Numerical Methods

(Problems and Solutions)

M. K. Jain
S.R.K. Iyengar
R.K. Jain

NEW AGE
NEW AGE INTERNATIONAL PUBLISHERS

Numerical Methods

(Problems and Solutions)

This page intentionally left blank

NUMERICAL METHODS

(Problems and Solutions)

Revised Second Edition

M.K. JAIN
S.R.K. IYENGAR
R.K. JAIN
Department of Mathematics
Indian Institute of Technology Delhi
India

PUBLISHING FOR ONE WORLD

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
New Delhi • Bangalore • Chennai • Cochin • Guwahati • Hyderabad
Jalandhar • Kolkata • Lucknow • Mumbai • Ranchi
Visit us at **www.newagepublishers.com**

This page intentionally left blank

PREFACE

We thank the faculty and the students of various universities, Engineering colleges and others for sending their suggestions for improving this book. Based on their suggestions, we have made the follwoing changes.

(*i*) New problems have been added and detailed solutions for many problems are given.

(*ii*) C-programs of frequently used numerical methods are given in the Appendix. These programs are written in a simple form and are user friendly. Modifications to these programs can be made to suit individual requirements and also to make them robust.

We look forward to more suggestions from the faculty and the students. We are thankful to New Age International Limited for bringing out this Second Edition.

New Delhi

M.K. Jain

S.R.K. Iyengar

R.K. Jain

This page intentionally left blank

CONTENTS

CHAPTER 1

Transcendental and Polynomial Equations

1.1 INTRODUCTION

We consider the methods for determining the roots of the equation

$$f(x) = 0 \tag{1.1}$$

which may be given explicitly as a polynomial of degree n in x or $f(x)$ may be defined implicitly as a transcendental function. A transcendental equation (1.1) may have no root, a finite or an infinite number of real and / or complex roots while a polynomial equation (1.1) has exactly n (real and / or complex) roots. If the function $f(x)$ changes sign in any one of the intervals $[x^* - \varepsilon, x^*]$, $[x^*, x^* + \varepsilon]$, then x^* defines an approximation to the root of $f(x)$ with accuracy ε. This is known as *intermediate value* theorem. Hence, if the interval $[a, b]$ containing x^* and ξ where ξ is the exact root of (1.1), is sufficiently small, then

$$| x^* - \xi | \le b - a$$

can be used as a measure of the error.

There are two types of methods that can be used to find the roots of the equation (1.1).

(*i*) *Direct methods* : These methods give the exact value of the roots (in the absence of round off errors) in a finite number of steps. These methods determine all the roots at the same time.

(*ii*) *Iterative methods* : These methods are based on the idea of successive approximations. Starting with one or more initial approximations to the root, we obtain a sequence of iterates $\{x_k\}$ which in the limit converges to the root. These methods determine one or two roots at a time.

Definition 1.1 A sequence of iterates $\{x_k\}$ is said to *converge* to the root ξ if

$$\lim_{k \to \infty} | x_k - \xi | = 0.$$

If $x_k, x_{k-1}, \ldots, x_{k-m+1}$ are m approximates to a root, then we write an iteration method in the form

$$x_{k+1} = \phi(x_k, x_{k-1}, \ldots, x_{k-m+1}) \tag{1.2}$$

where we have written the equation (1.1) in the equivalent form

$$x = \phi(x).$$

The function ϕ is called the *iteration function.* For $m = 1$, we get the one-point iteration method

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots \tag{1.3}$$

If $\phi(x)$ is continuous in the interval $[a, b]$ that contains the root and $| \phi'(x) | \le c < 1$ in this interval, then for any choice of $x_0 \in [a, b]$, the sequence of iterates $\{x_k\}$ obtained from (1.3) converges to the root of $x = \phi(x)$ or $f(x) = 0$.

Thus, for any iterative method of the form (1.2) or (1.3), we need the iteration function $\phi(x)$ and one or more initial approximations to the root.

In practical applications, it is not always possible to find ξ exactly. We therefore attempt to obtain an approximate root x_{k+1} such that

$$| f(x_{k+1}) | < \varepsilon \tag{1.4}$$

and / or

$$| x_{k+1} - x_k | < \varepsilon \tag{1.5}$$

where x_k and x_{k+1} are two consecutive iterates and ε is the prescribed error tolerance.

Definition 1.2 An iterative method is said to be of *order p* or has the *rate of convergence p*, if p is the largest positive real number for which

$$| \varepsilon_{k+1} | \le c \, | \varepsilon_k |^p \tag{1.6}$$

where $\varepsilon_k = x_k - \xi$ is the error in the kth iterate.

The constant c is called the *asymptotic error constant*. It depends on various order derivatives of $f(x)$ evaluated at ξ and is independent of k. The relation

$$\varepsilon_{k+1} = c\varepsilon_k^{\,p} + O(\varepsilon_k^{\,p+1})$$

is called the *error equation*.

By substituting $x_i = \xi + \varepsilon_i$ for all i in any iteration method and simplifying we obtain the error equation for that method. The value of p thus obtained is called the order of this method.

1.2 ITERATIVE METHODS FOR SIMPLE ROOTS

A root ξ is called a simple root of $f(x) = 0$, if $f(\xi) = 0$ and $f'(\xi) \ne 0$. Then, we can also write $f(x) = (x - \xi)\, g(x)$, where $g(x)$ is bounded and $g(\xi) \ne 0$.

Bisection Method

If the function $f(x)$ satisfies $f(a_0) f(b_0) < 0$, then the equation $f(x) = 0$ has atleast one real root or an odd number of real roots in the interval (a_0, b_0). If $m_1 = \frac{1}{2}(a_0 + b_0)$ is the mid point of this interval, then the root will lie either in the interval (a_0, m_1) or in the interval (m_1, b_0) provided that $f(m_1) \ne 0$. If $f(m_1) = 0$, then m_1 is the required root. Repeating this procedure a number of times, we obtain the bisection method

$$m_{k+1} = a_k + \frac{1}{2}(b_k - a_k), \quad k = 0, 1, \ldots \tag{1.7}$$

where

$$(a_{k+1}, b_{k+1}) = \left| \begin{array}{ll} (a_k, m_{k+1}), & \text{if } f(a_k)\, f(m_{k+1}) < 0, \\ (m_{k+1}, b_k), & \text{if } f(m_{k+1})\, f(b_k) < 0. \end{array} \right.$$

We take the midpoint of the last interval as an approximation to the root. This method always converges, if $f(x)$ is continuous in the interval $[a, b]$ which contains the root. If an error tolerance ε is prescribed, then the approximate number of the iterations required may be determined from the relation

$$n \ge [\log(b_0 - a_0) - \log \varepsilon] / \log 2.$$

Secant Method

In this method, we approximate the graph of the function $y = f(x)$ in the neighbourhood of the root by a straight line (secant) passing through the points (x_{k-1}, f_{k-1}) and (x_k, f_k), where $f_k = f(x_k)$ and take the point of intersection of this line with the x-axis as the next iterate. We thus obtain

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k, \quad k = 1, 2, ...$$

or

$$x_{k+1} = \frac{x_{k-1} f_k - x_k f_{k-1}}{f_k - f_{k-1}}, \quad k = 1, 2, ... \tag{1.8}$$

where x_{k-1} and x_k are two consecutive iterates. In this method, we need two initial approximations x_0 and x_1. This method is also called the *chord* method. The order of the method (1.8) is obtained as

$$p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.62.$$

If the approximations are chosen such that $f(x_{k-1}) f(x_k) < 0$ for each k, then the method is known as *Regula-Falsi* method and has linear (first order) rate of convergence. Both these methods require one function evaluation per iteration.

Newton-Raphson method

In this method, we approximate the graph of the function $y = f(x)$ in the neighbourhood of the root by the tangent to the curve at the point (x_k, f_k) and take its point of intersection with the x-axis as the next iterate. We have the Newton-Raphson method as

$$x_{k+1} = x_k - \frac{f_k}{f_k'}, \quad k = 0, 1, ... \tag{1.9}$$

and its order is $p = 2$. This method requires one function evaluation and one first derivative evaluation per iteration.

Chebyshev method

Writing $f(x) = f(x_k + x - x_k)$ and approximating $f(x)$ by a second degree Taylor series expansion about the point x_k, we obtain the method

$$x_{k+1} = x_k - \frac{f_k}{f_k'} - \frac{1}{2}(x_{k+1} - x_k)^2 \frac{f_k''}{f_k'}$$

Replacing $x_{k+1} - x_k$ on the right hand side by $(-f_k / f_k')$, we get the Chebyshev method

$$x_{k+1} = x_k - \frac{f_k}{f_k'} - \frac{1}{2}\left(\frac{f_k}{f_k'}\right)^2 \frac{f_k''}{f_k'}, \quad k = 0, 1, ... \tag{1.10}$$

whose order is $p = 3$. This method requires one function, one first derivative and one second derivative evaluation per iteration.

Multipoint iteration methods

It is possible to modify the Chebyshev method and obtain third order iterative methods which do not require the evaluation of the second order derivative. We give below two multipoint iteration methods.

(*i*)
$$x^*_{k+1} = x_k - \frac{1}{2}\frac{f_k}{f_k'}$$
$$x_{k+1} = x_k - \frac{f_k}{f'(x^*_{k+1})} \tag{1.11}$$
order $p = 3$.

This method requires one function and two first derivative evaluations per iteration.

(*ii*)
$$x^*_{k+1} = x_k - \frac{f_k}{f_k'}$$
$$x_{k+1} = x^*_{k+1} - \frac{f(x^*_{k+1})}{f_k'} \tag{1.12}$$
order $p = 3$.

This method requires two functions and one first derivative evaluation per iteration.

Müller Method

This method is a generalization of the secant method. In this method, we approximate the graph of the function $y = f(x)$ in the neighbourhood of the root by a second degree curve and take one of its points of intersection with the x axis as the next approximation.

We have the method as
$$x_{k+1} = x_k + (x_k - x_{k-1})\,\lambda_{k+1}, \quad k = 2, 3, \ldots \tag{1.13}$$
where $h_k = x_k - x_{k-1}$, $h_{k-1} = x_{k-1} - x_{k-2}$,
$$\lambda_k = h_k / h_{k-1}, \quad \delta_k = 1 + \lambda_k,$$
$$g_k = \lambda_k^2\, f(x_{k-2}) - \delta_k^2\, f(x_{k-1}) + (\lambda_k + \delta_k)\, f(x_k),$$
$$c_k = \lambda_k(\lambda_k f(x_{k-2}) - \delta_k f(x_{k-1}) + f(x_k)),$$
$$\lambda_{k+1} = -\frac{2\delta_k\, f(x_k)}{g_k \pm \sqrt{g_k^2 - 4\delta_k\, c_k\, f(x_k)}}.$$
The sign in the denominator is chosen so that λ_{k+1} has the smallest absolute value, *i.e.*, the sign of the square root in the denominator is that of g_k.

Alternative

We have the method as
$$x_{k+1} = x_k - \frac{2a_2}{a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}, \quad k = 2, 3, \ldots \tag{1.14}$$
where $a_2 = f_k$, $h_1 = x_k - x_{k-2}$, $h_2 = x_k - x_{k-1}$, $h_3 = x_{k-1} - x_{k-2}$,
$$a_1 = \frac{1}{D}[h_1^2\,(f_k - f_{k-1}) - h_2^2\,(f_k - f_{k-2})],$$
$$a_0 = \frac{1}{D}[h_1(f_k - f_{k-1}) - h_2(f_k - f_{k-2})],$$
$$D = h_1h_2h_3.$$
The sign in the denominator is chosen so that λ_{k+1} has the smallest absolute value, *i.e.*, the sign of the square root in the denominator is that of a_1.

This method requires three initial approximations to the root and one function evaluation per iteration. The order of the method is $p = 1.84$.

Derivative free methods

In many practical applications, only the data regarding the function $f(x)$ is available. In these cases, methods which do not require the evaluation of the derivatives can be applied.

We give below two such methods.

(*i*)
$$x_{k+1} = x_k - \frac{f_k}{g_k}, \quad k = 0, 1, ... \tag{1.15}$$
$$g_k = \frac{f(x_k + f_k) - f_k}{f_k},$$
order $p = 2$.

This method requires two function evaluations per iteration.

(*ii*)
$$x_{k+1} = x_k - w_1(x_k) - w_2(x_k), \quad k = 0, 1, ... \tag{1.16}$$
$$w_1(x_k) = \frac{f_k}{g_k}$$
$$w_2(x_k) = \frac{f(x_k - w_1(x_k))}{g_k}$$
$$g_k = \frac{f(x_k + \beta f_k) - f_k}{\beta f_k}$$
where $\beta \neq 0$ is arbitrary and order $p = 3$.

This method requires three function evaluations per iteration.

Aitken Δ^2-process

If x_{k+1} and x_{k+2} are two approximations obtained from a general linear iteration method
$$x_{k+1} = \phi(x_k), \quad k = 0, 1, ...$$
then, the error in two successive approximations is given by
$$\varepsilon_{k+1} = a_1 \varepsilon_k$$
$$\varepsilon_{k+2} = a_1 \varepsilon_{k+1}, \quad a_1 = \phi'(\xi).$$
Eliminating a_1 from the above equations, we get
$$\varepsilon_{k+1}^2 = \varepsilon_k \varepsilon_{k+2}$$
Using $\varepsilon_k = \xi - x_k$, we obtain
$$\xi \approx x_k^* = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k} \tag{1.17}$$
$$= x_k - \frac{(\Delta x_k)^2}{\Delta^2 x_k}$$
which has second order convergence.

A Sixth Order Method

A one-parameter family of sixth order methods for finding simple zeros of $f(x)$, which require three evaluations of $f(x)$ and one evaluation of the derivative $f'(x)$ are given by
$$w_n = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$z_n = w_n - \frac{f(w_n)}{f'(x_n)}\left[\frac{f(x_n)+Af(w_n)}{f(x_n)+(A-2)f(w_n)}\right]$$

$$x_{n+1} = z_n - \frac{f(z_n)}{f'(x_n)}\left[\frac{f(x_n)-f(w_n)+Df(z_n)}{f(x_n)-3f(w_n)+Df(z_n)}\right] \quad n = 0, 1, \ldots$$

with error term

$$\varepsilon_{n+1} = \frac{1}{144}\,[2F_3^2F_2 - 3(2A+1)F_2^3F_3]\,\varepsilon_n^{\,6} + \ldots$$

where $F^{(i)} = f^{(i)}(\xi) / f'(\xi)$.

The order of the methods does not depend on D and the error term is simplified when $A = -1/2$. The simplified formula for $D = 0$ and $A = -1/2$ is

$$w_n = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$z_n = w_n - \frac{f(w_n)}{f'(x_n)}\left[\frac{2f(x_n)-f(w_n)}{2f(x_n)-5f(w_n)}\right]$$

$$x_{n+1} = z_n - \frac{f(z_n)}{f'(x_n)}\left[\frac{f(x_n)-f(w_n)}{f(x_n)-3f(w_n)}\right], \; n = 0, 1, \ldots$$

1.3 ITERATIVE METHODS FOR MULTIPLE ROOTS

If the root ξ of (1.1) is a repeated root, then we may write (1.1) as

$$f(x) = (x-\xi)^m g(x) = 0$$

where $g(x)$ is bounded and $g(\xi) \neq 0$. The root ξ is called a *multiple root of multiplicity m*. We obtain from the above equation

$$f(\xi) = f'(\xi) = \cdots = f^{(m-1)}(\xi) = 0, \; f^{(m)}(\xi) \neq 0.$$

The methods listed in Section 1.2 do not retain their order while determining a multiple root and the order is reduced atleast by one. If the multiplicity m of the root is known in advance, then some of these methods can be modified so that they have the same rate of convergence as that for determining simple roots. We list some of the modified methods.

Newton-Raphson method

$$x_{k+1} = x_k - m\,\frac{f_k}{f_k'}, \quad k = 0, 1, \ldots \tag{1.18}$$

order $p = 2$.

Chebyshev method

$$x_{k+1} = x_k - \frac{m(3-m)}{2}\frac{f_k}{f_k'} - \frac{m^2}{2}\left[\frac{f_k}{f_k'}\right]^2\frac{f_k''}{f_k'}, \quad k = 0, 1, \ldots \tag{1.19}$$

order $p = 3$.

Alternatively, we apply the methods given in Section 1.2 to the equation

$$G(x) = 0 \tag{1.20}$$

where $$G(x) = \frac{f(x)}{f'(x)}$$

has a simple root ξ regardless of the multiplicity of the root of $f(x) = 0$. Thus, the Newton-Raphson method (1.9), when applied to (1.20) becomes

$$x_{k+1} = x_k - \frac{G(x_k)}{G'(x_k)}$$

or
$$x_{k+1} = x_k - \frac{f(x_k)\, f'(x_k)}{[f'(x_k)]^2 - f(x_k)\, f''(x_k)}, \quad k = 0, 1, 2, \ldots \tag{1.21}$$

The secant method for (1.20) can be written as

$$x_{k+1} = \frac{x_{k-1} f_k f'_{k-1} - x_k f_{k-1} f_k'}{f_k f'_{k-1} - f_{k-1} f_k'}.$$

Derivative free method

$$x_{k+1} = x_k - W_1(x_k) - W_2(x_k) \tag{1.22}$$

$$W_1(x_k) = \frac{F(x_k)}{g(x_k)}$$

$$W_2(x_k) = \frac{F(x_k - W_1(x_k))}{g(x_k)}$$

$$g(x_k) = \frac{F(x_k + \beta F(x_k)) - F(x_k)}{\beta F(x_k)}$$

where
$$F(x) = -\frac{f^2(x)}{f(x - f(x)) - f(x)}$$

and $\beta \neq 0$ is an arbitrary constant. The method requires six function evaluations per iteration and has order 3.

1.4 ITERATIVE METHODS FOR A SYSTEM OF NONLINEAR EQUATIONS

Let the given system of equations be

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ &\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned} \tag{1.23}$$

Starting with the initial approximations $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$, we obtain the sequence of iterates, using the Newton-Raphson method as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{J}^{-1}\, \mathbf{f}^{(k)}, \quad k = 0, 1, \ldots \tag{1.24}$$

where
$$\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T$$
$$\mathbf{f}^{(k)} = (f_1^{(k)}, f_2^{(k)}, \ldots, f_n^{(k)})^T$$
$$f_i^{(k)} = f_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$$

and $\mathbf{J}$ is the *Jacobian matrix* of the functions $f_1, f_2, \ldots, f_n$ evaluated at $(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$. The method has second order rate of convergence.

Alternatively, we may write (1.24) in the form

$$\mathbf{J}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = -\,\mathbf{f}^{(k)}$$

and may solve it as a linear system of equations. Very often, for systems which arise while solving ordinary and partial differential equations, $\mathbf{J}$ is of some special form like a tridiagonal, five diagonal or a banded matrix.

1.5 COMPLEX ROOTS

We write the given equation

$$f(z) = 0, \quad z = x + iy$$

in the form $\quad u(x, y) + iv(x, y) = 0,$

where $u(x, y)$ and $v(x, y)$ are the real and imaginary parts of $f(z)$ respectively. The problem of finding a complex root of $f(z) = 0$ is equivalent to finding a solution (x, y) of the system of two equations

$$u(x, y) = 0,$$
$$v(x, y) = 0.$$

Starting with $(x^{(0)}, y^{(0)})$, we obtain a sequence of iterates $\{x^{(k)}, y^{(k)}\}$ using the Newton-Raphson method as

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} - \mathbf{J}^{-1} \begin{pmatrix} u(x^{(k)}, y^{(k)}) \\ v(x^{(k)}, y^{(k)}) \end{pmatrix}, \quad k = 0, 1, \ldots \tag{1.25}$$

where
$$\mathbf{J} = \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix}_{(x^{(k)}, y^{(k)})}$$

is the Jacobian matrix of $u(x, y)$ and $v(x, y)$ evaluated at $(x^{(k)}, y^{(k)})$.

Alternatively, we can apply directly the Newton-Raphson method (1.9) to solve $f(z) = 0$ in the form

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)}, \quad k = 0, 1, \ldots, \tag{1.26}$$

and use complex arithmetic. The initial approximation z_0 must also be complex. The secant method can also be applied using complex arithmetic.

After one root z_1 is obtained, Newton's method should be applied on the deflated polynomial

$$f^*(z) = \frac{f(z)}{z - z_1}.$$

This procedure can be repeated after finding every root. If k roots are already obtained, then the new iteration can be applied on the function

$$f^*(z) = \frac{f(z)}{(z - z_1)(z - z_2) \ldots (z - z_k)}.$$

The new iteration is

$$z_{k+1} = z_k - \frac{f^*(z_k)}{f^{*\prime}(z_k)}.$$

The computation of $f^*(z_k) / f^{*\prime}(z_k)$ can be easily performed as follows

$$\frac{f^{*\prime}}{f^*} = \frac{d}{dz}(\log f^*) = \frac{d}{dz}[\log f(z) - \log (z - z_1)]$$
$$= \frac{f'}{f} - \frac{1}{z - z_1}.$$

Hence, computations are carried out with

$$\frac{f^{*\prime}(z_k)}{f^*(z_k)} = \frac{f'(z_k)}{f(z_k)} - \frac{1}{z_k - z_1}$$

Further, the following precautions may also be taken :

(*i*) Any zero obtained by using the deflated polynomial should be refined by applying Newton's method to the original polynomial with this zero as the starting approximation.

(*ii*) The zeros should be computed in the increasing order of magnitude.

1.6 ITERATIVE METHODS FOR POLYNOMIAL EQUATIONS

The methods discussed in the previous sections can be directly applied to obtain the roots of a polynomial of degree n

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n = 0 \tag{1.27}$$

where $a_0, a_1, \ldots, a_n$ are real numbers. Most often, we are interested to determine all the roots (real or complex, simple or multiple) of the polynomial and we need to know

(*i*) the exact number of real and complex roots along with their multiplicities.

(*ii*) the interval in which each real roots lies.

We can obtain this information using *Sturm sequences.*

Let $f(x)$ be the given polynomial of degree n and let $f_1(x)$ denote its first order derivative. Denote by $f_2(x)$ the remainder of $f(x)$ divided by $f_1(x)$ taken with reverse sign and by $f_3(x)$ the remainder of $f_1(x)$ divided by $f_2(x)$ with the reverse sign and so on until a constant remainder is obtained. The sequence of the functions $f(x), f_1(x), f_2(x), \ldots, f_n(x)$ is called the *Sturm sequence.* The number of real roots of the equation $f(x) = 0$ in (a, b) equals the difference between the number of sign changes in the Sturm sequence at $x = a$ and $x = b$ provided $f(a) \neq 0$ and $f(b) \neq 0$.

We note that if any function in the Sturm sequence becomes 0 for some value of x, we give to it the sign of the immediate preceding term.

If $f(x) = 0$ has a multiple root, we obtain the Sturm sequence $f(x), f_1(x), \ldots, f_r(x)$ where $f_{r-1}(x)$ is exactly divisible by $f_r(x)$. In this case, $f_r(x)$ will not be a constant. Since $f_r(x)$ gives the greatest common divisor of $f(x)$ and $f'(x)$, the multiplicity of the root of $f(x) = 0$ is one more than that of the root of $f_r(x) = 0$. We obtain a new Sturm sequence by dividing all the functions $f(x), f_1(x), \ldots, f_r(x)$ by $f_r(x)$. Using this sequence, we determine the number of real roots of the equation $f(x) = 0$ in the same way, without taking their multiplicity into account.

While obtaining the Sturm sequence, any positive constant common factor in any Sturm function $f_i(x)$ can be neglected.

Since a polynomial of degree n has exactly n roots, the number of complex roots equals (n–number of real roots), where a real root of multiplicity m is counted m times.

If $x = \xi$ is a real root of $P_n(x) = 0$ then $x - \xi$ must divide $P_n(x)$ exactly. Also, if $x = \alpha + i\beta$ is a complex root of $P_n(x) = 0$, then its complex conjugate $\alpha - i\beta$ is also a root. Hence

$$\begin{aligned}\{x - (\alpha + i\beta)\}\,\{x - (\alpha - i\beta)\} &= (x - \alpha)^2 + \beta^2 \\ &= x^2 - 2\alpha x + \alpha^2 + \beta^2 \\ &= x^2 + px + q\end{aligned}$$

for some real p and q must divide $P_n(x)$ exactly.

The quadratic factor $x^2 + px + q = 0$ may have a pair of real roots or a pair of complex roots.

Hence, the iterative methods for finding the real and complex roots of $P_n(x) = 0$ are based on the philosophy of extracting linear and quadratic factors of $P_n(x)$.

We assume that the polynomial $P_n(x)$ is complete, that is, it has $(n + 1)$ terms. If some term is not present, we introduce it at the proper place with zero coefficient.

Birge-Vieta method

In this method, we seek to determine a real number p such that $x - p$ is a factor of $P_n(x)$. Starting with p_0, we obtain a sequence of iterates $\{ p_k \}$ from

$$p_{k+1} = p_k - \frac{P_n(p_k)}{P_n'(p_k)}, \quad k = 0, 1, ... \tag{1.28}$$

or

$$p_{k+1} = p_k - \frac{b_n}{c_{n-1}}, \quad k = 0, 1, ... \tag{1.29}$$

which is same as the Newton-Raphson method.

The values of b_n and c_{n-1} are obtained from the recurrence relations

$$b_i = a_i + p_k b_{i-1}, \quad i = 0, 1, ... , n$$
$$c_i = b_i + p_k c_{i-1}, \quad i = 0, 1, ... , n - 1$$

with $\quad c_0 = b_0 = a_0, \quad b_{-1} = 0 = c_{-1}.$

We can also obtain b_i's and c_i's by using *synthetic division method* as given below :

p_k	a_0	a_1	a_2	...	a_{n-1}	a_n
		p_kb_0	p_kb_1	...	p_kb_{n-2}	p_kb_{n-1}
	b_0	b_1	b_2	...	b_{n-1}	b_n
		p_kc_0	p_kc_1	...	p_kc_{n-2}	
	c_0	c_1	c_2	...	c_{n-1}	

where $\quad b_0 = a_0 \quad$ and $\quad c_0 = b_0 = a_0.$

We have

$$\lim_{k \to \infty} b_n = 0 \quad \text{and} \quad \lim_{k \to \infty} p_k = p.$$

The order of this method is 2.

When p has been determined to the desired accuracy, we extract the next linear factor from the *deflated polynomial*

$$Q_{n-1}(n) = \frac{P_n(x)}{x - p} = b_0 x^{n-1} + b_1 x^{n-2} + ... + b_{n-1}$$

which can also be obtained from the first part of the synthetic division.

Synthetic division procedure for obtaining b_n is same as Horner's method for evaluating the polynomial $P_n(p_k)$, which is the most efficient way of evaluating a polynomial.

We can extract a multiple root of multiplicity m, using the Newton-Raphson method

$$p_{k+1} = p_k = - m \frac{b_n}{c_{n-1}}, \quad k = 0, 1, 2, ...$$

In this case, care should be taken while finding the deflated polynomial. For example, if $m = 2$, then as $k \to \infty$, $f(x) \approx b_n \to 0$ and $f'(x) \approx c_{n-1} \to 0$. Hence, the deflated polynomial is given by

$$c_0 x^{n-2} + c_1 x^{n-3} + ... + c_{n-2} = 0.$$

Bairstow method

This method is used to find two real numbers p and q such that $x^2 + px + q$ is a factor of $P_n(x)$. Starting with p_0, q_0, we obtain a sequence of iterates $\{(p_k, q_k)\}$ from

$$p_{k+1} = p_k + \Delta p_k,$$
$$q_{k+1} = q_k + \Delta q_k, \quad k = 0, 1, \ldots \tag{1.30}$$

where $$\Delta p_k = -\frac{b_n c_{n-3} - b_{n-1} c_{n-2}}{c_{n-2}^2 - c_{n-3}\,(c_{n-1} - b_{n-1})}$$

$$\Delta q_k = -\frac{b_{n-1}(c_{n-1} - b_{n-1}) - b_n c_{n-2}}{c_{n-2}^2 - c_{n-3}\,(c_{n-1} - b_{n-1})}$$

The values of b_i's and c_i's are obtained from the recurrence relations

$$b_i = a_i - p_k\, b_{i-1} - q_k\, b_{i-2}, \quad i = 1, 2, \ldots, n,$$
$$c_i = b_i - p_k\, c_{i-1} - q_k\, c_{i-2}, \quad i = 1, 2, \ldots, n-1,$$

with $$c_0 = b_0 = a_0, \quad c_{-1} = b_{-1} = 0.$$

We can also obtain the values of b_i's and c_i's using the synthetic division method as given below :

$$\begin{array}{c|cccccc}
-p_k & a_0 & a_1 & a_2 & \ldots & a_{n-1} & a_n \\
 & & -p_k b_0 & -p_k b_1 & \ldots & -p_k b_{n-2} & -p_k b_{n-1} \\
-q_k & & & -q_k b_0 & \ldots & -q_k b_{n-3} & -q_k b_{n-2} \\
\hline
 & b_0 & b_1 & b_2 & \ldots & b_{n-1} & b_n \\
 & & -p_k c_0 & -p_k c_1 & \ldots & -p_k c_{n-2} & \\
 & & & -q_k c_0 & \ldots & -q_k c_{n-3} & \\
\hline
 & c_0 & c_1 & c_2 & \ldots & c_{n-1} &
\end{array}$$

where $b_0 = a_0$ and $c_0 = b_0 = a_0$.

We have

$$\lim_{k \to \infty} b_n = 0, \qquad \lim_{k \to \infty} b_{n-1} = 0,$$
$$\lim_{k \to \infty} p_k = p, \qquad \lim_{k \to \infty} q_k = q.$$

The order of this method is 2.

When p and q have been obtained to the desired accuracy we obtain the next quadratic factor from the *deflated polynomial*

$$Q_{n-2}(x) = b_0 x^{n-2} + b_1 x^{n-3} + \ldots + b_{n-3} x + b_{n-2}$$

which can be obtained from the first part of the above synthetic division method.

Laguerre method

Define $$A = -P_n'(x_k) \,/\, P_n(x_k),$$
$$B = A^2 - P_n''(x_k) \,/\, P_n(x_k).$$

Then, the method is given by

$$x_{k+1} = x_k + \frac{n}{A \pm \sqrt{(n-1)\,(nB - A^2)}}. \tag{1.31}$$

The values $P_n(x_k)$, $P_n'(x_k)$ and $P_n''(x_k)$ can be obtained using the synthetic division method. The sign in the denominator on the right hand side of (1.31) is taken as the sign of A to make the denominator largest in magnitude. The order of the method is 2.

Graeffe's Root Squaring method

This is a direct method and is used to find all the roots of a polynomial with real coefficients. The roots may be real and distinct, real and equal or complex. We separate the roots of the equation (1.27) by forming another equation, whose roots are very high powers of the roots of (1.27) with the help of root squaring process.

Let $\xi_1, \xi_2, \ldots, \xi_n$ be the roots of (1.27). Separating the even and odd powers of x in (1.27) and squaring we get

$$(a_0 x^n + a_2 x^{n-2} + \ldots)^2 = (a_1 x^{n-1} + a_3 x^{n-3} + \ldots)^2.$$

Simplifying, we obtain

$$a_0^2 x^{2n} - (a_1^2 - 2a_0 a_2) x^{2n-2} + \ldots + (-1)^n a_n^2 = 0.$$

Substituting $z = -x^2$, we get

$$b_0 z^n + b_1 z^{n-1} + \ldots + b_{n-1} z + b_n = 0 \tag{1.32}$$

which has roots $-\xi_1^2, -\xi_2^2, \ldots, -\xi_n^2$. The coefficients b_k's are obtained from :

$$\begin{array}{|cccccc}
a_0 & a_1 & a_2 & a_3 & \ldots & a_n \\
\hline
a_0^2 & a_1^2 & a_2^2 & a_3^2 & \ldots & a_n^2 \\
 & -2a_0a_2 & -2a_1a_3 & -2a_2a_4 & & \\
 & & +2a_0a_4 & +2a_1a_5 & \ldots & \\
 & & & \vdots & & \\
\hline
b_0 & b_1 & b_2 & b_3 & \ldots & b_n
\end{array}.$$

The $(k + 1)$th column in the above table is obtained as explained below:

The terms in each column alternate in sign starting with a positive sign. The first term is square of the $(k + 1)$th coefficient a_k. The second term is twice the product of the nearest neighbouring pair a_{k-1} and a_{k+1}. The next term is twice the product of the next neighbouring pair a_{k-2} and a_{k+2}. This procedure is continued until there are no available coefficients to form the cross products.

After repeating this procedure m times we obtain the equation

$$B_0 x^n + B_1 x^{n-1} + \ldots + B_{n-1} x + B_n = 0 \tag{1.33}$$

whose roots are $R_1, R_2, \ldots, R_n$, where

$$R_i = -\xi_i^{2m}, \quad i = 1, 2, \ldots, n.$$

If we assume

$$|\xi_1| > |\xi_2| > \ldots > |\xi_n|,$$

then $$|R_1| >> |R_2| >> \ldots >> |R_n|.$$

We obtain from (1.33)

$$|R_i| \approx \frac{|B_i|}{|B_{i-1}|} = |\xi_i|^{2^m}$$

or $$\log(|\xi_i|) = 2^{-m} [\log |B_i| - \log |B_{i-1}|].$$

This determines the magnitude of the roots and substitution in the original equation (1.27) will give the sign of the roots.

We stop the squaring process when another squaring process produces new coefficients that are almost the squares of the corresponding coefficients B_k's, *i.e.*, when the cross product terms become negligible in comparison to square terms.

After few squarings, if the magnitude of the coefficient B_k is half the square of the magnitude of the corresponding coefficient in the previous equation, then it indicates that ξ_k is a double root. We can find this double root by using the following procedure. We have

$$R_k \simeq -\frac{B_k}{B_{k-1}} \quad \text{and} \quad R_{k+1} \simeq -\frac{B_{k+1}}{B_k}$$

$$R_k R_{k+1} \simeq R_k^2 \simeq \left| \frac{B_{k+1}}{B_{k-1}} \right|$$

or

$$| R_k^2 | = | \xi_k |^{2(2^m)} = \left| \frac{B_{k+1}}{B_{k-1}} \right|.$$

This gives the magnitude of the double root. Substituting in the given equation, we can find its sign. This double root can also be found directly since R_k and R_{k+1} converge to the same root after sufficient squarings. Usually, this convergence to the double root is slow. By making use of the above observation, we can save a number of squarings.

If ξ_k and ξ_{k+1} form a complex pair, then this would cause the coefficients of x^{n-k} in the successive squarings to fluctuate both in magnitude and sign. If $\xi_k, \xi_{k+1} = \beta_k \exp(\pm i\phi_k)$ is the complex pair, then the coefficients would fluctuate in magnitude and sign by an amount $2\beta_k^m \cos(m\phi_k)$. A complex pair can be spotted by such an oscillation. For m sufficiently large, $2\beta_k$ can be determined from the relation

$$\beta_k^{2(2^m)} \simeq \left| \frac{B_{k+1}}{B_{k-1}} \right|$$

and ϕ is suitably determined from the relation

$$2\beta_k^m \cos(m\phi_k) \simeq \frac{B_{k+1}}{B_{k-1}}.$$

If the equation has only one complex pair, then we can first determine all the real roots. The complex pair can be written as $\xi_k, \xi_{k+1} = p \pm iq$. The sum of the roots then gives

$$\xi_1 + \xi_2 + \ldots + \xi_{k-1} + 2p + \xi_{k+2} + \ldots + \xi_n = -a_1.$$

This determines p. We also have $| \beta_k |^2 = p^2 + q^2$. Since $| \beta_k |$ is already determined, this equation gives q.

1.7 PROBLEMS AND SOLUTIONS

Bisection method

1.1 Find the interval in which the smallest positive root of the following equations lies :

(a) $\tan x + \tanh x = 0$

(b) $x^3 - x - 4 = 0$.

Determine the roots correct to two decimal places using the bisection method.

Solution

(a) Let $f(x) = \tan x + \tanh x$.

Note that $f(x)$ has no root in the first branch of $y = \tan x$, that is, in the interval $(0, \pi / 2)$. The root is in the next branch of $y = \tan x$, that is, in the interval $(\pi / 2, 3\pi / 2)$.

We have $f(1.6) = -33.31$, $f(2.0) = -1.22$,

$f(2.2) = -0.40$, $f(2.3) = -0.1391$, $f(2.4) = 0.0676$.

Therefore, the root lies in the interval (2.3, 2.4). The sequence of intervals using the bisection method (1.7) are obtained as

k	a_{k-1}	b_{k-1}	m_k	$f(m_k) f(a_{k-1})$
1	2.3	2.4	2.35	> 0
2	2.35	2.4	2.375	< 0
3	2.35	2.375	2.3625	> 0
4	2.3625	2.375	2.36875	< 0

After four iterations, we find that the root lies in the interval (2.3625, 2.36875). Hence, the approximate root is $m = 2.365625$. The root correct to two decimal places is 2.37.

(*b*) For $f(x) = x^3 - x - 4$, we find $f(0) = -4$, $f(1) = -4$, $f(2) = 2$.

Therefore, the root lies in the interval (1, 2). The sequence of intervals using the bisection method (1.7) is obtained as

k	a_{k-1}	b_{k-1}	m_k	$f(m_k) f(a_{k-1})$
1	1	2	1.5	> 0
2	1.5	2	1.75	> 0
3	1.75	2	1.875	< 0
4	1.75	1.875	1.8125	> 0
5	1.75	1.8125	1.78125	> 0
6	1.78125	1.8125	1.796875	< 0
7	1.78125	1.796875	1.7890625	> 0
8	1.7890625	1.796875	1.792969	> 0
9	1.792969	1.796875	1.794922	> 0
10	1.794922	1.796875	1.795898	> 0.

After 10 iterations, we find that the root lies in the interval (1.795898, 1.796875). Therefore, the approximate root is $m = 1.796387$. The root correct to two decimal places is 1.80.

Iterative Methods

1.2 Find the iterative methods based on the Newton-Raphson method for finding $\sqrt{N}$, $1/N$, $N^{1/3}$, where N is a positive real number. Apply the methods to $N = 18$ to obtain the results correct to two decimal places.

Solution

(*a*) Let $x = N^{1/2}$ or $x^2 = N$.

We have therefore $f(x) = x^2 - N$, $f'(x) = 2x$.

Using Newton-Raphson method (1.9), we obtain the iteration scheme

$$x_{n+1} = x_n - \frac{x_n^2 - N}{2x_n}, \quad n = 0, 1, \ldots$$

or

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{N}{x_n}\right), \quad n = 0, 1, \ldots$$

For $N = 18$ and $x_0 = 4$, we obtain the sequence of iterates

$x_1 = 4.25$, $x_2 = 4.2426$, $x_3 = 4.2426$, ...

The result correct to two decimal places is 4.24.

(*b*) Let $x = 1 / N$ or $1 / x = N$.

We have therefore

$$f(x) = (1 / x) - N, \quad f'(x) = -1 / x^2.$$

Using Newton-Raphson method (1.9), we obtain the iteration scheme

$$x_{n+1} = x_n - \frac{(1/x_n) - N}{(-1/x_n^2)}, \quad n = 0, 1, \ldots$$

or $\quad x_{n+1} = x_n(2 - Nx_n), \; n = 0, 1, \ldots$

For $N = 18$ and $x_0 = 0.1$, we obtain the sequence of iterates

$$x_1 = 0.02, \quad x_2 = 0.0328, \quad x_3 = 0.0462,$$
$$x_4 = 0.0540, \quad x_5 = 0.0555, \quad x_6 = 0.0556.$$

The result correct to two decimals is 0.06.

(*c*) Let $x = N^{1/3}$ or $x^3 = N$.

We have therefore $f(x) = x^3 - N, \quad f'(x) = 3x^2.$

Using the Newton-Raphson method (1.9) we get the iteration scheme

$$x_{n+1} = x_n - \frac{x_n^3 - N}{3x_n^2} = \frac{1}{3}\left(2x_n + \frac{N}{x_n^2}\right), \quad n = 0, 1, \ldots$$

For $N = 18$ and $x_0 = 2$, we obtain the sequence of iterates

$$x_1 = 2.8333, \quad x_2 = 2.6363,$$
$$x_3 = 2.6208, \quad x_4 = 2.6207.$$

The result correct to two decimals is 2.62.

1.3 Given the following equations :

(*i*) $x^4 - x - 10 = 0$, (*ii*) $x - e^{-x} = 0$

determine the initial approximations for finding the smallest positive root. Use these to find the root correct to three decimal places with the following methods:

(*a*) Secant method, (*b*) Regula-Falsi method,

(*c*) Newton-Raphson method.

Solution

(*i*) For $f(x) = x^4 - x - 10$, we find that

$$f(0) = -10, f(1) = -10, f(2) = 4.$$

Hence, the smallest positive root lies in the interval (1, 2).

The Secant method (1.8) gives the iteration scheme

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k, \quad k = 1, 2, \ldots$$

With $x_0 = 1$, $x_1 = 2$, we obtain the sequence of iterates

$$x_2 = 1.7143, \quad x_3 = 1.8385, \quad x_4 = 1.8578,$$
$$x_5 = 1.8556, \quad x_6 = 1.8556.$$

The root correct to three decimal places is 1.856.

The Regula-Falsi method (1.8) gives the iteration scheme

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k, \quad k = 1, 2, \ldots$$

and $\quad f_k f_{k-1} < 0.$

With $x_0 = 1$, $x_1 = 2$, we obtain the sequence of iterates

$$\begin{aligned}
&x_2 = 1.7143, && f(x_2) = -3.0776, && \xi \in (x_1, x_2),\\
&x_3 = 1.8385, && f(x_3) = -0.4135, && \xi \in (x_1, x_3),\\
&x_4 = 1.8536, && f(x_4) = -0.0487, && \xi \in (x_1, x_4),\\
&x_5 = 1.8554, && f(x_5) = -0.0045, && \xi \in (x_1, x_5),\\
&x_6 = 1.8556.
\end{aligned}$$

The root correct to three decimal places is 1.856.

The Newton-Raphson method (1.9) gives the iteration scheme

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, ...$$

With $x_0 = 2$, we obtain the sequence of iterates

$$x_1 = 1.8710, \quad x_2 = 1.8558, \quad x_3 = 1.8556.$$

Hence, the root correct to three decimal places is 1.856.

(*ii*) For $f(x) = x - e^{-x}$, we find that $f(0) = -1$, $f(1) = 0.6321$.

Therefore, the smallest positive root lies in the interval (0, 1). For $x_0 = 0$, $x_1 = 1$, the Secant method gives the sequence of iterates

$$x_2 = 0.6127, \quad x_3 = 0.5638, \quad x_4 = 0.5671, \quad x_5 = 0.5671.$$

For $x_0 = 0$, $x_1 = 1$, the Regula-Falsi method gives the sequence of iterates

$$\begin{aligned}
&x_2 = 0.6127, && f(x_2) = 0.0708, && \xi \in (x_0, x_2),\\
&x_3 = 0.5722, && f(x_3) = 0.0079, && \xi \in (x_0, x_3),\\
&x_4 = 0.5677, && f(x_4) = 0.0009, && \xi \in (x_0, x_4),\\
&x_5 = 0.5672, && f(x_5) = 0.00009.
\end{aligned}$$

For $x_0 = 1$, the Newton-Raphson method gives the sequence of iterates

$$x_1 = 0.5379, \quad x_2 = 0.5670, \quad x_3 = 0.5671.$$

Hence, the root correct to three decimals is 0.567.

1.4 Use the Chebyshev third order method with $f(x) = x^2 - a$ and with $f(x) = 1 - a / x^2$ to obtain the iteration method converging to $a^{1/2}$ in the form

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right) - \frac{1}{8x_k}\left(x_k - \frac{a}{x_k}\right)^2$$

and

$$x_{k+1} = \frac{1}{2} x_k \left(3 - \frac{x_k^2}{a}\right) + \frac{3}{8} x_k \left(1 - \frac{x_k^2}{a}\right)^2.$$

Perform two iterations with these methods to find the value of $\sqrt{6}$.

Solution

(*i*) Taking $\quad f(x) = x^2 - a, \quad f'(x) = 2x, \quad f''(x) = 2$

and using the Chebyshev third order method (1.10) we obtain on simplification

$$\begin{aligned}
x_{k+1} &= x_k - \frac{x_k^2 - a}{2x_k} - \frac{1}{2}\left(\frac{x_k^2 - a}{2x_k}\right)^2 \left(\frac{1}{x_k}\right)\\
&= \frac{1}{2}\left(x_k + \frac{a}{x_k}\right) - \frac{1}{8x_k}\left(x_k - \frac{a}{x_k}\right)^2, \quad k = 0, 1, ...
\end{aligned}$$

For $a = 6$ and $x_0 = 2$, we get $x_1 = 2.4375$, $x_2 = 2.4495$.

(*ii*) Taking $\quad f(x) = 1 - \dfrac{a}{x^2}, \quad f'(x) = \dfrac{2a}{x^3}, \quad f''(x) = -\dfrac{6a}{x^4}$

and using the Chebyshev third order method (1.10), we obtain

$$x_{k+1} = x_k - \frac{1}{2}\left(\frac{x_k^3}{a} - x_k\right) - \frac{1}{8}\left(\frac{x_k^3}{a} - x_k\right)^2 \left(-\frac{3}{x_k}\right)$$

$$= \frac{1}{2} x_k \left(3 - \frac{x_k^2}{a}\right) + \frac{3}{8} x_k \left(1 - \frac{x_k^2}{a}\right)^2, \quad k = 0, 1, ...$$

For $a = 6$ and $x_0 = 2$, we get $x_1 = 2.4167$, $x_2 = 2.4495$.

1.5 Perform two iterations using the sixth order method, to find a root of the following equations :

(*i*) $x^4 - x - 10 = 0$, $x_0 = 2.0$; (*ii*) $x - e^{-x} = 0$, $x_0 = 1.0$.

Solution

(*i*) *First iteration*

$$f(x) = x^4 - x - 10, f'(x) = 4x^3 - 1,$$
$$x_0 = 2, f(x_0) = 4, f'(x_0) = 31, w_0 = 1.870968, f(w_0) = 0.382681,$$
$$z_0 = 1.855519, f(z_0) = -0.001609, x_1 = 1.855585.$$

Second iteration

$$f(x_1) = 0.000012, f'(x_1) = 24.556569, w_1 = 1.855585,$$
$$f(w_1) = 0.000012, z_1 = 1.855585, x_2 = 1.855585.$$

(*ii*) *First iteration*

$$f(x) = x - e^{-x}, f'(x) = 1 + e^{-x},$$
$$x_0 = 1.0, f(x_0) = 0.632121, f'(x_0) = 1.367879,$$
$$w_0 = 0.537882, f(w_0) = -0.046102,$$
$$z_0 = 0.567427, f(z_0) = 0.000445, x_1 = 0.567141.$$

Second iteration

$$f(x_1) = -0.000004, f'(x_1) = 1.567145, w_1 = 0.567144,$$
$$f(w_1) = 0.000001, z_1 = 0.567144, x_2 = 0.567144.$$

1.6 Perform 2 iterations with the Müller method (Eqs. (1.13), (1.14)) for the following equations :

(*a*) $x^3 - \frac{1}{2} = 0$, $x_0 = 0$, $x_1 = 1$, $x_2 = \frac{1}{2}$,

(*b*) $\log_{10} x - x + 3 = 0$, $x_0 = \frac{1}{4}$, $x_1 = \frac{1}{2}$, $x_2 = 1$.

Solution

(*a*) Using the Müller method (1.13) for $f(x) = x^3 - \frac{1}{2}$, we obtain

First iteration

$$x_0 = 0, x_1 = 1, x_2 = 0.5,$$
$$f_0 = -0.5, f_1 = 0.5, f_2 = -0.375,$$
$$h_2 = x_2 - x_1 = -0.5, h_1 = x_1 - x_0 = 1.0,$$
$$\lambda_2 = h_2 / h_1 = -0.5, \delta_2 = 1 + \lambda_2 = 0.5,$$

$$g_2 = \lambda_2^2 f_0 - \delta_2^2 f_1 + (\delta_2 + \lambda_2) f_2 = -0.25,$$
$$c_2 = \lambda_2(\lambda_2 f_0 - \delta_2 f_1 + f_2) = 0.1875,$$
$$\lambda_3 = \frac{-2\delta_2 f_2}{g_2 \pm \sqrt{g_2^2 - 4\delta_2 f_2 c_2}}.$$

Taking minus sign in the denominator (sign of g_2) we obtain

$$\lambda_3 = -0.5352,$$
$$x_3 = x_2 + (x_2 - x_1)\lambda_3 = 0.7676.$$

Second iteration

$$x_0 = 1,\ x_1 = 0.5,\ x_2 = 0.7676,$$
$$f_0 = 0.5,\ f_1 = -0.375,\ f_2 = -0.0477,$$
$$h_2 = 0.2676,\ h_1 = -0.5,\ \lambda_2 = -0.5352,$$
$$\delta_2 = 0.4648,\ g_2 = 0.2276,$$
$$c_2 = 0.0755,\ \lambda_3 = 0.0945,$$
$$x_3 = 0.7929.$$

Alternative

First iteration

$$x_0 = 0,\ x_1 = 1,\ x_2 = 0.5,\ x_2 - x_1 = -0.5,\ x_2 - x_0 = 0.5,\ x_1 - x_0 = 1.0,$$
$$f_0 = -0.5,\ f_1 = 0.5,\ f_2 = -0.375,$$
$$D = (x_2 - x_1)(x_2 - x_0)(x_1 - x_0) = -0.25,\ a_2 = f_2 = -0.375,$$
$$a_1 = \{[(x_2 - x_0)^2 (f_2 - f_1) - (x_2 - x_1)^2 (f_2 - f_0)] \ / \ D\} = 1,$$
$$a_0 = \{[(x_2 - x_0)(f_2 - f_1) - (x_2 - x_1)(f_2 - f_0)] \ / \ D\} = 1.5,$$
$$x_3 = x_2 - \frac{2a_2}{a_1 + \sqrt{a_1^2 - 4a_0 a_2}} = 0.7676.$$

Second iteration

$$x_0 = 1,\ x_1 = 0.5,\ x_2 = 0.7676,\ x_2 - x_1 = 0.2676,\ x_2 - x_0 = -0.2324,$$
$$x_1 - x_0 = -0.5,\ f_0 = 0.5,\ f_1 = -0.375,\ f_2 = -0.0477,$$
$$D = 0.0311,\ a_2 = -0.0477,\ a_1 = 1.8295,\ a_0 = 2.2669,\ x_3 = 0.7929.$$

(*b*) Using Müller method (1.13) with $f(x) = \log_{10} x - x + 3$, we obtain

First iteration

$$x_0 = 0.25,\ x_1 = 0.5,\ x_2 = 1.0,$$
$$f_0 = 2.147940,\ f_1 = 2.198970,\ f_2 = 2.0,$$
$$h_2 = 0.5,\ h_1 = 0.25,\quad \lambda_2 = 2.0,$$
$$\delta_2 = 3.0,\quad g_2 = -1.198970,\ c_2 = -0.602060,$$
$$\lambda_3 = 2.314450,\ x_3 = 2.157225,$$

Second iteration

$$x_0 = 0.5,\ x_1 = 1.0,\ x_2 = 2.157225,$$
$$f_0 = 2.198970,\ f_1 = 2.0,\ f_2 = 1.176670,$$
$$h_1 = 0.5,\quad h_2 = 1.157225,$$
$$\lambda_2 = 2.314450,\ \delta_2 = 3.314450,$$
$$g_2 = -3.568624,\ c_2 = -0.839738,$$
$$\lambda_3 = 0.901587,\ x_3 = 3.200564.$$

Alternative

First iteration

$$x_0 = 0.25,\ x_1 = 0.5,\ x_2 = 1.0,\ x_2 - x_1 = 0.5,\ x_2 - x_0 = 0.75,\ x_1 - x_0 = 0.25,$$
$$f_0 = 2.147940,\ f_1 = 2.198970,\ f_2 = 2.0,\ D = 0.09375,\ a_2 = 2.0,$$
$$a_1 = -0.799313,\ a_0 = -0.802747,$$
$$x_3 = x_2 - \frac{2a_2}{a_1 - \sqrt{a_1^2 - 4a_0a_2}} = 2.157225.$$

Second iteration

$$x_0 = 0.5,\ x_1 = 1.0,\ x_2 = 2.157225,\ x_2 - x_1 = 1.157225,$$
$$x_2 - x_0 = 1.657225,\ x_1 - x_0 = 0.5,\ f_0 = 2.198970,$$
$$f_1 = 2.0,\ f_2 = 1.176670,\ D = 0.958891,\ a_2 = 1.176670,$$
$$a_1 = -0.930404,\ a_0 = -0.189189,\ x_3 = 3.200564.$$

1.7 The equation $x = f(x)$ is solved by the iteration method $x_{k+1} = f(x_k)$, and a solution is wanted with a maximum error not greater than 0.5×10^{-4}. The first and second iterates were computed as : $x_1 = 0.50000$ and $x_2 = 0.52661$. How many iterations must be performed further, if it is known that $|f'(x)| \le 0.53$ for all values of x.

Solution

For the general iteration method $x_{k+1} = f(x_k)$, the error equation satisfies

$$|\varepsilon_{n+1}| \le c\,|\varepsilon_n|, \qquad (\text{where } c = |f'(\xi)| \text{ and } 0 < c < 1).$$

Hence,
$$\begin{aligned} |\xi - x_{n+1}| &\le c\,|\xi - x_n| \\ &= c\,|\xi - x_n + x_{n+1} - x_{n+1}| \\ &\le c\,|\xi - x_{n+1}| + c\,|x_{n+1} - x_n|. \end{aligned}$$

Thus, we get

$$|\xi - x_{n+1}| \le \frac{c}{1-c}\,|x_{n+1} - x_n|,\ n = 0, 1, \ldots$$

For $n = 1$, we have

$$|\xi - x_2| \le \frac{c}{1-c}\,|x_2 - x_1| = 0.03001$$

where, we have used $c = 0.53$.

We also have

$$|\xi - x_{n+2}| \le c^n\,|\xi - x_2| \le (0.53)^n\,(0.03001).$$

Now choose n such that

$$(0.53)^n\,(0.03001) \le 5 \times 10^{-5}.$$

We find $n \ge 11$.

1.8 A root of the equation $f(x) = x - F(x) = 0$ can often be determined by combining the iteration method with Regula-Falsi :

(*i*) With a given approximate value x_0, we compute

$$x_1 = F(x_0), \quad x_2 = F(x_1).$$

(*ii*) Observing that $f(x_0) = x_0 - x_1$ and $f(x_1) = x_1 - x_2$, we find a better approximation x' using Regula-Falsi method on the points $(x_0, x_0 - x_1)$ and $(x_1, x_1 - x_2)$.

(*iii*) The last x' is taken as a new x_0 and we start from (*i*) all over again.

Compute the smallest root of the equation $x - 5 \log_e x = 0$ with an error less than 0.5×10^{-4} starting with $x_0 = 1.3$.(Inst. Tech. Stockholm, Sweden, BIT 6 (1966), 176)

Solution

From $x = F(x)$, we have $F(x) = 5 \log_e x$.

First iteration

$$x_0 = 1.3, \quad x_1 = F(x_0) = 1.311821,$$
$$x_2 = F(x_1) = 1.357081,$$
$$f_0 = x_0 - x_1 = -0.011821, \quad f_1 = x_1 - x_2 = -0.045260.$$

Using Regula-Falsi method (1.8) on the points (1.3, – 0.011821) and (1.311821, – 0.045260), we obtain

$$x' = 1.295821.$$

Second iteration

$$x_0 = x' = 1.295821,$$
$$x_1 = 1.295722, f_0 = 0.000099,$$
$$x_2 = 1.295340, f_1 = 0.000382.$$

Using Regula-Falsi method (1.8) on the points (1.295821, 0.000099) and (1.295722, 0.000382) we get

$$x'' = 1.295854$$

which is the required root and satisfies the given error criteria.

1.9 The root of the equation $x = (1 / 2) + \sin x$ by using the iteration method

$$x_{k+1} = \frac{1}{2} + \sin x_k, \quad x_0 = 1$$

correct to six decimals is $x = 1.497300$. Determine the number of iteration steps required to reach the root by linear iteration. If the Aitken Δ^2-process is used after three approximations are available, how many iterations are required.

Solution

We have $\xi = 1.497300$ and $x_0 = 1$, $g(x) = (1 / 2) + \sin x$. The linear iteration method satisfies the error relation

$$| \varepsilon_n | < c^n | \varepsilon_0 |.$$

We now have $c = | g'(\xi) | = | \cos \xi | = 0.073430$

and $\varepsilon_0 = \xi - x_0 = 0.497300.$

Choose n such that $c^n | \varepsilon_0 | \le 5 \times 10^{-7}$ or

$$(0.07343)^n (0.4973) \le 5 \times 10^{-7}$$

which gives $n \ge 6$.

Starting with $x_0 = 1$, we obtain from the linear iteration formula

$$x_{k+1} = (1 / 2) + \sin x_k = g(x_k), k = 0, 1, ...$$
$$x_1 = 1.34147098, \quad x_2 = 1.47381998.$$

Using Aitken Δ^2-process (1.17) we get

$$x_0^* = 1.55758094,$$
$$x_1^* = g(x_0^*) = 1.49991268,$$
$$x_2^* = g(x_1^*) = 1.49748881.$$

Using Aitken Δ^2-process (1.17) we get

$$x_0^{**} = 1.49738246,$$
$$x_1^{**} = g(x_0^{**}) = 1.49730641,$$
$$x_2^{**} = g(x_1^{**}) = 1.49730083.$$

Using Aitken Δ^2-process (1.17) we get

$$x_0^{***} = 1.49730039,$$
$$x_1^{***} = g(x_0^{***}) = 1.49730039,$$
$$x_2^{***} = g(x_1^{***}) = 1.49730039.$$

The Aitken Δ^2-process gives the root as $\xi = 1.49730039$, which satisfies the given error criteria. Hence, three such iterations are needed in this case.

1.10 (*a*) Show that the equation $\log_e x = x^2 - 1$ has exactly two real roots, $\alpha_1 = 0.45$ and $\alpha_2 = 1$.

(*b*) Determine for which initial approximation x_0, the iteration

$$x_{n+1} = \sqrt{1 + \log_e x_n}$$

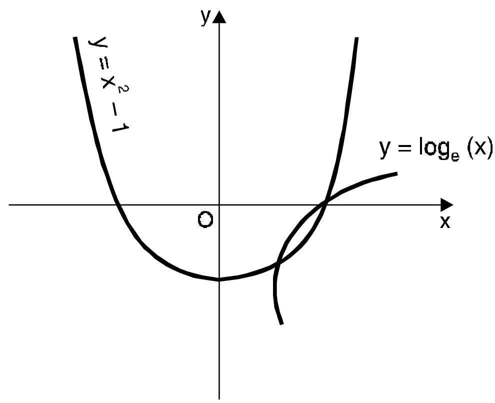
converges to α_1 or α_2. (Uppsala Univ., Sweden, BIT 10 (1970), 115).

Solution

(*a*) From the equation $\log_e x = x^2 - 1$, we find that the roots are the points of intersection of the curves

$$y = \log_e x, \quad \text{and} \quad y = x^2 - 1.$$

Since the curves intersect exactly at two points $x = 0.45$ and $x = 1$, the equation has exactly two roots $\alpha_1 = 0.45$ and $\alpha_2 = 1$.



(*b*) We write the given iteration formula as

$$x_{n+1} = g(x_n)$$

where $$g(x) = \sqrt{1 + \log_e x}.$$

We have $$g'(x) = \frac{1}{2x\sqrt{1 + \log_e x}}.$$

For convergence, we require $| g'(x) | < 1$. We find that for

$$x_0 < \alpha_1, \quad | g'(x) | > 1, \text{ hence no convergence,}$$
$$x_0 < \alpha_2, \quad | g'(x) | < 1, \text{ hence converges to } \alpha_2.$$

For $x_0 > x^*$, where x^* is a root of $4x^2(1 + \log_e x) - 1 = 0$, $| g'(x) | < 1$, hence the root converges to α_2.

1.11 If an attempt is made to solve the equation $x = 1.4 \cos x$ by using the iteration formula

$$x_{n+1} = 1.4 \cos x_n$$

it is found that for large n, x_n alternates between the two values A and B.

(*i*) Calculate A and B correct to three decimal places.

(*ii*) Calculate the correct solution of the equation to 4 decimal places.

(Lund Univ., BIT 17 (1977), 115)

Solution

Using the given iteration formula and starting with $x_0 = 0$, we obtain the sequence of iterates

$$x_1 = 1.4, \; x_2 = 0.2380, \; x_3 = 1.3605.$$

For $x_k > 0.79$ (approx.), the condition for convergence, $| 1.4 \sin x_k | < 1$, is violated. However, $x_k = 1.3$ when substituted in the given formula gives $x_{k+1} = 0.374$, that is bringing back to a value closer to the other end. After 28 iterations, we find

$$x_{28} = 0.3615, x_{29} = 1.3095, x_{30} = 0.3616.$$

Hence, the root alternates between two values $A = 0.362$ and $B = 1.309$.

Using the Newton-Raphson method (1.9) with the starting value $x_0 = 0$, we obtain

$$x_1 = 1.4, \quad x_2 = 0.91167, \quad x_3 = 0.88591,$$
$$x_4 = 0.88577, \quad x_5 = 0.88577.$$

Hence, the root correct to four decimal places is 0.8858.

1.12 We consider the multipoint iteration method

$$x_{k+1} = x_k - \alpha \frac{f(x_k)}{f'(x_k - \beta f(x_k)/f'(x_k))}$$

where α and β are arbitrary parameters, for solving the equation $f(x) = 0$. Determine α and β such that the multipoint method is of order as high as possible for finding ξ, a simple root of $f(x) = 0$.

Solution

We have

$$\frac{f(x_k)}{f'(x_k)} = \frac{f(\xi + \varepsilon_k)}{f'(\xi + \varepsilon_k)} = \left[\varepsilon_k + \frac{c_2}{2}\varepsilon_k^2 + \ldots\right][1 + c_2\varepsilon_k + \ldots]^{-1}$$
$$= \varepsilon_k - \frac{1}{2} c_2 \varepsilon_k^2 + O(\varepsilon_k^3)$$

where $c_i = f^{(i)}(\xi)/f'(\xi)$.

We also have

$$f'\left(x_k - \beta \frac{f(x_k)}{f'(x_k)}\right) = f'\left((\xi) + (1-\beta)\,\varepsilon_k + \frac{1}{2}\beta c_2 \varepsilon_k^2 + O(\varepsilon_k^3)\right)$$
$$= f'(\xi) + \left[(1-\beta)\,\varepsilon_k + \frac{1}{2}\beta c_2\varepsilon_k^2 + \ldots\right] f''(\xi)$$
$$+ \frac{1}{2}[(1-\beta)^2\varepsilon_k^2 + \ldots]\, f'''(\xi) + \ldots$$

Substituting these expressions in the given formula and simplifying, we obtain the error equation as

$$\varepsilon_{k+1} = \varepsilon_k - \alpha\left[\varepsilon_k + \frac{c_2}{2}\varepsilon_k^2 + \ldots\right]\left[1 + (1-\beta)\,c_2\,\varepsilon_k + \ldots\right]^{-1}$$
$$= (1-\alpha)\varepsilon_k - \alpha\left[\frac{1}{2} - (1-\beta)\right]c_2\varepsilon_k^2 + O(\varepsilon_k^3),$$

Thus, for $\alpha = 1, \beta \neq 1/2$, we have second order methods and for $\alpha = 1, \beta = 1/2$, we have a third order method.

1.13 The equation

$$2e^{-x} = \frac{1}{x+2} + \frac{1}{x+1}$$

has two roots greater than -1. Calculate these roots correct to five decimal places.

(Inst. Tech., Lund, Sweden, BIT 21 (1981), 136)

Solution

From $f(x) = 2e^{-x} - \dfrac{1}{x+2} - \dfrac{1}{x+1}$ we find that

$$f(-0.8) = -1.38, f(0) = 0.5, f(1.0) = -0.0976.$$

Hence, the two roots of $f(x) = 0$ which are greater than -1 lie in the intervals $(-0.8, 0)$ and $(0, 1)$. We use Newton-Raphson method (1.9) to find these roots. We have

$$f(x) = 2e^{-x} - \frac{1}{x+2} - \frac{1}{x+1}$$

$$f'(x) = -2e^{-x} + \frac{1}{(x+2)^2} + \frac{1}{(x+1)^2}$$

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

First root

Starting with $x_0 = -0.6$, we obtain the sequence of iterates as

$$x_1 = -0.737984, \quad x_2 = -0.699338,$$
$$x_3 = -0.690163, \quad x_4 = -0.689753, \quad x_5 = -0.689752.$$

Hence, the root correct to five decimals is -0.68975.

Second root

Starting with $x_0 = 0.8$, we obtain the sequence of iterates

$$x_1 = 0.769640, \quad x_2 = 0.770091, \quad x_3 = 0.770091.$$

Hence, the root correct to five decimals is 0.77009.

1.14 Find the positive root of the equation

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} e^{0.3x}$$

correct to five decimal places.

(Royal Inst. Tech. Stockholm, Sweden, BIT 21 (1981), 242)

Solution

From $$f(x) = e^x - 1 - x - \frac{x^2}{2} - \frac{x^3}{6} e^{0.3x}$$

we find $$f(0) = 0, \quad f(1) = -0.0067, \quad f(2) = -0.0404, \quad f(3) = 0.5173.$$

Hence, the positive root lies in the interval $(2, 3)$. Starting with $x_0 = 2.5$ and using the Newton-Raphson method (1.9)

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

we obtain the sequence of iterates

$$x_1 = 2.392307, \quad x_2 = 2.364986, \quad x_3 = 2.363382,$$
$$x_4 = 2.363376, \quad x_5 = 2.363376.$$

Hence, the root correct to five decimals is $2.363376 \pm 0.5 \times 10^{-6}$.

1.15 Assuming that Δx, in the Taylor expansion of $f(x + \Delta x)$, can be approximated by $a_1 f(x_0) + a_2 f^2(x_0) + a_3 f^3(x_0) + \ldots$, where $a_1, a_2, \ldots$ are arbitrary parameters to be determined, derive the Chebyshev methods of third and fourth orders for finding a root of $f(x) = 0$.

Solution

We have

$$f(x_0 + \Delta x) = f(x_0) + \Delta x\, f'(x_0) + \frac{1}{2}(\Delta x)^2 f''(x_0) + \frac{1}{6}(\Delta x)^3 f'''(x_0) + \ldots \equiv 0$$

Substituting

$$\Delta x = a_1 f(x_0) + a_2 f^2(x_0) + a_3 f^3(x_0) + \ldots$$

in the above expression and simplifying, we get

$$[1 + a_1 f'(x_0)]\, f(x_0) + \left[a_2 f'(x_0) + \frac{1}{2} a_1^2 f''(x_0)\right] f^2(x_0) + \left[a_3 f'(x_0) + a_1 a_2 f''(x_0) + \frac{1}{6} a_1^3 f'''(x_0)\right] f^3(x_0) + \ldots \equiv 0$$

Equating the coefficients of various powers of $f(x_0)$ to zero, we get

$$1 + a_1 f'(x_0) = 0 \tag{1.34}$$

$$a_2 f'(x_0) + \frac{1}{2} a_1^2 f''(x_0) = 0 \tag{1.35}$$

$$a_3 f'(x_0) + a_1 a_2 f''(x_0) + \frac{1}{6} a_1^3 f'''(x_0) = 0. \tag{1.36}$$

Solving for a_1, a_2 from (1.34) and (1.35), we obtain

$$a_1 = -\frac{1}{f'(x_0)}, \quad a_2 = -\frac{1}{2} \frac{f''(x_0)}{[f'(x_0)]^3},$$

$$\Delta x = -\frac{f(x_0)}{f'(x_0)} - \frac{1}{2} \frac{f''(x_0)}{[f'(x_0)]^3} f^2(x_0),$$

$$x_1 = x + \Delta x$$

which is the Chebyshev third order method (1.10).

Solving the equations (1.34), (1.35) and (1.36), we find the same values for a_1 and a_2 as given above and

$$a_3 = -\frac{[f''(x_0)]^2}{2[f'(x_0)]^5} + \frac{1}{6} \frac{f'''(x_0)}{[f'(x_0)]^4}.$$

Hence, $$\Delta x = -\frac{f(x_0)}{f'(x_0)} - \frac{1}{2} \frac{f''(x_0) f^2(x_0)}{[f'(x_0)]^3} + \left[\frac{1}{6} \frac{f'''(x_0)}{f'(x_0)} - \frac{1}{2} \left\{\frac{f''(x_0)}{f'(x_0)}\right\}^2\right] \left[\frac{f(x_0)}{f'(x_0)}\right]^3$$

and $$x_1 = x_0 + \Delta x$$

which is the Chebyshev fourth order method.

1.16 (*a*) Newton-Raphson's method for solving the equation $f(x) = c$, where c is a real valued constant, is applied to the function

$$f(x) = \begin{cases} \cos x, & \text{when } |x| \le 1 \\ \cos x + (x^2 - 1)^2, & \text{when } |x| \ge 1, \end{cases}.$$

For which c is $x_n = (-1)^n$, when $x_0 = 1$ and the calculation is carried out with no error ?

(*b*) Even in high precision arithmetics, say 10 decimals, the convergence is troublesome. Explain ? (Uppsala Univ., Sweden, BIT 24 (1984), 129)

Solution

(*a*) When we apply the Newton-Raphson method (1.9) to the equation $f(x) = c$, we get the iteration scheme

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

Starting with $x_0 = 1$, we obtain

$$x_1 = 1 - \frac{\cos 1 - c}{-\sin 1} = -1$$

which gives $c = \cos 1 + 2 \sin 1$.

With this value of c, we obtain

$$x_2 = 1, \quad x_3 = -1, \ldots$$

and hence $x_n = (-1)^n$.

(*b*) Since $f'(x) = 0$, between x_0 and the roots and also at $x = 0$, the convergence will be poor inspite of high-precision arithmetic.

1.17 The equation $f(x) = 0$, where

$$f(x) = 0.1 - x + \frac{x^2}{(2!)^2} - \frac{x^3}{(3!)^2} + \frac{x^4}{(4!)^2} - \cdots$$

has one root in the interval (0, 1). Calculate this root correct to 5 decimals.

(Inst. Tech., Linköping, Sweden, BIT 24 (1984), 258)

Solution

We use the Newton-Raphson method (1.9)

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

where $$f(x) = 0.1 - x + \frac{x^2}{4} - \frac{x^3}{36} + \frac{x^4}{576} - \frac{x^5}{14400} + \ldots$$

$$f'(x) = -1 + \frac{x}{2} - \frac{x^2}{12} + \frac{x^3}{144} - \frac{x^4}{2880} + \ldots$$

With $x_0 = 0.2$, we obtain the sequence of iterates

$$x_1 = 0.100120, \quad x_2 = 0.102600, \quad x_3 = 0.102602.$$

Hence, the root correct to 5 decimals is 0.10260.

1.18 Show that the equation

$$f(x) = \cos\left(\frac{\pi(x+1)}{8}\right) + 0.148x - 0.9062 = 0$$

has one root in the interval (– 1, 0) and one in (0, 1). Calculate the negative root correct to 4 decimals. (Inst. Tech., Lyngby, Denmark, BIT 25 (1985), 299)

Solution

We have from the given function

$$f(-1) = -0.0542, \quad f(0) = 0.0177, \quad f(1) = -0.0511.$$

Hence, one root lies in the interval (– 1, 0) and one root in the interval (0, 1). To obtain the negative root, we use the Newton-Raphson method (1.9)

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

where $$f'(x) = -\left(\frac{\pi}{8}\right)\sin\left(\frac{\pi(x+1)}{8}\right) + 0.148,$$

With $x_0 = -0.5$, we obtain the following sequence of iterates :

$$x_1 = -0.508199, \quad x_2 = -0.508129, \quad x_3 = -0.508129.$$

Hence, the root correct to four decimals is – 0.5081.

1.19 The equation $x = 0.2 + 0.4 \sin (x/b)$, where b is a parameter, has one solution near $x = 0.3$. The parameter is known only with some uncertainty : $b = 1.2 \pm 0.05$. Calculate the root with an accuracy reasonable with respect to the uncertainty of b.

(Royal Inst. Tech. Stockholm, Sweden, BIT 26 (1986), 398)

Solution

Taking $b = 1.2$, we write the iteration scheme in the form

$$x_{n+1} = 0.2 + 0.4 \sin\left(\frac{x_n}{1.2}\right), \; n = 0, 1, ...$$

Starting with $x_0 = 0.3$, we obtain

$$x_1 = 0.298962, \quad x_2 = 0.298626, \quad x_3 = 0.298518.$$

Hence, the root correct to three decimals is 0.299.

1.20 Find all positive roots to the equation

$$10 \int_0^x e^{-x^2} \, dt = 1$$

with six correct decimals. (Uppsala Univ., Sweden, BIT 27 (1987), 129)

Solution

We have from the function

$$f(x) = 10xe^{-x^2} - 1,$$

$$f(0) = -1, \quad f(1) = 2.6788, \quad f(2) = -0.6337,$$

and $$f(a) < 0 \quad \text{for} \quad a > 2.$$

Hence, the given equation $f(x) = 0$ has two positive roots, one in the interval (0, 1), and the other in the interval (1, 2).

We use the Newton-Raphson method (1.9)

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

where $$f'(x) = 10(1 - 2x^2)\, e^{-x^2}.$$

With $x_0 = 0.1$, we obtain the following sequence of iterates

$$x_1 = 0.10102553, \quad x_2 = 0.10102585, \quad x_3 = 0.10102585.$$

Hence, the root correct to six decimals is 0.101026.

With $x_0 = 1.6$, we obtain the following sequence of iterates

$$x_1 = 1.67437337, \quad x_2 = 1.67960443,$$
$$x_3 = 1.67963061, \quad x_4 = 1.67963061.$$

Hence, the root correct to six decimals is 1.679631.

1.21 Find all the roots of $\cos x - x^2 - x = 0$ to five decimal places.

(Lund Univ., Sweden, BIT 27 (1987), 285)

Solution

For $f(x) = \cos x - x^2 - x$, we have

$$f(a) < 0 \quad \text{for} \quad a < -2, \quad f(-2) = -2.4161,$$
$$f(-1) = 0.5403, f(0) = 1.0, \quad f(1) = -1.4597,$$
$$f(b) < 0 \quad \text{for} \quad b > 1.$$

Hence, $f(x) = 0$ has a real root in the interval $(-2, -1)$ and another root in the interval $(0, 1)$.

We use the Newton-Raphson method (1.9)

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

where $\qquad f'(x) = -(\sin x + 2x + 1).$

Starting with $x_0 = 0.5$ and $x_0 = -1.5$ we obtain the following sequences of iterates :

$$\begin{aligned} x_0 &= 0.5, & x_0 &= -1.5, \\ x_1 &= 0.55145650, & x_1 &= -1.27338985, \\ x_2 &= 0.55001049, & x_2 &= -1.25137907, \\ x_3 &= 0.55000935, & x_3 &= -1.25115186, \\ & & x_4 &= -1.25115184. \end{aligned}$$

Hence, the roots correct to five decimals are 0.55001 and – 1.25115.

1.22 Find a catenary $y = c \cosh((x - a)/c)$ passing through the points (1, 1) and (2, 3).

(Royal Inst. Tech., Stockholm, Sweden, BIT 29 (1989), 375)

Solution

Since the catenary $y = c \cosh((x - a)/c)$ passes through the points (1, 1) and (2, 3), we have

$$c \cosh[(1 - a)/c] = 1$$
$$c \cosh[(2 - a)/c] = 3$$

which can be rewritten as

$$a = 1 - c \cosh^{-1}(1/c), \quad c = \frac{2 - a}{\cosh^{-1}(3/c)}.$$

On eliminating a from the above equations, we get

$$c = \frac{1 + c \cosh^{-1}(1/c)}{\cosh^{-1}(3/c)} = g(c).$$

Define $f(c) = c - g(c)$. We find that, $f(0.5) = -0.1693$, $f(1.0) = 0.4327$. There is a root of $f(c) = 0$ in (0.5, 1.0).

Using the iteration scheme

$$c_{n+1} = g(c_n), \; n = 0, 1, \ldots$$

with $c_0 = 0.5$, we obtain the sequence of iterates as

$$\begin{aligned} c_1 &= 0.66931131, & c_2 &= 0.75236778, & c_3 &= 0.77411374, \\ c_4 &= 0.77699764, & c_5 &= 0.77727732, & c_6 &= 0.77730310, \end{aligned}$$

$$c_7 = 0.77730547, \quad c_8 = 0.77730568, \quad c_9 = 0.77730570.$$

With the value $c = 0.7773057$, we get $a = 0.42482219$.

1.23 The factorial function n ! was first only defined for positive integers n or 0. For reasonably great values of n, a good approximation of n ! is $f(n)$ where

$$f(x) = (2\pi)^{1/2} x^{x + 1/2} e^{-x} \left(1 + \frac{1}{12x} + \frac{1}{288x^2}\right).$$

Calculate x to four decimals so that $f(x) = 1000$.

(Lund Univ., Sweden, BIT 24 (1984), 257)

Solution

Here, the problem is to find x such that

$$(2\pi)^{1/2} x^{x + 1/2} e^{-x} \left(1 + \frac{1}{12x} + \frac{1}{288x^2}\right) = 1000.$$

Taking logarithms on both sides, we get

$$f(x) = \frac{1}{2} \ln (2\pi) + \left(x + \frac{1}{2}\right) \ln x - x + \ln \left(1 + \frac{1}{12x} + \frac{1}{288x^2}\right) - 3 \ln 10 = 0$$

and
$$f'(x) = \frac{1}{2x} + \ln x - \frac{2(1 + 12x)}{x + 24x^2 + 288x^3}.$$

Use the Newton-Raphson method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots$$

Since 6 ! = 720, we take the initial approximation as $x_0 = 6.0$. We get

$$x_0 = 6.0, \qquad f(x_0) = -0.328492, \qquad f'(x_0) = 1.872778,$$
$$x_1 = 6.175404, \qquad f(x_1) = 0.002342, \qquad f'(x_1) = 1.899356,$$
$$x_2 = 6.174171, \qquad f(x_2) = 0.00000045.$$

Hence, the root correct to four decimal places is 6.1742.

Multiple roots

1.24 Apply the Newton-Raphson method with $x_0 = 0.8$ to the equation

$$f(x) = x^3 - x^2 - x + 1 = 0$$

and verify that the convergence is only of first order. Then, apply the Newton-Raphson method

$$x_{n+1} = x_n - m \frac{f(x_n)}{f'(x_n)}$$

with $m = 2$ and verify that the convergence is of second order.

Solution

Using Newton-Raphson method (1.9), we obtain the iteration scheme

$$x_{n+1} = x_n - \frac{x_n^3 - x_n^2 - x_n + 1}{3x_n^2 - 2x_n - 1}, \quad n = 0, 1, \ldots$$

Starting with $x_0 = 0.8$, we obtain

$$x_1 = 0.905882, \quad x_2 = 0.954132, \quad x_3 = 0.977338, \quad x_4 = 0.988734.$$

Since the exact root is 1, we have

$$| \varepsilon_0 | = | \xi - x_0 | = 0.2 = 0.2 \times 10^0$$
$$| \varepsilon_1 | = | \xi - x_1 | = 0.094118 \approx 0.94 \times 10^{-1}$$
$$| \varepsilon_2 | = | \xi - x_2 | = 0.045868 \approx 0.46 \times 10^{-1}$$
$$| \varepsilon_3 | = | \xi - x_3 | = 0.022662 \approx 0.22 \times 10^{-1}$$
$$| \varepsilon_4 | = | \xi - x_4 | = 0.011266 \approx 0.11 \times 10^{-1}$$

which shows only linear rate of convergence. Using the modified Newton-Raphson method

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

we obtain the sequence of iterates

$$x_0 = 0.8, \qquad x_1 = 1.011765$$
$$x_2 = 1.000034, \qquad x_3 = 1.000000.$$

We now have

$$| \varepsilon_0 | = | \xi - x_0 | = 0.2 \approx 0.2 \times 10^0$$
$$| \varepsilon_1 | = | \xi - x_1 | = 0.011765 \approx 0.12 \times 10^{-1}$$
$$| \varepsilon_2 | = | \xi - x_2 | = 0.000034 \approx 0.34 \times 10^{-4}$$

which verifies the second order convergence.

1.25 The multiple root ξ of multiplicity two of the equation $f(x) = 0$ is to be determined. We consider the multipoint method

$$x_{k+1} = x_k - \frac{f(x_k + 2f(x_k)/f'(x_k))}{2f'(x_k)}.$$

Show that the iteration method has third order rate of convergence. Hence, solve the equation

$$9x^4 + 30x^3 + 34x^2 + 30x + 25 = 0 \quad \text{with} \quad x_0 = -1.4$$

correct to three decimals.

Solution

Since the root ξ has multiplicity two, we have

$$f(\xi) = f'(\xi) = 0 \text{ and } f''(\xi) \neq 0.$$

Using these conditions, we get

$$\frac{f(x_k)}{f'(x_k)} = \frac{f(\xi + \varepsilon_k)}{f'(\xi + \varepsilon_k)} = \varepsilon_k \left[\frac{1}{2} + \frac{1}{6}\varepsilon_k c_3 + \frac{1}{24}\varepsilon_k^2 c_4 + \ldots\right]\left[1 + \frac{1}{2}\varepsilon_k c_3 + \frac{1}{6}\varepsilon_k^2 c_4 + \ldots\right]^{-1}$$
$$= \frac{1}{2}\varepsilon_k - \frac{1}{12}c_3\varepsilon_k^2 + \frac{1}{24}(c_3^2 - c_4)\varepsilon_k^3$$
$$+ \left(\frac{5}{144}c_3c_4 - \frac{1}{48}c_3^3 - \frac{1}{80}c_5\right)\varepsilon_k^4 + \ldots$$

where $\qquad c_i = f^{(i)}(\xi) / f''(\xi).$

Similarly, we get

$$f\left(x_k + 2\frac{f(x_k)}{f'(x_k)}\right) = f''(\xi)\left[2\varepsilon_k^2 + c_3\varepsilon_k^3 + \frac{1}{72}(36c_4 - 11c_3^2)\varepsilon_k^4 + \ldots\right]$$

and
$$\frac{f\left(x_k + 2\dfrac{f(x_k)}{f'(x_k)}\right)}{f'(x_k)} = 2\varepsilon_k + \left(\frac{1}{6}c_4 - \frac{11}{72}c_3^2\right)\varepsilon_k^3 + O(\varepsilon_k^4).$$

Substituting these expansions in the given multipoint method, we obtain the error equation

$$\varepsilon_{k+1} = \left(\frac{11}{144}c_3^2 - \frac{1}{12}c_4\right)\varepsilon_k^3 + O(\varepsilon_k^4).$$

Hence, the method has third order rate of convergence.

Taking $f(x) = 9x^4 + 30x^3 + 34x^2 + 30x + 25$

and using the given method with $x_0 = -1.4$ we obtain the sequence of iterates

$$x_0 = -1.4,\ f_0 = 1.8944,\ f_0' = 12.4160,$$

$$x_0^* = x_0 + \frac{2f_0}{f_0'} = -1.09485,\ f_0^* = 6.47026,$$

$$x_1 = x_0 - \frac{1}{2}\frac{f_0^*}{f_0'} = -1.66056.$$

Similarly, we get $x_2 = -1.66667$, $x_3 = -1.66667$.

Therefore, the root correct to three decimals is – 1.667.

Rate of Convergence

1.26 The equation $x^2 + ax + b = 0$ has two real roots α and β. Show that the iteration method

(*i*) $x_{k+1} = -(ax_k + b)/x_k$ is convergent near $x = \alpha$ if $|\alpha| > |\beta|$.

(*ii*) $x_{k+1} = -b/(x_k + a)$ is convergent near $x = \alpha$ if $|\alpha| < |\beta|$.

(*iii*) $x_{k+1} = -(x_k^2 + b)/a$ is convergent near $x = \alpha$ if $2|\alpha| < |\alpha + \beta|$.

Solution

The method is of the form $x_{k+1} = g(x_k)$. Since α and β are the two roots, we have

$$\alpha + \beta = -a, \quad \alpha\beta = b.$$

We now obtain

(*i*) $g(x) = -a - b/x$, $g'(x) = b/x^2$.

For convergence to α, we need $|g'(\alpha)| < 1$. We thus have

$$|g'(\alpha)| = \left|\frac{b}{\alpha^2}\right| = \left|\frac{\alpha\beta}{\alpha^2}\right| = \left|\frac{\beta}{\alpha}\right| < 1$$

which gives $|\beta| < |\alpha|$.

(*ii*) $g(x) = -b/(a + x)$, $g'(x) = b/(a + x)^2$.

For convergence to α, we require

$$|g'(\alpha)| = \left|\frac{\alpha\beta}{(a+\alpha)^2}\right| = \left|\frac{\alpha\beta}{\beta^2}\right| = \left|\frac{\alpha}{\beta}\right| < 1$$

which gives $|\alpha| < |\beta|$.

(*iii*) $g(x) = -(x^2 + b)/a$, $g'(x) = -2x/a$.

For convergence to α, we require

$$|g'(\alpha)| = \left|\frac{2\alpha}{a}\right| = \left|\frac{2\alpha}{(\alpha+\beta)}\right| < 1$$

which gives $2|\alpha| < |\alpha + \beta|$.

