5	3	3.5	4	4.5	5
y	1.98	1.59	1.27	1.02	0.82	0.66	0.53

to fit the equation $y = ae^{bx}$.

[Answer: $y(x) = 4.7583e^{-0.4393x}$, $\|d\| = 0.006829$]

6. Fit the equation $y = a + bx^4$ to the data and then find y *and* y' when $x = 1.2$.

x	0.25	0.5	0.75	1	1.25	1.5	1.75	2
y	10.23	10.20	10.08	9.73	9.01	7.70	5.54	2.23

7. Fit the equation $y = mx^3 + c$ to the following readings

x	0.25	0.5	1.25	2	3	0.25
y	2.36	2.50	4.76	12.22	35.66	2.36

[Answer: $y(x) = 1.2339x^3 + 2.3461$, $\|d\| = 0.007336$]

8. Fit the curve $y = ae^{-bx^2}$ to the data:

x	1	2	3	4	5	8
y	9.51	8.19	6.38	4.49	2.87	0.41

Find y at $x = 3.5$ and x at $y = 7.5$.

9. Fit the equation $y = a\sqrt[3]{x} + b$ to the following readings

x	1	3	5	7	9
y	2.1	2.7	3.1	3.4	3.7

$\left[\text{Answer: } y(x) = 1.4692\sqrt[3]{x} + 0.6066\,,\ \|d\| = 0.057286\right]$

10. Use the following readings to fit the equation $y = \sqrt{ax + b}$.

x	0	2	4	6	9
y	2.1	2.7	3.1	3.4	3.7

11. Fit the following readings for the curve $y = \dfrac{1}{ax + b}$.

x	1	2	3	4	5
y	0.25	0.17	0.11	0.1	0.08

$\left[\text{Answer: } y(x) = \dfrac{1}{2.1118x + 1.9594}\,,\ \|d\| = 0.014635\right]$

12. Fit the following readings for the curve $y = \dfrac{x}{ax + b}$.

x	1	2	4	6
y	0.16	0.26	0.32	0.40

13. Fit the curve $y = \sqrt{ax} + b$ to the readings:

x	2.5	3	3.5	4	4.5
y	2.07	2.75	3.37	3.94	4.49

$\left[\text{Answer: } y(x) = \sqrt{20.00236x} - 4.9994\,,\ \|d\| = 007191\right]$

14. Given the table of readings:

x	1.0	1.5	2.0	2.5	3
y	0.12	0.52	1.47	3.24	6.21

Fit these readings for the curve $y = ax^b$.

15. Fit the curve $y = me^{cx^2}$ to the readings

x	1	1.25	1.5	1.75	2
y	4.7	4.2	3.7	3.1	2.6

Find the $L_2 - norm$ of the error.

[Answer: $y(x) = 5.7362e^{-0.1984x^2}$, $\|d\| = 0.039334$]

16. Fit the equation $y = \frac{1}{a+b\cos x}$ to the following readings

x	30°	45°	60°	80°
y	0.35	0.43	0.52	0.73

17. Fit the following readings to the curve $y = \left(\frac{\sqrt{x} + a}{b\sqrt{x}}\right)^2$,

x	2.5	3	3.5	4	4.5
y	7.45	6.30	5.47	4.84	4.35

and then find the $L_2 - norm$ of the error.

$$\left[\text{Answer: } y(x) = \left(\frac{\sqrt{x} + 19.90393}{4.977601\sqrt{x}}\right)^2, \quad \|d\| = 0.004248\right]$$

18. Fit the readings

x	1	2	3	4	5	6
y	2.85	5.16	11.04	20.92	34.86	43.20

to the curve $y = ax^2 + bx + c$.

19. Given the table of readings:

x	2	4	6	8	10
y	33.20	38.04	41.20	43.60	45.56

Fit the above readings to the curve

$$y = ax^b + 1,$$

and compute the $L_2 - norm$ of the error.

[Answer: $y(x) = 0.2004x^{0.2996} + 1$, $\|d\| = 0.004889$]

20. Fit the curve $y = a + b\sin^2 x$ to the data

$$(0, 42), \left(\frac{\pi}{6}, 58\right), \left(\frac{\pi}{4}, 83\right), \left(\frac{\pi}{3}, 99\right), \left(\frac{\pi}{2}, 125\right)$$