1.27 Show that the following two sequences have convergence of the second order with the same limit $\sqrt{a}$.

$$(i)\ x_{n+1} = \frac{1}{2} x_n \left(1 + \frac{a}{x_n^2}\right), \qquad (ii)\ x_{n+1} = \frac{1}{2} x_n \left(3 - \frac{x_n^2}{a}\right).$$

If x_n is a suitably close approximation to $\sqrt{a}$, show that the error in the first formula for x_{n+1} is about one-third of that in the second formula, and deduce that the formula

$$x_{n+1} = \frac{1}{8} x_n \left(6 + \frac{3a}{x_n^2} - \frac{x_n^2}{a}\right)$$

gives a sequence with third-order convergence.

Solution

Taking the limit as $n \to \infty$ and noting that $\lim_{n\to\infty} x_n = \xi$, $\lim_{n\to\infty} x_{n+1} = \xi$, where ξ is the exact root, we obtain from all the three methods $\xi^2 = a$. Thus, all the three methods determine $\sqrt{a}$, where a is any positive real number.

Substituting $x_n = \xi + \varepsilon_n$, $x_{n+1} = \xi + \varepsilon_{n+1}$ and $a = \xi^2$, we get

$$\begin{aligned}(i)\ \xi + \varepsilon_{n+1} &= (\xi + \varepsilon_n)\,[1 + \xi^2/(\xi + \varepsilon_n)^2]/2\\ &= (\xi + \varepsilon_n)\,[1 + (1 + \varepsilon_n/\xi)^{-2}]/2\\ &= (\xi + \varepsilon_n)\,[2 - 2(\varepsilon_n/\xi) + 3(\varepsilon_n^2/\xi^2) - \ldots]/2\end{aligned}$$

which gives $\varepsilon_{n+1} = \varepsilon_n^2/(2\xi) + O(\varepsilon_n^3)$. (1.37*a*)

Hence the method has second order convergence, with the error constant $c = 1/(2\xi)$.

$$\begin{aligned}(ii)\ \xi + \varepsilon_{n+1} &= (\xi + \varepsilon_n)\,[3 - (\xi + \varepsilon_n)^2/\xi^2]/2\\ &= (\xi + \varepsilon_n)\,[1 - (\varepsilon_n/\xi) - \varepsilon_n^2/\xi^2]/2\end{aligned}$$

which gives $\varepsilon_{n+1} = -\dfrac{3}{2\xi}\varepsilon_n^2 + O(\varepsilon_n^2)$. (1.37*b*)

Hence, the method has second order convergence with the error constant $c^* = -3/(2\xi)$. Therefore, the error, in magnitude, in the first formula is about one-third of that in the second formula.

If we multiply (1.37*a*) by 3 and add to (1.37*b*), we find that

$$\varepsilon_{n+1} = O(\varepsilon_n^3) \tag{1.38}$$

It can be verified that $O(\varepsilon_n^3)$ term in (1.38) does not vanish.

Adding 3 times the first formula to the second formula, we obtain the new formula

$$x_{n+1} = \frac{1}{8} x_n \left(6 + \frac{3a}{x_n^2} - \frac{x_n^2}{a}\right)$$

which has third order convergence.

1.28 Let the function $f(x)$ be four times continuously differentiable and have a simple zero ξ. Successive approximations x_n, $n = 1, 2, \ldots$ to ξ are computed from

$$x_{n+1} = \frac{1}{2}(x'_{n+1} + x''_{n+1})$$

where $$x'_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x''_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

$$g(x) = \frac{f(x)}{f'(x)}$$

Prove that if the sequence $\{x_n\}$ converges to ξ, then the convergence is cubic.

(Lund Univ., Sweden, BIT 8 (1968), 59)

Solution

We have $$g(x) = \frac{f(x)}{f'(x)}$$

$$g'(x) = \frac{(f'(x))^2 - f(x) f''(x)}{(f'(x))^2}$$

$$x'_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$x''_{n+1} = x_n - \frac{f(x_n)/f'(x_n)}{1 - [f(x_n) f''(x_n)/(f'(x_n))^2]}$$

$$= x_n - \frac{f(x_n)}{f'(x_n)}\left[1 + \frac{f(x_n) f''(x_n)}{(f'(x_n))^2} + \left\{\frac{f(x_n) f''(x_n)}{(f'(x_n))^2}\right\}^2 + \ldots\right].$$

From the formula

$$x_{n+1} = \frac{1}{2}(x'_{n+1} + x''_{n+1})$$

we obtain $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2}\left[\frac{f(x_n)}{f'(x_n)}\right]^2 \frac{f''(x_n)}{f'(x_n)} - \frac{1}{2}\left[\frac{f(x_n)}{f'(x_n)}\right]^3 \left[\frac{f''(x_n)}{f'(x_n)}\right]^2 + \ldots \tag{1.39}$$

Using $$x_n = \xi + \varepsilon_n, \quad \text{and} \quad c_i = \frac{f^{(i)}(\xi)}{f'(\xi)}, \quad i = 1, 2, 3, \ldots$$

we find $$\frac{f(x_n)}{f'(x_n)} = \varepsilon_n - \frac{1}{2} c_2 \varepsilon_n^2 + \left(\frac{1}{2} c_2^2 - \frac{1}{3} c_3\right)\varepsilon_n^3 + \ldots$$

$$\frac{f''(x_n)}{f'(x_n)} = c_2 + (c_3 - c_2^2)\,\varepsilon_n + \ldots$$

Using these expressions in (1.39), we obtain the error equation, on simplification, as

$$\varepsilon_{n+1} = \varepsilon_n - \left[\varepsilon_n - \frac{1}{2} c_2 \varepsilon_n^2 + \left(\frac{1}{2} c_2^2 - \frac{1}{3} c_3\right)\varepsilon_n^3 + \ldots\right]$$
$$- \frac{1}{2}[\varepsilon_n^2 - c_2\varepsilon_n^3 + \ldots]\,[c_2 + (c_3 - c_2^2)\,\varepsilon_n + \ldots]$$
$$- \frac{1}{2}[\varepsilon_n^3 + \ldots]\,[c_2^2 + 2c_2(c_3 - c_2^2)\,\varepsilon_n + \ldots] + \ldots$$

$$= -\frac{1}{6} c_3 \varepsilon_n^3 + O(\varepsilon_n^4).$$

Hence, the method has cubic convergence.

1.29 Determine the order of convergence of the iterative method

$$x_{k+1} = (x_0 f(x_k) - x_k f(x_0)) / (f(x_k) - f(x_0))$$

for finding a simple root of the equation $f(x) = 0$.

Solution

We write the method in the equivalent form

$$x_{k+1} = x_k - \frac{(x_k - x_0) f(x_k)}{f(x_k) - f(x_0)} \tag{1.40}$$

Substituting $x_k = \xi + \varepsilon_k$, $x_{k+1} = \xi + \varepsilon_{k+1}$, $x_0 = \xi + \varepsilon_0$ in (1.40) we get

$$\varepsilon_{k+1} = \varepsilon_k - \frac{[\varepsilon_k - \varepsilon_0] f(\xi + \varepsilon_k)}{f(\xi + \varepsilon_k) - f(\xi + \varepsilon_0)}. \tag{1.41}$$

Expanding $f(\xi + \varepsilon_k)$, $f(\xi + \varepsilon_0)$ in Taylor series about the point ξ and using $f(\xi) = 0$, we obtain from (1.41)

$$\varepsilon_{k+1} = \varepsilon_k - \frac{(\varepsilon_k - \varepsilon_0)[\varepsilon_k f'(\xi) + \frac{1}{2}\varepsilon_k^2 f''(\xi) + \ldots]}{(\varepsilon_k - \varepsilon_0) f'(\xi) + \frac{1}{2}(\varepsilon_k^2 - \varepsilon_0^2) f''(\xi) + \ldots}$$

$$= \varepsilon_k - \left[\varepsilon_k + \frac{1}{2}\varepsilon_k^2 c_2 + \ldots\right] \times \left[1 + \frac{1}{2}(\varepsilon_k + \varepsilon_0) c_2 + \ldots\right]^{-1}$$

$$= \varepsilon_k - \left[\varepsilon_k + \frac{1}{2}\varepsilon_k^2 c_2 + \ldots\right] \times \left[1 - \frac{1}{2}(\varepsilon_k + \varepsilon_0) c_2 + \ldots\right]$$

$$= \frac{1}{2}\varepsilon_k \varepsilon_0 c_2 + O(\varepsilon_k^2 \varepsilon_0 + \varepsilon_k \varepsilon_0^2)$$

where $\qquad c_2 = f''(\xi) / f'(\xi).$

Thus, the method has linear rate of convergence, since ε_0 is independent of k.

1.30 Find the order of convergence of the *Steffensen* method

$$x_{k+1} = x_k - \frac{f_k}{g_k}, \quad k = 0, 1, 2, \ldots$$

$$g_k = \frac{f(x_k + f_k) - f_k}{f_k}$$

where $f_k = f(x_k)$. Use this method to determine the non-zero root of the equation

$$f(x) = x - 1 + e^{-2x} \quad \text{with } x_0 = 0.7$$

correct to three decimals.

Solution

Write the given method as

$$x_{k+1} = x_k - \frac{f_k^2}{f(x_k + f_k) - f_k}.$$

Using $x_k = \xi + \varepsilon_k$, we obtain

$$f_k = f(\xi + \varepsilon_k) = \varepsilon_k f'(\xi) + \frac{1}{2}\varepsilon_k^2 f''(\xi) + \ldots$$

$$f(x_k + f_k) = f\left(\xi + (1 + f'(\xi))\varepsilon_k + \frac{1}{2} f''(\xi)\varepsilon_k^2 + \ldots\right)$$

$$= (1 + f'(\xi)) f'(\xi)\, \varepsilon_k + \frac{1}{2} f''(\xi)\ [1 + 3f'(\xi) + (f'(\xi))^2]\ \varepsilon_k^2 + \ldots$$

Substituting these expressions in the given formula, we get the error equation as

$$\varepsilon_{k+1} = \varepsilon_k - \frac{\varepsilon_k^2 (f'(\xi))^2 + \varepsilon_k^3 f'(\xi) f''(\xi) + \ldots}{\varepsilon_k (f'(\xi))^2 + \frac{1}{2}(3 + f'(\xi)) f'(\xi) f''(\xi)\varepsilon_k^2 + \ldots}$$

$$= \frac{1}{2}[1 + f'(\xi)] \frac{f''(\xi)}{f'(\xi)} \varepsilon_k^2 + O(\varepsilon_k^3).$$

Hence, the method has second order rate of convergence.

For $f(x) = x - 1 + e^{-2x}$ and $x_0 = 0.7$, we get

$$f_0 = -0.05340, \quad f(x_0 + f_0) = -0.07901,$$
$$g_0 = 0.47959, \quad x_1 = 0.81135,$$
$$f_1 = 0.00872, \quad f(x_1 + f_1) = 0.01402,$$
$$g_1 = 0.60780, \quad x_2 = 0.79700,$$
$$f2 = 0.00011, \quad f(x_2 + f2) = 0.00018,$$
$$g_2 = 0.63636, \quad x_3 = 0.79683.$$

The root correct to three decimal places is 0.797.

1.31 Let $x = \xi$ be a simple root of the equation $f(x) = 0$. We try to find the root by means of the iteration formula

$$x_{i+1} = x_i - (f(x_i))^2 / (f(x_i) - f(x_i - f(x_i)))$$

Find the order of convergence and compare the convergence properties with those of Newton-Raphson's method. (Bergen Univ., Sweden, BIT 20 (1980), 262)

Solution

Substituting $x_i = \xi + \varepsilon_i$, we get

$$f(x_i) = f(\xi + \varepsilon_i) = \varepsilon_i f'(\xi) + \frac{1}{2}\varepsilon_i^2 f''(\xi) + \ldots$$

$$f(x_i - f(x_i)) = f\left(\xi + \left\{(1 - f'(\xi))\varepsilon_i - \frac{1}{2}\varepsilon_i^2 f''(\xi) + \ldots\right\}\right)$$

$$= \left\{(1 - f'(\xi))\varepsilon_i - \frac{1}{2}\varepsilon_i^2 f''(\xi) + \ldots\right\} f'(\xi)$$

$$+ \frac{1}{2}\left\{(1 - f'(\xi))\varepsilon_i - \frac{1}{2}\varepsilon_i^2 f''(\xi) + \ldots\right\}^2 f''(\xi) + \ldots$$

$$= \{1 - f'(\xi)\} f'(\xi)\varepsilon_i + \frac{1}{2}\{1 - 3f'(\xi) + (f'(\xi))^2\} f''(\xi)\,\varepsilon_i^2 + \ldots$$

Substituting $x_{i+1} = \xi + \varepsilon_{i+1}$ and the above expressions for $f(x_i)$ and $f(x_i - f(x_i))$ in the given formula, we obtain on simplification

$$\varepsilon_{i+1} = \varepsilon_i - \frac{\varepsilon_i^2 (f'(\xi))^2 + \varepsilon_i^3 f'(\xi) f''(\xi) + \ldots}{\varepsilon_i (f'(\xi))^2 + \frac{1}{2}\{3 - f'(\xi)\} f'(\xi) \varepsilon_i^2 + \ldots}$$

$$= \varepsilon_i - \left[\varepsilon_i + \varepsilon_i^2 \frac{f''(\xi)}{f'(\xi)} + \ldots\right] \left[1 + \left\{\frac{3}{2} - \frac{1}{2} f'(\xi)\right\} \frac{f''(\xi)}{f'(\xi)} \varepsilon_i + \ldots\right]^{-1}$$

$$= \varepsilon_i - \left[\varepsilon_i + \varepsilon_i^2 \frac{f''(\xi)}{f'(\xi)} + \ldots\right] \left[1 - \left\{\frac{3}{2} - \frac{1}{2} f'(\xi)\right\} \frac{f''(\xi)}{f'(\xi)} \varepsilon_i + \ldots\right]$$

$$= \frac{1}{2}(1 - f'(\xi)) \frac{f''(\xi)}{f'(\xi)} \varepsilon_i^2 + O(\varepsilon_i^3).$$

Hence, the method has second order convergence if $f'(\xi) \neq 1$. The error constant is $(1 - f'(\xi)) f''(\xi)/(2f'(\xi))$.

The error constant for the Newton-Raphson method is $f''(\xi)/(2f'(\xi))$.

1.32 A root of the equation $f(x) = 0$ can be obtained by combining the Newton-Raphson method and the Regula-Falsi method. We start from $x_0 = \xi + \varepsilon$, where ξ is the true solution of $f(x) = 0$. Further, $y_0 = f(x_0)$, $x_1 = x_0 - f_0 / f_0'$ and $y_1 = f_1$ are computed. Lastly, a straight line is drawn through the points (x_1, y_1) and $((x_0 + x_1) / 2, y_0 / 2)$. If ε is sufficiently small, the intersection of the line and the x-axis gives a good approximation to ξ. To what power of ε is the error term proportional. Use this method to compute the positive root of the equation $x^4 - x - 10 = 0$, correct to three decimal places.

Solution

We have $\quad x_0 = \xi + \varepsilon_0, \; x_1 = \xi + \varepsilon_1$

$$y_0 = f(\xi + \varepsilon_0) = \varepsilon_0 f'(\xi) + \frac{1}{2} \varepsilon_0^2 f''(\xi) + \frac{1}{6} \varepsilon_0^3 f'''(\xi) + \frac{1}{24} \varepsilon_0^4 f^{iv}(\xi) + \ldots$$

$$x_1 = x_0 - \frac{f_0}{f_0'} = \xi + \varepsilon_0 - \frac{f(\xi + \varepsilon_0)}{f'(\xi + \varepsilon_0)}$$

$$\varepsilon_1 = \varepsilon_0 - \frac{\varepsilon_0 f'(\xi) + \frac{1}{2} \varepsilon_0^2 f''(\xi) + \frac{1}{6} \varepsilon_0^3 f'''(\xi) + \frac{1}{24} \varepsilon_0^4 f^{iv}(\xi) + \ldots}{f'(\xi) + \varepsilon_0 f''(\xi) + \frac{1}{2} \varepsilon_0^2 f'''(\xi) + \frac{1}{6} \varepsilon_0^3 f^{iv}(\xi) + \ldots}$$

We obtain on simplification

$$\varepsilon_1 = \frac{1}{2} c_2 \varepsilon_0^2 + \left(\frac{1}{3} c_3 - \frac{1}{2} c_2^2\right) \varepsilon_0^3 + \left(\frac{1}{2} c_2^3 - \frac{7}{12} c_2 c_3 + \frac{1}{8} c_4\right) \varepsilon_0^4 + \ldots$$

where $\quad c_i = \dfrac{f^{(i)}(\xi)}{f'(\xi)}, \quad i = 2, 3, \ldots$

Equation of the straight line through the points (x_1, y_1) and $((x_0 + x_1)/2, y_0/2)$ is

$$y - y_1 = \frac{y_0 - 2y_1}{x_0 - x_1} (x - x_1).$$

It intersects the x-axis at $\; x = x_1 - \dfrac{x_0 - x_1}{y_0 - 2y_1} y_1.$

The error equation is given by

$$\varepsilon = \varepsilon_1 - \frac{\varepsilon_0 - \varepsilon_1}{y_0 - 2y_1} y_1 \tag{1.42}$$

We have

$$\begin{aligned} y_1 &= f(\xi + \varepsilon_1) \\ &= \varepsilon_1 f'(\xi) + \frac{1}{2}\varepsilon_1^2 f''(\xi) + \frac{1}{6}\varepsilon_1^3 f'''(\xi) + \ldots \\ &= f'(\xi) + \left[\left\{\frac{1}{2}c_2\varepsilon_0^2 + \left(\frac{1}{3}c_3 - \frac{1}{2}c_2^2\right)\varepsilon_0^3 + \left(\frac{1}{2}c_2^3 - \frac{7}{12}c_2c_3 + \frac{1}{8}c_4\right)\varepsilon_0^4 + \ldots\right\}\right. \\ &\qquad \left. + \frac{1}{2}\left\{\frac{1}{4}c_2^2\varepsilon_0^4 + \ldots\right\}c_2 + \ldots\right] \\ &= f'(\xi)\left[\frac{1}{2}c_2\varepsilon_0^2 + \left(\frac{1}{3}c_3 - \frac{1}{2}c_2^2\right)\varepsilon_0^3 + \left(\frac{5}{8}c_2^3 - \frac{7}{12}c_2c_3 + \frac{1}{8}c_4\right)\varepsilon_0^4 + \ldots\right]. \end{aligned}$$

Similarly, we have

$$y_0 = f'(\xi)\left[\varepsilon_0 + \frac{1}{2}\varepsilon_0^2 c_2 + \frac{1}{6}\varepsilon_0^3 c_3 + \frac{1}{24}\varepsilon_0^4 c_4 + \ldots\right]$$

and $$y_0 - 2y_1 = f'(\xi)\left[\varepsilon_0 - \frac{1}{2}c_2\varepsilon_0^2 + \left(c_2^2 - \frac{1}{2}c_3\right)\varepsilon_0^3 + \left(\frac{7}{6}c_2c_3 - \frac{5}{24}c_4 - \frac{5}{4}c_2^3\right)\varepsilon_0^4 + \ldots\right]$$

$$\varepsilon_0 - \varepsilon_1 = \varepsilon_0 - \frac{1}{2}c_2\varepsilon_0^2 + \left(\frac{1}{2}c_2^2 - \frac{1}{3}c_3\right)\varepsilon_0^3 + \left(\frac{7}{12}c_2c_3 - \frac{1}{8}c_4 - \frac{1}{2}c_2^3\right)\varepsilon_0^4 + \ldots$$

$$\frac{\varepsilon_0 - \varepsilon_1}{y_0 - 2y_1} = \frac{1}{f'(\xi)}\left[1 + \left(\frac{1}{6}c_3 - \frac{1}{2}c_2^2\right)\varepsilon_0^2 + \left(\frac{1}{2}c_2^3 - \frac{1}{2}c_2c_3 + \frac{1}{24}c_4\right)\varepsilon_0^3 + \ldots\right].$$

Substituting these expressions in (1.42) and simplifying, we get

$$\varepsilon = \left(\frac{1}{8}c_2^3 - \frac{1}{12}c_2c_3\right)\varepsilon_0^4 + O(\varepsilon_0^5).$$

Hence, the error term is proportional to ε_0^4 and the method has fourth order rate of convergence.

For the equation $f(x) = x^4 - x - 10$, we find that $f(1) < 0$, $f(2) > 0$ and a root lies in (1, 2). Let $x_0 = 2$.

First iteration

$$x_0 = 2, \quad y_0 = f(x_0) = 4, \quad f'(x_0) = 31,$$

$$x_1 = x_0 - \frac{f_0}{f_0'} = 1.870968, \quad y_1 = f(x_1) = 0.382681,$$

$$x = x_1 - \frac{x_0 - x_1}{y_0 - 2y_1}y_1 = 1.855703.$$

Second iteration

$$x_0 = 1.855703, \; y_0 = f(x_0) = 0.002910, \quad f'(x_0) = 24.561445,$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1.855585, \quad y_1 = f(x_1) = 0.000012$$

$$x = x_1 - \frac{x_0 - x_1}{y_0 - 2y_1}y_1 = 1.855585.$$

Hence, the root correct to three decimal places is 1.856.

1.33 Determine p, q and r so that the order of the iterative method

$$x_{n+1} = px_n + \frac{qa}{x_n^2} + \frac{ra^2}{x_n^5}$$

for $a^{1/3}$ becomes as high as possible. For this choice of p, q and r, indicate how the error in x_{n+1} depends on the error in x_n. (Lund Univ., Sweden, BIT 8 (1968), 138)

Solution

We have $x = a^{1/3}$ or $x^3 = a$. We take

$$f(x) = x^3 - a.$$

Since ξ is the exact root, we have $\xi^3 = a$.

Substituting $x_n = \xi + \varepsilon_n$, $x_{n+1} = \xi + \varepsilon_{n+1}$ and $a = \xi^3$ in the given method, we obtain

$$\xi + \varepsilon_{n+1} = p(\xi + \varepsilon_n) + \frac{qa}{\xi^2}\left(1 + \frac{\varepsilon_n}{\xi}\right)^{-2} + \frac{ra^2}{\xi^5}\left(1 + \frac{\varepsilon_n}{\xi}\right)^{-5}$$

$$= p(\xi + \varepsilon_n) + \frac{qa}{\xi^2}\left(1 - \frac{2\varepsilon_n}{\xi} + \frac{3\varepsilon_n^2}{\xi^2} - 4\frac{\varepsilon_n^3}{\xi^3} + \ldots\right)$$

$$+ \frac{ra^2}{\xi^5}\left(1 - 5\frac{\varepsilon_n}{\xi} + 15\frac{\varepsilon_n^2}{\xi^2} - 35\frac{\varepsilon_n^3}{\xi^3} + \ldots\right)$$

$$= p(\xi + \varepsilon_n) + q\xi\left(1 - 2\frac{\varepsilon_n}{\xi} + 3\frac{\varepsilon_n^2}{\xi^2} - 4\frac{\varepsilon_n^3}{\xi^3} + \ldots\right)$$

$$+ r\xi\left(1 - 5\frac{\varepsilon_n}{\xi} + 15\frac{\varepsilon_n^2}{\xi^2} - 35\frac{\varepsilon_n^3}{\xi^3} + \ldots\right)$$

or

$$\varepsilon_{n+1} = (p + q + r - 1)\xi + (p - 2q - 5r)\varepsilon_n$$

$$+ \frac{1}{\xi}(3q + 15r)\varepsilon_n^2 - \frac{1}{\xi^2}(4q + 35r)\varepsilon_n^3 + \ldots$$

For the method to be of order three, we have

$$p + q + r = 1$$
$$p - 2q - 5r = 0$$
$$3q + 15r = 0$$

which gives $p = 5/9$, $q = 5/9$, $r = -1/9$.

The error equation becomes

$$\varepsilon_{n+1} = \frac{5}{3\xi^2}\varepsilon_n^3 + O(\varepsilon_n^4).$$

1.34 Given the equation $f(x) = 0$, obtain an iteration method using the rational approximation

$$f(x) = \frac{x - a_0}{b_0 + b_1 x}$$

where the coefficients a_0, b_0 and b_1 are determined by evaluating $f(x)$ at x_k, x_{k-1} and x_{k-2}.

(*i*) Find the order of convergence of this method.

(*ii*) Carry out two iterations using this method for the equation

$$f(x) = 2x^3 - 3x^2 + 2x - 3 = 0 \text{ with } x_0 = 0,\ x_1 = 1,\ x_2 = 2.$$

Solution

We have, from the given approximation

$$x - a_0 - (b_0 + b_1 x) f(x) = 0. \tag{1.43}$$

Substituting $x = x_k$, x_{k-1} and x_{k-2} in the above equation, we get

$$x_k - a_0 - (b_0 + b_1 x_k) f_k = 0 \tag{1.44}$$

$$x_{k-1} - a_0 - (b_0 + b_1 x_{k-1}) f_{k-1} = 0 \tag{1.45}$$

$$x_{k-2} - a_0 - (b_0 + b_1 x_{k-2}) f_{k-2} = 0 \tag{1.46}$$

Eliminating b_0 from the above equations, we obtain

$$g_2 + a_0 g_1 + b_1 x^* f_k f_{k-1} = 0 \tag{1.47}$$

$$h_2 + a_0 h_1 + b_1 x' f_k f_{k-2} = 0 \tag{1.48}$$

where $\quad x^* = x_{k-1} - x_k, \qquad x' = x_{k-2} - x_k,$

$g_1 = f_k - f_{k-1}, \qquad h_1 = f_k - f_{k-2},$

$g_2 = x_k f_{k-1} - x_{k-1} f_k, \qquad h_2 = x_k f_{k-2} - x_{k-2} f_k.$

Eliminating b_1 from (1.47) and (1.48) and solving for a_0, we get

$$a_0 = -\frac{x' g_2 f_{k-2} - x^* h_2 f_{k-1}}{x' g_1 f_{k-2} - x^* h_1 f_{k-1}} = x_k + \frac{x' x^* f_k (f_{k-1} - f_{k-2})}{x^* h_1 f_{k-1} - x' g_1 f_{k-2}} \tag{1.49}$$

The exact root is obtained from

$$f(\xi) = \frac{\xi - a_0}{b_0 + b_1 \xi} \equiv 0 \quad \text{or} \quad \xi = a_0.$$

Thus, we obtain the iteration formula

$$x_{k+1} = a_0 \tag{1.50}$$

where a_0 is given by (1.49).

(*i*) To find the order of convergence of the method, we write (1.50) as

$$x_{k+1} = x_k + \frac{\text{NUM}}{\text{DEN}} \tag{1.51}$$

Substituting $x_k = \xi + \varepsilon_k$ and simplifying, we get

$$\text{NUM} = D\varepsilon_k \left[1 + \frac{1}{2}(\varepsilon_k + \varepsilon_{k-1} + \varepsilon_{k-2})c_2 + \frac{1}{6}(\varepsilon_k^2 + \varepsilon_{k-1}^2 + \varepsilon_{k-2}^2 + \varepsilon_{k-1}\varepsilon_{k-2})c_3 \right.$$
$$\left. + \frac{1}{4}(\varepsilon_{k-1} + \varepsilon_{k-2})\varepsilon_k c_2^2 + \ldots \right]$$

$$\text{DEN} = (\varepsilon_k - \varepsilon_{k-1})(\varepsilon_k - \varepsilon_{k-2})(\varepsilon_{k-1} - \varepsilon_{k-2})(f'(\xi))^2$$
$$- D \left[1 + \frac{1}{2}(\varepsilon_k + \varepsilon_{k-1} + \varepsilon_{k-2})c_2 + \frac{1}{6}(\varepsilon_k^2 + \varepsilon_{k-1}^2 + \varepsilon_{k-2}^2)c_3 \right.$$
$$\left. + \frac{1}{4}(\varepsilon_k\varepsilon_{k-1} + \varepsilon_k\varepsilon_{k-2} + \varepsilon_{k-1}\varepsilon_{k-2})c_2^2 + \ldots \right]$$

where $\qquad c_i = \dfrac{f^{(i)}(\xi)}{f'(\xi)}, \quad i = 2, 3, \ldots$

Substituting the expressions for NUM and DEN in (1.51), taking DEN to the numerator and simplifying we get error equation

$$\varepsilon_{n+1} = c\, \varepsilon_n \varepsilon_{n-1}\, \varepsilon_{n-2} \tag{1.52}$$

where $$c = \frac{1}{4} c_2^2 - \frac{1}{6} c_3 .$$

From the definition, we have

$$\varepsilon_{n+1} = A \varepsilon_n^p$$

$$\varepsilon_n = A \varepsilon_{n-1}^p \quad \text{or} \quad \varepsilon_{n-1} = A^{-1/p}\, \varepsilon_n^{1/p}$$

$$\varepsilon_{n-1} = A\, \varepsilon_{n-2}^p = A^{-1/p}\, \varepsilon_n^{1/p}$$

or $$\varepsilon_{n-2} = A^{-(1/p)-(1/p^2)} \varepsilon_n^{1/p^2} .$$

Substituting the values of ε_{n+1}, ε_{n-1} and ε_{n-2} in terms of ε_n in the error equation (1.52), we obtain

$$A \varepsilon_n^p = c \varepsilon_n \{A^{-1/p} \varepsilon_n^{1/p}\} \{A^{-(1/p)-(1/p^2)} \varepsilon_n^{1/p^2}\}$$

which gives $$\varepsilon_n^p = c A^{-1-(2/p)-(1/p^2)} \varepsilon_n^{1+(1/p)+(1/p^2)} .$$

Comparing the powers of ε_n on both sides, we get

$$p = 1 + \frac{1}{p} + \frac{1}{p^2} \quad \text{or} \quad p^3 - p^2 - p - 1 = 0$$

which has the smallest positive root 1.84.

Hence the order of the method is 1.84.

(*ii*) For $f(x) = 2x^3 - 3x^2 + 2x - 3$ and $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, we obtain

First iteration

$$x^* = x_1 - x_2 = -1,\ x' = x_0 - x_2 = -2,\ f_0 = -3,\ f_1 = -2,$$
$$f_2 = 5,\ g_1 = f_2 - f_1 = 7,\ h_1 = f_2 - f_0 = 8,$$
$$x_3 = x_2 + \frac{x' x^* f_2 (f_1 - f_0)}{x^* h_1 f_1 - x' g_1 f_0} = 1.6154.$$

Second iteration

$$x_0 = 1,\quad x_1 = 2,\ x_2 = 1.6154,\quad f_0 = -2,\quad f_1 = 5,\quad f_2 = 0.8331,$$
$$x^{*4} = 0.3846,\quad x' = -0.6154,\ g_1 = -4.1669,\quad h_1 = 2.8331\ \ g_1,\quad x_4 = 1.4849.$$

1.35 The equation $x^4 + x = \varepsilon$, where ε is a small number, has a root which is close to ε. Computation of this root is done by the expression

$$\xi = \varepsilon - \varepsilon^4 + 4\varepsilon^7.$$

(*i*) Find an iterative formula $x_{n+1} = F(x_n)$, $x_0 = 0$, for the computation. Show that we get the expression above after three iterations when neglecting terms of higher order.

(*ii*) Give a good estimate (of the form $N\varepsilon^k$, where N and k are integers) of the maximal error when the root is estimated by the expression above.

(Inst. Tech. Stockholm, Sweden, BIT 9 (1969), 87)

Solution

(*i*) We write the given equation $x^4 + x = \varepsilon$ in the form

$$x = \frac{\varepsilon}{x^3 + 1}$$

and consider the formula

$$x_{n+1} = \frac{\varepsilon}{x_n^3 + 1}.$$

The root is close to ε. Starting with $x_0 = 0$, we obtain

$$x_1 = \varepsilon$$

$$x_2 = \frac{\varepsilon}{1+\varepsilon^3} = \varepsilon(1 + \varepsilon^3)^{-1} = \varepsilon(1 - \varepsilon^3 + \varepsilon^6 + ...)$$

$$= \varepsilon - \varepsilon^4 + \varepsilon^7, \quad \text{neglecting higher powers of } \varepsilon,$$

$$x_3 = \frac{\varepsilon}{1+(\varepsilon - \varepsilon^4 + \varepsilon^7)^3} = \varepsilon - \varepsilon^4 + 4\varepsilon^7 + ...$$

$$x_4 = \frac{\varepsilon}{1+(\varepsilon - \varepsilon^4 + 4\varepsilon^7)^3} = \varepsilon - \varepsilon^4 + 4\varepsilon^7 + ...$$

(*ii*) Taking $\xi = \varepsilon - \varepsilon^4 + 4\varepsilon^7$, we find that

$$\text{Error} = \xi^4 + \xi - \varepsilon = (\varepsilon - \varepsilon^4 + 4\varepsilon^7)^4 + (\varepsilon - \varepsilon^4 + 4\varepsilon^7) - \varepsilon$$

$$= 22\varepsilon^{10} + \text{higher powers of } \varepsilon.$$

1.36 Consider the iteration method

$$x_{k+1} = \phi\,(x_k), \quad k = 0, 1, ...$$

for solving the equation $f(x) = 0$. We choose the iteration function in the form

$$\phi(x) = x - \gamma_1 f(x) - \gamma_2 f^2(x) - \gamma_3 f^3(x)$$

where γ_1, γ_2, γ_3 are arbitrary parameters to be determined. Find the γ's such that the iteration method has the orders (*i*) three (*ii*) four. Apply these methods to determine a root of the equation $x = e^x / 5$ with $x_0 = 0.4$ correct to three decimal places.

Solution

Substituting $x_k = \xi + \varepsilon_k$, $\varepsilon_{k+1} = \xi + \varepsilon_{k+1}$ in the iteration method

$$x_{k+1} = x_k - \gamma_1 f_k - \gamma_2 f_k^2 - \gamma_3 f_k^3$$

and expanding f_k in Taylor series about the point ξ, we obtain

$$\varepsilon_{k+1} = \varepsilon_k - \gamma_1 \left[f'(\xi)\varepsilon_k + \frac{1}{2} f''(\xi)\,\varepsilon_k^2 + \frac{1}{6} f'''(\xi)\,\varepsilon_k^3 + \frac{1}{24} f^{iv}(\xi)\,\varepsilon_k^4 + ...\right]$$

$$- \gamma_2 \left[(f'(\xi))^2\,\varepsilon_k^2 + f'(\xi)\,f''(\xi)\,\varepsilon_k^3 + \left(\frac{1}{4}(f''(\xi))^2 + \frac{1}{3} f'(\xi) f'''(\xi)\right)\varepsilon_k^4 + ...\right]$$

$$- \gamma_3 \left[(f'(\xi))^3\,\varepsilon_k^3 + \frac{3}{2}(f'(\xi))^3\, f''(\xi)\,\varepsilon_k^4 + ...\right]$$

$$= [1 - \gamma_1 f'(\xi)]\,\varepsilon_k - \left[\frac{1}{2}\gamma_1 f''(\xi) + \gamma_2 (f'(\xi))^2\right]\varepsilon_k^2$$

$$- \left[\frac{1}{6}\gamma_1 f'''(\xi) + \gamma_2 f'(\xi) f''(\xi) + \gamma_3 (f'(\xi))^3\right]\varepsilon_k^3$$

$$- \left[\frac{1}{24}\gamma_1 f^{iv}(\xi) + \gamma_2 \left(\frac{1}{4}(f''(\xi))^2 + \frac{1}{3} f'(\xi)\, f'''(\xi)\right) + \frac{3}{2}\gamma_3 (f'(\xi))^2 f''(\xi)\right]\varepsilon_k^4$$

$$+ \ ...$$

If the method is of third order, we have

$$1 - \gamma_1 f'(\xi) = 0$$

$$\frac{1}{2}\gamma_1 f''(\xi) + \gamma_2 (f'(\xi))^2 = 0$$

which gives $$\gamma_1 = \frac{1}{f'(\xi)}, \quad \gamma_2 = -\frac{1}{2}\frac{f''(\xi)}{(f'(\xi))^3}.$$

Replacing ξ by x_k, we obtain the third order method

$$x_{k+1} = x_k - \frac{f_k}{f_k'} + \frac{1}{2}\frac{f_k'' f_k^2}{(f_k')^3}. \tag{1.53}$$

If the method is of fourth order, we have

$$1 - \gamma_1 f'(\xi) = 0$$

$$\frac{1}{2}\gamma_1 f''(\xi) + \gamma_2 (f'(\xi))^2 = 0$$

$$\frac{1}{6}\gamma_1 f'''(\xi) + \gamma_2 f'(\xi) f''(\xi) + \gamma_3 (f'(\xi))^3 = 0$$

which give $$\gamma_1 = \frac{1}{f'(\xi)},$$

$$\gamma_2 = -\frac{1}{2}\frac{f''(\xi)}{(f'(\xi))^3}$$

$$\gamma_3 = \left[\frac{1}{2}\frac{f''(\xi)}{(f'(\xi))^2} - \frac{1}{6}\frac{f'''(\xi)}{f'(\xi)}\right] \Big/ (f'(\xi))^3.$$

Replacing ξ by x_k, we obtain the fourth order method

$$x_{k+1} = x_k - \frac{f_k}{f_k'} + \frac{1}{2}\frac{f_k'' f_k^2}{(f_k')^3} - \left[\frac{1}{2}\frac{f_k''}{(f_k')^2} - \frac{1}{6}\frac{f_k'''}{f_k'}\right]\frac{f_k^3}{(f_k')^3}. \tag{1.54}$$

For the function $f(x) = x - e^x / 5$, we have

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

Using the third order method (1.53), we obtain

$x_0 = 0.4$, $f_0 = 0.1016$, $f_0' = 0.7016$, $f_0'' = -0.2984$,

$x_1 = 0.2507$, $f_1 = -0.0063$, $f_1' = 0.7430$, $f_1'' = -0.2570$,

$x_2 = 0.2592$, $f_2 = 0.00002$, $f_2' = 0.7408$ $f_2'' = -0.2592$,

$x_3 = 0.2592$.

Hence, the root exact to three decimal places is 0.259.

Using the fourth order method (1.54), we obtain

$x_0 = 0.4$, $f_0 = 0.1016$, $f_0' = 0.7016$, $f_0'' = -0.2984$, $f_0''' = -0.2984$

$x_1 = 0.2514$, $f_1 = -0.0058$, $f_1' = 0.7428$, $f_1'' = -0.2572$, $f_1''' = -0.2572$

$x_2 = 0.2592$

Hence, the root correct to three decimal places is 0.259.

1.37 The equation

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

has one root near $x = 4$, which is to be computed by the iteration

$$x_0 = 4$$

$$x_{n+1} = \frac{3 + (k-4)x_n + 5x_n^2 - x_n^3}{k}, \quad k \text{ integer}$$

(*a*) Determine which value of k will give the fastest convergence.

(*b*) Using this value of k, iterate three times and estimate the error in x_3.

(Royal Inst. Tech. Stockholm, Sweden, BIT 11 (1971), 125)

Solution

(*a*) Let ξ be the exact root of the given equation. Hence, we get

$$\xi^3 - 5\xi^2 + 4\xi - 3 = 0$$

From the iteration formula

$$kx_{n+1} = 3 + (k-4)\, x_n + 5x_n^2 - x_n^3$$

we get, on substituting $x_n = \xi + \varepsilon_n$ and $x_{n+1} = \xi + \varepsilon_{n+1}$

$$k\varepsilon_{n+1} = (3 - 4\xi + 5\xi^2 - \xi^3) + (k - 4 + 10\xi - 3\xi^2)\varepsilon_n + O(\varepsilon_n^2) \tag{1.55}$$

Since the root is near $x = 4$, we can choose $\xi = 4 + \delta$.

Substituting $\xi = 4 + \delta$ in (1.55), we obtain

$$k\varepsilon_{n+1} = (k - 12)\varepsilon_n + O(\delta\varepsilon_n).$$

Hence, highest rate of convergence, is obtained when $k = 12$.

For $k = 12$, we obtain the iteration formula

$$x_{n+1} = \frac{1}{12}\,(3 + 8x_n + 5x_n^2 - x_n^3).$$

(*b*) Starting with $x_0 = 4$, we obtain the sequence of iterates

$$x_1 = 4.25, \qquad x_2 = 4.2122, \qquad x_3 = 4.2230, \quad x_4 = 4.2201.$$

Since the root is correct to two decimal places, maximum absolute error is 0.005.

1.38 A sequence $\{x_n\}_1^\infty$ is defined by

$$x_0 = 5$$

$$x_{n+1} = \frac{1}{16}x_n^4 - \frac{1}{2}x_n^3 + 8x_n - 12$$

Show that it gives cubic convergence to $\xi = 4$.

Calculate the smallest integer n for which the inequality

$$|\, x_n - \xi \,| < 10^{-6}$$

in valid. (Uppsala Univ., Sweden, BIT 13 (1973), 493)

Solution

As $n \to \infty$, the method converges to

$$\xi^4 - 8\xi^3 + 112\xi - 192 = 0.$$

Hence, the method finds a solution of the equation

$$f(x) = x^4 - 8x^3 + 112x - 192 = 0.$$

Substituting $x_n = \xi + \varepsilon_n$ and $x_{n+1} = \xi + \varepsilon_{n+1}$ in the given iteration formula, we get the error equation

$$\varepsilon_{n+1} = \left(\frac{1}{16}\xi^4 - \frac{1}{2}\xi^3 + 7\xi - 12\right) + \left(\frac{1}{4}\xi^3 - \frac{3}{2}\xi^2 + 8\right)\varepsilon_n$$
$$+ \left(\frac{3}{8}\xi^2 - \frac{3}{2}\xi\right)\varepsilon_n^2 + \left(\frac{1}{4}\xi - \frac{1}{2}\right)\varepsilon_n^3 + \frac{1}{16}\varepsilon_n^4.$$

For $\xi = 4$, we get $\qquad \varepsilon_{n+1} = \frac{1}{2}\varepsilon_n^3 + O(\varepsilon_n^4).$

Hence, the method has cubic rate of convergence.

Taking the error equation as

$$\varepsilon_{n+1} = c\varepsilon_n^3, \quad c = 1/2$$

we find $\qquad \varepsilon_n = c\varepsilon_{n-1}^3 = c\,(c\varepsilon_{n-2}^3)^3 = \ldots = c \,.\, c^3 \,.\, c^{3^2} \,.\, c^{3^{n-1}}\, \varepsilon_0^{3^n} = c^p\, \varepsilon_0^{3^n}.$

where $\qquad p = (3^n - 1)/2.$

Since $\varepsilon_0 = |\ \xi - x_0\ | = 1$, we have

$$\varepsilon_n = c^p = (1/2)^p.$$

Choosing n such that $(1/2)^{(3^n-1)/2} < 10^{-6}$ we obtain $n \geq 4$.

1.39 We wish to compute the root of the equation

$$e^{-x} = 3 \log_e x,$$

using the formula

$$x_{n+1} = x_n - \frac{3 \log_e x_n - \exp(-x_n)}{p}.$$

Show that, $p = 3$ gives rapid convergence.

(Stockholm Univ., Sweden, BIT 14 (1974), 254)

Solution

Substituting $x_n = \xi + \varepsilon_n$ and $x_{n+1} = \xi + \varepsilon_{n+1}$ in the given iteration method, we get

$$\varepsilon_{n+1} = \varepsilon_n - \frac{3 \log \varepsilon_e (\xi + \varepsilon_n) - \exp(-\xi - \varepsilon_n)}{p}$$
$$= \varepsilon_n - \frac{1}{p}\left[3 \log_e \xi + 3 \log_e \left(1 + \frac{\varepsilon_n}{\xi}\right) - \exp(-\xi) \exp(-\varepsilon_n)\right]$$
$$= \varepsilon_n - \frac{1}{p}\left[3 \log_e \xi + 3\left(\frac{\varepsilon_n}{\xi} - \frac{\varepsilon_n^2}{2\xi^2} + O(\varepsilon_n^3)\right) - \exp(-\xi)\left(1 - \varepsilon_n + \frac{\varepsilon_n^2}{2} - \ldots\right)\right].$$

Since ξ is the exact root, $e^{-\xi} - 3 \log_e \xi = 0$, and we obtain the error equation as

$$\varepsilon_{n+1} = \left[1 - \frac{1}{p}\left(\frac{3}{\xi} + e^{-\xi}\right)\right]\varepsilon_n + O(\varepsilon_n^2).$$

The method will have rapid convergence if

$$p = \frac{3}{\xi} + e^{-\xi} \tag{1.56}$$

where ξ is the root of $e^{-x} - 3 \log_e x = 0$. The root lies in (1, 2). Applying the Newton-Raphson method (1.9) to this equation with $x_0 = 1.5$, we obtain

$$x_1 = 1.053213, \quad x_2 = 1.113665, \quad x_3 = 1.115447, \quad x_4 = 1.115448.$$

Taking $\xi = 1.1154$, we obtain from (1.56), $p = 2.9835$. Hence $p \approx 3$.

1.40 How should the constant α be chosen to ensure the fastest possible convergence with the iteration formula

$$x_{n+1} = \frac{\alpha x_n + x_n^{-2} + 1}{\alpha + 1}$$

(Uppsala Univ., Sweden, BIT 11 (1971), 225)

Solution

Since $\lim_{n \to \infty} x_n = \lim_{n \to \infty} x_{n+1} = \xi$

we obtain from the given iteration formula

$$f(x) = \xi^3 - \xi^2 - 1 = 0.$$

Thus, the formula is being used to find a root of

$$f(x) = x^3 - x^2 - 1 = 0.$$

Substituting $x_n = \xi + \varepsilon_n$, $x_{n+1} = \xi + \varepsilon_{n+1}$, we obtain

$$(1 + \alpha)(\xi + \varepsilon_{n+1}) = \alpha(\xi + \varepsilon_n) + \frac{1}{\xi^2}\left(1 + \frac{\varepsilon_n}{\xi}\right)^{-2} + 1$$

which gives $(1 + \alpha)\,\varepsilon_{n+1} = \left(\alpha - \dfrac{2}{\xi^3}\right)\varepsilon_n + O(\varepsilon_n^2)$.

For fastest convergence, we must have $\alpha = 2 / \xi^3$.

We can determine the approximate value of ξ by using Newton-Rephson method to the equation $x^3 - x^2 - 1 = 0$. The root lies in (1, 2). Starting with $x_0 = 1.5$, we obtain $\xi \approx 1.4656$. Hence, $\alpha \approx 0.6353$.

System of nonlinear equations

1.41 Perform three iterations of the Newton-Raphson method directly or using (1.25) for solving the following equations :

(*i*) $1 + z^2 = 0$, $\quad z_0 = (1 + i) / 2$.

(*ii*) $z^3 - 4iz^2 - 3e^z = 0$, $\quad z_0 = -0.53 - 0.36i$.

Solution

(*i*) Separating the given equation into real and imaginary parts, we get

$$u(x, y) = 1 + x^2 - y^2, \quad v(x, y) = 2xy, \quad x_0 = 1/2, y_0 = 1/2$$

$$u_x = 2x, \; u_y = -2y, \quad v_x = 2y, \quad v_y = 2x.$$

Using the method (1.25), we obtain

$$x_{k+1} = x_k - [(uv_y - vu_y)_k] / D,$$

$$y_{k+1} = y_k - [(u_x v - v_x u)_k] / D,$$

$$D = (u_x v_y - u_y v_x)_k, \; k = 0, 1, 2, \ldots$$

We obtain

$$\begin{array}{llll} u_0 = 1.0, & v_0 = 0.5, & x_1 = -0.25, & y_1 = 0.75, \\ u_1 = 0.5, & v_1 = -0.375, & x_2 = 0.075, & y_2 = 0.975, \\ u_2 = 0.055, & v_2 = 0.14625, & x_3 = -0.00172, & y_3 = 0.9973. \end{array}$$

(*ii*) We can proceed exactly as in part (*i*), or can use the method

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)}$$

directly. Starting with $z_0 = (-0.53, -0.36)$ and using complex arithmetic, we obtain

$$z_1 = (-0.5080, -0.3864), \quad z_2 = (-0.5088, -0.3866),$$
$$z_3 = (-0.5088, -0.3867).$$

1.42 It is required to solve the two simultaneous equations

$$x = f(x, y), \; y = g(x, y)$$

by means of an iteration sequence. Show that the sequence

$$x_{n+1} = f(x_n, y_n), \; y_{n+1} = g(x_n, y_n) \tag{1.57}$$

will converge to a solution if the roots of the quadratic

$$\lambda^2 - (f_x + g_x)\,\lambda + (f_x g_y - f_y g_x) = 0$$

are less than unity in modulus, the derivatives being evaluated at the solution.
Obtain the condition that the iterative scheme

$$x_{n+1} = f(x_n, y_n), \quad y_{n+1} = g(x_{n+1}, y_n) \tag{1.58}$$

will converge. Show further that if $f_x = g_y = 0$ and both sequences converge, then the second sequence converges more rapidly than the first.

Solution

Let (ξ, η) be the exact solution and $(\varepsilon_{n+1}, \eta_{n+1})$ be the error in the $(n + 1)$th iteration. We have

$$\begin{array}{ll} x_{n+1} = f(x_n, y_n), & \xi = f(\xi, \eta), \\ y_{n+1} = g(x_n, y_n), & \eta = g(\xi, \eta) \end{array}$$

and the error equations

$$\xi + \varepsilon_{n+1} = f(\xi + \varepsilon_n, \eta + \eta_n) \approx \xi + \varepsilon_n f_x + \eta_n f_y$$
$$\eta + \eta_{n+1} = g(\xi + \varepsilon_n, \eta + \eta_n) \approx \eta + \varepsilon_n g_x + \eta_n g_y$$

where the derivatives are being evaluated at (ξ, η). Hence, we have

$$\begin{pmatrix} \varepsilon_{n+1} \\ \eta_{n+1} \end{pmatrix} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} \begin{pmatrix} \varepsilon_n \\ \eta_n \end{pmatrix}, \; n = 0, 1, \ldots$$

which can also be written in the form

$$\mathbf{E}_{n+1} = \mathbf{AE}_n, \; n = 0, 1, \ldots$$

where $\quad \mathbf{E}_n = \begin{pmatrix} \varepsilon_n \\ \eta_n \end{pmatrix}$ and $\mathbf{A} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}$.

The characteristic equation associated with $\mathbf{A}$ is given by

$$\lambda^2 - (f_x + g_y)\lambda + (f_x g_y - f_y g_x) = 0 \tag{1.59}$$

Using (1.59) we find that the necessary and sufficient condition for the convergence of the iterative sequence (1.57) is that the roots of (1.59) must be less than unity in modulus. When $f_x = g_y = 0$, the roots of (1.63) are obtained as

$$\lambda = \pm \sqrt{|f_y g_x|} \quad \text{or} \quad \rho(\mathbf{A}) = \sqrt{|f_y g_x|}.$$

The error equations for the iterative scheme (1.58) are obtained as

$$\xi + \varepsilon_{n+1} = f(\xi + \varepsilon_n, \eta + \eta_n) \approx \xi + \varepsilon_n f_x + \eta_n f_y$$
$$\eta + \eta_{n+1} = g(\xi + \varepsilon_{n+1}, \eta + \eta_n) \approx \eta + \varepsilon_{n+1} g_x + \eta_n g_y.$$

We get from above

$$\varepsilon_{n+1} = \varepsilon_n f_x + \eta_n f_y$$
$$\eta_{n+1} = \varepsilon_n g_x f_x + \eta_n (g_x f_y + g_y)$$

or
$$\begin{pmatrix} \varepsilon_{n+1} \\ \eta_{n+1} \end{pmatrix} = \begin{pmatrix} f_x & f_y \\ g_x f_x & g_x f_y + g_y \end{pmatrix} \begin{pmatrix} \varepsilon_n \\ \eta_n \end{pmatrix},$$

or
$$\mathbf{E}_{n+1} = \mathbf{B}\mathbf{E}_n, \; n = 0, 1, \ldots$$

where
$$\mathbf{B} = \begin{pmatrix} f_x & f_y \\ g_x f_x & g_x f_y + g_y \end{pmatrix}.$$

The necessary and sufficient condition for the convergence of the iteration sequence (1.58) is that the roots of the characteristic equation associated with **B**, that is,

$$\lambda^2 - \lambda(f_x + g_y + g_x f_y) + f_x g_y = 0 \tag{1.60}$$

are less than unity in modulus. When $f_x = g_y = 0$, the roots of (1.60) are obtained as

$$\lambda = 0, g_x f_y \quad \text{and} \quad \rho(\mathbf{B}) = |g_x f_y| = [\rho(\mathbf{A})]^2.$$

Hence, the iterative sequence (1.58) is at least two times faster than the iterative sequence (1.57).

1.43 The system of equations

$$y \cos(xy) + 1 = 0$$
$$\sin(xy) + x - y = 0$$

has one solution close to $x = 1$, $y = 2$ Calculate this solution correct to 2 decimal places.

(Umea Univ., Sweden, BIT 19 (1979), 552)

Solution

We obtain from

$$f_1(x, y) = y \cos(x y) + 1, \quad f_2(x, y) = \sin(x y) + x - y$$

the Jacobian matrix as

$$\boldsymbol{J}(x_n, y_n) = \begin{bmatrix} -y_n^2 \sin(x_n y_n) & \cos(x_n y_n) - x_n y_n \sin(x_n y_n) \\ y_n \cos(x_n y_n) + 1 & x_n \cos(x_n y_n) - 1 \end{bmatrix}$$

and
$$\boldsymbol{J}^{-1}(x_n, y_n) = \frac{1}{D} \begin{bmatrix} x_n \cos(x_n y_n) - 1 & x_n y_n \sin(x_n y_n) - \cos(x_n y_n) \\ -1 - y_n \cos(x_n y_n) & -y_n^2 \sin(x_n y_n) \end{bmatrix}$$

where
$$D = (x_n + y_n) y_n \sin(x_n y_n) - \cos(x_n y_n)\, [y_n \cos(x_n y_n) + 1]$$

Using the method

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - J^{-1}(x_n, y_n) \begin{pmatrix} f_1(x_n, y_n) \\ f_2(x_n, y_n) \end{pmatrix}, \; n = 0, 1, \ldots$$

and starting with the initial approximation $x_0 = 1, y_0 = 2$, we obtain the following sequence of iterates

First iteration $D = 5.5256$,

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} -0.2563 & 0.4044 \\ -0.0304 & -0.6582 \end{pmatrix} \begin{pmatrix} 0.1677 \\ -0.0907 \end{pmatrix} = \begin{pmatrix} 1.0797 \\ 1.9454 \end{pmatrix}.$$

Second iteration $D = 5.0873$,

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1.0797 \\ 1.9454 \end{pmatrix} - \begin{pmatrix} -0.3038 & 0.4556 \\ -0.0034 & -0.6420 \end{pmatrix} \begin{pmatrix} 0.0171 \\ -0.0027 \end{pmatrix} = \begin{pmatrix} 1.0861 \\ 1.9437 \end{pmatrix}.$$

Third iteration $D = 5.0504$,

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1.0861 \\ 1.9437 \end{pmatrix} - \begin{pmatrix} -0.3086 & 0.4603 \\ -0.00005 & -0.6415 \end{pmatrix} \begin{pmatrix} 0.00025 \\ -0.00002 \end{pmatrix} = \begin{pmatrix} 1.0862 \\ 1.9437 \end{pmatrix}.$$

Hence, the solution correct to 2 decimal places is $x = 1.09$, $y = 1.94$.

1.44 The system of equations

$$\log_e (x^2 + y) - 1 + y = 0$$

$$\sqrt{x} + xy = 0$$

has one approximate solution $(x_0, y_0) = (2.4, -0.6)$. Improve this solution and estimate the accuracy of the result. (Lund Univ., Sweden, BIT 18 (1978), 366)

Solution

We have from

$$f_1(x, y) = \log_e (x^2 + y) - 1 + y,$$

$$f_2(x, y) = \sqrt{x} + xy$$

the Jacobian matrix as

$$\boldsymbol{J}(x_n, y_n) = \begin{bmatrix} \dfrac{2x_n}{(x_n^2 + y_n)} & \dfrac{(x_n^2 + y_n + 1)}{(x_n^2 + y_n)} \\ \dfrac{(1 + 2y_n\sqrt{x_n})}{(2\sqrt{x_n})} & x_n \end{bmatrix}$$

and
$$\boldsymbol{J}^{-1}(x_n, y_n) = \frac{1}{D} \begin{bmatrix} x_n & \dfrac{-(x_n^2 + y_n + 1)}{(x_n^2 + y_n)} \\ \dfrac{-(1 + 2y_n\sqrt{x_n})}{(2\sqrt{x_n})} & \dfrac{2x_n}{(x_n^2 + y_n)} \end{bmatrix}$$

where
$$D = \frac{4x_n^{5/2} - (1 + 2y_n\sqrt{x_n})(x_n^2 + y_n + 1)}{(x_n^2 + y_n)(2\sqrt{x_n})}.$$

Using the method

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \boldsymbol{J}^{-1}(x_n, y_n) \begin{pmatrix} f_1(x_n, y_n) \\ f_2(x_n, y_n) \end{pmatrix}, \quad n = 0, 1, \ldots$$

and starting with $(x_0, y_0) = (2.4, -0.6)$, we obtain the following sequence of iterates

First iteration $D = 2.563540$,

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 2.4 \\ -0.6 \end{pmatrix} - \begin{pmatrix} 0.936205 & -0.465684 \\ 0.108152 & 0.362870 \end{pmatrix} \begin{pmatrix} 0.040937 \\ 0.109193 \end{pmatrix} = \begin{pmatrix} 2.412524 \\ -0.644050 \end{pmatrix}.$$

Second iteration $D = 2.633224$,

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2.412524 \\ -0.644050 \end{pmatrix} - \begin{pmatrix} 0.916186 & -0.453129 \\ 0.122337 & 0.353998 \end{pmatrix} \begin{pmatrix} 0.000025 \\ -0.000556 \end{pmatrix} = \begin{pmatrix} 2.412249 \\ -0.643856 \end{pmatrix}$$

Third iteration $D = 2.632964$,

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2.412249 \\ -0.643856 \end{pmatrix} - \begin{pmatrix} 0.916172 & -0.453190 \\ 0.122268 & 0.354070 \end{pmatrix} \begin{pmatrix} 0.0000006 \\ -0.0000007 \end{pmatrix} = \begin{pmatrix} 2.412249 \\ -0.643856 \end{pmatrix}$$

Since the result is exact upto six decimal places, we have the solution

$$x = 2.412249 \pm 10^{-6}, \quad y = -0.643856 \pm 10^{-6}.$$

1.45 Calculate all solutions of the system

$$x^2 + y^2 = 1.12, \; xy = 0.23$$

correct to three decimal places. (Lund Univ., Sweden, BIT 20 (1980), 389)

Solution

From the system

$$f_1(x, y) = x^2 + y^2 - 1.12, \quad f_2(x, y) = xy - 0.23$$

we have the Jacobian matrix

$$\boldsymbol{J}(x_n, y_n) = \begin{bmatrix} 2x_n & 2y_n \\ y_n & x_n \end{bmatrix}$$

and $$\boldsymbol{J}^{-1}(x_n, y_n) = \frac{1}{2(x_n^2 - y_n^2)} \begin{bmatrix} x_n & -2y_n \\ -y_n & 2x_n \end{bmatrix}$$

Using the method

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \boldsymbol{J}^{-1}(x_n, y_n) \begin{pmatrix} f_1(x_n, y_n) \\ f_2(x_n, y_n) \end{pmatrix}, \; n = 0, 1, \ldots$$

and starting with $x_0 = 1$, $y_0 = 0.23$, we obtain the following sequence of iterates

First iteration

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.23 \end{pmatrix} - \begin{pmatrix} 0.52793 & -0.24285 \\ -0.12142 & 1.05585 \end{pmatrix} \begin{pmatrix} -0.0671 \\ 0.0 \end{pmatrix} = \begin{pmatrix} 1.03542 \\ 0.22185 \end{pmatrix}.$$

Second iteration

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1.03542 \\ 0.22185 \end{pmatrix} - \begin{pmatrix} 0.50613 & -0.21689 \\ -0.10844 & 1.01226 \end{pmatrix} \begin{pmatrix} 0.00131 \\ -0.00029 \end{pmatrix} = \begin{pmatrix} 1.03469 \\ 0.22229 \end{pmatrix}.$$

Third iteration

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1.03469 \\ 0.22229 \end{pmatrix} - \begin{pmatrix} 0.50662 & -0.21768 \\ -0.10884 & 1.01324 \end{pmatrix} \begin{pmatrix} -0.000004 \\ 0.000001 \end{pmatrix} = \begin{pmatrix} 1.03469 \\ 0.22229 \end{pmatrix}.$$

Hence, the solution correct to three decimal places is obtained as

$$x = 1.035, \; y = 0.222.$$

Hence, all solutions of the system are $\pm$ (1.035, 0.222).

1.46 Describe how, in general, suitable values of a, b, c and d may be estimated so that the sequence of values of x and y determined from the recurrence formula

$$x_{n+1} = x_n + a\,f(x_n, y_n) + bg(x_n, y_n)$$
$$y_{n+1} = y_n + c\,f(x_n, y_n) + dg(x_n, y_n)$$

will converge to a solution of

$$f(x, y) = 0, \quad g(x, y) = 0.$$

Illustrate the method by finding a suitable initial point and a recurrence relation of find the solution of

$$y = \sin (x + y), \quad x = \cos (y - x).$$

Solution

Let (ξ, η) be the exact solution of the system of equations

$$f(x, y) = 0, \quad g(x, y) = 0.$$

Substituting $x_n = \xi + \varepsilon_n$, $y_n = \eta + \eta_n$ in the iteration method

$$x_{n+1} = x_n + a\, f(x_n, y_n) + b\, g(x_n, y_n)$$
$$y_{n+1} = y_n + c\, f(x_n, y_n) + d\, g(x_n, y_n)$$

we obtain the error equations

$$\varepsilon_{n+1} = \left(1 + a\frac{\partial f}{\partial x} + b\frac{\partial g}{\partial x}\right)\varepsilon_n + \left(a\frac{\partial f}{\partial y} + b\frac{\partial g}{\partial y}\right)\eta_n + \ldots$$

$$\eta_{n+1} = \left(c\frac{\partial f}{\partial x} + d\frac{\partial g}{\partial x}\right)\varepsilon_n + \left(1 + c\frac{\partial f}{\partial y} + d\frac{\partial g}{\partial y}\right)\eta_n + \ldots$$

Convergence is obtained when

$$1 + a\frac{\partial f}{\partial x} + b\frac{\partial g}{\partial x} = 0,$$
$$a\frac{\partial f}{\partial y} + b\frac{\partial g}{\partial y} = 0,$$
$$c\frac{\partial f}{\partial x} + d\frac{\partial g}{\partial x} = 0,$$
$$1 + c\frac{\partial f}{\partial y} + d\frac{\partial g}{\partial y} = 0.$$

Solving the above system of equations, we obtain

$$a = -\frac{1}{D}\frac{\partial g}{\partial y}, \qquad b = \frac{1}{D}\frac{\partial f}{\partial y}, \qquad c = \frac{1}{D}\frac{\partial g}{\partial x}, \qquad d = -\frac{1}{D}\frac{\partial f}{\partial x},$$

where $$D = \frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial g}{\partial x}\frac{\partial f}{\partial y}.$$

Hence we get the iteration method

$$x_{n+1} = x_n + \frac{1}{D}\left[-f_n\frac{\partial g}{\partial y} + g_n\frac{\partial f}{\partial y}\right]$$

$$y_{n+1} = y_n + \frac{1}{D}\left[f_n\frac{\partial g}{\partial x} - g_n\frac{\partial f}{\partial x}\right], \quad n = 0, 1, \ldots$$

where the partial derivatives are to be evaluated at (x_n, y_n).

To find the initial approximation for the given system of equations

$$y = \sin (x + y), \quad x = \cos (y - x),$$

we approximate

$$\sin (x + y) \approx x + y$$

$$\cos (y - x) \approx 1 - \frac{1}{2}(y - x)^2$$

and obtain $y = x + y$ or $x = 0$,

$x = 1 - \frac{1}{2}(y - x)^2$ or $y = \sqrt{2}$.

We have $f(x, y) = y - \sin(x + y)$, $g(x, y) = x - \cos(y - x)$, and $f_x = -\cos(x + y)$,

$f_y = 1 - \cos(x + y)$, $g_x = 1 - \sin(y - x)$, $g_y = \sin(y - x)$.

1.47 Consider the system of equations $f(x, y) = 0$, $g(x, y) = 0$. Let $x = x_0 + \Delta x$ and $y = y_0 + \Delta y$, where (x_0, y_0) is an initial approximation to the solution. Assume

$$\Delta x = A_1(x_0, y_0) + A_2(x_0, y_0) + A_3(x_0, y_0) + \ldots$$
$$\Delta y = B_1(x_0, y_0) + B_2(x_0, y_0) + B_3(x_0, y_0) + \ldots$$

where $A_1(x_0, y_0)$, $B_1(x_0, y_0)$ are linear in f_0, g_0 ; $A_2(x_0, y_0)$, $B_2(x_0, y_0)$ are quadratic in f_0, g_0 and so on. Use Taylor series method to derive iterative methods of second and third order.

Solution

We have $f(x_0 + \Delta x, y_0 + \Delta y) \equiv 0$,

$g(x_0 + \Delta x, y_0 + \Delta y) \equiv 0$.

Expanding f and g in Taylor's series about the point (x_0, y_0), we get

$$f(x_0, y_0) + [\Delta x\, f_x + \Delta y\, f_y) + \frac{1}{2}[(\Delta x)^2 f_{xx} + 2\Delta x\, \Delta y\, f_{xy} + (\Delta y)^2 f_{yy}] + \ldots \equiv 0$$

$$g(x_0, y_0) + [\Delta x\, g_x + \Delta y g_y) + \frac{1}{2}[(\Delta x)^2 g_{xx} + 2\Delta x\, \Delta y\, g_{xy} + (\Delta y)^2 g_{yy}] + \ldots \equiv 0 \quad (1.61)$$

where partial derivatives are evaluated at (x_0, y_0).

Substituting

$$\Delta x = A_1 + A_2 + A_3 + \ldots$$
$$\Delta y = B_1 + B_2 + B_3 + \ldots$$

where A_i's and B_i's are arbitrary, we obtain

$$f_0 + (A_1 f_x + B_1 f_y) + \left[\frac{1}{2} A_1^2 f_{xx} + \frac{1}{2} B_1^2 f_{yy} + A_1 B_1 f_{xy} + A_2 f_x + B_2 f_y\right] + \ldots \equiv 0 \quad (1.62)$$

$$g_0 + (A_1 g_x + B_1 g_y) + \left[\frac{1}{2} A_1^2 g_{xx} + \frac{1}{2} B_1^2 g_{yy} + A_1 B_1 g_{xy} + A_2 g_x + B_2 g_y\right] + \ldots \equiv 0. \quad (1.63)$$

Setting the linear terms to zero, we get

$$A_1 f_x + B_1 f_y + f_0 = 0$$
$$A_1 g_x + B_1 g_y + g_0 = 0$$

which gives $$A_1 = -\left(\frac{f g_y - g f_y}{f_x g_y - g_x f_y}\right)_{(x_0, y_0)}$$

$$B_1 = -\left(\frac{g f_x - f g_x}{f_x g_y - g_x f_y}\right)_{(x_0, y_0)}$$

Hence, we obtain the second order method

$$x_1 = x_0 + (A_1)_0 \qquad \text{or} \qquad x_{k+1} = x_k + (A_1)_k,$$
$$y_1 = y_0 + (B_1)_0 \qquad\qquad\quad y_{k+1} = x_k + (B_1)_k.$$

Setting the quadratic terms in (1.62) and (1.63) to zero, we obtain

$$A_2 f_x + B_2 f_y + f_2 = 0$$
$$A_2 g_x + B_2 g_y + g_2 = 0$$

where $$f_2 = \frac{1}{2}(A_1^2 f_{xx} + 2A_1B_1 f_{xy} + B_1^2 f_{yy})$$
$$g_2 = \frac{1}{2}(A_1^2 g_{xx} + 2A_1B_1 g_{xy} + B_1^2 g_{yy})$$

are known values.

Solving the above equations, for A_2 and B_2, we get

$$A_2 = -\frac{f_2 g_y - g_2 f_y}{f_x g_y - g_x f_y}, \; B_2 = -\frac{g_2 f_x - f_2 g_x}{f_x g_y - g_x f_y}.$$

Hence we obtain the third order method

$$x_1 = x_0 + A_1 + A_2,$$
$$y_1 = y_0 + B_1 + B_2,$$

or $$x_{k+1} = x_k + A_1(x_k, y_k) + A_2(x_k, y_k),$$
$$y_{k+1} = x_k + B_1(x_k, y_k) + B_2(x_k, y_k).$$

1.48 Calculate the solution of the system of equations

$$x^3 + y^3 = 53$$
$$2y^3 + z^4 = 69$$
$$3x^5 + 10z^2 = 770$$

which is close to $x = 3, \quad y = 3, \quad z = 2.$

(Stockholm Univ., Sweden, BIT 19 (1979), 285)

Solution

Taking $$f_1(x, y, z) = x^3 + y^3 - 53,$$
$$f_2(x, y, z) = 2y^3 + z^4 - 69,$$
$$f_3(x, y, z) = 3x^5 + 10z^2 - 770,$$

we obtain the Jacobian matrix as

$$\boldsymbol{J} = \begin{bmatrix} 3x^2 & 3y^2 & 0 \\ 0 & 6y^2 & 4z^3 \\ 15x^4 & 0 & 20z \end{bmatrix}.$$

We write the Newton's method as

$$\mathbf{J^{(k)}} \; \Delta\mathbf{x} = -\,\mathbf{f}^{\,(k)}, \text{ where } \Delta\mathbf{x} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}.$$

For $k = 0$, with $x_0 = 3, y_0 = 3, z_0 = 2$, we get

$$\begin{bmatrix} 27 & 27 & 0 \\ 0 & 54 & 32 \\ 1215 & 0 & 40 \end{bmatrix} \Delta\mathbf{x} = -\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

The solution of the system is $\Delta\mathbf{x} = [-0.000195 \quad -0.036842 \quad 0.030921]^T$.

and $$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x} = [2.999805 \quad 2.963158 \quad 2.030921]^T.$$

For $k = 1$, we have $\mathbf{J}^{(1)} \; \Delta\mathbf{x} = -\,\mathbf{f}^{\,(1)}$ as

$$\begin{bmatrix} 26.998245 & 26.340916 & 0 \\ 0 & 52.681832 & 33.507273 \\ 1214.684131 & 0 & 40.618420 \end{bmatrix} \Delta\mathbf{x} = -\begin{bmatrix} 0.012167 \\ 0.047520 \\ 0.009507 \end{bmatrix}.$$

The solution of the system is

$$\Delta \mathbf{x} = [0.000014 \quad -0.000477 \quad -0.000669]^T.$$

and $\qquad \mathbf{x}^{(2)}\mathbf{x}^{(1)} + \Delta x = [2.999819 \quad 2.962681 \quad 2.030252]^T.$

1.49 (*a*) Take one step from a suitable point with Newton-Raphson's method applied to the system

$$\begin{aligned} 10x + \sin(x+y) &= 1 \\ 8y - \cos^2(z-y) &= 1 \\ 12z + \sin z &= 1 \end{aligned}$$

(*b*) Suggest some explicit method of the form $\mathbf{x}^{(k+1)} = \mathbf{F}(\mathbf{x}^{(k)})$ where no inversion is needed for **F**, and estimate how many iterations are required to obtain a solution correct to six decimal places from the starting point in (*a*).

(Uppsala Univ., Sweden, BIT 19 (1979), 139)

Solution

(*a*) To obtain a suitable starting point, we use the approximations

$$\sin(x+y) \approx 0, \ \cos(z-y) \approx 1, \ \sin(z) \approx 0,$$

and obtain from the given equations $x_0 = 1/10$, $y_0 = 1/4$, $z_0 = 1/12$.

Taking,
$$\begin{aligned} f_1(x, y, z) &= 10x + \sin(x+y) - 1, \\ f_2(x, y, z) &= 8y - \cos^2(z-y) - 1, \\ f_3(x, y, z) &= 12z + \sin z - 1, \end{aligned}$$

we write the Newton-Raphson method as

$$\mathbf{J}^{(\mathbf{k})} \Delta \mathbf{x} = -\mathbf{f}(\mathbf{x}^{(\mathbf{k})})$$

where
$$\mathbf{J} = \begin{bmatrix} 10 + \cos(x+y) & \cos(x+y) & 0 \\ 0 & 8 - \sin(2(z-y)) & \sin(2(z-y)) \\ 0 & 0 & 12 + \cos z \end{bmatrix}.$$

Taking the initial approximation as $x_0 = 1/10$, $y_0 = 1/4$, $z_0 = 1/12$, we get

$$\begin{bmatrix} 10.939373 & 0.939373 & 0 \\ 0 & 8.327195 & -0.327195 \\ 0 & 0 & 12.996530 \end{bmatrix} \Delta \mathbf{x} = -\begin{bmatrix} 0.342898 \\ 0.027522 \\ 0.083237 \end{bmatrix}.$$

We solve the third equation for Δx_3, then the second equation for Δx_2 and then the first equation for Δx_1. We obtain the solution as

$$\Delta \mathbf{x} = [-0.031040 \quad -0.003557 \quad -0.006405]^T,$$

and $\qquad \mathbf{x}^{(1)} = [0.068960 \quad 0.246443 \quad 0.076928]^T.$

(*b*) We write the explicit method in the form

$$\begin{aligned} x_{n+1} &= \frac{1}{10}[1 - \sin(x_n + y_n)] = f_1(x_n, y_n, z_n), \\ y_{n+1} &= \frac{1}{8}[1 + \cos^2(z_n - y_n)] = f_2(x_n, y_n, z_n), \\ z_{n+1} &= \frac{1}{12}[1 + \sin(z_n)] = f_3(x_n, y_n, z_n). \end{aligned}$$

Starting with the initial point $\mathbf{x}^{(0)} = [1/10, 1/4, 1/12]^T$, we obtain the following sequence of iterates

$$\mathbf{x}^{(1)} = [0.065710,\ 0.246560,\ 0.076397]^T$$
$$\mathbf{x}^{(2)} = [0.069278,\ 0.246415,\ 0.076973]^T$$
$$\mathbf{x}^{(3)} = [0.068952,\ 0.246445,\ 0.076925]^T$$
$$\mathbf{x}^{(4)} = [0.068980,\ 0.246442,\ 0.076929]^T$$
$$\mathbf{x}^{(5)} = [0.068978,\ 0.246442,\ 0.076929]^T.$$

Hence, the solution correct to six decimal places is obtained after five iterations.

Polynomial equations

1.50 Obtain the number of real roots between 0 and 3 of the equation

$$P(x) = x^4 - 4x^3 + 3x^2 + 4x - 4 = 0$$

using Sturm's sequence.

Solution

For the given polynomial, we obtain the Sturm's sequence

$$f(x) = x^4 - 4x^3 + 3x^2 + 4x - 4,$$
$$f_1(x) = 2x^3 - 6x^2 + 3x + 2,$$
$$f_2(x) = x^2 - 3x + 2,$$
$$f_3(x) = x - 2,$$
$$f_4(x) = 0.$$

Since $f_4(x) = 0$, we find that $x = 2$ is a multiple root of $f(x) = 0$ with multiplicity 2. Dividing the elements in the Sturm's sequence by $x - 2$, we obtain the new Sturm's sequence as

$$f^*(x) = x^3 - 2x^2 - x + 2,$$
$$f_1^*(x) = 2x^2 - 2x - 1,$$
$$f_2^*(x) = x - 1,$$
$$f_3^*(x) = 1.$$

The changes in signs of f_i^* are given in the following table.

x	f^*	f_1^*	f_2^*	f_3^*	V
0	+	–	–	+	2
3	+	+	+	+	0

Using Sturm's theorem, we find that there are two real roots between 0 and 3.

Hence, the polynomial $f(x) = 0$ has 3 real roots between 0 and 3. One root is 2 which is a double root.

1.51 Determine the multiplicity of the root $\xi = 1$, of the polynomial

$$P(x) = x^5 - 2x^4 + 4x^3 - x^2 - 7x + 5 = 0$$

using synthetic division. Find also $P'(2)$ and $P''(2)$.

Solution

Using the synthetic division method, we obtain

1	1	– 2	4	– 1	– 7	5	
		1	– 1	3	2	– 5	
	1	– 1	3	2	– 5	0 = P (1)	
		1	0	3	5		
	1	0	3	5	0 = $P\,'(1)$		
		1	1	4			
	1	1	4	9 = $P\,''(1) / 2$			

Since $P(1) = P\,'(1) = 0$ and $P\,''(1) \neq 0$, the root 1 is a double root of $P(x) = 0$.

To find $P\,'(2)$ and $P\,''(2)$, we again use the synthetic division method

2	1	– 2	4	– 1	– 7	5	
		2	0	8	14	14	
	1	0	4	7	7	19 = P (2)	
		2	4	16	46		
	1	2	8	23	53 = $P\,'(2)$		
		2	8	32			
	1	4	16	55 = $P\,''(2) / 2$			

Hence, we get $P\,'(2) = 53$ and $P\,''(2) = 110$.

1.52 Use the Birge-Vieta method to find a real root correct to three decimals of the following equations :

(*i*) $x^3 - 11x^2 + 32x - 22 = 0$, $p = 0.5$, (*ii*) $x^5 - x + 1 = 0$, $p = -1.5$,

(*iii*) $x^6 - x^4 - x^3 - 1 = 0$, $p = 1.5$

Find the deflated polynomial in each case.

Solution

Using the Birge-Vieta method (1.29),

$$p_{k+1} = p_k - \frac{b_n}{c_{n-1}}, \quad k = 0, 1, ...$$

(where $b_n = f(x_k)$, $c_{n-1} = f'(x_{k-1})$) and the synthetic division, we obtain the following approximations.

(*i*) *First iteration* $p_0 = 0.5$.

0.5	1	– 11	32	– 22
		0.5	– 5.25	13.375
	1	– 10.5	26.75	– 8.625
		0.5	– 5.00	
	1	– 10.0	21.75	

$$p_1 = 0.5 + \frac{8.625}{21.75} = 0.8966.$$

Second iteration $p_1 = 0.8966$.

0.8966	1	– 11	32	– 22
		0.8966	– 9.0587	20.5692
	1	– 10.1034	22.9413	– 1.4308
		0.8966	– 8.2548	
	1	– 9.2068	14.6865	

$$p_2 = 0.8966 + \frac{1.4308}{14.6865} = 0.9940.$$

Third iteration $p_2 = 0.9940$.

0.9940	1	– 11	32	– 22
		0.9940	– 9.9460	21.9217
	1	– 10.0060	22.0540	– 0.0783
		0.9940	– 8.9579	
	1	– 9.0120	13.0961	

$$p_3 = 0.9940 + \frac{0.0783}{13.0961} = 0.99998.$$

The root correct to three decimals is 1.00.

Deflated polynomial

1	1	– 11	32	– 22
		1	– 10	
	1	– 10	22	

The deflated polynomial is $x^2 - 10x + 22$.

(*ii*) *First iteration* $p_0 = -1.5$.

– 1.5	1	0	0	0	– 1	1
		– 1.5	2.25	– 3.375	5.0625	– 6.0938
	1	– 1.5	2.25	– 3.375	4.0625	– 5.0938
		– 1.5	4.5	– 10.125	20.25	
	1	– 3	6.75	– 13.5	24.3125	

$$p_1 = -1.5 + \frac{5.0938}{24.3125} = -1.2905,$$

Second iteration $p_1 = -1.2905$.

– 1.2905	1	0	0	0	– 1	1
		– 1.2905	1.6654	– 2.1492	2.7735	– 2.2887
	1	– 1.2905	1.6654	– 2.1492	1.7735	– 1.2887
		– 1.2905	3.3308	– 6.4476	11.0941	
	1	– 2.5810	4.9962	– 8.5968	12.8676	

$$p_2 = -1.2905 + \frac{1.2887}{12.8676} = -1.1903.$$

Third iteration $p_2 = -1.1903$

– 1.1903	1	0	0	0	–1	1
		– 1.1903	1.4168	– 1.6864	2.0073	– 1.1990
	1	– 1.1903	1.4168	– 1.6864	1.0073	– 0.1990
		– 1.1903	2.8336	– 5.0593	8.0294	
	1	– 2.3806	4.2504	– 6.7457	8.0367	

$$p_3 = -1.1903 + \frac{0.1990}{9.0367} = -1.1683.$$

Fourth iteration $p_3 = -1.1683$.

– 1.1683	1	0	0	0	–1	1
		– 1.1683	1.3649	– 1.5946	1.8630	– 1.0082
	1	– 1.1683	1.3649	– 1.5946	0.8630	– 0.0082
		– 1.1683	2.7298	– 4.7838	7.4519	
	1	– 2.3366	4.0947	– 6.3784	9.3149	

$$p_3 = -1.1683 + \frac{0.0082}{8.3149} = -1.1673.$$

The root correct to three decimals is – 1.167.

Deflated polynomial

– 1.167	1	0	0	0	– 1	1
		– 1.167	1.3619	– 1.5893	1.8547	
	1	– 1.167	1.3619	– 1.5893	0.8547	

The deflated polynomial is given by

$$x^4 - 1.167x^3 + 1.3619x^2 - 1.5893x + 0.8547 = 0.$$

(*iii*) *First iteration* $p_0 = 1.5$.

1.5	1	0	– 1	– 1	0	0	– 1
		1.5	2.25	1.875	1.3125	1.9688	2.9532
	1	1.5	1.25	0.875	1.3125	1.9688	1.9532
		1.5	4.5	8.625	14.25	23.3438	
	1	3	5.75	9.5	15.5625	25.3126	

$$p_1 = 1.5 - \frac{1.9532}{25.3126} = 1.4228.$$

Second iteration $p_1 = 1.4228$.

1.4228	1	0	– 1	– 1	0	0	– 1
		1.4228	2.0244	1.4575	0.6509	0.9261	1.3177
	1	1.4228	1.0244	0.4575	0.6509	0.9261	0.3177
		1.4228	4.0487	7.2180	10.9207	16.4641	
	1	2.8456	5.0731	7.6755	11.5716	17.3902	

$$p_2 = 1.4228 - \frac{0.3177}{17.3902} = 1.4045.$$

Third iteration $p_2 = 1.4045$.

1.4045	1	0	– 1	– 1	0	0	– 1
		1.4045	1.9726	1.3660	0.5140	0.7219	1.0139
	1	1.4045	0.9726	0.3660	0.5140	0.7219	0.0139
		1.4045	3.9452	6.9071	10.2151	15.0690	
	1	2.8090	4.9178	7.2731	10.7291	15.7909	

$$p_3 = 1.4045 - \frac{0.0139}{15.7909} = 1.4036.$$

The root correct to three decimals is 1.404.

Deflated polynomial

1.404	1	0	– 1	– 1	0	0	– 1
		1.404	1.9712	1.3636	0.5105	0.7167	
	1	1.404	0.9712	0.3636	0.5105	0.7167	

The deflated polynomial is given by

$$x^5 + 1.404\, x^4 + 0.9712\, x^3 + 0.3636\, x^2 + 0.5105x + 0.7167 = 0.$$

1.53 Find to two decimals the real and complex roots of the equation $x^5 = 3x - 1$.

Solution

Taking $f(x) = x^5 - 3x + 1$, we find that $f(x)$ has two sign changes in the coefficients and thus can have a maximum of two positive real roots. Also, $f(-x) = -x^5 + 3x + 1$ has only one change of sign and hence has one negative real root.

We find that

$$f(-2) < 0, f(-1) > 0, f(0) > 0, f(1) < 0 \text{ and } f(2) > 0.$$

Thus, $f(x) = 0$ has one negative real root in $(-2, -1)$ and two positive real roots in the intervals $(0, 1)$ and $(1, 2)$ respectively. Hence, the given polynomial has three real roots and a pair of complex roots.

We first determine the real roots using the Birge-Vieta method (1.29) :

$$p_{k+1} = p_k - \frac{b_n}{c_{n-1}}, \quad k = 0, 1, \ldots$$

First real root. Let $p_0 = 0$.

0	1	0	0	0	– 3	1
		0	0	0	0	0
	1	0	0	0	– 3	$1 = b_n$
		0	0	0	0	
	1	0	0	0	$- 3 = c_{n-1}$	

$$p_1 = 0.3333.$$

0.3333	1	0	0	0	– 3	1
		0.3333	0.1111	0.0370	0.0123	– 0.9958
	1	0.3333	0.1111	0.0370	– 2.9877	$0.0042 = b_n$
		0.3333	0.2222	0.1111	0.0494	
	1	0.6666	0.3333	0.1481	$- 2.9383 = c_{n-1}$	

$$p_2 = 0.3333 + \frac{0.0042}{2.9383} = 0.3347.$$

Hence, the root correct to two decimals is 0.33. Let the root be taken as 0.3347.

First deflated polynomial

0.3347	1	0	0	0	– 3	1
		0.3347	0.1120	0.0375	0.0125	– 0.9999
	1	0.3347	0.1120	0.0375	– 2.9875	0.0001

The deflated polynomial is

$$x^4 + 0.3347x^3 + 0.1120x^2 + 0.0375x - 2.9875 = 0$$

with the error in satisfying the original equation as $f(0.3347) = 0.0001$.

We now find the second root using the deflated polynomial.

Second real root. Let $p_0 = 1.2$

1.2	1	0.3347	0.1120	0.0375	– 2.9875
		1.2	1.8416	2.3444	2.8583
	1	1.5347	1.9536	2.3819	$- 0.1292 = b_n$
		1.2	3.2816	6.2822	
	1	2.7347	5.2352	$8.6641 = c_{n-1}$	

$$p_1 = 1.2 + \frac{0.1292}{8.6641} = 1.2149.$$

1.2149	1	0.3347	0.1120	0.0375	– 2.9875
		1.2149	1.8826	2.4233	2.9896
	1	1.5496	1.9946	2.4608	$0.0021 = b_n$
		1.2149	3.3586	6.5036	
	1	2.7645	5.3532	$8.9644 = c_{n-1}$	

$$p_2 = 1.2149 - \frac{0.0021}{8.9644} = 1.2147.$$

Hence, the root correct to two decimals is 1.21.

Let the root be taken as 1.2147.

Second deflated polynomial

1.2147	1	0.3347	0.1120	0.0375	– 2.9875
		1.2147	1.8821	2.4222	2.9878
	1	1.5494	1.9941	2.4597	0.0003

The deflated polynomial now is

$$x^3 + 1.5494x^2 + 1.9941x + 2.4597 = 0$$

with the error P_4 (1.2147) = 0.0003.

We now find the third root using this deflated polynomial.

Third real root. Let $p_0 = -1.4$.

– 1.4	1	1.5494	1.9941	2.4597
		– 1.4	– 0.2092	– 2.4989
	1	0.1494	1.7849	– 0.0392 = b_n
		– 1.4	1.7508	
	1	– 1.2506	3.5357 = c_{n-1}	

$$p_1 = -1.4 + \frac{0.0392}{3.5357} = -1.3889.$$

– 1.3889	1	1.5494	1.9941	2.4597
		– 1.3889	– 0.2229	– 2.4600
	1	0.1605	1.7712	0.0003 = b_n
		– 1.3889	1.7061	
	1	– 1.2284	3.4773 = c_{n-1}	

$$p_2 = -1.3889 + \frac{0.0003}{3.4773} = -1.3888.$$

Hence, the root correct to two decimals is – 1.39.

Let the root be taken as – 1.3888.

We now determine the next deflated polynomial

– 1.3888	1	1.5494	1.9941	2.4597
		– 1.3888	– 0.2230	– 2.4597
	1	0.1606	1.7711	0.0000

The final deflated polynomial is

$$x^2 + 0.1606x + 1.7711 = 0$$

whose roots are $-0.0803 \pm 1.3284i$.

Hence, the roots are

$$0.3347,\ 1.2147,\ -1.3888,\ -0.0803 \pm 1.3284i.$$

Rounding to two places, we may have the roots as

$$0.33,\ 1.21,\ -1.39,\ -0.08 \pm 1.33i.$$

1.54 Carry out two iterations of the Chebyshev method, the multipoint methods (1.11) and (1.12) for finding the root of the polynomial equation $x^3 - 2 = 0$ with $x_0 = 1$, using synthetic division.

Solution

Chebyshev method (1.10) is given by

$$x_{k+1} = x_k - \frac{f_k}{f_k'} - \frac{1}{2}\frac{f_k^2 f_k''}{(f_k')^3}, \quad k = 0, 1, \ldots$$

We use synthetic division method to find f_k, f_k' and f_k''.

(*i*) *First iteration* $x_0 = 1$

$$\begin{array}{c|cccc}
1 & 1 & 0 & 0 & -2 \\
 & & 1 & 1 & 1 \\
\hline
 & 1 & 1 & 1 & -1 = f_k \\
 & & 1 & 2 & \\
\cline{2-4}
 & 1 & 2 & 3 = f_k' & \\
 & & 1 & & \\
\cline{2-3}
 & 1 & 3 = f_k''/2 & &
\end{array}$$

$$x_1 = 1 + \frac{1}{3} - \frac{1}{2}\left(\frac{6}{27}\right) = 1.2222.$$

Second iteration

$$\begin{array}{c|cccc}
1.2222 & 1 & 0 & 0 & -2 \\
 & & 1.2222 & 1.4938 & 1.8257 \\
\hline
 & 1 & 1.2222 & 1.4938 & -0.1743 = f_k \\
 & & 1.2222 & 2.9875 & \\
\cline{2-4}
 & 1 & 2.4444 & 4.4813 = f_k' & \\
 & & 1.2222 & & \\
\cline{2-3}
 & 1 & 3.6666 = f_k''/2 & &
\end{array}$$

$$x_2 = 1.2222 + \frac{0.1743}{4.4813} - \frac{1}{2} \cdot \frac{(-0.1743)^2 (7.3332)}{(4.4813)^3} = 1.2599.$$

(*ii*) Multipoint method (1.11) gives the iteration scheme

$$x_{k+1}^{*} = x_k - \frac{1}{2}\frac{f_k}{f_k'}$$

$$x_{k+1} = x_k - \frac{f_k}{f'(x_{k+1}^{*})}$$

We calculate f_k, f_k', $f'(x_{k+1}^{*})$ using synthetic division method.

First iteration $x_0 = 1$.

The values of f_k, f_k' can be taken from the first iteration of Chebyshev method. We have

$$x_1^* = 1 + \frac{1}{2} \cdot \frac{1}{3} = 1.1667,$$

$$\begin{array}{r|llll}
1.1667 & 1 & 0 & 0 & -2 \\
 & & 1.1667 & 1.3612 & 1.5881 \\
\hline
 & 1 & 1.1667 & 1.3612 & -0.4119 = f(x_{k+1}^*) \\
 & & 1.1667 & 2.7224 & \\
\hline
 & 1 & 2.3334 & 4.0836 = f'(x_{k+1}^*) &
\end{array}$$

$$x_1 = 1 + \frac{1.0}{4.0836} = 1.2449.$$

Second iteration $x_1 = 1.2449$.

$$\begin{array}{r|llll}
1.2449 & 1 & 0 & 0 & -2 \\
 & & 1.2449 & 1.5498 & 1.9293 \\
\hline
 & 1 & 1.2449 & 1.5498 & -0.0707 = f_k \\
 & & 1.2449 & 3.0996 & \\
\hline
 & 1 & 2.4898 & 4.6494 = f_k' &
\end{array}$$

$$x_2^* = 1.2449 + \frac{1}{2} \cdot \frac{0.0707}{4.6494} = 1.2525.$$

$$\begin{array}{r|llll}
1.2525 & 1 & 0 & 0 & -2 \\
 & & 1.2525 & 1.5688 & 1.9649 \\
\hline
 & 1 & 1.2525 & 1.5688 & -0.0351 = f(x_{k+1}^*) \\
 & & 1.2525 & 3.1375 & \\
\hline
 & 1 & 2.5050 & 4.7063 = f'(x_{k+1}^*) &
\end{array}$$

$$x_2 = 1.2449 + \frac{0.0707}{4.7063} = 1.2599.$$

(*iii*) Multipoint method (1.12) gives the iteration scheme

$$x_{k+1}^* = x_k - \frac{f_k}{f_k'}$$

$$x_{k+1} = x_{k+1}^* - \frac{f(x_{k+1}^*)}{f_k'}.$$

First iteration $x_0 = 1$.

The values of f_k, f_k' can be taken from the first iteration of Chebyshev method, We have

$$x_1^* = 1 + \frac{1}{3} = 1.3333.$$

$$\begin{array}{r|rrrr}
1.3333 & 1 & 0 & 0 & -2 \\
 & & 1.3333 & 1.7777 & 2.3702 \\
\hline
 & 1 & 1.3333 & 1.7777 & 0.3702 = f_1^*
\end{array}$$

$$x_1 = 1.3333 - \frac{0.3702}{3} = 1.2099.$$

Second iteration $x_1 = 1.2099$.

$$\begin{array}{r|rrrr}
1.2099 & 1 & 0 & 0 & -2 \\
 & & 1.2099 & 1.4639 & 1.7712 \\
\hline
 & 1 & 1.2099 & 1.4639 & -0.2288 = f_k \\
 & & 1.2099 & 2.9277 & \\
\hline
 & 1 & 2.4198 & 4.3916 = f_k' &
\end{array}$$

$$x_2^* = 1.2099 + \frac{0.2288}{4.3916} = 1.2620$$

$$\begin{array}{r|rrrr}
1.2620 & 1 & 0 & 0 & -2 \\
 & & 1.2620 & 1.5926 & 2.0099 \\
\hline
 & 1 & 1.2620 & 1.5926 & 0.0099 = f(x_{k+1}^*)
\end{array}$$

$$x_2 = 1.2620 - \frac{0.0099}{4.3916} = 1.2597.$$

1.55 It is given that the polynomial equation

$$9x^4 + 12x^3 + 13x^2 + 12x + 4 = 0$$

has a double root near -0.5. Perform three iterations to find this root using (*i*) Birge-Vieta method, (*ii*) Chebyshev method, for multiple roots. Find the deflated polynomial in each case.

Solution

(*i*) *Birge-Vieta method* $p_0 = -0.5$

First iteration

$$\begin{array}{r|rrrrr}
-0.5 & 9 & 12 & 13 & 12 & 4 \\
 & & -4.5 & -3.75 & -4.625 & -3.6875 \\
\hline
 & 9 & 7.5 & 9.25 & 7.375 & 0.3125 = b_4 \\
 & & -4.5 & -1.5 & -3.875 & \\
\hline
 & 9 & 3.0 & 7.75 & 3.5 = c_3 &
\end{array}$$

$$p_1 = p_0 - 2\left(\frac{b_4}{c_3}\right) = -0.5 - \frac{2(0.3125)}{3.5} = -0.6786.$$

Second iteration

– 0.6786	9	12	13	12	4
		– 6.1074	– 3.9987	– 6.1083	– 3.9981
	9	5.8926	9.0013	5.8917	0.0019 = b_4
		– 6.1074	0.1458	– 6.2072	
	9	– 0.2148	9.1471	– 0.3155 = c_3	

$$p_2 = p_1 - 2\left(\frac{b_4}{c_3}\right) = -0.6786 - 2\left(\frac{0.0019}{-0.3155}\right) = -0.6666.$$

Third iteration

– 0.6666	9	12	13	12	4
		– 5.9994	– 4.0000	– 5.9994	– 4.0000
	9	6.0006	9.0000	6.0006	0.0 = b_4
		– 5.9994	– 0.0008	– 5.9989	
	9	0.0012	8.9992	0.0017 = c_3	

Since, $b_4 = 0$, the root is – 0.6666. Again, since $b_4 = 0$ and $c_3 \approx 0$, the deflated polynomial is

$$9x^2 + 0.0012x + 8.9992 = 0.$$

(*ii*) *Chebyshev method* $p_0 = -0.5$.

First iteration

– 0.5	9	12	13	12	4
		– 4.5	– 3.75	– 4.625	– 3.6875
	9	7.5	9.25	7.375	0.3125 = b_4
		– 4.5	– 1.5	– 3.875	
	9	3.0	7.75	3.5 = c_3	
		– 4.5	0.75		
	9	– 1.5	8.50 = $d_2 = P''(p_0)/2$		

Using (1.19) for $m = 2$, we have the method as

$$p_1 = p_0 - \frac{b_4}{c_3} - 4\left(\frac{b_4}{c_3}\right)^2\left(\frac{d_2}{c_3}\right)$$

$$= -0.5 - \left(\frac{0.3125}{3.5}\right) - 4\left(\frac{0.3125}{3.5}\right)^2\left(\frac{8.5}{3.5}\right) = -0.6667.$$

Second iteration

– 0.6667	9	12	13	12	.4
		– 6.0003	– 4.0000	– 6.0003	– 4.0000
	9	5.9997	9.0000	5.9997	0.0 = b_4
		– 6.0003	0.0004	– 6.0006	
	9	– 0.0006	9.0004	– 0.0009 = c_3	

Since $b_4 = 0$, the root is – 0.6667. Again, since $b_4 = 0$ and $c_3 \approx 0$, the deflated polynomial is given by

$$9x^2 - 0.0006x + 9.0004 = 0.$$

or
$$x^2 - 0.0007x + 1.00004 = 0.$$

The exact deflated polynomial equation is $x^2 + 1 = 0$.

1.56 Given the two polynomial

$$P(x) = x^6 - 4.8x^4 + 3.3x^2 - 0.05$$

and
$$Q(x, h) = x^6 - (4.8 - h)x^4 + (3.3 + h)x^2 - (0.05 - h)$$

(*a*) Calculate all the roots of P.

(*b*) When $h << 1$, the roots of Q are close to those of P. Estimate the difference between the smallest positive root of P and the corresponding root of Q.

(Denmark Tekniske Hojskole, Denmark, BIT 19 (1979), 139)

Solution

(*a*) Writing $x^2 = t$, we have the polynomial

$$P(t) = t^3 - 4.8t^2 + 3.3t - 0.05.$$

Using Graeffe's root squaring method, we obtain

0	1	– 4.8	3.3	– 0.05
	1	23.04 – 6.6	10.89 – 0.48	0.0025
1	1	16.44	10.41	0.0025
	1	270.2736 – 20.82	108.3681 – 0.0822	0.6250×10^{-5}
2	1	249.4536	108.2859	0.6250×10^{-5}
	1	62227.0986 – 216.5718	11725.8361 – 0.0031	0.3906×10^{-10}
3	1	62010.5268	11725.8330	0.3906×10^{-10}

Hence, the roots of $P(t)$ are obtained as

$$t_1^{(1)} = 4.0546, \quad t_2^{(1)} = 0.7957, \quad t_3^{(1)} = 0.0155,$$
$$t_1^{(2)} = 3.9742, \quad t_2^{(2)} = 0.8117, \quad t_3^{(2)} = 0.0155,$$
$$t_1^{(3)} = 3.9724, \quad t_2^{(3)} = 0.8121, \quad t_3^{(3)} = 0.0155.$$

Substituting in $P(t)$, we find that all the roots are positive.

Hence, the roots of the given polynomial $P(x)$ may be taken as

$$\pm 1.9931, \pm 0.9012, \pm 0.1245.$$

(*b*) When $h << 1$, the roots of Q are close to those of P. The approximation to the smallest positive root x_1 of P and x_1^* of Q can be approximated from

$$P = 3.3x^2 - 0.05 = 0$$
$$Q = (3.3 + h)x^2 - (0.05 - h) = 0.$$

We obtain $\qquad x_1 = \sqrt{\dfrac{0.05}{3.3}},\ x_1^* = \sqrt{\dfrac{0.05-h}{3.3+h}}.$

Hence
$$x_1 - x_1^* = \sqrt{\frac{0.05}{3.3}} - \sqrt{\frac{0.05-h}{3.3+h}}$$
$$= \sqrt{\frac{0.05}{3.3}}\left[1-\sqrt{\left(1-\frac{h}{0.05}\right)\left(1+\frac{h}{3.3}\right)^{-1}}\right]$$
$$= \sqrt{\frac{0.05}{3.3}}\left[1-\left(1-\frac{6.7}{0.66}h+\ldots\right)\right] \approx \sqrt{\frac{0.05}{3.3}}\cdot\frac{6.7}{0.66}h \approx 1.25h$$

1.57 Using Bairstow's method obtain the quadratic factor of the following equations (Perform two iterations)

(*i*) $x^4 - 3x^3 + 20x^2 + 44x + 54 = 0$ with $(p, q) = (2, 2)$

(*ii*) $x^4 - x^3 + 6x^2 + 5x + 10 = 0$ with $(p, q) = (1.14, 1.42)$

(*iii*) $x^3 - 3.7x^2 + 6.25x - 4.069 = 0$ with $(p, q) = (-2.5, 3)$.

Solution

Bairstow's method (1.30) for finding a quadratic factor of the polynomial of degree n is given by
$$p_{k+1} = p_k + \Delta p$$
$$q_{k+1} = q_k + \Delta q,\ k = 0.1, \ldots$$
where
$$\Delta p = -\frac{b_n c_{n-3} - b_{n-1}\, c_{n-2}}{c_{n-2}^2 - c_{n-3}\,(c_{n-1} - b_{n-1})}$$
$$\Delta q = -\frac{b_{n-1}\,(c_{n-1} - b_{n-1}) - b_n c_{n-2}}{c_{n-2}^2 - c_{n-3}\,(c_{n-1} - b_{n-1})}$$
We use the synthetic division method to determine b_i's and c_i's.

(*i*) *First iteration* $p_0 = 2$, $q_0 = 2$

– 2	1	– 3	20	44	54
– 2		– 2	10	– 56	4
			– 2	10	– 56
	1	– 5	28	– 2 = b_{n-1}	2 = b_n
		– 2	14	– 80	
			– 2	14	
	1	– 7	40	– 68 = c_{n-1}	

$\Delta p = -0.0580$, $\Delta q = -0.0457$,

$p_1 = 1.9420$, $q_1 = 1.9543$.

Second iteration $p_1 = 1.9420$, $q_1 = 1.9543$

– 1.9420	1	– 3	20	44	54
– 1.9543		– 1.9420	9.597364	– 53.682830	0.047927
			– 1.9543	9.658151	– 54.022840

1	– 4.9420	27.643064	– 0.024679 = b_{n-1}	0.025087 = b_n
	– 1.9420	13.368728	– 75.849649	
		– 1.9543	13.453401	
1	– 6.8840	39.057492	– 62.420927 = c_{n-1}	

$\Delta p = -0.00072$, $\Delta q = -0.000511$ $p_2 = 1.9413$, $q_2 = 1.9538$.

(*ii*) *First iteration* $p_0 = 1.14$, $q_0 = 1.42$.

– 1.14	1	– 1	6	5	10
– 1.42		– 1.14	2.4396	– 8.0023	– 0.0416
			– 1.42	3.0388	– 9.9678
	1	– 2.14	7.0196	0.0365 = b_{n-1}	– 0.0094 = b_n
		– 1.14	3.7392	– 10.6462	
			– 1.42	4.6576	
	1	– 3.28	9.3388	– 5.9521 = c_{n-1}	

$\Delta p = 0.0046$, $\Delta q = 0.0019$, $p_1 = 1.1446$, $q_1 = 1.4219$.

Second iteration $p_1 = 1.1446$, $q_1 = 1.4219$

– 1.1446	1	– 1	6	5	10
– 1.4219		– 1.1446	2.4547	– 8.0498	0.0005
			– 1.4219	3.0494	– 10.0
	1	– 2.1446	7.0328	– 0.0004 = b_{n-1}	0.0005 = b_n
		– 1.1446	3.7648	– 10.7314	
			– 1.4219	4.6769	
	1	– 3.2892	9.3757	– 6.0549 = c_{n-1}	

$\Delta p = -0.00003$, $\Delta q = 0.00003$, $p_2 = 1.1446$, $q_2 = 1.4219$.

(*iii*) *First iteration* $p_0 = -2.5$, $q_0 = 3.0$.

2.5	1	– 3.7	6.25	– 4.069
– 3.0		2.5	– 3.0	0.625
			– 3.0	3.6
	1	– 1.2	0.25 = b_{n-1}	0.156 = b_n
		2.5	3.25	
			– 3.0	
	1	1.3	0.50 = c_{n-1}	

$\Delta p = 0.1174$, $\Delta q = 0.0974$, $p_1 = -2.3826$, $q_1 = 3.0974$.

Second iteration $p_1 = -2.3826$, $q_1 = 3.0974$.

$$
\begin{array}{r|llll}
2.3826 & 1 & -3.7 & 6.25 & -4.069 \\
-3.0974 & & 2.3826 & -3.1388 & 0.0329 \\
 & & & -3.0974 & 4.0805 \\
\hline
 & 1 & -1.3174 & 0.0138 = b_{n-1} & 0.0444 = b_n \\
 & & 2.3826 & 2.5379 & \\
 & & & -3.0974 & \\
\hline
 & 1 & 1.0652 & -0.5457 = c_{n-1} &
\end{array}
$$

$$\Delta p = -0.0175, \Delta q = 0.0325, \; p_2 = -2.4001, q_2 = 3.1299.$$

1.58 Find all the roots of the polynomial

$$x^3 - 6x^2 + 11x - 6 = 0$$

using the Graeffe's root squaring method.

The coefficients of the successive root squarings are given below.

Coefficients in the root squarings by Graeffe's method

m	2^m				
0	1	1	– 6	11	– 6
		1	36	121	36
			– 22	– 72	
1	2	1	14	49	36
		1	196	2401	1296
			– 98	– 1008	
2	4	1	98	1393	1296
		1	9604	1940449	1679616
			– 2786	– 254016	
3	8	1	6818	1686433	1679616
		1	46485124	2.8440562(12)	2.8211099(12)
			– 3372866	– 2.2903243(10)	
4	16	1	43112258	2.8211530(12)	2.8211099(12)

Successive approximations to the roots are given below. The exact roots of the equation are 3, 2, 1.

Approximations to the roots

m	α_1	α_2	α_3
1	3.7417	1.8708	0.8571
2	3.1463	1.9417	0.9821
3	3.0144	1.9914	0.9995
4	3.0003	1.9998	1.0000

1.59 Apply the Graeffe's root squaring method to find the roots of the following equations correct to two decimals :

(*i*) $x^3 - 2x + 2 = 0$, (*ii*) $x^3 + 3x^2 - 4 = 0$.

Solution

(*i*) Using Graeffe's root squaring method, we get the following results :

m	2^m				
0	1	1	0	– 2	2
		1	0	4	4
			4	0	
1	2	1	4	4	4
		1	16	16	16
			– 8	– 32	
2	4	1	8	– 16	16
		1	64	256	256
			32	– 256	
3	8	1	96	0	256
		1	9216	0	65536
			0	– 49152	
4	16	$1 = B_0$	$9216 = B_1$	$-49152 = B_2$	$65536 = B_3$

Since B_2 is alternately positive and negative, we have a pair of complex roots based on B_1, B_2, B_3.

One real root is $|\xi_1|^{16} = 9216$ or $|\xi_1| = 1.7692$. On substituting into the given polynomial, we find that root must be negative. Hence, one real is $\xi_1 = -1.7692$.

To find the pair of complex roots $p \pm iq$, we have

$$|\beta|^{32} = \left|\frac{B_3}{B_1}\right| \quad \text{or} \quad \beta = 1.0632 = \sqrt{p^2 + q^2}\,.$$

Also, $\xi_1 + 2p = 0$ or $p = 0.8846$,

$q^2 = \beta^2 - p^2$ or $q = 0.5898$.

Hence, roots are $0.8846 \pm 0.5898i$.

(*ii*)

m	2^m				
0	1	1	3	0	– 4
		1	9	0	16
			0	24	
1	2	1	9	24	16
		1	81	576	256
			– 48	– 288	

2	4	1	33	288	256
		1	1089 – 576	82944 – 16896	65536
3	8	1	513	66048	65536
		1	263169 – 132096	4362338304 – 67239936	16^8
4	16	1	131073 almost half	4295098368	16^8
		$= B_0$	$= B_1$	$= B_2$	$= B_3$

Since B_1 is almost half of the corresponding value in the previous squaring, it indicates that there is a double root based on B_0, B_1 and B_2. Thus, we obtain one double root as

$$|\,\xi_1\,|^{32} = |\,\xi_2\,|^{32} = |\,B_2\,|$$

which gives $|\,\xi_1\,| = 2.0000$. Substituting in the given equation we find that this root is negative. Hence, $\xi_1 = -\,2.0$.

One simple real root : $|\,\xi_3\,|^{16} = |\,B_3 \,/\, B_2\,|$

which gives $\xi_3 = 1.0000$. Substituting in the given equation, we find that the root is positive.

Hence, the roots are 1.0000, – 2.0000, – 2.0000.

1.60 Consider the equation $P(x) = 10x^{10} + x^5 + x - 1 = 0$.

Compute the largest positive real root with an error less than 0.02 using the Laguerre method.

Solution

Let $x_0 = 0.5$. We have

$$p'(x) = 100x^9 + 5x^4 + 1,\ p''(x) = 900x^8 + 20x^3.$$

First iteration

$$A = -\frac{P'(0.5)}{P(0.5)} = 3.28511,$$

$$B = A^2 - \frac{P''(0.5)}{P(0.5)} = 23.89833,$$

$$x_1 = 0.5 + \frac{10}{A + \sqrt{9(10B - A^2)}} = 0.5 + 0.20575 \approx 0.706.$$

Second iteration

$$A = -\frac{P'(0.706)}{P(0.706)} = -\,34.91187$$

$$B = A^2 - \frac{P''(0.706)}{P(0.706)} = 887.75842$$

$$x_2 = 0.706 + \frac{10}{A - \sqrt{9(10B - A^2)}} \approx 0.6724.$$

Third iteration

$$A = -\frac{P'(0.6724)}{P(0.6724)} = 3928.21138$$

$$B = A^2 - \frac{P''(0.6724)}{P(0.6724)} = 15466362.32$$

$$x_3 = 0.6724 + \frac{10}{A + \sqrt{9(10B - A^2)}} = 0.672654.$$

The root correct to two decimals is 0.67.

CHAPTER 2

Linear Algebraic Equations and Eigenvalue Problems

2.1 INTRODUCTION

Let the given system of n equations be written as

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n.
\end{aligned} \tag{2.1}
$$

In matrix notation, we can write (2.1) as

$$\mathbf{Ax} = \mathbf{b} \tag{2.2}$$

where
$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \cdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = (a_{ij})$$

$$\mathbf{x} = [x_1 \quad x_2 \quad \ldots \quad x_n]^T \quad \text{and} \quad \mathbf{b} = [b_1 \quad b_2 \quad \ldots \quad b_n]^T.$$

Definitions 2.1 A real matrix **A** is

nonsingular if $|\mathbf{A}| \neq 0$,
singular if $|\mathbf{A}| = 0$
symmetric if $\mathbf{A} = \mathbf{A}^T$,
skew symmetric if $\mathbf{A} = -\mathbf{A}^T$,
null if $a_{ij} = 0$, $i, j = 1(1)n$,
diagonal if $a_{ij} = 0$, $i \neq j$,
unit matrix if $a_{ij} = 0$, $i \neq j$, $a_{ii} = 1$, $i = 1(1)n$,
lower triangular if $a_{ij} = 0$, $j > i$,
upper triangular if $a_{ij} = 0$, $i > j$,
band matrix if $a_{ij} = 0$, for $j > i + p$ and
$i > j + q$, with band width $p + q + 1$,
tridiagonal if $a_{ij} = 0$, for $|i - j| > 1$,

diagonally dominant if $|a_{ii}| \geq \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|, \quad i = 1(1)n,$

orthogonal if $\mathbf{A}^{-1} = \mathbf{A}^T$.

A complex matrix $\mathbf{A}$ is

Hermitian, denoted by $\mathbf{A}^*$ or $\mathbf{A}^H$, if $\mathbf{A} = (\overline{\mathbf{A}})^T$ where $\overline{\mathbf{A}}$ is the complex conjugate of $\mathbf{A}$,

unitary if $\mathbf{A}^{-1} = (\overline{\mathbf{A}})^T$,

normal if $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$.

Definition 2.2 A matrix $\mathbf{A}$ is said to be a permutation matrix if it has exactly one 1 in each row and column and all other entries are 0.

Definition 2.3 A matrix $\mathbf{A}$ is reducible if there exists a permutation matrix $\mathbf{P}$ such that

$$\mathbf{PAP}^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} \quad \text{or} \quad \mathbf{PAP}^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tag{2.3}$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square submatrices.

Definition 2.4 A real matrix $\mathbf{M}$ is said to have '*property A*' if there exists a permutation matrix $\mathbf{P}$ such that

$$\mathbf{PMP}^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tag{2.4}$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are diagonal matrices.

Definition 2.5 A matrix $\mathbf{M}$ is positive definite, if $\mathbf{x}^*\mathbf{Mx} > 0$ for any vector $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}^* = (\overline{\mathbf{x}})^T$. Further, $\mathbf{x}^*\mathbf{Mx} = 0$ if $\mathbf{x} = \mathbf{0}$.

If $\mathbf{A}$ is a Hermitian, strictly diagonal dominant matrix with positive real diagonal entries, then $\mathbf{A}$ is positive definite.

Positive definite matrices have the following important properties :

(*i*) If $\mathbf{A}$ is nonsingular and positive definite, then $\mathbf{B} = \mathbf{A}^*\mathbf{A}$ is Hermitian and positive definite.

(*ii*) The eigenvalues of a positive definite matrix are all real and positive.

(*iii*) All the leading minors of $\mathbf{A}$ are positive.

The solution of the system of equations (2.2) exists and is unique if $|\mathbf{A}| \neq 0$. It has nonzero solution if at least one of b_i is not zero. The solution of (2.2) may then be written as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{2.5}$$

The homogeneous system ($b_i = 0$, $i = 1(1)n$) possesses only a trivial solution $x_1 = x_2 = \ldots = x_n = 0$ if $|\mathbf{A}| \neq 0$. Consider a homogeneous system in which a parameter λ occurs. The problem then is to determine the values of λ, called the *eigenvalues*, for which the system has nontrivial solution. These solutions are called the *eigenvectors* or the *eigenfunctions* and the entire system is called an *eigenvalue problem*. The eigenvalue problem may therefore be written as

$$\mathbf{Ax} = \lambda\mathbf{x} \quad \text{or} \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{2.6}$$

This system has nontrivial solutions if

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \tag{2.7}$$

which is a polynomial of degree n in λ and is called the characteristic equation. The n roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ are called the eigenvalues of $\mathbf{A}$. The largest eigenvalue in magnitude is called the *spectral radius* of **A** and is denoted by $\rho(\mathbf{A})$. Corresponding to each eigenvalue λ_i, there exists an eigenvector $\mathbf{x}_i$ which is the nontrivial solution of

$$(\mathbf{A} - \lambda_i \mathbf{I})\, \mathbf{x}_i = \mathbf{0}. \tag{2.8}$$

If the n eigenvalues λ_i, $i = 1(1)n$ are distinct, then the n independent eigenvectors $\mathbf{x}_i$, $i = 1(1)n$ constitute a complete system and can be taken as a basis of an n-dimensional space. In this space, any vector **v** can be expressed as

$$\mathbf{v} = c_1\,\mathbf{x}_1 + c_2\,\mathbf{x}_2 + \ldots + c_n\,\mathbf{x}_n. \tag{2.9}$$

Let the n eigenvalues λ_i, $i = 1(1)n$ be distinct and **S** denote the matrix of the corresponding eigenvectors

$$\mathbf{S} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_n].$$

Then,
$$\mathbf{S}^{-1}\,\mathbf{A}\mathbf{S} = \mathbf{D} \tag{2.10}$$

where **D** is diagonal matrix and the eigenvalues of **A** are located on the diagonal of **D**. Further, **S** is an orthogonal matrix. This result is true even if the eigenvalues are not distinct but the problem has the complete system of eigenvectors.

Norm of a vector x

(*i*) Absolute norm (l_1 norm)

$$\| \mathbf{x} \|_1 = \sum_{i=1}^{n} | x_i |. \tag{2.11}$$

(*ii*) Euclidean norm

$$\| \mathbf{x} \|_2 = (\mathbf{x}^*\mathbf{x})^{1/2} = \left(\sum_{i=1}^{n} | x_i |^2 \right)^{1/2}. \tag{2.12}$$

(*iii*) Maximum norm (l_∞ norm)

$$\| \mathbf{x} \|_\infty = \max_{1 \le i \le n} | x_i |. \tag{2.13}$$

Norm of a matrix A

(*i*) Frobenius or Euclidean norm

$$F(\mathbf{A}) = \left(\sum_{i, j=1}^{n} | a_{ij} |^2 \right)^{1/2}. \tag{2.14}$$

(*ii*) Maximum norm

$$\| \mathbf{A} \|_\infty = \max_i \sum_k | a_{ik} | \tag{2.15}$$

(maximum absolute row sum).

$$\| \mathbf{A} \|_1 = \max_k \sum_i | a_{ik} | \tag{2.16}$$

(maximum absolute column sum).

(*iii*) *Hilbert norm* or *spectral norm*

$$\| \mathbf{A} \|_2 = \sqrt{\lambda} \tag{2.17}$$

where $\lambda = \rho(\mathbf{A}^*\mathbf{A})$. If **A** is Hermitian or real and symmetric, then

$$\lambda = \rho(\mathbf{A}^2) = [\,\rho(\mathbf{A})]^2 \quad \text{and} \quad \| \mathbf{A} \|_2 = \rho(\mathbf{A}). \tag{2.18}$$

Theorem 2.1 No eigenvalue of a matrix **A** exceeds the norm of a matrix

$$\| \mathbf{A} \| \geq \rho(\mathbf{A}). \tag{2.19}$$

Theorem 2.2 Let **A** be a square matrix. Then

$$\lim_{n \to \infty} \mathbf{A}^n = \mathbf{0}$$

if $\| \mathbf{A} \| < 1$, or if and only if $\rho(\mathbf{A}) < 1$.

Theorem 2.3 The infinite series

$$\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots \tag{2.20}$$

converges if $\lim_{m \to \infty} \mathbf{A}^m = \mathbf{0}$. The series converges to $(\mathbf{I} - \mathbf{A})^{-1}$.

Consider now the system of equations (2.2) $\mathbf{Ax} = \mathbf{b}$.

(*i*) If $\mathbf{A} = \mathbf{D}$, *i.e.*, $\mathbf{Dx} = \mathbf{b}$, then the solution of the system is given by

$$x_i = \frac{b_i}{a_{ii}}, \quad i = 1(1)n \tag{2.21}$$

where $a_{ii} \neq 0$.

(*ii*) If **A** is a lower triangular matrix, *i.e.*, $\mathbf{Lx} = \mathbf{b}$, then, the solution is obtained as

$$x_k = \left(b_k - \sum_{j=1}^{k-1} a_{kj} x_j \right) \Big/ a_{kk}, \quad k = 1, 2, \ldots, n \tag{2.22}$$

where $a_{kk} \neq 0$, $k = 1(1)n$. This method is known as the *forward substitution method.*

(*iii*) If **A** is an upper triangular matrix, *i.e.*, $\mathbf{Ux} = \mathbf{b}$, then, the solution is given by

$$x_k = \left(b_k - \sum_{j=k+1}^{n} a_{kj} x_j \right) \Big/ a_{kk}, \quad k = n, n-1, \ldots, 1 \tag{2.23}$$

where $a_{kk} \neq 0$, $k = 1(1)n$. This method is known as the *backward substitution method.*

2.2 DIRECT METHODS

Gauss Elimination Method

Consider the *augmented matrix* $[\mathbf{A} \mid \mathbf{b}]$ of the system of equations $\mathbf{Ax} = \mathbf{b}$. Using *elementary row transformations*, Gauss elimination method reduces the matrix **A** in the augmented matrix to an upper triangular form

$$[\mathbf{A} \mid \mathbf{b}] \xrightarrow[\text{elimination}]{\text{Gauss}} [\mathbf{U} \mid \mathbf{c}]. \tag{2.24}$$

Back substitution, using (2.23), then gives the solution vector **x**. For large n, the operational count is $\approx n^3 / 3$. The successive elements after each elimination procedure are obtained as follow :

$$\text{Set} \qquad b_i^{(k)} = a_{i, n+1}^{(k)}, \quad i, k = 1(1)n \tag{2.25}$$

$$\text{with} \qquad b_i^{(1)} = b_i, \quad i = 1(1)n.$$

The elements $a_{ij}^{(k)}$ with $i, j \geq k$ are given by

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}$$

$$i = k + 1, k + 2, ..., n \,; j = k + 1, ..., n, n + 1$$

$$a_{ij}^{(1)} = a_{ij}. \tag{2.26}$$

The elements $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{nn}^{(n)}$ are called the *pivots.*

To avoid division by zero and to reduce roundoff error, *partial pivoting* is normally used.

The pivot is chosen as follows :

Choose j, the smallest integer for which

$$\left| a_{jk}^{(k)} \right| = \max_i \left| a_{ik}^{(k)} \right|, \quad k \le i \le n \tag{2.27}$$

and interchange rows k and j. It is called *partial pivoting*.

If at the kth step, we interchange both the rows and columns of the matrix so that the largest number in magnitude in the remaining matrix is used as pivot, *i.e.*, after pivoting

$$| a_{kk} | = \max | a_{ij} |, i, j = k, k + 1, ..., n,$$

then, it is called *complete pivoting*.

Note that, when we interchange two columns, the position of the corresponding elements in the solution vector is also changed.

Complete pivoting is safe as errors are never magnified unreasonably. The magnification factor is less than or equal to

$$f_n = [(n - 1) \times 2 \times 3^{1/2} \times 4^{1/3} \times ... \times n^{1/(n-1)}]^{1/2}$$

for $n \times n$ system of equations. For example, we have the magnification factors

n	5	10	20	100
f_n	5.74	18.30	69.77	3552.41

which reveals that the growth is within limits. Eventhough, the bound for the magnification factor in the case of partial pivoting cannot be given by an expression, it is known (experimentally) that the magnification error is almost eight times, in most cases, the magnification factor for complete pivoting. Complete pivoting approximately doubles the cost, while the partial pivoting costs negligibly more than the Gauss elimination.

Gauss elimination with or without partial pivoting are same for diagonally dominant matrices.

Gauss-Jordan Method

Starting with the augmented matrix, the coefficient matrix **A** is reduced to a diagonal matrix rather than an upper triangular matrix. This means that elimination is done not only in the equations below but also in the equations above, producing the solution without using the back substitution method.

$$[\mathbf{A} \,|\, \mathbf{b}] \xrightarrow[\text{Jordan}]{\text{Gauss}} [\mathbf{I} \,|\, \mathbf{d}]. \tag{2.28}$$

This method is more expensive from the computation view point compared to the Gauss elimination method. For large n, the operational count is $\approx n^3/2$. However, this method is useful in finding the inverse of a non singular square matrix.

$$[\mathbf{A} \,|\, \mathbf{I}] \xrightarrow[\text{Jordan}]{\text{Gauss}} [\mathbf{I} \,|\, \mathbf{A}^{-1}]. \tag{2.29}$$

Triangularization Method

In this method, the coefficient matrix **A** in (2.2) is decomposed into the product of a lower triangular matrix **L** and an upper triangular matrix **U**. We write

$$\mathbf{A} = \mathbf{LU} \tag{2.30}$$

where $l_{ij} = 0, j > i$; $u_{ij} = 0, i > j$ and $u_{ii} = 1$.

This method is also called the *Crout's method.* Instead of $u_{ii} = 1$, if we take $l_{ii} = 1$, then the method is also called the *Doolittle's method*.

Comparing the elements of the matrices on both sides, we obtain n^2 equations in n^2 unknowns, which uniquely determines **L** and **U**. We get

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}, \quad i \geq j,$$

$$u_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \right) \Big/ l_{ii}, \quad i < j,$$

$$u_{ii} = 1.$$

The system of equations (2.2) becomes

$$\mathbf{LUx} = \mathbf{b}. \tag{2.31}$$

We rewrite this system as

$$\mathbf{Ux} = \mathbf{z}, \tag{2.32i}$$

$$\mathbf{Lz} = \mathbf{b}. \tag{2.32ii}$$

We first find **z** from (2.32*ii*) using forward substitution and then find **x** from (2.32*i*) using the back substitution.

Alternately, from (2.32*ii*) we have

$$\mathbf{z} = \mathbf{L}^{-1}\mathbf{b} \tag{2.33i}$$

and from (2.32*i*) we have

$$\mathbf{x} = \mathbf{U}^{-1} \mathbf{z}. \tag{2.33ii}$$

The inverse of **A** can be obtained from

$$\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}. \tag{2.34}$$

Triangularization is used more often than the Gauss elimination. The operational count is same as in the Gauss elimination.

LU decomposition is not always guaranteed for arbitrary matrices. Decomposition is guaranteed when the matrix **A** is positive definite.

Cholesky Method (Square Root Method)

If the coefficient matrix in (2.2) is symmetric and positive definite, then **A** can be decomposed as

$$\mathbf{A} = \mathbf{LL}^T \tag{2.35}$$

where $l_{ij} = 0, j > i$.

The elements of **L** are given by

$$l_{ii} = \left(a_{ii} - \sum_{j=1}^{i-1} l_{ij}^2 \right)^{1/2}, \ i = 1(1)n$$

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{jk}\, l_{ik} \right) \bigg/ l_{jj}\,,$$
$$i = j+1, j+2, \ldots, n\,;\, j = 1(1)n$$
$$l_{ij} = 0, \quad i < j. \tag{2.36}$$

Corresponding to equations (2.33*i,ii*) we have

$$\mathbf{z} = \mathbf{L}^{-1}\mathbf{b}, \tag{2.37i}$$
$$\mathbf{x} = (\mathbf{L}^T)^{-1}\,\mathbf{z} = (\mathbf{L}^{-1})^T\mathbf{z}. \tag{2.37ii}$$

The inverse is obtained as

$$\mathbf{A}^{-1} = (\mathbf{L}^T)^{-1}\,\mathbf{L}^{-1} = (\mathbf{L}^{-1})^T\,\mathbf{L}^{-1}. \tag{2.38}$$

The operational count for large n, in this case, is $\approx n^3 / 6$.

Instead of $\mathbf{A} = \mathbf{LL}^T$, we can also decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{UU}^T$.

Partition Method

This method is usually used to find the inverse of a large nonsingular square matrix by partitioning. Let $\mathbf{A}$ be partitioned as

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{E} & \mathbf{D} \end{array}\right] \tag{2.39}$$

where **B, C, E, D,** are of orders $r \times r$, $r \times s$, $s \times r$ and $s \times s$ respectively, with $r + s = n$. Similarly, we partition $\mathbf{A}^{-1}$ as

$$\mathbf{A}^{-1} = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{Z} & \mathbf{V} \end{array}\right] \tag{2.40}$$

where **X, Y, Z** and **V** are of the same orders as **B, C, E** and **D** respectively. Using the identity

$$\mathbf{AA}^{-1} = \left[\begin{array}{c|c} \mathbf{I}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_2 \end{array}\right]$$

we obtain
$$\mathbf{V} = (\mathbf{D} - \mathbf{EB}^{-1}\,\mathbf{C})^{-1}, \quad \mathbf{Y} = -\,\mathbf{B}^{-1}\,\mathbf{CV},$$
$$\mathbf{Z} = -\,\mathbf{VEB}^{-1}, \quad \mathbf{X} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\,\mathbf{CZ},$$

where we have assumed that $\mathbf{B}^{-1}$ exists. If $\mathbf{B}^{-1}$ does not exist but $\mathbf{D}^{-1}$ exists then the equations can be modified suitably. This procedure requires finding the inverse of two lower order matrices, $\mathbf{B}^{-1}$ and $(\mathbf{D} - \mathbf{EB}^{-1}\,\mathbf{C})^{-1}$.

Condition Numbers

Sometimes, one comes across a system of equations which are very sensitive to round off errors. That is, one gets different solutions when the elements are rounded to different number of digits. In such cases, the system is called an *ill-conditioned* system of equations. The measure of the ill-conditionedness is given by the value of the *condition number* of the matrix **A**. The condition number is defined as

$$\text{cond}\,(\mathbf{A}) = K(\mathbf{A}) = \|\,\mathbf{A}\,\|\;\|\,\mathbf{A}^{-1}\,\| \tag{2.41}$$

where $\|\,.\,\|$ is any suitable norm. This number is usually referred to as standard condition number.

If $K(\mathbf{A})$ is large, then small changes in **A** or **b** produces large relative changes in **x**, and the system of equations $\mathbf{Ax} = \mathbf{b}$ is ill-conditioned. If $K(\mathbf{A}) \approx 1$, then the system (2.2) is well conditioned. If $\|\,.\,\|$ is the spectral norm, then

$$K(\mathbf{A}) = \| \mathbf{A} \|_2 \, \| \mathbf{A}^{-1} \|_2 = \sqrt{\frac{\lambda}{\mu}} \tag{2.42}$$

where λ and μ are the largest and smallest eigenvalues in modulus of $\mathbf{A^*A}$. If $\mathbf{A}$ is Hermitian or real and symmetric, we have

$$K(\mathbf{A}) = \frac{\lambda^*}{\mu^*} \tag{2.43}$$

where λ^*, μ^* are the largest and smallest eigenvalues in modulus of $\mathbf{A}$.

Another important condition number is the *Aird-Lynch* estimate. This estimate gives both the lower and upper bounds for the error magnification. We have the estimate as

$$\frac{\| \mathbf{cr} \|}{\| \mathbf{x} \| (1+T)} \le \frac{\| \delta \mathbf{x} \|}{\| \mathbf{x} \|} \le \frac{\| \mathbf{cr} \|}{\| \mathbf{x} \| (1-T)}$$

where $\mathbf{c}$ is the appropriate inverse of $\mathbf{A}$ (usually the outcome of Gauss elimination) ; $\mathbf{r} = \mathbf{Ax} - \mathbf{b}$, $\mathbf{x}$ is the computed solution and $T = \| \mathbf{cA} - \mathbf{I} \| < 1$.

2.3 ITERATION METHODS

A general linear iterative method for the solution of the system of equations (2.2) may be defined in the form

$$\mathbf{x}^{(k+1)} = \mathbf{Hx}^{(k)} + \mathbf{c} \tag{2.44}$$

where $\mathbf{x}^{(k+1)}$ and $\mathbf{x}^{(k)}$ are the approximations for $\mathbf{x}$ at the $(k + 1)$th and kth iterations, respectively. $\mathbf{H}$ is called the *iteration matrix* depending on $\mathbf{A}$ and $\mathbf{c}$ is a column vector. In the limiting case, when $k \to \infty$, $\mathbf{x}^{(k)}$ converges to the exact solution

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b}. \tag{2.45}$$

Theorem 2.4 The iteration method of the form (2.44) for the solution of (2.2) converges to the exact solution for any initial vector, if $\| \mathbf{H} \| < 1$ or iff $\rho(\mathbf{H}) < 1$.

Let the coefficient matrix $\mathbf{A}$ be written as

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U} \tag{2.46}$$

where $\mathbf{L}, \mathbf{D}, \mathbf{U}$ are the strictly lower triangular, diagonal and strictly upper triangular parts of $\mathbf{A}$ respectively. Write (2.2) as

$$(\mathbf{L} + \mathbf{D} + \mathbf{U})\mathbf{x} = \mathbf{b}. \tag{2.47}$$

Jacobi Iteration Method

We rewrite (2.47) as

$$\mathbf{Dx} = -(\mathbf{L} + \mathbf{U})\,\mathbf{x} + \mathbf{b}$$

and define an iterative procedure as

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}. \tag{2.48}$$

The iteration matrix is given by

$$\mathbf{H} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}). \tag{2.49}$$

The method (2.48) is called the *Jacobi Iteration method.*

We write (2.48) as

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - [\mathbf{I} + \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})]\,\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \\ &= \mathbf{x}^{(k)} - \mathbf{D}^{-1}\,[\mathbf{D} + \mathbf{L} + \mathbf{U}]\,\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \end{aligned}$$

or

$$\mathbf{x}^{(x+1)} - \mathbf{x}^{(k)} = \mathbf{D}^{-1}\,(\mathbf{b} - \mathbf{Ax}^{(k)})$$

or $$\mathbf{v}^{(k)} = \mathbf{D}^{-1}\,\mathbf{r}^{(k)}, \quad \text{or} \quad \mathbf{D}\mathbf{v}^{(k)} = \mathbf{r}^{(k)} \tag{2.50}$$

where $\mathbf{v}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ is the error vector and $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$ is the *residual vector*. From the computational view point, (2.50) may be preferred as we are dealing with the errors and not the solutions.

Gauss-Seidel Iteration Method

In this case, we define the iterative procedure as

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

or $$\mathbf{x}^{(k+1)} = -\,(\mathbf{D} + \mathbf{L})^{-1}\,\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b} \tag{2.51}$$

where $\mathbf{H} = -\,(\mathbf{D} + \mathbf{L})^{-1}\,\mathbf{U}$ is the iteration matrix. In terms of the error vector, we can write the procedure as

$$\mathbf{v}^{(k+1)} = (\mathbf{D} + \mathbf{L})^{-1}\,\mathbf{r}^{(k)}, \quad \text{or} \quad (\mathbf{D} + \mathbf{L})\,\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)}. \tag{2.52}$$

Successive Over Relaxation (SOR) Method

This method is often used when the coefficient matrix **A** of the system of equations is symmetric and has 'property *A*'. The iterative procedure is given by

$$\mathbf{x}^{(k+1)} = (\mathbf{D} + w\mathbf{L})^{-1}\,[(1 - w)\mathbf{D} - w\mathbf{U}]\,\mathbf{x}^{(k)} + w(\mathbf{D} + w\mathbf{L})^{-1}\,\mathbf{b} \tag{2.53}$$

where w is the *relaxation parameter*. In terms of the error vector **v**, we can rewrite (2.53) as

$$\mathbf{v}^{(k+1)} = w(\mathbf{D} + w\mathbf{L})^{-1}\,\mathbf{r}^{(k)}, \quad \text{or} \quad (\mathbf{D} + w\mathbf{L})\,\mathbf{v}^{(k+1)} = w\mathbf{r}^{(k)}. \tag{2.54}$$

When $w = 1$, eq. (2.53) reduces to the Gauss-Seidel method (2.51). The relaxation parameter w satisfies the condition $0 < w < 2$. If $w > 1$ then the method is called an *over relaxation method* and if $w < 1$, it is called an *under relaxation method.* Maximum convergence of SOR is obtained when

$$w = w_{\text{opt}} \approx \frac{2}{\mu^2}\,[1 - \sqrt{1 - \mu^2}\,] = \frac{2}{1 + \sqrt{1 - \mu^2}} \tag{2.55}$$

where $\mu = \rho(\mathbf{H}_{\text{Jacobi}})$ and w_{opt} is rounded to the next digit.

The rate of convergence of an iterative method is defined as

$$v = -\ln\,(\rho(\mathbf{H})), \quad \text{or also as} \quad v = -\log_{10}(\,\rho(\mathbf{H})). \tag{2.56}$$

where **H** is the iteration matrix.

The spectral radius of the SOR method is $W_{\text{opt}} - 1$ and its rate of convergence is

$$v = -\ln\,(W_{\text{opt}} - 1) \text{ or } V = -\log_{10}\,(W_{\text{opt}} - 1).$$

Extrapolation Method

There are some powerful acceleration procedures for iteration methods. One of them is the *extrapolation* method. We write the given iteration formula as

$$\mathbf{x}^{(k+1)} = \mathbf{H}\mathbf{x}^{(k)} + \mathbf{c}$$

and consider one parameter family of extrapolation methods as

$$\mathbf{x}^{(k+1)} = \gamma\,[\mathbf{H}\mathbf{x}^{(k)} + \mathbf{c}] + (1 - \gamma)\,\mathbf{x}^{(k)} = \mathbf{H}_\gamma\mathbf{x}^{(k)} + \gamma\mathbf{c}$$

where $\mathbf{H}_\gamma = \gamma\mathbf{H} + (1 - \gamma)\,\mathbf{I}$. Suppose that, we know that all the eigenvalues of **H** lie in an interval $[a, b]$, $1 \notin [a, b]$, on the real line. Then

$$\rho(\mathbf{H}_\gamma) \le 1 - |\,\gamma\,|d$$

where d is the distance from 1 to $[a, b]$. The optimal value of γ which gives maximum rate of convergence is $\gamma = 2\,/\,(2 - a - b)$.

2.4 EIGENVALUE PROBLEMS

Consider the eigenvalue problem

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{2.57}$$

Theorem 2.5 (*Gerschgorin*) The largest eigenvalue in modulus of a square matrix **A** cannot exceed the largest sum of the moduli of the elements in any row or column.

Theorem 2.6 (*Brauer*) Let P_k be the sum of the moduli of the elements along the kth row excluding the diagonal element a_{kk}. Then, every eigenvalue of **A** lies inside or on the boundary of atleast one of the circles

$$| \lambda - a_{kk} | = P_k, \quad k = 1(1)n.$$

We have, therefore

(*i*) $| \lambda_i | \le \max\limits_i \sum\limits_{k=1}^{n} | a_{ik} |$ (maximum absolute row sum).

(*ii*) $| \lambda_i | \le \max\limits_k \sum\limits_{j=1}^{n} | a_{jk} |$ (maximum absolute column sum).

(*iii*) All the eigenvalues lie in the union of the circles

$$| \lambda_i - a_{kk} | \le \sum_{\substack{j=1 \\ j \ne k}}^{n} | a_{kj} |$$

(*iv*) All the eigenvalues lie in the union of the circles

$$| \lambda_i - a_{kk} | \le \sum_{\substack{j=1 \\ j \ne k}}^{n} | a_{jk} | \tag{2.58i}$$

These four bounds are independent. Hence, the required bound is the intersection of these four bounds.

If **A** is symmetric, then the circles become intervals on the real line.

These bounds are referred to as *Gerschgorin bounds* or *Gerschgorin circles.*

Theorem 2.7 If the matrix **A** is diagonalized by the similarity transformation $\mathbf{S}^{-1}\mathbf{AS}$, and if **B** is any matrix, then the eigenvalues μ_i of **A + B** lie in the union of the disks

$$| \mu - \lambda_i | \le \text{cond}_\infty (\mathbf{S}) \, \| \mathbf{B} \|_\infty \tag{2.58ii}$$

where $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of **A** and $\text{cond}_\infty(\mathbf{S})$ is the condition number of **S**.

Usually, **B** is a permutation matrix.

Let $\mathbf{S}^{-1}\mathbf{AS} = \mathbf{D}$. Then we have,

$$\begin{aligned} \text{spectrum } (\mathbf{A} + \mathbf{B}) &= \text{spectrum } [\mathbf{S}^{-1}(\mathbf{A} + \mathbf{B})\mathbf{S}] \\ &= \text{spectrum } [\mathbf{D} + \mathbf{S}^{-1}\mathbf{BS}] \\ &= \text{spectrum } [\mathbf{D} + \mathbf{Q}] \end{aligned}$$

where $\mathbf{Q} = (q_{ij}) = \mathbf{S}^{-1}\mathbf{BS}$, and **D** is a diagonal matix.

By applying Gerschgorin theorem, the eigenvalues of **A + B** lie in the union of disks

$$| \mu - \lambda_i - q_{ii} | \le \sum_{\substack{j=1 \\ i \ne j}}^{n} | q_{ij} |$$

Further, if **A** is Hermitian, then condition (2.58*ii*) simplifies to

$$| \mu - \lambda_i | \le n \, \| \mathbf{B} \|_\infty$$

Let us now consider methods for finding all the eigenvalues and eigenvectors of the given matrix **A**.

Jacobi Method for Symmetric Matrices

Let **A** be a real symmetric matrix. **A** is reduced to a diagonal matrix by a series of orthogonal transformations $\mathbf{S}_1$, $\mathbf{S}_2$, ... in 2 × 2 subspaces. When the diagonalization is completed, the eigenvalues are located on the diagonal and the orthogonal matrix of eigenvectors is obtained as the product of all the orthogonal transformations.

Among the off-diagonal elements, let $| a_{ik} |$ be the numerically largest element. The orthogonal transformation in the 2 × 2 subspace spanned by a_{ii}, a_{ik}, a_{ki}, a_{kk} is done using the matrix

$$\mathbf{S}_1{}^* = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

The value of θ is obtained such that $(\mathbf{S}_1{}^*)^{-1} \mathbf{A} \, \mathbf{S}_1{}^* = (\mathbf{S}_1{}^*)^T \mathbf{A}\mathbf{S}_1{}^*$ is diagonalized. We find

$$\tan 2\theta = \frac{2a_{ik}}{a_{ii} - a_{kk}}, \quad -\frac{\pi}{4} \le \theta \le \frac{\pi}{4}. \tag{2.59}$$

If $a_{ii} = a_{kk}$, then

$$\theta = \left| \begin{array}{ll} \pi/4, & a_{ik} > 0, \\ -\pi/4, & a_{ik} < 0. \end{array} \right. \tag{2.60}$$

The minimum number of rotations required to bring **A** into a diagonal form is $n(n - 1) / 2$. A disadvantage of the Jacobi method is that the elements annihilated by a plane rotation may not necessarily remain zero during subsequent transformations.

Givens Method for Symmetric Matrices

Let **A** be a real symmetric matrix. Givens proposed an algorithm using plane rotations, which preserves the zeros in the off-diagonal elements, once they are created. Eigenvalues and eigenvectors are obtained using the following procedure :

(*a*) reduce **A** to a tridiagonal form **B,** using plane rotations,

(*b*) form a *Sturm* sequence for the characteristic equation of **B**, study the changes in signs in the sequences and find the intervals which contain the eigenvalues of **B**, which are also the eigenvalues of **A.**

(*c*) using any iterative method, find the eigenvalues to the desired accuracy.

(*d*) find the eigenvectors of **B** and then the eigenvectors of **A**.

The reduction to the tridiagonal form is achieved by using orthogonal transformations as in Jacobi method using the (2, 3), (2, 4), ..., (2, *n*), (3, 4), ..., (4, 5), ... subspaces. When reduction with respect to the (2, 3) subspace is being done, θ is obtained by setting $a_{13}' = a_{31}' = 0$, which gives

$$\tan\theta = \frac{a_{13}}{a_{12}}, \quad -\frac{\pi}{4} \le \theta \le \frac{\pi}{4} \tag{2.61}$$

where **A′** is the transformed matrix. This value of θ, produces zeros in the (3, 1) and (1, 3) locations. The value of θ, obtained by setting $a_{14}{}^* = a_{41}{}^* = 0$, when working in the (2, 4) subspace, that is $\tan\theta = a_{14}' \,/\, a'_{12}$, produces zeros in the (4, 1) and (1, 4) locations. The total number of plane rotations required to bring a matrix of order n to its tridiagonal form is $(n - 1)(n - 2) / 2$. We finally obtain

$$\mathbf{B} = \begin{bmatrix} b_1 & c_1 & & & & \mathbf{0} \\ c_1 & b_2 & c_2 & & & \\ & c_2 & b_3 & c_3 & & \\ & \ddots & \ddots & \ddots & & \\ & & & c_{n-2} & b_{n-1} & c_{n-1} \\ \mathbf{0} & & & & c_{n-1} & b_n \end{bmatrix} \tag{2.62}$$

A and **B** have the same eigenvalues. If $c_i \neq 0$, then the eigenvalues are distinct. The characteristic equation of **B** is

$$f_n = |\ \lambda\mathbf{I} - \mathbf{B}\ | = 0. \tag{2.63}$$

Expanding by minors, we obtain the sequence $\{ f_i \}$

$$f_0 = 1, f_1 = \lambda - b_1$$

and

$$f_r = (\lambda - b_r) f_{r-1} - (c_{r-1})^2 f_{r-2} \ ; \ 2 \le r \le n. \tag{2.64}$$

If none of the c_i, $i = 1, 2, ..., n - 1$ vanish, then $\{ f_i \}$ is a *Sturm sequence*. If any of $c_i = 0$, then the system degenerates. For example, if any of the $c_i = 0$, then **B** given by (2.62) is of the form

$$\mathbf{B} = \begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix} \tag{2.65}$$

and the characteristic equation of B is

$$f_n = (\text{ch. equation of } \mathbf{P}) \ (\text{ch. equation of } \mathbf{Q}). \tag{2.66}$$

Let $V(x)$ denote the number of changes in signs in the sequence $\{ f_i \}$ for a given number x. Then, the number of zeros of f_n in (a, b) is $|\ V(a) - V(b)\ |$ (provided a or b is not a zero of f_n). Repeated application, using the bisection method, produces the eigenvalues to any desired accuracy.

Let $\mathbf{v}_i$ be the eigenvector of **B** corresponding to λ_i. Then, the eigenvector $\mathbf{u}_i$ of **A** is given by

$$\mathbf{u}_i = \mathbf{S}\mathbf{v}_i \tag{2.67}$$

where $\mathbf{S} = \mathbf{S}_1\mathbf{S}_2 \ldots \mathbf{S}_j$ is the product of the orthogonal matrices used in the plane rotations.

Householder's Method for Symmetric Matrices

In Householder's method, **A** is reduced to the tridiagonal form by orthogonal transformations representing reflections. This reduction is done in exactly $n - 2$ transformations. The orthogonal transformations are of the form

$$\mathbf{P} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^T \tag{2.68}$$

where $\mathbf{w} \in R^n$, such that $\mathbf{w} = [x_1\ x_2 \ldots x_n]^T$ and

$$\mathbf{w}^T \mathbf{w} = x_1^2 + x_2^2 + \ldots + x_n^2 = 1. \tag{2.69}$$

P is symmetric and orthogonal. The vectors **w** are constructed with the first $(r - 1)$ components as zeros, that is

$$\mathbf{w}_r^T = (0, 0, \ldots, 0, x_r, x_{r+1}, \ldots, x_n) \tag{2.70}$$

with $x_r^2 + x_{r+1}^2 + \ldots + x_n^2 = 1$. With this choice of $\mathbf{w}_r$, form the matrices

$$\mathbf{P}_r = \mathbf{I} - 2\mathbf{w}_r\ \mathbf{w}_r^T. \tag{2.71}$$

The similarity transformation is given by

$$\mathbf{P}_r^{-1}\ \mathbf{A}\mathbf{P}_r = \mathbf{P}_r^T\ \mathbf{A}\mathbf{P}_r = \mathbf{P}_r\mathbf{A}\mathbf{P}_r. \tag{2.72}$$

Put $\mathbf{A} = \mathbf{A}_1$ and form successively

$$\mathbf{A}_r = \mathbf{P}_r\mathbf{A}_{r-1}\ \mathbf{P}_r, \quad r = 2, 3, \ldots, n - 1. \tag{2.73}$$

At the first transformation, we find x_r's such that we get zeros in the positions (1, 3), (1, 4), ..., (1, n) and in the corresponding positions in the first column. In the second transformation, we find x_r's such that we get zeros in the positions (2, 4), (2, 5), ..., (2, n) and in the corresponding positions in the second column. In $(n - 2)$ transformations, **A** is reduced to the tridiagonal form. The remaining procedure is same as in Givens method.

For example, consider

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{2.74}$$

For the first transformation, choose

$$\mathbf{w}_2^T = [0 \; x_2 \; x_3 \; x_4]$$

$$x_2^2 + x_3^2 + x_4^2 = 1. \tag{2.75}$$

We find
$$s_1 = \sqrt{a_{12}^2 + a_{13}^2 + a_{14}^2}$$

$$x_2^2 = \frac{1}{2}\left(1 + \frac{a_{12}\ \text{sign}\,(a_{12})}{s_1}\right)$$

$$x_3 = \frac{a_{13}\ \text{sign}\,(a_{12})}{2 s_1 x_2}, \quad x_4 = \frac{a_{14}\ \text{sign}\,(a_{12})}{2 s_1 x_2}. \tag{2.76}$$

This transformation produces two zeros in the first row and first column. One more transformation produces zeros in the (2, 4) and (4, 2) positions.

Rutishauser Method for Arbitrary Matrices

Set $\mathbf{A} = \mathbf{A}_1$ and decompose $\mathbf{A}_1$ as

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1 \tag{2.77}$$

with $l_{ii} = 1$. Then, form $\mathbf{A}_2 = \mathbf{U}_1\mathbf{L}_1$. Since $\mathbf{A}_2 = \mathbf{U}_1\mathbf{L}_1 = \mathbf{U}_1\mathbf{A}_1\ \mathbf{U}_1^{-1}$, $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same eigenvalues. We again write

$$\mathbf{A}_2 = \mathbf{L}_2\mathbf{U}_2 \tag{2.78}$$

with $l_{ii} = 1$. Form $\mathbf{A}_3 = \mathbf{U}_2\mathbf{L}_2$ so that $\mathbf{A}_2$ and $\mathbf{A}_3$ have the same eigenvalues. Proceeding this way, we get a sequence of matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, ... which in general reduces to an upper triangular matrix. If the eigenvalues are real, then they all lie on the diagonal. The procedure is very slow, if **A** has multiple eigenvalues, and it will not converge if **A** has complex eigenvalues.

Power Method

This method is normally used to determine the largest eigenvalue in magnitude and the corresponding eigenvector of **A**. Fastest convergence is obtained when λ_i's are distinct and far separated. Let **v** be any vector (non orthogonal to **x**) in the space spanned by the eigenvectors. Then, we have the algorithm

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{A}\mathbf{v}_k \\ \mathbf{v}_{k+1} &= \mathbf{y}_{k+1} / m_{k+1} \end{aligned} \tag{2.79}$$

where
$$m_{k+1} = \max_r \; | \, (\mathbf{y}_{k+1})_r \, |.$$

Then,
$$\lambda_1 = \lim_{k \to \infty} \frac{(\mathbf{y}_{k+1})_r}{(\mathbf{v}_k)_r}, \quad r = 1, 2, ..., n \tag{2.80}$$

and $\mathbf{v}_{k+1}$ is the required eigenvector.

Inverse Power Method

Inverse power method can give approximation to any eigenvalue. However, it is used usually to find the smallest eigenvalue in magnitude and the corresponding eigenvector of a given matrix **A**. The eigenvectors are computed very accurately by this method. Further, the method is powerful to calculate accurately the eigenvectors, when the eigenvalues are not well separated. In this case, power method converges very slowly.

If λ is an eigenvalue of **A**, then $1 / \lambda$ is an eigenvalue of $\mathbf{A}^{-1}$ corresponding to the same eigenvector. The smallest eigenvalue in magnitude of **A** is the largest eigenvalue in magnitude of $\mathbf{A}^{-1}$. Choose an arbitrary vector $\mathbf{y}_0$ (non-orthogonal to **x**). Applying the power method on $\mathbf{A}^{-1}$, we have

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{A}^{-1}\mathbf{y}_k \\ \mathbf{y}_{k+1} &= \mathbf{z}_{k+1} / m_{k+1} \end{aligned} \tag{2.81}$$

where m_{k+1} has the same meaning as in power method. We rewrite (2.81) as

$$\mathbf{A}\mathbf{z}_{k+1} = \mathbf{y}_k \tag{2.81a}$$

$$\mathbf{y}_{k+1} = \mathbf{z}_{k+1} / m_{k+1} \tag{2.81b}$$

We find $\mathbf{z}_{k+1}$ by solving the linear system (2.81*a*). The coefficient matrix is same for all iterations.

Shift of Origin

Power and inverse power methods can be used with a shift of origin. We have the following methods :

Shifted Power Method

$$\mathbf{z}^{(k+1)} = (\mathbf{A} - q\mathbf{I})\, \mathbf{z}^{(k)}$$

It can be used to find an eigenvalue farthest from a given number q.

Shifted inverse power method

$$\mathbf{z}^{(k+1)} = (\mathbf{A} - q\mathbf{I})^{-1}\, \mathbf{z}^{(k)} \quad \text{or} \quad (\mathbf{A} - q\mathbf{I})\, \mathbf{z}^{(k+1)} = \mathbf{z}^{(k)}$$

It can be used to find an eigenvalue closest to a given number q.

In both cases, normalization is done according to (2.81*b*).

2.5 SPECIAL SYSTEM OF EQUATIONS

Solution of tridiagonal system of equations

Consider the system of equations

$$\mathbf{Ax} = \mathbf{b}$$

where
$$\mathbf{A} = \begin{bmatrix} q_1 & -r_1 & & & \mathbf{0} \\ -p_2 & q_2 & -r_2 & & \\ & -p_3 & q_3 & -r_3 & \\ & & \ddots & \ddots & \ddots \\ \mathbf{0} & & & -p_n & q_n \end{bmatrix}$$

A special case of tridiagonal system of equations arise in the numerical solution of the differential equations. The tridiagonal system is of the form

$$-p_j x_{j-1} + q_j x_j - r_j x_{j+1} = b_j, \quad 1 \le j \le n \tag{2.82}$$

where p_1, r_n are given and x_0, x_{n+1} are known from the boundary conditions of the given problem. Assume that

$$p_j > 0,\ q_j > 0,\ r_j > 0 \quad \text{and} \quad q_j \geq p_j + r_j \tag{2.83}$$

for $1 \leq j \leq n$ (that is **A** is diagonally dominant). However, this requirement is a sufficient condition. For the solution of (2.82) consider the difference relation

$$x_j = \alpha_j x_{j+1} + \beta_j,\ 0 \leq j \leq n. \tag{2.84}$$

From (2.84) we have

$$x_{j-1} = a_{j-1} x_j + \beta_{j-1}. \tag{2.85}$$

Eliminating x_{j-1} from (2.82) and (2.85), we get

$$x_j = \frac{r_j}{q_j - p_j \alpha_{j-1}} x_{j+1} + \frac{b_j + p_j \beta_{j-1}}{q_j - p_j \alpha_{j-1}}. \tag{2.86}$$

Comparing (2.84) and (2.86), we have

$$\alpha_j = \frac{r_j}{q_j - p_j \alpha_{j-1}}, \quad \beta_j = \frac{b_j + p_j \beta_{j-1}}{q_j - p_j \alpha_{j-1}} \tag{2.87}$$

If $x_0 = A$, then $\alpha_0 = 0$ and $\beta_0 = A$, so that the relation

$$x_0 = \alpha_0 x_1 + \beta_0 \tag{2.88}$$

holds for all x_1. The remaining α_j, β_j, $1 \leq j \leq n$, can be calculated from (2.87).

$$\alpha_1 = \frac{r_1}{q_1}, \qquad \beta_1 = \frac{b_1 + p_1 A}{q_1}$$

$$\alpha_2 = \frac{r_2}{q_2 - p_2 \alpha_1}, \qquad \beta_2 = \frac{b_2 + p_2 \beta_1}{q_2 - p_2 \alpha_1}$$

$$\ldots \qquad \ldots$$

$$\alpha_n = \frac{r_n}{q_n - p_n \alpha_{n-1}}, \qquad \beta_n = \frac{b_n + p_n \beta_{n-1}}{q_n - p_n \alpha_{n-1}}.$$

If $x_{n+1} = B$ is the prescribed value, then the solution of the tridiagonal system (2.82) is given as

$$\begin{aligned} x_n &= \alpha_n B + \beta_n \\ x_{n-1} &= \alpha_{n-1} x_n + \beta_{n-1} \\ \ldots & \quad \ldots \\ x_1 &= \alpha_1 x_2 + \beta_1. \end{aligned} \tag{2.89}$$

The procedure converges if $|\alpha_j| \leq 1$. This method is equivalent to the Gauss elimination and also minimizes the storage in the machine computations as only three diagonals are to be stored.

If the problem is to solve only the tridiagonal system, then, set $A = 0$, $B = 0$ in the above algorithm. This gives, from (2.88), $\alpha_0 = 0$, $\beta_0 = 0$. The remaining procedure is the same as above.

Solution of five diagonal system of equations

Another system of algebraic equations that is commonly encountered in the solution of the fourth order differential equations is the five diagonal system

$$\mathbf{Ax} = \mathbf{b}$$

where

$$\mathbf{A} = \begin{bmatrix} r_1 & s_1 & t_1 & & & & \\ q_2 & r_2 & s_2 & t_2 & & & \mathbf{0} \\ p_3 & q_3 & r_3 & s_3 & t_3 & & \\ & \cdots & & \cdots & & & \\ & & & p_{n-1} & q_{n-1} & r_{n-1} & s_{n-1} \\ \mathbf{0} & & & & p_n & q_n & r_n \end{bmatrix}$$

This leads to the recurrence relation

$$p_j x_{j-2} + q_j x_{j-1} + r_j x_j + s_j x_{j+1} + t_j x_{j+2} = b_j \tag{2.90}$$

$2 \le j \le n-2$. For the solution, assume the recurrence relation

$$x_j = \alpha_j - \beta_j x_{j+1} - \gamma_j x_{j+2}, \quad 0 \le j \le n. \tag{2.91}$$

From (2.91), we have

$$x_{j-1} = \alpha_{j-1} - \beta_{j-1} x_j - \gamma_{j-1} x_{j+1}$$

and
$$x_{j-2} = \alpha_{j-2} - \beta_{j-2} x_{j-1} - \gamma_{j-2} x_j .$$

Substituting these expressions in (2.90) and simplifying we get

$$x_j = \frac{1}{r^*}[(b_j - p^*) - (s_j - \gamma_{j-1} q^*) x_{j+1} - t_j x_{j+2}] \tag{2.92}$$

where
$$q^* = q_j - p_j \beta_{j-2}, \; p^* = p_j \alpha_{j-2} + \alpha_{j-1} q^*$$
$$r^* = r_j - p_j \gamma_{j-2} - \beta_{j-1} q^*$$

Comparing (2.92) with (2.91), we have

$$\alpha_j = (b_j - p^*) / r^*,$$
$$\beta_j = (s_j - \gamma_{j-1} q^*) / r^*,$$
$$\gamma_j = t_j / r^*. \tag{2.93}$$

Setting $j = 0$ in (2.91) we have

$$x_0 = \alpha_0 - \beta_0 x_1 - \gamma_0 x_2. \tag{2.94}$$

This equation is satisfied for all x_1, x_2 only if $x_0 = \alpha_0$, $\beta_0 = 0 = \gamma_0$.

If x_0 is prescribed then α_0 is known. If only a given system is to be solved then we set $\alpha_0 = x_0 = 0$. Setting $j = 1$ in (2.91), we get

$$x_1 = \alpha_1 - \beta_1 x_2 - \gamma_1 x_3. \tag{2.95}$$

This equation should be identical with the first equation of the system

$$x_1 = \frac{1}{r_1}[b_1 - s_1 x_2 - t_1 x_3]. \tag{2.96}$$

Comparing (2.95) and (2.96) we have

$$\alpha_1 = \frac{b_1}{r_1}, \; \beta_1 = \frac{s_1}{r_1} \quad \text{and} \quad \gamma_1 = \frac{t_1}{r_1}. \tag{2.97}$$

The remaining values $\alpha_i, \beta_i, \gamma_i$, $i = 2, \ldots n$ are obtained from (2.93). Setting $j = n$ in (2.91), we get

$$x_n = \alpha_n - \beta_n x_{n+1} - \gamma_n x_{n+2}. \tag{2.98}$$

Set $\gamma_n = 0$. If the problem is derived from a boundary value problem in which the values at the end points are prescribed, then $x_{n+1} = g_{n+1}$ is given. Otherwise, set $\beta_n = 0$. Then (2.98) gives either

$$x_n = \alpha_n - \beta_n x_{n+1} \quad \text{or} \quad x_n = \alpha_n.$$

The values $x_{n-1}, x_{n-2}, \ldots, x_1$ are obtained by back substitution in the equation (2.91).

2.6 PROBLEMS AND SOLUTIONS

2.1 Show that the matrix

$$\begin{bmatrix} 12 & 4 & -1 \\ 4 & 7 & 1 \\ -1 & 1 & 6 \end{bmatrix}$$

is positive definite.

Solution

Let $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^T$. Then

$$\mathbf{x}^*\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^T \mathbf{A}\mathbf{x}$$

$$= [\overline{x}_1 \quad \overline{x}_2 \quad \overline{x}_3]\begin{bmatrix} 12 & 4 & -1 \\ 4 & 7 & 1 \\ -1 & 1 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= 12|x_1|^2 + 4(\overline{x}_1 x_2 + \overline{x}_2 x_1) - (\overline{x}_1 x_3 + x_1 \overline{x}_3) + (\overline{x}_2 x_3 + x_2 \overline{x}_3) + 7|x_2|^2 + 6|x_3|^2$$

Let $x_1 = p_1 + iq_1, x_2 = p_2 + iq_2,$ and $x_3 = p_3 + iq_3$.

Then,

$$\begin{aligned} \mathbf{x}^*\mathbf{A}\mathbf{x} &= 12(p_1^2 + q_1^2) + 8(p_1p_2 + q_1q_2) - 2(p_1p_3 + q_1q_3) \\ &\quad + 2(p_2p_3 + q_2q_3) + 7(p_2^2 + q_2^2) + 6(p_3^2 + q_3^2) \\ &= (p_1 - p_3)^2 + (q_1 - q_3)^2 + 4(p_1 + p_2)^2 + 4(q_1 + q_2)^2 \\ &\quad + (p_2 + p_3)^2 + (q_2 + q_3)^2 + 7(p_1^2 + q_1^2) + 2(p_2^2 + q_2^2) + 4(p_3^2 + q_3^2) > 0 \end{aligned}$$

Hence A is positive definite.

2.2 Show that the matrix

$$\begin{bmatrix} 15 & 4 & -2 & 9 & 0 \\ 4 & 7 & 1 & 1 & 1 \\ -2 & 1 & 18 & 6 & 6 \\ 9 & 1 & 6 & 19 & 3 \\ 0 & 1 & 6 & 3 & 11 \end{bmatrix}$$

is positive definite. (Gothenburg Univ., Sweden, BIT 6 (1966), 359)

Solution

A matrix $\mathbf{A}$ is positive definite if $\mathbf{x}^*\mathbf{A}\mathbf{x} > 0$, $\mathbf{x} \neq 0$.

Let $\mathbf{x} = [\,p \quad q \quad r \quad s \quad t]^T$ where

$p = p_1 + ip_2, \quad q = q_1 + iq_2$ etc.

We have

$$\begin{aligned} \mathbf{x}^*\mathbf{A}\mathbf{x} &= [\overline{p} \quad \overline{q} \quad \overline{r} \quad \overline{s} \quad \overline{t}]\,\mathbf{A}\,[p \quad q \quad r \quad s \quad t]^T \\ &= 15|p|^2 + 4(\overline{p}q + p\overline{q}) - 2(\overline{p}r + p\overline{r}) \\ &\quad + 9(\overline{p}s + p\overline{s}) + 7|q|^2 + (\overline{q}r + q\overline{r}) \\ &\quad + (\overline{q}s + q\overline{s}) + (\overline{q}t + q\overline{t}) + 18|r|^2 \\ &\quad + 6(\overline{r}s + r\overline{s}) + 6(\overline{r}t + r\overline{t}) + 19|s|^2 \\ &\quad + 3(\overline{s}t + s\overline{t}) + 11|t|^2 \end{aligned}$$

Substituting $p = p_1 + ip_2$ etc. and simplifying we get

$$\begin{aligned} \mathbf{x}^*\mathbf{A}\mathbf{x} &= 4(p_1 + q_1)^2 + 2(p_1 - r_1)^2 + 9(p_1 + s_1)^2 \\ &\quad + 4(p_2 + q_2)^2 + 2(p_2 - r_2)^2 + 9(p_2 + s_2)^2 \\ &\quad + (q_1 + r_1)^2 + (q_1 + s_1)^2 + (q_1 + t_1)^2 \\ &\quad + (q_2 + r_2)^2 + (q_2 + s_2)^2 + (q_2 + t_2)^2 \\ &\quad + 6(r_1 + s_1)^2 + 6(r_1 + t_1)^2 + 6(r_2 + s_2)^2 \\ &\quad + 6(r_2 + t_2)^2 + 3(s_1 + t_1)^2 + 3(s_2 + t_2)^2 \\ &\quad + 3r_1^2 + 3r_2^2 + t_1^2 + t_2^2 \\ &> 0 \end{aligned}$$

Hence, $\mathbf{A}$ is positive definite.

2.3 The matrix

$$\mathbf{A} = \begin{pmatrix} 1+s & -s \\ s & 1-s \end{pmatrix}$$

is given. Calculate p and q such that $\mathbf{A}^n = p\mathbf{A} + q\mathbf{I}$ and determine $e^{\mathbf{A}}$.

(Lund Univ., Sweden, BIT 28 (1988), 719)

Solution

We have $$\mathbf{A}^2 = \begin{pmatrix} 1+s & -s \\ s & 1-s \end{pmatrix}\begin{pmatrix} 1+s & -s \\ s & 1-s \end{pmatrix} = \begin{pmatrix} 1+2s & -2s \\ 2s & 1-2s \end{pmatrix}$$

$$= 2\begin{pmatrix} 1+s & -s \\ s & 1-s \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 2\mathbf{A} - \mathbf{I}$$

$$\mathbf{A}^3 = \mathbf{A}(2\mathbf{A} - \mathbf{I}) = 2\mathbf{A}^2 - \mathbf{A} = 2(2\mathbf{A} - \mathbf{I}) - \mathbf{A} = 3\mathbf{A} - 2\mathbf{I}$$

$$\mathbf{A}^4 = \mathbf{A}(3\mathbf{A} - 2\mathbf{I}) = 3\mathbf{A}^2 - 2\mathbf{A} = 4\mathbf{A} - 3\mathbf{I}$$

By induction, we get

$$\mathbf{A}^n = n\mathbf{A} + (1-n)\mathbf{I}$$

Hence, $p = n$ and $q = 1 - n$.

We have $$e^{\mathbf{A}} = \mathbf{I} + \frac{\mathbf{A}}{1!} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!} + \cdots$$

$$= \mathbf{I} + \frac{\mathbf{A}}{1!} + \frac{1}{2!}[2\mathbf{A} + (1-2)\mathbf{I}] + \frac{1}{3!}[3\mathbf{A} + (1-3)\mathbf{I}] + \ldots$$

$$= \mathbf{A}\left[1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots\right] + \mathbf{I}\left[1 + \frac{1}{2!}(1-2) + \frac{1}{3!}(1-3) + \cdots\right]$$

$$= e\mathbf{A} + \mathbf{I}\left[\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots\right) - \left(\frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots\right)\right]$$

$$= e\mathbf{A} + \mathbf{I}\,[(e-1) - (e-1)] = e\mathbf{A}.$$

2.4 Solve the following system of equations

(*a*)
$$4x_1 + x_2 + x_3 = 4$$
$$x_1 + 4x_2 - 2x_3 = 4$$
$$3x_1 + 2x_2 - 4x_3 = 6$$

(*b*)
$$x_1 + x_2 - x_3 = 2$$
$$2x_1 + 3x_2 + 5x_3 = -3$$
$$3x_1 + 2x_2 - 3x_3 = 6$$

(*i*) by the Gauss elimination method with partial pivoting,

(*ii*) by the decomposition method with $u_{11} = u_{22} = u_{33} = 1$.

Solution

(*a*) (*i*) Consider the augmented matrix $(\mathbf{A} \mid \mathbf{b})$. Using elementary row transformations, we get

$$(\mathbf{A} \mid \mathbf{b}) = \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 1 & 4 & -2 & 4 \\ 3 & 2 & -4 & 6 \end{array}\right] \sim \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 0 & 15/4 & -9/4 & 3 \\ 0 & 5/4 & -19/4 & 3 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|c} 4 & 1 & 1 & 4 \\ 0 & 15/4 & -9/4 & 3 \\ 0 & 0 & -4 & 2 \end{array}\right]$$

Back substitution gives the solution

$$x_3 = -1/2, \quad x_2 = 1/2 \quad \text{and} \quad x_1 = 1$$

(*ii*) Writing **A = LU**, with $u_{ii} = 1$, we have

$$\begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & -2 \\ 3 & 2 & -4 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Comparing the elements on both sides and solving, we get

$$\mathbf{L} = \begin{bmatrix} 4 & 0 & 0 \\ 1 & 15/4 & 0 \\ 3 & 5/4 & -4 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & 1/4 & 1/4 \\ 0 & 1 & -3/5 \\ 0 & 0 & 1 \end{bmatrix}$$

Solving **Lz = b**, by forward substitution, we get

$$\mathbf{z} = [1 \quad 4/5 \quad -1/2]^T.$$

Solving **Ux = z**, by backward substitution, we have

$$\mathbf{x} = [1 \quad 1/2 \quad -1/2]^T.$$

(*b*) (*i*) Using the elementary row operations on the augmented matrix, we get

$$(\mathbf{A} \mid \mathbf{b}) = \left[\begin{array}{ccc|c} 1 & 1 & -1 & 2 \\ 2 & 3 & 5 & -3 \\ 3 & 2 & -3 & 6 \end{array}\right] \sim \left[\begin{array}{ccc|c} 3 & 2 & -3 & 6 \\ 2 & 3 & 5 & -3 \\ 1 & 1 & -1 & 2 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|c} 3 & 2 & -3 & 6 \\ 0 & 5/3 & 7 & -7 \\ 0 & 1/3 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 3 & 2 & -3 & 6 \\ 0 & 5/3 & 7 & -7 \\ 0 & 0 & -7/5 & 7/5 \end{array}\right]$$

Using, backward substitution, we obtain

$$x_3 = -1, \; x_2 = 0 \quad \text{and} \quad x_1 = 1.$$

(*ii*) Writing **A** = **LU**, with $u_{ii} = 1$, we have

$$\begin{bmatrix} 1 & 1 & -1 \\ 2 & 3 & 5 \\ 3 & 2 & -3 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Comparing the elements on both sides and solving, we get

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -1 & 7 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 7 \\ 0 & 0 & 1 \end{bmatrix}$$

Solving **Lz** = **b**, we get **z** = $[2 \quad -7 \quad -1]^T$

Solving **Ux** = **z**, we get **x** = $[1 \quad 0 \quad -1]^T$.

2.5 Find the inverse of the matrix

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix}$$

by the Gauss-Jordan method.

Solution

Consider the augmented matrix (**A** | **I**). We have

$$(\mathbf{A} \mid \mathbf{I}) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 2 & 3 & -1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -1 & -3 & -2 & 1 & 0 \\ 0 & -5 & 1 & -2 & 0 & 1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & -1 & 0 \\ 0 & 0 & 16 & 8 & -5 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & -5 & -3 & 2 & 0 \\ 0 & 1 & 3 & 2 & -1 & 0 \\ 0 & 0 & 16 & 8 & -5 & 1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1/2 & 7/16 & 5/16 \\ 0 & 1 & 0 & 1/2 & -1/16 & -3/16 \\ 0 & 0 & 1 & 1/2 & -5/16 & 1/16 \end{array}\right)$$

The required inverse is

$$\frac{1}{16}\begin{pmatrix} -8 & 7 & 5 \\ 8 & -1 & -3 \\ 8 & -5 & 1 \end{pmatrix}.$$

2.6 Find the inverse of coefficient matrix of the system

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix}$$

by the Gauss-Jordan method with partial pivoting and hence solve the system.

Solution

Using the augmented matrix [**A** | **I**], we obtain

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 4 & 3 & -1 & 0 & 1 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 4 & 3 & -1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 5 & 3 & 0 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \\ 0 & 11/4 & 15/4 & 0 & -3/4 & 1 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 11/4 & 15/4 & 0 & -3/4 & 1 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 1 & 3/4 & -1/4 & 0 & 1/4 & 0 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 1/4 & 5/4 & 1 & -1/4 & 0 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & -14/11 & 0 & 5/11 & -3/11 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 0 & 10/11 & 1 & -2/11 & -1/11 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & -14/11 & 0 & 5/11 & -3/11 \\ 0 & 1 & 15/11 & 0 & -3/11 & 4/11 \\ 0 & 0 & 1 & 11/10 & -1/5 & -1/10 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 7/5 & 1/5 & -2/5 \\ 0 & 1 & 0 & -3/2 & 0 & 1/2 \\ 0 & 0 & 1 & 11/10 & -1/5 & -1/10 \end{array}\right]$$

Therefore, the solution of the system is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7/5 & 1/5 & -2/5 \\ -3/2 & 0 & 1/2 \\ 11/10 & -1/5 & -1/10 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ -1/2 \end{bmatrix}$$

2.7 Show that the following matrix is nonsingular but it cannot be written as the product of lower and upper triangular matrices, that is, as **LU**.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ -1 & 0 & 2 \end{bmatrix}$$

Solution

We have $|\mathbf{A}| = 10 \neq 0$. Hence, **A** is nonsingular.

Write $$\mathbf{A} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Comparing, we get

$$l_{11} = 1,\ l_{21} = 2,\ l_{31} = -1,$$
$$u_{12} = 2,\ u_{13} = 3,\ l_{22} = 0.$$

Since the pivot $l_{22} = 0$, the next equation for u_{23} is inconsistent and **LU** decomposition of **A** is not possible.

2.8 Calculate the inverse of the n-rowed square matrix **L**

$$\mathbf{L} = \begin{bmatrix} 1 & & & & \mathbf{0} \\ -1/2 & 1 & & & \\ & -2/3 & 1 & & \\ & & \ddots & & \\ \mathbf{0} & & & -(n-1)/n & 1 \end{bmatrix}$$

(Lund Univ., Sweden, BIT 10 (1970), 515)

Solution

Since the inverse of a lower triangular matrix is also lower triangular, we write $\mathbf{L}^{-1} = (p_{ij})$ and

$$\begin{bmatrix} 1 & & & & \mathbf{0} \\ -1/2 & 1 & & & \\ & -2/3 & 1 & & \\ & \ddots & & \ddots & \\ \mathbf{0} & & & -(n-1)/n & 1 \end{bmatrix} \begin{bmatrix} p_{11} & & & \mathbf{0} \\ p_{21} & p_{22} & & \\ \cdots & & \cdots & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} = \mathbf{I}$$

Comparing the elements on both sides, we obtain

$$p_{11} = 1\ ;\ -\frac{1}{2}p_{11} + p_{21} = 0 \quad \text{or} \quad p_{21} = \frac{1}{2},$$
$$p_{22} = 1\ ;\ -\frac{2}{3}\,p_{21} + p_{31} = 0 \quad \text{or} \quad p_{31} = \frac{1}{3},$$
$$-\frac{2}{3}p_{22} + p_{32} = 0 \quad \text{or} \quad p_{32} = \frac{2}{3}\ ;\ p_{33} = 1 \text{ etc.}$$

We find $$p_{ij} = j / i, \quad i \geq j.$$

Hence,
$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & & & & \mathbf{0} \\ 1/2 & 1 & & & \\ 1/3 & 2/3 & 1 & & \\ \vdots & \vdots & \vdots & & \\ 1/n & 2/n & \cdots & (n-1)/n & 1 \end{bmatrix}$$

2.9 Given the system of equations
$$\begin{bmatrix} 2 & 3 & 0 & 0 \\ 2 & 4 & 1 & 0 \\ 0 & 2 & 6 & A \\ 0 & 0 & 4 & B \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 4 \\ C \end{bmatrix}$$
State the solvability and uniqueness conditions for this system. Give the solution when it exists. (Trondheim Univ., Sweden, BIT 26 (1986), 398)

Solution

Applying elementary row transformations on the augmented matrix, we obtain
$$\left(\begin{array}{cccc|c} 2 & 3 & 0 & 0 & 1 \\ 2 & 4 & 1 & 0 & 2 \\ 0 & 2 & 6 & A & 4 \\ 0 & 0 & 4 & B & C \end{array}\right) \sim \left(\begin{array}{cccc|c} 2 & 3 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 2 & 6 & A & 4 \\ 0 & 0 & 4 & B & C \end{array}\right)$$
$$\sim \left(\begin{array}{cccc|c} 2 & 3 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 4 & A & 2 \\ 0 & 0 & 4 & B & C \end{array}\right) \sim \left(\begin{array}{cccc|c} 2 & 3 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 4 & A & 2 \\ 0 & 0 & 0 & B-A & C-2 \end{array}\right)$$
We conclude that

the solution exists and is unique if $B \neq A$,

there is no solution if $B = A$ and $C \neq 2$,

a one parameter family of solutions exists if $B = A$ and $C = 2$,

For $B \neq A$, the solution is
$$x_1 = (8A - 2B - 3AC) / (8(B - A)),$$
$$x_2 = (2B - 4A + AC) / (4(B - A)),$$
$$x_3 = (2B - AC) \ / (4(B - A)),$$
$$x_4 = (C - 2) / (B - A).$$
For $B = A$ and $C = 2$, we have the solution
$$\mathbf{x} = (-0.25, 0.5, 0.5, 0)^T + t(-0.375A, 0.25A - 0.25A, 1)^T,$$
where t is arbitrary.

2.10 We want to solve the tridiagonal system $\mathbf{Ax = b}$, where $\mathbf{A}$ is $(N-1) \times (N-1)$ and
$$\mathbf{A} = \begin{bmatrix} -3 & 1 & & & & & \\ 2 & -3 & 1 & & & & 0 \\ & 2 & -3 & 1 & & & \\ & \cdots & \cdots & \cdots & & & \\ 0 & & & 2 & -3 & 1 \\ & & & & 2 & -3 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$
State the difference equation which replaces this matrix formulation of the problem, and find the solution. (Umea Univ., Sweden, BIT 24 (1984), 257)

Solution

The difference equation is

$$2x_{n-1} - 3x_n + x_{n+1} = 0, \; n = 1, 2, ..., N - 1$$

with $x_0 = -0.5$ and $x_N = 0$. The solution of this constant coefficient difference equation is

$$x_n = A1^n + B\,2^n.$$

Substituting $x_N = 0$, we get $A = -B\,2^N$. Hence

$$x_n = B(2^n - 2^N).$$

We determine B from the first difference equation $-3x_1 + x_2 = 1$. We have

$$-3B(2 - 2^N) + B(2^2 - 2^N) = 1.$$

The solution is
$$B = \frac{1}{2^{N+1} - 2}.$$

Hence,
$$x_n = \frac{2^n - 2^N}{2^{N+1} - 2} = \frac{2^{n-1} - 2^{N-1}}{2^N - 1}, \; n = 1, 2, ..., N - 1.$$

2.11 Given

$$\mathbf{A} = \begin{bmatrix} 5.5 & 0 & 0 & 0 & 0 & 3.5 \\ 0 & 5.5 & 0 & 0 & 0 & 1.5 \\ 0 & 0 & 6.25 & 0 & 3.75 & 0 \\ 0 & 0 & 0 & 5.5 & 0 & 0.5 \\ 0 & 0 & 3.75 & 0 & 6.25 & 0 \\ 3.5 & 1.5 & 0 & 0.5 & 0 & 5.5 \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

(*a*) Find the lower triangular matrix **L** of the Cholesky factorization,

(*b*) Solve the system **Ax = b.** (Inst. Tech. Lyngby, Denmark, BIT 24 (1984), 128)

Solution

(*a*) Write

$$\mathbf{L} = \begin{pmatrix} l_{11} & & & & & \\ l_{21} & l_{22} & & & \mathbf{0} & \\ l_{31} & l_{32} & l_{33} & & & \\ l_{41} & l_{42} & l_{43} & l_{44} & & \\ l_{51} & l_{52} & l_{53} & l_{54} & l_{55} & \\ l_{61} & l_{62} & l_{63} & l_{64} & l_{65} & l_{66} \end{pmatrix}$$

Using $\mathbf{LL}^T = \mathbf{A}$ and comparing we get l_{ij}. We obtain

$$\mathbf{L} = \begin{pmatrix} p & & & & & \\ 0 & p & & & \mathbf{0} & \\ 0 & 0 & 2.5 & & & \\ 0 & 0 & 0 & p & & \\ 0 & 0 & 1.5 & 0 & 2 & \\ 7/(2p) & 3/(2p) & 0 & 1/(2p) & 0 & q \end{pmatrix}$$

where $p = \sqrt{5.5}$ and $q = \sqrt{31/11}$.

(*b*) We have $\mathbf{LL}^T\,\mathbf{x} = \mathbf{b}.$

Set $\mathbf{L}^T\mathbf{x} = \mathbf{z}.$

Solving $\mathbf{Lz} = \mathbf{b}$, we get $\mathbf{z} = (1/p \quad 1/p \quad 0.4 \quad 1/p \quad 0.2 \quad 0)^T.$

Solving $\mathbf{L}^T\mathbf{x} = \mathbf{z}$, we get $\mathbf{x} = (2/11 \quad 2/11 \quad 0.1 \quad 2/11 \quad 0.1 \quad 0)^T.$

2.12 Calculate $\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}$ when $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, with

$$\mathbf{L} = \begin{bmatrix} 1 & & & \mathbf{0} \\ 1 & 2 & & \\ 1 & 2 & 3 & \\ 1 & 2 & 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 \\ 5 \\ 14 \\ 30 \end{bmatrix}.$$

(Umea Univ., Sweden, BIT 24 (1984), 398)

Solution

$$\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B} = \mathbf{C}^T(\mathbf{L}\mathbf{L}^T)^{-1}\,\mathbf{B} = \mathbf{C}^T(\mathbf{L}^T)^{-1}\mathbf{L}^{-1}\mathbf{B}$$
$$= \mathbf{C}^T(\mathbf{L}^{-1})^T\mathbf{L}^{-1}\mathbf{B} = (\mathbf{L}^{-1}\mathbf{C})^T\mathbf{L}^{-1}\mathbf{B}.$$

Since $\mathbf{L}$ is lower triangular, we have

$$\begin{bmatrix} 1 & & & \mathbf{0} \\ 1 & 2 & & \\ 1 & 2 & 3 & \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} l_{11} & & & \mathbf{0} \\ l_{21} & l_{22} & & \\ l_{31} & l_{32} & l_{33} & \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

We find

$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & & & \\ -1/2 & 1/2 & & \mathbf{0} \\ 0 & -1/3 & 1/3 & \\ 0 & 0 & -1/4 & 1/4 \end{bmatrix}$$

and
$$\mathbf{L}^{-1}\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 2 & 2 & 2 \\ 4/3 & 4/3 & 4/3 & 4/3 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{L}^{-1}\mathbf{C} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Hence,
$$\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B} = (\mathbf{L}^{-1}\mathbf{C})^T\,\mathbf{L}^{-1}\mathbf{B}$$
$$= (13 \quad 14 \quad 15 \quad 16).$$

2.13 Find the inverse of the following $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} 1 & & & & & \\ x & 1 & & & \mathbf{0} & \\ x^2 & x & 1 & & & \\ x^3 & x^2 & x & 1 & & \\ \vdots & & \vdots & & & \\ x^{n-1} & x^{n-2} & & \cdots & x^2 & x & 1 \end{bmatrix}$$

(Lund Univ. Sweden, BIT 11 (1971), 338)

Solution

The inverse of a lower triangular matrix is also a lower triangular matrix. Let the inverse of the given matrix $\mathbf{A}$ be $\mathbf{L}$. Using the identity $\mathbf{AL} = \mathbf{I}$, we get

$$\begin{bmatrix} 1 & & & & & \\ x & 1 & & & & \mathbf{0} \\ x^2 & x & 1 & & & \\ \vdots & & \vdots & & & \\ x^{n-1} & & \cdots & x^2 & x & 1 \end{bmatrix} \begin{bmatrix} l_{11} & & & & \mathbf{0} \\ l_{21} & l_{22} & & & \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & & & & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Comparing elements on both sides, we get

$$l_{11} = 1,\ x l_{11} + l_{21} = 0, \quad \text{or} \quad l_{21} = -x,\ l_{22} = 1 \text{ etc.}$$

We find that
$$l_{ij} = \begin{vmatrix} 1, & \text{if } i = j \\ -x, & \text{if } i = j+1 \\ 0, & \text{otherwise.} \end{vmatrix}$$

Hence, we obtain

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & & & & \\ -x & 1 & & & \mathbf{0} \\ 0 & -x & 1 & & \\ \vdots & & & & \\ 0 & \cdots & 0 & -x & 1 \end{bmatrix}$$

2.14 Find the inverse of the matrix

$$\begin{bmatrix} 2 & -1 & 2 \\ -1 & 1 & -1 \\ 2 & -1 & 3 \end{bmatrix}$$

by the Cholesky method.

Solution

Using the Cholesky method, write

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

Comparing the coefficients, we get

$$l_{11}^2 = 2,\ l_{11} = \sqrt{2}\ ;$$
$$l_{21} = -1 / \sqrt{2}\ ; l_{31} = 2 / \sqrt{2}\ ;$$
$$l_{22}^2 = 1/2,\ l_{22} = 1 / \sqrt{2}\ ;$$
$$l_{32} = 0\ ;\ l_{33} = 1.$$

Hence,
$$\mathbf{L} = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ \sqrt{2} & 0 & 1 \end{bmatrix}$$

Since $\mathbf{L}^{-1}$ is also a lower triangular matrix, write

$$\begin{bmatrix} \sqrt{2} & 0 & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ \sqrt{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} l_{11}^* & 0 & 0 \\ l_{21}^* & l_{22}^* & 0 \\ l_{31}^* & l_{32}^* & l_{33}^* \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We find
$$\mathbf{L}^{-1} = \begin{bmatrix} 1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & \sqrt{2} & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Hence,
$$\mathbf{A}^{-1} = (\mathbf{L}\mathbf{L}^T)^{-1} = (\mathbf{L}^T)^{-1}\,\mathbf{L}^{-1} = (\mathbf{L}^{-1})^T\,\mathbf{L}^{-1} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

2.15 Find the Cholesky factorization of

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

(Oslo Univ., Norway, BIT 20 (1980), 529)

Solution

Let $\mathbf{L} = (l_{ij})$ where $l_{ij} = 0$ for $i < j$.

Writing the given matrix as $\mathbf{LL}^T$, we obtain

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} & 0 \\ l_{51} & l_{52} & l_{53} & l_{54} & l_{55} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} & l_{41} & l_{51} \\ 0 & l_{22} & l_{32} & l_{42} & l_{52} \\ 0 & 0 & l_{33} & l_{43} & l_{53} \\ 0 & 0 & 0 & l_{44} & l_{54} \\ 0 & 0 & 0 & 0 & l_{55} \end{bmatrix}$$

On comparing corresponding elements on both sides and solving, we get

$$\begin{aligned} &l_{11} = 1,\ l_{21} = -1,\ l_{i1} = 0,\ i = 3, 4, 5, \\ &l_{22} = 1,\ l_{32} = -1,\ l_{i2} = 0,\ i = 4, 5 \\ &l_{33} = 1,\ l_{43} = -1,\ l_{53} = 0, \\ &l_{44} = 1,\ l_{54} = -1,\ l_{55} = 1. \end{aligned}$$

Hence,
$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

2.16 Determine the inverse of the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix}$$

using the partition method. Hence, find the solution of the system of equations

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 4x_1 + 3x_2 - x_3 &= 6 \\ 3x_1 + 5x_2 + 3x_3 &= 4 \end{aligned}$$

Solution

Let the matrix $\mathbf{A}$ be partitioned as

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{E} & \mathbf{D} \end{array}\right] = \left[\begin{array}{cc|c} 1 & 1 & 1 \\ 4 & 3 & -1 \\ \hline 3 & 5 & 3 \end{array}\right] \quad \text{and} \quad \mathbf{A}^{-1} = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{Z} & \mathbf{V} \end{array}\right]$$

Now,
$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & 1 \\ 4 & 3 \end{bmatrix}^{-1} = -\begin{bmatrix} 3 & -1 \\ -4 & 1 \end{bmatrix}$$

$$\mathbf{D} - \mathbf{E}\mathbf{B}^{-1}\mathbf{C} = 3 + [3 \quad 5] \begin{bmatrix} 3 & -1 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -10$$

$$\mathbf{V} = (\mathbf{D} - \mathbf{E}\mathbf{B}^{-1}\mathbf{C}^{-1}) = -\frac{1}{10}$$

$$\mathbf{Y} = -\mathbf{B}^{-1}\mathbf{C}\mathbf{V} = \begin{bmatrix} 3 & -1 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \left(-\frac{1}{10}\right) = -\frac{1}{10} \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

$$\mathbf{Z} = -\mathbf{V}\,\mathbf{E}\,\mathbf{B}^{-1} = -\frac{1}{10} [3 \quad 5] \begin{bmatrix} 3 & -1 \\ -4 & 1 \end{bmatrix} = -\frac{1}{10} [-11 \quad 2]$$

$$\mathbf{X} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\,\mathbf{C}\mathbf{Z}$$

$$= \begin{bmatrix} -3 & 1 \\ 4 & -1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 3 & -1 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} [-11 \quad 2]$$

$$= \begin{bmatrix} -3 & 1 \\ 4 & -1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} -44 & 8 \\ 55 & -10 \end{bmatrix} = \begin{bmatrix} 1.4 & 0.2 \\ -1.5 & 0 \end{bmatrix}$$

Hence,
$$\mathbf{A}^{-1} = \begin{bmatrix} 1.4 & 0.2 & -0.4 \\ -1.5 & 0 & 0.5 \\ 1.1 & -0.2 & -0.1 \end{bmatrix}$$

The solution of the given system of equations is

$$\mathbf{x} = \begin{bmatrix} 1.4 & 0.2 & -0.4 \\ -1.5 & 0 & 0.5 \\ 1.1 & -0.2 & -0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \\ -0.5 \end{bmatrix}$$

2.17 Find the inverse of the matrix

$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

by the partition method.

Solution

We partition the given matrix as

$$\mathbf{A} = \left[\begin{array}{cc|cc} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ \hline 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{E} & \mathbf{D} \end{array}\right]$$

and write the inverse matrix in the form

$$\mathbf{A}^{-1} = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{Z} & \mathbf{V} \end{array}\right]$$

Using the fact that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, we obtain

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{E} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{BX} + \mathbf{CZ} & \mathbf{BY} + \mathbf{CV} \\ \mathbf{EX} + \mathbf{DZ} & \mathbf{EY} + \mathbf{DV} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Hence,
$$\mathbf{BX} + \mathbf{CZ} = \mathbf{I}, \quad \mathbf{BY} + \mathbf{CV} = \mathbf{0},$$
$$\mathbf{EX} + \mathbf{DZ} = \mathbf{0}, \quad \mathbf{EY} + \mathbf{DV} = \mathbf{I}.$$

We find $$\mathbf{B}^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Solving the above matrix equations, we get

$$\mathbf{V} = (\mathbf{D} - \mathbf{EB}^{-1}\,\mathbf{C})^{-1} = \begin{bmatrix} 4/3 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \frac{3}{5}\begin{bmatrix} 2 & -1 \\ -1 & 4/3 \end{bmatrix};$$

$$\mathbf{Y} = -\,\mathbf{B}^{-1}\,\mathbf{CV} = -\frac{1}{5}\begin{bmatrix} -2 & 1 \\ 4 & -2 \end{bmatrix};$$

$$\mathbf{Z} = -\,\mathbf{VEB}^{-1} = -\frac{1}{5}\begin{bmatrix} -2 & 4 \\ 1 & -2 \end{bmatrix};$$

$$\mathbf{X} = \mathbf{B}^{-1}\,(\mathbf{I} - \mathbf{CZ}) = \frac{1}{5}\begin{bmatrix} 4 & -3 \\ -3 & 6 \end{bmatrix}.$$

Thus we obtain

$$\mathbf{A}^{-1} = \frac{1}{5}\begin{bmatrix} 4 & -3 & 2 & -1 \\ -3 & 6 & -4 & 2 \\ 2 & -4 & 6 & -3 \\ -1 & 2 & -3 & 4 \end{bmatrix}.$$

2.18 (*a*) What is wrong in the following computation ?

$$\begin{bmatrix} 1 & 0.01 \\ 1 & 1 \end{bmatrix}^n = \left\{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} + 10^{-2}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right\}^n$$

$$= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^n + n \times 10^{-2}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{n-1}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

since $$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}^k = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for} \quad k \ge 2.$$

(*b*) Compute $\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}^{10}$ exactly. (Lund Univ., Sweden, BIT 12 (1972), 589)

Solution

(*a*) We know that

$$(\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2 = \mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$$

if and only if $\mathbf{AB} = \mathbf{BA}$.

In the given example, let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = 10^{-2}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Then, $$\mathbf{AB} = 10^{-2}\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{BA} = 10^{-2}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

Hence, $\mathbf{AB} \ne \mathbf{BA}$ and therefore the given expression is not valid.

(*b*) The given matrix is symmetric. Hence, there exists an orthogonal similarity matrix **S** which reduces **A** to its diagonal form **D**.

Let $$\mathbf{S} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Then, $$\mathbf{S}^{-1}\mathbf{AS} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$= \begin{pmatrix} 1+0.1\sin 2\theta & 0.1\cos 2\theta \\ 0.1\cos 2\theta & 1-0.1\sin 2\theta \end{pmatrix} = \mathbf{D}.$$

Since **D** is a diagonal matrix, we choose θ such that $0.1 \cos 2\theta = 0$, which gives $\theta = \pi / 4$. Therefore,

$$\mathbf{A} = \mathbf{SDS}^{-1} \quad \text{where } \mathbf{D} = \begin{pmatrix} 1.1 & 0 \\ 0 & 0.9 \end{pmatrix}$$

Hence, $$\mathbf{A}^{10} = \mathbf{SD}^{10}\mathbf{S}^{-1}$$

$$= \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} a+b & a-b \\ a-b & a+b \end{pmatrix}$$

where $a = (1.1)^{10}$ and $b = (0.9)^{10}$.

2.19 Compute $\mathbf{A}^{10}$ where

$$\mathbf{A} = \frac{1}{9}\begin{bmatrix} 4 & 1 & -8 \\ 7 & 4 & 4 \\ 4 & -8 & 1 \end{bmatrix}$$

(Uppsala Univ., Sweden, BIT 14 (1974), 254)

Solution

We find $$\mathbf{A}^2 = \mathbf{AA} = \frac{1}{9}\begin{bmatrix} -1 & 8 & -4 \\ 8 & -1 & -4 \\ -4 & -4 & -7 \end{bmatrix}$$

$$\mathbf{A}^4 = \mathbf{A}^2\mathbf{A}^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}.$$

Hence, $$\mathbf{A}^8 = \mathbf{A}^4\mathbf{A}^4 = \mathbf{I},$$

$$\mathbf{A}^{10} = \mathbf{A}^8\mathbf{A}^2 = \mathbf{A}^2 = \frac{1}{9}\begin{bmatrix} -1 & 8 & -4 \\ 8 & -1 & -4 \\ -4 & -4 & -7 \end{bmatrix}.$$

2.20. For a linear system of equations of the kind

$$(\mathbf{I} - \mathbf{UV}^T)\mathbf{x} = \mathbf{b}$$

Shermann-Morrisons's formula gives the solution

$$\mathbf{x} = \left[\mathbf{I} + \frac{\mathbf{UV}^T}{1-\mathbf{V}^T\mathbf{U}}\right]\mathbf{b}.$$

Let $\mathbf{U}^T = [0 \quad 1 \quad 2]$, $\mathbf{V}^T = [1 \quad 0 \quad 1]$ and $\mathbf{b}^T = [1 \quad -1 \quad -3]$.
Use Shermann-Morrisons formula to solve the system

$$\mathbf{Ax} = \mathbf{b} \quad \text{when } \mathbf{A} = \mathbf{I} - \mathbf{UV}^T.$$

(Royal Inst. Tech., Stockholm, Sweden, BIT 26 (1986), 135)

Solution

We have $$\mathbf{UV}^T = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix}, \quad \mathbf{V}^T\mathbf{U} = 2.$$

Therefore, $$\mathbf{x} = [\mathbf{I} - \mathbf{UV}^T]\,\mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & -1 \\ -2 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

2.21 The matrix $\mathbf{A}$ is rectangular with m rows and n columns, $n < m$. The matrix $\mathbf{A}^T\mathbf{A}$ is regular. Let $\mathbf{X} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$. Show that $\mathbf{AXA} = \mathbf{A}$ and $\mathbf{XAX} = \mathbf{X}$. Show that in the sense of the method of least squares, the solution of the system $\mathbf{Ax} = \mathbf{b}$ can be written as $\mathbf{x} = \mathbf{Xb}$. Calculate $\mathbf{X}$ when

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$ (Lund Univ., Sweden, BIT 25 (1985), 428)

Solution

Note that $\mathbf{A}^T \mathbf{A}$ is an $n \times n$ regular matrix. We have

$$\mathbf{AXA} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T\mathbf{A} = \mathbf{AI} = \mathbf{A}$$

and $$\mathbf{XAX} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T\mathbf{A}\,(\mathbf{A}^T\mathbf{A})^{-1} \mathbf{A}^T = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T = \mathbf{X}.$$

The given system is

$$\mathbf{Ax} = \mathbf{b}$$

$$(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T\mathbf{Ax} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$$

or $\mathbf{x} = \mathbf{Xb}$, which is the least square solution of the given problem, as described by

$$\mathbf{A}^T (\mathbf{Ax} - \mathbf{b}) = \mathbf{0}.$$

For the given $\mathbf{A}$, we have

$$\mathbf{X} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$$

$$= \frac{1}{6}\begin{bmatrix} 29 & -9 \\ -9 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 11 & 2 & -7 \\ -3 & 0 & 3 \end{bmatrix}.$$

NORMS AND APPLICATIONS

2.22 $\mathbf{A}$ is a given nonsingular $n \times n$ matrix, $\mathbf{u}$ is a given $n \times 1$ vector, and $\mathbf{v}^T$ is a given $1 \times n$ vector

(*a*) Show that

$$(\mathbf{A} - \mathbf{uv}^T)^{-1} \approx \mathbf{A}^{-1} + \alpha \mathbf{A}^{-1} \mathbf{uv}^T \mathbf{A}^{-1}$$

where α is a scalar. Determine α and give conditions for the existence of the inverse on the left hand side.

(*b*) Discuss the possibility of a breakdown of the algorithm even though $\mathbf{A}$ is neither singular nor ill-conditioned, and describe how such difficulties may be overcome.

(Stockholm Univ., Sweden, BIT 4 (1964), 61)

Solution

(*a*) We write

$$(\mathbf{A} - \mathbf{uv}^T)^{-1} = [\mathbf{A}(\mathbf{I} - \mathbf{A}^{-1} \mathbf{uv}^T)]^{-1} = (\mathbf{I} - \mathbf{A}^{-1} \mathbf{uv}^T)^{-1} \mathbf{A}^{-1}.$$

The required inverse exists if

$$\| \mathbf{A}^{-1} \mathbf{uv}^T \| < 1, \quad \text{or} \quad \text{iff} \quad \rho(\mathbf{A}^{-1} \mathbf{uv}^T) < 1.$$

If $\| \mathbf{A}^{-1} \mathbf{uv}^T \| < 1$, then

$$(\mathbf{I} - \mathbf{A}^{-1}\mathbf{uv}^T)^{-1} = \mathbf{I} + \mathbf{A}^{-1} \mathbf{uv}^T + (\mathbf{A}^{-1} \mathbf{uv}^T)^2 + \ldots$$

Hence,

$$(\mathbf{A} - \mathbf{uv}^T)^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\,\mathbf{uv}^T\,\mathbf{A}^{-1} + \ldots \approx \mathbf{A}^{-1} + \mathbf{A}^{-1}\,\mathbf{uv}^T\,\mathbf{A}^{-1}$$

Therefore, $\alpha = 1$.

(*b*) The above algorithm may fail if

(*i*) $\mathbf{uv}^T = \mathbf{A}$, or

(*ii*) $|\,\mathbf{I} - \mathbf{A}^{-1}\,\mathbf{uv}^T\,| = 0$, or

(*iii*) $\|\,\mathbf{A}^{-1}\,\mathbf{uv}^T\,\| \ge 1$.

However, if $\mathbf{uv}^T$ (an $n \times n$ matrix) is nonsingular, then we may write the expansion as

$$(\mathbf{A} - \mathbf{uv}^T)^{-1} = ((\mathbf{uv}^T)\,[(\mathbf{uv}^T)^{-1}\,\mathbf{A} - \mathbf{I}])^{-1} = -\,[\mathbf{I} - (\mathbf{uv}^T)^{-1}\,\mathbf{A}]^{-1}\,(\mathbf{uv}^T)^{-1}.$$

Setting $\mathbf{H} = \mathbf{uv}^T$, we obtain

$$(\mathbf{A} - \mathbf{H})^{-1} = -\,[\mathbf{I} - \mathbf{H}^{-1}\,\mathbf{A}]^{-1}\,\mathbf{H}^{-1} = -\,[\mathbf{I} + \mathbf{H}^{-1}\,\mathbf{A} + (\mathbf{H}^{-1}\,\mathbf{A})^2 + \ldots]\,\mathbf{H}^{-1}$$

if $\qquad \|\,\mathbf{H}^{-1}\,\mathbf{A}\,\| < 1.$

Hence, we obtain

$$(\mathbf{A} - \mathbf{uv}^T)^{-1} \approx -\,[\mathbf{I} + \mathbf{H}^{-1}\mathbf{A}]\mathbf{H}^{-1} = -\,[\mathbf{I} + (\mathbf{uv}^T)^{-1}\,\mathbf{A}]\,(\mathbf{uv}^T)^{-1}.$$

2.23 The matrix **B** is defined as

$$\mathbf{B} = \mathbf{I} + ir\mathbf{A}^2$$

where **I** is the identity matrix, **A** is a Hermitian matrix and $i^2 = -1$. Show that $\|\,\mathbf{B}\,\| > 1$, for all real $r \ne 0$. $\|\,.\,\|$ denotes the Hilbert norm. (Lund Univ., Sweden BIT 9 (1969), 87)

Solution

We have $\qquad \mathbf{B} = \mathbf{I} + ir\mathbf{A}^2.$

Since **A** is Hermitian, we have

$$\mathbf{B}^*\mathbf{B} = (\mathbf{I} - ir\mathbf{A}^2)(\mathbf{I} + ir\mathbf{A}^2) = \mathbf{I} + r^2\mathbf{A}^4.$$

Using the Hilbert norm, we obtain

$$\|\,\mathbf{B}\,\| = \sqrt{\rho(\mathbf{B}^*\mathbf{B})} = \sqrt{\rho(\mathbf{I} + r^2\mathbf{A}^4)} = \sqrt{1 + r^2\lambda^4} > 1$$

where $\lambda = \rho(\mathbf{A})$ and $r \ne 0$.

2.24 Let **R** be a $n \times n$ triangular matrix with unit diagonal elements and with the absolute value of non-diagonal elements less than or equal to 1. Determine the maximum possible value of the maximum norm $\|\,\mathbf{R}^{-1}\,\|$. (Stockholm Univ., Sweden, BIT 8(1968), 59)

Solution

Without loss of generality, we assume that **R** is a lower triangular matrix, $\mathbf{R} = (r_{ij})$, where

$$r_{ii} = 1,\ r_{ij} = 0 \text{ for } i < j \quad \text{and} \quad |\,r_{ij}\,| \le 1 \text{ for } i > j.$$

Let $\qquad \mathbf{R}^{-1} = (l_{ij}),\ l_{ij} = 0 \text{ for } i < j.$

Since $\mathbf{RR}^{-1} = \mathbf{I}$, we have

$$\begin{bmatrix} 1 & & & & \\ r_{21} & 1 & & & \mathbf{0} \\ r_{31} & r_{32} & 1 & & \\ \vdots & & & & \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \mathbf{0} \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & & & & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & 1 \end{bmatrix}$$

Comparing the corresponding elements on both sides, we get

$$l_{ii} = 1,\ r_{21} + l_{21} = 0 \quad \text{or} \quad l_{21} = -\, r_{21},$$
$$r_{31} + r_{32}l_{21} + l_{31} = 0 \quad \text{or} \quad l_{31} = -\,(r_{31} + r_{32}l_{21}),$$
$$r_{32} + l_{32} = 0 \quad \text{or} \quad l_{32} = -\, r_{32},$$
$$r_{41} + r_{42}l_{21} + r_{43}l_{31} + l_{41} = 0 \quad \text{or} \quad l_{41} = -\,(r_{41} + r_{42}l_{21} + r_{43}l_{31}) \text{ etc.}$$

Hence, we have

$$| l_{21} | \le 1,\ | l_{31} | \le 2,$$
$$| l_{32} | \le 1,\ | l_{41} | \le 2^2, \ldots,$$
$$| l_{n1} | \le 2^{n-2}.$$

Using the maximum norm, we get

$$\| \mathbf{R}^{-1} \| \le 1 + 1 + 2 + 2^2 + \ldots + 2^{n-2}$$
$$= 2 + 2(1 + 2 + \ldots + 2^{n-3}) = 2^{n-1}.$$

Hence, the maximum possible value of the maximum norm of $\mathbf{R}^{-1}$ is 2^{n-1}.

2.25 The $n \times n$ matrix $\mathbf{A}$ satisfies

$$\mathbf{A}^4 = -\, 1.6\, \mathbf{A}^2 - 0.64\, \mathbf{I}$$

Show that $\lim_{m \to \infty} \mathbf{A}^m$ exists and determine this limit.

(Inst. Tech., Gothenburg, Sweden, BIT 11 (1971), 455)

Solution

From the given matrix equation, we have

$$(\mathbf{A}^2 + 0.8\, \mathbf{I})^2 = \mathbf{0}.$$

We get
$$\mathbf{A}^2 = -\, 0.8\, \mathbf{I}$$
$$\mathbf{A}^3 = -\, 0.8\, \mathbf{A},$$
$$\mathbf{A}^4 = -\, 0.8\, \mathbf{A}^2 = (-\, 0.8)^2\, \mathbf{I},$$

etc. Hence,
$$\mathbf{A}^m = (-\, 0.8)^{m/2}\, \mathbf{I}, \text{ if } m \text{ is even},$$
$$\mathbf{A}^m = (-\, 0.8)^{(m-1)/2}\, \mathbf{A}, \text{ if } m \text{ is odd}.$$

As $m \to \infty$, we have in both cases that $\lim_{m \to \infty} \mathbf{A}^m = \mathbf{0}$.

2.26 Compute $[\ \ln (\mathbf{I} + \frac{1}{4}\mathbf{A})]\ \mathbf{Y}$, when

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

correct to four decimals. (Uppsala Univ. Sweden, BIT 27 (1987), 129)

Solution

Since $\frac{1}{4} \| \mathbf{A} \| = \frac{1}{2}$, we have

$$\left[\ln \left(\mathbf{I} + \frac{1}{4} \mathbf{A} \right) \right] \mathbf{Y} = \left[\frac{(\mathbf{A}/4)}{1} - \frac{(\mathbf{A}/4)^2}{2} + \frac{(\mathbf{A}/4)^3}{3} - \frac{(\mathbf{A}/4)^4}{4} + \cdots \right] \mathbf{Y}$$

We get,
$$\mathbf{AY} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\mathbf{A}^2\, \mathbf{Y} = 3(2) \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \mathbf{A}^3\, \mathbf{Y} = 3(2^2) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ldots,$$

$$\mathbf{A}^m \mathbf{Y} = 3(2^{m-1}) \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Hence, $$\left[\ln\left(\mathbf{I} + \frac{1}{4}\mathbf{A}\right)\right]\mathbf{Y} = \frac{3}{2}\left[\frac{1}{2} - \frac{1}{8} + \frac{1}{24} - \frac{1}{64} + \cdots\right]\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{3}{2}\left[\frac{(1/2)}{1} - \frac{(1/2)^2}{2} + \frac{(1/2)^3}{3} - \cdots\right]\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{3}{2}\ln\left(1 + \frac{1}{2}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{3}{2}\ln\left(\frac{3}{2}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0.6082\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

2.27 The matrix **A** is defined by $a_{ij} = 1$, when $i + j$ is even and $a_{ij} = 0$, when $i + j$ is odd. The order of the matrix is $2n$. Show that

$$\| \mathbf{A} \|_F = \| \mathbf{A} \|_\infty = n,$$

where $\| \mathbf{A} \|_F$ is the Frobenius norm, and that

$$\sum_{k=1}^{\infty} \left(\frac{1}{2n}\right)^k \mathbf{A}^k = \frac{1}{n}\mathbf{A}.$$

(Uppsala Univ. Sweden, BIT 27 (1987), 628)

Solution

Note that **A** is of order $2n$ and is of the form

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & \cdots & 1 & 0 \\ \cdots & & & \cdots & & & & \\ 1 & 0 & 1 & 0 & 1 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Also, **A** is symmetric. We have

$$\| \mathbf{A} \|_\infty = \max_i \sum_k |a_{ik}| = n, \quad \| \mathbf{A} \|_F = \sqrt{n^2} = n.$$

We have, by multiplying

$$\mathbf{A}^2 = n\,\mathbf{A}, \quad \mathbf{A}^3 = n^2\,\mathbf{A}, \ldots, \quad \mathbf{A}^k = n^{k-1}\,\mathbf{A}.$$

Hence,

$$\sum_{k=1}^{n} \left(\frac{1}{2n}\right)^k \mathbf{A}^k = \frac{1}{2n}\mathbf{A} + \left(\frac{1}{2n}\right)^2 \mathbf{A}^2 + \left(\frac{1}{2n}\right)^3 \mathbf{A}^3 + \cdots$$

$$= \frac{1}{2n}\mathbf{A} + \frac{1}{2^2 n}\mathbf{A} + \frac{1}{2^3 n}\mathbf{A} + \cdots = \frac{1}{2n}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right)\mathbf{A} = \frac{1}{n}\mathbf{A}.$$

2.28 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & 1 & -1 \\ -1 & 1 & 2 & -1 \\ 1 & -1 & -1 & 2 \end{bmatrix}$$

(*a*) Determine the spectral norm $\rho(\mathbf{A})$.

(*b*) Determine a vector $\mathbf{x}$ with $\| \mathbf{x} \|_2 = 1$ satisfying $\| \mathbf{Ax} \|_2 = \rho(\mathbf{A})$.

(Inst. Tech., Lund, Sweden, BIT 10 (1970), 288)

Solution

(*a*) The given matrix $\mathbf{A}$ is symmetric. Hence, $\| \mathbf{A} \|_2 = \rho(\mathbf{A})$.

The eigenvalues of $\mathbf{A}$ are given by

$$| \mathbf{A} - \lambda \mathbf{I} | = (1 - \lambda)^2 (\lambda^2 - 6\lambda + 5) = 0$$

which gives $\lambda = 1, 1, 1, 5$. Hence, $\| \mathbf{A} \|_2 = \rho(\mathbf{A}) = 5$.

(*b*) For $\lambda = 5$. We have the eigensystem

$$\begin{bmatrix} -3 & -1 & -1 & 1 \\ -1 & -3 & 1 & -1 \\ -1 & 1 & -3 & -1 \\ 1 & -1 & -1 & -3 \end{bmatrix} \mathbf{x} = \mathbf{0}$$

Solving this system, we get $\mathbf{x} = [1 \quad -1 \quad -1 \quad 1]^T$. Normalizing, such that

$$\| \mathbf{x} \|_2 = (\Sigma |\, x_i \,|^2)^{1/2} = 1,$$

we obtain the eigenvector as

$$\mathbf{x} = [1/2 \quad -1/2 \quad -1/2 \quad 1/2]^T.$$

2.29 Determine the condition number of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 9 \\ 4 & 9 & 16 \\ 9 & 16 & 25 \end{bmatrix}$$

using the (*i*) maximum absolute row sum norm, and (*ii*) spectral norm.

Solution

(*i*) We have

$$\mathbf{A}^{-1} = -\frac{1}{8} \begin{bmatrix} -31 & 44 & -17 \\ 44 & -56 & 20 \\ -17 & 20 & -7 \end{bmatrix} = \begin{bmatrix} 31/8 & -44/8 & 17/8 \\ -44/8 & 56/8 & -20/8 \\ 17/8 & -20/8 & 7/8 \end{bmatrix}$$

$$\| \mathbf{A} \|_\infty = \textit{maximum absolute row sum norm} \text{ for } \mathbf{A}$$
$$= \max \{14, \quad 29, \quad 50\} = 50.$$
$$\| \mathbf{A}^{-1} \|_\infty = \textit{maximum absolute row sum norm} \text{ for } \mathbf{A}^{-1}$$
$$= \max \left\{ \left(\frac{31}{8} + \frac{44}{8} + \frac{17}{8} \right), \quad \left(\frac{44}{8} + \frac{56}{8} + \frac{20}{8} \right), \quad \left(\frac{17}{8} + \frac{20}{8} + \frac{7}{8} \right) \right\}$$
$$= \max \left\{ \frac{92}{8}, \quad 15, \quad \frac{44}{8} \right\} = 15.$$

Therefore, $\kappa(\mathbf{A}) = \| \mathbf{A} \|_\infty \| \mathbf{A}^{-1} \|_\infty = 750.$

(*ii*) The given matrix is real and symmetric. Therefore, $\kappa(\mathbf{A}) = \lambda^* / \mu^*$, where λ^* and μ^* are the largest and the smallest eigenvalues in modulus of $\mathbf{A}$.

The characteristic equation of $\mathbf{A}$ is given by

$$| \mathbf{A} - \lambda \mathbf{I} | = \begin{vmatrix} 1-\lambda & 4 & 9 \\ 4 & 9-\lambda & 16 \\ 9 & 16 & 25-\lambda \end{vmatrix} = -\lambda^3 + 35\lambda^2 + 94\lambda - 8 = 0.$$

A root lies in (0, 0.1). Using the Newton-Raphson method

$$\lambda_{k+1} = \lambda_k - \frac{\lambda_k^3 - 35\lambda_k^2 - 94\lambda_k + 8}{3\lambda_k^2 - 70\lambda_k - 94}, \quad k = 0, 1, 2, \ldots$$

with $\lambda_0 = 0.1$, we get $\lambda_1 = 0.08268$, $\lambda_2 = 0.08257$, $\lambda_3 = 0.08257$. The root correct to five places is 0.08257. Dividing the characteristic equation by $(x - 0.08257)$, we get the deflated polynomial as

$$x^2 - 34.91743x - 96.88313 = 0$$

whose roots are 37.50092, – 2.58349. Hence,

$$\kappa(\mathbf{A}) = \frac{37.50092}{0.08257} \approx 454.17.$$

2.30 Let $\mathbf{A}(\alpha) = \begin{bmatrix} 0.1\alpha & 0.1\alpha \\ 1.0 & 1.5 \end{bmatrix}$

Determine α such that cond $(\mathbf{A}(\alpha))$ is minimized. Use the maximum norm.

(Uppsala Univ. Sweden, BIT 16 (1976), 466)

Solution

For the matrix $\mathbf{A}$, cond $(\mathbf{A}) = \| \mathbf{A} \| \, \| \mathbf{A}^{-1} \|$.

Here, we have

$$\mathbf{A}(\alpha) = \begin{bmatrix} 0.1\alpha & 0.1\alpha \\ 1.0 & 1.5 \end{bmatrix}$$

and $$\mathbf{A}^{-1}(\alpha) = \frac{1}{0.05\alpha} \begin{bmatrix} 1.5 & -0.1\alpha \\ -1.0 & 0.1\alpha \end{bmatrix}.$$

Using maximum norm, we get

$$\| \mathbf{A}(\alpha) \| = \max\,[0.2\,|\alpha|, 2.5],$$

$$\| \mathbf{A}^{-1}(\alpha) \| = \max \left[\frac{2|\alpha| + 30}{|\alpha|}, \frac{2|\alpha| + 20}{|\alpha|} \right] = \frac{2|\alpha| + 30}{|\alpha|}.$$

We have, $$\text{cond}\,(\mathbf{A}(\alpha)) = \frac{1}{|\alpha|}\,[2\,|\alpha| + 30][\max\,[0.2\,|\alpha|, 2.5]]$$

We want to determine α such that cond $(\mathbf{A}(\alpha))$ is minimum. We have

$$\text{cond}\,(\mathbf{A}(\alpha)) = \max \left[0.4\,|\alpha| + 6,\ 5 + \frac{75}{|\alpha|} \right] = \text{minimum}.$$

Choose α such that

$$0.4\,|\alpha| + 6 = 5 + \frac{75}{|\alpha|}$$

which gives $|\alpha| = 12.5$. The minimum value of cond $(\mathbf{A}(\alpha)) = 11$.

2.31 Estimate the effect of a disturbance $[\varepsilon_1, \varepsilon_2]^T$ on the right hand side of the system of equations

$$\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

if $|\varepsilon_1|, |\varepsilon_2| \le 10^{-4}$.

(Uppsala Univ. Sweden, BIT 15 (1975), 335)

Solution

The solution of the system of equations $\mathbf{Ax} = \mathbf{b}$, is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$

if $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ is the solution when the disturbance $\delta\mathbf{b} = [\varepsilon_1 \quad \varepsilon_2]^T$ is present on the right hand side, we obtain

$$\hat{\mathbf{x}} = \mathbf{A}^{-1}(\mathbf{b} + \delta\mathbf{b}).$$

Therefore, we get, $\delta\mathbf{x} = \mathbf{A}^{-1}\,\delta\mathbf{b}$, or $\| \delta\mathbf{x} \| \le \| \mathbf{A}^{-1} \| \, \| \delta\mathbf{b} \|$.

Since, $\mathbf{A}^{-1} = \dfrac{1}{5}\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$, we have

$$\| \mathbf{A}^{-1} \| = \rho(\mathbf{A}^{-1}) = \sqrt{0.2}.$$

We also have $\| \delta\mathbf{b} \| \le \sqrt{2}\,\varepsilon$, where $\varepsilon = \max\,[|\,\varepsilon_1\,|, |\,\varepsilon_2\,|]$.

Hence, we obtain $\| \delta\mathbf{x} \| \le \sqrt{0.4}\,\varepsilon = \sqrt{0.4} \times 10^{-4}$.

2.32 Solve the system

$$x_1 + 1.001x_2 = 2.001$$
$$x_1 + x_2 = 2.$$

Compute the residual $\mathbf{r} = \mathbf{Ay} - \mathbf{b}$ for $\mathbf{y} = [2 \;\; 0]^T$ and compare the relative size $\| \mathbf{x} - \mathbf{y} \| / \| \mathbf{x} \|$ of the error in the solution with the size $\| \mathbf{r} \| / \| \mathbf{b} \|$ of the residual relative to the right side.

Solution

The exact solution is $\mathbf{x} = [1 \quad 1]^T$. For $\mathbf{y} = [2 \quad 0]^T$, the residual is

$$\mathbf{r} = \mathbf{Ay} - \mathbf{b}$$
$$= \begin{bmatrix} 1 & 1.001 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 2.001 \\ 2 \end{bmatrix} = \begin{bmatrix} -0.001 \\ 0 \end{bmatrix}$$

$$\| \mathbf{r} \| = 0.001, \quad \| \mathbf{x} \| = \sqrt{2}, \quad \| \mathbf{b} \| \approx 2.829,$$

$$\| \mathbf{x} - \mathbf{y} \| = \sqrt{2},$$

$$\frac{\| \mathbf{x} - \mathbf{y} \|}{\| \mathbf{x} \|} = \frac{\sqrt{2}}{\sqrt{2}} = 1. \quad \text{Also,} \quad \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} = \frac{0.001}{2.829} = 0.00035.$$

Eventhough, $\| \mathbf{r} \| / \| \mathbf{b} \|$ is very small, $\mathbf{y}$ is not a solution of the problem.

2.33 Given the system of equations $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \\ 1/4 & 1/5 & 1/6 \end{bmatrix}$$

the vector $\mathbf{b}$ consists of three quantities measured with an error bounded by ε. Derive error bounds for

(*a*) The components of $\mathbf{x}$.

(*b*) The sum of the components $y = x_1 + x_2 + x_3$.

(Royal Inst. Tech., Stockholm, Sweden, BIT 8 (1968), 343)

Solution

(*a*) Let $\hat{x}$ be the computed solution, when the right hand side vector is in error by $\delta\mathbf{b}$. Writing $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$, we have

$$\hat{\mathbf{x}} = \mathbf{A}^{-1}(\mathbf{b} + \delta\mathbf{b}) = \mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}\,\delta\mathbf{b}.$$

Hence, $\delta \mathbf{x} = \mathbf{A}^{-1} \delta \mathbf{b}.$

We find
$$\mathbf{A}^{-1} = 12 \begin{bmatrix} 6 & -20 & 15 \\ -20 & 75 & -60 \\ 15 & -60 & 50 \end{bmatrix}$$

Hence,
$$\delta x_1 = 12\,[6\,\delta b_1 - 20\,\delta b_2 + 15\,\delta b_3],$$
$$\delta x_2 = 12\,[-20\,\delta b_1 + 75\,\delta b_2 - 60\,\delta b_3],$$
$$\delta x_3 = 12\,[15\,\delta b_1 - 60\,\delta b_2 + 50\,\delta b_3],$$
$$|\,\delta x_1\,| \le 12(41\varepsilon) = 492\varepsilon,$$
$$|\,\delta x_2\,| \le 12(155\varepsilon) = 1860\varepsilon,$$
$$|\,\delta x_3\,| \le 12(125\varepsilon) = 1500\varepsilon.$$

(*b*) The error for the sum of the components, $y = x_1 + x_2 + x_3$, is given by
$$\delta y = \delta x_1 + \delta x_2 + \delta x_3 = 12(\delta b_1 - 5\,\delta b_2 + 5\,\delta b_3).$$
Hence, the error bound is obtained as
$$|\,\Delta y\,| \le 12(1 + 5 + 5) \max_i |\,\delta b_i\,| \le 132\varepsilon.$$

EIGENVALUES AND APPLICATIONS

2.34 Show that the matrix
$$\begin{bmatrix} 2 & 4 & & & & \\ 1 & 2 & 4 & & & \mathbf{0} \\ & 1 & 2 & 4 & & \\ & & \ddots & \ddots & \ddots & \\ \mathbf{0} & & & 1 & 2 & 4 \\ & & & & 1 & 2 \end{bmatrix}$$
has real eigenvalues. (Lund Univ., Sweden, BIT 12 (1972), 435)

Solution

We use a diagonal matrix to transform **A** into a symmetric matrix.

Write
$$\mathbf{D} = \begin{pmatrix} d_1 & & & \mathbf{0} \\ & d_2 & & \\ & & \ddots & \\ \mathbf{0} & & & d_n \end{pmatrix}$$
and consider the similarity transformation $\mathbf{B} = \mathbf{D}^{-1}\,\mathbf{AD}$.

$$\mathbf{B} = \mathbf{D}^{-1}\,\mathbf{AD}$$
$$= \begin{pmatrix} 1/d_1 & & & \mathbf{0} \\ & 1/d_2 & & \\ & & \ddots & \\ \mathbf{0} & & & 1/d_n \end{pmatrix} \begin{pmatrix} 2 & 4 & & \mathbf{0} \\ 1 & 2 & 4 & \\ & \ddots & \ddots & \ddots \\ \mathbf{0} & 1 & 2 & 4 \\ & & 1 & 2 \end{pmatrix} \begin{pmatrix} d_1 & & & \mathbf{0} \\ & d_2 & & \\ & & \ddots & \\ \mathbf{0} & & & d_n \end{pmatrix}$$
$$= \begin{pmatrix} 1/d_1 & & & \mathbf{0} \\ & 1/d_2 & & \\ & & \ddots & \\ \mathbf{0} & & & 1/d_n \end{pmatrix} \begin{pmatrix} 2d_1 & 4d_2 & & \\ d_1 & 2d_2 & 4d_3 & \mathbf{0} \\ \ddots & \ddots & \ddots & \\ \mathbf{0} & d_{n-2} & 2d_{n-1} & 4d_n \\ & & d_{n-1} & 2d_n \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 4d_2/d_1 & & & \mathbf{0} \\ d_1/d_2 & 2 & 4d_3/d_2 & & \\ \ddots & \ddots & \ddots & & \\ & d_{n-2}/d_{n-1} & 2 & 4d_n/d_{n-1} & \\ \mathbf{0} & & d_{n-1}/d_n & 2 & \end{pmatrix}$$

The matrix **B** is symmetric if

$$\frac{d_1}{d_2} = 4\frac{d_2}{d_1}, \frac{d_2}{d_3} = 4\frac{d_3}{d_2}, \cdots, \frac{d_{n-1}}{d_n} = 4\frac{d_n}{d_{n-1}}.$$

or $$d_1^2 = 4d_2^2, d_2^2 = 4d_3^2, \ldots, d_{n-1}^2 = 4d_n^2.$$

Without loss of generality, we may take $d_n = 1$. Then, we get

$$d_{n-1} = 2, d_{n-2} = 2^2, \quad d_{n-3} = 2^3, \ldots, d_1 = 2^{n-1}.$$

Therefore, **A** can be reduced to a symmetric form. Since **B** is symmetric, it has real eigenvalues. Hence, **A** has real eigenvalues.

2.35 Compute the spectral radius of the matrix $\mathbf{A}^{-1}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

(Gothenburg Univ., Sweden, BIT 7(1967), 170)

Solution

The given matrix is symmetric. Consider the eigenvalue problem

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}.$$

The three term recurrence relation satisfying this equation is

$$x_{j-1} - \lambda x_j + x_{j+1} = 0$$

with $x_0 = 0$ and $x_7 = 0$. Setting $\lambda = 2 \cos \theta$ and $x_j = \xi^j$, we get

$$1 - 2 (\cos \theta)\xi + \xi^2 = 0$$

whose solution is $\xi = \cos \theta \pm i \sin \theta = e^{\pm i\theta}$. Hence, the solution is

$$x_j = C \cos j\theta + D \sin j\theta.$$

Using the boundary conditions, we get

$$x_0 = 0 = C$$
$$x_7 = 0 = D \sin (7\theta) = \sin(k\pi).$$

We get $$\theta = \frac{k\pi}{7}, k = 1, 2, 3, 4, 5, 6.$$

The eigenvalues of **A** are $2 \cos (\pi/7)$, $2 \cos (2\pi/7)$, $2 \cos (3\pi/7)$, $2 \cos (4\pi/7)$, $2 \cos (5\pi/7)$ and $2 \cos (6\pi/7)$. The smallest eigenvalue in magnitude of **A** is $2 \cos (3\pi/7) = 2 \,|\cos (4\pi/7)|$. Hence,

$$\rho(\mathbf{A}^{-1}) = \frac{1}{2 \cos (3\pi/7)}.$$

2.36 Which of the following matrices have the spectral radius < 1 ?

$(a)\begin{bmatrix} 0 & 1/3 & 1/4 \\ -1/3 & 0 & 1/2 \\ -1/4 & -1/2 & 0 \end{bmatrix}$, $\quad (b)\begin{bmatrix} 1/2 & 1/4 & -1/4 \\ 1/2 & 0 & -1/4 \\ -1/4 & 1/2 & -1/4 \end{bmatrix}$,

$(c)\begin{bmatrix} \cos\alpha & 0 & \sin\alpha \\ 0 & 0.5 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{bmatrix}$, $\alpha = 5\pi/8$ $\quad (d)\begin{bmatrix} 0.5 & -0.25 & 0.75 \\ 0.25 & 0.25 & 0.5 \\ -0.5 & 0.5 & 1.0 \end{bmatrix}$

(Uppsala Univ. Sweden, BIT 12 (1972), 272)

Solution

(*a*) Using the Gerschgorin theorem, we find that

$$|\lambda| \le \max\left[\frac{7}{12}, \frac{5}{6}, \frac{3}{4}\right] = \frac{5}{6}.$$

Hence, $\rho(\mathbf{A}) < 1$.

(*b*) Using the Gerschgorin theorem, we obtain the independent bounds as

(*i*) $|\lambda| \le 1$.

(*ii*) Union of the circles

$$\left|\lambda - \frac{1}{2}\right| \le \frac{1}{2}, \quad |\lambda| \le \frac{3}{4}, \quad \left|\lambda + \frac{1}{4}\right| \le \frac{3}{4}.$$

From (*ii*) we find that there are no complex eigenvalues with magnitude 1. We also find that $|\mathbf{A} - \mathbf{I}| \ne 0$ and $|\mathbf{A} + \mathbf{I}| \ne 0$.

Hence, $\lambda = \pm 1$ are not the eigenvalues. Therefore, $\rho(\mathbf{A}) < 1$.

(*c*) By actual computation, we find that the eigenvalues of the given matrix are 0.5 and $e^{\pm i\alpha}$. Therefore, $\rho(\mathbf{A}) = 1$.

(*d*) By actual computation, we obtain the characteristic equation of the given matrix as

$$16\lambda^3 - 28\lambda^2 + 17\lambda - 5 = 0$$

which has a real root $\lambda = 1$ and a complex pair whose magnitude is less than 1.

Hence, $\rho(\mathbf{A}) = 1$.

2.37 Give a good upper estimate of the eigenvalues of the matrix **A** in the complex number plane. Also, give an upper estimate of the matrix norm of **A**, which corresponds to the Euclidean vector norm

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 1+2i \\ 0 & 2 & 1-i \\ 1-2i & 1+i & 0 \end{bmatrix}$$ (Uppsala Univ. Sweden, BIT 9 (1969), 294)

Solution

Since the given matrix **A** is an Hermitian matrix, its eigenvalues are real. By Gerschgorin theorem, we have

$$|\lambda| \le \max[\sqrt{5} + 1, 2 + \sqrt{2}, \sqrt{5} + \sqrt{2}] = \sqrt{5} + \sqrt{2}.$$

Hence, the eigenvalues lie in the interval $[-(\sqrt{5} + \sqrt{2}), (\sqrt{5} + \sqrt{2})]$, *i.e.* in the interval $(-3.65, 3.65)$.

The Euclidean norm of **A** is

$$\| \mathbf{A} \| = (\Sigma \mid a_{ij} \mid^2)^{1/2} = (1 + 5 + 4 + 2 + 5 + 2)^{1/2} = \sqrt{19}.$$

2.38 (*a*) **A** and **B** are (2 × 2) matrices with spectral radii $\rho(\mathbf{A}) = 0$ and $\rho(\mathbf{B}) = 1$. How big can $\rho(\mathbf{AB})$ be ?

(*b*) Let $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} \beta_1 & 1 \\ 0 & \beta_2 \end{bmatrix}$.

For which β_1, β_2 does $(\mathbf{AB})^k \to 0$ as $k \to \infty$?

(Gothenburg Univ., Sweden, BIT 9 (1969), 294)

Solution

(*a*) Since $\rho(\mathbf{A}) = 0$, it implies that the eigenvalues are 0, 0 and that $| \mathbf{A} | = 0$ and trace $(\mathbf{A}) = 0$. Therefore, **A** must be of the form

$$\mathbf{A} = \begin{pmatrix} a & a \\ -a & -a \end{pmatrix}.$$

Hence, eventhough $\rho(\mathbf{B}) = 1$, it is not possible to bound $\rho(\mathbf{AB})$ as a can take any value.

(*b*)
$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \beta_1 & 1 \\ 0 & \beta_2 \end{bmatrix}.$$

We have
$$\mathbf{AB} = \begin{pmatrix} \beta_1 & 1+\beta_2 \\ \beta_1 & 1+\beta_2 \end{pmatrix}$$

which has eigenvalues 0 and $1 + \beta_1 + \beta_2$.

Now, $\rho(\mathbf{AB}) = | 1 + \beta_1 + \beta_2 |$.

Hence, for $| 1 + \beta_1 + \beta_2 | < 1$, $(\mathbf{AB})^k \to \mathbf{0}$ as $k \to \infty$.

2.39 Calculate $f(\mathbf{A}) = e^{\mathbf{A}} - e^{-\mathbf{A}}$, where **A** is the matrix

$$\begin{bmatrix} 2 & 4 & 0 \\ 6 & 0 & 8 \\ 0 & 3 & -2 \end{bmatrix}$$

(Stockholm Univ., Sweden, BIT 18 (1978), 504)

Solution

The eigenvalues of **A** are $\lambda_1 = 0$, $\lambda_2 = 2\sqrt{13}$, $\lambda_3 = -2\sqrt{13}$.

Let **S** be the matrix having its columns as eigenvectors corresponding to the eigenvalues of **A**. Then, we have

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{D}, \quad \text{and} \quad \mathbf{SDS}^{-1} = \mathbf{A}.$$

where **D** is the diagonal matrix with the eigenvalues of **A** as the diagonal entries.

We have, when m is odd

$$\mathbf{S}^{-1}\mathbf{A}^m\mathbf{S} = \mathbf{D}^m = \begin{pmatrix} 0 & 0 & 0 \\ 0 & (2\sqrt{13})^m & 0 \\ 0 & 0 & (-2\sqrt{13})^m \end{pmatrix}$$

$$= (2\sqrt{13})^{m-1} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2\sqrt{13} & 0 \\ 0 & 0 & -2\sqrt{13} \end{pmatrix} = (2\sqrt{13})^{m-1}\mathbf{D}.$$

Hence,
$$\mathbf{A}^m = (2\sqrt{13})^{m-1}\mathbf{SDS}^{-1} = (2\sqrt{13})^{m-1}\mathbf{A}.$$

Now, $$f(\mathbf{A}) = e^{\mathbf{A}} - e^{-\mathbf{A}} = 2\left[\mathbf{A} + \frac{1}{3!}\mathbf{A}^3 + \frac{1}{5!}\mathbf{A}^5 + \cdots\right]$$

$$= 2\left[\mathbf{A} + \frac{(2\sqrt{13})^2}{3!}\mathbf{A} + \frac{(2\sqrt{13})^4}{5!}\mathbf{A} + \cdots\right]$$

$$= 2\left[1 + \frac{(2\sqrt{13})^2}{3!} + \frac{(2\sqrt{13})^4}{5!} + \cdots\right]\mathbf{A}$$

$$= \frac{2}{2\sqrt{13}}\left[2\sqrt{13} + \frac{(2\sqrt{13})^3}{3!} + \frac{(2\sqrt{13})^5}{5!} + \cdots\right]\mathbf{A}$$

$$= \frac{1}{\sqrt{13}} \sinh(2\sqrt{13})\,\mathbf{A}.$$

2.40 The matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 3 \\ 6 & -13 & 18 \\ 4 & -10 & 14 \end{bmatrix}$$

is transformed to diagonal form by the matrix

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 4 \\ 2 & 2 & 3 \end{bmatrix}, \quad \textit{i.e.} \quad \mathbf{T}^{-1}\mathbf{AT}.$$

Calculate the eigenvalues and the corresponding eigenvectors of **A**.

(Uppsala Univ. Sweden, BIT 9 (1969), 174)

Solution

We have the equation

$$\mathbf{T}^{-1}\mathbf{AT} = \mathbf{D}$$

or **AT = TD** where $\mathbf{D} = \text{diag}\,[\lambda_1\ \lambda_2\ \lambda_3]$, and $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of **A**.

We have, $$\begin{bmatrix} 1 & -2 & 3 \\ 6 & -13 & 18 \\ 4 & -10 & 14 \end{bmatrix}\begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 4 \\ 2 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 4 \\ 2 & 2 & 3 \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Comparing the corresponding elements on both sides, we obtain $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = 2$. Since **T** transforms **A** to the diagonal form, **T** is the matrix of the corresponding eigenvectors. Hence, the eigenvalues are 1, – 1, 2 and the corresponding eigenvectors are $[1\ 3\ 2]^T$, $[0\ 3\ 2]^T$, and $[1\ 4\ 3]^T$ respectively.

2.41 Show that the eigenvalues of the tridiagonal matrix

$$\mathbf{A} = \begin{bmatrix} a & b_1 & & & \mathbf{0} \\ c_1 & a & b_2 & & \\ & c_2 & a & b_3 & \\ & \ddots & \ddots & \ddots & \\ \mathbf{0} & & c_{n-1} & a & \end{bmatrix}$$

satisfy the inequality

$$|\lambda - a| < 2\sqrt{\left(\max_j |b_j|\right)\left(\max_j |c_j|\right)}$$

(Uppsala Univ. Sweden, BIT 8 (1968), 246)

Solution

The ith equation of the eigenvalue system is

$$c_{j-1}\, x_{j-1} - (\lambda - a)\, x_j + b_j x_{j+1} = 0$$

with $x_0 = 0$ and $x_{n+1} = 0$.

Setting $\lambda - a = r \cos \theta$ and $x_j = \xi^j$, we get

$$b_j \xi^2 - (r \cos \theta)\xi + c_{j-1} = 0$$

whose solution is

$$\xi = [r \cos \theta \pm \sqrt{r^2 \cos^2 \theta - 4b_j\, c_{j-1}}\,]\big/(2b_j),$$

Requiring ξ to be complex, we get

$$r^2 \cos^2 \theta < 4b_j c_{j-1}.$$

The solution is then given by

$$\xi = p \pm iq$$

where $$p = \frac{r \cos \theta}{2b_j}, \quad q = \frac{\sqrt{4b_j c_{j-1} - r^2 \cos^2 \theta}}{2b_j}$$

and $$x_j = p^j\, [A \cos q\, j + B \sin q\, j].$$

Substituting the conditions, we get

$$x_0 = 0 = A$$
$$x_{n+1} = 0 = p^{n+1} B \sin [(n+1)q] = \sin k\pi.$$

Hence, $$q = \frac{k\pi}{n+1}, \; k = 1, 2, ..., n.$$

The required bounds are given by

$$|\, \lambda - a \,|^2 = |\, r \cos \theta \,|^2 < 4\, |\, b_j c_{j-1} \,|.$$

Hence, $$|\, \lambda - a \,| < 2\sqrt{\max_j |\, b_j \,| \max_j |\, c_j \,|}\;.$$

2.42 Let $P_n(\lambda) = \det (\mathbf{A}_n - \lambda \mathbf{I})$, where

$$\mathbf{A}_n = \begin{bmatrix} a & 0 & \cdots & 0 & a_n \\ 0 & a & \cdots & 0 & a_{n-1} \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & a & a_2 \\ a_n & a_{n-1} & \cdots & a_2 & a_1 \end{bmatrix}$$

Prove the recurrence relation

$$P_n(\lambda) = (a - \lambda)\, P_{n-1}(\lambda) - a_n^2 (a - \lambda)^{n-2},$$
$$P_1(\lambda) = a_1 - \lambda$$

and determine all eigenvalues of $\mathbf{A}_n$.

(Royal Inst. Tech., Stockholm, Sweden, BIT 8 (1968), 243)

Solution

We have $$P_n(\lambda) = \begin{vmatrix} a-\lambda & 0 & \cdots & 0 & a_n \\ 0 & a-\lambda & \cdots & 0 & a_{n-1} \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & a-\lambda & a_2 \\ a_n & a_{n-1} & \cdots & a_2 & a_1 - \lambda \end{vmatrix} = 0, \quad n \geq 0.$$

Expanding the determinant by the first column, we get

$$P_n(\lambda) = (a - \lambda) \begin{vmatrix} a-\lambda & 0 & \cdots & 0 & a_{n-1} \\ 0 & a-\lambda & \cdots & 0 & a_{n-2} \\ \vdots & \vdots & & \vdots & \\ a_{n-1} & a_{n-2} & \cdots & a_2 & a_1 - \lambda \end{vmatrix}$$

$$+ (-1)^{n-1} a_n \begin{vmatrix} 0 & 0 & \cdots & 0 & a_n \\ a-\lambda & 0 & \cdots & 0 & a_{n-1} \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & a-\lambda & a_2 \end{vmatrix}$$

Expanding the second determinant by the first row, we get

$$P_n(\lambda) = (a-\lambda)P_{n-1}(\lambda) + (-1)^{2n-3} a_n^2 \begin{vmatrix} q & 0 & \dots & 0 \\ 0 & q & \dots & 0 \\ \dots & \dots & \dots & \\ 0 & 0 & \dots & q \end{vmatrix}$$

$$= (a-\lambda)\, P_{n-1}(\lambda) + (-1)^{2n-3} a_n^2 (a-\lambda)^{n-2}$$

$$n = 2, 3, \ldots,$$

where $\quad q = a - \lambda \quad$ and $\quad P_1(\lambda) = a_1 - \lambda.$

We have
$$P_n = (a-\lambda)\,[P_{n-1} - a_n^2 (a-\lambda)^{n-3}]$$
$$= (a-\lambda)^2\,[P_{n-2} - (a_n^2 + a_{n-1}^2)(a-\lambda)^{n-4}]$$
$$\vdots$$
$$= (a-\lambda)^{n-2}\,[P_2 - (a_n^2 + a_{n-1}^2 + \ldots + a_3^2)]$$
$$= (a-\lambda)^{n-2}\,[\lambda^2 - \lambda(a + a_1) + aa_1 - (a_n^2 + \ldots + a_2^2)]$$

Hence, the eigenvalues are

$$\lambda_i = a, \; i = 1, 2, \ldots, n-2,$$

and
$$\lambda = \frac{1}{2}\left[(a + a_1) \pm \sqrt{(a_1 - a)^2 + 4(a_n^2 + a_{n-1}^2 + \ldots + a_2^2)}\right]$$

2.43 Let
$$\mathbf{A} = \begin{bmatrix} -2 & -1 & 2 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} -1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \end{bmatrix}$$

$\lambda_i(\varepsilon)$ are the eigenvalues of $\mathbf{A} + \varepsilon\,\mathbf{B}$, $\varepsilon \ge 0$.

Estimate $|\, \lambda_i(\varepsilon) - \lambda_i(0)\,|$, $i = 1, 2, 3$. (Gothenburg Univ., Sweden, BIT 9 (1969), 174)

Solution

The eigenvalues of **A** are 1, – 1, 0 and the matrix of eigenvectors is

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 1/2 \\ 1 & -1 & -1 \\ 1/2 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{S}^{-1} = \begin{bmatrix} 0 & 0 & 2 \\ 2 & 1 & -2 \\ -2 & -2 & 4 \end{bmatrix}.$$

We have $\mathbf{S}^{-1}(\mathbf{A} + \varepsilon\mathbf{B})\mathbf{S} = \mathbf{S}^{-1}\mathbf{AS} + \varepsilon\mathbf{S}^{-1}\mathbf{BS} = \mathbf{D} + \varepsilon\mathbf{P}$

where **D** is a diagonal matrix with 1, – 1, 0 on the diagonal and

$$\mathbf{P} = \mathbf{S}^{-1}\mathbf{BS} = \begin{bmatrix} 1 & -4 & -3 \\ -1/2 & 2 & 3/2 \\ 2 & -8 & -6 \end{bmatrix}$$

The eigenvalues of $A + \varepsilon \mathbf{B}$, ($\varepsilon << 1$) lie in the union of the disks

$$| \lambda(\varepsilon) - \lambda_i(0) | \le \varepsilon \text{ cond }_\infty (\mathbf{S}) \| \mathbf{P} \|_\infty$$

Since, $\text{cond}_\infty (\mathbf{S}) = \| \mathbf{S} \| \| \mathbf{S}^{-1} \| = 24$ and $\| \mathbf{P} \|_\infty = 16$, we have the union of the disks as

$$| \lambda(\varepsilon) - \lambda_i (0) | \le 384\varepsilon$$

where $\lambda_1(0) = 1$, $\lambda_2(0) = -1$ and $\lambda_3(0) = 0$.

A more precise result is obtained using the Gerschgorin theorem. We have the union of disks as

$$| \lambda(\varepsilon) - \lambda_i(0) - \varepsilon p_{ii} | \le \varepsilon \sum_{\substack{j=1 \\ i \ne j}}^{n} | p_{ij} |, \quad \text{or} \quad | \lambda(\varepsilon) - 1 - \varepsilon | \le 7\varepsilon,$$

$$| \lambda(\varepsilon) + 1 - 2\varepsilon | \le 2\varepsilon, \quad \text{and} \quad | \lambda(\varepsilon) + 6\varepsilon | \le 10\varepsilon.$$

The eigenvalues of $\mathbf{A}$ are real and $\varepsilon \mathbf{B}$ represents a perturbation. Hence, we assume that the eigenvalues of $\mathbf{A} + \varepsilon \mathbf{B}$ are also real. We now have the bounds for the eigenvalues as

$$-6\varepsilon \le \lambda_1(\varepsilon) - \lambda_1(0) \le 8\varepsilon,$$

$$0 \le \lambda_2(\varepsilon) - \lambda_2(0) \le 4\varepsilon,$$

and

$$-16\varepsilon \le \lambda_3(\varepsilon) - \lambda_3(0) \le 4\varepsilon.$$

Alternately, we have that the eigenvalues lie in the interval

$$-16\varepsilon \le \lambda(\varepsilon) - \lambda_i(0) \le 8\varepsilon.$$

2.44 Using the Gerschgorin's theorem, find bounds for the eigenvalues λ of the real $n \times n$ matrix $\mathbf{A}$ ($n \ge 3$)

$$\mathbf{A} = \begin{bmatrix} a & -1 & & & \mathbf{0} \\ -1 & a & -1 & & \\ & -1 & a & -1 & \\ & & \ddots & & \\ \mathbf{0} & & & -1 & a \end{bmatrix}$$

Show that the components x_i of the eigenvector $\mathbf{x}$ obey a linear difference equation, and find all the eigenvalues and eigenvectors. (Bergen Univ., Norway, BIT 5 (1965), 214)

Solution

By the Gerschgorin's theorem, the bound of the eigenvalues is given by

(*i*) $| \lambda - a | \le 2$, and (*ii*) $| \lambda | \le | a | + 2$.

The ith equation of the eigenvalue system is

$$-x_{j-1} + (a - \lambda)x_j - x_{j+1} = 0 \text{ with } x_0 = 0 \text{ and } x_{n+1} = 0.$$

Setting $a - \lambda = 2 \cos \theta$ and $x_j = \xi^j$, we get

$$\xi^2 - (2 \cos \theta)\xi + 1 = 0$$

$$\xi = \frac{2 \cos \theta \pm \sqrt{4 \cos^2 \theta - 4}}{2} = \cos \theta \pm i \sin \theta = e^{\pm i\theta}.$$

The solution of the difference equation is

$$x_j = A e^{(i\theta)j} + B e^{(-i\theta)j} = C \cos j\theta + D \sin j\theta.$$

Using the boundary conditions, we have

$$x_0 = 0 = C$$

and

$$x_{n+1} = 0 = \sin k\pi = D \sin [(n + 1)\theta]$$

Therefore, $\theta = \dfrac{k\pi}{n+1}, \quad k = 1, 2, ..., n.$

Hence, the eigenvalues are given by

$$\lambda_k = a - 2\cos\theta = a - 2\cos\left(\frac{k\pi}{n+1}\right), \; k = 1, 2, ..., n$$

and the eigenvectors are $x_{jk} = \sin\left(\dfrac{jk\pi}{n+1}\right), \quad j, k = 1, 2, ..., n.$

2.45 Use Gerschgorin's theorem to estimate

$$|\lambda_i - \bar{\lambda}_i|, \quad i = 1, 2, 3$$

where λ_i are eigenvalues of

$$\mathbf{A} = \begin{bmatrix} 2 & 3/2 & 0 \\ 1/2 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

and $\bar{\lambda}_i$ are eigenvalues of

$$\tilde{A} = A + 10^{-2}\begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$$

(Lund Univ., Sweden, BIT 11 (1971), 225)

Solution

The eigenvalues of **A** are 1/2, 5/2 and – 1. The corresponding eigenvectors are found to be

$$[1\; -1\;\; 0]^T, \quad [3\;\; 1\;\; 0]^T, \text{ and } [0\;\; 0\;\; 1]^T.$$

Hence, the matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 3 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

reduces **A** to its diagonal form.

We have

$$\mathbf{S}^{-1}\tilde{\mathbf{A}}\mathbf{S} = \mathbf{S}^{-1}(\mathbf{A} + 10^{-2}\mathbf{B})\,\mathbf{S} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} + 10^{-2}\mathbf{S}^{-1}\mathbf{B}\mathbf{S}$$

where $\mathbf{B} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$

We also have $\mathbf{S}^{-1} = \dfrac{1}{4}\begin{bmatrix} 1 & -3 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad \mathbf{S}^{-1}\mathbf{B}\mathbf{S} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & 0 \\ 2 & 2 & 1 \end{bmatrix}$

Therefore, $\mathbf{S}^{-1}\tilde{\mathbf{A}}\mathbf{S} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 5/2 & 0 \\ 0 & 0 & -1 \end{bmatrix} + 10^{-2}\begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & 0 \\ 2 & 2 & 1 \end{bmatrix}$

By Gerschgorin theorem, we obtain that the eigenvalues of $\tilde{\mathbf{A}}$ lies in the union of the circles

$$\left|\bar{\lambda}-\left(\frac{1}{2}+2\times 10^{-2}\right)\right| \le 3 \times 10^{-2},$$

$$\left|\bar{\lambda}-\frac{5}{2}\right| = 0,$$

$$|\,\bar{\lambda} - (-1 + 10^{-2})\,| \le 4 \times 10^{-2}$$

which are disjoint bounds.

Hence, we have

$$\lambda_1 = \frac{1}{2},\ |\,\bar{\lambda}_i - \lambda_1\,| \le 5 \times 10^{-2},$$

$$\lambda_2 = \frac{5}{2},\ |\,\bar{\lambda}_2 - \lambda_2\,| = 0,$$

$$\lambda_3 = -1,\ |\,\bar{\lambda}_3 - \lambda_3\,| \le 5 \times 10^{-2}.$$

ITERATIVE METHODS

2.46 Given

$$\mathbf{A} = \begin{bmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{bmatrix}$$

For which values of α does the vector sequence $\{\,\mathbf{y}_n\,\}_0^\infty$ defined by

$$\mathbf{y}_n = (\mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2)\mathbf{y}_{n+1}, \quad n = 1, 2, \ldots$$

$\mathbf{y}_0$ arbitrary, converges to $\mathbf{0}$ as $n \to \infty$? (Uppsala Univ. Sweden, BIT 14 (1974), 366)

Solution

From the given equation $\mathbf{y}_n = (\mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2)\,\mathbf{y}_{n-1}$

we get $\mathbf{y}_n = (\mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2)^n\,\mathbf{y}_0$

where $\mathbf{y}_0$ is arbitrary.

Hence, $\lim\limits_{n\to\infty} \mathbf{y}_n \to \mathbf{0}$ if and only if $\rho(\mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2) < 1$.

The eigenvalues of

$$\mathbf{A} = \begin{bmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{bmatrix}$$

are 1 and 2. Hence, the eigenvalues of $\mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2$ are $1 + \alpha + \alpha^2$ and $1 + 2\alpha + 4\alpha^2$. We require that

$$|\,1 + \alpha + \alpha^2\,| < 1, \quad \text{and} \quad |\,1 + 2\alpha + 4\alpha^2\,| < 1.$$

The first inequality gives

$$-1 < 1 + \alpha + \alpha^2 < 1, \quad \text{or} \quad -2 < \alpha\,(1 + \alpha) < 0.$$

This gives, $\alpha < 0,\ \alpha + 1 > 0, \quad \text{or} \quad \alpha \in (-1, 0).$

The second inequality gives

$$-1 < 1 + 2\alpha + 4\alpha^2 < 1, \quad \text{or} \quad -2 < 2\alpha\,(1 + 2\alpha) < 0.$$

This given, $\alpha < 0$, $1 + 2\alpha > 0$, or $\alpha \in (-1/2, 0)$.

Hence, the required interval is $(-1/2, 0)$.

2.47 The system $\mathbf{Ax} = \mathbf{b}$ is to be solved, where $\mathbf{A}$ is the fourth order Hilbert matrix, the elements of which are $a_{ij} = 1/(i + j)$ and $\mathbf{b}^T = (1 \quad 1 \quad 1 \quad 1)$. Since $\mathbf{A}$ is ill-conditioned, the matrix $\mathbf{B}$, a close approximation to (the unknown) $\mathbf{A}^{-1}$, is used to get an approximate solution $\mathbf{x}_0 = \mathbf{Bb}$

$$\mathbf{B} = \begin{bmatrix} 202 & -1212 & 2121 & -1131 \\ -1212 & 8181 & -15271 & 8484 \\ 2121 & -15271 & 29694 & -16968 \\ -1131 & 8484 & -16968 & 9898 \end{bmatrix}$$

It is known that the given system has an integer solution, however $\mathbf{x}_0$ is not the correct one. Use iterative improvement (with $\mathbf{B}$ replacing $\mathbf{A}^{-1}$) to find the correct integer solution.

(Royal Inst. Tech., Stockholm, Sweden, BIT 26 (1986), 540)

Solution

Let $\overline{\mathbf{x}}$ be a computed solution of $\mathbf{Ax} = \mathbf{b}$ and let $\mathbf{r} = \mathbf{b} - \mathbf{A}\overline{\mathbf{x}}$ be the residual. Then,

$$\mathbf{A}(\mathbf{x} - \overline{\mathbf{x}}) = \mathbf{Ax} - \mathbf{A}\overline{\mathbf{x}} = \mathbf{b} - \mathbf{A}\overline{\mathbf{x}} = \mathbf{r}, \quad \text{or} \quad \mathbf{A}\delta\mathbf{x} = \mathbf{r}.$$

Inverting $\mathbf{A}$, we have

$$\delta\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{r} \approx \mathbf{Br}.$$

The next approximation to the solution is then given by $\mathbf{x} = \overline{\mathbf{x}} + \delta\mathbf{x}$.

We have in the present problem

$$\mathbf{x}_0 = \mathbf{Bb} = [-20 \;\; 182 \; -424 \;\; 283]^T,$$
$$\mathbf{r} = \mathbf{b} - \mathbf{Ax}_0 = [-0.2667 \;\; -0.2 \;\; -0.1619 \;\; -0.1369]^T,$$
$$\delta\mathbf{x} = \mathbf{Br} = [-0.0294 \;\; -2.0443 \;\; 3.9899 \;\; -3.0793]^{\mathrm{T}},$$
$$\mathbf{x} = \mathbf{x}_0 + \delta\mathbf{x} = [-20.0294 \;\;\; 179.9557 \;\; -420.0101 \;\; 279.9207]^T$$
$$\approx [-20 \;\; 180 \; -420 \;\; 280]^T$$

since an integer solution is required. It can be verified that this is the exact solution.

2.48 The system of equations $\mathbf{Ax} = \mathbf{y}$, where

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

can be solved by the following iteration

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha(\mathbf{Ax}^{(n)} - \mathbf{y}),$$
$$\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

How should the parameter α be chosen to produce optimal convergence ?

(Uppsala Univ. Sweden, BIT 10 (1970), 228)

Solution

The given iteration scheme is

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha(\mathbf{Ax}^{(n)} - \mathbf{y}) = (\mathbf{I} + \alpha\mathbf{A})\,\mathbf{x}^{(n)} - \alpha\mathbf{y}.$$

Setting $n = 0, 1, 2, \ldots$, we obtain

$$\mathbf{x}^{(n+1)} = \mathbf{q}^{n+1}\,\mathbf{x}^{(0)} - \alpha\,[\mathbf{I} + \mathbf{q} + \ldots + \mathbf{q}^n]\,\mathbf{y}$$

where $\qquad \mathbf{q} = \mathbf{I} + \alpha\mathbf{A}.$

The iteration scheme will converge if and only if $\rho(\mathbf{I} + \alpha\mathbf{A}) < 1$.

The eigenvalues of $[\mathbf{I} + \alpha\mathbf{A}]$ are $\lambda_1 = 1 + \alpha$ and $\lambda_2 = 1 + 4\alpha$.

We choose α such that

$$|\,1 + \alpha\,| = |\,1 + 4\alpha\,|$$

which gives $\qquad \alpha = -\,0.4.$

2.49 (*a*) Let $\mathbf{A} = \mathbf{B} - \mathbf{C}$ where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are nonsingular matrices and set

$$\mathbf{Bx}^{(m)} = \mathbf{Cx}^{(m-1)} + \mathbf{y},\; m = 1, 2, \ldots$$

Give a necessary and sufficient condition so that

$$\lim_{m \to \infty} \mathbf{x}^{(m)} = \mathbf{A}^{-1}\mathbf{y}$$

for every choice of $\mathbf{x}^{(0)}$.

(*b*) Let $\mathbf{A}$ be an $n \times n$ matrix with real positive elements a_{ij} fulfilling the condition

$$\sum_{j=1}^{n} a_{ij} = 1, \quad i = 1, 2, \ldots n$$

Show that $\lambda = 1$ is an eigenvalue of the matrix $\mathbf{A}$, and give the corresponding eigenvector. Then, show that the spectral radius $\rho(\mathbf{A}) \le 1$. (Lund Univ., Sweden, BIT 9 (1969), 174)

Solution

(*a*) We write the given iteration scheme in the form

$$\begin{aligned}\mathbf{x}^{(m)} &= \mathbf{B}^{-1}\,\mathbf{C}\mathbf{x}^{(m-1)} + \mathbf{B}^{-1}\mathbf{y}\\ &= (\mathbf{B}^{-1}\mathbf{C})^{(m)}\,\mathbf{x}^{(0)} + [\mathbf{I} + \mathbf{B}^{-1}\mathbf{C} + (\mathbf{B}^{-1}\mathbf{C})^2 + \ldots + (\mathbf{B}^{-1}\mathbf{C})^{m-1}]\,\mathbf{B}^{-1}\mathbf{y}\end{aligned}$$

If $\| \mathbf{B}^{-1}\mathbf{C} \| < 1$, or $\rho(\mathbf{B}^{-1}\mathbf{C}) < 1$, we have

$$\begin{aligned}\lim_{m \to \infty} \mathbf{x}^{(m)} &= (\mathbf{I} - \mathbf{B}^{-1}\mathbf{C})^{-1}\,\mathbf{B}^{-1}\mathbf{y} = [\mathbf{B}^{-1}\,(\mathbf{B} - \mathbf{C})]^{-1}\,\mathbf{B}^{-1}\mathbf{y}\\ &= (\mathbf{B}^{-1}\mathbf{A})^{-1}\,\mathbf{B}^{-1}\mathbf{y} = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^{-1}\mathbf{y} = \mathbf{A}^{-1}\mathbf{y}.\end{aligned}$$

Hence, $\rho(\mathbf{B}^{-1}\mathbf{C}) < 1$ is the necessary and sufficient condition. $\| \mathbf{B}^{-1}\mathbf{C} \| < 1$ is a sufficient condition.

(*b*) Let the $n \times n$ matrix $\mathbf{A} = (a_{ij})$, $a_{ij} > 0$, with $\sum_{j=1}^{n} a_{ij} = 1$, $i = 1, 2, \ldots, n$ be

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

We have $$| \mathbf{A} - \lambda\mathbf{I} | = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Adding to the first column, all the remaining columns and using $\sum_{j=1}^{n} a_{ij} = 1$, we obtain

$$\begin{aligned}| \mathbf{A} - \lambda\mathbf{I} | &= \begin{vmatrix} 1-\lambda & a_{12} & \cdots & a_{1n} \\ 1-\lambda & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ 1-\lambda & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0\\ &= (1 - \lambda) \begin{vmatrix} 1 & a_{12} & \cdots & a_{1n} \\ 1 & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ 1 & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0\end{aligned}$$

which shows that $\lambda = 1$ is an eigenvalue of $\mathbf{A}$.

Since $\lambda = 1$ is an eigenvalue and $\sum_{j=1}^{n} a_{ij} = 1$, $i = 1, 2, ..., n$, it is obvious that the corresponding eigenvector is $[1\ 1\ ...\ 1]^T$. Using the Gerschgorin theorem, we have

$$|\lambda| \le \max_i \left[\sum_{j=1}^{n} a_{ij} \right] \le 1$$

Hence, $\rho(A) \le 1$.

2.50 Show that if $\mathbf{A}$ is strictly diagonally dominant in $\mathbf{Ax} = \mathbf{b}$, then the Jacobi iteration always converges.

Solution

The Jacobi scheme is

$$\begin{aligned}\mathbf{x}^{(k+1)} &= -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}\\ &= -\mathbf{D}^{-1}(\mathbf{A} - \mathbf{D})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{A})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}.\end{aligned}$$

The scheme converges if $\| \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} \| < 1$. Using absolute row sum criterion, we have

$$\frac{1}{|a_{ii}|} \sum_{j=1, i \ne j}^{n} |a_{ij}| < 1, \quad \text{for all } i$$

or $$|a_{ii}| > \sum_{j=1, i \ne j}^{n} |a_{ij}|, \quad \text{for all } i.$$

This proves that if $\mathbf{A}$ is strictly diagonally dominant, then the Jacobi iteration converges.

2.51 Show if $\mathbf{A}$ is a strictly diagonally dominant matrix, then the Gauss-Seidel iteration scheme converges for any initial starting vector.

Solution

The Gauss-Seidel iteration scheme is given by

$$\begin{aligned}\mathbf{x}^{(k+1)} &= -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b}\\ &= -(\mathbf{D} + \mathbf{L})^{-1}[\mathbf{A} - (\mathbf{D} + \mathbf{L})]\mathbf{x}^{(k)} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b}\\ &= [\mathbf{I} - (\mathbf{D} + \mathbf{L})^{-1}\mathbf{A}]\,\mathbf{x}^{(k)} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b}.\end{aligned}$$

Therefore, the iteration scheme will be convergent if

$$\rho[\mathbf{I} - (\mathbf{D} + \mathbf{L})^{-1}\mathbf{A}] < 1.$$

Let λ be an eigenvalue of $\mathbf{I} - (\mathbf{D} + \mathbf{L})^{-1}\mathbf{A}$. Therefore,

$$(\mathbf{I} - (\mathbf{D} + \mathbf{L})^{-1}\mathbf{A})\,\mathbf{x} = \lambda\,\mathbf{x} \quad \text{or} \quad (\mathbf{D} + \mathbf{L})\,\mathbf{x} - \mathbf{A}\,\mathbf{x} = \lambda\,(\mathbf{D} + \mathbf{L})\,\mathbf{x}$$

or $$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda \sum_{j=1}^{i} a_{ij}x_j, \; 1 \le i \le n$$

or $$\lambda a_{ii}\,x_i = -\sum_{j=i+1}^{n} a_{ij}x_j - \lambda \sum_{j=1}^{i-1} a_{ij}x_j$$

or $$|\lambda a_{ii}\,x_i| \le \sum_{j=i+1}^{n} |a_{ij}|\,|x_j| + |\lambda| \sum_{j=1}^{i-1} |a_{ij}|\;|x_j|.$$

Since $\mathbf{x}$ is an eigenvector, $\mathbf{x} \ne \mathbf{0}$. Without loss of generality, we assume that $\| \mathbf{x} \|_\infty = 1$. Choose an index i such that

$$|x_i| = 1 \quad \text{and} \quad |x_j| \le 1 \quad \text{for all} \quad j \ne i.$$

Hence, $$|\lambda||a_{ii}| \le \sum_{j=i+1}^{n} |a_{ij}| + |\lambda| \sum_{j=1}^{i-1} |a_{ij}|$$

or $$|\lambda| \left[|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \right] \le \sum_{j=i+1}^{n} |a_{ij}|$$

Therefore, $$|\lambda| \le \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1$$

which is true, since **A** is strictly diagonally dominant.

2.52 Solve the system of equations

$$4x_1 + 2x_2 + x_3 = 4$$
$$x_1 + 3x_2 + x_3 = 4$$
$$3x_1 + 2x_2 + 6x_3 = 7$$

Using the Gauss-Jacobi method, directly and in error format. Perform three iterations using the initial approximation, $\mathbf{x}^{(0)} = [0.1 \quad 0.8 \quad 0.5]^T$.

Solution

Gauss-Jacobi method in error format is given by

$$\mathbf{Dv}^{(k)} = \mathbf{r}^{(k)}, \text{ where } \quad \mathbf{v}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(k)},$$

and **D** is the diagonal part of **A.**

We have the following approximations.

$$\mathbf{r}^{(0)} = \begin{pmatrix} 4 \\ 4 \\ 7 \end{pmatrix} - \begin{pmatrix} 4 & 2 & 1 \\ 1 & 3 & 1 \\ 3 & 2 & 6 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.8 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 1.5 \\ 1.0 \\ 2.1 \end{pmatrix}; \quad \mathbf{v}^{(0)} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/6 \end{pmatrix} \begin{pmatrix} 1.5 \\ 1.0 \\ 2.1 \end{pmatrix} = \begin{pmatrix} 0.375 \\ 0.3333 \\ 0.350 \end{pmatrix},$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{v}^{(0)} = \begin{pmatrix} 0.1 \\ 0.8 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0.375 \\ 0.3333 \\ 0.350 \end{pmatrix} = \begin{pmatrix} 0.4750 \\ 1.1333 \\ 0.850 \end{pmatrix}; \mathbf{r}^{(1)} = \begin{pmatrix} -1.0166 \\ -0.7249 \\ -1.7916 \end{pmatrix}; \mathbf{v}^{(1)} = \begin{pmatrix} -0.25415 \\ -0.2416 \\ -0.2986 \end{pmatrix},$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{v}^{(1)} = \begin{pmatrix} 0.2209 \\ 0.8917 \\ 0.5514 \end{pmatrix}; \mathbf{r}^{(2)} = \begin{pmatrix} 0.7816 \\ 0.5526 \\ 1.2455 \end{pmatrix}; \mathbf{v}^{(2)} = \begin{pmatrix} 0.1954 \\ 0.1842 \\ 0.2075 \end{pmatrix},$$

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + \mathbf{v}^{(2)} = \begin{pmatrix} 0.4163 \\ 1.0759 \\ 0.7590 \end{pmatrix}.$$

Direct method We write

$$x_1^{(k+1)} = \frac{1}{4} [4 - 2x_2^{(k)} - x_3^{(k)}], \quad x_2^{(k+1)} = \frac{1}{3} [4 - x_1^{(k)} - x_3^{(k)}],$$

$$x_3^{(k+1)} = \frac{1}{6} [7 - 3x_1^{(k)} - 2x_2^{(k)}].$$

Using $\mathbf{x}^{(0)} = [0.1 \quad 0.8 \quad 0.5]^T$, we obtain the following approximations.

$$x_1^{(1)} = 0.475, \quad x_2^{(1)} = 1.1333, \quad x_3^{(1)} = 0.85,$$
$$x_1^{(2)} = 0.2209, \quad x_2^{(2)} = 0.8917, \quad x_3^{(2)} = 0.5514,$$
$$x_1^{(3)} = 0.4163, \quad x_2^{(3)} = 1.0759, \quad x_3^{(3)} = 0.7590.$$

2.53 Solve the system of equations

$$4x_1 + 2x_2 + x_3 = 4$$
$$x_1 + 3x_2 + x_3 = 4$$
$$3x_1 + 2x_2 + 6x_3 = 7$$

using the Gauss-Seidel method, directly and in error format. Perform three iterations using the initial approximation, $\mathbf{x}^{(0)} = [0.1 \quad 0.8 \quad 0.5]^T$.

Solution

Gauss-Seidel method, in error format, is given by

$$(\mathbf{D} + \mathbf{L})\, \mathbf{v}^{(k)} = \mathbf{r}^{(k)}, \quad \text{where } \mathbf{v}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)},\ \mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\, \mathbf{x}^{(k)}.$$

We have the following approximations.

$$\mathbf{r}^{(0)} = \begin{pmatrix} 1.5 \\ 1.0 \\ 2.1 \end{pmatrix}; \quad \begin{bmatrix} 4 & 0 & 0 \\ 1 & 3 & 0 \\ 3 & 2 & 6 \end{bmatrix} \mathbf{v}^{(0)} = \begin{pmatrix} 1.5 \\ 1.0 \\ 2.1 \end{pmatrix}, \quad \mathbf{v}^{(0)} = \begin{pmatrix} 0.375 \\ 0.2083 \\ 0.0931 \end{pmatrix}$$

(By forward substitution),

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{v}^{(0)} = \begin{pmatrix} 0.1 \\ 0.8 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0.375 \\ 0.2083 \\ 0.0931 \end{pmatrix} = \begin{pmatrix} 0.475 \\ 1.0083 \\ 0.5931 \end{pmatrix}; \quad \mathbf{r}^{(1)} = \begin{pmatrix} -0.5097 \\ -0.093 \\ -0.0002 \end{pmatrix};$$

$$\begin{bmatrix} 4 & 0 & 0 \\ 1 & 3 & 0 \\ 3 & 2 & 6 \end{bmatrix} \mathbf{v}^{(1)} = \begin{pmatrix} -0.5097 \\ -0.093 \\ -0.0002 \end{pmatrix}, \quad \mathbf{v}^{(1)} = \begin{pmatrix} -0.1274 \\ 0.0115 \\ 0.0598 \end{pmatrix};$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{v}^{(1)} = \begin{pmatrix} 0.3476 \\ 1.0198 \\ 0.6529 \end{pmatrix}, \mathbf{r}^{(2)} = \begin{pmatrix} -0.0829 \\ -0.0599 \\ 0.0002 \end{pmatrix};$$

$$\begin{bmatrix} 4 & 0 & 0 \\ 1 & 3 & 0 \\ 3 & 2 & 6 \end{bmatrix} \mathbf{v}^{(2)} = \begin{pmatrix} -0.0829 \\ -0.0599 \\ 0.0002 \end{pmatrix}, \mathbf{v}^{(2)} = \begin{pmatrix} -0.0207 \\ -0.0131 \\ 0.0148 \end{pmatrix}; \mathbf{x}^{(3)} = \mathbf{x}^{(2)} + \mathbf{v}^{(2)} = \begin{pmatrix} 0.3269 \\ 1.0067 \\ 0.6677 \end{pmatrix}.$$

Direct method We write

$$x_1^{(k+1)} = \frac{1}{4}[4 - 2x_2^{(k)} - x_3^{(k)}], \; x_2^{(k+1)} = \frac{1}{3}\,[4 - x_1^{(k+1)} - x_3^{(k)}]$$
$$x_3^{(k+1)} = \frac{1}{6}[7 - 3x_1^{(k+1)} - 2x_2^{(k+1)}].$$

Using $\mathbf{x}^{(0)} = [0.1 \quad 0.8 \quad 0.5]^T$, we obtain the following approximations.

$$x_1^{(1)} = 0.475, \quad x_2^{(1)} = 1.0083, \quad x_3^{(1)} = 0.5931,$$
$$x_1^{(2)} = 0.3476, \quad x_2^{(2)} = 1.0198, \quad x_3^{(2)} = 0.6529,$$
$$x_1^{(3)} = 0.3269, \quad x_2^{(3)} = 1.0067, \quad x_3^{(3)} = 0.6677.$$

2.54 The system of equations $\mathbf{Ax} = \mathbf{b}$ is to be solved iteratively by

$$\mathbf{x}_{n+1} = \mathbf{M}\mathbf{x}_n + \mathbf{b}$$

Suppose $$\mathbf{A} = \begin{bmatrix} 1 & k \\ 2k & 1 \end{bmatrix}, \quad k \neq \sqrt{2}/2,\ k \text{ real},$$

(*a*) Find a necessary and sufficient condition on k for convergence of the Jacobi method.

(*b*) For $k = 0.25$ determine the optimal relaxation factor w, if the system is to be solved with relaxation method. (Lund Univ., Sweden, BIT 13 (1973), 375)

Solution

(*a*) The Jacobi method for the given system is

$$\mathbf{x}_{n+1} = -\begin{bmatrix} 0 & k \\ 2k & 0 \end{bmatrix} \mathbf{x}_n + \mathbf{b} = \mathbf{M}\mathbf{x}_n + \mathbf{b}.$$

The necessary and sufficient condition for convergence of the Jacobi method is $\rho(\mathbf{M}) < 1$. The eigenvalues of $\mathbf{M}$ are given by the equation

$$\lambda^2 - 2k^2 = 0.$$

Hence, $$\rho(\mathbf{M}) = \sqrt{2}\ |\ k\ |.$$

The required condition is therefore

$$\sqrt{2}\ |\ k\ | < 1 \quad \text{or} \quad |\ k\ | < 1/\sqrt{2}.$$

(*b*) The optimal relaxation factor is

$$w_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \mu^2}} = \frac{2}{1 + \sqrt{1 - 2k^2}}$$

$$= \frac{2}{1 + \sqrt{(7/8)}} \approx 1.033 \quad \text{for} \quad k = 0.25.$$

2.55 Suppose that the system of linear equations $\mathbf{Mx} = \mathbf{y}$, is given. Suppose the system can be partitioned in the following way.

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_1 & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{A}_3 & \mathbf{B}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_3 & \mathbf{A}_4 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix},$$

$\mathbf{A}_i$ and $\mathbf{B}_i$ are $p \times p$ matrices and $\mathbf{x}_i$ and $\mathbf{y}_i$ are column vectors ($p \times 1$). Suppose that $\mathbf{A}_i$, $i = 1, 2, 3, 4$ are strictly diagonally dominant and tridiagonal. In that case, systems of the type $\mathbf{A}_i\mathbf{v} = \mathbf{w}$ are easily solved. For system $\mathbf{Mx} = \mathbf{y}$, we therefore propose the following iterative method

$$\begin{aligned} \mathbf{A}_1\mathbf{x}_1^{(n+1)} &= \mathbf{y}_1 - \mathbf{B}_1\,\mathbf{x}_2^{(n)} \\ \mathbf{A}_2\mathbf{x}_2^{(n+1)} &= \mathbf{y}_2 - \mathbf{B}_1\,\mathbf{x}_1^{(n)} - \mathbf{B}_2\,\mathbf{x}_3^{(n)} \\ \mathbf{A}_3\mathbf{x}_3^{(n+1)} &= \mathbf{y}_3 - \mathbf{B}_2\,\mathbf{x}_2^{(n)} - \mathbf{B}_3\,\mathbf{x}_4^{(n)} \\ \mathbf{A}_4\mathbf{x}_4^{(n+1)} &= \mathbf{y}_4 - \mathbf{B}_3\,\mathbf{x}_3^{(n)} \end{aligned}$$

(*i*) if $p = 1$, do you recognize the method ?

(*ii*) Show that for $p > 1$, the method converges if $\| \mathbf{A}_i^{-1} \| < 1/2$ and $\| \mathbf{B}_i \| < 1$.

Solution

(*i*) When $p = 1$, it reduces to the Jacobi iterative method.

(*ii*) The given iteration system can be written as

$$\mathbf{A}\mathbf{x}^{(n+1)} = -\mathbf{B}\mathbf{x}^{(n)} + \mathbf{y}$$

where
$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_4 \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_1 & \mathbf{0} & \mathbf{B}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{0} & \mathbf{B}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_3 & \mathbf{0} \end{bmatrix}$$

Therefore,
$$\mathbf{x}^{(n+1)} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{x}^{(n)} + \mathbf{A}^{-1}\mathbf{y} = \mathbf{H}\mathbf{x}^{(n)} + \mathbf{C}$$

where
$$\mathbf{H} = -\mathbf{A}^{-1}\mathbf{B} = -\begin{bmatrix} \mathbf{0} & \mathbf{A}_1^{-1}\mathbf{B}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{A}_2^{-1}\mathbf{B}_1 & \mathbf{0} & \mathbf{A}_2^{-1}\mathbf{B}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_3^{-1}\mathbf{B}_2 & \mathbf{0} & \mathbf{A}_3^{-1}\mathbf{B}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_4^{-1}\mathbf{B}_3 & \mathbf{0} \end{bmatrix}$$

The iteration converges if $\| \mathbf{H} \| < 1$. This implies that it is sufficient to have $\| \mathbf{A}_i^{-1} \| < 1/2$ and $\| \mathbf{B}_i \| < 1$.

2.56 (*a*) Show that the following matrix formula (where q is a real number) can be used to calculate $\mathbf{A}^{-1}$ when the process converges :

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + q(\mathbf{A}\mathbf{x}^{(n)} - \mathbf{I}).$$

(*b*) When $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, give the values of q for which the process in (*a*) can be used. Which q yields the fastest convergence ?

(*c*) Let $\mathbf{A}$ be a symmetric and positive definite $n \times n$ matrix with smallest eigenvalue λ_1, and greatest eigenvalue λ_2. Find q to get as fast convergence as possible.

(Royal Inst. Tech., Stockholm, Sweden, BIT 24 (1984), 398)

Solution

(*a*) When the process converges, $\mathbf{x}^{(n)} \to \mathbf{x}$ and $\mathbf{x}^{(n+1)} \to \mathbf{x}$.

Then, we have

$$\mathbf{x} = \mathbf{x} + q(\mathbf{A}\mathbf{x} - \mathbf{I}),\ q \neq 0, \quad \text{or} \quad \mathbf{A}\mathbf{x} = \mathbf{I}, \quad \text{or} \quad \mathbf{x} = \mathbf{A}^{-1}.$$

(*b*) The iteration converges if and only if $\rho(\mathbf{I} + q\mathbf{A}) < 1$.

The eigenvalues of $\mathbf{I} + q\mathbf{A}$ are obtained from

$$\begin{vmatrix} 1 + 2q - \lambda & q \\ q & 1 + 2q - \lambda \end{vmatrix} = 0$$

which gives $\lambda = 1 + 3q,\ 1 + q$.

$| \lambda | < 1$ gives the condition $-(2/3) < q < 0$. The minimum value of $\rho(\mathbf{I} + q\mathbf{A})$ is obtained when $| 1 + 3q | = | 1 + q |$, which gives $q = -1/2$. The minimum value is 0.5.

(*c*) $\mathbf{A}$ is a symmetric and positive definite matrix. Hence, $\lambda_i > 0$. The eigenvalues of the iteration matrix $\mathbf{I} + q\mathbf{A}$ are $1 + q\lambda_i$. The iteration converges if and only if

$$-1 < 1 + q\lambda_i < 1, \quad \text{or} \quad -2 < q\lambda_i < 0.$$

Further, since $q < 0$, the smallest and largest eigenvalues of $\mathbf{I} + q\mathbf{A}$ are $1 + q\lambda_1$ and $1 + q\lambda_2$ respectively or vice-versa. Hence, fastest convergence is obtained when

$$| 1 + q\lambda_2 | = | 1 + q\lambda_1 |$$

which gives $q = -2/(\lambda_1 + \lambda_2)$.

2.57 Given the matrix $\mathbf{A} = \mathbf{I} + \mathbf{L} + \mathbf{U}$ where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}$$

$\mathbf{L}$ and $\mathbf{U}$ are strictly lower and upper triangular matrices respectively, decide whether (*a*) Jacobi and (*b*) Gauss-Seidel methods converge to the solution of $\mathbf{Ax} = \mathbf{b}$.

(Royal Inst. Tech., Stockholm, Sweden, BIT 29 (1989), 375)

Solution

(*a*) The iteration matrix of the Jacobi method is

$$\mathbf{H} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) = -(\mathbf{L} + \mathbf{U})$$

$$= -\begin{bmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix}$$

The characteristic equation of $\mathbf{H}$ is

$$|\lambda\mathbf{I} - \mathbf{H}| = \begin{vmatrix} \lambda & 2 & -2 \\ 1 & \lambda & 1 \\ 2 & 2 & \lambda \end{vmatrix} = \lambda^3 = 0.$$

The eigenvalues of $\mathbf{H}$ are $\lambda = 0, 0, 0$ and $\rho(\mathbf{H}) < 1$. The iteration converges.

(*b*) The iteration matrix of the Gauss-Seidel method is

$$\mathbf{H} = -(\mathbf{D} + \mathbf{L})^{-1}\,\mathbf{U}$$

$$= -\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= -\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = -\begin{bmatrix} 0 & 2 & -2 \\ 0 & -2 & 3 \\ 0 & 0 & -2 \end{bmatrix}$$

The eigenvalues of $\mathbf{H}$ are $\lambda = 0, 2, 2$ and $\rho(\mathbf{H}) > 1$.

The iteration diverges.

2.58 Solve the system of equations

$$\begin{aligned} 2x - y &= 1 \\ -x + 2y - z &= 0 \\ -y + 2z - w &= 0 \\ -z + 2w &= 1 \end{aligned}$$

using Gauss-Seidel iteration scheme with $\mathbf{x}^{(0)} = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]^T$. Iterate three times. Obtain the iteration matrix and determine its eigenvalues. Use the extrapolation method and iterate three times. Compare the maximum absolute error and the rate of convergence of the methods.

Solution

We solve the system of equations directly.

$$x^{(k+1)} = \tfrac{1}{2}[1 + y^{(k)}], \quad y^{(k+1)} = \tfrac{1}{2}[x^{(k+1)} + z^{(k)}],$$

$$z^{(k+1)} = \tfrac{1}{2}[y^{(k+1)} + w^{(k)}], \quad w^{(k+1)} = \tfrac{1}{2}[1 + z^{(k+1)}].$$

With $\mathbf{x}^{(0)} = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]^T$, we obtain the following approximate values.

$$\mathbf{x}^{(1)} = [0.75 \quad 0.625 \quad 0.5625 \quad 0.78125]^T,$$
$$\mathbf{x}^{(2)} = [0.8125 \quad 0.6875 \quad 0.7344 \quad 0.8672]^T,$$
$$\mathbf{x}^{(3)} = [0.8434 \quad 0.7889 \quad 0.8281 \quad 0.9140].$$

The Gauss-Seidel method is $\mathbf{x}^{(k+1)} = \mathbf{H}\,\mathbf{x}^{(k)} + \mathbf{c}$, where

$$\mathbf{H} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U} = -\begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 0 & 8 & 0 & 0 \\ 0 & 4 & 8 & 0 \\ 0 & 2 & 4 & 8 \\ 0 & 1 & 2 & 4 \end{bmatrix}$$

$$\mathbf{c} = (\mathbf{D}+\mathbf{L})^{-1}\mathbf{b} = \frac{1}{16}\,[8 \quad 4 \quad 2 \quad 9]^T.$$

The eigenvalues of $\mathbf{H}$ are given by the equation

$$|\,\mathbf{H} - \lambda\mathbf{I}\,| = \lambda^2\left[\lambda^2 - \frac{3}{4}\lambda + \frac{1}{16}\right] = 0,$$

whose solution is $\lambda = 0, 0, (3 \pm \sqrt{5})/8$. The eigenvalues lie in the interval

$$[a, b] = \left[\frac{3-\sqrt{5}}{8}, \frac{3+\sqrt{5}}{8}\right].$$

We have $\rho(\mathbf{H}_{\mathbf{GS}}) = \dfrac{3+\sqrt{5}}{8}$ and rate of convergence $(G\text{-}S) = -\log_{10}\left(\dfrac{3+\sqrt{5}}{8}\right) = 0.1841.$

We have
$$\gamma = \frac{2}{2-a-b} = \frac{2}{2-(3/4)} = \frac{8}{5} = 1.6, \text{ and}$$

$$\mathbf{H}_\gamma = \gamma\mathbf{H} + (1-\gamma)\,\mathbf{I} = -0.6\mathbf{I} + 1.6\mathbf{H} = \begin{bmatrix} -0.6 & 0.8 & 0 & 0 \\ 0 & -0.2 & 0.8 & 0 \\ 0 & 0.2 & -0.2 & 0.8 \\ 0 & 0.1 & 0.2 & -0.2 \end{bmatrix}$$

$$\gamma\mathbf{c} = [0.8 \quad 0.4 \quad 0.2 \quad 0.9]^T.$$

The extrapolation iteration scheme is given by $\mathbf{x}^{(k+1)} = \mathbf{H}_\gamma\mathbf{x}^{(k)} + \gamma\,\mathbf{c}$.

With $\mathbf{x}^{(0)} = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]^T$, we get

$$\mathbf{x}^{(1)} = [0.9 \quad 0.7 \quad 0.6 \quad 0.95]^T,$$
$$\mathbf{x}^{(2)} = [0.82 \quad 0.74 \quad 0.98 \quad 0.9]^T,$$
$$\mathbf{x}^{(3)} = [0.9 \quad 1.036 \quad 0.872 \quad 0.99]^T.$$

We also have $\rho(\mathrm{H}_\gamma) = 1 - |\gamma|\,d$, where d is the distance of 1 from $[a, b] = \left[\dfrac{3-\sqrt{5}}{8}, \dfrac{3+\sqrt{5}}{8}\right]$

which is equal to 0.3455. Hence,

$$\rho(\mathrm{H}_\gamma) = 1 - (1.6)\,(0.3455) = 0.4472,$$

and rate of convergence $= -\log_{10}(0.4472) = 0.3495$. The maximum absolute errors in the Gauss-Seidel method and the extrapolation are respectively (after three iterations) 0.2109 and 0.1280.

2.59 (*a*) Determine the convergence factor for the Jacobi and Gauss-Seidel methods for the system

$$\begin{bmatrix} 4 & 0 & 2 \\ 0 & 5 & 2 \\ 5 & 4 & 10 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ -3 \\ 2 \end{bmatrix}$$

(*b*) This system can also be solved by the relaxation method. Determine w_{opt} and write down the iteration formula exactly. (Lund Univ., Sweden, BIT 13 (1973), 493)

Solution

(*a*) We write the iteration method in the form

$$\mathbf{x}^{(n+1)} = \mathbf{M}\mathbf{x}^{(n)} + \mathbf{c}.$$

For Jacobi method, we have

$$\mathbf{M}_J = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}).$$

For Gauss-Seidel method, we have

$$\mathbf{M}_{\text{GS}} = -(\mathbf{D} + \mathbf{L})^{-1}\,\mathbf{U}.$$

The iteration method converges if and only if $\rho(\mathbf{M}) < 1$.

For Jacobi method, we find

$$\mathbf{M}_J = -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/10 \end{bmatrix}\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1/2 \\ 0 & 0 & -2/5 \\ -1/2 & -2/5 & 0 \end{bmatrix}$$

The eigenvalues of $\mathbf{M}_J$ are $\mu = 0$ and $\mu = \pm\sqrt{0.41}$.

The convergence factor (rate of convergence) of Jacobi method is

$$\nu = -\log_{10}(\rho(\mathbf{M}_J)) = -\log_{10}(\sqrt{0.41}) = 0.194.$$

For Gauss-Seidel method, we find

$$\mathbf{M}_{\mathbf{GS}} = -\begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 5 & 4 & 10 \end{bmatrix}^{-1}\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} = -\frac{1}{200}\begin{bmatrix} 50 & 0 & 0 \\ 0 & 40 & 0 \\ -25 & -16 & 20 \end{bmatrix}\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= -\frac{1}{200}\begin{bmatrix} 0 & 0 & 100 \\ 0 & 0 & 80 \\ 0 & 0 & -82 \end{bmatrix}$$

Eigenvalues of $\mathbf{M}_{\text{GS}}$ are 0, 0, 0.41.

Hence, the convergence factor (rate of convergence) for Gauss-Seidel method is

$$\nu = -\log_{10}(\rho(\mathbf{M}_{\text{GS}})) = -\log_{10}(0.41) = 0.387.$$

(*b*)

$$w_{\text{opt}} = \frac{2}{\mu^2}\left(1 - \sqrt{1-\mu^2}\right), \text{ where } \mu = \rho(\mathbf{M}_J)$$

$$= \frac{2}{0.41}\left(1 - \sqrt{1 - 0.41}\right) \approx 1.132.$$

The SOR method becomes

$$\mathbf{x}^{(n+1)} = \mathbf{M}\mathbf{x}^{(n)} + \mathbf{c}$$

where

$$\mathbf{M} = (\mathbf{D} + w_{\text{opt}}\mathbf{L})^{-1}\,[(1 - w_{\text{opt}})\,\mathbf{D} - w_{\text{opt}}\,\mathbf{U}]$$

$$= \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 5.660 & 4.528 & 10 \end{bmatrix}^{-1}\begin{bmatrix} -0.528 & 0 & -2.264 \\ 0 & -0.660 & -2.264 \\ 0 & 0 & -1.320 \end{bmatrix}$$

$$= \frac{1}{200}\begin{bmatrix} 50 & 0 & 0 \\ 0 & 40 & 0 \\ -28.3 & -18.112 & 20 \end{bmatrix}\begin{bmatrix} -0.528 & 0 & -2.264 \\ 0 & -0.660 & -2.264 \\ 0 & 0 & -1.320 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1320 & 0 & -0.5660 \\ 0 & -0.1320 & -0.4528 \\ 0.0747 & 0.0598 & 0.3944 \end{bmatrix}$$

and $\quad \mathbf{c} = w_{\text{opt}} \, (\mathbf{D} + w_{\text{opt}} \, \mathbf{L})^{-1} \, \mathbf{b}$

$$= \frac{1.132}{200} \begin{bmatrix} 50 & 0 & 0 \\ 0 & 40 & 0 \\ -28.3 & -18.112 & 20 \end{bmatrix} \begin{bmatrix} 4 \\ -3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1.132 \\ -0.6792 \\ -0.1068 \end{bmatrix}$$

2.60 The following system of equations is given

$$4x + y + 2z = 4$$
$$3x + 5y + z = 7$$
$$x + y + 3z = 3$$

(*a*) Set up the Jacobi and Gauss-Seidel iterative schemes for the solution and iterate three times starting with the initial vector $\mathbf{x}^{(0)} = \mathbf{0}$. Compare with the exact solution.

(*b*) Find the spectral radii of the iteration matrices and hence find the rate of convergence of these schemes. (Use the Newton-Raphson method, to find the spectral radius of the iteration matrix of the Jacobi method).

Solution

(*a*) For the given system of equations, we obtain :

Jacobi iteration scheme

$$\mathbf{x}^{(n+1)} = -\begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 3 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \mathbf{x}^{(n)} + \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \begin{pmatrix} 4 \\ 7 \\ 3 \end{pmatrix}$$

$$= -\begin{pmatrix} 0 & 1/4 & 1/2 \\ 3/5 & 0 & 1/5 \\ 1/3 & 1/3 & 0 \end{pmatrix} \mathbf{x}^{(n)} + \begin{pmatrix} 1 \\ 7/5 \\ 1 \end{pmatrix}$$

Starting with $\mathbf{x}^{(0)} = \mathbf{0}$, we get

$$\mathbf{x}^{(1)} = (1 \quad 1.4 \quad 1)^T,$$
$$\mathbf{x}^{(2)} = (0.15 \quad 0.6 \quad 0.2)^T,$$
$$\mathbf{x}^{(3)} = (0.75 \quad 1.27 \quad 0.75)^T.$$

Gauss-Seidel iteration scheme

$$\mathbf{x}^{(n+1)} = -\begin{pmatrix} 4 & 0 & 0 \\ 3 & 5 & 0 \\ 1 & 1 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{x}^{(n)} + \begin{pmatrix} 4 & 0 & 0 \\ 3 & 5 & 0 \\ 1 & 1 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ 7 \\ 3 \end{pmatrix}$$

$$= -\frac{1}{60} \begin{pmatrix} 0 & 15 & 30 \\ 0 & -9 & -6 \\ 0 & -2 & -8 \end{pmatrix} \mathbf{x}^{(n)} + \frac{1}{60} \begin{pmatrix} 60 \\ 48 \\ 24 \end{pmatrix}$$

Starting with $\mathbf{x}^{(0)} = \mathbf{0}$, we get

$$\mathbf{x}^{(1)} = (1.0 \quad 0.8 \quad 0.4)^T, \; \mathbf{x}^{(2)} = (0.6 \quad 0.96 \quad 0.48)^T,$$
$$\mathbf{x}^{(3)} = (0.52 \quad 0.992 \quad 0.496)^T.$$

Exact solution of the given system of equations is $[0.5 \quad 1 \quad 0.5]^T$.

(*b*) The Jacobi iteration matrix is

$$\mathbf{M}_J = \begin{bmatrix} 0 & -1/4 & -1/2 \\ -3/5 & 0 & -1/5 \\ -1/3 & -1/3 & 0 \end{bmatrix}$$

The characteristic equation of $\mathbf{M}_J$ is given by

$$60\lambda^3 - 23\lambda + 7 = 0.$$

The equation has one real root in (– 0.8, 0) and a complex pair. The real root can be obtained by the Newton-Raphson method

$$\lambda_{k+1} = \lambda_k - \frac{60\lambda_k^3 - 23\lambda_k + 7}{180\lambda_k^2 - 23}, \; k = 0, 1, 2, \ldots$$

Starting with $\lambda_0 = -0.6$, we obtain the successive approximations to the root as – 0.7876, – 0.7402, – 0.7361, – 0.7361. The complex pair are the roots of $60\lambda^2 - 44.1660\lambda + 9.5106 = 0$, which is obtained as $0.3681 \pm 0.1518i$. The magnitude of this pair is 0.3981. Hence, $\rho(\mathbf{M}_J) = 0.7876$ and $\nu = -\log_{10}(0.7876) = 0.1037$.

The Gauss-Seidel iteration matrix is

$$\mathbf{M}_{GS} = \begin{bmatrix} 0 & -1/4 & -1/2 \\ 0 & 3/20 & 1/10 \\ 0 & 1/30 & 2/15 \end{bmatrix}$$

The characteristic equation of $\mathbf{M}_{GS}$ is obtained as

$$60\lambda^3 - 17\lambda^2 + \lambda = 0$$

whose roots are 0, 1/12, 1/5. Hence,

$$\rho(\mathbf{M}_{GS}) = 0.2,$$

and $$\nu_{GS} = -\log_{10}(0.2) = 0.6990.$$

2.61 Given a system of equations $\mathbf{Ax} = \mathbf{b}$ where

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Prove that the matrix **A** has 'property *A*' and find the optimum value of the relaxation factor *w* for the method of successive over-relaxation.

(Gothenburg Univ., Sweden, BIT 8 (1968), 138)

Solution

Choose the permutation matrix as

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then, $$\mathbf{PAP}^T = \mathbf{P} \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} \mathbf{P}^T$$

$$= \mathbf{P}\begin{bmatrix} 0 & -1 & 0 & 2 & 0 \\ 0 & 2 & -1 & -1 & 0 \\ -1 & -1 & 2 & 0 & 0 \\ 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 0 & 0 & 2 \end{bmatrix} = \left[\begin{array}{cc|ccc} 2 & 0 & -1 & 0 & -1 \\ 0 & 2 & -1 & -1 & 0 \\ \hline -1 & -1 & 2 & 0 & 0 \\ 0 & -1 & 0 & 2 & 0 \\ -1 & 0 & 0 & 0 & 2 \end{array}\right]$$

Hence, the matrix **A**, has 'property *A*'. The Jacobi iteration matrix is

$$\mathbf{H}_J = \left[\begin{array}{cc|ccc} 0 & 0 & -1/2 & 0 & -1/2 \\ 0 & 0 & -1/2 & -1/2 & 0 \\ \hline -1/2 & -1/2 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 0 & 0 \end{array}\right]$$

The eigenvalues of $\mathbf{H}_J$ are $\mu^* = 0, \pm 1/2, \pm \sqrt{3}/2$. Therefore,

$$\rho(\mathbf{H}_J) = \sqrt{3}/2 = \mu.$$

$$w_{\text{opt}} = \frac{2}{1+\sqrt{1-\mu^2}} = \frac{4}{3}.$$

2.62 The following system of equations is given

$$3x + 2y = 4.5$$
$$2x + 3y - z = 5$$
$$-y + 2z = -0.5$$

Set up the SOR iteration scheme for the solution.

(*a*) Find the optimal relaxation factor and determine the rate of convergence.

(*b*) Using the optimal relaxation factor, iterate five times with the above scheme with $\mathbf{x}^{(0)} = \mathbf{0}$.

(*c*) Taking this value of the optimal relaxation factor, iterate five times, using the error format of the SOR scheme, with $\mathbf{x}^{(0)} = \mathbf{0}$. Compare with the exact solution.

Solution

(*a*) The iteration matrix of the Jacobi method is given by

$$\mathbf{M}_J = -\begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}\begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2/3 & 0 \\ -2/3 & 0 & 1/3 \\ 0 & 1/2 & 0 \end{pmatrix}$$

Eigenvalues of $\mathbf{M}_J$ are $\lambda = 0, \pm\sqrt{11/18}$. Therefore

$$\rho(\mathbf{M}_J) = \mu = \sqrt{11/18}.$$

The optimal relaxation parameter for SOR method is obtained as

$$w_{\text{opt}} = \frac{2}{1+\sqrt{1-\mu^2}} \approx 1.23183,$$

$$\rho(SOR) = w_{\text{opt}} - 1 = 0.23183.$$

Hence, rate of convergence of SOR method is

$$\nu(SOR) = -\log_{10}(0.23183) = 0.6348.$$

(*b*) The SOR iteration scheme can be written as

$$\mathbf{x}^{(n+1)} = \mathbf{M}\mathbf{x}^{(n)} + \mathbf{c}$$

where, with $w = w_{\text{opt}} = 1.23183$, we have

$$\mathbf{M} = (\mathbf{D} + w\mathbf{L})^{-1} [(1 - w)\, \mathbf{D} - w\mathbf{U}]$$

$$= \begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix}^{-1} \begin{bmatrix} -0.6954 & -2.4636 & 0 \\ 0 & -0.6954 & 1.2318 \\ 0 & 0 & -0.4636 \end{bmatrix}$$

$$= \frac{1}{18} \begin{bmatrix} 6 & 0 & 0 \\ -4.9272 & 6 & 0 \\ -3.0347 & 3.6954 & 9 \end{bmatrix} \begin{bmatrix} -0.6954 & -2.4636 & 0 \\ 0 & -0.6954 & 1.2318 \\ 0 & 0 & -0.4636 \end{bmatrix}$$

$$= \begin{bmatrix} -0.2318 & -0.8212 & 0 \\ 0.1904 & 0.4426 & 0.4106 \\ 0.1172 & 0.2726 & 0.0211 \end{bmatrix}$$

and $\mathbf{c} = w(\mathbf{D} + w\mathbf{L})^{-1}\, \mathbf{b}$

$$= \frac{1.2318}{18} \begin{pmatrix} 6 & 0 & 0 \\ -4.9272 & 6 & 0 \\ -3.0347 & 3.6954 & 9 \end{pmatrix} \begin{pmatrix} 4.5 \\ 5 \\ -0.5 \end{pmatrix} = \begin{pmatrix} 1.8477 \\ 0.5357 \\ 0.0220 \end{pmatrix}$$

Hence, we have the iteration scheme

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} -0.2318 & -0.8212 & 0 \\ 0.1904 & 0.4426 & 0.4106 \\ 0.1172 & 0.2726 & 0.0211 \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} 1.8477 \\ 0.5357 \\ 0.0220 \end{bmatrix}$$

$$k = 0, 1, 2, \ldots$$

Starting with $\mathbf{x}^{(0)} = \mathbf{0}$, we obtain

$$\mathbf{x}^{(1)} = [1.8477 \quad 0.5357 \quad 0.0220]^T$$
$$\mathbf{x}^{(2)} = [0.9795 \quad 1.1336 \quad 0.3850]^T$$
$$\mathbf{x}^{(3)} = [0.6897 \quad 1.3820 \quad 0.4539]^T$$
$$\mathbf{x}^{(4)} = [0.5529 \quad 1.4651 \quad 0.4891]^T$$
$$\mathbf{x}^{(5)} = [0.5164 \quad 1.4902 \quad 0.4965]^T$$

(*c*) $(\mathbf{D} + w_{\text{opt}}\, \mathbf{L})\, \mathbf{v}^{(k+1)} = w_{\text{opt}}\, \mathbf{r}^{(k)}$

or $(\mathbf{D} + 1.2318\, \mathbf{L})\, \mathbf{v}^{(k+1)} = 1.2318\mathbf{r}^{(k)}$ where $\mathbf{v}^{(k+1)} = \mathbf{x}^{(x+1)} - \mathbf{x}^{(k)}$, and $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$.

We have

$$\begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix} \mathbf{v}^{(k+1)} = 1.2318 \begin{bmatrix} 4.5 - 3x^{(k)} - 2y^{(k)} \\ 5.0 - 2x^{(k)} - 3y^{(k)} + z^{(k)} \\ -0.5 + y^{(k)} - 2z^{(k)} \end{bmatrix}$$

With $\mathbf{x}^{(0)} = \mathbf{0}$, we obtain the following iterations. The equations are solved by forward substitution.

First iteration

$$\mathbf{v}^{(1)} = \mathbf{x}^{(1)} = [1.8477 \quad 0.5357 \quad 0.0220]^T.$$

Second iteration

$$\begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix} \mathbf{v}^{(2)} = \begin{bmatrix} -2.6046 \\ -0.3455 \\ -0.0102 \end{bmatrix},$$

which gives, $\mathbf{v}^{(2)} = [-0.8682 \quad 0.5978 \quad 0.3631]^T$,

and $\mathbf{x}^{(2)} = \mathbf{v}^{(2)} + \mathbf{x}^{(1)} = [0.9795 \quad 1.1335 \quad 0.3851]^T$.

Third iteration

$$\begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix} \mathbf{v}^{(3)} = \begin{bmatrix} -0.8690 \\ 0.0315 \\ -0.1684 \end{bmatrix},$$

which gives, $\mathbf{v}^{(3)} = [-0.2897 \quad 0.2484 \quad 0.0688]^T$,

and $\mathbf{x}^{(3)} = \mathbf{v}^{(3)} + \mathbf{x}^{(2)} = [0.6898 \quad 1.3819 \quad 0.4539]^T$.

Fourth iteration

$$\begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix} \mathbf{v}^{(4)} = \begin{bmatrix} -0.4104 \\ -0.0880 \\ -0.0319 \end{bmatrix},$$

which gives, $\mathbf{v}^{(4)} = [-0.1368 \quad 0.0830 \quad 0.0352]^T$,

and $\mathbf{x}^{(4)} = \mathbf{v}^{(4)} + \mathbf{x}^{(3)} = [0.5530 \quad 1.4649 \quad 0.4891]^T$.

Fifth iteration

$$\begin{bmatrix} 3 & 0 & 0 \\ 2.4636 & 3 & 0 \\ 0 & -1.2318 & 2 \end{bmatrix} \mathbf{v}^{(5)} = \begin{bmatrix} -0.1094 \\ -0.0143 \\ -0.0164 \end{bmatrix},$$

which gives $\mathbf{v}^{(5)} = [-0.0365 \quad 0.0252 \quad 0.0073]^T$,

and $\mathbf{x}^{(5)} = \mathbf{v}^{(5)} + \mathbf{x}^{(4)} = [0.5165 \quad 1.4901 \quad 0.4964]^T$.

Exact solution is $\mathbf{x} = [0.5 \quad 1.5 \quad 0.5]^T$.

EIGENVALUE PROBLEMS

2.63 Using the Jacobi method find all the eigenvalues and the corresponding eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & \sqrt{2} & 2 \\ \sqrt{2} & 3 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{bmatrix}$$

Solution

The largest off-diagonal element is $a_{13} = a_{31} = 2$. The other two elements in this 2×2 submatrix are $a_{11} = 1$ and $a_{33} = 1$.

$$\theta = \frac{1}{2} \tan^{-1}\left(\frac{4}{0}\right) = \pi/4$$

$$\mathbf{S}_1 = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}$$

The *first rotation* gives

$$\mathbf{B}_1 = \mathbf{S}_1^{-1} \mathbf{A} \mathbf{S}_1$$

$$= \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & \sqrt{2} & 2 \\ \sqrt{2} & 3 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

The largest off-diagonal element in magnitude in $\mathbf{B}_1$ is $a_{12} = a_{21} = 2$. The other elements are $a_{11} = 3$, $a_{22} = 3$.

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{4}{0}\right) = \pi/4 \quad \text{and} \quad \mathbf{S}_2 = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The *second rotation* gives

$$\mathbf{B}_2 = \mathbf{S}_2^{-1}\mathbf{B}_1\mathbf{S}_2 = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

We have the matrix of eigenvectors as

$$\mathbf{S} = \mathbf{S}_1\mathbf{S}_2 = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/2 & -1/2 & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/2 & -1/2 & 1/\sqrt{2} \end{bmatrix}$$

The eigenvalues are 5, 1, – 1 and the corresponding eigenvectors are the columns of **S**.

2.64 Find all the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} 2 & 3 & 1 \\ 3 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

by the Jacobi method.

Solution

This example illustrates the fact that in the Jacobi method, zeros once created may be disturbed and thereby the number of iterations required are increased. We have the following results.

First rotation

Largest off diagonal element in magnitude = $a_{12} = 3$.

$$\tan 2\theta = \frac{2a_{12}}{a_{11} - a_{22}} = \frac{6}{0}, \quad \theta = \frac{\pi}{4}$$

$$\mathbf{S}_1 = \begin{bmatrix} 0.707106781 & -0.707106781 & 0 \\ 0.707106781 & 0.707106781 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}_1 = \mathbf{S}_1^{-1}\,\mathbf{AS}_1 = \mathbf{S}_1^{T}\,\mathbf{AS}_1$$

$$= \begin{bmatrix} 5.0 & 0 & 2.121320343 \\ 0 & -1.0 & 0.707106781 \\ 2.121320343 & 0.707106781 & 1.0 \end{bmatrix}$$

Second rotation

Largest off diagonal element in magnitude = a_{13}.

$$\tan 2\theta = \frac{2a_{13}}{a_{11} - a_{33}}.$$

We get $\theta = 0.407413458.$

$$\mathbf{S}_2 = \begin{bmatrix} 0.918148773 & 0.0 & -0.396235825 \\ 0.0 & 1.0 & 0.0 \\ 0.396235825 & 0.0 & 0.918148773 \end{bmatrix}$$

$$\mathbf{A}_2 = \mathbf{S}_2^T \mathbf{A}_1 \mathbf{S}_2$$

$$= \begin{bmatrix} 5.915475938 & 0.280181038 & 0.0 \\ 0.280181038 & -1.0 & 0.649229223 \\ 0.0 & 0.649229223 & 0.08452433 \end{bmatrix}$$

Notice now that the zero in the (1, 2) position is disturbed. After six iterations, we get

$$\mathbf{A}_6 = \mathbf{S}_6^T\, \mathbf{A}_5 \mathbf{S}_6$$

$$= \begin{bmatrix} 5.9269228 & -0.000089 & 0.0 \\ -0.000089 & -1.31255436 & 0.0 \\ 0 & 0 & 0.38563102 \end{bmatrix}$$

Hence, the approximate eigenvalues are 5.92692, – 1.31255 and 0.38563. The orthogonal matrix of eigenvectors is given by $\mathbf{S} = \mathbf{S}_1\mathbf{S}_2\mathbf{S}_3\mathbf{S}_4\mathbf{S}_5\mathbf{S}_6$. We find that the corresponding eigenvectors are

$$\mathbf{x}_1 = [-0.61853 \quad -0.67629 \quad -0.40007]^T,$$
$$\mathbf{x}_2 = [0.54566 \quad -0.73605 \quad 0.40061]^T,$$
$$\mathbf{x}_3 = [0.56540 \quad -0.29488 \quad -0.82429]^T.$$

2.65 Transform the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & -1 \\ 3 & -1 & 1 \end{bmatrix}$$

to tridiagonal form by Given's method. Use exact arithmetic.

Solution

Perform the orthogonal rotation with respect to $a_{22}, a_{23}, a_{32}, a_{33}$ submatrix. We get

$$\tan\theta = \frac{a_{13}}{a_{12}} = \frac{3}{2}, \quad \cos\theta = \frac{2}{\sqrt{13}},\ \sin\theta = \frac{3}{\sqrt{13}}.$$

Hence, $\mathbf{B} = \mathbf{S}^{-1}\mathbf{M}\mathbf{S} = \mathbf{S}^T\,\mathbf{M}\mathbf{S}$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2/\sqrt{3} & 3/\sqrt{3} \\ 0 & -3/\sqrt{13} & 2/\sqrt{13} \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & -1 \\ 3 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2/\sqrt{3} & -3/\sqrt{3} \\ 0 & 3/\sqrt{13} & 2/\sqrt{13} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \sqrt{3} & 0 \\ \sqrt{3} & 1/13 & 5/13 \\ 0 & 5/13 & 25/13 \end{bmatrix}$$

is the required tridiagonal form.

2.66 Transform the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}$$

to tridiagonal form by Givens method. Find the eigenvector corresponding to the largest eigenvalue from the eigenvectors of the tridiagonal matrix.

(Uppsala Univ. Sweden, BIT 6 (1966), 270)

Solution

Using the Given's method, we have

$$\tan \theta = \frac{a_{13}}{a_{12}} = 1 \quad \text{or} \quad \theta = \frac{\pi}{4}$$

$$\mathbf{A}_1 = \mathbf{S}^{-1}\,\mathbf{AS}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

which is the required tridiagonal form.

The characteristic equation of $\mathbf{A}_1$ is given by

$$f_n = |\,\lambda\mathbf{I} - \mathbf{A}\,| = \begin{vmatrix} \lambda - 1 & -2\sqrt{2} & 0 \\ -2\sqrt{2} & \lambda - 3 & 0 \\ 0 & 0 & \lambda + 1 \end{vmatrix} = 0$$

The Sturm sequence $\{ f_n \}$ is defined as

$$f_0 = 1,$$

$$f_1 = \lambda - 1,$$

$$f_2 = (\lambda - 3)\, f_1 - (-2\sqrt{2})^2 f_0 = \lambda^2 - 4\lambda - 5,$$

$$f_3 = (\lambda + 1)\, f_2 - (0)^2 f_1 = (\lambda + 1)(\lambda + 1)(\lambda - 5).$$

Since, $f_3(-1) = 0$ and $f_3(5) = 0$, the eigenvalues of $\mathbf{A}$ are -1, -1 and 5. The largest eigenvalue in magnitude is 5.

The eigenvector corresponding to $\lambda = 5$ of $\mathbf{A}_1$ is $\mathbf{v}_1 = [1 \quad \sqrt{2} \quad 0]^T$.

Hence, the corresponding eigenvector of $\mathbf{A}$ is

$$\mathbf{v} = \mathbf{S}\mathbf{v}_1 = [1 \quad 1 \quad 1]^T.$$

2.67 Transform, using Givens method, the symmetric matrix $\mathbf{A}$, by a sequence of orthogonal transformations to tridiagonal form. Use exact arithmetic.

$$\mathbf{A} = \begin{bmatrix} 1 & \sqrt{2} & \sqrt{2} & 2 \\ \sqrt{2} & -\sqrt{2} & -1 & \sqrt{2} \\ \sqrt{2} & -1 & \sqrt{2} & \sqrt{2} \\ 2 & \sqrt{2} & \sqrt{2} & -3 \end{bmatrix}$$

(Inst. Tech., Lund, Sweden, BIT 4 (1964), 261)

Solution

Using the Given's method, we obtain the following.

First rotation

$$\tan \theta_1 = \frac{a_{13}}{a_{12}} = 1, \quad \theta_1 = \frac{\pi}{4},$$

$$\mathbf{S}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{S}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\mathbf{A}_1 = \mathbf{S}_1^{-1}\mathbf{A}\mathbf{S}_1 = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 2 & -1 & \sqrt{2} & 2 \\ 0 & \sqrt{2} & 1 & 0 \\ 2 & 2 & 0 & -3 \end{bmatrix} = (a'_{ij}).$$

Second rotation

$$\tan \theta_2 = \frac{a'_{14}}{a'_{12}} = 1, \quad \theta_2 = \frac{\pi}{4},$$

$$\mathbf{S}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 0 & 1 & 0 \\ 0 & 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}; \quad \mathbf{S}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 0 & 1 & 0 \\ 0 & -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix};$$

$$\mathbf{A}_2 = \mathbf{S}_2^{-1}\mathbf{A}_1\mathbf{S}_2 = \begin{bmatrix} 1 & 2\sqrt{2} & 0 & 0 \\ 2\sqrt{2} & 0 & 1 & -1 \\ 0 & 1 & 1 & -1 \\ 0 & -1 & -1 & -4 \end{bmatrix} = (a_{ij}^*)$$

Third rotation

$$\tan \theta_3 = \frac{a_{24}^*}{a_{23}^*} = -1, \quad \theta_3 = -\frac{\pi}{4},$$

$$\mathbf{S}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}; \quad \mathbf{S}_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix};$$

$$\mathbf{A}_3 = \mathbf{S}_3^{-1}\mathbf{A}_2\mathbf{S}_3 = \begin{bmatrix} 1 & 2\sqrt{2} & 0 & 0 \\ 2\sqrt{2} & 0 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & -1/2 & 5/2 \\ 0 & 0 & 5/2 & -5/2 \end{bmatrix}$$

which is the required tridiagonal form.

2.68 Find all the eigenvalues of the matrix

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 2 \\ -1 & 2 & 1 \end{bmatrix}$$

using the Householder method.

Solution

Choose $\mathbf{w}_2^T = [0 \quad x_2 \quad x_3]$ such that $x_2^2 + x_3^2 = 1$. The parameters in the first Householder transformation are obtained as follows :

$$s_1 = \sqrt{a_{12}^2 + a_{13}^2} = \sqrt{5},$$

$$x_2^2 = \frac{1}{2}\left[1 + \frac{a_{12}}{s_1} \text{ sign } (a_{12})\right] = \frac{1}{2}\left(1 + \frac{2}{\sqrt{5}}\right) = \frac{\sqrt{5}+2}{2\sqrt{5}},$$

$$x_3 = \frac{a_{13} \text{ sign } (a_{12})}{2s_1x_2} = -\frac{1}{2s_1x_2},$$

$$x_2 x_3 = -\frac{1}{2\sqrt{5}},$$

$$\mathbf{P}_2 = \mathbf{I} - 2\mathbf{w}_2\mathbf{w}_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2/\sqrt{5} & 1/\sqrt{5} \\ 0 & 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix}$$

The required Householder transformation is

$$\mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 1 & -\sqrt{5} & 0 \\ -\sqrt{5} & -3/5 & -6/5 \\ 0 & -6/5 & 13/5 \end{bmatrix}$$

Using the Given's method, we obtain the Sturm's sequence as

$$f_0 = 1, f_1 = \lambda - 1,$$

$$f_2 = \lambda^2 - \frac{2}{5}\lambda - \frac{28}{5},$$

$$f_3 = \lambda^3 - 3\lambda^2 - 6\lambda + 16.$$

Let $V(\lambda)$ denote the number of changes in sign in the Sturm sequence. We have the following table giving $V(\lambda)$

λ	f_0	f_1	f_2	f_3	$V(\lambda)$
– 3	+	–	+	–	3
– 2	+	–	–	+	2
– 1	+	–	–	+	2
0	+	–	–	+	2
1	+	+	–	+	2
2	+	+	–	0	1
3	+	+	+	–	1
4	+	+	+	+	0

Since $f_3 = 0$ for $\lambda = 2$, $\lambda = 2$ is an eigenvalue. The remaining two eigenvalues lie in the intervals (– 3, – 2) and (3, 4). Repeated bisection and application of the Sturm's theorem gives the eigenvalues as $\lambda_2 = -2.372$ and $\lambda_3 = 3.372$. Exact eigenvalues are 2, $(1 \pm \sqrt{33})/2$.

2.69 Use the Householder's method to reduce the given matrix **A** into the tridiagonal form

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

Solution

First transformation :

$$\mathbf{w}_2 = [0 \quad x_2 \quad x_3 \quad x_4]^T,$$

$$s_1 = \sqrt{a_{12}^2 + a_{13}^2 + a_{14}^2} = 3,$$

$$x_2^2 = \frac{1}{2}\left[1 + \frac{(-1)(-1)}{3}\right] = \frac{2}{3}\ ; \quad x_2 = \sqrt{\frac{2}{3}},$$

$$x_3 = \frac{2}{2(3)}\sqrt{\frac{3}{2}} = \frac{1}{\sqrt{6}}\ ; \quad x_4 = -\frac{1}{\sqrt{6}},$$

$$\mathbf{P}_2 = \mathbf{I} - 2\mathbf{w}_2\mathbf{w}_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1/3 & -2/3 & 2/3 \\ 0 & -2/3 & 2/3 & 1/3 \\ 0 & 2/3 & 1/3 & 2/3 \end{bmatrix}$$

$$\mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & 16/3 & 2/3 & 1/3 \\ 0 & 2/3 & 16/3 & -1/3 \\ 0 & 1/3 & -1/3 & 4/3 \end{bmatrix}$$

Second transformation :

$$\mathbf{w}_3 = [0 \quad 0 \quad x_3 \quad x_4]^T,$$

$$s_1 = \sqrt{a_{23}^2 + a_{24}^2} = \frac{\sqrt{5}}{3},$$

$$x_3^2 = \frac{1}{2}\left[1 + \frac{2/3}{\sqrt{5}/3}\right] = \left(\frac{\sqrt{5}+2}{2\sqrt{5}}\right) = a,$$

$$x_4^2 = 1 - x_3^2 = 1 - \frac{\sqrt{5}+2}{2\sqrt{5}} = \frac{\sqrt{5}-2}{2\sqrt{5}} = \frac{1}{20a},$$

$$\mathbf{P}_3 = \mathbf{I} - 2\mathbf{w}_3\,\mathbf{w}_3^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-2a & -1/\sqrt{5} \\ 0 & 0 & -1/\sqrt{5} & 1-1/(10a) \end{bmatrix}$$

$$\mathbf{A}_3 = \mathbf{P}_3\mathbf{A}_2\mathbf{P}_3 = \begin{bmatrix} 4 & 3 & 0 & 0 \\ 3 & 16/3 & -5/(3\sqrt{5}) & 0 \\ 0 & -5/(3\sqrt{5}) & 16/3 & 9/5 \\ 0 & 0 & 9/5 & 12/5 \end{bmatrix}$$

is the required tridiagonal form

2.70 Find approximately the eigenvalues of the matrix

$$\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

Using the Rutishauser method. Apply the procedure until the elements of the lower triangular part are less than 0.005 in magnitude.

We have the following decompositions :

$$\mathbf{A}_1 = \mathbf{A} = \mathbf{L}_1\mathbf{U}_1 = \begin{bmatrix} 1 & 0 \\ 1/3 & 1 \end{bmatrix}\begin{bmatrix} 3 & 1 \\ 0 & 2/3 \end{bmatrix}$$

$$\mathbf{A}_2 = \mathbf{U}_1\mathbf{L}_1 = \begin{bmatrix} 10/3 & 1 \\ 2/9 & 2/3 \end{bmatrix}$$

$$= \mathbf{L}_2\,\mathbf{U}_2 = \begin{bmatrix} 1 & 0 \\ 1/15 & 1 \end{bmatrix}\begin{bmatrix} 10/3 & 1 \\ 0 & 3/5 \end{bmatrix}$$

$$\mathbf{A}_3 = \mathbf{U}_2\mathbf{L}_2 = \begin{bmatrix} 17/5 & 1 \\ 1/25 & 3/5 \end{bmatrix}$$

$$= \mathbf{L}_3\mathbf{U}_3 = \begin{bmatrix} 1 & 0 \\ 1/85 & 1 \end{bmatrix}\begin{bmatrix} 17/5 & 1 \\ 0 & 10/17 \end{bmatrix}$$

$$\mathbf{A}_4 = \mathbf{U}_3\mathbf{L}_3 = \begin{bmatrix} 58/17 & 1 \\ 2/289 & 10/17 \end{bmatrix}$$

$$= \mathbf{L}_4\mathbf{U}_4 = \begin{bmatrix} 1 & 0 \\ 1/493 & 1 \end{bmatrix}\begin{bmatrix} 58/17 & 1 \\ 0 & 289/493 \end{bmatrix}$$

$$\mathbf{A}_5 = \mathbf{U}_4\mathbf{L}_4 = \begin{bmatrix} 3.4138 & 1 \\ 0.0012 & 0.5862 \end{bmatrix}$$

To the required accuracy the eigenvalues are 3.4138 and 0.5862. The exact eigenvalues are $2 \pm \sqrt{2}$.

2.71 Find all the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 1 & 3 & 2 \end{bmatrix}$$

using the Rutishauser method. Iterate till the elements of the lower triangular part are less than 0.05 in magnitude.

Solution

We have

$$\mathbf{A}_1 = \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{L}_1\mathbf{U}_1$$

$$\mathbf{A}_2 = \mathbf{U}_1\,\mathbf{L}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 & 1 \\ -2 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/4 & 7/6 & 1 \end{bmatrix} \begin{bmatrix} 4 & -1 & 1 \\ 0 & -3/2 & 1/2 \\ 0 & 0 & 1/6 \end{bmatrix} = \mathbf{L}_2\mathbf{U}_2$$

$$\mathbf{A}_3 = \mathbf{U}_2\,\mathbf{L}_2 = \begin{bmatrix} 19/4 & 1/6 & 1 \\ 7/8 & -11/12 & 1/2 \\ 1/24 & 7/36 & 1/6 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 7/38 & 1 & 0 \\ 1/114 & -11/54 & 1 \end{bmatrix} \begin{bmatrix} 19/4 & 1/6 & 1 \\ 0 & -18/19 & 6/19 \\ 0 & 0 & 38/171 \end{bmatrix} = \mathbf{L}_3\mathbf{U}_3$$

$$\mathbf{A}_4 = \mathbf{U}_3\mathbf{L}_3 = \begin{bmatrix} 4.789474 & -0.037037 & 1 \\ -0.171745 & -1.011696 & 0.315789 \\ 0.001949 & -0.045267 & 0.222222 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ -0.035859 & 1 & 0 \\ 0.000407 & 0.044670 & 1 \end{bmatrix} \begin{bmatrix} 4.789474 & -0.037037 & 1 \\ 0 & -1.013024 & 0.351648 \\ 0 & 0 & 0.206107 \end{bmatrix} = \mathbf{L}_4\mathbf{U}_4$$

$$\mathbf{A}_5 = \mathbf{U}_4\mathbf{L}_4 = \begin{bmatrix} 4.791209 & 0.007633 & 1 \\ 0.036469 & -0.997316 & 0.351648 \\ 0.000084 & 0.009207 & 0.206107 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0.007612 & 1 & 0 \\ 0.000018 & -0.009231 & 1 \end{bmatrix} \begin{bmatrix} 4.791209 & 0.007633 & 1 \\ 0 & -0.997374 & 0.344036 \\ 0 & 0 & 0.209265 \end{bmatrix} = \mathbf{L}_5\mathbf{U}_5$$

$$\mathbf{A}_6 = \mathbf{U}_5\mathbf{L}_5 = \begin{bmatrix} 4.791285 & -0.001598 & 1 \\ -0.007586 & -1.000550 & 0.344036 \\ 0.000004 & -0.001932 & 0.209265 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ -0.001583 & 1 & 0 \\ 0.000001 & 0.001931 & 1 \end{bmatrix} \begin{bmatrix} 4.791285 & -0.001598 & 1 \\ 0 & -1.000553 & 0.345619 \\ 0 & 0 & 0.208597 \end{bmatrix} = \mathbf{L}_6\,\mathbf{U}_6$$

$$\mathbf{A}_7 = \mathbf{U}_6\,\mathbf{L}_6 = \begin{bmatrix} 4.791289 & 0.000333 & 1 \\ 0.001584 & -0.999886 & 0.345619 \\ 0 & 0.000403 & 0.208597 \end{bmatrix}$$

Hence, the eigenvalues are approximately 4.791289, – 0.999886 and 0.208597. The exact eigenvalues are

$$\lambda = (5 + \sqrt{21})\,/\,2 = 4.791288,\ \lambda = -1 \text{ and } \lambda = (5 - \sqrt{21})\,/\,2 = 0.208712$$

2.72 Find the largest eigenvalue of the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

using power method. (Stockholm Univ., Sweden, BIT 7 (1967), 81)

Solution

Starting with $\mathbf{v}_0 = [1 \quad 1 \quad 1 \quad 1]^T$ and using the algorithm for power method we obtain

$$\mathbf{y}_1 = \mathbf{A}\mathbf{v}_0 = [4 \quad 3 \quad 3 \quad 4]^T,$$

$$\mathbf{v}_1 = \frac{\mathbf{y}_1}{m_1} = [1 \quad 3/4 \quad 3/4 \quad 1]^T,$$

$$\mathbf{y}_2 = \mathbf{A}\mathbf{v}_1 = [7/2 \quad 11/4 \quad 11/4 \quad 7/2]^T,$$

$$\mathbf{v}_2 = \frac{\mathbf{y}_2}{m_2} = [1 \quad 11/14 \quad 11/14 \quad 1]^T,$$

$$\mathbf{y}_3 = \mathbf{A}\mathbf{v}_2 = [25/7 \quad 39/14 \quad 39/14 \quad 25/7]^T,$$

$$\mathbf{v}_3 = \frac{\mathbf{y}_3}{m_3} = [1 \quad 39/50 \quad 39/50 \quad 1]^T,$$

$$\ldots \qquad \ldots$$

$$\mathbf{y}_5 = \mathbf{A}\mathbf{v}_4 = [317/89 \quad 495/178 \quad 495/178 \quad 317/89]^T,$$

$$\mathbf{v}_5 = [1 \quad 495/634 \quad 495/634 \quad 1]^T,$$

$$\mathbf{y}_6 = \mathbf{A}\mathbf{v}_5 = [1129/317 \quad 1763/634 \quad 1763/634 \quad 1129/317]^T,$$

$$\mathbf{v}_6 = [1 \quad 1763/2258 \quad 1763/2258 \quad 1]^T,$$

After six iterations, the ratios

$$(\mathbf{y}_6)_r / (\mathbf{v}_5)_r, \; r = 1, 2, 3, 4$$

are 3.5615, 3.5616, 3.5616 and 3.5615. Hence, the largest eigenvalue in magnitude is 3.5615. The corresponding eigenvector is $[1 \quad 0.7808 \quad 0.7808 \quad 1]^T$.

2.73 Determine the largest eigenvalue and the corresponding eigenvector of the matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 20 & 1 \\ 0 & 1 & 4 \end{bmatrix}$$

to 3 correct decimal places using the power method.

(Royal Inst. Tech., Stockholm, Sweden, BIT 11 (1971), 125)

Solution

Starting with $\mathbf{v}_0 = [1 \quad 1 \quad 1]^T$ and using power method we obtain the following :

$$\mathbf{y}_1 = \mathbf{A}\mathbf{v}_0 = [5 \quad 22 \quad 5]^T,$$

$$\mathbf{v}_1 = \frac{\mathbf{y}_1}{m_1} = [5/22 \quad 1 \quad 5/22]^T$$

$$\mathbf{y}_2 = \mathbf{A}\mathbf{v}_1 = [21/11 \quad 225/11 \quad 21/11]^T,$$

$$\mathbf{v}_2 = \frac{\mathbf{y}_2}{m_2} = [21/225 \quad 1 \quad 21/225]^T$$

$$\ldots \qquad \ldots$$

$$\mathbf{y}_7 = \mathbf{A}\mathbf{v}_6 = [1.24806 \quad 20.12412 \quad 1.24824]^T,$$

$$\mathbf{v}_7 = \frac{\mathbf{y}_7}{m_7} = [0.06202 \quad 1 \quad 0.06202]^T,$$

$$\mathbf{y}_8 = \mathbf{A}\mathbf{v}_7 = [1.24806 \quad 20.12404 \quad 1.24806]^T,$$

$$\mathbf{v}_8 = \frac{\mathbf{y}_8}{m_8} = [0.06202 \quad 1 \quad 0.06202]^T$$

After 8 iterations, the ratios $(\mathbf{y}_8)_r / (\mathbf{v}_7)_r$, $r = 1, 2, 3$ are 20.1235, 20.1240 and 20.1235. The largest eigenvalue in magnitude correct to 3 decimal places is 20.124 and the corresponding eigenvector is

$$[0.06202 \quad 1 \quad 0.06202]^T.$$

2.74 Compute with an iterative method the greatest charateristic number λ of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

with four correct decimal places. (Lund Univ., Sweden, BIT 4 (1964), 131)

Solution

Starting with $\mathbf{v}_0 = [1 \quad 1 \quad 1 \quad 1 \quad 1]^T$ and using the power method, we obtain the following :

$$\mathbf{y}_1 = \mathbf{A}\mathbf{v}_0 = [2 \quad 2 \quad 3 \quad 2 \quad 3]^T$$

$$\mathbf{v}_1 = \frac{\mathbf{y_1}}{m_1} = [0.666667 \quad 0.666667 \quad 1 \quad 0.666667 \quad 1]^T$$

$$\mathbf{y}_2 = \mathbf{A}\mathbf{v}_1 = [1.666667 \quad 2 \quad 2.333334 \quad 1.666667 \quad 2.333334]^T$$

$$\mathbf{v}_2 = [0.714286 \quad 0.857143 \quad 1 \quad 0.714286 \quad 1]^T$$

$$\ldots \qquad \ldots$$

$$\mathbf{y}_{13} = \mathbf{A}\mathbf{v}_{12} = [1.675145 \quad 2 \quad 2.481239 \quad 1.675145 \quad 2.481239]^T$$

$$\mathbf{v}_{13} = [0.675124 \quad 0.806049 \quad 1 \quad 0.675124 \quad 1]^T$$

$$\mathbf{y}_{14} = \mathbf{A}\mathbf{v}_{13} = [1.675124 \quad 2 \quad 2.481173 \quad 1.675124 \quad 2.481173]^T$$

$$\mathbf{v}_{14} = [0.675124 \quad 0.806070 \quad 1 \quad 0.675134 \quad 1]^T$$

After 14 iterations, the ratios $(\mathbf{y}_{14})_r / (\mathbf{v}_{13})_r$, $r = 1, 2, 3, 4, 5$ are 2.481209, 2.481238, 2.481173, 2.481238 and 2.481173. Hence, the largest eigenvalue in magnitude may be taken as 2.4812.

2.75 Calculate an approximation to the least eigenvalue of $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, where

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

using one step of inverse iteration. Choose the vector $(6 \quad -7 \quad 3)^T$ as a first approximation to the corresponding eigenvector. Estimate the error in the approximate eigenvalue.

(Univ. and Inst. Tech., Linköping, BIT 28 (1988), 373)

Solution

The inverse power method is defined by

$$\mathbf{z}_{k+1} = \mathbf{A}^{-1}\, \mathbf{y}_k,$$

$$\mathbf{y}_{k+1} = \mathbf{z}_{k+1} / m_{k+1},$$

where m_{k+1} is the maximal element in magnitude of $\mathbf{z}_{k+1}$ and $\mathbf{y}_0$ is the initial approximation to the eigenvector. We have alternately,

$$\mathbf{A}\mathbf{z}_{k+1} = \mathbf{L}\mathbf{L}^T\mathbf{z}_{k+1} = \mathbf{y}_k,$$

$$\mathbf{y}_{k+1} = \mathbf{z}_{k+1} / m_{k+1}.$$

Set $\mathbf{L}^T\mathbf{z}_{k+1} = \mathbf{t}_{k+1}$. Solve $\mathbf{L}\mathbf{t}_{k+1} = \mathbf{y}_k$ and then solve $\mathbf{L}^T\mathbf{z}_{k+1} = \mathbf{t}_{k+1}$.

Solving $\mathbf{L}\mathbf{t}_1 = \mathbf{y}_0$, where $\mathbf{y}_0 = [6 \quad -7 \quad 3]^T$, we get $\mathbf{t}_1 = [6 \quad -13 \quad 10]^T$.

Solving $\mathbf{L}^T\mathbf{z}_1 = \mathbf{t}_1$ we get

$$\mathbf{z}_1 = [19 \quad -23 \quad 10]^T \quad \text{and} \quad \mathbf{y}_1 = [19/23 \quad -1 \quad 10/23]^T.$$

Hence, the ratios approximating the largest eigenvalue of $\mathbf{A}^{-1}$ are 19 / 6, 23 / 7, 10 / 3, *i.e.*, 3.167, 3.286 and 3.333. The approximation to the smallest eigenvalue in magnitude of **A** may be taken as 3.2. The exact eigenvalue is 5.0489 (approximately).

2.76 Find the smallest eigenvalue in magnitude of the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

using four iterations of the inverse power method.

Solution

The smallest eigenvalue in magnitude of **A** is the largest eigenvalue in magnitude of $\mathbf{A}^{-1}$. We have

$$\mathbf{A}^{-1} = \begin{bmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{bmatrix}$$

Using $\quad \mathbf{y}^{(k+1)} = \mathbf{A}^{-1}\,\mathbf{v}^{(k)}, \; k = 0, 1 \ldots$

and $\quad \mathbf{v}^{(0)} = [1, \quad 1, \quad 1]^T$, we obtain

$$\mathbf{y}^{(1)} = [1.5, \quad 2, \quad 1.5]^T, \quad \mathbf{v}^{(1)} = [0.75, \quad 1, \quad 0.75]^T$$

$$\mathbf{y}^{(2)} = [1.25, \quad 1.75, \quad 1.25]^T, \; \mathbf{v}^{(2)} = [0.7143, \quad 1, \quad 0.7143]^T$$

$$\mathbf{y}^{(3)} = [1.2143, \quad 1.7143, \quad 1.2143]^T, \; \mathbf{v}^{(3)} = [0.7083, \quad 1, \quad 0.7083]^T$$

$$\mathbf{y}^{(4)} = [1.2083, 1.7083, 1.2083]^T, \; \mathbf{v}^{(4)} = [0.7073, \quad 1, \quad 0.7073]^T.$$

After four iterations, we obtain the ratios as

$$\mu = \frac{[\mathbf{y}^{(4)}]_r}{[\mathbf{v}^{(3)}]_r} = (1.7059, \quad 1.7083 \quad 1.7059).$$

Therefore, $\quad \mu = 1.71 \quad \text{and} \quad \lambda = 1/\mu \approx 0.5848.$

Since $|\,\mathbf{A} - 0.5848\ \mathbf{I}\,| \approx 0$, $\lambda = 0.5848$ is the required eigenvalue. The corresponding eigenvector is $[0.7073, \quad 1, \quad 0.7043]^T$

The smallest eigenvalue of **A** is $2 - \sqrt{2} = 0.5858$.

Alternately, we can write

$$\mathbf{A}\mathbf{y}^{(k+1)} = \mathbf{v}^{(k)}, \; k = 0, 1 \ldots$$

or
$$\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & -2/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{bmatrix} \mathbf{y}^{(k+1)} = \mathbf{v}^{(k)}$$

Writing the above system as

$$\mathbf{L}\mathbf{z}^{(k)} = \mathbf{v}^{(k)} \quad \text{and} \quad \mathbf{U}\mathbf{y}^{(k+1)} = \mathbf{z}^{(k)}$$

we obtain for $\quad \mathbf{v}^{(0)} = [1, \quad 1, \quad 1]^T, \; \mathbf{z}^{(0)} = [1, \quad 1.5, \quad 2]^T \,.\; \mathbf{y}^{(1)} = [1.5, \quad 2, \quad 1.5]^T.$

We obtain the same successive iterations as before.

2.77 Find the eigenvalue nearest to 3 for the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

using the power method. Perform five iterations. Take the initial approximate vector as $\mathbf{v}^{(0)} = [1, \quad 1, \quad 1]^T$. Also obtain the corresponding eigenvector.

Solution

The eigenvalue of $\mathbf{A}$ which is nearest to 3 is the smallest eigenvalue (in magnitude) of $\mathbf{A} - 3\mathbf{I}$. Hence it is the largest eigenvalue (in magnitude) of $(\mathbf{A} - 3\mathbf{I})^{-1}$. We have

$$\mathbf{A} - 3\mathbf{I} = \begin{bmatrix} -1 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -1 \end{bmatrix}, \quad (\mathbf{A} - 3\,\mathbf{I})^{-1} = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix}.$$

Using $\mathbf{y}^{(k+1)} = (\mathbf{A} - 3\mathbf{I})^{-1}\,\mathbf{v}^{(k)}$, $k = 0, 1, \ldots$ and $\mathbf{v}^{(0)} = [1, \quad 1, \quad 1]^T$, we obtain

$$\mathbf{y}^{(1)} = [0, -1, 0]^T, \mathbf{v}^{(1)} = [0, -1, \quad 0]^T$$

$$\mathbf{y}^{(2)} = [1, \quad -1, \quad 1]^T, \mathbf{v}^{(2)} = [1, \quad -1, \quad 1]^T$$

$$\mathbf{y}^{(3)} = [2, \quad -3, \quad 2]^T, \mathbf{v}^{(3)} = [0.6667, \quad -1, \quad 0.6667]^T$$

$$\mathbf{y}^{(4)} = [1.6667, \quad -2.3334, \quad 1.6667]^T$$

$$\mathbf{v}^{(4)} = [0.7143, \quad -1, \quad 0.7143]^T$$

$$\mathbf{y}^{(5)} = [1.7143, \quad -2.4286, \quad 1.7143]^T.$$

After five iterations, we obtain the ratios as

$$\mu = \frac{[\mathbf{y}^{(5)}]_r}{[\mathbf{v}^{(4)}]_r} = [2.4000, \quad 2.43, \quad 2.4000].$$

Therefore, $\mu = 2.4$ and $\lambda = 3 \pm (1/\mu) = 3 \pm 0.42$. Since $\lambda = 2.58$ does not satisfy $|\,\mathbf{A} - 2.58\,\mathbf{I}\,| = 0$, the correct eigenvalue nearest to 3 is 3.42 and the corresponding eigenvector is $[0.7143, -1, 0.7143]^T$. The exact eigenvalues of $\mathbf{A}$ are $2 + \sqrt{2} = 3.42$, 2 and $2 - \sqrt{2} \approx 0.59$.

Chapter 3

Interpolation and Approximation

3.1 INTRODUCTION

We know that for a function $f(x)$ that has continuous derivatives upto and including the $(n + 1)$st order, the Taylor formula in the neighbourhood of the point $x = x_0$, $x_0 \in [a, b]$ may be written as

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0)$$

$$+ \ldots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) + R_{n+1}(x) \tag{3.1}$$

where the remainder term $R_{n+1}(x)$ is of the form

$$R_{n+1}(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \; x_0 < \xi < x. \tag{3.2}$$

Neglecting $R_{n+1}(x)$ in (3.1), we obtain a polynomial of degree n :

$$P(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0)$$

$$+ \ldots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0). \tag{3.3}$$

The polynomial $P(x)$ may be called an *interpolating polynomial* satisfying the $(n + 1)$ conditions

$$f^{(v)}(x_0) = P^{(v)}(x_0), \; v = 0, 1, 2, \ldots, n \tag{3.4}$$

which are called the *interpolating conditions*. The conditions (3.4) may be replaced by more general conditions such as the values of $P(x)$ and / or its certain order derivatives coincide with the corresponding values of $f(x)$ and the same order derivatives, at one or more distinct tabular points, $a \le x_0 < x_1 < \ldots < x_{n-1} < x_n \le b$. In general, the deviation or remainder due to replacement of a function $f(x)$ by another function $P(x)$ may be written as

$$E(f, x) = f(x) - P(x). \tag{3.5}$$

In *approximation,* we measure the deviation of the given function $f(x)$ from the approximating function $P(x)$ for all values of $x \in [a, b]$.

We now give a few methods for constructing the interpolating polynomials and approximating functions for a given function $f(x)$.

Taylor Series Interpolation

If the polynomial $P(x)$ is written as the Taylor's expansion, for the function $f(x)$ about a point x_0, $x_0 \in [a, b]$, in the form

$$P(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(x_0) + \ldots + \frac{1}{n!}(x - x_0)^n f^{(n)}(x_0)$$

then, $P(x)$ may be regarded as an interpolating polynomial of degree n, satisfying the conditions

$$P^{(k)}(x_0) = f^{(k)}(x_0),\ k = 0, 1, \ldots n.$$

The term

$$R_{n+1} = \frac{1}{(n+1)!}(x - x_0)^{n+1} f^{(n+1)}(\xi),\ x_0 < \xi < x$$

which has been neglected in the Taylor expansion is called the **remainder** or the **truncation error**.

The number of terms to be included in the Taylor expansion may be determined by the acceptable error. If this error is $\varepsilon > 0$ and the series is truncated at the term $f^{(n)}(x_0)$, then, we can write

$$\frac{1}{(n+1)!} \mid x - x_0 \mid^{n+1} \mid f^{(n+1)}(\xi) \mid \le \varepsilon$$

or

$$\frac{1}{(n+1)!} \mid x - x_0 \mid^{n+1} M_{n+1} \le \varepsilon$$

where, $\quad M_{n+1} = \max\limits_{a \le x \le b} \mid f^{(n+1)}(x) \mid.$

Assume that the value of M_{n+1} or its estimate is available.

For a given ε and x, we can determine n, and if n and x are prescribed, we can determine ε. When both n and ε are given, we can find an upper bound on $(x - x_0)$, that is, it will give an interval about x_0 in which this Taylor's polynomial approximates $f(x)$ to the prescribed accuracy.

3.2 LAGRANGE AND NEWTON INTERPOLATIONS

Given the values of a function $f(x)$ at $n + 1$ distinct points $x_0, x_1, \ldots, x_n$, such that $x_0 < x_1 < x_2 < \ldots < x_n$, we determine a unique polynomial $P(x)$ of degree n which satisfies the conditions

$$P(x_i) = f(x_i),\ i = 0, 1, 2, \ldots, n. \tag{3.6}$$

Lagrange Interpolation

The maximum degree of the polynomial satisfying the $n + 1$ conditions (3.6) will be n. We assume the polynomial $P(x)$ in the form

$$P(x) = l_0(x) f(x_0) + l_1(x) f(x_1) + \ldots + l_n(x) f(x_n) \tag{3.7}$$

where $l_i(x)$, $0 \le i \le n$ are polynomials of degree n. The polynomials (3.7) will satisfy the interpolating conditions (3.6) if and only if

$$l_i(x_j) = \left| \begin{array}{ll} 0, & i \ne j, \\ 1, & i = j. \end{array} \right. \tag{3.8}$$

The polynomial $l_i(x)$ satisfying the conditions (3.8) can be written as

$$l_i(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)} \tag{3.9}$$

or
$$l_i(x) = \frac{w(x)}{(x - x_i)w'(x_i)}$$

where
$$w(x) = (x - x_0)\,(x - x_1)(x - x_n).$$

The functions $l_i(x)$, $i = 0(1)n$ are called the *Lagrange fundamental polynomials* and (3.7) is the *Lagrange interpolation* polynomial.

The truncation error in the Lagrange interpolation is given by

$$E_n(f\,;x) = f(x) - P(x).$$

Since $E_n(f; x) = 0$ at $x = x_i$, $i = 0, 1, \ldots, n$, then for $x \in [a, b]$ and $x \neq x_i$, we define a function $g(t)$ as

$$g(t) = f(t) - P(t) - [f(x) - P(x)]\,\frac{(t - x_0)(t - x_1)\ldots(t - x_n)}{(x - x_0)(x - x_1)\ldots(x - x_n)}.$$

We observe that $g(t) = 0$ at $t = x$ and $t = x_i$, $i = 0, 1, \ldots, n$.

Applying the Rolle's theorem repeatedly for $g(t)$, $g'(t)$, ..., and $g^{(n)}(t)$, we obtain $g^{(n+1)}(\xi) = 0$ where ξ is some point such that

$$\min(x_0, x_1, \ldots, x_n, x) < \xi < \max(x_0, x_1, \ldots, x_n, x).$$

Differentiating $g(t)$, $n + 1$ times with respect to t, we get

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{(n+1)!\,[f(x) - P(x)]}{(x - x_0)(x - x_1)\ldots(x - x_n)}.$$

Setting $g^{(n+1)}(\xi) = 0$ and solving for $f(x)$, we get

$$f(x) = P(x) + \frac{w(x)}{(n+1)!}\,f^{(n+1)}(\xi).$$

Hence, the truncation error in Lagrange interpolation is given by

$$E_n(f\,;x) = \frac{w(x)}{(n+1)!}\,f^{(n+1)}(\xi) \tag{3.10}$$

where $\min(x_0, x_1, \ldots, x_n, x) < \xi < \max(x_0, x_1, \ldots, x_n, x)$.

Iterated Interpolation

The iterated form of the Lagrange interpolation can be written as

$$I_{0,1,2,\ldots,n}(x) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} I_{0,1,\ldots,n-1}(x) & x_{n-1} - x \\ I_{0,1,\ldots,n-2,n}(x) & x_n - x \end{vmatrix} \tag{3.11}$$

The interpolating polynomials appearing on the right side of (3.11) are any two independent $(n - 1)$th degree polynomials which could be constructed in a number of ways. In the *Aitken* method, we construct the successive iterated polynomials as follows :

$$I_0(x) = f(x_0),\; I_1(x) = f(x_1),$$

$$I_{0,1}(x) = \frac{1}{x_1 - x_0} \begin{vmatrix} I_0(x) & x_0 - x \\ I_1(x) & x_1 - x \end{vmatrix}$$

$$I_{0,1,2}(x) = \frac{1}{x_2 - x_1} \begin{vmatrix} I_{0,1}(x) & x_1 - x \\ I_{0,2}(x) & x_2 - x \end{vmatrix}$$

$$I_{0,1,2,3}(x) = \frac{1}{x_3 - x_2} \begin{vmatrix} I_{0,1,2}(x) & x_2 - x \\ I_{0,1,3}(x) & x_3 - x \end{vmatrix}$$

This interpolation is identical with the Lagrange interpolation polynomial but it is much simpler to construct.

Newton Divided Difference Interpolation

An interpolation polynomial satisfying the conditions (3.6) can also be written in the form

$$P(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \dots$$
$$+ (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x_0, x_1, \dots x_n] \tag{3.12}$$

where

$$f[x_0] = f(x_0),$$
$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$
$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0},$$
$$f[x_0, x_1, \dots x_k] = \frac{f[x_1, x_2, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0}, \tag{3.13}$$

are the zeroth, first, second and kth order divided differences respectively. The polynomial (3.12) is called the *Newton divided difference interpolation* polynomial. The function $f(x)$ may be written as

$$f(x) = P(x) + R_{n+1}(x) \tag{3.14}$$

where $R_{n+1}(x)$ is the remainder.

Since $P(x)$ is a polynomial of degree $\le n$ and satisfies the conditions

$$f(x_k) = P(x_k),\ k = 0, 1, \dots, n,$$

the remainder term R_{n+1} vanishes at $x = x_k$, $k = 0(1)n$. It may be noted that the interpolation polynomial satisfying the conditions (3.6) is unique, and the polynomials given in (3.7), (3.11) and (3.12) are different forms of the same interpolation polynomial. Therefore, $P(x)$ in (3.12) must be identical with the Lagrange interpolation polynomial. Hence, we have

$$R_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x). \tag{3.15}$$

When a data item is added at the beginning or at the end of the tabular data and if it is possible to derive an interpolating polynomial by adding one more term to the previously calculated interpolating polynomial, then such an interpolating polynomial is said to possess **permanence property**. Obviously, Lagrange interpolating polynomial does not possess this property. Interpolating polynomials based on divided differences have the permanence property. If one more data item (x_{n+1}, f_{n+1}) is added to the given data (x_i, f_i), $i = 0, 1, \dots, n$, then in the case of Newton's divided difference formula, we need to add the term

$$(x - x_0)(x - x_1) \dots (x - x_n) f[x_0, x_1, \dots, x_{n+1}]$$

to the previously calculated nth degree interpolating polynomial.

3.3 GREGORY-NEWTON INTERPOLATIONS

Assume that the tabular points $x_0, x_1, \dots, x_n$ are equispaced, that is

$$x_i = x_0 + ih,\ i = 0, 1, \dots, n$$

with the step size h.

Finite Difference Operators

We define

$$E f(x_i) = f(x_i + h) \qquad \text{The shift operator}$$

$$\Delta f(x_i) = f(x_i + h) - f(x_i) \qquad \text{The forward difference operator}$$

$$\nabla f(x_i) = f(x_i) - f(x_i - h) \qquad \text{The backward difference operator}$$

$$\delta f(x_i) = f\left(x_i + \frac{h}{2}\right) - f\left(x_i - \frac{h}{2}\right) \qquad \text{The central difference operator}$$

$$\mu f(x_i) = \frac{1}{2}\left[f\left(x_i + \frac{h}{2}\right) + f\left(x_i - \frac{h}{2}\right)\right] \qquad \text{The averaging operator} \tag{3.16}$$

Repeated application of the difference operators give the following higher order differences :

$$E^n f(x_i) = f(x_i + nh)$$

$$\Delta^n f(x_i) = \Delta^{n-1} f_{i+1} - \Delta^{n-1} f_i = (E - 1)^n f_i$$

$$= \sum_{k=0}^{n} (-1)^k \frac{n!}{(n-k)!k!} f_{i+n-k}$$

$$\nabla^n f(x_i) = \nabla^{n-1} f_i - \nabla^{n-1} f_{i-1} = (1 - E^{-1})^n f_i$$

$$= \sum_{k=0}^{n} (-1)^k \frac{n!}{(n-k)!k!} f_{i-k}$$

$$\delta^n f(x_i) = \delta^{n-1} f_{i+1/2} - \delta^{n-1} f_{i-1/2} = (E^{1/2} - E^{-1/2})^n f_i$$

$$= \sum_{k=0}^{n} (-1)^k \frac{n!}{(n-k)!k!} f_{i+(n/2)-k} \tag{3.17}$$

where, $f_i = f(x_i)$.

We also have

$$f[x_0, x_1, ..., x_k] = \frac{1}{k!\,h^k} \Delta^k f_0 = \frac{1}{k!\,h^k} \nabla^k f_k. \tag{3.18}$$

$\Delta f(x) = h f'(x)$, or $f'(x) = [\Delta f(x)] / h$. Error in $f'(x) : O(h)$.

$\Delta^2 f(x) = h^2 f''(x)$, or $f''(x) = [\Delta^2 f(x)] / h^2$. Error in $f''(x) : O(h)$.

$\nabla f(x) = h f'(x)$, or $f'(x) = [\nabla f(x)] / h$. Error in $f'(x) : O(h)$.

$\nabla^2 f(x) = h^2 f''(x)$, or $f''(x) = [\nabla^2 f(x)] / h^2$. Error in $f''(x) : O(h)$.

$\delta^2 f(x) = h^2 f''(x)$, or $f''(x) = [\delta^2 f(x)] / h^2$. Error in $f''(x) : O(h^2)$.

$$f[x_0, x_1] = \frac{1}{h} [\Delta f_0] = f_0',$$

$$f[x_0, x_1, x_2] = \frac{1}{2!\,h^2} \Delta^2 f_0 = \frac{1}{2} f_0''.$$

The results can be generalized to higher order derivations.

Table: Relationship among the operaters

	E	Δ	∇	δ
E	E	$\Delta + 1$	$(1-\nabla)^{-1}$	$1+\frac{1}{2}\delta^2+\delta\sqrt{\left(1+\frac{1}{4}\delta^2\right)}$
Δ	$E-1$	Δ	$(1-\nabla)^{-1}-1$	$\frac{1}{2}\delta^2+\delta\sqrt{\left(1+\frac{1}{4}\delta^2\right)}$
∇	$1-E^{-1}$	$1-(1+\Delta)^{-1}$	∇	$-\frac{1}{2}\delta^2+\delta\sqrt{\left(1+\frac{1}{4}\delta^2\right)}$
δ	$E^{1/2}-E^{-1/2}$	$\Delta(1+\Delta)^{-1/2}$	$\nabla(1-\nabla)^{-1/2}$	δ
μ	$\frac{1}{2}(E^{1/2}+E^{-1/2})$	$\left(1+\frac{1}{2}\Delta\right)(1+\Delta)^{1/2}$	$\left(1-\frac{1}{2}\nabla\right)(1-\nabla)^{-1/2}$	$\sqrt{\left(1+\frac{1}{4}\delta^2\right)}$

Gregory-Newton Forward Difference Interpolation

Replacing the divided differences in (3.12) by the forward differences, we get

$$P(x) = f_0 + \frac{(x-x_0)}{h}\Delta f_0 + \frac{(x-x_0)(x-x_1)}{2!\,h^2}\Delta^2 f_0 + \ldots + \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{n!\,h^n}\Delta^n f_0 \tag{3.19}$$

or
$$P(x_0+hs) = f_0 + s\Delta f_0 + \frac{s(s-1)}{2!}\Delta^2 f_0 + \ldots + \frac{s(s-1)\ldots(s-n+1)}{n!}\Delta^n f_0 \tag{3.20}$$

$$= \sum_{i=0}^{n}\binom{s}{i}\Delta^i f_0$$

where, $s = (x-x_0)/h$. Note that $s > 0$.

The error of interpolation is

$$E_n(f\,;x) = \binom{s}{n+1}h^{n+1}f^{(n+1)}(\xi).$$

Gregory-Newton Backward Difference Interpolation

We have

$$P(x) = f_n + \frac{(x-x_n)}{h}\nabla f_n + \frac{(x-x_n)(x-x_{n-1})}{2!\,h^2}\nabla^2 f_n + \ldots + \frac{(x-x_n)(x-x_{n-1})\ldots(x-x_1)}{n!\,h^n}\nabla^n f_n \tag{3.21}$$

or
$$P_n(x_n+hs) = f_n + s\nabla f_n + \frac{s(s+1)}{2!}\nabla^2 f_n + \ldots + \frac{s(s+1)\ldots(s+n-1)}{n!}\nabla^n f_n \tag{3.22}$$

$$= \sum_{i=0}^{n}(-1)^i\binom{-s}{i}\nabla^i f_n$$

where, $s = (x-x_n)/h$. Note that $s < 0$.

The error of interpolation is

$$E_n(f\,;x) = (-1)^{n+1} \binom{-s}{n+1} h^{n+1} f^{n+1}(\xi).$$

3.4 HERMITE INTERPOLATION

Given the values of $f(x)$ and $f'(x)$ at the distinct points x_i, $i = 0, 1, ..., n$, $x_0 < x_1 < x_2 < ... < x_n$, we determine a unique polynomial of degree $\le 2n + 1$ which satisfies the conditions

$$P(x_i) = f_i,$$
$$P'(x_i) = f_i', \; i = 0, 1, ..., n. \tag{3.23}$$

The required polynomial is given by

$$P(x) = \sum_{i=0}^{n} A_i(x) f(x_i) + \sum_{i=0}^{n} B_i(x) f'(x_i) \tag{3.24}$$

where $A_i(x)$, $B_i(x)$ are polynomials of degree $2n + 1$, and are given by

$$A_i(x) = [1 - 2(x - x_i) l_i'(x_i)]\; l_i^2(x),$$

$$B_i(x) = (x - x_i)\; l_i^2(x),$$

and $l_i(x)$ is the Lagrange fundamental polynomial (3.9).

The error of interpolation in (3.24) is given by

$$E_{2n+1}(f\,;x) = \frac{w^2(x)}{(2n+2)!} f^{(2n+2)}(\xi),\; x_0 < \xi < x_n. \tag{3.25}$$

3.5 PIECEWISE AND SPLINE INTERPOLATION

In order to keep the degree of the interpolating polynomial small and also to obtain accurate results, we use *piecewise interpolation*. We divide the interval $[a, b]$ containing the tabular points $x_0, x_1, ..., x_n$ where $x_0 = a$ and $x_n = b$ into a number of subintervals $[x_{i-1}, x_i]$, $i = 1, 2, ..., n$ and replace the function $f(x)$ by some lower degree interpolating polynomial in each subinterval.

Piecewise Linear Interpolation

If we replace $f(x)$ on $[x_{i-1}, x_i]$ by the Lagrange linear polynomial, we obtain

$$F_1(x) = P_{i1}(x) = \frac{x - x_i}{x_{i-1} - x_i} f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} f_i, \tag{3.26}$$
$$i = 1, 2, ..., n.$$

Piecewise cubic Hermite Interpolation

Let the values of $f(x)$, $f'(x)$ be given at the points $x_0, x_1, ..., x_n$.

If we replace $f(x)$ on $[x_{i-1}, x_i]$, by the cubic Hermite interpolation polynomial, we obtain

$$F_3(x) = P_{i3}(x) = A_{i-1}(x) f_{i-1} + A_i(x) f_i + B_{i-1}(x) f_{i-1}' + B_i(x) f_i', \tag{3.27}$$
$$i = 1, 2, ..., n$$

where $$A_{i-1} = \left[1 - \frac{2(x - x_{i-1})}{x_{i-1} - x_i}\right] \frac{(x - x_i)^2}{(x_{i-1} - x_i)^2},$$

$$A_i = \left[1 - \frac{2(x - x_i)}{x_i - x_{i-1}}\right] \frac{(x - x_{i-1})^2}{(x_i - x_{i-1})^2},$$

$$B_{i-1} = \frac{(x - x_{i-1})(x - x_i)^2}{(x_{i-1} - x_i)^2},$$

$$B_i = \frac{(x - x_i)(x - x_{i-1})^2}{(x_i - x_{i-1})^2}. \tag{3.28}$$

We note that piecewise cubic Hermite interpolation requires prior knowledge of $f'(x_i)$, $i = 0, 1, ..., n$. If we only use f_i, $i = 0, 1, ..., n$, the resulting piecewise cubic polynomial will still interpolate $f(x)$ at $x_0, x_1, ..., x_n$ regardless of the choice of $m_i = f'(x_i)$, $i = 0, 1, ..., n$. Since $P_3(x)$ is twice continuously differentiable on $[a, b]$, we determine m_i's using these continuity conditions. Such an interpolation is called *spline interpolation*. We assume that the tabular points are equispaced.

Cubic Spline Interpolation (Continuity of second derivative)

We assume the continuity of the second derivative. Write (3.27) in the intervals $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$, differentiate two times with respect to x and use the continuity of second order derivatives at x_i, that is

$$\lim_{\varepsilon \to 0} F''(x_i + \varepsilon) = \lim_{\varepsilon \to 0} F''(x_i - \varepsilon). \tag{3.29}$$

We obtain,

$$\lim_{\varepsilon \to 0} F''(x_i + \varepsilon) = \frac{6}{h_{i+1}^2}(f_{i+1} - f_i) - \frac{4}{h_{i+1}} f_i' - \frac{2}{h_{i+1}} f_{i+1}' \qquad [3.29\ (i)]$$

$$\lim_{\varepsilon \to 0} F''(x_i - \varepsilon) = \frac{6}{h_i^2}(f_{i-1} - f_i) + \frac{2}{h_i} f_{i-1}' + \frac{4}{h_i} f_i'. \qquad [3.29\ (ii)]$$

Equating the right hand sides, we obtain

$$\frac{1}{h_i} f_{i-1}' + \left(\frac{2}{h_i} + \frac{2}{h_{i+1}}\right) f_i' + \frac{1}{h_{i+1}} f_{i+1}'$$
$$= -\frac{3(f_{i-1} - f_i)}{h_i^2} + \frac{3(f_{i+1} - f_i)}{h_{i+1}^2} \quad i = 1, 2, ..., n - 1. \qquad [3.29\ (iii)]$$

These are $n - 1$ equations in $n + 1$ unknowns $f_0', f_1', ..., f_n'$. If f_0'' and f_n'' are prescribed, then from [3.29 (i)] and [3.29 (ii)] for $i = 0$ and $i = n$ respectively, we obtain

$$\frac{2}{h_1} f_0' + \frac{1}{h_1} f_1' = \frac{3(f_1 - f_0)}{h_1^2} - \frac{1}{2} f_0'' \qquad [3.29\ (iv)]$$

$$\frac{1}{h_n} f_{n-1}' + \frac{2}{h_n} f_n' = \frac{3(f_n - f_{n-1})}{h_n^2} + \frac{1}{2} f_n''. \qquad [3.29\ (v)]$$

The derivatives f_i', $i = 0, 1, ..., n$ are determined by solving the equations [3.29 (iii)] to [3.29 (v)]. If f_0' and f_n' are specified, then we determine $f_1', f_2', ..., f_{n-1}'$ from the equation [3.29 (iii)].

For equispaced points, equations [3.29 (iii)] to [3.29 (v)] become, respectively,

$$f_{i-1}' + 4f_i' + f_{i+1}' = \frac{3}{h}(f_{i+1} - f_{i-1}), \; i = 1, 2, ..., n - 1 \tag{3.30}$$

$$2f_0' + f_1' = \frac{3}{h}(f_1 - f_0) - \frac{h}{2} f_0'' \qquad [3.31\ (i)]$$

$$f_{n-1}' + 2f_n' = \frac{3}{h}(f_n - f_{n-1}) + \frac{h}{2} f_n'' \qquad [3.31\ (ii)]$$

where $\quad x_i - x_{i-1} = h,\ i = 1(1)n.$

The above procedure gives the values of f_i', $i = 0, 1, ..., n$. Substituting f_i and f_i', $i = 0, 1, ..., n$ in the piecewise cubic Hermite interpolating polynomial (3.27), we obtain the required cubic spline interpolation. It may be noted that we need to solve only an $(n - 1) \times (n - 1)$ or an $(n + 1) \times (n + 1)$ tridiagonal system of equations for the solution of f_i'. This method is computationally much less expensive than the direct method.

Cubic Spline Interpolation (Continuity of first derivative)

We assume the continuity of the first derivative. Since $F(x)$ is a cubic polynomial, $F''(x)$ must be a linear function. We write $F''(x)$ on $[x_{i-1}, x_i]$ in the form

$$F''(x) = \frac{x_i - x}{x_i - x_{i-1}} F''(x_{i-1}) + \frac{(x - x_{i-1})}{x_i - x_{i-1}} F''(x_i). \qquad (3.32)$$

Integrating (3.32) two times with respect to x, we get

$$F(x) = \frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i + c_1 x + c_2 \qquad (3.33)$$

where $M_i = F''(x_i)$ and c_1 and c_2 are arbitrary constants to be determined by using the conditions $F(x_{i-1}) = f(x_{i-1})$ and $F(x_i) = f(x_i)$. We obtain

$$\begin{aligned}
F(x) &= \frac{1}{6h_i}(x_i - x)^3 M_{i-1} + \frac{1}{6h_i}(x - x_{i-1})^3 M_i + \frac{x}{h_i}(f_i - f_{i-1}) \\
&\quad - \frac{x}{6}(M_i - M_{i-1})h_i + \frac{1}{h_i}(x_i f_{i-1} - x_{i-1} f_i) - \frac{1}{6}(x_i M_{i-1} - x_{i-1} M_i)h_i \\
&= \frac{1}{6h_i}[(x_i - x)\{(x_i - x)^2 - h_i^2\}] M_{i-1} \\
&\quad + \frac{1}{6h_i}[(x - x_{i-1})\{(x - x_{i-1})^2 - h_i^2\}] M_i \\
&\quad + \frac{1}{h_i}(x_i - x) f_{i-1} + \frac{1}{h_i}(x - x_{i-1}) f_i \qquad (3.34)
\end{aligned}$$

where, $\quad x_{i-1} \le x \le x_i$.

Now, we require that the derivative $F'(x)$ be continuous at $x = x_i \pm \varepsilon$ as $\varepsilon \to 0$. Letting $F'(x_i - \varepsilon) = F'(x_i + \varepsilon)$ as $\varepsilon \to 0$, we get

$$\frac{h_i}{6} M_{i-1} + \frac{h_i}{3} M_i + \frac{1}{h_i}(f_i - f_{i-1}) = -\frac{h_{i+1}}{3} M_i - \frac{h_{i+1}}{6} M_{i+1} + \frac{1}{h_{i+1}}(f_{i+1} - f_i)$$

which may be written as

$$\frac{h_i}{6} M_{i-1} + \frac{h_i + h_{i+1}}{3} M_i + \frac{h_{i+1}}{6} M_{i+1} = \frac{1}{h_{i+1}}(f_{i+1} - f_i) - \frac{1}{h_i}(f_i - f_{i-1}),\ i = 1, 2, ..., n - 1. \qquad (3.35)$$

For equispaced knots $h_i = h$ for all i, equations (3.34) and (3.35) reduce to

$$F(x) = \frac{1}{6h}\,[(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i] + \frac{1}{h}(x_i - x)\left(f_{i-1} - \frac{h^2}{6} M_{i-1}\right)$$

$$+ \frac{1}{h}(x - x_{i-1})\left(f_i - \frac{h^2}{6} M_i\right) \qquad \text{...[3.36 } (i)]$$

and $$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}\,(f_{i+1} - 2f_i + f_{i-1}). \qquad \text{...[3.36 } (ii)]$$

This gives a system of $n - 1$ linear equations in $n + 1$ unknowns $M_0, M_1, ..., M_n$. The two additional conditions may be taken in one of the following forms.

(*i*) $M_0 = M_n = 0$. (**natural spline**)

(*ii*) $M_0 = M_n$, $M_1 = M_{n+1}$, $f_0 = f_n$, $f_1 = f_{n+1}$, $h_1 = h_{n+1}$.

(A spline satisfying these conditions is called a **Periodic spline**)

(*iii*) For a **non-periodic spline**, we use the conditions

$$F'(a) = f'(a) = f_0' \text{ and } F'(b) = f'(b) = f_n'.$$

For $i = 0$ and $i = n$, we get

$$2M_0 + M_1 = \frac{6}{h_1}\left(\frac{f_1 - f_0}{h_1} - f_0'\right)$$

$$M_{n-1} + 2M_n = \frac{6}{h_n}\left(f_n' - \frac{f_n - f_{n-1}}{h_n}\right). \qquad (3.37)$$

This method gives the values of $M_i = f''(x_i)$, $i = 1, 2, ..., N - 1$, while in method 1, we were determining $f'(x_i)$. The solutions obtained for M_i, $i = 1, 2, ..., N - 1$ are substituted in (3.34) or [3.36 (*i*)] to find the cubic spline interpolation. It may be noted that in this method also, we need to solve only an $(n - 1) \times (n - 1)$ tridiagonal system of equations for finding M_i.

Splines usually provide a better approximation of the behaviour of functions that have abrupt local changes. Further, splines perform better than higher order polynomial approximations.

3.6 BIVARIATE INTERPOLATION

Lagrange Bivariate Interpolation

If the values of the function $f(x, y)$ at $(m + 1)(n + 1)$ distinct point (x_i, y_j), $i = 0(1)m, j = 0(1)n$ are given, then the polynomial $P(x, y)$ of degree atmost m in x and n in y which satisfies the conditions

$$P(x_i, y_j) = f(x_i, y_j) = f_{i,j}, \; i = 0(1)m, \; j = 0(1)n$$

is given by $$P_{m,n}(x, y) = \sum_{j=0}^{n} \sum_{i=0}^{m} X_{m,i}(x) Y_{n,j}(y) f_{i,j} \qquad (3.38)$$

where $$X_{m,i}(x) = \frac{w(x)}{(x - x_i)\, w'(x_i)}, \; Y_{n,j}(y) = \frac{w^*(y)}{(y - y_j)\, w'^*(y_j)}$$

and
$$w(x) = (x - x_0)(x - x_1) \ldots (x - x_m)$$
$$w^*(y) = (y - y_0)(y - y_1) \ldots (y - y_n).$$

Newton's Bivariate Interpolation for Equispaced Points

With equispaced points, with spacing h in x and k in y, we define

$$\Delta_x f(x, y) = f(x + h, y) - f(x, y) = (E_x - 1) f(x, y)$$
$$\Delta_y f(x, y) = f(x, y + k) - f(x, y) = (E_y - 1) f(x, y)$$
$$\Delta_{xx} f(x, y) = \Delta_x f(x + h, y) - \Delta_x f(x, y) = (E_x - 1)^2 f(x, y)$$
$$\Delta_{yy} f(x, y) = \Delta_y f(x, y + k) - \Delta_y f(x, y) = (E_y - 1)^2 f(x, y)$$
$$\begin{aligned}\Delta_{xy} f(x, y) &= \Delta_x [f(x, y + k) - f(x, y)] = \Delta_x \Delta_y f(x, y)\\ &= (E_x - 1)(E_y - 1) f(x, y) = (E_y - 1)(E_x - 1) f(x, y)\\ &= \Delta_y \Delta_x f(x, y) = \Delta_{yx} f(x, y)\end{aligned}$$

Now, $f(x_0 + mh, y_0 + nk) = E_x^m \; E_y^n \; f(x_0, y_0) = (1 + \Delta_x)^m (1 + \Delta_y)^n f(x_0, y_0)$

$$= \left[1 + \binom{m}{1}\Delta_x + \binom{m}{2}\Delta_{xx} + \ldots\right]\left[1 + \binom{n}{1}\Delta_y + \binom{n}{2}\Delta_{yy} + \ldots\right] f(x_0, y_0)$$

$$= \left[1 + \binom{m}{1}\Delta_x + \binom{n}{1}\Delta_y + \binom{m}{2}\Delta_{xx} + \binom{m}{1}\binom{n}{1}\Delta_{xy} + \binom{n}{2}\Delta_{yy} + \ldots\right] f(x_0, y_0) \qquad [3.39\ (i)]$$

Let $x = x_0 + mh$ and $y = y_0 + nk$. Hence, $m = (x - x_0) \,/\, h$ and $n = (y - y_0) \,/\, k$. Then, from [3.39 (*i*)] we have the interpolating polynomial

$$P(x, y) = f(x_0, y_0) + \left[\frac{1}{h}(x - x_0)\,\Delta_x + \frac{1}{k}(y - y_0)\,\Delta_y\right] f(x_0, y_0)$$

$$+ \frac{1}{2!}\left[\frac{1}{h^2}(x - x_0)(x - x_1)\,\Delta_{xx} + \frac{2}{hk}(x - x_0)(y - y_0)\,\Delta_{xy} + \frac{1}{k^2}(y - y_0)(y - y_1)\,\Delta_{yy}\right] f(x_0, y_0) + \ldots \qquad [3.39\ (ii)]$$

This is called the **Newton's bivariate interpolating polynomial** for equispaced points.

3.7 APPROXIMATION

We approximate a given continuous function $f(x)$ on $[a, b]$ by an expression of the form

$$f(x) \approx P(x, c_0, c_1, \ldots, c_n) = \sum_{i=0}^{n} c_i\, \phi_i(x) \qquad (3.40)$$

where $\phi_i(x)$, $i = 0, 1, \ldots, n$ are $n + 1$ appropriately chosen linearly independent functions and $c_0, c_1, \ldots, c_n$ are parameters to be determined such that

$$E(f\,; c) = \left\| f(x) - \sum_{i=0}^{n} c_i \phi_i(x) \right\| \qquad (3.41)$$

is minimum, where $\|\,.\,\|$ is a well defined norm. By using different norms, we obtain different types of approximations. Once a particular norm is chosen, the function which minimizes the error norm (3.41) is called the *best approximation*. The functions $\phi_i(x)$ are called *coordinate functions* and are usually taken as $\phi_i(x) = x^i$, $i = 0(1)n$ for polynomial approximation.

Least Squares Approximation

We determine the parameters $c_0, c_1, ..., c_n$ such that

$$I(c_0, c_1, ..., c_n) = \sum_{k=0}^{N} W(x_k)\left[f(x_k) - \sum_{i=0}^{n} c_i \phi_i(x_k)\right]^2 = \text{minimum}. \tag{3.42}$$

Here, the values of $f(x)$ are given at $N + 1$ distinct points $x_0, x_1, ..., x_N$.

For functions which are continuous on $[a, b]$, we determine $c_0, c_1, ..., c_n$ such that

$$I(c_0, c_1, ..., c_n) = \int_a^b W(x)\left[f(x) - \sum_{i=0}^{n} c_i \phi_i(x)\right]^2 dx = \text{minimum} \tag{3.43}$$

where $W(x) > 0$ is the weight function.

The necessary conditions for (3.42) or (3.43) to have a minimum value is that

$$\frac{\partial I}{\partial c_i} = 0, \quad i = 0, 1, ..., n \tag{3.44}$$

which give a system of $(n + 1)$ linear equations in $(n + 1)$ unknowns $c_0, c_1, ..., c_n$ in the form

$$\sum_{k=0}^{N} W(x_k)\left[f(x_k) - \sum_{i=0}^{n} c_i \phi_i(x_k)\right] \phi_j(x_k) = 0 \tag{3.45}$$

or

$$\int_a^b W(x)\left[f(x) - \sum_{i=0}^{n} c_i \phi_i(x)\right] \phi_j(x) dx = 0, \tag{3.46}$$

$$j = 0, 1, ..., n.$$

The equations (3.45) or (3.46) are called the *normal equations*. For large n, normal equations become ill-conditioned, when $\phi_i(x) = x^i$. This difficulty can be avoided if the functions $\phi_i(x)$ are so chosen that they are *orthogonal* with respect to the weight function $W(x)$ over $[a, b]$ that is

$$\sum_{k=0}^{N} W(x_k)\phi_i(x_k)\, \phi_j(x_k) = 0, \quad i \neq j, \tag{3.47}$$

or

$$\int_a^b W(x)\, \phi_i(x)\, \phi_j(x) dx = 0, \quad i \neq j. \tag{3.48}$$

If the functions $\phi_i(x)$ are orthogonal, then we obtain from (3.45)

$$c_i = \frac{\sum_{k=0}^{N} W(x_k)\phi_i(x_k) f(x_k)}{\sum_{k=0}^{N} W(x_k)\phi_i^2(x_k)}, \; i = 0, 1, 2, ..., n \tag{3.49}$$

and from (3.46), we obtain

$$c_i = \frac{\int_a^b W(x)\phi_i(x) f(x) dx}{\int_a^b W(x)\phi_i^2(x) dx}, \; i = 0, 1, 2, ..., n. \tag{3.50}$$

Gram-Schmidt Orthogonalizing Process

Given the polynomials $\phi_i(x)$, the polynomials $\phi_i^*(x)$ of degree i which are orthogonal on $[a, b]$ with respect to the weight function $W(x)$ can be generated recursively.

We have, $\phi_0^*(x) = 1$

$$\phi_i^*(x) = x^i - \sum_{r=0}^{i-1} a_{ir}\phi_r^*(x) \tag{3.51}$$

where
$$a_{ir} = \frac{\int_a^b W(x)x^i\phi_r^*(x)dx}{\int_a^b W(x)(\phi_r^*(x))^2 dx}, i = 0, 1, 2, ..., n. \tag{3.52}$$

Over a discrete set of points, we replace the integrals by summation in (3.52).

Uniform (minimax) Polynomial Approximation

Taking the approximating polynomials for a continuous function $f(x)$ on $[a, b]$ in the form

$$P_n(x) = c_0 + c_1x + ... + c_nx^n \tag{3.53}$$

we determine $c_0, c_1, ..., c_n$ such that the deviation

$$E_n(f, c_0, c_1, ..., c_n) = f(x) - P_n(x) \tag{3.54}$$

satisfies the condition

$$\max_{a \le x \le b} |E_n(f, c_0, c_1, ..., c_n) = \min_{a \le x \le b} |E_n(f, c_0, c_1, ..., c_n)|. \tag{3.55}$$

If we denote

$$\varepsilon_n(x) = f(x) - P_n(x),$$
$$E_n(f, x) = \max_{a \le x \le b} | \varepsilon_n(x) |,$$

then there are atleast $n + 2$ points $a = x_0 < x_1 < x_2 ... < x_n < x_{n+1} = b$ where (*Chebyshev equi-oscillation theorem*)

(*i*) $\varepsilon(x_i) = \pm E_n$, $i = 0, 1, ..., n + 1$, [3.56 (*i*)]

(*ii*) $\varepsilon(x_i) = - \varepsilon(x_{i+1})$, $i = 0, 1, ..., n$. [3.56 (*ii*)]

The best uniform (minimax) polynomial approximation is uniquely determined under the conditions (3.56). It may be observed that [3.56 (*ii*)] implies that

$$\varepsilon'(x_i) = 0, \quad i = 1, 2, ..., n. \qquad [3.56\ (iii)]$$

For finding the best uniform approximation it is sufficient to use [3.56 (*ii*)] and [3.56 (*iii*)].

Chebyshev Polynomials

The Chebyshev polynomials of the first kind $T_n(x)$ defined on $[-1, 1]$ are given by

$$T_n(x) = \cos(n \cos^{-1} x) = \cos n\theta$$

where $\theta = \cos^{-1} x \quad$ or $\quad x = \cos \theta$.

These polynomials satisfy the differential equation

$$(1 - x^2) y'' - x y' + n^2 y = 0.$$

One independent solution gives $T_n(x)$ and the second independent solution is given by $u_n(x) = \sin(n\theta) = \sin(n \cos^{-1} x)$. We note that $u_n(x)$ is not a polynomial. The Chebyshev polynomials of second kind, denoted by $U_n(x)$ are defined by

$$U_n(x) = \frac{\sin [(n+1)\theta]}{\sin \theta} = \frac{\sin [(n+1) \cos^{-1} x]}{\sqrt{1-x^2}}$$

Note that $U_n(x)$ is a polynomial of degree n.

The Chebyshev polynomials $T_n(x)$ satisfy the recurrence relation

$$T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x)$$

$$T_0(x) = 1,\; T_1(x) = x.$$

Thus, we have

$$\begin{aligned}
T_0(x) &= 1, & 1 &= T_0(x),\\
T_1(x) &= x, & x &= T_1(x),\\
T_2(x) &= 2x^2 - 1, & x^2 &= [T_2(x) + T_0(x)] / 2,\\
T_3(x) &= 4x^3 - 3x, & x^3 &= [T_3(x) + 3T_1(x)] / 2^2,\\
T_4(x) &= 8x^4 - 8x^2 + 1, & x^4 &= [T_4(x) + 4T_2(x) + 3T_0(x)] / 2^3.
\end{aligned}$$

We also have

$$\begin{aligned}
T_n(x) &= \cos n\theta = \text{real part } (e^{in\theta}) = Re\, (\cos\theta + i \sin\theta)^n\\
&= Re \left[\cos^n \theta + \binom{n}{1} \cos^{n-1} \theta\, (i \sin\theta) + \binom{n}{2} \cos^{n-2} \theta\, (i \sin \theta)^2 + \ldots \right]\\
&= x^n + \binom{n}{2} x^{n-2} (x^2 - 1) + \binom{n}{4} x^{n-4} (x^2 - 1)^2 + \ldots\\
&= 2^{n-1} x^n + \text{terms of lower degree.}
\end{aligned}$$

The Chebyshev polynomials $T_n(x)$ possess the following properties :

(*i*) $T_n(x)$ is a polynomial of degree n. If n is even, $T_n(x)$ is an even polynomial and if n is odd, $T_n(x)$ is an odd polynomial.

(*ii*) $T_n(x)$ has n simple zeros $x_k = \cos\left(\frac{2k-1}{2n}\pi\right)$, $k = 1, 2, \ldots, n$ on the interval $[-1, 1]$.

(*iii*) $T_n(x)$ assumes extreme values at $n + 1$ points $x_k = \cos (k\pi / n)$, $k = 0, 1, \ldots, n$ and the extreme value at x_k is $(-1)^k$.

(*iv*) $|\, T_n(x)\, | \le 1$, $x \in [-1, 1]$.

(*v*) If $P_n(x)$ is any polynomial of degree n with leading coefficient unity (monic polynomial) and $\tilde{T}_n(x) = T_n(x) / 2^{n-1}$ is the monic Chebyshev polynomial, then

$$\max_{-1\le x\le 1} |\tilde{T}_n(x)| \le \max_{-1\le x\le 1} |P_n(x)|.$$

This property is called the **minimax property.**

(*vi*) $T_n(x)$ is orthogonal with respect to the weight function

$$W(x) = \frac{1}{\sqrt{1-x^2}}, \text{ and}$$

$$\int_{-1}^{1} \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}}\, dx = \begin{vmatrix} 0, & m \ne n\\ \dfrac{\pi}{2}, & m = n \ne 0\\ \pi, & m = n = 0. \end{vmatrix}$$

Chebyshev Polynomial Approximation and Lanczos Economization

Let the Chebyshev series expansion of $f(x) \in C[-1, 1]$ be

$$f(x) = \frac{a_0}{2} + \sum_{i=1}^{\infty} a_i T_i(x).$$

Then, the partial sum

$$P_n(x) = \frac{a_0}{2} + \sum_{i=0}^{n} a_i T_i(x) \tag{3.57}$$

is very nearly the solution of the mini-max problem

$$\max_{-1 \le x \le 1} \left| f(x) - \sum_{i=0}^{n} c_i x^i \right| = \text{minimum}.$$

To obtain the approximating polynomial $P_n(x)$, we follow the following steps :

1. Transform the interval $[a, b]$ to $[-1, 1]$ by using the linear transformation $x = [(b - a)t + (b + a)] / 2$, and obtain the new function $f(t)$ defined on $[-1, 1]$.
2. Obtain the power series expansion of $f(t)$ on $[-1, 1]$. Writing each term t^i in terms of Chebyshev polynomials we obtain

$$f(t) = \sum_{i=0}^{\infty} c_i T_i(t).$$

The partial sum

$$P_n(t) = \sum_{i=0}^{n} c_i T_i(t) \tag{3.58}$$

is a good uniform approximation to $f(t)$ in the sense

$$\max_{-1 \le t \le 1} | f(t) - P_n(t) | \le | c_{n+1} | + | c_{n+2} | + \ldots \le \varepsilon. \tag{3.59}$$

Given ε, it is possible to find the number of terms to be retained in (3.59).

This procedure is called the *Lanczos economization*. Replacing each $T_i(t)$ by its polynomial form, we obtain $P_n(t)$. Writing t in terms of x, we obtain the required economized Chebyshev polynomial approximation to $f(x)$ on $[a, b]$.

3.8 PROBLEMS AND SOLUTIONS

Taylor Series Interpolation

3.1 Obtain a second degree polynomial approximation to $f(x) = (1 + x)^{1/2}$ over $[0, 1]$ by means of the Taylor expansion about $x = 0$. Use the first three terms of the expansion to approximate $f(0.05)$. Obtain a bound of the error in the interval $[0, 1]$.

Solution

The Taylor expansion of $f(x) = (1 + x)^{1/2}$ about $x = 0$ is obtained as

$$f(x) = f(0) + x f'(0) + \frac{1}{2!} x^2 f''(0) + \frac{1}{3!} x^3 f'''(0) + \ldots = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \ldots$$

Taking terms upto x^2, we obtain the approximation

$$f(x) = P(x) = 1 + \frac{x}{2} - \frac{x^2}{8}.$$

We have $f(0.05) \approx P(0.05) = 1.0246875.$

The error of approximation is given by

$$\text{TE} = \frac{x^3}{3!} f'''(\xi).$$

Hence,

$$|\,\text{TE}\,| \le \max_{0 \le x \le 1} \left| \frac{x^3}{6} \right| \max_{0 \le x \le 1} |f'''(x)|$$

$$= \frac{1}{6} \max_{0 \le x \le 1} \left| \frac{3}{8(1+x)^{5/2}} \right| = \frac{1}{16} = 0.0625.$$

3.2 Expand $\ln(1 + x)$ in a Taylor expansion about $x_0 = 1$ through terms of degree 4. Obtain a bound on the truncation error when approximating ln 1.2 using this expansion.

Solution

The Taylor series expansion of $\ln(1 + x)$ about $x_0 = 1$ is obtained as

$$\ln(1 + x) = \ln 2 + \frac{1}{2}(x - 1) - \frac{1}{8}(x - 1)^2 + \frac{1}{24}(x - 1)^3 - \frac{1}{64}(x - 1)^4 + \ldots$$

Taking terms upto degree 4, we get

$$\ln(1.2) \approx 0.185414.$$

A bound on the error of approximation is given by

$$|\,\text{TE}\,| \le \max_{1 \le x \le 1.2} \left| \frac{(x-1)^5}{5!} f^{v}(x) \right|$$

$$= \max_{1 \le x \le 1.2} \left| \frac{(x-1)^5}{120} \cdot \frac{24}{(1+x)^5} \right| = 0.2 \times 10^{-5}.$$

3.3 Obtain the polynomial approximation to $f(x) = (1 - x)^{1/2}$ over [0, 1], by means of Taylor expansion about $x = 0$. Find the number of terms required in the expansion to obtain results correct to 5×10^{-3} for $0 \le x \le 1/2$.

Solution

We have the Taylor series expansion for $f(x) = (1 - x)^{1/2}$ about $x = 0$ as

$$f(x) = f(0) + x f'(0) + \ldots + \frac{x^{n-1}}{(n-1)!} f^{(n-1)}(0) + \frac{x^n}{n!} f^{(n)}(0) + \ldots$$

If we keep the first n terms, then the error of approximation is given by

$$\text{TE} = \frac{x^n}{n!} f^{(n)}(\xi),\ 0 \le \xi \le 1/2$$

where

$$f^{(n)}(x) = -\frac{1}{2}\left(\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{(2n-3)}{2}\right)(1-x)^{-(2n-1)/2}$$

$$= -\frac{(2n-2)!}{2^n (n-1)!\, 2^{n-1}} (1-x)^{-(2n-1)/2} = -\frac{(2n-2)!}{2^{2n-1}(n-1)!} \frac{1}{(1-x)^{(2n-1)/2}}$$

and

$$f^{(n)}(0) = -\frac{(2n-2)!}{2^{2n-1}(n-1)!}$$

Hence, we find n such that

$$\max_{0 \le x \le 1/2} |\text{TE}| \le \left| \frac{2^{-n}}{n!} \cdot \frac{(2n-2)!}{2^{2n-1}(n-1)!} 2^{n-\frac{1}{2}} \right| \le 0.005$$

or
$$\frac{(2n-2)!\sqrt{2}}{n!(n-1)!\,2^{2n}} \le 0.005$$

which gives $n \ge 12$. Therefore, atleast 13 terms are required in the Taylor expansion.

3.4 If we use somewhat unsuitable method of Taylor expansion around $x = 0$ for computation of $\sin x$ in the interval $[0, 2\pi]$ and if we want 4 accurate decimal places, how many terms are needed. If instead we use the fact $\sin(\pi + x) = -\sin x$, we only need the expansion in $[0, \pi]$. How many terms do we then need for the same accuracy.

(Lund Univ., Sweden, BIT 16 (1976), 228)

Solution

Taylor expansion of $\sin x$ about $x = 0$ is given by

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots + \frac{(-1)^{n-1}x^{2n-1}}{(2n-1)!}$$

with the error term

$$\text{TE} = \frac{(-1)^n x^{2n+1}}{(2n+1)!} M, 0 < \xi < x$$

where $M = \pm\cos(\xi)$ and $\max|M| = 1$.

For $x \in [0, 2\pi]$, we choose the smallest n so that

$$\max_{0 \le x \le 2\pi} \left| \frac{(-1)^n x^{2n+1}}{(2n+1)!} M \right| \le 0.00005$$

or
$$\frac{(2\pi)^{2n+1}}{(2n+1)!} \le 0.00005$$

which gives $n \ge 12$.

For $x \in [0, \pi]$, we choose the smallest n so that

$$\max_{0 \le x \le \pi} \left| \frac{(-1)^n x^{2n+1}}{(2n+1)!} M \right| \le 0.00005$$

or
$$\frac{(\pi)^{2n+1}}{(2n+1)!} \le 0.00005$$

which gives $n \ge 7$.

3.5 Determine the constants a, b, c and d such that the interpolating polynomial

$$y_s = y(x_0 + sh) = ay_0 + by_1 + h^2(cy_0'' + dy_1'')$$

becomes correct to the highest possible order.

Solution

The interpolation error is given by

$$\text{TE} = y(x_s) - a\,y(x_0) - b\,y(x_1) - h^2(cy''(x_0) + dy''(x_1))$$

Expanding each term in Taylor series about x_0, we obtain

$$\text{TE} = y_0 + shy_0' + \frac{s^2h^2}{2!}y_0'' + \frac{s^3h^3}{3!}y_0''' + \frac{s^4h^4}{4!}y_0^{iv} + \ldots$$

$$- \left[(a+b)y_0 + bhy_0' + h^2\left(\frac{b}{2} + c + d\right)y_0'' + h^3\left(\frac{b}{6} + d\right)y_0''' + h^4\left(\frac{b}{24} + \frac{d}{2}\right)y_0^{iv} + \ldots\right]$$

Setting the coefficients of various powers of h to zero, we get the system of equations

$$a + b = 1,$$
$$b = s,$$
$$\frac{b}{2} + c + d = \frac{s^2}{2},$$
$$\frac{b}{6} + d = \frac{s^3}{6},$$

which give $a = 1 - s$, $b = s$, $c = \dfrac{-s(s-1)(s-2)}{6}$, $d = \dfrac{s(s^2-1)}{6}$.

The error term is given by

$$\text{TE} = \left(\frac{s^4}{24} - \frac{b}{24} - \frac{d}{2}\right)h^4 y^{iv}(\xi) = \frac{1}{24}(s^4 - 2s^3 + s)h^4 y^{iv}(\xi).$$

3.6 Determine the constants a, b, c and d such that the interpolating polynomial

$$y(x_0 + sh) = ay(x_0 - h) + by(x_0 + h) + h[cy'(x_0 - h) + dy'(x_0 + h)]$$

becomes correct to the highest possible order. Find the error term.

Solution

The interpolating error is written as

$$\text{TE} = y(x_0 + sh) - ay(x_0 - h) - by(x_0 + h) - h[cy'(x_0 - h) + dy'(x_0 + h)].$$

Expanding each term in Taylor series about x_0, we obtain

$$\text{TE} = y_0 + shy_0' + \frac{s^2h^2}{2}y_0'' + \frac{s^3h^3}{6}y_0''' + \frac{s^4h^4}{24}y_0^{iv} + \ldots$$

$$- \left[(a+b)y_0 + h(-a+b+c+d)y_0' + \frac{h^2}{2}(a+b-2c+2d)y_0''\right.$$

$$\left. + \frac{h^3}{6}(-a+b+3c+3d)y_0''' + \frac{h^4}{24}(a+b-4c+4d)y_0^{iv} + \ldots\right].$$

Putting the coefficients of various powers of h to zero, we get the system of equations

$$a + b = 1,$$
$$-a + b + c + d = s,$$
$$a + b - 2c + 2d = s^2,$$
$$-a + b + 3c + 3d = s^3,$$

which has the solution $a = (s-1)(s^2+s-2)/4$, $b = (s+1)(2+s-s^2)/4$,
$c = (s+1)(s-1)^2/4$, $d = (s-1)(s+1)^2/4$.

The error term is given by

$$\text{TE} = \frac{h^4}{24} (s^4 - a - b + 4c - 4d)\, y^{iv}(\xi)$$
$$= \frac{1}{24} (s^4 - 2s^2 + 1) h^4 y^{iv}(\xi).$$

3.7 Determine the parameters in the formula

$$P(x) = a_0(x - a)^3 + a_1(x - a)^2 + a_2(x - a) + a_3$$

such that

$$P(a) = f(a),\ P'(a) = f'(a),$$
$$P(b) = f(b),\ P'(b) = f'(b).$$

Solution

Using the given conditions, we obtain the system of equations

$$f(a) = a_3,$$
$$f'(a) = a_2,$$
$$f(b) = a_0(b - a)^3 + a_1(b - a)^2 + a_2(b - a) + a_3,$$
$$f'(b) = 3a_0(b - a)^2 + 2a_1(b - a) + a_2,$$

which has the solution

$$a_0 = \frac{2}{(b-a)^3} [f(a) - f(b)] + \frac{1}{(b-a)^2} [f'(a) + f'(b)],$$
$$a_1 = \frac{3}{(b-a)^2} [f(b) - f(a)] - \frac{1}{(b-a)} [2f'(a) + f'(b)],$$
$$a_2 = f'(a),\ a_3 = f(a).$$

3.8 Obtain the unique polynomial $P(x)$ of degree 5 or less, approximating the function $f(x)$, where

$$f(x_0) = 1,\ f'(x_0) = 2,$$
$$f''(x_0) = 1,\ f(x_1) = 3,$$
$$f'(x_1) = 0,\ f''(x_1) = -2,\ x_1 = x_0 + h.$$

Also find $P((x_0 + x_1) / 2)$.

Solution

We take the polynomial in the form

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + a_4(x - x_0)^4 + a_5(x - x_0)^5.$$

Using the given conditions, with $h = x_1 - x_0$, we obtain the system of equations

$$a_0 = 1,\ a_1 = 2,\ a_2 = 1 / 2,$$
$$a_0 + ha_1 + h^2a_2 + h^3a_3 + h^4a_4 + h^5a_5 = 3,$$
$$a_1 + 2ha_2 + 3h^2a_3 + 4h^3a_4 + 5h^4a_5 = 0,$$
$$2a_2 + 6ha_3 + 12h^2a_4 + 20h^2a_5 = -2,$$

which has the solution

$$a_0 = 1,\ a_1 = 2,\ a_2 = 1 / 2,$$
$$a_3 = \frac{1}{2h^3} (40 - 24h - 5h^2),$$
$$a_4 = \frac{1}{2h^4} (-60 + 32h + 7h^2),$$

$$a_5 = \frac{3}{2h^5}(8 - 4h - h^2).$$

Substituting in the given polynomial, we obtain

$$P\left(\frac{x_0 + x_1}{2}\right) = \frac{1}{64}(128 + 20h - h^2).$$

Lagrange and Newton Interpolation

3.9 For the data (x_i, f_i), $i = 0, 1, 2, ..., n$, construct the Lagrange fundamental polynomials $l_i(x)$ using the information that they satisfy the conditions $l_i(x_j) = 0$, for $i \neq j$ and $= 1$ for $i = j$.

Solution

Since $l_i(x_j) = 0$ for $i \neq j$; $(x - x_0), (x - x_1), ..., (x - x_{i-1}), (x - x_{i+1}), ..., (x - x_n)$ are factors of $l_i(x)$. Now, $l_i(x)$ is a polynomial of degree n and $(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)$ is also a polynomial of degree n. Hence, We can write

$$l_i(x) = A(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n),$$

where A is a constant. Since, $l_i(x_i) = 1$, we get

$$l_i(x_i) = 1 = A(x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n).$$

This determines A. Therefore, the Lagrange fundamental polynomials are given by

$$l_i(x) = \frac{(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}.$$

3.10 Let $f(x) = \ln(1 + x)$, $x_0 = 1$ and $x_1 = 1.1$. Use linear interpolation to calculate an approximate value of $f(1.04)$ and obtain a bound on the truncation error.

Solution

We have
$$f(x) = \ln(1 + x),$$
$$f(1.0) = \ln(2) = 0.693147,$$
$$f(1.1) = \ln(2.1) = 0.741937.$$

The Lagrange interpolating polynomial is obtained as

$$P_1(x) = \frac{x - 1.1}{1.0 - 1.1}(0.693147) + \frac{x - 1}{1.1 - 1.0}(0.741937)$$

which gives $\quad P_1(1.04) = 0.712663.$

The error in linear interpolation is given by

$$\text{TE} = \frac{1}{2!}(x - x_0)(x - x_1) f''(\xi), \quad x_0 < \xi < x_1.$$

Hence, we obtain the bound on the error as

$$|\text{ TE }| \leq \frac{1}{2} \max_{1 \leq x \leq 1.1} |(x - x_0)(x - x_1)| \max_{1 \leq x \leq 1.1} |f''(x)|$$

Since the maximum of $(x - x_0)(x - x_1)$ is obtained at $x = (x_0 + x_1) / 2$ and $f''(x) = -1 / (1 + x)^2$, we get

$$|\text{ TE }| \leq \frac{1}{2} \frac{(x_1 - x_0)^2}{4} \max_{1 \leq x \leq 1.1} \left| \frac{1}{(1 + x)^2} \right|$$

$$= \frac{(0.1)^2}{8} \cdot \frac{1}{4} = 0.0003125.$$

3.11 Determine an appropriate step size to use, in the construction of a table of $f(x) = (1 + x)^6$ on [0, 1]. The truncation error for linear interpolation is to be bounded by 5×10^{-5}.

Solution

The maximum error in linear interpolation is given by $h^2 M_2 / 8$, where

$$M_2 = \max_{0 \le x \le 1} |f''(x)| = \max_{0 \le x \le 1} |30(1+x)^4| = 480.$$

We choose h so that

$$60h^2 \le 0.00005$$

which gives $\quad h \le 0.00091.$

3.12 (*a*) Show that the truncation error of quadratic interpolation in an equidistant table is bounded by

$$\left(\frac{h^3}{9\sqrt{3}}\right) \max |f'''(\xi)|.$$

(*b*) We want to set up an equidistant table of the function $f(x) = x^2 \ln x$ in the interval $5 \le x \le 10$. The function values are rounded to 5 decimals. Give the step size h which is to be used to yield a total error less than 10^{-5} on quadratic interpolation in this table. (Bergen Univ., Sweden, BIT 25 (1985), 299)

Solution

(*a*) Error in quadratic interpolation based on the points x_{i-1}, x_i and x_{i+1} is given by

$$\text{TE} = \frac{(x - x_{i-1})(x - x_i)(x - x_{i+1})}{3!} f'''(\xi), \quad x_{i-1} < \xi < x_{i+1}.$$

Writing $(x - x_i) / h = t$, we obtain

$$\text{TE} = \frac{(t-1)(t)(t+1)}{6} h^3 f'''(\xi), \quad -1 < \xi < 1.$$

The extreme values of $g(t) = (t - 1)t(t + 1) = t^3 - t$ occur at $t = \pm 1/\sqrt{3}$. Now,

$\max | g(t) | = 2/(3\sqrt{3})$. Hence,

$$| \text{TE} | \le \frac{h^3}{9\sqrt{3}} \max |f'''(\xi)|.$$

(*b*) We have $f(x) = x^2 \ln(x)$, which gives

$$f'''(x) = \frac{2}{x} \quad \text{or} \quad \max_{5 \le x \le 10} |f'''(x)| = \frac{2}{5}.$$

Hence, we choose h such that

$$\frac{h^3}{9\sqrt{3}}\left(\frac{2}{5}\right) \le 0.000005$$

which gives $\quad h \le 0.0580.$

3.13 Determine the maximum step size that can be used in the tabulation of $f(x) = e^x$ in [0, 1], so that the error in the linear interpolation will be less than 5×10^{-4}. Find also the step size if quadratic interpolation is used.

Solution

We have

$$f(x) = e^x, f^{(r)}(x) = e^x, r = 1, 2, \ldots$$

Maximum error in linear interpolation is given by

$$\frac{h^2}{8} \max_{0 \le x \le 1} |e^x| = \frac{h^2 e}{8}.$$

We choose h so that

$$\frac{h^2 e}{8} \le 0.0005,$$

which gives $\qquad h \le 0.03836.$

Maximum error in quadratic interpolation is given by

$$\frac{h^3}{9\sqrt{3}} \max_{0 \le x \le 1} |e^x| = \frac{h^3 e}{9\sqrt{3}}$$

We choose h so that

$$\frac{h^3 e}{9\sqrt{3}} \le 0.0005$$

which gives $\qquad h \le 0.1420.$

3.14 By considering the limit of the three point Lagrange interpolation formula relative to x_0, $x_0 + \varepsilon$ and x_1 as $\varepsilon \to 0$, obtain the formula

$$f(x) = \frac{(x_1 - x)(x + x_1 - 2x_0)}{(x_1 - x_0)^2} f(x_0) + \frac{(x - x_0)(x_1 - x)}{(x_1 - x_0)} f'(x_0) + \frac{(x - x_0)^2}{(x_1 - x_0)} f(x_1) + E(x)$$

where $E(x) = \frac{1}{6}(x - x_0)^2(x - x_1) f'''(\xi)$.

Solution

The Lagrange interpolating polynomial relative to the points x_0, $x_0 + \varepsilon$ and x_1 is obtained as

$$P_2(x) = \frac{(x - x_0 - \varepsilon)(x - x_1)}{-\varepsilon(x_0 - x_1)} f(x_0) + \frac{(x - x_0)(x - x_1)}{\varepsilon(x_0 - x_1 + \varepsilon)} f(x_0 + \varepsilon) + \frac{(x - x_0)(x - x_0 - \varepsilon)}{(x_1 - x_0)(x_1 - x_0 - \varepsilon)} f(x_1).$$

Taking the limit as $\varepsilon \to 0$, we get

$$\lim_{\varepsilon \to 0} P_2(x) = \frac{(x - x_0)^2}{(x_1 - x_0)^2} f(x_1) + \lim_{\varepsilon \to 0} \left[-\frac{(x - x_1)(x - x_0 - \varepsilon)}{\varepsilon(x_0 - x_1)} f(x_0) + \frac{(x - x_0)(x - x_1)}{\varepsilon(x_0 - x_1 + \varepsilon)} (f(x_0) + \varepsilon f'(x_0) + O(\varepsilon^2)) \right]$$

$$= \frac{(x - x_0)^2}{(x_1 - x_0)^2} f(x_1) + \lim_{\varepsilon \to 0} \left[\frac{x - x_0}{\varepsilon(x_0 - x_1 + \varepsilon)} - \frac{x - x_0 - \varepsilon}{\varepsilon(x_0 - x_1)} \right] (x - x_1) f(x_0) + \frac{(x - x_0)(x - x_1)}{x_0 - x_1} f'(x_0)$$

$$= \frac{(x_1 - x)(x + x_1 - 2x_0)}{(x_1 - x_0)^2} f(x_0) + \frac{(x - x_0)(x - x_1)}{(x_0 - x_1)} f'(x_0) + \frac{(x - x_0)^2}{(x_1 - x_0)^2} f(x_1)$$

The error in quadratic interpolation is given by

$$\text{TE} = \frac{(x - x_0)(x - x_0 - \varepsilon)(x - x_1)}{3\,!} f'''(\xi)$$

which in the limit as $\varepsilon \to 0$ becomes

$$\text{TE} = \frac{(x - x_0)^2 (x - x_1)}{3\,!} f'''(\xi).$$

3.15 Denoting the interpolant of $f(x)$ on the set of (distinct) points $x_0, x_1, ..., x_n$ by $\sum_{k=0}^{n} l_k(x) f(x_k)$, find an expression for $\sum_{k=0}^{n} l_k(0) x_k^{n+1}$. (Gothenburg Univ., Sweden, BIT 15 (1975), 224)

Solution

We have

$$f(x) = \sum_{k=0}^{n} l_k(x) f(x_k) + \frac{(x - x_0)(x - x_1) \ldots (x - x_n)}{(n+1)\,!} f^{(n+1)}(\xi).$$

Letting $f(x) = x^{n+1}$, we get

$$x^{n+1} = \sum_{k=0}^{n} l_k(x) x_k^{n+1} + (x - x_0) \ldots (x - x_n).$$

Taking $x = 0$, we obtain

$$\sum_{k=0}^{n} l_k(0) x_k^{n+1} = (-1)^n x_0 x_1 \ldots x_n.$$

3.16 Find the unique polynomial $P(x)$ of degree 2 or less such that

$$P(1) = 1,\ P(3) = 27,\ P(4) = 64$$

using each of the following methods : (*i*) Lagrange interpolation formula, (*ii*) Newton-divided difference formula and (*iii*) Aitken's iterated interpolation formula. Evaluate $P(1.5)$.

Solution

(*i*) Using Lagrange interpolation (3.7), we obtain

$$P_2(x) = \frac{(x-4)(x-3)}{(1-4)(1-3)}(1) + \frac{(x-1)(x-4)}{(3-1)(3-4)}(27) + \frac{(x-1)(x-3)}{(4-1)(4-3)}(64)$$

$$= \frac{1}{6}(x^2 - 7x + 12) - \frac{27}{2}(x^2 - 5x + 4) + \frac{64}{3}(x^2 - 4x + 3)$$

$$= 8x^2 - 19x + 12.$$

(*ii*) We form the divided difference table.

x	$P(x)$		
1	1		
		13	
3	27		8
		37	
4	64		

Using Newton's divided difference formula (3.12), we obtain

$$P_2(x) = P[x_0] + (x - x_0) P[x_0, x_1] + (x - x_0)(x - x_1) P[x_0, x_1, x_2]$$
$$= 1 + (x - 1)(13) + (x - 1)(x - 3)(8) = 8x^2 - 19x + 12.$$

(*iii*) Using iterated interpolation (3.11), we obtain

$$I_{01}(x) = \frac{1}{x_1 - x_0} \begin{vmatrix} I_0(x) & x_0 - x \\ I_1(x) & x_1 - x \end{vmatrix} = \frac{1}{2} \begin{vmatrix} 1 & 1 - x \\ 27 & 3 - x \end{vmatrix} = 13x - 12.$$

$$I_{02}(x) = \frac{1}{x_2 - x_0} \begin{vmatrix} I_0(x) & x_0 - x \\ I_2(x) & x_2 - x \end{vmatrix} = \frac{1}{3} \begin{vmatrix} 1 & 1 - x \\ 64 & 4 - x \end{vmatrix} = 21x - 20.$$

$$I_{012}(x) = \frac{1}{x_2 - x_1} \begin{vmatrix} I_{01}(x) & x_1 - x \\ I_{02}(x) & x_2 - x \end{vmatrix} = \begin{vmatrix} 13x - 12 & 3 - x \\ 21x - 20 & 4 - x \end{vmatrix}$$
$$= 8x^2 - 19x + 12.$$

We obtain $\qquad P_2(1.5) = 1.5.$

3.17 Suppose $f'(x) = e^x \cos x$ is to be approximated on [0, 1] by an interpolating polynomial on $n + 1$ equally spaced points $0 = x_0 < x_1 < x_2 \ldots < x_n = 1$. Determine n so that the truncation error will be less than 0.0001 in this interval.

Solution

The nodal points are given by

$$x_r = r / n, r = 0, 1, \ldots, n.$$

On [0, 1], the maximum of $\left|\left(x - \frac{r}{n}\right)\left(x - \frac{n-r}{n}\right)\right|$ occurs at $x = 1 / 2$. Hence,

$$\max_{0 \le x \le 1} \left|\left(x - \frac{r}{n}\right)\left(x - \frac{n-r}{n}\right)\right| = \left(\frac{1}{2} - \frac{r}{n}\right)^2 = \left(\frac{1}{2} - x_r\right)^2 \le \frac{1}{4}$$

since x_r is any point in [0, 1].

Using this result and combining the first and last terms, second and last but one terms, etc., we get

$$\max_{0 \le x \le 1} \left| x\left(x - \frac{1}{n}\right) \ldots \left(x - \frac{n-1}{n}\right)(x - 1)\right| \le \frac{1}{2^{n+1}}$$

We have $\qquad f(x) = e^x \cos x = Re[e^{(1+i)x}]$

where *Re* stands for the real part. We have

$$f^{(r)}(x) = Re\,[(1 + i)^r e^{(1+i)x}]$$
$$= Re\left[2^{r/2}\left(\cos\frac{r\pi}{4} + i \sin\frac{r\pi}{4}\right)(\cos x + i \sin x)e^x\right]$$
$$= 2^{r/2} \cos\left(\frac{r\pi}{4} + x\right)e^x.$$

The maximum truncation error is given by

$$| \text{TE} | = \max_{0 \le x \le 1} \left| \frac{(x-0)(x-\frac{1}{n})(x-\frac{2}{n})\ldots(x-1)}{(n+1)!} \right| \max_{0 \le x \le 1} | f^{(n+1)}(x) |$$

$$\le \frac{1}{2^{n+1}(n+1)!} \max_{0 \le x \le 1} \left| 2^{(n+1)/2} \cos \left((n+1)\frac{\pi}{4} + x \right) e^x \right|$$

$$\le \frac{e}{2^{(n+1)/2}\,(n+1)!}$$

For $| \text{TE} | \le 0.0001$, we get $n \ge 6$.

3.18 If $f(x) = e^{ax}$, show that

$$\Delta^n f(x) = (e^{ah} - 1)^n e^{ax}.$$

Solution

We establish the result by induction. Since

$$\Delta f(x) = e^{a(x+h)} - e^{ax} = (e^{ah} - 1)\, e^{a},$$

the result is true for $n = 1$.

We assume that the result holds for $n = m$ that is

$$\Delta^m f(x) = (e^{ah} - 1)^m e^{ax}.$$

Then, we have

$$\Delta^{m+1} f(x) = (e^{ah} - 1)^m [e^{a(x+h)} - e^{a}] = (e^{ah} - 1)^{m+1} e^{a}$$

and the result also holds for $n = m + 1$.

Hence, the result holds for all values of n.

3.19 Calculate the nth divided difference of $f(x) = 1 / x$.

Solution

We have $\qquad f[x_0, x_1] = \left[\dfrac{1}{x_1} - \dfrac{1}{x_0} \right] \Big/ (x_1 - x_0) = -1 / (x_0 x_1).$

$$f[x_0, x_1, x_2] = \left[-\frac{1}{x_1 x_2} + \frac{1}{x_0 x_1} \right] \Big/ (x_2 - x_0) = (-1)^2 / (x_0 x_1 x_2).$$

Let the result be true for $n = k$. That is

$$f[x_0, x_1, \ldots, x_k] = \frac{(-1)^k}{x_0 x_1 \ldots x_k}.$$

We have for $n = k + 1$

$$f[x_0, x_1, \ldots, x_{k+1}]$$

$$= \frac{1}{(x_{k+1} - x_0)} \left(f[x_1, x_2, \ldots, x_{k+1}] - f[x_0, \ldots, x_k] \right)$$

$$= \frac{1}{(x_{k+1} - x_0)} \left[\frac{(-1)^k}{x_1 x_2 \ldots x_{k+1}} - \frac{(-1)^k}{x_0 x_1 \ldots x_k} \right] = \frac{(-1)^{k+1}}{x_0 x_1 x_2 \ldots x_{k+1}}.$$

Hence, $\qquad f[x_0, x_1, \ldots, x_n] = (-1)^n / (x_0 x_1 \ldots x_n).$

3.20 If $f(x) = U(x)V(x)$, show that

$$f[x_0, x_1] = U[x_0] V[x_0, x_1] + U[x_0, x_1] V[x_1].$$

Solution

We have
$$f[x_0, x_1] = [U(x_1)V(x_1) - U(x_0)V(x_0)] / (x_1 - x_0)$$
$$= [V(x_1)\{U(x_1) - U(x_0)\} + U(x_0)\{V(x_1) - V(x_0)\}] / (x_1 - x_0)$$
$$= V(x_1)U[x_0, x_1] + U(x_0)V[x_0, x_1].$$

3.21 Prove the relations

(*i*) $\nabla - \Delta = -\Delta\nabla$.

(*ii*) $\Delta + \nabla = \Delta / \nabla - \nabla / \Delta$.

(*iii*) $\sum_{k=0}^{n-1} \Delta^2 f_k = \Delta f_n - \Delta f_0$.

(*iv*) $\Delta(f_i g_i) = f_i \Delta g_i + g_{i+1} \Delta f_i$.

(*v*) $\Delta f_i^2 = (f_i + f_{i+1})\Delta f_i$.

(*vi*) $\Delta(f_i / g_i) = (g_i \Delta f_i - f_i \Delta g_i) / g_i g_{i+1}$.

(*vii*) $\Delta(1 / f_i) = -\Delta f_i / (f_i f_{i+1})$.

Solution

(*i*) L.H.S. $= (1 - E^{-1}) - (E - 1) = -(E + E^{-1} - 2)$.

R.H.S. $= -(E - 1)(1 - E^{-1}) = -(E + E^{-1} - 2)$.

(*ii*) L.H.S. $= (E - 1) + (1 - E^{-1}) = E - E^{-1}$.

R.H.S. $= (E - 1) / (1 - E^{-1}) - (1 - E^{-1}) / (E - 1) = E - E^{-1}$.

(*iii*) L.H.S. $= \sum_{k=0}^{n-1} \Delta^2 f_k = \sum_{k=0}^{n-1} (\Delta f_{k+1} - \Delta f_k)$

$= (\Delta f_1 - \Delta f_0) + (\Delta f_2 - \Delta f_1) + \ldots + (\Delta f_n - \Delta f_{n-1}) = \Delta f_n - \Delta f_0$.

(*iv*) L.H.S. $= f_{i+1} g_{i+1} - f_i g_i = g_{i+1}(f_{i+1} - f_i) + f_i(g_{i+1} - g_i) = g_{i+1} \Delta f_i + f_i \Delta g_i$.

(*v*) L.H.S. $= f_{i+1}^2 - f_i^2 = (f_{i+1} - f_i)(f_{i+1} + f_i) = (f_{i+1} + f_i)\Delta f_i$.

(*vi*) L.H.S. $= \dfrac{f_{i+1}}{g_{i+1}} - \dfrac{f_i}{g_i} = \dfrac{g_i f_{i+1} - f_i g_{i+1}}{g_i g_{i+1}}$

$= [g_i(f_{i+1} - f_i) - f_i(g_{i+1} - g_i)] / (g_i g_{i+1}) = [g_i \Delta f_i - f_i \Delta g_i] / (g_i g_{i+1})$.

(*vii*) L.H.S. $= \dfrac{1}{f_{i+1}} - \dfrac{1}{f_i} = \dfrac{1}{f_i f_{i+1}} [f_i - f_{i+1}] = -\Delta f_i / (f_i f_{i+1})$.

3.22 Use the Lagrange and the Newton-divided difference formulas to calculate $f(3)$ from the following table :

x	0	1	2	4	5	6
$f(x)$	1	14	15	5	6	19

Solution

Using Lagrange interpolation formula (3.7) we obtain

$$P_5(x) = \frac{1}{240}(x - 1)(x - 2)(x - 4)(x - 5)(x - 6)$$
$$+ \frac{14}{60}(x)(x - 2)(x - 4)(x - 5)(x - 6)$$
$$- \frac{15}{48}(x)(x - 1)(x - 4)(x - 5)(x - 6)$$
$$+ \frac{5}{48}(x)(x - 1)(x - 2)(x - 5)(x - 6)$$

$$- \frac{6}{60} (x)(x-1)(x-2)(x-4)(x-6)$$

$$+ \frac{19}{240} (x)(x-1)(x-2)(x-4)(x-5)$$

which gives $f(3) = P_5(3) = 10.$

To use the Newton divided difference interpolation formula (3.12), we first construct the divided difference table

x	$f(x)$					
0	1					
1	14	13				
2	15	1	– 6			
4	5	– 5	– 2	1		
5	6	1	2	1	0	
6	19	13	6	1	0	0

We obtain the Newton divided difference interpolating polynomial as

$$P_5(x) = 1 + 13x - 6x(x-1) + x(x-1)(x-2)$$
$$= x^3 - 9x^2 + 21x + 1$$

which gives $f(3) = P_5(3) = 10.$

3.23 The following data are part of a table for $g(x) = \sin x \;/\; x^2$.

x	0.1	0.2	0.3	0.4	0.5
$g(x)$	9.9833	4.9667	3.2836	2.4339	1.9177

Calculate $g(0.25)$ as accurately as possible

(a) by interpolating directly in this table,

(b) by first tabulating $x\, g(x)$ and then interpolating in that table,

(c) explain the difference between the results in (a) and (b) respectively.

(Umea Univ., Sweden, BIT 19 (1979), 285)

Solution

(a) First we construct the forward difference table from the given data

x	$g(x)$	Δg	$\Delta^2 g$	$\Delta^3 g$	$\Delta^4 g$
0.1	9.9833				
		– 5.0166			
0.2	4.9667		3.3335		
		– 1.6831		– 2.5001	
0.3	3.2836		0.8334		2.0002
		– 0.8497		– 0.4999	
0.4	2.4339		0.3335		
		– 0.5162			
0.5	1.9177				

Using these differences, we obtain the interpolating polynomial using the first four points as

$$P_3(x) = 9.9833 + \frac{(x-0.1)}{0.1}(-5.0166) + \frac{(x-0.1)(x-0.2)}{2(0.1)^2}(3.3335) + \frac{(x-0.1)(x-0.2)(x-0.3)}{6(0.1)^3}(-2.5001)$$

which gives $\quad g(0.25) \approx P_3(0.25) = 3.8647.$

We write the error term as

$$\text{Error} \approx \frac{(x-0.1)(x-0.2)(x-0.3)(x-0.4)}{4\,!(0.1)^4}\Delta^4 f_0$$

since $f^{(4)}(\xi) \approx \Delta^4 f_0 \,/\, h^4$, and obtain the error at $x = 0.25$ as

$$|\text{ Error }| = 0.0469 \approx 0.05.$$

Hence, we have $g(0.25) = 3.87 \pm 0.05$.

(*b*) We first form the table for $f(x) = x\, g(x)$ and then compute the forward differences.

x	$f = x\,g(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0.1	0.99833				
		– 0.00499			
0.2	0.99334		– 0.00327		
		– 0.00826		0.00001	
0.3	0.98508		– 0.00326		0.00006
		– 0.01152		0.00007	
0.4	0.97356		– 0.00319		
		– 0.01471			
0.5	0.95885				

Using the first four points and these forward differences, we obtain the interpolating polynomial as

$$P_3(x) = 0.99833 + \frac{(x-0.1)}{0.1}(-0.00499) + \frac{(x-0.1)(x-0.2)}{2(0.1)^2}(-0.00327) + \frac{(x-0.1)(x-0.2)(x-0.3)}{6(0.1)^3}(0.00001)$$

which gives $\quad (0.25)g(0.25) \approx P_3(0.25) = 0.989618,$

or $\quad g(0.25) = 3.958472.$

We write the error term in $0.25g(0.25)$ as

$$\frac{(x-0.1)(x-0.2)(x-0.3)(x-0.4)}{4\,!(0.1)^4}\Delta^4 f_0$$

which gives error in $0.25g(0.25)$ as 0.000001406 and therefore, error in $g(0.25)$ as 0.000005625.

Hence, we have

$$g(0.25) = 3.95847 \pm 0.000006.$$

(*c*) Since the differences in (*a*) are oscillating and are not decreasing fast, the resulting error in interpolation would be large.

Since differences in part (*b*) tend to become smaller in magnitude, we expect more accurate results in this case.

3.24 In a computer program, quick access to the function 2^x is needed, $0 \le x \le 1$. A table with step size h is stored into an array and the function values are calculated by interpolation in this table.

(*a*) Which is the maximal step size to be used when function values are wanted correct to 5 decimal places by linear interpolation ?

(The precision of the computer arithmetic is much better than so.)

(*b*) The same question when quadratic interpolation is used.

(Royal Inst. Tech., Stockholm, Sweden, BIT 26 (1986), 541)

Solution

We have $f(x) = 2^x, f^{(r)}(x) = 2^x (\ln 2)^r, r = 1, 2, ...$

The maximum errors in linear and quadratic interpolation are given by $h^2M_2 / 8$ and $h^3M_3 / (9\sqrt{3})$ respectively, where

$$M_r = \max_{0 \le x \le 1} | f^{(r)}(x) |.$$

Since $0 \le x \le 1$, we have

$$M_2 = 2 (\ln 2)^2 \quad \text{and} \quad M_3 = 2(\ln 2)^3.$$

(*a*) We choose h such that

$$\frac{2h^2}{8} (\ln 2)^2 \le 0.000005$$

which gives $h \le 0.00645.$

(*b*) We choose h such that

$$\frac{2h^3}{9\sqrt{3}} (\ln 2)^3 \le 0.000005$$

which gives $h \le 0.04891.$

3.25 The error function *erf* (x) is defined by the integral

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

(*a*) Approximate *erf* (0.08) by linear interpolation in the given table of correctly rounded values. Estimate the total error.

x	0.05	0.10	0.15	0.20
$erf(x)$	0.05637	0.11246	0.16800	0.22270

(*b*) Suppose that the table were given with 7 correct decimals and with the step size 0.001 of the abscissas. Find the maximal total error for linear interpolation in the interval $0 \le x \le 0.10$ in this table. (Linköping Univ., Sweden, BIT 26(1986), 398)

Solution

(*a*) Using linear interpolation based on the points 0.05 and 0.10, we have

$$P_1(x) = \frac{x - 0.10}{0.05 - 0.10} (0.05637) + \frac{x - 0.05}{0.10 - 0.05} (0.11246)$$

We obtain $erf(0.08) \approx P_1(0.08) = 0.09002.$

The maximum error of interpolation is given by

$$|\text{TE}| = \frac{h^2}{8} M_2$$

where $$M_2 = \max_{0.05 \le x \le 0.10} |f''(x)| = \max_{0.05 \le x \le 0.10} \left| \frac{-4x}{\sqrt{\pi}} e^{-x^2} \right| = 0.2251.$$

Hence, $$|\text{TE}| = \frac{(0.05)^2}{8} (0.2251) = 0.000070 = 7.0 \times 10^{-5}.$$

(*b*) In this case, $h = 0.001$ and

$$M_2 = \max_{0 \le x \le 0.10} \left| \frac{-4x}{\sqrt{\pi}} e^{-x^2} \right| = \frac{0.4}{\sqrt{\pi}} = 0.2256758.$$

Hence, we have

$$|\text{TE}| = \frac{(0.001)^2}{8} (0.2256758) = 3.0 \times 10^{-8}.$$

3.26 The function f is displayed in the table, rounded to 5 correct decimals. We know that $f(x)$ behaves like $1 / x$ when $x \to 0$. We want an approximation of $f(0.55)$. Either we use quadratic interpolation in the given table, or we set up a new table for $g(x) = xf(x)$, interpolate in that table and finally use the connection $f(x) = g(x) / x$. Choose the one giving the smallest error, calculate $f(0.55)$ and estimate the error.

x	$f(x)$	x	$f(x)$
0.1	20.02502	0.6	3.48692
0.2	10.05013	0.7	3.03787
0.3	6.74211	0.8	2.70861
0.4	5.10105	0.9	2.45959
0.5	4.12706	1.0	2.26712

(Bergen Univ., Sweden, BIT 24(1984), 397)

Solution

The second procedure must be chosen as in that case $g(x)$ is a well behaved function as $x \to 0$ and the interpolation would have the smallest possible error. However, to illustrate the difference between the two procedures, we obtain the solution using both the methods.

We form the forward difference table based on the points 0.5, 0.6, 0.7 and 0.8 for the function $f(x)$, so that quadratic interpolation can be used.

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
0.5	4.12706			
		– 0.64014		
0.6	3.48692		0.19109	
		– 0.44905		
0.7	3.03787			– 0.07130
		– 0.32926	0.11979	
0.8	2.70861			

We obtain the quadratic interpolating polynomial based on the points 0.5, 0.6 and 0.7 as

$$P_2(x) = 4.12706 + \frac{(x-0.5)}{0.1}(-0.64014) + \frac{(x-0.5)(x-0.6)}{2!(0.1)^2}(0.19109)$$

which gives $f(0.55) \approx P_2(0.55) = 3.783104.$

The error term is given by

$$\text{Error }(x) = \frac{(x-0.5)(x-0.6)(x-0.7)}{3!(0.1)^3}(-0.07130)$$

since $f'''(\xi) \approx \Delta^3 f_0 / h^3$. Hence,

$$|\text{ Error }(0.55)\ | = 4.5 \times 10^{-3}.$$

Now, we set up a new table for $g(x) = x f(x)$ and form the forward difference table based on the points 0.5, 0.6, 0.7 and 0.8 for $g(x)$ as

x	$g(x)$	Δg	$\Delta^2 g$	$\Delta^3 g$
0.5	2.063530			
		0.028622		
0.6	2.092152		0.005735	
		0.034357		0.000287
0.7	2.126509		0.006022	
		0.040379		
0.8	2.166888			

The quadratic interpolating polynomial for $g(x)$ based on the points 0.5, 0.6 and 0.7 is obtained as

$$P_2(x) = 2.063530 + \frac{(x-0.5)}{0.1}(0.028622) + \frac{(x-0.5)(x-0.6)}{2!(0.1)^2}(0.005735)$$

which gives

$$0.55 f(0.55) = g(0.55) \approx P_2(0.55) = 2.077124,$$

or

$$f(0.55) = 3.776589.$$

$$\text{Error in } g(0.55) = \frac{(0.55-0.5)(0.55-0.6)(0.55-0.7)}{6(0.1)^3}(0.000287) = 0.000018.$$

Hence, $|$ error in $f(0.55)\ | = 3.3 \times 10^{-5}$.

3.27 The graph of a function f is almost a parabolic segment attaining its extreme values in an interval (x_0, x_2). The function values $f_i = f(x_i)$ are known at equidistant abscissas x_0, x_1, x_2. The extreme value is searched. Use the quadratic interpolation to derive x coordinate of the extremum. (Royal Inst. Tech., Stockholm, Sweden, BIT 26(1986), 135)

Solution

Replacing $f(x)$ by the quadratic interpolating polynomial, we have

$$P_2(x) = f_0 + \frac{(x-x_0)}{h}\Delta f_0 + \frac{(x-x_0)(x-x_1)}{2!h^2}\Delta^2 f_0 .$$

The extremum is attained when

$$P_2'(x) = 0 = \frac{1}{h}\Delta f_0 + \frac{(2x-x_0-x_1)}{2h^2}\Delta^2 f_0$$

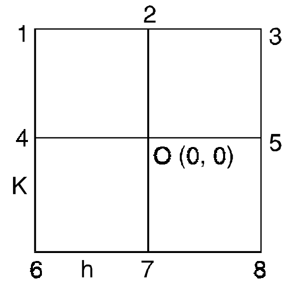
which gives $$x_{\text{extremum}} = \frac{1}{2}(x_0+x_1) - h\frac{\Delta f_0}{\Delta^2 f_0}.$$

Piecewise and Spline Interpolation

3.28 Determine the piecewise quadratic approximating function of the form

$$S_\Delta(x, y) = \sum_{i=0}^{8} N_i \, f_i$$

for the following configuration of the rectangular network.



Solution

We have

$$N_0 = (x^2 - h^2)(y^2 - k^2) \,/\, d, \qquad N_1 = x(x - h)y(y + k) \,/\, (4d),$$
$$N_2 = -\,(x^2 - h^2)y(y + k) \,/\, (2d), \qquad N_3 = x(x + h)y(y + k) \,/\, (4d),$$
$$N_4 = -\,x(x - h)(y^2 - k^2) \,/\, (2d), \qquad N_5 = -\,x(x + h)(y^2 - k^2) \,/\, (2d),$$
$$N_6 = x(x - h)y(y - k) \,/\, (4d), \qquad N_7 = -\,(x^2 - h^2)(y)(y - k) \,/\, (2d),$$
$$N_8 = x(x + h)y(y - k) \,/\, (4d), \qquad d = h^2k^2.$$

3.29 Determine the piecewise quadratic fit $P(x)$ to $f(x) = (1 + x^2)^{-1/2}$ with knots at $-1, -1/2, 0, 1/2, 1$. Estimate the error $|\, f - P \,|$ and compare this with full Lagrange polynomial fit.

Solution

We have the following data values

x	-1	$-1/2$	0	$1/2$	1
f	$1/\sqrt{2}$	$2/\sqrt{5}$	1	$2/\sqrt{5}$	$1/\sqrt{2}$

We obtain the quadratic interpolating polynomial based on the points $-1, -1/2$ and 0, as

$$P_2(x) = \frac{(x + 1/2)x}{(-1/2)(-1)}\left(\frac{1}{\sqrt{2}}\right) + \frac{(x + 1)x}{(1/2)(-1/2)}\left(\frac{2}{\sqrt{5}}\right) + \frac{(x + 1)(x + 1/2)}{(1)\,(1/2)}\,(1)$$

$$= \frac{1}{\sqrt{10}}\,[(2\sqrt{5} - 8\sqrt{2} + 2\sqrt{10}\,)x^2 + (\sqrt{5} - 8\sqrt{2} + 3\sqrt{10}\,)x + \sqrt{10}\,]$$

Similarly, the quadratic interpolating polynomial based on the points $0, 1/2$ and 1 is obtained as

$$P_2(x) = \frac{(x - 1/2)(x - 1)}{(-1/2)(-1)}\,(1) + \frac{x(x - 1)}{(1/2)(-1/2)}\left(\frac{2}{\sqrt{5}}\right) + \frac{x(x - 1/2)}{(1)(1/2)}\left(\frac{1}{\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{10}}\,[(2\sqrt{10} - 8\sqrt{2} + 2\sqrt{5})x^2 + (8\sqrt{2} - 3\sqrt{10} - \sqrt{5})x + \sqrt{10}\,].$$

The maximum error of quadratic interpolation is given by

$$|\text{ TE }| \le \frac{h^3 M_3}{9\sqrt{3}}$$

where $h = 0.5$ and $M_3 = \max\limits_{-1 \le x \le 0} |\, f'''(x) \,|$.

We find $f'''(x) = (9x - 6x^3)(1 + x^2)^{-7/2}$,

and $$M_3 = \max_{-1 \le x \le 1} |\, 9x - 6x^3 \,| \max_{-1 \le x \le 1} |\, (1 + x^2)^{-7/2} \,|.$$

Now, $F(x) = 9x - 6x^3$ attains its extreme values at $x^2 = 1/2$.

We get $$\max_{-1 \le x \le 1} |9x - 6x^3| = \frac{6}{\sqrt{2}}.$$

Hence, we have $$|\text{ TE }| \le \frac{1}{12\sqrt{6}} = 0.0340.$$

Maximum error occurs at two points, $x = \pm 1/\sqrt{2}$, in $[-1, 1]$.

The Lagrangian fourth degree polynomial based on the points $-1, -1/2, 0, 1/2, 1$ is obtained as

$$P_4(x) = \left(x + \frac{1}{2}\right) x \left(x - \frac{1}{2}\right)(x - 1)\left(\frac{2}{3\sqrt{2}}\right) - (x + 1)x\left(x - \frac{1}{2}\right)(x - 1)\left(\frac{16}{3\sqrt{5}}\right)$$
$$+ (x + 1)\left(x + \frac{1}{2}\right)\left(x - \frac{1}{2}\right)(x - 1)(4) - (x + 1)\left(x + \frac{1}{2}\right)x(x - 1)\left(\frac{16}{3\sqrt{5}}\right)$$
$$+ (x + 1)\left(x + \frac{1}{2}\right)x\left(x - \frac{1}{2}\right)\left(\frac{2}{3\sqrt{2}}\right)$$
$$= \frac{1}{6\sqrt{5}}[(4\sqrt{10} - 64 + 24\sqrt{5})x^4 + (-\sqrt{10} + 64 - 30\sqrt{5})x^2 + 6\sqrt{5}]$$

The error of interpolation is given by

$$\text{TE} = \frac{1}{5!}(x + 1)\left(x + \frac{1}{2}\right)x\left(x - \frac{1}{2}\right)(x - 1)f^{v}(\xi), \quad -1 < \xi < 1.$$

Hence, $$|\text{ TE }| \le \frac{1}{120}\max_{-1 \le x \le 1}|G(x)| M_5$$

where, $$G(x) = (x + 1)\left(x + \frac{1}{2}\right)x\left(x - \frac{1}{2}\right)(x - 1)$$
$$= x(x^2 - 1)\left(x^2 - \frac{1}{4}\right) = x^5 - \frac{5}{4}x^3 + \frac{x}{4}.$$

$G(x)$ attains extreme values when

$$G'(x) = 5x^4 - \frac{15}{4}x^2 + \frac{1}{4} = 0, \quad -1 \le x \le 1.$$

whose solution is $$x^2 = \frac{15 \pm \sqrt{145}}{40} \approx 0.0740, 0.6760.$$

Now, max $|G(x)|$ is obtained for $x^2 = 0.6760$.

We obtain, $\max\limits_{-1 \le x \le 1} |G(x)| = 0.1135$.

$$M_5 = \max_{-1 \le x \le 1} |f^{v}(x)|$$
$$= \max_{-1 \le x \le 1} |(-120x^5 + 600x^3 - 225x)(1 + x^2)^{-11/2}|$$
$$= \max_{-1 \le x \le 1} |H(x)| \max_{-1 \le x \le 1}\left|\frac{1}{(1 + x^2)^{11/2}}\right|$$

$H(x)$ attains extreme values when

$$H'(x) = 600x^4 - 1800x^2 + 225 = 0, -1 \le x \le 1.$$

We obtain $x^2 = 0.1307$, which is the only possible value. We have $|H(\pm\sqrt{0.1307})| = 53.7334$. We also have $|H(\pm 1)| = 255$. Hence, $M_5 = 255$, and

$$\text{maximum error} = \frac{0.1135}{120}(255) = 0.2412.$$

3.30 $S_3(x)$ is the piecewise cubic Hermite interpolating approximant of $f(x) = \sin x \cos x$ in the abscissas 0, 1, 1.5, 2, 3. Estimate the error $\max\limits_{0 \le x \le 3} | f(x) - S_3(x) |$.

(Uppsala Univ., Sweden, BIT 19(1979), 425)

Solution

Error in piecewise cubic Hermite interpolation is given by

$$\text{TE} = \frac{1}{4!}(x - x_{i-1})^2 (x - x_i)^2 f^{iv}(\xi), \quad x_{i-1} < \xi < x_i.$$

Hence,

$$| \text{TE} | \le \max_{x_{i-1} \le x \le x_i} \left| \frac{1}{4!}(x - x_{i-1})^2 (x - x_i)^2 \right| \max_{x_{i-1} \le x \le x_i} |f^{iv}(x)|$$

$$= \frac{1}{384}(x_i - x_{i-1})^4 \max_{x_{i-1} \le x \le x_i} |f^{iv}(x)|$$

Since,

$$f(x) = \tfrac{1}{2} \sin 2x, \; f'(x) = \cos 2x,$$
$$f''(x) = -2 \sin 2x \; f'''(x) = -4 \cos 2x,$$
$$f^{iv}(x) = 8 \sin 2x,$$

we have

$$\text{on } [0, 1] : | \text{Error} | \le \frac{1}{384} \max_{0 \le x \le 1} | 8 \sin 2x | = 0.0208,$$

$$\text{on } [1, 1.5] : | \text{Error} | \le \frac{1}{384 \times 16} \max_{1 \le x \le 1.5} | 8 \sin 2x | = 0.0012,$$

$$\text{on } [1.5, 2] : | \text{Error} | \le \frac{1}{16 \times 384} \max_{1.5 \le x \le 2} | 8 \sin 2x | = 0.00099,$$

$$\text{on } [2, 3] : | \text{Error} | \le \frac{1}{384} \max_{2 \le x \le 3} | 8 \sin 2x | = 0.0208.$$

Hence, maximum error on [0, 3] is 0.0208.

3.31 Suppose $f_i = x_i^{-2}$ and $f_i' = -2x_i^{-3}$ where $x_i = i / 2$, $i = 1(1)4$ are given. Fit these values by the piecewise cubic Hermite polynomial.

Solution

We have the data

i	1	2	3	4
x_i	1 / 2	1	3 / 2	2
f_i	4	1	4 / 9	1 / 4
f_i'	– 16	– 2	– 16 / 27	– 1 / 4

Cubic Hermite interpolating polynomial on $[x_{i-1}, x_i]$ is given by

$$P_3(x) = \frac{(x - x_i)^2}{(x_{i-1} - x_i)^2}\left[1 + \frac{2(x_{i-1} - x)}{(x_{i-1} - x_i)}\right] f_{i-1} + \frac{(x - x_{i-1})^2}{(x_i - x_{i-1})^2}\left[1 + \frac{2(x_i - x)}{x_i - x_{i-1}}\right] f_i$$

$$+ \frac{(x - x_{i-1})(x - x_i)^2}{(x_{i-1} - x_i)^2} f'_{i-1} + \frac{(x - x_i)(x - x_{i-1})^2}{(x_{i-1} - x_i)^2} f'_i$$

On [1 / 2, 1], we obtain

$$P_3(x) = 4(x - 1)^2 [1 - 4((1/2) - x)](4) + 4(x - (1/2))^2 [1 - 4(x - 1)](1)$$
$$+ 4(x - (1/2))(x - 1)^2(-16) + 4(x - 1)(x - (1/2))^2(-2)$$
$$= -24x^3 + 68x^2 - 66x + 23.$$

On [1, 3 / 2], we obtain

$$P_3(x) = 4(x - (3/2))^2[1 - 4(1 - x)](1) + 4(x - 1)^2 [1 - 4(x - (3/2))](4/9)$$
$$+ 4(x - 1)(x - (3/2))^2(-2) + 4(x - (3/2))(x - 1)^2(-16/27)$$
$$= [-40x^3 + 188x^2 - 310x + 189] / 27.$$

On [3 / 2, 2], we have

$$P_3(x) = 4(x - 2)^2[1 - 4((3/2) - x)](4/9) + 4(x - (3/2))^2 [1 - 4(x - 2)] (1/4)$$
$$+ 4(x - (3/2))(x - 2)^2(-16/27) + 4(x - 2)(x - (3/2))^2(-1/4)$$
$$= [-28x^3 + 184x^2 - 427x + 369] / 108.$$

3.32 Find whether the following functions are splines or not.

(*i*) $f(x) = \begin{vmatrix} x^2 - x + 1, & 1 \le x \le 2 \\ 3x - 3, & 2 \le x \le 3 \end{vmatrix}$ (*ii*) $f(x) = \begin{vmatrix} -x^2 - 2x^3, & -1 \le x \le 0 \\ -x^2 + 2x^3, & 0 \le x \le 1 \end{vmatrix}$

(*iii*) $f(x) = \begin{vmatrix} -x^2 - 2x^3, & -1 \le x \le 0 \\ x^2 + 2x^3, & 0 \le x \le 1 \end{vmatrix}.$

Solution

(*i*) $f(x)$ defines a second order polynomial. Since $f(x)$ and $f'(x)$ are continuous in each of the intervals [1, 2] and [2, 3], the given function is a quadratic spline.

(*ii*) $f(x)$ defines a third degree polynomial. Since $f(x)$, $f'(x)$ and $f''(x)$ are continuous in each of the intervals [– 1, 0] and [0, 1], the given function is a cubic spline.

(*iii*) $f(x)$ defines a third degree polynomial. Since $f''(x)$ is not continuous at $x = 0$, the given function is not a spline.

3.33 Fit a cubic spline, $s(x)$ to the function $f(x) = x^4$ on the interval $-1 \le x \le 1$ corresponding to the partition $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ and satisfying the conditions $s'(-1) = f'(-1)$ and $s'(1) = f'(1)$.

Solution

We have the data

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

with $m_0 = f'(-1) = -4$ and $m_2 = f'(1) = 4.$

The nodal points are equispaced with $h = 1$. We obtain the equation

$$m_0 + 4m_1 + m_2 = 3(f_2 - f_0) = 0$$

which gives $m_1 = 0.$

Spline interpolation becomes

On the interval $[x_0, x_1] : x_0 = -1, x_1 = 0, h = 1$

$$\begin{aligned} s(x) &= (x - x_1)^2 [1 - 2(x_0 - x)] f_0 + (x - x_0)^2 [1 - 2(x - x_1)] f_1 \\ &\quad + (x - x_0)(x - x_1)^2 m_0 + (x - x_1)(x - x_0)^2 m_1 \\ &= x^2(3 + 2k)(1) + (x - 1)^2 (1 - 2x)(0) + (x + 1)x^2(-4) + x(x + 1)^2 (0) \\ &= -2x^3 - x^2. \end{aligned}$$

On the interval $[x_1, x_2] : x_1 = 0, x_2 = 1, h = 1$

$$\begin{aligned} s(x) &= (x - x_2)^2 [1 - 2(x_1 - x)] f_1 + (x - x_1)^2 [1 - 2(x - x_2)] f_2 \\ &\quad + (x - x_1)(x - x_2)^2 m_1 + (x - x_2)(x - x_1)^2 m_2 \\ &= (x - 1)^2(1 + 2x)(0) + x^2(3 - 2x)(1) + x(x - 1)^2 (0) + (x - 1) x^2 (4) \\ &= 2x^3 - x^2. \end{aligned}$$

3.34 Obtain the cubic spline fit for the data

x	-1	0	1	2
$f(x)$	5	-2	-7	2

with the conditions $f'(-1) = f'(2) = 1$.

Solution

Here, the points are equispaced with $h = 1$. We have the system of equations

$$m_{i-1} + 4m_i + m_{i+1} = 3(f_{i+1} - f_{i-1}),\ i = 1, 2$$

with $m_0 = m_3 = 1$.

Using the given data, we obtain the system of equations

$$\begin{aligned} 4m_1 + m_2 &= 3(f_2 - f_0) - m_0 = -36 \\ m_1 + 4m_2 &= 3(f_3 - f_1) - m_3 = 12 \end{aligned}$$

which give $m_1 = -53 / 5$ and $m_2 = 27 / 5$.

Spline interpolation becomes :

On the interval $[x_0, x_1].\ x_0 = -1, x_1 = 0, h = 1$

$$\begin{aligned} P(x) &= (x - x_1)^2 [1 - 2(x_0 - x)] f_0 + (x - x_0)^2 [1 - 2(x - x_1)] f_1 \\ &\quad + (x - x_0)(x - x_1)^2 m_0 + (x - x_1)(x - x_0)^2 m_1 \\ &= x^2(3 + 2x)(5) + (x + 1)^2 (1 - 2x)(-2) + x^2(x + 1)(1) + (x + 1)^2 x(-53 / 5) \\ &= \frac{1}{5} [22x^3 + 4x^2 - 53x - 10]. \end{aligned}$$

On the interval $[x_1, x_2] : x_1 = 0, x_2 = 1, h = 1$

$$\begin{aligned} P(x) &= (x - x_2)^2 [1 - 2(x_1 - x)] f_1 + (x - x_1)^2 [1 - 2(x - x_2)] f_2 \\ &\quad + (x - x_1)(x - x_2)^2 m_1 + (x - x_2)(x - x_1)^2 m_2 \\ &= (x - 1)^2 (1 + 2x) (-2) + x^2 (3 - 2x)(-7) + x(x - 1)^2(-53 / 5) + (x - 1) x^2 (27 / 5) \\ &= \frac{1}{5} [24x^3 + 4x^2 - 53x - 10]. \end{aligned}$$

On the interval $[x_2, x_3] : x_2 = 1, x_3 = 2, h = 1$

$$\begin{aligned} P(x) &= (x - x_3)^2 [1 - 2(x_2 - x)] f_2 + (x - x_2)^2 [1 - 2(x - x_3)] f_3 \\ &\quad + (x - x_2)(x - x_3)^2 m_2 + (x - x_3)(x - x_2)^2 m_3 \end{aligned}$$

$$= (x-2)^2(-1+2x)(-7) + (x-1)^2(5-2x)(2) + (x-1)(x-2)^2(27/5) + (x-2)(x-1)^2$$

$$= \frac{1}{5}[-58x^3 + 250x^2 - 299x + 72].$$

3.35 Obtain the cubic spline fit for the data

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

under the end conditions $f''(0) = 0 = f''(3)$ and valid in the interval [1, 2]. Hence, obtain the estimate of $f(1.5)$.

Solution

Here the points are equispaced with $h = 1$. We have the system of equations

$$M_{i-1} + 4M_i + M_{i+1} = 6(f_{i+1} - 2f_i + f_{i-1}), \quad i = 1, 2$$

with $$M_0 = M_3 = 0.$$

Using the given data, we obtain the system of equations

$$4M_1 + M_2 = 6(10 - 8 + 1) = 18$$

$$M_1 + 4M_2 = 6(8 - 20 + 4) = -48$$

which give $$M_1 = 8, M_2 = -14.$$

Spline interpolation on the interval $[x_1, x_2]$, where $x_1 = 1$, $x_2 = 2$ becomes

$$P(x) = \frac{1}{6}(x_2 - x)[(x_2 - x)^2 - 1]M_1 + \frac{1}{6}(x - x_1)[(x - x_1)]^2 - 1]M_2 + (x_2 - x)f_1 + (x - x_1)f_2$$

$$= \frac{1}{6}(2 - x)[(2 - x)^2 - 1](8) + \frac{1}{6}(x - 1)[(x - 1)^2 - 1](-14) + (2 - x)(4) + (x - 1)(10)$$

$$= \frac{1}{3}[-11x^3 + 45x^2 - 40x + 18].$$

We get $f(1.5) \approx P(1.5) = 7.375$.

3.36 Fit the following four points by the cubic splines

i	0	1	2	3
x_i	1	2	3	4
y_i	1	5	11	8

Use the end conditions $y_0'' = y_3'' = 0$. Hence, compute (*i*) $y(1.5)$ and (*ii*) $y'(2)$.

Solution

Here, the points are equispaced with $h = 1$. We have the system of equations

$$M_{i-1} + 4M_i + M_{i+1} = 6(y_{i+1} - 2y_i + y_{i-1}), \; i = 1, 2$$

with $$M_0 = M_3 = 0.$$

We obtain from the given data

$$4M_1 + M_2 = 6(11 - 10 + 1) = 12,$$

$$M_1 + 4M_2 = 6(8 - 22 + 5) = -54,$$

which give $$M_1 = 102/15, M_2 = -228/15.$$

Spline interpolation becomes :

On the interval $[x_0, x_1] : x_0 = 1, x_1 = 2, h = 1.$

$$P_1(x) = \frac{1}{6}(x_1 - x)[(x_1 - x)^2 - 1]\, M_0 + \frac{1}{6}(x - x_0)[(x - x_0)^2 - 1]\, M_1 + (x_1 - x)y_0 + (x - x_0)y_1$$

$$= \frac{1}{6}(x - 1)[(x - 1)^2 - 1]\left(\frac{102}{15}\right) + (2 - x)(1) + (x - 1)(5)$$

$$= \frac{1}{15}[17x^3 - 51x^2 + 94x - 45].$$

On the interval $[x_1, x_2] : x_1 = 2, x_2 = 3, h = 1.$

$$P_2(x) = \frac{1}{6}(x_2 - x)[(x_2 - x)^2 - 1]\, M_1 + \frac{1}{6}(x - x_1)[(x - x_1)^2 - 1]\, M_2 + (x_2 - x)y_1 + (x - x_1)y_2$$

$$= \frac{1}{6}(3 - x)[(3 - x)^2 - 1]\left(\frac{102}{15}\right) + \frac{1}{6}(x - 2)[(x - 2)^2 - 1]\left(-\frac{228}{15}\right) + (3 - x)(5) + (x - 2)(11)$$

$$= \frac{1}{15}[-55x^3 + 381x^2 - 770x + 531].$$

On the interval $[x_2, x_3] : x_2 = 3, x_3 = 4, h = 1.$

$$P_3(x) = \frac{1}{6}(x_3 - x)[(x_3 - x)^2 - 1]\, M_2 + \frac{1}{6}(x - x_2)[(x - x_2)^2 - 1]\, M_3 + (x_3 - x)y_2 + (x - x_2)y_3$$

$$= \frac{1}{6}(4 - x)[(4 - x)^2 - 1]\left(-\frac{228}{15}\right) + (4 - x)(11) + (x - 3)(8)$$

$$= \frac{1}{15}[38x^3 - 456x^2 + 1741x - 1980].$$

Since $1.5 \in [1, 2]$ and $2 \in [1, 2]$, we have

$$y(1.5) \approx P_1(1.5) = \frac{103}{40} = 2.575.$$

$$y'(2.0) \approx P_1'(2.0) = \frac{94}{15} = 6.267.$$

Bivariate Interpolation

3.37 The following data represents a function $f(x, y)$

$y \backslash x$	0	1	4
0	1	4	49
1	1	5	53
3	1	13	85

Obtain the bivariate interpolating polynomial which fits this data.

Solution

Since the nodal points are not equispaced, we determine Lagrange bivariate interpolating polynomial. We have

$$X_{20} = \frac{(x-1)(x-4)}{(0-1)(0-4)}, \quad X_{21} = \frac{(x-0)(x-4)}{(1-0)(1-4)}, \quad X_{22} = \frac{(x-0)(x-1)}{(4-0)(4-1)},$$

$$Y_{20} = \frac{(y-1)(y-3)}{(0-1)(0-3)}, \quad Y_{21} = \frac{(y-0)(y-3)}{(1-0)(1-3)}, \quad Y_{22} = \frac{(y-0)(y-1)}{(3-0)(3-1)}$$

and
$$P_2(x, y) = \sum_{i=0}^{2} \sum_{j=0}^{2} X_{2i} Y_{2j} f_{ij}$$
$$= X_{20}(Y_{20} f_{00} + Y_{21} f_{01} + Y_{22} f_{02})$$
$$+ X_{21}(Y_{20} f_{10} + Y_{21} f_{11} + Y_{22} f_{12}) + X_{22}(Y_{20} f_{20} + Y_{21} f_{21} + Y_{22} f_{22})$$

Using the given data, we obtain

$$Y_{20} f_{00} + Y_{21} f_{01} + Y_{22} f_{02} = \frac{1}{3}(y^2 - 4y + 3) - \frac{1}{2}(y^2 - 3y) + \frac{1}{6}(y^2 - y) = 1$$

$$Y_{20} f_{10} + Y_{21} f_{11} + Y_{22} f_{12} = \frac{4}{3}(y^2 - 4y + 3) - \frac{5}{2}(y^2 - 3y) + \frac{13}{6}(y^2 - y) = y^2 + 4$$

$$Y_{20} f_{20} + Y_{21} f_{21} + Y_{22} f_{22} = \frac{49}{3}(y^2 - 4y + 3) - \frac{53}{2}(y^2 - 3y) + \frac{85}{6}(y^2 - y) = 4y^2 + 49$$

Hence, we get

$$P_2(x, y) = \frac{1}{4}(x^2 - 5x + 4)(1) - \frac{1}{3}(x^2 - 4x)(y^2 + 4) + \frac{1}{12}(x^2 - x)(4y^2 + 49)$$
$$= 1 + 3x^2 + xy^2$$

3.38 Obtain the Newton's bivariate interpolating polynomial that fits the following data

y \ x	1	2	3
1	4	18	56
2	11	25	63
3	30	44	82

Solution

We have $\quad h = k = 1 \quad$ and

$$P_2(x, y) = f_{00} + [(x - x_0)\,\Delta_x + (y - y_0)\,\Delta_y] f_{00}$$
$$+ \tfrac{1}{2}\,[(x - x_0)(x - x_1)\,\Delta_{xx} - 2(x - x_0)(y - y_0)\,\Delta_{xy} + (y - y_0)(y - y_1)\,\Delta_{yy}] f_{00}$$

Using the given data, we obtain

$$\Delta_x f_{00} = f_{10} - f_{00} = 18 - 4 = 14$$
$$\Delta_y f_{00} = f_{01} - f_{00} = 11 - 4 = 7$$
$$\Delta_{xx} f_{00} = f_{20} - 2f_{10} + f_{00} = 56 - 36 + 4 = 24$$
$$\Delta_{xy} f_{00} = f_{11} - f_{10} - f_{01} + f_{00} = 25 - 18 - 11 + 4 = 0$$
$$\Delta_{yy} f_{00} = f_{02} - 2f_{01} + f_{00} = 30 - 22 + 4 = 12.$$

Therefore,

$$P_2(x, y) = 4 + [14(x - 1) + 7(y - 1)] + \tfrac{1}{2}\,[24(x - 1)(x - 2) + 12(y - 1)(y - 2)]$$
$$= 12x^2 + y^2 - 22x + 4y + 9.$$

3.39 Using the following data, obtain the (*i*) Lagrange and (*ii*) Newton's bivariate interpolating polynomials.

y \ x	0	1	2
0	1	3	7
1	3	6	11
2	7	11	17

Solution

(*i*) We have

$$X_{20} = \frac{(x-1)(x-2)}{(-1)(-2)},\ X_{21} = \frac{x(x-2)}{(1)(-1)},\ X_{22} = \frac{x(x-1)}{(2)(1)},$$

$$Y_{20} = \frac{(y-1)(y-2)}{(-1)(-2)},\ Y_{21} = \frac{y(y-2)}{(1)(-1)},\ Y_{22} = \frac{y(y-1)}{(2)(1)},$$

and
$$P_2(x, y) = \sum_{i=0}^{2} \sum_{j=0}^{2} X_{2i}\, Y_{2j}\, f_{ij}$$
$$= X_{20}(Y_{20} f_{00} + Y_{21} f_{01} + Y_{22} f_{02})$$
$$+ X_{21}(Y_{20} f_{10} + Y_{21} f_{11} + Y_{22} f_{12}) + X_{22}(Y_{20} f_{20} + Y_{21} f_{21} + Y_{22} f_{22}).$$

Hence, we get

$$P_2(x, y) = \tfrac{1}{2}\,(x-1)(x-2)(y^2 + y + 1) - x(x-2)(y^2 + 2y + 3)$$
$$+ \tfrac{1}{2}\, x(x-1)(y^2 + 3y + 7)$$
$$= 1 + x + y + x^2 + xy + y^2.$$

(*ii*) We have

$$P_2(x, y) = f_{00} + (x\,\Delta_x + y\,\Delta_y) f_{00} + \tfrac{1}{2}\,[x(x-1)\,\Delta_{xx} + 2xy\,\Delta_{xy} + y(y-1)\,\Delta_{yy}] f_{00}$$

We obtain

$$\Delta_x f_{00} = f_{10} - f_{00} = 2,$$
$$\Delta_y f_{00} = f_{01} - f_{00} = 2,$$
$$\Delta_{xx} f_{00} = f_{20} - 2f_{10} + f_{00} = 2,$$
$$\Delta_{yy} f_{00} = f_{11} - 2f_{01} + f_{00} = 2,$$
$$\Delta_{xy} f_{00} = f_{11} - f_{10} - f_{01} + f_{00} = 1.$$

Hence,
$$P_2(x, y) = 1 + [2x + 2y] + \tfrac{1}{2}\,[2(x-1)x + 2xy + 2(y-1)y]$$
$$= 1 + x + y + x^2 + xy + y^2.$$

Least Squares Approximation

3.40 Determine the least squares approximation of the type $ax^2 + bx + c$, to the function 2^x at the points $x_i = 0, 1, 2, 3, 4$. (Royal Inst. Tech., Stockholm, Sweden, BIT 10(1970), 398)

Solution

We determine a, b and c such that

$$I = \sum_{i=0}^{4} [2^{x_i} - ax_i^2 - bx_i - c]^2 = \text{minimum}.$$

We obtain the normal equations as

$$\sum_{i=0}^{4} [2^{x_i} - ax_i^2 - bx_i - c] = 0,$$

$$\sum_{i=0}^{4} [2^{x_i} - ax_i^2 - bx_i - c]\, x_i = 0,$$

$$\sum_{i=0}^{4} [2^{x_i} - ax_i^2 - bx_i - c]\, x_i^2 = 0,$$

or

$$30a + 10b + 5c = 31,$$
$$100a + 30b + 10c = 98,$$
$$354a + 100b + 30c = 346,$$

which has the solution

$$a = 1.143,\ b = -0.971,\ c = 1.286.$$

Hence, the least squares approximation to 2^x is

$$y = 1.143x^2 - 0.971x + 1.286.$$

3.41 Obtain an approximation in the sense of the principle of least squares in the form of a polynomial of the degree 2 to the function $1 / (1 + x^2)$ in the range $-1 \le x \le 1$.

Solution

We approximate the function $y = 1 / (1 + x^2)$ by a polynomial of degree 2, $P_2(x) = a + bx + cx^2$, such that

$$I = \int_{-1}^{1} \left[\frac{1}{1+x^2} - a - bx - cx^2 \right]^2 dx = \text{minimum}.$$

We obtain the normal equations as

$$\int_{-1}^{1} \left[\frac{1}{1+x^2} - a - bx - cx^2 \right] dx = 0,$$

$$\int_{-1}^{1} \left[\frac{1}{1+x^2} - a - bx - cx^2 \right] x\, dx = 0,$$

$$\int_{-1}^{1} \left[\frac{1}{1+x^2} - a - bx - cx^2 \right] x^2\, dx = 0.$$

Integrating, we get the equations

$$2a + \frac{2c}{3} = \frac{\pi}{2},$$

$$\frac{2b}{3} = 0,$$

$$\frac{2a}{3} + \frac{2c}{5} = 2 - \frac{\pi}{2},$$

whose solution is $a = 3(2\pi - 5) / 4,\ b = 0,\ c = 15(3 - \pi) / 4.$

The least squares approximation is

$$P_2(x) = \frac{1}{4}\,[3(2\pi - 5) + 15(3 - \pi)x^2].$$

3.42 The following measurements of a function f were made :

x	-2	-1	0	1	3
$f(x)$	7.0	4.8	2.3	2	13.8

Fit a third degree polynomial $P_3(x)$ to the data by the least squares method. As the value for $x = 1$ is known to be exact and $f'(1) = 1$, we demand that $P_3(1) = 2$ and $P_3'(1) = 1$.

(Linköping Univ., Sweden, BIT 28(1988), 904)

Solution

We take the polynomial as

$$P_3(x) = a_3(x-1)^3 + a_2(x-1)^2 + a_1(x-1) + a_0$$

Since, $\quad P_3(1) = 2$ and $P_3'(1) = 1$

we obtain $\quad a_0 = 2,\ a_1 = 1.$

Hence, we determine a_3 and a_2 such that

$$\sum_{i=1}^{5} [f(x_i) - a_3(x_i-1)^3 - a_2(x_i-1)^2 - (x_i-1) - 2]^2 = \text{minimum}.$$

The normal equations are

$$\sum_{i=1}^{5}(x_i-1)^3 f_i - a_3\sum_{i=1}^{5}(x_i-1)^6 - a_2\sum_{i=1}^{5}(x_i-1)^5 - \sum_{i=1}^{5}(x_i-1)^4 - 2\sum_{i=1}^{5}(x_i-1)^3 = 0,$$

$$\sum_{i=1}^{5}(x_i-1)^2 f_i - a_3\sum_{i=1}^{5}(x_i-1)^5 - a_2\sum_{i=1}^{5}(x_i-1)^4 - \sum_{i=1}^{5}(x_i-1)^3 - 2\sum_{i=1}^{5}(x_i-1)^2 = 0.$$

Using the given data values, we obtain

$$858a_3 - 244a_3 = -177.3,$$
$$244a_3 - 114a_2 = -131.7,$$

which has the solution, $a_3 = 0.3115$, $a_2 = 1.8220$.

Hence, the required least squares approximation is

$$P_3(x) = 2 + (x-1) + 1.822(x-1)^2 + 0.3115(x-1)^3.$$

3.43 A person runs the same race track for five consecutive days and is timed as follows :

days (x)	1	2	3	4	5
times (y)	15.30	15.10	15.00	14.50	14.00

Make a least square fit to the above data using a function $a + b / x + c / x^2$.

(Uppsala Univ., Sweden, BIT 18(1978), 115)

Solution

We determine the values of a, b and c such that

$$I = \sum_{i=0}^{4} \left[y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2} \right]^2 = \text{minimum.}$$

The normal equations are obtained as

$$\sum_{i=0}^{4} y_i - 5a - b \sum_{i=0}^{4} \frac{1}{x_i} - c \sum_{i=0}^{4} \frac{1}{x_i^2} = 0,$$

$$\sum_{i=0}^{4} \frac{y_i}{x_i} - a \sum_{i=0}^{4} \frac{1}{x_i} - b \sum_{i=0}^{4} \frac{1}{x_i^2} - c \sum_{i=0}^{4} \frac{1}{x_i^3} = 0,$$

$$\sum_{i=0}^{4} \frac{y_i}{x_i^2} - a \sum_{i=0}^{4} \frac{1}{x_i^2} - b \sum_{i=0}^{4} \frac{1}{x_i^3} - c \sum_{i=0}^{4} \frac{1}{x_i^4} = 0.$$

Using the given data values, we get

$$5a + 2.283333b + 1.463611c = 73.90,$$
$$2.283333a + 1.463611b + 1.185662c = 34.275,$$
$$1.463611a + 1.185662b + 1.080352c = 22.207917,$$

which has the solution, $a = 13.0065,\ b = 6.7512,\ c = -4.4738.$

The least squares approximation is

$$f(x) = 13.0065 + \frac{6.7512}{x} - \frac{4.4738}{x^2}.$$

3.44 Use the method of least squares to fit the curve $y = c_0 / x + c_1 \sqrt{x}$ to the table of values :

x	0.1	0.2	0.4	0.5	1	2
y	21	11	7	6	5	6

(Royal Inst. Tech., Stockholm, Sweden, BIT 26(1986), 399)

Solution

We determine the values of c_0 and c_1 such that

$$\sum_{i=1}^{6} [y_i - c_0 / x_i - c_1 \sqrt{x_i}\,]^2 = \text{minimum.}$$

We obtain the normal equations as

$$\sum_{i=0}^{6} \frac{y_i}{x_i} - c_0 \sum_{i=1}^{6} \frac{1}{x_i^2} - c_1 \sum_{i=1}^{6} \frac{1}{\sqrt{x_i}} = 0,$$

$$\sum_{i=1}^{6} y_i \sqrt{x_i} - c_0 \sum_{i=1}^{6} \frac{1}{\sqrt{x_i}} - c_1 \sum_{i=1}^{6} x_i = 0.$$

Using the given data values, we obtain

$$136.5\, c_0 + 10.100805\, c_1 = 302.5,$$
$$10.100805\, c_0 + 4.2\, c_1 = 33.715243,$$

which has the solution, $c_0 = 1.9733,\ c_1 = 3.2818.$

Hence, the required least squares approximation is

$$y = \frac{1.9733}{x} + 3.2818\sqrt{x}.$$

3.45 A function $f(x)$ is given at four points according to the table :

x	0	0.5	1	2
$f(x)$	1	3.52	3.73	– 1.27

Compute the values of a, b and the natural number n such that the sum

$$\sum_{i=1}^{4} [f(x_i) - a \sin(nx_i) - b]^2$$

is minimized. (Uppsala Univ., Sweden, BIT 27(1987) 628)

Solution

Using the method of least squares, the normal equations are obtained as

$$\sum_{i=1}^{4} [f(x_i) - a \sin(nx_i) - b] = 0,$$

$$\sum_{i=1}^{4} [f(x_i) - a \sin(nx_i) - b] \sin(nx_i) = 0,$$

$$\sum_{i=1}^{4} [f(x_i) - a \sin(nx_i) - b]\, x_i \cos(nx_i) = 0.$$

Substituting the values from the table of values, we get the equations

$$ap_3 + 4b = 6.98,$$
$$ap_2 + bp_3 = p_1,$$
$$aq_2 + bq_3 = q_1,$$

where
$$p_1 = 3.52 \sin(n/2) + 3.73 \sin(n) - 1.27 \sin(2n),$$
$$p_2 = \sin^2(n/2) + \sin^2(n) + \sin^2(2n),$$
$$p_3 = \sin(n/2) + \sin(n) + \sin(2n),$$
$$q_1 = 1.76 \cos(n/2) + 3.73 \cos(n) - 2.54 \cos(2n),$$
$$q_2 = \frac{1}{4} \sin(n) + \frac{1}{2} \sin(2n) + \sin(4n),$$
$$q_3 = \frac{1}{2} \cos(n/2) + \cos(n) + 2\cos(2n).$$

Solving for a and b from the second and third equations, we get

$$a = (p_1 q_3 - q_1 p_3) / (p_2 q_3 - q_2 p_3),$$
$$b = (p_2 q_1 - p_1 q_2) / (p_2 q_3 - q_2 p_3).$$

Substituting in the first equation, we get

$$f(n) = 6.98(p_2 q_3 - p_3 q_2) - p_3(p_1 q_3 - p_3 q_1) - 4(p_2 q_1 - p_1 q_2) = 0.$$

This is a nonlinear equation in n, whose solution (smallest natural number) may be obtained by the Newton-Raphson method. We have

$$n_{k+1} = n_k - \frac{f(n_k)}{f'(n_k)}.$$

It is found that the solution $n \to 0$, if the initial approximation $n_0 < 1.35$. Starting with $n_0 = 1.5$, we have the iterations as

k	n_k	f_k	f_k'
1	2.11483	– 9.3928	15.277
2	1.98029	1.4792	10.995
3	1.99848	– 0.2782	15.302
4	1.99886	– 0.5680(– 2)	14.677
5	1.99886	– 0.4857(– 5)	14.664

Hence, the smallest natural number is $n = 2$. The corresponding values of a and b are

$$a = 3.0013 \approx 3.0 \text{ and } b = 0.9980 \approx 1.0.$$

The approximation is

$$P(x) = 3 \sin (2x) - 1.$$

3.46 Let $l(x)$ be a straight line which is the best approximation of $\sin x$ in the sense of the method of least squares over the interval $[-\pi / 2, \pi / 2]$. Show that the residual $d(x) = \sin x - l(x)$ is orthogonal to any second degree polynomial. The scalar product is given by

$$(f, g) = \int_{-\pi/2}^{\pi/2} \bar{f}(x) g(x) dx.$$ (Uppsala Univ., Sweden, BIT 14(1974), 122)

Solution

Let $l(x) = a + bx$. We determine a and b such that

$$\int_{-\pi/2}^{\pi/2} [\sin x - a - bx]^2 \, dx = \text{minimum}.$$

We obtain the normal equations as

$$\int_{-\pi/2}^{\pi/2} [\sin x - a - bx] \, dx = 0,$$

$$\int_{-\pi/2}^{\pi/2} [\sin x - a - bx] \, x \, dx = 0$$

which give $a = 0$, $b = 24 / \pi^3$.

Hence, we find that $l(x) = \dfrac{24}{\pi^3} x$, and $d(x) = \sin x - \dfrac{24}{\pi^3} x$.

Using the given scalar product and taking $P_2(x) = Ax^2 + Bx + C$, it is easy to verify that

$$(f, g) = \int_{-\pi/2}^{\pi/2} (\bar{A}x^2 + \bar{B}x + \bar{C}) \left(\sin x - \frac{24x}{\pi^3} \right) dx$$

is always zero. Hence the result.

3.47 Experiments with a periodic process gave the following data :

$t°$	0	50	100	150	200	250	300	350
y	0.754	1.762	2.041	1.412	0.303	– 0.484	– 0.380	0.520

Estimate the parameters a and b in the model $y = b + a \sin t$, using the least squares approximation. (Lund Univ., Sweden, BIT 21(1981), 242)

Solution

We determine a and b such that

$$\sum_{i=1}^{8} [y_i - b - a \sin t_i]^2 = \text{minimum}.$$

The normal equations are given by

$$8b + a \sum_{i=1}^{8} \sin t_i = \sum_{i=1}^{8} y_i ,$$

$$b \sum_{i=1}^{8} \sin t_i + a \sum_{i=1}^{8} \sin^2 t_i = \sum_{i=1}^{8} y_i \sin t_i .$$

Using the given data values (after converting degrees to radians) we obtain

$$8b - 0.070535\, a = 5.928$$

$$0.070535\, b - 3.586825\, a = -4.655735$$

which give $\quad a = 1.312810, \; b = 0.752575.$

The least squares approximation is

$$y = 0.752575 + 1.31281 \sin t.$$

3.48 A physicist wants to approximate the following data :

x	0.0	0.5	1.0	2.0
$f(x)$	0.00	0.57	1.46	5.05

using a function $a\, e^{bx} + c$. He believes that $b \approx 1$.

(*i*) Compute the values of a and c that give the best least squares approximation assuming $b = 1$.

(*ii*) Use these values of a and c to obtain a better value of b.

(Uppsala Univ., Sweden, BIT 17(1977), 369)

Solution

(*i*) We take $b = 1$ and determine a and c such that

$$\sum_{i=1}^{4} [f(x_i) - a e^{x_i} - c]^2 = \text{minimum}.$$

We obtain the normal equations as

$$\sum_{i=1}^{4} f(x_i) - a \sum_{i=1}^{4} e^{x_i} - 4c = 0,$$

$$\sum_{i=1}^{4} f(x_i) e^{x_i} - a \sum_{i=1}^{4} e^{2x_i} - c \sum_{i=1}^{4} e^{x_i} = 0.$$

Using the given data values, we get

$$12.756059a + 4c = 7.08$$

$$65.705487a + 12.756059c = 42.223196$$

which has the solution, $\quad a = 0.784976, \; c = -0.733298.$

The approximation is

$$f(x) = 0.784976\, e^x - 0.733298.$$

(*ii*) Taking the approximation as

$$f(x) = 0.784976\, e^{bx} - 0.733298,$$

we now determine b such that

$$\sum_{i=1}^{4} [f(x_i) - pe^{bx_i} - q]^2 = \text{minimum}$$

where $p = 0.784976$ and $q = -0.733298$. We obtain the normal equation

$$\sum_{i=1}^{4} [f(x_i) - pe^{bx_i} - q]\, px_i e^{bx_i} = 0$$

or $$\sum_{i=1}^{4} x_i \{f(x_i) - q\} e^{bx_i} - p \sum_{i=1}^{4} x_i e^{2bx_i} = 0$$

which becomes on simplification

$$F(b) = 0.651649\, e^{b/2} + 1.80081\, e^{b} + 10.78162\, e^{2b} - 1.569952\, e^{4b} = 0$$

We shall determine b by the Newton-Raphson method. Starting with $b_0 = 1$, we obtain

$$b_1 = 1 - \frac{F(1)}{F'(1)} = 1 + \frac{0.080983}{-178.101606} = 0.9995.$$

Hence, $\quad b = 0.9995.$

3.49 We are given the following values of a function f of the variable t :

t	0.1	0.2	0.3	0.4
f	0.76	0.58	0.44	0.35

Obtain a least squares fit of the form $f = ae^{-3t} + be^{-2t}$.

(Royal Inst. Tech., Stockholm, Sweden, BIT 17(1977), 115)

Solution

We determine a and b such that

$$I = \sum_{i=1}^{4} [f_i - ae^{-3t_i} - be^{-2t_i}]^2 = \text{minimum.}$$

We obtain the following normal equations

$$\sum_{i=1}^{4} (f_i - ae^{-3t_i} - be^{-2t_i}) e^{-3t_i} = 0,$$

$$\sum_{i=1}^{4} (f_i - ae^{-3t_i} - be^{-2t_i}) e^{-2t_i} = 0,$$

or $$a \sum_{i=1}^{4} e^{-6t_i} + b \sum_{i=1}^{4} e^{-5t_i} = \sum_{i=1}^{4} f_i\, e^{-3t_i},$$

$$a \sum_{i=1}^{4} e^{-5t_i} + b \sum_{i=1}^{4} e^{-4t_i} = \sum_{i=1}^{4} f_i\, e^{-2t_i}.$$

Using the given data values, we obtain

$$1.106023a + 1.332875b = 1.165641,$$
$$1.332875a + 1.622740b = 1.409764,$$

which has the solution $a = 0.68523,\ b = 0.30593.$

The least squares solution is

$$f(t) = 0.68523\, e^{-3t} + 0.30593\, e^{-2t}.$$

3.50 The second degree polynomial $f(x) = a + bx + cx^2$ is determined from the condition

$$d = \sum_{i=m}^{n} [f(x_i) - y_i]^2 = \text{minimum}$$

where (x_i, y_i), $i = m(1)n$, $m < n$, are given real numbers. Putting $X = x - \xi$, $Y = y - \eta$, $X_i = x_i - \xi$, $Y_i = y_i - \eta$, we determine $F(x) = A + BX + CX^2$ from the condition

$$D = \sum_{i=m}^{n} [F(X_i) - Y_i]^2 = \text{minimum}.$$

Show that $F(X) = f(x) - \eta$. Also derive explicit formula for $F'(0)$ expressed in Y_i when $X_i = ih$ and $m = -n$. (Bergen Univ., Sweden, BIT 7(1967), 247)

Solution

We have
$$\begin{aligned} f(x) - y &= a + bx + cx^2 - y \\ &= a + b(X + \xi) + c(X + \xi)^2 - Y - \eta \\ &= (a + b\xi + c\xi^2) + (b + 2c\xi)X + cX^2 - Y - \eta \\ &= (A + BX + CX^2 - \eta) - Y = F(X) - Y \end{aligned}$$

Hence,
$$\begin{aligned} F(X) &= A + BX + CX^2 - \eta \\ &= (a + b\xi + c\xi^2) + (b + 2c\xi)X + CX^2 - \eta = f(x) - \eta. \end{aligned}$$

Now, $F'(0) = B.$

The normal equations are, when $m = -n$,

$$(2n + 1)A + B \sum X_i + C \sum X_i^2 = \sum Y_i,$$
$$A \sum X_i + B \sum X_i^2 + C \sum X_i^3 = \sum X_i Y_i,$$
$$A \sum X_i^2 + B \sum X_i^3 + C \sum X_i^4 = \sum X_i^2 Y_i.$$

Since $X_i = ih$, we have

$$\sum_{-n}^{n} X_i = 0, \sum_{-n}^{n} X_i^3 = 0 \text{ and}$$

$$\sum_{-n}^{n} X_i^2 = \sum_{-n}^{n} i^2 h^2 = 2h^2 \sum_{1}^{n} i^2 = \frac{1}{3} h^2 n(n + 1)(2n + 1).$$

The second normal equation gives

$$B \sum X_i^2 = \sum ihY_i$$

$$B = \frac{3 \sum iY_i}{hn(n+1)(2n+1)}.$$

3.51 A function is approximated by a piecewise linear function in the sense that

$$\int_0^1 \left[f(x) - \sum_{i=0}^{10} a_i \phi_i(x) \right]^2 dx$$

is minimized, where the shape functions ϕ_i are defined by

$$\phi_0 = \begin{vmatrix} 1 - 10x, & 0 \le x \le 0.1, \\ 0, & \text{otherwise.} \end{vmatrix}$$

$$\phi_{10} = \begin{vmatrix} 10x - 9, & 0.9 \le x \le 1.0, \\ 0, & \text{otherwise.} \end{vmatrix}$$

$$\phi_i = \begin{vmatrix} 10(x - x_{i-1}), & x_{i-1} \le x \le x_i \\ 10(x_{i+1} - x), & x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise} \end{vmatrix}$$

$x_i = 0.1i$, $i = 1(1)9$.

Write down the coefficient matrix of the normal equations.

(Uppsala Univ., Sweden, BIT 19(1979), 552)

Solution

We obtain the normal equations as

$$\int_0^1 \left[f(x) - \sum_{i=0}^{10} a_i \phi_i(x) \right] \phi_j(x) dx = 0, j = 0, 1, ..., 10.$$

For $j = 0$, we have

$$\int_0^{0.1} [f(x) - a_0(1 - 10x) - a_1(10x)](1 - 10x) dx = 0$$

which gives

$$a_0 \int_0^{0.1} (1 - 10x)^2 dx + 10a_1 \int_0^{0.1} x(1 - 10x) dx = \int_0^{0.1} (1 - 10x) f(x) dx$$

or
$$\frac{1}{30} a_0 + \frac{1}{60} a_1 = I_0 = \int_0^{0.1} f(x)(1 - 10x) dx.$$

For $j = 2, 3, ..., 8$, we have

$$\int_{x_{j-1}}^{x_j} [f(x) - a_{j-1}\phi_{j-1}(x) - a_j \phi_j(x) - a_{j+1}\phi_{j+1}(x)] \phi_j(x) dx$$

$$+ \int_{x_j}^{x_{j+1}} [f(x) - a_{j-1}\phi_{j-1}(x) - a_j\phi_j(x) - a_{j+1}\phi_{j+1}(x)]\phi_j(x) dx = 0.$$

Using the given expressions for ϕ_j, we obtain

$$\int_{x_{j-1}}^{x_j} [f(x) - 10a_{j-1}(x_j - x) - 10a_j(x - x_{j-1})] \, 10(x - x_{j-1}) dx$$

$$+ \int_{x_j}^{x_{j+1}} [f(x) - 10a_j(x_{j+1} - x) - 10a_{j+1}(x - x_j)] \, 10(x_{j+1} - x) dx = 0$$

or
$$100a_{j-1} \int_{x_{j-1}}^{x_j} (x_j - x)(x - x_{j-1}) dx + 100a_j \int_{x_{j-1}}^{x_j} (x - x_{j-1})^2 dx$$

$$+ 100a_j \int_{x_j}^{x_{j+1}} (x_{j+1} - x)^2\, dx + 100a_{j+1} \int_{x_j}^{x_{j+1}} (x - x_j)(x_{j+1} - x)dx$$

$$= 10 \int_{x_{j-1}}^{x_j} f(x)(x - x_{j-1}) + 10 \int_{x_j}^{x_{j+1}} f(x)(x_{j+1} - x)dx$$

which gives

$$\frac{1}{60} a_{j-1} + \frac{4}{60} a_j + \frac{1}{60} a_{j+1} = I_j .$$

For $j = 1$ and $j = 9$, we obtain respectively

$$\frac{1}{60} a_0 + \frac{4}{60} a_1 + \frac{1}{60} a_2 = I_1, \quad \text{and} \quad \frac{1}{60} a_8 + \frac{4}{60} a_9 + \frac{1}{60} a_{10} = I_9$$

Similarly, for $j = 10$, we have

$$\int_{0.9}^{1.0} [f(x) - a_9(10 - 10x) - a_{10}(10x - 9)](10x - 9)dx = 0$$

or $$a_9 \int_{0.9}^{1.0} (10 - 10x)(10x - 9)dx + a_{10} \int_{0.9}^{1.0} (10x - 9)^2\, dx = \int_{0.9}^{1.0} f(x)(10x - 9)dx = I_{10}$$

or $$\frac{1}{60} a_9 + \frac{1}{30} a_{10} = I_{10} .$$

Assembling the above equations for $j = 0, 1, ..., 10$, we obtain

$$\mathbf{Aa = b}$$

where $$\mathbf{a} = (a_0, a_1, ..., a_{10})^T, \mathbf{b} = (I_0, I_1, ..., I_{10})^T,$$

and $$\mathbf{A} = \frac{1}{60} \begin{bmatrix} 2 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 4 & 1 & 0 & \ldots & 0 \\ 0 & 1 & 4 & 1 & \ldots & 0 \\ \vdots & & & & & \\ 0 & 0 & \ldots & 1 & 4 & 1 \\ 0 & 0 & \ldots & 0 & 1 & 2 \end{bmatrix}$$

3.52 Polynomials $P_r(x)$, $r = 0(1)n$, are defined by

$$\sum_{j=0}^{n} P_r(x_j)P_s(x_j) \begin{vmatrix} = 0, & r \neq s \\ \neq 0, & r = s \end{vmatrix} \quad r, s \leq n$$

$$x_j = -1 + \frac{2j}{n}, j = 0(1)n$$

subject also to $P_r(x)$ being a polynomial of degree r with leading term x^r. Derive a recurrence relation for these polynomials and obtain $P_0(x)$, $P_1(x)$, $P_2(x)$ when $n = 4$. Hence, obtain coefficients a_0, a_1, a_2 which minimize

$$\sum_{j=0}^{4} [(1 + x_j^2)^{-1} - (a_0 + a_1 x_j + a_2 x_j^2)]^2.$$

Solution

As $xP_k(x)$ is a polynomial of degree $k + 1$, we can write it in the form

$$xP_k(x) = d_0 P_0(x) + d_1 P_1(x) + ... + d_{k+1} P_{k+1}(x) \tag{3.60}$$

$$= \sum_{r=0}^{k+1} d_r P_r(x)$$

where $$d_{k+1} = 1.$$

We obtain from (3.60)

$$\sum_{j=0}^{n} x_j P_k(x_j) P_s(x_j) = \sum_{j=0}^{n} \sum_{r=0}^{k+1} d_r P_r(x_j) P_s(x_j). \tag{3.61}$$

Using the orthogonality conditions, we get from (3.61)

$$d_s = \sum_{j=0}^{n} x_j P_k(x_j) P_s(x_j) \Big/ \sum_{j=0}^{n} P_s^2(x_j). \tag{3.62}$$

If $s < k - 1$, then $xP_s(x)$ is a polynomial of degree $s + 1 < k$ and because of orthogonality conditions, we get

$$d_s = 0, \; s < k - 1.$$

Hence, we have from (3.60)

$$xP_k(x) = d_{k-1}P_{k-1}(x) + d_k P_k(x) + d_{k+1}P_{k+1}(x). \tag{3.63}$$

Since $d_{k+1} = 1$, we can also write the recurrence relation (3.63) in the form

$$P_{k+1}(x) = (x - b_k)P_k(x) - c_k P_{k-1}(x) \tag{3.64}$$

where $$b_k = d_k = \sum_{j=0}^{n} x_j P_k^2(x_j) \Big/ \sum_{j=0}^{n} P_k^2(x_j) \tag{3.65}$$

$$c_k = d_{k-1} = \sum_{j=0}^{n} x_j P_k(x_j) P_{k-1}(x_j) \Big/ \sum_{j=0}^{n} P_{k-1}^2(x_j). \tag{3.66}$$

We also have from (3.64)

$$P_k(x) = (x - b_{k-1})P_{k-1}(x) - c_{k-1}P_{k-2}(x)$$

and $$\sum_{j=0}^{n} P_k^2(x_j) = \sum_{j=0}^{n} (x_j - b_{k-1})P_{k-1}(x_j)P_k(x_j) - c_{k-1} \sum_{j=0}^{n} P_{k-2}(x_j)P_k(x_j)$$

which gives $\displaystyle\sum_{j=0}^{n} x_j P_{k-1}(x_j) P_k(x_j) = \sum_{j=0}^{n} P_k^2(x_j).$

Using (3.66), we have

$$c_k = \sum_{j=0}^{n} P_k^2(x_j) \Big/ \sum_{j=0}^{n} P_{k-1}^2(x_j). \tag{3.67}$$

Thus, (3.64) is the required recurrence relation, where b_k and c_k are given by (3.65) and (3.67) respectively.

For $n = 4$, we have

$$x_j = -1 + \frac{2j}{4}, j = 0, 1, ..., 4$$

or $$x_0 = -1, x_1 = -\frac{1}{2}, x_2 = 0, x_3 = \frac{1}{2}, x_4 = 1.$$

$$\sum_{j=0}^{4} x_j = \sum_{j=0}^{4} x_j^3 = 0 \quad \text{and} \quad \sum_{j=0}^{4} x_j^2 = \frac{5}{2}.$$

Using the recurrence relation (3.64) together with $P_0(x) = 1$, $P_{-1}(x) = 0$, we obtain $P_1(x_0) = x - b_0$

where $$b_0 = \sum_{j=0}^{4} x_j P_0^2(x_j) \Big/ \sum_{j=0}^{4} P_0^2(x_j) = \frac{1}{5} \sum_{j=0}^{4} x_j = 0.$$

Hence, we have $P_1(x) = x$.

Similarly, we obtain

$$P_2(x) = (x - b_1)\, x - c_1$$

where $$b_1 = \sum_{j=0}^{4} x_j P_1^2(x_j) \Big/ \sum_{j=0}^{4} P_1^2(x_j) = \sum_{j=0}^{4} x_j^3 \Big/ \sum_{j=0}^{4} x_j^2 = 0,$$

$$c_1 = \sum_{j=0}^{4} P_1^2(x_j) \Big/ \sum_{j=0}^{4} P_0^2(x_j) = \frac{1}{5} \sum_{j=0}^{4} x_j^2 = \frac{1}{2}.$$

Hence, we get $P_2(x) = x^2 - \dfrac{1}{2}$.

For the problem

$$\sum_{j=0}^{4} [(1 + x_j^2)^{-1} - (d_0 P_0(x_j) + d_1 P_1(x_j) + d_2 P_2(x_j))]^2 = \text{minimum},$$

we obtain the normal equations as

$$\sum_{j=0}^{4} [(1 + x_j^2)^{-1} - (d_0 P_0(x_j) + d_1 P_1(x_j) + d_2 P_2(x_j))]\, P_0(x_j) = 0,$$

$$\sum_{j=0}^{4} [(1 + x_j^2)^{-1} - (d_0 P_0(x_j) + d_1 P_1(x_j) + d_2 P_2(x_j))]\, P_1(x_j) = 0,$$

$$\sum_{j=0}^{4} [(1 + x_j^2)^{-1} - (d_0 P_0(x_j) + d_1 P_1(x_j) + d_2 P_2(x_j))]\, P_2(x_j) = 0.$$

The solution of this system is

$$d_0 = \sum_{j=0}^{4} \left[\frac{P_0(x_j)}{1 + x_j^2}\right] \Big/ \sum_{j=0}^{4} P_0^2(x_j) = \frac{18}{25},$$

$$d_1 = \sum_{j=0}^{4} \left[\frac{P_1(x_j)}{1 + x_j^2}\right] \Big/ \sum_{j=0}^{4} P_1^2(x_j) = 0,$$

$$d_2 = \sum_{j=0}^{4} \left[\frac{P_2(x_j)}{1 + x_j^2}\right] \Big/ \sum_{j=0}^{4} P_2^2(x_j) = -\frac{16}{35}.$$

The approximation is given by

$$d_0 P_0(x) + d_1 P_1(x) + d_2 P_2(x) = \frac{18}{25} - \frac{16}{35}\left(x^2 - \frac{1}{2}\right) = \frac{166}{175} - \frac{16}{35} x^2.$$

Hence, we have $a_0 = 166/175$, $a_1 = 0$, $a_2 = -16/35$, and the polynomial is

$$P(x) = (166 - 80x^2)/175.$$

3.53 Find suitable values of $a_0, ..., a_4$ so that $\sum_{r=0}^{4} a_r T_r^*(x)$ is a good approximation to $1/(1+x)$ for $0 \le x \le 1$. Estimate the maximum error of this approximation.

(Note : $T_r^*(x) = \cos r\theta$ where $\cos \theta = 2x - 1$).

Solution

It can be easily verified that $T_r^*(x)$ are orthogonal with respect to the weight function

$$W(x) = 1 / \sqrt{x(1-x)} \text{ and}$$

$$\int_0^1 \frac{T_m^*(x) T_n^*(x) dx}{\sqrt{x(1-x)}} = \begin{vmatrix} 0, & m \ne n, \\ \pi/2, & m = n \ne 0, \\ \pi, & m = n = 0. \end{vmatrix}$$

Writing

$$\frac{1}{1+x} = a_0 T_0^*(x) + a_1 T_1^*(x) + ... + a_4 T_4^*(x)$$

we obtain
$$a_r = \int_0^1 \frac{T_r^*(x) dx}{(1+x)\sqrt{x(1-x)}} \Big/ \int_0^1 \frac{[T_r^*(x)]^2 dx}{\sqrt{x(1-x)}}, \; r = 0, 1, ..., 4$$

which gives
$$a_0 = \frac{1}{\pi} \int_0^1 \frac{T_0^*(x) dx}{(1+x)\sqrt{x(1-x)}}$$

$$a_r = \frac{2}{\pi} \int_0^1 \frac{T_r^*(x) dx}{(1+x)\sqrt{x(1-x)}}, \; r = 1, 2, 3, 4.$$

We have
$$a_0 = \frac{1}{\pi} \int_0^1 \frac{dx}{(1+x)\sqrt{x(1-x)}} = \frac{2}{\pi} \int_0^{\pi/2} \frac{d\theta}{1 + \sin^2 \theta} = \frac{2}{\pi} I$$

where
$$I = \int_0^{\pi/2} \frac{1}{1 + \sin^2 \theta} d\theta = \frac{\pi}{2\sqrt{2}}.$$

Hence, we have $a_0 = 1/\sqrt{2}$.

Similarly, we have,

$$a_1 = \frac{2}{\pi} \int_0^1 \frac{(2x-1) dx}{(1+x)\sqrt{x(1-x)}} = \frac{4}{\pi} \int_0^{\pi/2} \frac{2 \sin^2 \theta - 1}{(1 + \sin^2 \theta)} d\theta$$

$$= \frac{8}{\pi} \left[\int_0^{\pi/2} d\theta - \frac{3}{2} I \right] = 4 - 3\sqrt{2},$$

$$a_2 = \frac{2}{\pi} \int_0^1 \frac{[2(2x-1)^2 - 1]}{(1+x)\sqrt{x(1-x)}} dx = \frac{4}{\pi} \int_0^{\pi/2} \frac{[8 \sin^4 \theta - 8 \sin^2 \theta + 1]}{1 + \sin^2 \theta} d\theta$$

$$= \frac{4}{\pi} \left[\int_0^{\pi/2} (8 \sin^2 \theta - 16) d\theta + 17 I \right] = 17\sqrt{2} - 24,$$

$$a_3 = \frac{2}{\pi}\int_0^1 \frac{[4(2x-1)^3 - 3(2x-1)]}{(1+x)\sqrt{x(1-x)}}\,dx$$

$$= \frac{4}{\pi}\int_0^{\pi/2} \frac{[4(2\sin^2\theta - 1)^3 - 3(2\sin^2\theta - 1)]}{1+\sin^2\theta}\,d\theta$$

$$= \frac{4}{\pi}\left[\int_0^{\pi/2}(32\sin^4\theta - 80\sin^2\theta + 98)d\theta - 99I\right] = 140 - 99\sqrt{2},$$

$$a_4 = \frac{2}{\pi}\int_0^1 \frac{[8(2x-1)^4 - 8(2x-1)^2 + 1]}{(1+x)\sqrt{x(1-x)}}\,dx$$

$$= \frac{4}{\pi}\int_0^{\pi/2} \frac{[8(2\sin^2\theta - 1)^4 - 8(2\sin^2\theta - 1)^2 + 1]}{1+\sin^2\theta}\,d\theta$$

$$= \frac{4}{\pi}\left[\int_0^{\pi/2}(128\sin^6\theta - 384\sin^4\theta + 544\sin^2\theta - 576)d\theta + 577I\right]$$

$$= 577\sqrt{2} - 816.$$

The maximum error in the approximation is given by $|\, a_5 \,|$. We have

$$a_5 = \frac{2}{\pi}\int_0^1 \frac{T_5^*(x)dx}{(1+x)\sqrt{x(1-x)}}$$

$$= \frac{4}{\pi}\int_0^{\pi/2} \frac{16(2\sin^2\theta - 1)^5 - 20(2\sin^2\theta - 1)^3 + 5(2\sin^2\theta - 1)}{1+\sin^2\theta}\,d\theta$$

$$= \frac{4}{\pi}\left[\int_0^{\pi/2}(512\sin^8\theta - 1792\sin^6\theta + 2912\sin^4\theta - 3312\sin^2\theta + 3362)d\theta - 3363I\right]$$

$$= 4756 - 3363\sqrt{2}.$$

Hence, $|\, a_5 \,| = 3363\sqrt{2} - 4756 \approx 0.0002.$

Chebyshev Polynomials

3.54 Develop the function $f(x) = \ln[(1 + x) / (1 - x)] / 2$ in a series of Chebyshev polynomials.

(Lund Univ., Sweden, BIT 29 (1989), 375)

Solution

We write $$f(x) = \sum_{r=0}^{\infty}{}' a_r T_r(x)$$

where Σ' denotes a summation whose first term is halved. Using the orthogonal properties of the Chebyshev polynomials, we obtain

$$a_r = \frac{2}{\pi}\int_{-1}^{1}(1-x^2)^{-1/2} f(x)T_r(x)dx = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos r\theta\, d\theta.$$

Since, $$f(x) = \frac{1}{2}\ln\frac{1+x}{1-x},$$

we have $$f(\cos\theta) = \ln(\cot\theta / 2).$$

Hence, we have $$a_r = \frac{2}{\pi}\int_0^{\pi}\ln(\cot(\theta/2))\cos r\theta\, d\theta.$$

For $r = 0$, we get

$$a_0 = \frac{2}{\pi}\int_0^\pi \ln\left(\cot\left(\frac{\theta}{2}\right)\right)d\theta = \frac{2}{\pi}\int_0^\pi \ln\left(\tan\left(\frac{\theta}{2}\right)\right)d\theta = -\frac{2}{\pi}\int_0^\pi \ln\left[\cot\left(\frac{\theta}{2}\right)\right]d\theta$$

Hence, $a_0 = 0$.

Integrating by parts, we have (for $r \neq 0$)

$$a_r = \frac{2}{\pi}\left[\left\{\frac{1}{r}\sin r\theta \ln\cot\left(\frac{\theta}{2}\right)\right\}_0^\pi + \frac{1}{2r}\int_0^\pi \sin r\theta \tan\left(\frac{\theta}{2}\right)\text{cosec}^2\left(\frac{\theta}{2}\right)d\theta\right]$$

$$= \frac{2}{\pi r}\int_0^\pi \frac{\sin r\theta}{\sin\theta}\,d\theta = \frac{2}{\pi r} I_r$$

We get $I_1 = \pi$, and $a_1 = 2$. For $r \geq 2$, we have

$$I_r = \int_0^\pi \frac{\sin r\theta}{\sin\theta}\,d\theta = \int_0^\pi \frac{\sin(r-1)\theta\cos\theta + \cos(r-1)\theta\sin\theta}{\sin\theta}\,d\theta$$

$$= \frac{1}{2}\int_0^\pi \frac{2\sin(r-1)\theta\cos\theta}{\sin\theta}\,d\theta + \int_0^\pi \cos(r-1)\theta\,d\theta$$

$$= \frac{1}{2}\int_0^\pi \frac{\sin r\theta + \sin(r-2)\theta}{\sin\theta}\,d\theta = \frac{1}{2}I_r + \frac{1}{2}I_{r-2}.$$

Hence, we get

$$I_r = I_{r-2} = I_{r-4} = \ldots = I_0, \text{ if } r \text{ is even,}$$

and $\quad I_r = I_{r-2} = I_{r-4} = \ldots = I_1$, if r is odd.

Hence, $a_r = 0$ if r is even and $a_r = 2 / r$ if r is odd.

Therefore, we get

$$\frac{1}{2}\ln\frac{1+x}{1-x} = 2\left[T_1 + \frac{1}{3}T_3 + \frac{1}{5}T_5 + \ldots\right]$$

3.55 Let $T_n(x)$ denote the Chebyshev polynomial of degree n. Values of an analytic function, $f(x)$ in the interval $-1 \leq x \leq 1$, are calculated from the series expansion

$$f(x) = \sum_{r=0}^{\infty} a_r T_{2r}(x)$$

the first few coefficients of which are displayed in the table :

r	a_r	r	a_r
0	0.3155	5	– 0.0349
1	– 0.0087	6	0.0048
2	0.2652	7	– 0.0005
3	– 0.3701	8	0.0003
4	0.1581		

(*a*) Calculate $f(1)$ with the accuracy allowed by the table.

(*b*) Show the relation $T_{rs}(x) = T_r(T_s(x))$, $-1 \leq x \leq 1$. With $s = 2$, the series above can be written as

$$f(x) = \sum_{r=0}^{\infty} a_r T_r(2x^2 - 1).$$

Use this series to calculate $f(\sqrt{2} / 2)$ and after differentiation, $f'(\sqrt{2} / 2)$.

(*c*) $f(x)$ has got a zero close to $x_0 = \sqrt{2} / 2$.

Use Newton-Raphson method and the result of (*b*) to get a better approximation to this zero. (Inst. Tech., Lyngby, Denmark, BIT 24(1984), 397)

Solution

(*a*) We are given that

$$T_{2r}(x) = \cos(2r \cos^{-1} x).$$

Hence, we have

$$T_{2r}(1) = \cos(2r \cos^{-1} 1) = \cos(0) = 1 \text{ for all } r.$$

We have therefore,

$$f(1) = \sum_{r=0}^{\infty} a_r = 0.3297.$$

(*b*) We have

$$T_s(x) = \cos(s \cos^{-1} x)$$

$$T_r(T_s(x)) = \cos(r \cos^{-1}\{\cos(s \cos^{-1} x)\})$$

$$= \cos(rs \cos^{-1} x) = T_{rs}(x).$$

We get from

$$f(x) = \sum_{r=0}^{\infty} a_r T_r(2x^2 - 1)$$

$$f\left(\frac{\sqrt{2}}{2}\right) = \sum_{r=0}^{\infty} a_r T_r(0) = \sum_{r=0}^{\infty} a_r \cos(\pi r / 2)$$

$$= a_0 - a_2 + a_4 - \ldots = 0.2039.$$

We have

$$f'(x) = \sum_{r=0}^{\infty} a_r T_r'(2x^2 - 1)(4x)$$

and

$$f'\left(\frac{\sqrt{2}}{2}\right) = \frac{4}{\sqrt{2}} \sum_{r=0}^{\infty} a_r T_r'(0).$$

We also have

$$T_r(x) = \cos r\theta, \quad \theta = \cos^{-1} x$$

$$T_r'(x) = \frac{r \sin r\theta}{\sin \theta}.$$

Hence, we obtain

$$T_r'(0) = \frac{r \sin(r\pi/2)}{\sin(\pi/2)} = r \sin\left(\frac{r\pi}{2}\right).$$

We thus obtain

$$f'\left(\frac{\sqrt{2}}{2}\right) = \frac{4}{\sqrt{2}} [a_1 - 3a_3 + 5a_5 - \ldots] = 2.6231.$$

(*c*) Taking $x_0 = \sqrt{2}/2$ and using Newton-Raphson method, we have

$$x^* = x_0 - \frac{f(x_0)}{f'(x_0)} = \frac{\sqrt{2}}{2} - \frac{0.2039}{2.6231} = 0.6296.$$

Uniform (minimax) Approximation

3.56 Determine as accurately as possible a straight line $y = ax + b$, approximating $1 / x^2$ in the Chebyshev sense on the interval [1, 2]. What is the maximal error ? Calculate a and b to two correct decimals. (Royal Inst. Tech., Stockholm, Sweden, BIT 9(1969), 87)

Solution

We have the error of approximation as

$$\varepsilon(x) = \frac{1}{x^2} - ax - b.$$

Choosing the points 1, α and 2 and using Chebyshev equi-oscillation theorem, we have

$$\varepsilon(1) + \varepsilon(\alpha) = 0,$$
$$\varepsilon(\alpha) + \varepsilon(2) = 0,$$
$$\varepsilon'(\alpha) = 0,$$

or

$$1 - a - b + \frac{1}{\alpha^2} - a\alpha - b = 0,$$

$$\frac{1}{4} - 2a - b + \frac{1}{\alpha^2} - a\alpha - b = 0,$$

$$\frac{2}{\alpha^3} + a = 0.$$

Subtracting the first two equations, we get $a = -3/4$. From the third equation, we get $\alpha^3 = 8/3$. From the first equation, we get

$$2b = \frac{7}{4} + \frac{3}{4}\alpha + \frac{1}{\alpha^2} = \frac{7}{4} + \frac{3}{4}\alpha + \frac{3}{8}\alpha = \frac{7}{4} + \frac{9}{8}\alpha.$$

Hence, $b \approx 1.655$.

The maximal error in magnitude is $|\varepsilon(1)| = |\varepsilon(\alpha)| = |\varepsilon(2)|$.

Hence, max. error $= \varepsilon(1) = 1 - a - b \approx 0.095$.

3.57 Suppose that we want to approximate a continuous function $f(x)$ on $|x| \le 1$ by a polynomial $P_n(x)$ of degree n. Suppose further that we have found

$$f(x) - P_n(x) = \alpha_{n+1} T_{n+1}(x) + r(x),$$

where $T_{n+1}(x)$ denotes the Chebyshev polynomial of degree $n + 1$ with

$$\frac{1}{2^{n+1}} \le |\alpha_{n+1}| \le \frac{1}{2^n},$$

and

$$|r(x)| \le |\alpha_{n+1}| / 10, \quad |x| \le 1.$$

Show that

$$\frac{0.4}{2^n} \le \max_{|x| \le 1} |f(x) - P_n^*(x)| \le \frac{1.1}{2^n}$$

where, finally $P_n^*(x)$ denotes the optimal polynomial of degree n for $f(x)$ on $|x| \le 1$.

(Uppsala Univ., Sweden, BIT 13 (1973), 375)

Solution

We have $$\frac{1}{2^{n+1}} \le |\alpha_{n+1}| \le \frac{1}{2^n}; \quad |T_{n+1}(x)| \le 1$$

and $$|\ r(x)\ | \le \frac{1}{2^n(10)}, \quad \text{or} \quad -\frac{0.1}{2^n} \le r(x) \le \frac{0.1}{2^n}.$$

From the given equation

$$\max_{|x|\le 1} |\ f(x) - P_n(x)\ | \ge |\ \alpha_{n+1}\ | - |\ r(x)\ | \ge \frac{1}{2^{n+1}} - \frac{1}{2^n(10)}$$

Also, $\max\limits_{|x|\le 1} |\ f(x) - P_n(x)\ | \le |\ \alpha_{n+1}\ | + |\ r(x)\ | \le \dfrac{1}{2^n} + \dfrac{1}{2^n(10)}$.

Hence, $$\frac{1}{2^{n+1}} - \frac{0.1}{2^n} \le \max_{|x|\le 1} |f(x) - P_n^*(x)| \le \frac{1}{2^n} + \frac{0.1}{2^n}$$

which gives $$\frac{0.4}{2^n} \le \max_{|x|\le 1} |f(x) - P_n^*(x)| \le \frac{1.1}{2^n}.$$

3.58 Determine the polynomial of second degree, which is the best approximation in maximum norm to $\sqrt{x}$ on the point set $\{0, 1/9, 4/9, 1\}$.

(Gothenburg Univ., Sweden, BIT 8 (1968), 343)

Solution

We have the error function as

$$\varepsilon(x) = ax^2 + bx + c - \sqrt{x}.$$

Using the Chebyshev equioscillation theorem, we have

$$\varepsilon(0) + \varepsilon(1/9) = 0,$$
$$\varepsilon(1/9) + \varepsilon(4/9) = 0,$$
$$\varepsilon(4/9) + \varepsilon(1) = 0,$$

which give $$\frac{a}{81} + \frac{1}{9}b + 2c = \frac{1}{3},$$

$$\frac{17}{81}a + \frac{5}{9}b + 2c = 1,$$

$$\frac{97}{81}a + \frac{13}{9}b + 2c = \frac{5}{3}.$$

The solution of this system is

$$a = -9/8, b = 2, c = 1/16.$$

Hence, the best polynomial approximation is

$$P_2(x) = \frac{1}{16}(1 + 32x - 18x^2).$$

3.59 Two terms of the Taylor expansion of e^x around $x = a$ are used to approximate e^x on the interval $0 \le x \le 1$. How should a be chosen ao as to minimize the error in maximum norm ? Compute a correct to two decimal places.

(Lund Univ., Sweden, BIT 12 (1972), 589)

Solution

We have $$e^x \approx e^a + (x - a)e^a = Ax + B$$
where $A = e^a$ and $B = (1 - a)e^a$.

We approximate e^x by $Ax + B$ such that

$$\max_{0 \le x \le 1} | e^x - Ax - B | = \text{minimum}.$$

Defining the error function

$$\varepsilon(x) = e^x - Ax - B$$

on the points 0, α, 1 and using the Chebyshev equioscillation theorem, we get

$$\varepsilon(0) + \varepsilon(\alpha) = 0, \quad \text{or} \quad (1 - B) + (e^\alpha - A\alpha - B) = 0,$$
$$\varepsilon(\alpha) + \varepsilon(1) = 0, \quad \text{or} \quad (e^\alpha - A\alpha - B) + (e - A - B) = 0,$$
$$\varepsilon'(\alpha) = 0, \quad \text{or} \quad e^\alpha - A = 0.$$

Subtracting the first two equations, we get $A = e - 1$.

Third equation gives $e^\alpha = A = e - 1$. Since, $A = e^a$, we get $\alpha = a = \ln (e - 1) \approx 0.5412$.

First equation gives $B = \frac{1}{2}\,[1 + (1 - \alpha)e^a] \approx 0.8941$

Hence, the best approximation is

$$P(x) = (e - 1)x + 0.7882.$$

3.60 Calculate $\min_p \max_{0 \le x \le 1} | f(x) - P(x) |$ where P is a polynomial of degree at most 1 and

$$f(x) = \int_x^1 \frac{x^2}{y^3}\, dy$$ (Uppsala Univ., Sweden, BIT 18 (1978), 236)

Solution

We have $$f(x) = \int_x^1 \frac{x^2}{y^3}\, dy = -\frac{x^2}{2} + \frac{1}{2}.$$

Let $f(x)$ be approximated by $P_0(x) = C$, where $C = (m + M) / 2$, and

$$m = \min_{0 \le x \le 1} [\, f(x)\,] = 0, M = \max_{0 \le x \le 1} [\, f(x)\,] = \frac{1}{2}.$$

Hence, we get the approximation as $P_0(x) = 1 / 4$ and

$$\max_{0 \le x \le 1} | f(x) - P_0(x) | = \frac{1}{4}.$$

Now, we approximate $f(x)$ by $P_1(x) = a_0 + a_1 x$ such that

$$\max_{0 \le x \le 1} | f(x) - P_1(x) | = \text{minimum}.$$

Let $$\varepsilon(x) = \tfrac{1}{2}(1 - x^2) - a_0 - a_1 x.$$

Choosing the points as 0, α, 1 and using the Chebyshev equioscillation theorem, we have

$$\varepsilon(0) + \varepsilon(\alpha) = 0, \quad \text{or} \quad \left(\tfrac{1}{2} - a_0\right) + \tfrac{1}{2}(1 - \alpha^2) - a_0 - a_1\alpha = 0,$$
$$\varepsilon(\alpha) + \varepsilon(1) = 0, \quad \text{or} \quad \tfrac{1}{2}(1 - \alpha^2) - a_0 - a_1\alpha - a_0 - a_1 = 0,$$
$$\varepsilon'(\alpha) = 0, \quad \text{or} \quad \alpha + a_1 = 0.$$

Subtracting the first two equations, we get $a_1 = -1 / 2$. Third equation gives $\alpha = 1 / 2$. First equation gives $a_0 = 9 / 16$. Hence, we obtain the approximation as

$$P_1(x) = \frac{9}{16} - \frac{x}{2}.$$

We find, $\max\limits_{0 \le x \le 1} |f(x) - P_1(x)| = |\varepsilon(0)| = \left| \frac{1}{2} - a_0 \right| = \frac{1}{16}$.

Therefore, we have $\min\limits_{p} \max\limits_{0 \le x \le 1} |f(x) - P(x)| = \frac{1}{16}$.

3.61 Consider the following approximating polynomial :

Determine $\min\limits_{g} \| 1 - x - g(x) \|$ where $g(x) = ax + bx^2$ and a and b are real numbers. Determine a best approximation g if

$$\| f \|^2 = \int_0^1 f^2(x) dx$$

Is the approximation unique ? (Uppsala Univ., Sweden, BIT 10(1970), 515)

Solution

Using the given norm, we have

$$I = \int_0^1 [1 - x - ax - bx^2]^2 \, dx = \text{minimum}.$$

We obtain the normal equations as

$$\int_0^1 (1 - x - ax - bx^2) \, x \, dx = 0.$$

$$\int_0^1 (1 - x - ax - bx^2) \, x^2 \, dx = 0.$$

Integrating, we get

$$4a + 3b = 2,$$
$$15a + 12b = 5,$$

whose solution is $a = 3, b = -10/3$.

The unique least squares approximation is given by

$$g(x) = 3x - \frac{10}{3} x^2.$$

Chebyshev Polynomial Approximation (Lanczos Economization)

3.62 Suppose that we want to approximate the function $f(x) = (3 + x)^{-1}$ on the interval $-1 \le x \le 1$ with a polynomial $P(x)$ such that

$$\max_{|x| \le 1} | f(x) - P(x) | \le 0.021.$$

(*a*) Show that there does not exist a first degree polynomial satisfying this condition.

(*b*) Show that there exists a second degree polynomial satisfying this condition.

(Stockholm Univ., Sweden, BIT 14 (1974), 366)

Solution

We have, on $-1 \le x \le 1$

$$f(x) = \frac{1}{3 + x} = \frac{1}{3}\left(1 + \frac{1}{3} x\right)^{-1} = \frac{1}{3} - \frac{1}{9} x + \frac{1}{27} x^2 - \frac{1}{81} x^3 + \ldots$$

If we approximate $f(x)$ by $P_1(x) = (3 - x)/9$, then |error of approximation| is greater than $1/27 \approx 0.04$, which is more than the tolerable error.

If we approximate $f(x)$ by the second degree polynomial

$$P_2(x) = \frac{1}{3} - \frac{1}{9}x + \frac{1}{27}x^2$$

Then, | error of approximation | $\le \frac{1}{81} + \frac{1}{243} + \frac{1}{729} + \frac{1}{2187} + \ldots \approx 0.0185.$

Alternately,

$$|\text{ error of approximation }| \le \frac{1}{3^4} + \frac{1}{3^5} + \frac{1}{3^6} + \ldots = \frac{1}{3^4}\left(1 + \frac{1}{3} + \frac{1}{3^2} + \ldots\right)$$

$$= \frac{1}{81}\left[\frac{1}{1-(1/3)}\right] = \frac{1}{54} \approx 0.0185.$$

Expressing $P_2(x)$ in terms of Chebyshev polynomials, we get

$$P_2(x) = \frac{1}{3} - \frac{1}{9}x + \frac{1}{27}x^2 = \frac{1}{3}T_0 - \frac{1}{9}T_1 + \frac{1}{27}\cdot\frac{1}{2}(T_2 + T_0)$$

$$= \frac{19}{54}T_0 - \frac{1}{9}T_1 + \frac{1}{54}T_2.$$

If we truncate $P_2(x)$ at T_1, then max | error of approximation | is $1/54 = 0.0185$ and the total error becomes $0.0185 + 0.0185 = 0.0370$, which is again more than the tolerable error.

Hence, there does not exist a polynomial of first degree satisfying the given accuracy. $P_2(x)$ is the second degree polynomial satisfying the given condition.

3.63 (*a*) Approximate $f(x) = (2x - 1)^3$ by a straight line on the interval [0, 1], so that the maximum norm of the error function is minimized (use Lanczos economization).

(*b*) Show that the same line is obtained if f is approximated by the method of least squares with weight function $1/\sqrt{x(1-x)}$.

(*c*) Calculate the norm of the corresponding error functions in (*a*) and (*b*).

(Linköping Univ., Sweden, BIT 28(1988), 188)

Solution

(*a*) Substituting $x = (t + 1)/2$, we get $f(t) = t^3$ on the interval [– 1, 1].

We want to approximate $f(t) = t^3$ by a straight line on [– 1, 1]. We write

$$f(t) = t^3 = \frac{1}{4}(3T_1 + T_3)$$

where T_1, T_3 are Chebyshev polynomials.

Hence, linear approximation to $f(t)$ is $3T_1/4 = 3t/4$ or linear approximation to $f(x)$ is $3(2x - 1)/4$. The maximum absolute error is $1/4$.

(*b*) We take the approximation in the form

$$P_1(x) = a_0 + a_1(2x - 1)$$

and determine a_0 and a_1 such that

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}}\,[(2x-1)^3 - a_0 - a_1(2x-1)]^2\,dx = \text{minimum}.$$

We have the normal equations as

$$\int_0^1 \frac{[(2x-1)^3 - a_0 - a_1(2x-1)]dx}{\sqrt{x(1-x)}} = 0,$$

$$\int_0^1 \frac{[(2x-1)^3 - a_0 - a_1(2x-1)](2x-1)dx}{\sqrt{x(1-x)}} = 0,$$

which gives $$a_0 \int_0^1 \frac{dx}{\sqrt{x(1-x)}} + a_1 \int_0^1 \frac{(2x-1)dx}{\sqrt{x(1-x)}} = \int_0^1 \frac{(2x-1)^3}{\sqrt{x(1-x)}}\, dx,$$

$$a_0 \int_0^1 \frac{(2x-1)}{\sqrt{x(1-x)}}\, dx + a_1 \int_0^1 \frac{(2x-1)^2 dx}{\sqrt{x(1-x)}} = \int_0^1 \frac{(2x-1)^4}{\sqrt{x(1-x)}}\, dx.$$

We obtain

$$\int_0^1 \frac{dx}{\sqrt{x(1-x)}} = 2\int_0^1 \frac{dx}{\sqrt{(1-(2x-1)^2)}} = \int_{-\pi/2}^{\pi/2} d\theta = \pi,$$

$$\int_0^1 \frac{(2x-1)}{\sqrt{x(1-x)}}\, dx = 2\int_0^1 \frac{(2x-1)\, dx}{\sqrt{(1-(2x-1)^2)}} = \int_{-\pi/2}^{\pi/2} \sin\theta\, d\theta = 0,$$

$$\int_0^1 \frac{(2x-1)^2}{x(1-x)}\, dx = 2\int_0^1 \frac{(2x-1)^2 dx}{\sqrt{(1-(2x-1)^2)}} = \int_{-\pi/2}^{\pi/2} \sin^2\theta\, d\theta = \frac{\pi}{2},$$

$$\int_0^1 \frac{(2x-1)^3}{\sqrt{x(1-x)}}\, dx = 2\int_0^1 \frac{(2x-1)^3 dx}{\sqrt{(1-(2x-1)^2)}} = \int_{-\pi/2}^{\pi/2} \sin^3\theta\, d\theta = 0,$$

$$\int_0^1 \frac{(2x-1)^4}{\sqrt{x(1-x)}}\, dx = 2\int_0^1 \frac{(2x-1)^4 dx}{\sqrt{(1-(2x-1)^2)}} = \int_{-\pi/2}^{\pi/2} \sin^4\theta\, d\theta = \frac{3\pi}{8}.$$

Hence, we obtain $a_0 = 0$ and $a_1 = 3/4$.

The approximation is given by

$$P_1(x) = \frac{3}{4}(2x-1)$$

which is same as obtained in (a).

(c) Error norm in (a) is $1/4$.

Error norm in (b) can be obtained by evaluating E^2.

$$E^2 = \int_0^1 \frac{1}{\sqrt{x(1-x)}} \left[(2x-1)^3 - \frac{3}{4}(2x-1)\right]^2 dx$$

$$= \int_0^1 \frac{(2x-1)^6}{\sqrt{x(1-x)}\; x(1-x)} - \frac{3}{2}\int_0^1 \frac{(2x-1)^4}{\sqrt{x(1-x)}}\, dx + \frac{9}{16}\int_0^1 \frac{(2x-1)^2}{\sqrt{x(1-x)}}\, dx$$

$$= 2\int_0^{\pi/2} \sin^6\theta d\theta - 3\int_0^{\pi/2} \sin^4\theta d\theta + \frac{9}{8}\int_0^{\pi/2} \sin^2\theta d\theta$$

$$= \frac{\pi}{2}\left(\frac{15}{24} - \frac{9}{8} + \frac{9}{16}\right) = \frac{\pi}{32}.$$

Hence, $$E = \frac{\sqrt{\pi}}{4\sqrt{2}} = 0.3133.$$

3.64 The function $P_3(x) = x^3 - 9x^2 - 20x + 5$ is given. Find a second degree polynomial $P_2(x)$ such that

$$\delta = \max_{0 \le x < 4} | P_3(x) - P_2(x) |$$

becomes as small as possible. The value of δ and the values of x for which $| P_3(x) - P_2(x) | = \delta$ should also be given. (Inst. Tech., Lund, Sweden, BIT 7 (1967), 81)

Solution

Using the transformation, $x = 2(t + 1)$, we get

$$\begin{aligned} P_3(t) &= 8(t + 1)^3 - 36(t + 1)^2 - 40(t + 1) + 5 = 8t^3 - 12t^2 - 88t - 63 \\ &= - 63T_0 - 88T_1 - 6(T_2 + T_0) + 2(T_3 + 3T_1) \\ &= - 69T_0 - 82T_1 - 6T_2 + 2T_3 \end{aligned}$$

where $- 1 \le t \le 1$.

If we truncate the polynomial at T_2, we have

$$\max_{-1\le t\le 1} | P_3(t) - P_2(t) | = \max_{-1\le t\le 1} | P_3(t) - (- 69T_0 - 82T_1 - 6T_2) = \max_{-1\le t\le 1} | 2T_3 | = 2.$$

The required approximation is

$$P_2(t) = - 69T_0 - 82T_1 - 6T_2 = - 69 - 82t - 6(2t^2 - 1) = - 63 - 82t - 12t^2$$

which has the maximum absolute error $\delta = 2$.

Substituting $t = (x - 2) / 2$, we obtain

$$P_2(x) = - 3x^2 - 29x + 7.$$

We also have

$$| P_3(x) - P_2(x) | = | x^3 - 6x^2 + 9x - 2 | = 2$$

for $x = 0, 1, 3$ and 4.

3.65 Find a polynomial $P(x)$ of degree as low as possible such that

$$\max_{|x|\le 1} | e^{x^2} - P(x) | \le 0.05.$$ (Lund Univ., Sweden, BIT 15 (1975), 224)

Solution

We have $- 1 \le x \le 1$, and

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2} + \frac{x^6}{6} + \frac{x^8}{24} + \ldots \approx 1 + x^2 + \frac{x^4}{2} + \frac{x^6}{6} + \frac{x^8}{24} = P(x)$$

$$\begin{aligned} | \text{ error } | &= \left| \frac{x^{10}}{5!} + \frac{x^{12}}{6!} + \ldots \right| \le \frac{1}{5!} + \frac{1}{6!} + \ldots \\ &= e - \left(1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} \right) | = 0.00995. \end{aligned}$$

We now write

$$\begin{aligned} P(x) &= T_0 + \frac{1}{2}(T_2 + T_0) + \frac{1}{16}(T_4 + 4T_2 + 3T_0) \\ &\quad + \frac{1}{192}(T_6 + 6T_4 + 15T_2 + 10T_0) \\ &\quad + \frac{1}{3072}(T_8 + 8T_6 + 28T_4 + 56T_2 + 35T_0) \\ &= \frac{1}{3072}(T_8 + 24T_6 + 316T_4 + 2600T_2 + 5379T_0) \end{aligned}$$

Since

$$\left| \frac{1}{3072} (T_8 + 24T_6) \right| \le 0.00814$$

and the total error (0.00995 + 0.00814 = 0.01809) is less than 0.05, we get the approximation

$$\begin{aligned} e^{x^2} &\approx \frac{1}{3072} (316T_4 + 2600T_2 + 5379T_0) \\ &= \frac{1}{3072} [316(8x^4 - 8x^2 + 1) + 2600(2x^2 - 1) + 5379] \\ &= \frac{1}{3072} [2528x^4 + 2672x^2 + 3095] \\ &= 0.8229x^4 + 0.8698x^2 + 1.0075. \end{aligned}$$

3.66 The curve $y = e^{-x}$ is to be approximated by a straight line $y = b - ax$ such that $| b - ax - e^{-x} | \le 0.005$. The line should be chosen in such a way that the criterion is satisfied over as large an interval $(0, c)$ as possible (where $c > 0$). Calculate a, b and c to 3 decimal accuracy. (Inst. Tech., Lund, Sweden, BIT 5 (1965), 214)

Solution

Changing the interval $[0, c]$ to $[-1, 1]$ by the transformation $x = c(t + 1)/2$, we have the problem of approximating $exp[- c(t + 1) / 2]$ by $A + Bt$, satisfying the condition

$$\max_{-1 \le t \le 1} | A + Bt - exp[- c(t + 1) / 2] | \le 0.005.$$

We write $$f(t) = exp[- c(t + 1) / 2] \approx 1 - \frac{c(t+1)}{2} + \frac{c^2(t+1)^2}{8} = g(t)$$

with the approximate error of approximation $- c^3(t + 1)^3 / 48$ (where higher powers of c are neglected).

Writing each power of t in terms of Chebyshev polynomials, we obtain

$$\begin{aligned} g(t) &= \left(1 - \frac{c}{2} + \frac{c^2}{8}\right) + \left(\frac{c^2}{4} - \frac{c}{2}\right)t + \frac{c^2}{8} t^2 \\ &= \left(1 - \frac{c}{2} + \frac{c^2}{8}\right)T_0 + \left(\frac{c^2}{4} - \frac{c}{2}\right)T_1 + \frac{c^2}{16}(T_2 + T_0) \\ &= \left(1 - \frac{c}{2} + \frac{3c^2}{16}\right)T_0 + \left(\frac{c^2}{4} - \frac{c}{2}\right)T_1 + \frac{c^2}{16} T_2. \end{aligned}$$

If we truncate at T_1, then $g(t)$ has maximum absolute error $c^2 / 16$. We also have

$$\max_{-1 \le t \le 1} \left| \frac{-c^3}{48} (t+1)^3 \right| = \frac{c^3}{6}.$$

We choose c such that the total error

$$\frac{c^3}{6} + \frac{c^2}{16} \le 0.005.$$

We solve the equation

$$\frac{c^3}{6} + \frac{c^2}{16} = 0.005, \quad \text{or} \quad f(c) = 8c^3 + 3c^2 - 0.24 = 0$$

using the Newton-Raphson method. The smallest positive root lies in the interval (0, 0.25). Starting with $c_0 = 0.222$, we get

$$c_1 = 0.223837, \; c_2 = 0.223826.$$

Taking $c = 0.2238$, we obtain

$$f(t) = 0.8975T_0 - 0.0994T_1 = 0.8975 - 0.0994t$$

and $\qquad t = (2x - c) \,/\, c = 8.9366x - 1.$

Hence, $\qquad f(x) = 0.8975 - 0.0994(-1 + 8.9366x) = 0.9969 - 0.8883x.$

3.67 Find the lowest order polynomial which approximates the function

$$f(x) = \sum_{r=0}^{4} (-x)^r$$

in the range $0 \le x \le 1$, with an error less than 0.1.

Solution

We first change the interval [0, 1] to [– 1, 1] using the transformation $x = (1 + t) / 2$.

We have

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

$$\begin{aligned} F(t) &= 1 - \frac{1}{2}(1+t) + \frac{1}{4}(1+t)^2 - \frac{1}{8}(1+t)^3 + \frac{1}{16}(1+t)^4 \\ &= \frac{11}{16} - \frac{1}{8}t + \frac{1}{4}t^2 + \frac{1}{8}t^3 + \frac{1}{16}t^4 \\ &= \frac{11}{16}T_0 - \frac{1}{8}T_1 + \frac{1}{8}(T_2 + T_0) + \frac{1}{32}(T_3 + 3T_1) + \frac{1}{128}(T_4 + 4T_2 + 3T_0) \\ &= \frac{107}{128}T_0 - \frac{1}{32}T_1 + \frac{5}{32}T_2 + \frac{1}{32}T_3 + \frac{1}{128}T_4. \end{aligned}$$

Since $\left| \frac{1}{32}T_3 \right| + \left| \frac{1}{128}T_4 \right| \le 0.1$, we obtain the approximation

$$\begin{aligned} F(t) &= \frac{107}{128}T_0 - \frac{1}{32}T_1 + \frac{5}{32}T_2 = \frac{107}{128} - \frac{1}{32}t + \frac{5}{32}(2t^2 - 1) \\ &= \frac{5}{16}t^2 - \frac{t}{32} + \frac{87}{128}. \end{aligned}$$

Substituting $t = 2x - 1$, we obtain the polynomial approximation as

$$g(x) = \frac{1}{128}(160x^2 - 168x + 131).$$

3.68 Approximate

$$F(x) = \frac{1}{x}\int_0^x \frac{e^t - 1}{t}\, dt$$

by a third degree polynomial $P_3(x)$ so that

$$\max_{-1 \le x \le 1} \mid F(x) - P_3(x) \mid \; \le \; 3 \times 10^{-4}.$$

(Inst. Tech., Lund, Sweden, BIT 6 (1976), 270)

Solution

We have $$F(x) = \frac{1}{x}\int_0^x \left(1 + \frac{t}{2} + \frac{t^2}{6} + \frac{t^3}{24} + \frac{t^4}{120} + \frac{t^5}{720} + \frac{t^6}{5040} + \ldots\right) dt$$

$$= 1 + \frac{x}{4} + \frac{x^2}{18} + \frac{x^3}{96} + \frac{x^4}{600} + \frac{x^5}{4320} + \frac{x^6}{35280} + \ldots$$

We truncate the series at the x^5 term. Since, $-1 \le x \le 1$, the maximum absolute error is given by

$$\frac{1}{7(7!)} + \frac{1}{8(8!)} + \ldots \approx 3 \times 10^{-5}.$$

Now, $$F(x) = 1 + \frac{x}{4} + \frac{x^2}{18} + \frac{x^3}{96} + \frac{x^4}{600} + \frac{x^5}{4320}$$

$$= T_0 + \frac{1}{4}(T_1) + \frac{1}{36}(T_2 + T_0) + \frac{1}{384}(T_3 + 3T_1)$$

$$+ \frac{1}{4800}(T_4 + 4T_2 + 3T_5) + \frac{1}{69120}(T_5 + 5T_3 + 10T_1)$$

$$= \frac{14809}{14400} T_0 + \frac{1783}{6912} T_1 + \frac{103}{3600} T_2 + \frac{37}{13824} T_3 + \frac{1}{4800} T_4 + \frac{1}{69120} T_5.$$

If we truncate the right side at T_3, the neglected terms gtive the maximum error as

$$\left|\frac{T_4}{4800}\right| + \left|\frac{T_5}{69120}\right| \text{ £ } 0.00022.$$

The total error is $0.00003 + 0.00022 = 0.00025 < 3 \times 10^{-4}$. Hence, we have

$$P_3(x) = \frac{14809}{14400} + \frac{1783}{6912} x + \frac{103}{3600}(2x^2 - 1) + \frac{37}{13824}(4x^3 - 3x)$$

$$= \frac{37}{3456} x^3 + \frac{103}{1800} x^2 + \frac{3455}{13824} x + \frac{4799}{4800}.$$

3.69 The function $f(x)$ is defined by

$$f(x) = \frac{1}{x}\int_0^x \frac{1 - e^{-t^2}}{t^2} dt$$

Approximate $f(x)$ by a polynomial $P(x) = a + bx + cx^2$, such that

$$\max_{|x| \le 1} |f(x) - P(x)| \le 5 \times 10^{-3}.$$ (Lund Univ., Sweden, BIT 10 (1970), 228)

Solution

We have the given function as

$$f(x) = \frac{1}{x}\int_{0.}^x \left(1 - \frac{t^2}{2} + \frac{t^4}{6} - \frac{t^6}{24} + \frac{t^8}{120} - \frac{t^{10}}{720} + \ldots\right) dt$$

$$= 1 - \frac{x^2}{6} + \frac{x^4}{30} - \frac{x^6}{168} + \frac{x^8}{1080} - \frac{x^{10}}{7920} + \ldots$$

Truncate the series at x^8. Since $-1 \le x \le 1$, the maximum absolute error is given by

$$\frac{1}{11(6!)} + \frac{1}{13(7!)} + \ldots \approx 0.00014.$$

We get, $$P(x) = 1 - \frac{x^2}{6} + \frac{x^4}{30} - \frac{x^6}{168} + \frac{x^8}{1080}$$
$$= T_0 - \frac{1}{12}(T_2 + T_0) + \frac{1}{240}(T_4 + 4T_2 + 3T_0)$$
$$- \frac{1}{5376}(T_6 + 6T_4 + 15T_2 + 10T_0)$$
$$+ \frac{1}{138240}(T_8 + 8T_6 + 28T_4 + 56T_2 + 35T_0)$$
$$= 0.92755973T_0 - 0.06905175T_2 + 0.003253T_4 - 0.000128T_6 + 0.000007T_8.$$

Truncating the right hand side at T_2, we obtain
$$P(x) = 0.92755973T_0 - 0.06905175T_2 = 0.9966 - 0.1381x^2.$$

The maximum absolute error in the neglected terms is 0.00339.

The total error is 0.00353.

3.70 The function F is defined by
$$F(x) = \int_0^x \exp(-t^2/2)dt$$

Determine the coefficients of a fifth degree polynomial $P_5(x)$ for which
$$| F(x) - P_5(x) | \le 10^{-4} \text{ when } | x | \le 1$$
(the coefficients should be accurate to within $\pm 2 \times 10^{-5}$)

(Uppsala Univ., Sweden, BIT 5 (1965), 294)

Solution

We have the given function as
$$f(x) = \int_0^x \left(1 - \frac{t^2}{2} + \frac{t^4}{8} - \frac{t^6}{48} + \frac{t^8}{384} - \frac{t^{10}}{3840} + \ldots\right)dt$$
$$= x - \frac{x^3}{6} + \frac{x^5}{40} - \frac{x^7}{336} + \frac{x^9}{3456} - \frac{x^{11}}{42240} + \ldots$$

Truncate the series at x^9. Since $-1 \le x \le 1$, the maximum absolute error is given by
$$\frac{1}{11(2^2)(5!)} + \frac{1}{13(2^6)(6!)} + \ldots \approx 0.000025.$$

We get $$P(x) = x - \frac{x^3}{6} + \frac{x^5}{40} - \frac{x^7}{336} + \frac{x^9}{3456}$$
$$= T_1 - \frac{1}{24}(T_3 + 3T_1) + \frac{1}{640}(T_5 + 5T_3 + 10T_1)$$
$$- \frac{1}{21504}(T_7 + 7T_5 + 21T_3 + 35T_1)$$
$$+ \frac{1}{884736}(T_9 + 9T_7 + 36T_5 + 84T_3 + 105T_1)$$
$$= 0.889116T_1 - 0.034736T_3 + 0.001278T_5 - 0.000036T_7 + 0.000001T_9.$$

Neglecting T_7 and T_9 on the right hand side, we obtain
$$P(x) = 0.889116x - 0.034736(4x^3 - 3x) + 0.001278(16x^5 - 20x^3 + 5x)$$
$$= 0.0204x^5 - 0.1645x^3 + 0.9997x$$

The neglected terms have maximum absolute error 0.000037.

The total error is 0.000062.

3.71 Find the polynomial of degree 3 minimizing $\| q(x) - P(x) \|_2$ where the norm is defined by

$$(g, h) = \int_0^\infty g(x)h(x)e^{-x}\,dx \quad \text{and} \quad q(x) = x^5 - 3x^3 + x.$$

(Umea Univ., Sweden, BIT 19 (1979), 425)

Solution

The Laguerre polynomials defined by

$$L_{n+1}(x) = (-1)^{n+1} e^x \frac{d^{n+1}}{dx^{n+1}} [e^{-x} x^{n+1}]$$

are orthogonal on $[0, \infty)$ with respect to the weight function e^{-x}. We have

$$\begin{aligned} L_0(x) &= 1 \\ L_1(x) &= x - 1 \\ L_2(x) &= x^2 - 4x + 2 \\ L_3(x) &= x^3 - 9x^2 + 18x - 6 \\ L_4(x) &= x^4 - 16x^3 + 72x^2 - 96x + 24 \\ L_5(x) &= x^5 - 25x^4 + 200x^3 - 600x^2 + 600x - 120 \end{aligned}$$

and

$$\begin{aligned} 1 &= L_0(x) \\ x &= L_1(x) + L_0(x) \\ x^2 &= L_2(x) + 4L_1(x) + 2L_0(x) \\ x^3 &= L_3(x) + 9L_2(x) + 18L_1(x) + 6L_0(x) \\ x^4 &= L_4(x) + 16L_3(x) + 72L_2(x) + 96L_1(x) + 24L_0(x) \\ x^5 &= L_5(x) + 25L_4(x) + 200L_3(x) + 600L_2(x) + 600L_1(x) + 120L_0(x). \end{aligned}$$

The given polynomial can be written as

$$\begin{aligned} q(x) &= x^5 - 3x^3 + x \\ &= L_5(x) + 25L_4(x) + 197L_3(x) + 573L_2(x) + 547L_1(x) + 103L_0(x). \end{aligned}$$

Taking the approximating polynomial in the form

$$P_3(x) = a_0 L_0(x) + a_1 L_1(x) + a_2 L_2(x) + a_3 L_3(x)$$

and using the given norm, we want to determine a_0, a_1, a_2 and a_3 such that

$$\int_0^\infty [q(x) - P_3(x)]e^{-x} L_i(x)dx = 0, \; i = 0,1, 2, 3.$$

or $\displaystyle\int_0^\infty [L_5 + 25L_4 + (197 - a_3)L_3 + (573 - a_2)L_2 + (547 - a_1)L_1 + (103 - a_0)L_0]e^{-x} L_i(x)dx = 0.$

Setting $i = 0, 1, 2, 3$, and using the orthogonal properties of Laguerre polynomials

$$\int_0^\infty e^{-x} L_i(x)L_j(x)dx = 0, \; i \neq j,$$

we get $\quad a_0 = 103, a_1 = 547, a_2 = 573, a_3 = 197.$

Hence,

$$\begin{aligned} P_3(x) &= 103 + 547(x - 1) + 573(x^2 - 4x + 2) + 197(x^3 - 9x^2 + 18x - 6) \\ &= 197x^3 - 1200x^2 + 1801x - 480. \end{aligned}$$

Chapter 4

Differentiation and Integration

4.1 INTRODUCTION

Given a function $f(x)$ explicitly or defined at a set of $n + 1$ distinct tabular points, we discuss methods to obtain the approximate value of the rth order derivative $f^{(r)}(x)$, $r \geq 1$, at a tabular or a non-tabular point and to evaluate

$$\int_a^b w(x) f(x)\, dx,$$

where $w(x) > 0$ is the *weight* function and a and / or b may be finite or infinite.

4.2 NUMERICAL DIFFERENTIATION

Numerical differentiation methods can be obtained by using any one of the following three techniques :

(*i*) methods based on interpolation,

(*ii*) methods based on finite differences,

(*iii*) methods based on undetermined coefficients.

Methods Based on Interpolation

Given the value of $f(x)$ at a set of $n + 1$ distinct tabular points $x_0, x_1, \ldots, x_n$, we first write the interpolating polynomial $P_n(x)$ and then differentiate $P_n(x)$, r times, $1 \leq r \leq n$, to obtain $P_n^{(r)}(x)$. The value of $P_n^{(r)}(x)$ at the point x^*, which may be a tabular point or a non-tabular point gives the approximate value of $f^{(r)}(x)$ at the point $x = x^*$. If we use the Lagrange interpolating polynomial

$$P_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) \tag{4.1}$$

having the error term

$$\begin{aligned} E_n(x) &= f(x) - P_n(x) \\ &= \frac{(x - x_0)(x - x_1) \ldots (x - x_n)}{(n+1)!} f^{(n+1)}(\xi) \end{aligned} \tag{4.2}$$

we obtain $\qquad f^{(r)}(x^*) \approx P_n^{(r)}(x^*),\ 1 \leq r \leq n$

and $\qquad E_n^{(r)}(x^*) = f^{(r)}(x^*) - P_n^{(r)}(x^*) \qquad (4.3)$

is the error of differentiation. The error term (4.3) can be obtained by using the formula

$$\frac{1}{(n+1)!}\frac{d^j}{dx^j}[f^{(n+1)}(\xi)] = \frac{j!}{(n+j+1)!} f^{(n+j+1)}(\eta_j)$$

$$j = 1, 2, ..., r$$

where $\min(x_0, x_1, ..., x_n, x) < \eta_j < \max(x_0, x_1, ..., x_n, x)$.

When the tabular points are equispaced, we may use Newton's forward or backward difference formulas.

For $n = 1$, we obtain

$$f(x) = \frac{x - x_1}{x_0 - x_1} f_0 + \frac{x - x_0}{x_1 - x_0} f_1 \quad \text{...[4.4 } (i)]$$

and
$$f'(x_k) = \frac{f_1 - f_0}{x_1 - x_0} \quad k = 0, 1 \quad \text{...[4.4 } (ii)]$$

Differentiating the expression for the error of interpolation

$$E_1(x) = \frac{1}{2}(x - x_0)(x - x_1) f''(\xi), \quad x_0 < \xi < x_1$$

we get, at $x = x_0$ and $x = x_1$

$$E_1^{(1)}(x_0) = -E_1^{(1)}(x_1) = \frac{x_0 - x_1}{2} f''(\xi), \quad x_0 < \xi < x_1.$$

For $n = 2$, we obtain

$$f(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2 \quad \text{...[4.5 } (i)]$$

$$E_2(x) = \frac{1}{6}(x - x_0)(x - x_1)(x - x_2) f'''(\xi), \quad x_0 < \xi < x_2 \quad \text{...[4.5 } (ii)]$$

$$f'(x_0) = \frac{2x_0 - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)} f_2$$

...[4.5 (iii)]

with the error of differentiation

$$E_2^{(1)}(x_0) = \frac{1}{6}(x_0 - x_1)(x_0 - x_2) f'''(\xi), \quad x_0 < \xi < x_2.$$

Differentiating (4.5 i) and (4.5 ii) two times and setting $x = x_0$, we get

$$f''(x_0) = 2\left[\frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)}\right] \quad (4.6)$$

with the error of differentiation

$$E_2^{(2)}(x_0) = \frac{1}{3}(2x_0 - x_1 - x_2) f'''(\xi) + \frac{1}{24}(x_0 - x_1)(x_1 - x_2)[f^{iv}(\eta_1) + f^{iv}(\eta_2)]$$

where $x_0 < \xi, \eta_1, \eta_2 < x_2$.

For equispaced tabular points, the formulas [4.4 (ii)], [4.5 (iii)], and (4.6) become, respectively

$$f'(x_0) = (f_1 - f_0) / h, \quad (4.7)$$

$$f'(x_0) = (-3f_0 + 4f_1 - f_2) / (2h), \quad (4.8)$$

$$f''(x_0) = (f_0 - 2f_1 + f_2) / h^2, \quad (4.9)$$

with the respective error terms

$$E_1^{(1)}(x_0) = -\frac{h}{2} f''(\xi),\ x_0 < \xi < x_1,$$

$$E_2^{(1)}(x_0) = -\frac{h^2}{3} f'''(\xi),\ x_0 < \xi < x_2,$$

$$E_2^{(2)}(x_0) = -hf'''(\xi),\ x_0 < \xi < x_2.$$

If we write $\quad E_n^{(r)}(x_k) = |\ f^{(r)}(x_k) - P_n^{(r)}(x_k)\ |$

$$= c\, h^p + O(h^{p+1})$$

where c is a constant independent of h, then the method is said to be of *order p*. Hence, the methods (4.7) and (4.9) are of order 1, whereas the method (4.8) is of order 2.

Methods Based on Finite Differences

Consider the relation

$$Ef(x) = f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \ldots$$

$$= \left(1 + hD + \frac{h^2 D^2}{2!} + \ldots\right) f(x) = e^{hD} f(x) \tag{4.10}$$

where $D = d\ /\ dx$ is the differential operator.

Symbolically, we get from (4.10)

$$E = e^{hD}, \text{ or } hD = \ln E.$$

We have $\quad \delta = E^{1/2} - E^{-1/2} = e^{hD/2} - e^{-hD/2}$

$$= 2 \sinh (hD / 2).$$

Hence, $hD = 2 \sinh^{-1} (\delta / 2)$.

Thus, we have

$$hD = \ln E$$

$$= \left|\ \begin{aligned} &\ln (1 + \Delta) = \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \ldots \\ &-\ln (1 - \nabla) = \nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \ldots \\ &2 \sinh^{-1}\left(\frac{\delta}{2}\right) = \delta - \frac{1^2}{2^2 . 3!}\delta^3 + \ldots \end{aligned}\right. \tag{4.11}$$

Similarly, we obtain

$$h^r D^r = \left|\ \begin{aligned} &\Delta^r - \frac{1}{2} r\Delta^{r+1} + \frac{r(3r+5)}{24} \Delta^{r+2} - \ldots \\ &\nabla^r + \frac{1}{2} r\nabla^{r+1} + \frac{r(3r+5)}{24} \nabla^{r+2} + \ldots \\ &\mu\delta^r - \frac{r+3}{24}\mu\delta^{r+2} + \frac{5r^2 + 52r + 135}{5760}\mu\delta^{r+4} - \ldots,\ (r \text{ odd}) \\ &\delta^r - \frac{r}{24}\delta^{r+2} + \frac{r(5r+22)}{5760}\delta^{r+4} - \ldots,\ (r \text{ even}) \end{aligned}\right. \tag{4.12}$$

where, $\mu = \sqrt{\left(1 + \dfrac{\delta^2}{4}\right)}$ is the averaging operator and is used to avoid off-step points in the method.

Retaining various order differences in (4.12), we obtain different order methods for a given value of r. Keeping only one term in (4.12), we obtain for $r = 1$

$$f'(x_k) = \begin{vmatrix} (f_{k+1} - f_k) / h, & \text{...[4.13 } (i)] \\ (f_k - f_{k-1}) / h, & \text{...[4.13 } (ii)] \\ (f_{k+1} - f_{k-1}) / (2h), & \text{...[4.13 } (iii)] \end{vmatrix}$$

and for $r = 2$

$$f''(x_k) = \begin{vmatrix} (f_{k+2} - 2f_{k+1} + f_k) / h^2, & \text{...[4.14 } (i)] \\ (f_k - 2f_{k-1} + f_{k-2}) / h^2, & \text{...[4.14 } (ii)] \\ (f_{k+1} - 2f_k + f_{k-1}) / h^2. & \text{...[4.14 } (iii)] \end{vmatrix}$$

The methods (4.13*i*), (4.13*ii*), (4.14*i*), (4.14*ii*) are of first order, whereas the methods (4.13*iii*) and (4.14*iii*) are of second order.

Methods Based on Undetermined Coefficients

We write

$$h^r f^{(r)}(x_k) = \sum_{i=-m}^{m} a_i f(x_{k+i}) \tag{4.15}$$

for symmetric arrangement of tabular points and

$$h^r f^{(r)}(x_k) = \sum_{i=\pm m}^{n} a_i f(x_{k+i}) \tag{4.16}$$

for non symmetric arrangement of tabular points.

The error term is obtained as

$$E_r(x_k) = \frac{1}{h^r} [h^r f^{(r)}(x_k) - \Sigma a_i f(x_{k+i})]. \tag{4.17}$$

The coefficients a_i's in (4.15) or (4.16) are determined by requiring the method to be of a particular order. We expand each term in the right side of (4.15) or (4.16) in Taylor series about the point x_k and on equating the coefficients of various order derivatives on both sides, we obtain the required number of equations to determine the unknowns. The first non-zero term gives the error term.

For $m = 1$ and $r = 1$ in (4.15), we obtain

$$\begin{aligned} hf'(x_k) &= a_{-1} f(x_{k-1}) + a_0 f(x_k) + a_1 f(x_{k+1}) \\ &= (a_{-1} + a_0 + a_1) f(x_k) + (-a_{-1} + a_1) hf'(x_k) + \frac{1}{2}(a_{-1} + a_1) h^2 f''(x_k) \\ &\quad + \frac{1}{6}(-a_{-1} + a_1) h^3 f'''(x_k) + \ldots \end{aligned}$$

Comparing the coefficients of $f(x_k)$, $hf'(x_k)$ and $(h^2/2) f''(x_k)$ on both sides, we get

$$a_{-1} + a_0 + a_1 = 0, \quad -a_{-1} + a_1 = 1, \; a_{-1} + a_1 = 0$$

whose solution is $a_0 = 0$, $a_{-1} = -a_1 = -1/2$. We obtain the formula

$$h f_k' = \frac{1}{2}(f_{k+1} - f_{k-1}), \quad \text{or} \quad f_k' = \frac{1}{2h}(f_{k+1} - f_{k-1}). \tag{4.18}$$

The error term in approximating $f'(x_k)$ is given by $(-h^2/6) f'''(\xi)$, $x_{k-1} < \xi < x_{k+1}$.

For $m = 1$ and $r = 2$ in (4.15), we obtain

$$\begin{aligned} h^2 f''(x_k) &= a_{-1} f(x_{k-1}) + a_0 f(x_k) + a_1 f(x_{k+1}) \\ &= (a_{-1} + a_0 + a_1) f(x_k) + (-a_{-1} + a_1) hf'(x_k) \\ &\quad + \frac{1}{2}(a_{-1} + a_1) h^2 f''(x_k) + \frac{1}{6}(-a_{-1} + a_1) h^3 f'''(x_k) \\ &\quad + \frac{1}{24}(a_{-1} + a_1) h^4 f^{iv}(x_k) + \ldots\ldots \end{aligned}$$

Comparing the coefficients of $f(x_k)$, $hf'(x_k)$ and $h^2 f''(x_k)$ on both sides, we get

$$a_{-1} + a_0 + a_1 = 0, \quad -a_{-1} + a_1 = 0, \; a_{-1} + a_1 = 2$$

whose solution is $a_{-1} = a_1 = 1$, $a_0 = -2$. We obtain the formula

$$h^2 f_k'' = f_{k-1} - 2f_k + f_{k+1}, \quad \text{or} \quad f_k'' = \frac{1}{h^2}(f_{k-1} - 2f_k + f_{k+1}). \tag{4.19}$$

The error term in approximating $f''(x_k)$ is given by $(-h^2/12) f^{(4)}(\xi)$, $x_{k-1} < \xi < x_{k+1}$.

Formulas (4.18) and (4.19) are of second order.

Similarly, for $m = 2$ in (4.15) we obtain the fourth order methods

$$f'(x_k) = (f_{k-2} - 8f_{k-1} + 8f_{k+1} - f_{k+2}) / (12h) \tag{4.20}$$

$$f''(x_k) = (-f_{k-2} + 16f_{k-1} - 30 f_k + 16f_{k+1} - f_{k+2}) / (12h^2) \tag{4.21}$$

with the error terms $(h^4/30) f^{v}(\xi)$ and $(h^4/90) f^{vi}(\xi)$ respectively and $x_{k-2} < \xi < x_{k+2}$.

4.3 EXTRAPOLATION METHODS

To obtain accurate results, we need to use higher order methods which require a large number of function evaluations and may cause growth of roundoff errors. However, it is generally possible to obtain higher order solutions by combining the computed values obtained by using a certain lower order method with different step sizes.

If $g(x)$ denotes the quantity $f^{(r)}(x_k)$ and $g(h)$ and $g(qh)$ denote its approximate value obtained by using a certain method of order p with step sizes h and qh respectively, we have

$$g(h) = g(x) + ch^p + O(h^{p+1}), \tag{4.22}$$

$$g(qh) = g(x) + c\, q^p h^p + O(h^{p+1}). \tag{4.23}$$

Eliminating c from (4.22) and (4.23) we get

$$g(x) = \frac{q^p g(h) - g(qh)}{q^p - 1} + O(h^{p+1}) \tag{4.24}$$

which defines a method of order $p + 1$. This procedure is called *extrapolation* or *Richardson's extrapolation.*

If the error term of the method can be written as a power series in h, then by repeating the extrapolation procedure a number of times, we can obtain methods of higher orders. We often take the step sizes as h, $h/2$, $h/2^2$, ... If the error term of the method is of the form

$$E(x_k) = c_1 h + c_2 h^2 + \ldots \tag{4.25}$$

then, we have

$$g(h) = g(x) + c_1 h + c_2 h^2 + \ldots \tag{4.26}$$

Writing (4.26) for h, $h/2$, $h/2^2$, ... and eliminating c_i's from the resulting equations, we obtain the extrapolation scheme

$$g^{(p)}(h) = \frac{2^p g^{(p-1)}(h/2) - g^{(p-1)}(h)}{2^p - 1}, \quad p = 1, 2, \ldots \tag{4.27}$$

where $\qquad g^{(0)}(h) = g(h).$

The method (4.27) has order $p + 1$.

The extrapolation table is given below.

Table 4.1. Extrapolation table for (4.25).

Step \ Order	First	Second	Third	Fourth
h	$g(h)$	$g^{(1)}(h)$	$g^{(2)}(h)$	$g^{(3)}(h)$
$h/2$	$g(h/2)$	$g^{(1)}(h/2)$	$g^{(2)}(h/2)$	
$h/2^2$	$g(h/2^2)$	$g^{(1)}(h/2^2)$		
$h/2^3$	$g(h/2^3)$			

Similarly, if the error term of the method is of the form

$$E(x_k) = g(x) + c_1 h^2 + c_2 h^4 + \ldots \tag{4.28}$$

then, we have

$$g(h) = g(x) + c_1 h^2 + c_2 h^4 + \ldots \tag{4.29}$$

The extrapolation scheme is now given by

$$g^{(p)}(h) = \frac{4^p g^{(p-1)}(h/2) - g^{(p-1)}(h)}{4^p - 1}, \quad p = 1,2, \ldots \tag{4.30}$$

which is of order $2p + 2$.

The extrapolation table is given below.

Table 4.2. Extrapolation table for (4.28).

Order \ Step	Second	Fourth	Sixth	Eighth
h	$g(h)$	$g^{(1)}(h)$	$g^{(2)}(h)$	$g^{(3)}(h)$
$h/2$	$g(h/2)$	$g^{(1)}(h/2)$	$g^{(2)}(h/2)$	
$h/2^2$	$g(h/2^2)$	$g^{(1)}(h/2^2)$		
$h/2^3$	$g(h/2^3)$			

The extrapolation procedure can be stopped when

$$|\, g^{(k)}(h) - g^{(k-1)}(h/2) \,| < \varepsilon$$

where ε is the prescribed error tolerance.

4.4 PARTIAL DIFFERENTIATION

One way to obtain numerical partial differentiation methods is to consider only one variable at a time and treat the other variables as constants. We obtain

$$\left(\frac{\partial f}{\partial x}\right)_{(x_i, y_j)} = \left|\begin{array}{l} (f_{i+1,j} - f_{i,j})/h + O(h), \\ (f_{i,j} - f_{i-1,j})/h + O(h), \\ (f_{i+1,j} - f_{i-1,j})/(2h) + O(h^2), \end{array}\right. \tag{4.31}$$

$$\left(\frac{\partial f}{\partial y}\right)_{(x_i, y_j)} = \begin{cases} (f_{i,j+1} - f_{i,j})/k + O(k), \\ (f_{i,j} - f_{i,j-1})/k + O(k), \\ (f_{i,j+1} - f_{i,j-1})/(2k) + O(k^2), \end{cases} \tag{4.32}$$

where h and k are the step sizes in x and y directions respectively.

Similarly, we obtain

$$\left(\frac{\partial^2 f}{\partial x^2}\right)_{(x_i, y_j)} = (f_{i-1, j} - 2f_{i, j} + f_{i+1, j}) \,/\, h^2 + O(h^2),$$

$$\left(\frac{\partial^2 f}{\partial y^2}\right)_{(x_i, y_j)} = (f_{i, j+1} - 2f_{i, j} + f_{i, j-1}) \,/\, k^2 + O(k^2),$$

$$\left(\frac{\partial^2 f}{\partial x\, \partial y}\right)_{(x_i, y_j)} = (f_{i+1, j+1} - f_{i+1, j-1} - f_{i-1, j+1} + f_{i-1, j-1}) \,/\, (4hk) + O(h^2 + k^2).$$

(4.33)

4.5 OPTIMUM CHOICE OF STEP LENGTH

In numerical differentiation methods, error of approximation or the truncation error is of the form ch^p which tends to zero as $h \to 0$. However, the method which approximates $f^{(r)}(x)$ contains h^r in the denominator. As h is successively decreased to small values, the truncation error decreases, but the roundoff error in the method may increase as we are dividing by a smaller number. It may happen that after a certain critical value of h, the roundoff error may become more dominant than the truncation error and the numerical results obtained may start worsening as h is further reduced. When $f(x)$ is given in tabular form, these values may not themselves be exact. These values contain roundoff errors, that is $f(x_i) = f_i + \varepsilon_i$, where $f(x_i)$ is the exact value and f_i is the tabulated value. To see the effect of this roundoff error in a numerical differentiation method, we consider the method

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{h} - \frac{h}{2} f''(\xi), \quad x_0 < \xi < x_1. \tag{4.34}$$

If the roundoff errors in $f(x_0)$ and $f(x_1)$ are ε_0 and ε_1 respectively, then we have

$$f'(x_0) = \frac{f_1 - f_0}{h} + \frac{\varepsilon_1 - \varepsilon_0}{h} - \frac{h}{2} f''(\xi) \tag{4.35}$$

or

$$f'(x_0) = \frac{f_1 - f_0}{h} + \text{RE} + \text{TE} \tag{4.36}$$

where RE and TE denote the roundoff error and the truncation error respectively. If we take

$$\varepsilon = \max(\,|\,\varepsilon_1\,|, |\,\varepsilon_2\,|\,), \quad \text{and } M_2 = \max_{x_0 \le x \le x_1} |\, f''(x)\,|$$

then, we get

$$|\,\text{RE}\,| \le \frac{2\varepsilon}{h}, \quad \text{and } |\,\text{TE}\,| \le \frac{h}{2} M_2.$$

We may call that value of h as an optimal value for which one of the following criteria is satisfied :

(*i*) $|\,\text{RE}\,| = |\,\text{TE}\,|$ [4.37 (*i*)]

(*ii*) $|\,\text{RE}\,| + |\,\text{TE}\,| = \text{minimum}.$ [4.37 (*ii*)]

If we use the criterion [4.37(*i*)], then we have

$$\frac{2\varepsilon}{h} = \frac{h}{2} M_2$$

which gives $\qquad h_{\text{opt}} = 2\sqrt{\varepsilon/M_2}$, and $|\,\text{RE}\,| = |\,\text{TE}\,| = \sqrt{\varepsilon M_2}$.

If we use the criterion [4.37 (*ii*)], then we have

$$\frac{2\varepsilon}{h} + \frac{h}{2} M_2 = \text{minimum}$$

which gives $\qquad -\dfrac{2\varepsilon}{h^2} + \dfrac{1}{2} M_2 = 0, \quad \text{or} \quad h_{\text{opt}} = 2\sqrt{\varepsilon/M_2}$.

The minimum total error is $2(\varepsilon M_2)^{1/2}$.

This means that if the roundoff error is of the order 10^{-k} (say) and $M_2 \approx 0(1)$, then the accuracy given by the method may be approximately of the order $10^{-k/2}$. Since, in any numerical differentiation method, the local truncation error is always proportional to some power of h, whereas the roundoff error is inversely proportional to some power of h, the same technique can be used to determine an optimal value of h, for any numerical method which approximates $f^{(r)}(x_k)$, $r \geq 1$.

4.6 NUMERICAL INTEGRATION

We approximate the integral

$$I = \int_a^b w(x) f(x)\, dx \tag{4.38}$$

by a finite linear combination of the values of $f(x)$ in the form

$$I = \int_a^b w(x) f(x)\, dx = \sum_{k=0}^{n} \lambda_k f(x_k) \tag{4.39}$$

where x_k, $k = 0(1)n$ are called the *abscissas* or *nodes* which are distributed within the limits of integration $[a, b]$ and λ_k, $k = 0(1)n$ are called the *weights* of the integration method or the *quadrature rule* (4.39). $w(x) > 0$ is called the *weight function.* The error of integration is given by

$$R_n = \int_a^b w(x) f(x) dx - \sum_{k=0}^{n} \lambda_k f(x_k). \tag{4.40}$$

An integration method of the form (4.39) is said to be of *order p*, if it produces exact results ($R_n \equiv 0$), when $f(x)$ is a polynomial of degree $\leq p$.

Since in (4.39), we have $2n + 2$ unknowns ($n + 1$ nodes x_k's and $n + 1$ weights λ_k's), the method can be made exact for polynomials of degree $\leq 2n + 1$. Thus, the method of the form (4.39) can be of maximum order $2n + 1$. If some of the nodes are known in advance, the order will be reduced.

For a method of order m, we have

$$\int_a^b w(x) x^i dx - \sum_{k=0}^{n} \lambda_k x_k^i = 0, \; i = 0, 1, ..., m \tag{4.41}$$

which determine the weights λ_k's and the abscissas x_k's. The error of integration is obtained from

$$R_n = \frac{C}{(m+1)!} f^{(m+1)}(\xi), \quad a < \xi < b, \tag{4.42}$$

where
$$C = \int_a^b w(x) x^{m+1} \, dx - \sum_{k=0}^{n} \lambda_k x_k^{m+1}. \tag{4.43}$$

4.7 NEWTON-COTES INTEGRATION METHODS

In this case, $w(x) = 1$ and the nodes x_k's are uniformly distributed in $[a, b]$ with $x_0 = a$, $x_n = b$ and the spacing $h = (b - a) / n$. Since the nodes x_k's, $x_k = x_0 + kh$, $k = 0, 1, ..., n$, are known, we have only to determine the weights λ_k's, $k = 0, 1, ..., n$. These methods are known as *Newton-Cotes integration methods* and have the order n. When both the end points of the interval of integration are used as nodes in the methods, the methods are called *closed* type methods, otherwise, they are called *open* type methods.

Closed type methods

For $n = 1$ in (4.39), we obtain the *trapezoidal rule*

$$\int_a^b f(x)\, dx = \frac{h}{2} [f(a) + f(b)] \tag{4.44}$$

where $h = b - a$. The error term is given as

$$R_1 = -\frac{h^3}{12} f''(\xi), \quad a < \xi < b. \tag{4.45}$$

For $n = 2$ in (4.39), we obtain the *Simpson's rule*

$$\int_a^b f(x) dx = \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{4.46}$$

where $h = (b - a)2$. The error term is given by

$$R_2 = \frac{C}{3!} f'''(\xi), \quad a < \xi < b.$$

We find that in this case

$$C = \int_a^b x^3 \, dx - \frac{(b-a)}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right] = 0$$

and hence the method is exact for polynomials of degree 3 also. The error term is now given by

$$R_2 = \frac{C}{4!} f^{iv}(\xi), \quad a < \xi < b.$$

We find that
$$C = \int_a^b x^4 dx - \frac{(b-a)}{6}\left[a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right] = -\frac{(b-a)^5}{120}.$$

Hence, the error of approximation is given by

$$R_2 = -\frac{(b-a)^5}{2880} f^{iv}(\xi) = -\frac{h^5}{90} f^{iv}(\xi), \quad a < \xi < b. \tag{4.47}$$

since $h = (b - a) / 2$.

For $n = 3$ in (4.39), we obtain the *Simpson's* 3 / 8 *rule*

$$\int_a^b f(x)dx = \frac{3h}{8} [f(a) + 3f(a+h) + 3f(a+2h) + f(b)] \tag{4.48}$$

where $h = (b - a) / 3$. The error term is given by

$$R_3 = -\frac{3}{80} h^5 f^{iv}(\xi), \quad a < \xi < b, \tag{4.49}$$

and hence the method (4.49) is also a third order method.

The weights λ_k's of the Newton-Cotes rules for $n \le 5$ are given in Table 4.3. For large n, some of the weights become negative. This may cause loss of significant digits due to mutual cancellation.

Table 4.3. Weights of Newton-Cotes Integration Rule (4.39)

n	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5
1	1 / 2	1 / 2				
2	1 / 3	4 / 3	1 / 3			
3	3 / 8	9 / 8	9 / 8	3 / 8		
4	14 / 45	64 / 45	24 / 45	64 / 45	14 / 45	
5	95 / 288	375 / 288	250 / 288	250 / 288	375 / 288	95 / 288

Open type methods

We approximate the integral (4.38) as

$$I = \int_a^b f(x)dx = \sum_{k=1}^{n-1} \lambda_k f(x_k), \tag{4.50}$$

where the end points $x_0 = a$ and $x_n = b$ are excluded.

For $n = 2$, we obtain the *mid-point* rule

$$\int_a^b f(x)dx = 2h\, f(a+b) \tag{4.51}$$

where $h = (b - a) / 2$. The error term is given by

$$R_2 = \frac{h^3}{3} f''(\xi).$$

Similarly, for different values of n and $h = (b - a) / n$, we obtain

$$n = 3: \qquad I = \frac{3h}{2} [f(a+h) + f(a+2h)].$$

$$R_3 = \frac{3}{4} h^3 f''(\xi). \tag{4.52}$$

$$n = 4: \qquad I = \frac{4h}{3} [2f(a+h) - f(a+2h) + 2f(a+3h)].$$

$$R_4 = \frac{14}{45} h^5 f^{iv}(\xi). \tag{5.53}$$

$$n = 5: \qquad I = \frac{5h}{24} [11f(a+h) + f(a+2h) + f(a+3h) + 11f(a+4h)].$$

$$R_5 = \frac{95}{144} h^5 f^{iv}(\xi), \tag{4.54}$$

where $\quad a < \xi < b$.

4.8 GAUSSIAN INTEGRATION METHODS

When both the nodes and the weights in the integration method (4.39) are to be determined, then the methods are called *Gaussian integration methods.*

If the abscissas x_k's in (4.39) are selected as zeros of an orthogonal polynomial, orthogonal with respect to the weight function $w(x)$ on the interval $[a, b]$, then the method (4.39) has order $2n + 1$ and all the weights $\lambda_k > 0$.

The proof is given below.

Let $f(x)$ be a polynomial of degree less than or equal to $2n + 1$. Let $q_n(x)$ be the Lagrange interpolating polynomial of degree $\le n$, interpolating the data (x_i, f_i), $i = 0, 1, ..., n$

$$q_n(x) = \sum_{k=0}^{n} l_k(x)\, f(x_k)$$

with
$$l_k(x) = \frac{\pi(x)}{(x - x_k)\pi'(x_k)}.$$

The polynomial $[f(x) - q_n(x)]$ has zeros at $x_0, x_1, ... x_n$. Hence, it can be written as

$$f(x) - q_n(x) = p_{n+1}(x)\, r_n(x)$$

where $r_n(x)$ is a polynomial of degree atmost n and $p_{n+1}(x_i) = 0$, $i = 0, 1, 2, ... n$. Integrating this equation, we get

$$\int_a^b w(x)\ [f(x) - q_n(x)]\ dx = \int_a^b w(x)\ p_{n+1}(x)\ r_n(x)\ dx$$

or
$$\int_a^b w(x) f(x)\, dx = \int_a^b w(x)\, q_n(x)\, dx + \int_a^b w(x) p_{n+1}(x)\ r_n(x)\ dx.$$

The second integral on the right hand side is zero, if $p_{n+1}(x)$ is an orthogonal polynomial, orthogonal with respect to the weight function $w(x)$, to all polynomials of degree less than or equal to n.

We then have

$$\int_a^b w(x) f(x)\, dx = \int_a^b w(x)\, q_n(x)\, dx = \sum_{k=0}^{n} \lambda_k f(x_k)$$

where
$$\lambda_k = \int_a^b w(x)\, l_k(x)\, dx.$$

This proves that the formula (4.39) has precision $2n + 1$.

Observe that $l_j^2(x)$ is a polynomial of degree less than or equal to $2n$.

Choosing $f(x) = l_j^2(x)$, we obtain

$$\int_a^b w(x)\, l_j^2(x)\, dx = \sum_{k=0}^{n} \lambda_k\, l_j^2(x_k).$$

Since $l_j(x_k) = \delta_{jk}$, we get

$$\lambda_j = \int_a^b w(x)\, l_j^2(x)\, dx > 0.$$

Since any finite interval $[a, b]$ can be transformed to $[-1, 1]$, using the transformation

$$x = \frac{(b-a)}{2} t + \frac{(b+a)}{2}$$

we consider the integral in the form

$$\int_{-1}^{1} w(x) f(x) dx = \sum_{k=0}^{n} \lambda_k f(x_k). \tag{4.55}$$

Gauss-Legendre Integration Methods

We consider the integration rule

$$\int_{-1}^{1} f(x) dx = \sum_{k=0}^{n} \lambda_k f(x_k). \tag{4.56}$$

The nodes x_k's are the zeros of the *Legendre* polynomials

$$P_{n+1}(x) = \frac{1}{2^{n+1}(n+1)!} \frac{d^{n+1}}{dx^{n+1}} [(x^2 - 1)^{n+1}]. \tag{4.57}$$

The first few Legendre polynomials are given by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = (3x^2 - 1) / 2, P_3(x) = (5x^3 - 3x) / 2,$$
$$P_4(x) = (35x^4 - 30x^2 + 3) / 8.$$

The Legendre polynomials are orthogonal on $[-1, 1]$ with respect to the weight function $w(x) = 1$. The methods (4.56) are of order $2n + 1$ and are called *Gauss-Legendre integration methods.*

For $n = 1$, we obtain the method

$$\int_{-1}^{1} f(x) dx = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \tag{4.58}$$

with the error term $(1 / 135) f^{(4)}(\xi), -1 < \xi < 1$.

For $n = 2$, we obtain the method

$$\int_{-1}^{1} f(x) dx = \frac{1}{9} [5f(-\sqrt{3/5}) + 8f(0) + 5f(\sqrt{3/5})] \tag{4.59}$$

with the error term $(1 / 15750) f^{(6)}(\xi), -1 < \xi < 1$.

The nodes and the corresponding weights of the method (4.56) for $n \le 5$ are listed in Table 4.4.

Table 4.4. Nodes and Weights for the Gauss-Legendre Integration Methods (4.56)

n	*nodes* x_k	*weights* λ_k
1	± 0.5773502692	1.0000000000
2	0.0000000000	0.8888888889
	± 0.7745966692	0.5555555556
3	± 0.3399810436	0.6521451549
	± 0.8611363116	0.3478548451
4	0.0000000000	0.5688888889
	± 0.5384693101	0.4786286705
	± 0.9061798459	0.2369268851
5	± 0.2386191861	0.4679139346
	± 0.6612093865	0.3607615730
	± 0.9324695142	0.1713244924

Lobatto Integration Methods

In this case, $w(x) = 1$ and the two end points -1 and 1 are always taken as nodes. The remaining $n - 1$ nodes and the $n + 1$ weights are to be determined. The integration methods of the form

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(-1) + \sum_{k=1}^{n-1} \lambda_k f(x_k) + \lambda_n f(1) \tag{4.60}$$

are called the *Lobatto integration methods* and are of order $2n - 1$.

For $n = 2$, we obtain the method

$$\int_{-1}^{1} f(x)dx = \frac{1}{3}\,[f(-1) + 4f(0) + f(1)] \tag{4.61}$$

with the error term $(-1/90) f^{(4)}(\xi), -1 < \xi < 1$.

The nodes and the corresponding weights for the method (4.60) for $n \le 5$ are given in Table 4.5.

Table 4.5. Nodes and Weights for Lobatto Integration Method (4.60)

n	nodes x_k	weights λ_k
2	± 1.00000000	0.33333333
	0.00000000	1.33333333
3	± 1.00000000	0.16666667
	± 0.44721360	0.83333333
4	± 1.00000000	0.10000000
	± 0.65465367	0.54444444
	0.00000000	0.71111111
5	± 1.00000000	0.06666667
	± 0.76505532	0.37847496
	± 0.28523152	0.55485837

Radau Integration Methods

In this case, $w(x) = 1$ and the lower limit -1 is fixed as a node. The remaining n nodes and $n + 1$ weights are to be determined. The integration methods of the form

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(-1) + \sum_{k=1}^{n} \lambda_k f(x_k) \tag{4.62}$$

are called *Radau integration methods* and are of order $2n$.

For $n = 1$, we obtain the method

$$\int_{-1}^{1} f(x)dx = \frac{1}{2} f(-1) + \frac{3}{2} f\left(\frac{1}{3}\right) \tag{4.63}$$

with the error term $(2/27) f'''(\xi), -1 < \xi < 1$.

For $n = 2$, we obtain the method

$$\int_{-1}^{1} f(x)\,dx = \frac{2}{9} f(-1) + \frac{16+\sqrt{6}}{18} f\left(\frac{1-\sqrt{6}}{5}\right) + \frac{16-\sqrt{6}}{18} f\left(\frac{1+\sqrt{6}}{5}\right) \tag{4.64}$$

with the error term $(1/1125) f^{(5)}(\xi), -1 < \xi < 1$.

The nodes and the corresponding weights for the method (4.62) are given in Table 4.6.

Table 4.6. Nodes and Weights for Radau Integration Method (4.62)

n	nodes x_k	weights λ_k
1	– 1.0000000	0.5000000
	0.3333333	1.5000000
2	– 1.0000000	0.2222222
	– 0.2898979	1.0249717
	0.6898979	0.7528061
3	– 1.0000000	0.1250000
	– 0.5753189	0.6576886
	0.1810663	0.7763870
	0.8228241	0.4409244
4	– 1.0000000	0.0800000
	– 0.7204803	0.4462078
	0.1671809	0.6236530
	0.4463140	0.5627120
	0.8857916	0.2874271
5	– 1.0000000	0.0555556
	– 0.8029298	0.3196408
	– 0.3909286	0.4853872
	0.1240504	0.5209268
	0.6039732	0.4169013
	0.9203803	0.2015884

Gauss-Chebyshev Integration Methods

We consider the integral

$$\int_{-1}^{1} \frac{f(x)dx}{\sqrt{1-x^2}} = \sum_{k=0}^{n} \lambda_k f(x_k) \tag{4.65}$$

where $w(x) = 1 / \sqrt{1-x^2}$ is the weight function. The nodes x_k's are the zeros of the *Chebyshev* polynomial

$$T_{n+1}(x) = \cos((n + 1) \cos^{-1} x). \tag{4.66}$$

The first few Chebyshev polynomials are given by

$$T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = 2x^2 - 1,$$
$$T_3(x) = 4x^3 - 3x,\ T_4(x) = 8x^4 - 8x^2 + 1.$$

The Chebyshev polynomials are orthogonal on [– 1, 1] with respect to the weight function $w(x) = 1 / \sqrt{1-x^2}$. The methods of the form (4.65) are called *Gauss-Chebyshev integration methods* and are of order $2n + 1$.

We obtain from (4.66)

$$x_k = \cos\left(\frac{(2k+1)\pi}{2n+1}\right), k = 0.1, ..., n. \tag{4.67}$$

The weights λ_k's in (4.65) are equal and are given by

$$\lambda_k = \frac{\pi}{n+1}, k = 0, 1, ..., n. \tag{4.68}$$

For $n = 1$, we obtain the method

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx = \frac{\pi}{2}\left[f\left(-\frac{1}{\sqrt{2}}\right) + f\left(\frac{1}{\sqrt{2}}\right)\right] \tag{4.69}$$

with the error term $(\pi / 192) f^{(4)}(\xi), -1 < \xi < 1$.

For $n = 2$, we obtain the method

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x) dx = \frac{\pi}{3}\left[f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right)\right] \tag{4.70}$$

with the error term $(\pi / 23040) f^{(6)}(\xi), -1 < \xi < 1$.

Gauss-Laguerre Integration Methods

We consider the integral

$$\int_0^\infty e^{-x} f(x) dx = \sum_{k=0}^{n} \lambda_k f(x_k) \tag{4.71}$$

where $w(x) = e^{-x}$ is the weight function. The nodes x_k's are the zeros of the *Laguerre* polynomial

$$L_{n+1}(x) = (-1)^{n+1} e^x \frac{d^{n+1}}{dx^{n+1}} [e^{-x} x^{n+1}] \tag{4.72}$$

The first few Laguerre polynomials are given by

$$L_0(x) = 1, \quad L_1(x) = x - 1, L_2(x) = x^2 - 4x + 2,$$
$$L_3(x) = x^3 - 9x^2 + 18x - 6.$$

The Laguerre polynomials are orthogonal on $[0, \infty)$ with respect to the weight function e^{-x}. The methods of the form (4.71) are called *Gauss-Laguerre integration method* and are of order $2n + 1$.

For $n = 1$, we obtain the method

$$\int_0^\infty e^{-x} f(x) dx = \frac{2+\sqrt{2}}{4} f(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4} f(2+\sqrt{2}) \tag{4.73}$$

with the error term $(1 / 6) f^{(4)}(\xi), -1 < \xi < 1$.

The nodes and the weights of the method (4.71) for $n \le 5$ are given in Table 4.7.

Table 4.7. Nodes and Weights for Gauss-Laguerre Integration Method (4.71)

n	*nodes* x_k	*weights* λ_k
1	0.5857864376	0.8535533906
	3.4142135624	0.1464466094
2	0.4157745568	0.7110930099
	2.2942803603	0.2785177336
	6.2899450829	0.0103892565
3	0.3225476896	0.6031541043
	1.7457611012	0.3574186924
	4.5366202969	0.0388879085
	9.3950709123	0.0005392947
4	0.2635603197	0.5217556106
	1.4134030591	0.3986668111
	3.5964257710	0.0759424497
	7.0858100059	0.0036117587
	12.6408008443	0.0000233700
5	0.2228466042	0.4589646740
	1.1889321017	0.4170008308
	2.9927363261	0.1133733821
	5.7751435691	0.0103991975
	9.8374674184	0.0002610172
	15.9828739806	0.0000008955

Gauss-Hermite Integration Methods

We consider the integral

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx = \sum_{k=0}^{n} \lambda_k f(x_k) \tag{4.74}$$

where $w(x) = e^{-x^2}$ is the weight function. The nodes x_k's are the roots of the *Hermite* polynomial

$$H_{n+1}(x) = (-1)^{n+1}\, e^{-x^2} \frac{d^{n+1}}{dx^{n+1}} (e^{-x^2}). \tag{4.75}$$

The first few Hermite polynomials are given by

$$H_0(x) = 1,\ H_1(x) = 2x,\ H_2(x) = 2(2x^2 - 1),$$
$$H_3(x) = 4(2x^3 - 3x).$$

The Hermite polynomials are orthogonal on $(-\infty, \infty)$ with respect to the weight function $w(x) = e^{-x^2}$. Methods of the form (4.74) are called *Gauss-Hermite integration methods* and are of order $2n + 1$.

For $n = 1$, we obtain the method

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx = \frac{\sqrt{\pi}}{2}\left[f\left(-\frac{1}{\sqrt{2}}\right) + f\left(\frac{1}{\sqrt{2}}\right)\right] \tag{4.76}$$

with the error term $(\sqrt{\pi}/48) f^{(4)}(\xi),\ -\infty < \xi < \infty$.

For $n = 2$, we obtain the method

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx = \frac{\sqrt{\pi}}{6}\left[f\left(-\frac{\sqrt{6}}{2}\right) + 4f(0) + f\left(\frac{\sqrt{6}}{2}\right)\right] \tag{4.77}$$

with the error term $(\sqrt{\pi}/960) f^{(6)}(\xi)$, $-\infty < \xi < \infty$.

The nodes and the weights for the integration method (4.74) for $n \le 5$ are listed in Table 4.8.

Table 4.8. Nodes and Weights for Gauss-Hermite Integration Methods (4.74)

n	*nodes* x_k	*weights* λ_k
0	0.0000000000	1.7724538509
1	± 0.7071067812	0.8862269255
2	0.0000000000	1.1816359006
	± 1.2247448714	0.2954089752
3	± 0.5246476233	0.8049140900
	± 1.6506801239	0.0813128354
4	0.0000000000	0.9453087205
	± 0.9585724646	0.3936193232
	± 2.0201828705	0.0199532421
5	± 0.4360774119	0.7264295952
	± 1.3358490740	0.1570673203
	± 2.3506049737	0.0045300099

4.9 COMPOSITE INTEGRATION METHODS

To avoid the use of higher order methods and still obtain accurate results, we use the *composite integration methods*. We divide the interval $[a, b]$ or $[-1, 1]$ into a number of subintervals and evaluate the integral in each subinterval by a particular method.

Composite Trapezoidal Rule

We divide the interval $[a, b]$ into N subintervals $[x_{i-1}, x_i]$, $i = 1, 2, ..., N$, each of length $h = (b - a) / N$, $x_0 = a$, $x_N = b$ and $x_i = x_0 + ih$, $i = 1, 2, ..., N - 1$. We write

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{N-1}}^{x_N} f(x)dx . \tag{4.78}$$

Evaluating each of the integrals on the right hand side of (4.78) by the trapezoidal rule (4.44), we obtain the composite rule

$$\int_a^b f(x)dx = \frac{h}{2}\,[f_0 + 2(f_1 + f_2 + \ldots + f_{N-1}) + f_N] \tag{4.79}$$

where $f_i = f(x_i)$.

The error in the integration method (4.79) becomes

$$R_1 = -\frac{h^3}{12}\,[f''(\xi_1) + f''(\xi_2) + \ldots + f''(\xi_N)],\ x_{i-1} < \xi_i < x_i. \tag{4.80}$$

Denoting

$$f''(\eta) = \max_{a \le x \le b} |\, f''(x)\,|,\ a < \eta < b$$

we obtain from (4.80)

$$| R_1 | \le \frac{Nh^3}{12} f''(\eta) = \frac{(b-a)^3}{12N^2} f''(\eta) = \frac{(b-a)}{12} h^2 f''(\eta). \tag{4.81}$$

Composite Simpson's Rule

We divide the interval $[a, b]$ into $2N$ subintervals each of length $h = (b - a) / (2N)$. We have $2N + 1$ abscissas $x_0, x_1, ..., x_{2N}$ with $x_0 = a$, $x_{2N} = b$, $x_i = x_0 + ih$, $i = 1, 2, ..., 2N - 1$.

We write

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \ldots + \int_{x_{2N-2}}^{x_{2N}} f(x)dx. \tag{4.82}$$

Evaluating each of the integrals on the right hand side of (4.82) by the Simpson's rule (4.46), we obtain the composite rule

$$\int_a^b f(x)dx = \frac{h}{3}[f_0 + 4(f_1 + f_3 + \ldots + f_{2N-1}) + 2(f_2 + f_4 + \ldots + f_{2N-2}) + f_{2N}]. \tag{4.83}$$

The error in the integration method (4.83) becomes

$$R_2 = -\frac{h^5}{90} [f^{iv}(\xi_1) + f^{iv}(\xi_2) + \ldots + f^{iv}(\xi_N)],\ x_{2i-2} < \xi_i < x_{2i} \tag{4.84}$$

Denoting $\quad f^{iv}(\eta) = \max\limits_{a \le x \le b} | f^{iv}(x) |, a < \eta < b$

we obtain from (4.84)

$$| R_2 | \le \frac{Nh^5}{90} f^{iv}(\eta) = \frac{(b-a)^5}{2880N^4} f^{iv}(\eta) = \frac{(b-a)}{180} h^4 f^{iv}(\eta) \tag{4.85}$$

4.10 ROMBERG INTEGRATION

Extrapolation procedure of section 4.3, applied to the integration methods is called *Romberg integration*. The errors in the composite trapezoidal rule (4.79) and the composite Simpson's rule (4.83) can be obtained as

$$I = I_T + c_1h^2 + c_2h^4 + c_3h^6 + \ldots \tag{4.86}$$

$$I = \mathrm{I}_S + d_1h^4 + d_2h^6 + d_3h^8 + \ldots \tag{4.87}$$

respectively, where c_i's and d_i's are constants independent of h.

Extrapolation procedure for the trapezoidal rule becomes

$$I_T^{(m)}(h) = \frac{4^m I_T^{(m-1)}(h/2) - I_T^{(m-1)}(h)}{4^m - 1}, \ m = 1, 2, \ldots \tag{4.88}$$

where $\quad I_T^{(0)}(h) = I_T(h).$

The method (4.88) has order $2m + 2$.

Extrapolation procedure for the Simpson's rule becomes

$$I_S^{(m)}(h) = \frac{4^{m+1} I_S^{(m-1)}(h/2) - I_S^{(m-1)}(h)}{4^{m+1} - 1}, \ m = 1, 2, \ldots \tag{4.89}$$

where $\quad I_S^{(0)}(h) = \mathrm{I}_S(h).$

The method (4.89) has order $2m + 4$.

4.11 DOUBLE INTEGRATION

The problem of double integration is to evaluate the integral of the form

$$I = \int_a^b \int_c^d f(x, y) dx\, dy . \tag{4.90}$$

This integral can be evaluated numerically by two successive integrations in x any y directions respectively, taking into account one variable at a time.

Trapezoidal rule

If we evaluate the inner integral in (4.90) by the trapezoidal rule, we get

$$I_T = \frac{d-c}{2} \int_a^b [f(x, c) + f(x, d)]\, dx. \tag{4.91}$$

Using the trapezoidal rule again in (4.91) we get

$$I_T = \frac{(b-a)(d-c)}{4} [f(a, c) + f(b, c) + f(a, d) + f(b, d)]. \tag{4.92}$$

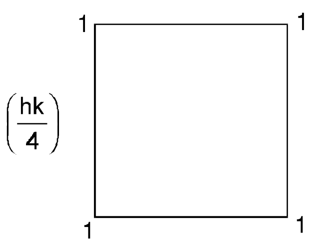
The composite trapezoidal rule for evaluating (4.90) can be written as

$$\begin{aligned} I_T = \frac{hk}{4} & [\{ f_{00} + f_{0M} + 2(f_{01} + f_{02} + \ldots + f_{0, M-1})\} \\ & + 2 \sum_{i=1}^{N-1} \{f_{i0} + f_{iM} + 2(f_{i1} + f_{i2} + \ldots + f_{i, M-1})\} \\ & + \{ f_{N0} + f_{NM} + 2(f_{N1} + f_{N2} + \ldots + f_{N, M-1})\}] \end{aligned} \tag{4.93}$$

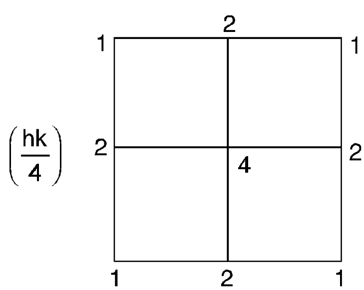
where h and k are the spacings in x and y directions respectively and

$$\begin{aligned} & h = (b-a) / N, k = (d-c) / M, \\ & x_i = x_0 + ih, i = 1, 2, \ldots, N-1, \\ & y_j = y_0 + jk, j = 1, 2, \ldots, M-1, \\ & x_0 = a, x_N = b, y_0 = c, y_M = d. \end{aligned}$$

The computational molecule of the method (4.93) for $M = N = 1$ and $M = N = 2$ can be written as



Trapezoidal rule



Composite trapezoidal rule

Simpson's rule

If we evaluate the inner integral in (4.90) by Simpson's rule then we get

$$I_S = \frac{k}{3} \int_a^b [f(x, c) + 4f(x, c+k) + f(x, d)]\, dx \tag{4.94}$$

where $k = (d - c) / 2$.

Using Simpson's rule again in (4.94), we get

$$I_S = \frac{hk}{9} [f(a, c) + f(a, d) + f(b, c) + f(b, d)$$
$$+ 4\{f(a + h, c) + f(a + h, d) + f(b, c + k)$$
$$+ f(a, c + k)\} + 16 f(a + h, c + k)] \tag{4.95}$$

where $h = (b - a) / 2$.

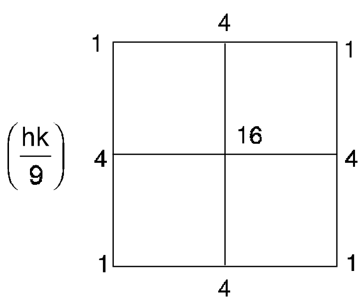
The composite Simpson's rule for evaluating (4.90) can be written as

$$I_S = \frac{hk}{9}\Bigg[\Bigg\{ f_{00} + 4 \sum_{i=1}^{N} f_{2i-1,0} + 2 \sum_{i=1}^{N-1} f_{2i,0} + f_{2N,0} \Bigg\}$$
$$+ 4 \sum_{j=1}^{M} \Bigg\{ f_{0,2j-1} + 4 \sum_{i=1}^{N} f_{2i-1,2j-1} + 2 \sum_{i=1}^{N-1} f_{2i,2j-1} + f_{2N,2j-1} \Bigg\}$$
$$+ 2 \sum_{j=1}^{M-1} \Bigg\{ f_{0,2j} + 4 \sum_{i=1}^{N} f_{2i-1,2j} + 2 \sum_{i=1}^{N-1} f_{2i,2j} + f_{2N,2j} \Bigg\}$$
$$+ \Bigg\{ f_{0,2M} + 4 \sum_{i=1}^{N} f_{2i-1,2M} + 2 \sum_{i=1}^{N-1} f_{2i,2M} + f_{2N,2M} \Bigg\} \tag{4.96}$$

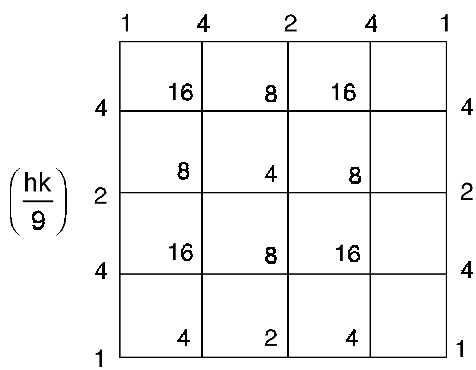
where h and k are the spacings in x and y directions respectively and

$$h = (b - a) / (2N),\ k = (d - c) / (2M),$$
$$x_i = x_0 + ih,\ i = 1, 2, ..., 2N - 1,$$
$$y_j = y_0 + jk,\ j = 1, 2, ..., 2M - 1,$$
$$x_0 = a,\ x_{2N} = b,\ y_0 = c,\ y_{2M} = d.$$

The computational module for $M = N = 1$ and $M = N = 2$ can be written as



Simpson's rule



Composite Simpson's rule

4.12 PROBLEMS AND SOLUTIONS

Numerical differentiation

4.1 A differentiation rule of the form

$$f'(x_0) = \alpha_0 f_0 + \alpha_1 f_1 + \alpha_2 f_2,$$

where $x_k = x_0 + kh$ is given. Find the values of α_0, α_1 and α_2 so that the rule is exact for $f \in P_2$. Find the error term.

Solution

The error in the differentiation rule is written as

$$\text{TE} = f'(x_0) - \alpha_0 f(x_0) - \alpha_1 f(x_1) - \alpha_2 f(x_2).$$

Expanding each term on the right side in Taylor's series about the point x_0, we obtain

$$\text{TE} = -(\alpha_0 + \alpha_1 + \alpha_2) f(x_0) + (1 - h(\alpha_1 + 2\alpha_2)) f'(x_0)$$

$$- \frac{h^2}{2}(\alpha_1 + 4\alpha_2) f''(x_0) - \frac{h^3}{6}(\alpha_1 + 8\alpha_2) f'''(x_0) - \ldots$$

We choose α_0, α_1 and α_2 such that

$$\alpha_0 + \alpha_1 + \alpha_2 = 0,$$

$$\alpha_1 + 2\alpha_2 = 1 / h,$$

$$\alpha_1 + 4\alpha_2 = 0.$$

The solution of this system is

$$\alpha_0 = -3/(2h), \ \alpha_1 = 4/(2h), \ \alpha_2 = -1/(2h).$$

Hence, we obtain the differentiation rule

$$f'(x_0) = (-3f_0 + 4f_1 - f_2)/(2h)$$

with the error term

$$\text{TE} = \frac{h^3}{6}(\alpha_1 + 8\alpha_2) f'''(\xi) = -\frac{h^2}{3} f'''(\xi), \ x_0 < \xi < x_2.$$

The error term is zero if $f(x) \in P_2$. Hence, the method is exact for all polynomials of degree ≤ 2.

4.2. Using the following data find $f'(6.0)$, error $= O(h)$, and $f''(6.3)$, error $= O(h^2)$

x	6.0	6.1	6.2	6.3	6.4
$f(x)$	0.1750	– 0.1998	– 0.2223	– 0.2422	– 0.2596

Solution

Method of $O(h)$ for $f'(x_0)$ is given by

$$f'(x_0) = \frac{1}{h}[f(x_0 + h) - f(x_0)]$$

With $x_0 = 6.0$ and $h = 0.1$, we get

$$f'(6.0) = \frac{1}{0.1}[f(6.1) - f(6.0)]$$

$$= \frac{1}{0.1}[-0.1998 - 0.1750] = -3.748.$$

Method of $O(h^2)$ for $f''(x_0)$ is given by

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)]$$

With $x_0 = 6.3$ and $h = 0.1$, we get

$$f''(6.3) = \frac{1}{(0.1)^2}[f(6.2) - 2f(6.3) + f(6.4)] = 0.25.$$

4.3 Assume that $f(x)$ has a minimum in the interval $x_{n-1} \le x \le x_{n+1}$ where $x_k = x_0 + kh$. Show that the interpolation of $f(x)$ by a polynomial of second degree yields the approximation

$$f_n - \frac{1}{8}\left(\frac{(f_{n+1} - f_{n-1})^2}{f_{n+1} - 2f_n + f_{n-1}}\right), \ f_k = f(x_k)$$

for this minimum value of $f(x)$. (Stockholm Univ., Sweden, BIT 4 (1964), 197)

Solution

The interpolation polynomial through the points (x_{n-1}, f_{n-1}), (x_n, f_n) and (x_{n+1}, f_{n+1}) is given as

$$f(x) = f(x_{n-1}) + \frac{1}{h}(x - x_{n-1})\,\Delta f_{n-1} + \frac{1}{2!\,h^2}\,(x - x_{n-1})(x - x_n)\,\Delta^2 f_{n-1}$$

Since $f(x)$ has a minimum, set $f'(x) = 0$.

Therefore $$f'(x) = \frac{1}{h}\Delta f_{n-1} + \frac{1}{2h^2}\,(2x - x_{n-1} - x_n)\,\Delta^2 f_{n-1} = 0$$

which gives $$x_{\min} = \frac{1}{2}\,(x_n + x_{n-1}) - h\,\frac{\Delta f_{n-1}}{\Delta^2 f_{n-1}}.$$

Hence, the minimum value of $f(x)$ is

$$f(x_{\min}) = f_{n-1} + \frac{1}{h}\left[\frac{1}{2}(x_n - x_{n-1}) - h\,\frac{\Delta f_{n-1}}{\Delta^2 f_{n-1}}\right]\Delta f_{n-1}$$

$$+ \frac{1}{2h^2}\left[\frac{1}{2}(x_n - x_{n-1}) - h\,\frac{\Delta f_{n-1}}{\Delta^2 f_{n-1}}\right]\left[\frac{1}{2}(x_{n-1} - x_n) - h\,\frac{\Delta f_{n-1}}{\Delta^2 f_{n-1}}\right]\Delta^2 f_{n-1}$$

Since $x_n - x_{n-1} = h$, we obtain

$$f_{\min} = f_{n-1} + \frac{1}{2}\,\Delta f_{n-1} - \frac{(\Delta f_{n-1})^2}{2\Delta^2 f_{n-1}} - \frac{1}{8}\,\Delta^2 f_{n-1}$$

$$= f_n - \Delta f_{n-1} + \frac{1}{8\Delta^2 f_{n-1}}\,[4\Delta f_{n-1}\,\Delta^2 f_{n-1} - 4(\Delta f_{n-1})^2 - (\Delta^2 f_{n-1})^2]$$

$$= f_n - \frac{1}{8\Delta^2 f_{n-1}}\,[(4\Delta f_{n-1} + \Delta^2 f_{n-1})\,\Delta^2 f_{n-1} + 4(\Delta f_{n-1})^2]$$

Using

$$\Delta f_{n-1} = f_n - f_{n-1}, \quad \Delta^2 f_{n-1} = f_{n+1} - 2f_n + f_{n-1},$$

and simplifying, we obtain

$$f_{\min} = f_n - \frac{f_{n+1}^2 - 2f_{n-1}f_{n+1} + f_{n-1}^2}{8(f_{n+1} - 2f_n + f_{n-1})} = f_n - \frac{1}{8}\left(\frac{(f_{n+1} - f_{n-1})^2}{f_{n+1} - 2f_n + f_{n-1}}\right).$$

4.4 Define

$$S(h) = \frac{-\,y(x + 2h) + 4y(x + h) - 3y(x)}{2h}$$

(*a*) Show that

$$y'(x) - S(h) = c_1 h^2 + c_2 h^3 + c_3 h^4 + \ldots$$

and state c_1.

(*b*) Calculate $y'(0.398)$ as accurately as possible using the table below and with the aid of the approximation $S(h)$. Give the error estimate (the values in the table are correctly rounded).

x	0.398	0.399	0.400	0.401	0.402
$f(x)$	0.408591	0.409671	0.410752	0.411834	0.412915

(Royal Inst. Tech. Stockholm, Sweden, BIT 19(1979), 285)

Solution

(*a*) Expanding each term in the formula

$$S(h) = \frac{1}{2h}[-y(x+2h) + 4y(x+h) - 3y(x)]$$

in Taylor series about the point x, we get

$$S(h) = y'(x) - \frac{h^2}{3}y'''(x) - \frac{h^3}{4}y^{iv}(x) - \frac{7h^4}{60}y^{v}(x) - \ldots$$
$$= y'(x) - c_1 h^2 - c_2 h^3 - c_3 h^4 - \ldots$$

Thus we obtain

$$y'(x) - S(h) = c_1 h^2 + c_2 h^3 + c_3 h^4 + \ldots$$

where $c_1 = y'''(x)/3$.

(*b*) Using the given formula with $x_0 = 0.398$ and $h = 0.001$, we obtain

$$y'(0.398) \approx \frac{1}{2(0.001)}[-y(0.400) + 4y(0.399) - 3y(0.398)]$$
$$= 1.0795.$$

The error in the approximation is given by

$$\text{Error} \approx c_1 h^2 = \frac{h^2}{3}y'''(x_0) \approx \frac{h^2}{3}\left(\frac{1}{h^3}\Delta^3 y_0\right)$$
$$= \frac{1}{3h}(y_3 - 3y_2 + 3y_1 - y_0)$$
$$= \frac{1}{3h}[y(0.401) - 3y(0.400) + 3y(0.399) - y(0.398)] = 0.$$

Hence, the error of approximation is given by the next term, which is

$$\text{Error} \approx c_2 h^3 = \frac{1}{4}h^3 y^{iv}(x_0) \approx \frac{h^3}{4}\left(\frac{1}{h^4}\Delta^4 f_0\right)$$
$$= \frac{1}{4h}(y_4 - 4y_3 + 6y_2 - 4y_1 + y_0)$$
$$= \frac{1}{4h}[y(0.402) - 4y(0.401) + 6y(0.400) - 4y(0.399) + y(0.398)]$$
$$= -0.0005.$$

4.5 Determine α, β, γ and δ such that the relation

$$y'\left(\frac{a+b}{2}\right) = \alpha\, y(a) + \beta\, y(b) + \gamma\, y''(a) + \delta\, y''(b)$$

is exact for polynomials of as high degree as possible. Give an asymptotically valid expression for the truncation error as $|\, b - a \,| \to 0$.

Solution

We write the error term in the form

$$\text{TE} = y'\left(\frac{a+b}{2}\right) - \alpha\, y(a) - \beta\, y(b) - \gamma\, y''(a) - \delta\, y''(b).$$

Letting $(a + b) / 2 = s$, $(b - a) / 2 = h / 2 = t$, in the formula, we get

$$\text{TE} = y'(s) - \alpha\, y(s - t) - \beta\, y(s + t) - \gamma\, y''(s - t) - \delta\, y''(s + t).$$

Expanding each term on the right hand side in Taylor series about s, we obtain

$$\begin{aligned}\text{TE} &= -(\alpha + \beta)y(s) + \{1 - t(\beta - \alpha)\}\, y'(s) \\ &\quad - \left\{\frac{t^2}{2}(\alpha+\beta) + \gamma + \delta\right\} y''(s) - \left\{\frac{t^3}{6}(\beta-\alpha) + t(\delta-\gamma)\right\} y'''(s) \\ &\quad - \left\{\frac{t^4}{24}(\beta+\alpha) + \frac{t^2}{2}(\delta+\gamma)\right\} y^{iv}(s) \\ &\quad - \left\{\frac{t^5}{120}(\beta-\alpha) + \frac{t^3}{6}(\delta-\gamma)\right\} y^{v}(s) - \ldots\end{aligned}$$

We choose α, β, γ and δ such that

$$\begin{aligned}\alpha + \beta &= 0, \\ -\alpha + \beta &= 1 / t = 2 / h, \\ \frac{h^2}{8}(\alpha + \beta) + \gamma + \delta &= 0, \\ \frac{h^3}{48}(-\alpha + \beta) + \frac{h}{2}(\delta - \gamma) &= 0.\end{aligned}$$

The solution of this system is

$$\alpha = -1 / h, \beta = 1 / h, \quad \gamma = h / 24 \quad \text{and} \quad \delta = -h / 24.$$

Since,

$$\frac{t^4}{24}(\beta+\alpha) + \frac{t^2}{2}(\delta+\gamma) = \left[\frac{h^4}{384}(\alpha+\beta) + \frac{h^2}{8}(\delta+\gamma)\right] = 0,$$

we obtain the error term as

$$\begin{aligned}\text{TE} &= -\left[\frac{h^5}{3840}(\beta-\alpha) + \frac{h^3}{48}(\delta-\gamma)\right] y^{v}(\xi) \\ &= -h^4\left(\frac{1}{1920} - \frac{1}{576}\right) y^{v}(\xi) = \frac{7}{5760} h^4 y^{v}(\xi),\ a < \xi < b.\end{aligned}$$

4.6 Find the coefficients a_s's in the expansion

$$D = \sum_{s=1}^{\infty} a_s\, \mu\delta^s$$

($h = 1$, D = differentiation operator, μ = mean value operator and δ = central difference operator) (Arhus Univ., Denmark, BIT 7 (1967), 81)

Solution

Since $\mu = [1 + \frac{1}{4}\delta^2]^{1/2}$, we get

$$hD = 2\sinh^{-1}\left(\frac{\delta}{2}\right) = \frac{2\mu}{\mu}\sinh^{-1}\left(\frac{\delta}{2}\right) = \frac{2\mu}{[1+(\delta^2/4)]^{1/2}}\sinh^{-1}\left(\frac{\delta}{2}\right)$$

$$= 2\mu\left[1+\frac{\delta^2}{4}\right]^{-1/2}\sinh^{-1}\left(\frac{\delta}{2}\right)$$

$$= \mu\left[1-\frac{1}{2}\frac{\delta^2}{4}+\frac{3}{8}\left(\frac{\delta^2}{4}\right)^2-\ldots\right]\left[\delta-\frac{1^2}{2^2(3!)}\delta^3+\ldots\right]$$

$$= \mu\left[\delta-\frac{1^2}{3!}\delta^3+\frac{(2!)^2}{5!}\delta^5-\ldots\right] \tag{4.97}$$

The given expression is

$$D = a_1\mu\delta + a_2\mu\delta^2 + a_3\mu\delta^3 + \ldots \tag{4.98}$$

Taking $h = 1$ and comparing the right hand sides in (4.97) and (4.98), we get

$$a_{2n} = 0,\ a_{2n+1} = \frac{(-1)^n (n!)^2}{(2n+1)!}.$$

4.7 (a) Determine the exponents k_i in the difference formula

$$f''(x_0) = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} + \sum_{i=1}^{\infty} a_i h^{k_i}$$

assuming that $f(x)$ has convergent Taylor expansion in a sufficiently large interval around x_0.

(b) Compute $f''(0.6)$ from the following table using the formula in (a) with $h = 0.4$, 0.2 and 0.1 and perform repeated Richardson extrapolation.

x	$f(x)$
0.2	1.420072
0.4	1.881243
0.5	2.128147
0.6	2.386761
0.7	2.657971
0.8	2.942897
1.0	3.559753

(Lund Univ., Sweden, BIT 13 (1973), 123)

Solution

(a) Expanding each term in Taylor series about x_0 in the given formula, we obtain

$$k_i = 2i,\ i = 1, 2, \ldots$$

(b) Using the given formula, we get

$$h = 0.4: \quad f''(0.6) = \frac{f(1.0) - 2f(0.6) + f(0.2)}{(0.4)^2} = 1.289394.$$

$$h = 0.2: \quad f''(0.6) = \frac{f(0.8) - 2f(0.6) + f(0.4)}{(0.2)^2} = 1.265450.$$

$$h = 0.1: \quad f''(0.6) = \frac{f(0.7) - 2f(0.6) + f(0.5)}{(0.1)^2} = 1.259600.$$

Applying the Richardson extrapolation

$$f''_{i,h} = \frac{4^i f''_{i-1,h} - f''_{i-1,2h}}{4^i - 1}$$

where i denotes the ith iterate, we obtain the following extrapolation table.

Extrapolation Table

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
0.4	1.289394		
		1.257469	
0.2	1.265450		1.257662
		1.257650	
0.1	1.259600		

4.8 (a) Prove that one can use repeated Richardson extrapolation for the formula

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

What are the coefficients in the extrapolation scheme ?

(b) Apply this to the table given below, and estimate the error in the computed $f''(0.3)$.

x	$f(x)$
0.1	17.60519
0.2	17.68164
0.3	17.75128
0.4	17.81342
0.5	17.86742

(Stockholm Univ., Sweden, BIT 9(1969), 400)

Solution

(a) Expanding each term in the given formula in Taylor series, we get

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = f''(x) + c_1 h^2 + c_2 h^4 + \ldots$$

If we assume that the step lengths form a geometric sequence with common ratio 1 / 2, we obtain the extrapolation scheme

$$f''_{i,h} = \frac{4^i f''_{i-1,h} - f''_{i-1,2h}}{4^i - 1}, \quad i = 1, 2, \ldots$$

where i denotes the ith iterate.

(*b*) Using the given formula, we obtain for $x = 0.3$

$$h = 0.2 : \quad f''(0.3) = \frac{f(0.5) - 2f(0.3) + f(0.1)}{(0.2)^2} = -\, 0.74875. \tag{4.99}$$

$$h = 0.1 : \quad f''(0.3) = \frac{f(0.4) - 2f(0.3) + f(0.2)}{(0.1)^2} = -\, 0.75. \tag{4.100}$$

Using extrapolation, we obtain

$$f''(0.3) = -\, 0.750417. \tag{4.101}$$

If the roundoff error in the entries in the given table is $\le 5 \times 10^{-6}$, then we have

roundoff error in (4.99) is $\le \dfrac{4 \times 5 \times 10^{-6}}{(0.2)^2} = 0.0005$,

roundoff error in (4.100) is $\le \dfrac{4 \times 5 \times 10^{-6}}{(0.1)^2} = 0.002$,

roundoff error in (4.101) is $\le \dfrac{4(0.002) + 0.0005}{3} = 0.0028$,

and the truncation error in the original formula is

$$\text{TE} \approx \frac{h^2}{12} f^{iv}(0.3) \approx \frac{1}{12h^2} \delta^4 f(0.3)$$

$$= \frac{1}{12h^2} [f(0.5) - 4f(0.4) + 6f(0.3) - 4f(0.2) + f(0.1)] = 0.000417.$$

4.9 By use of repeated Richardson extrapolation find $f'(1)$ from the following values :

x	$f(x)$
0.6	0.707178
0.8	0.859892
0.9	0.925863
1.0	0.984007
1.1	1.033743
1.2	1.074575
1.4	1.127986

Apply the approximate formula

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h}$$

with $h = 0.4, 0.2, 0.1$.

(Royal Inst. Tech., Stockhlom, Sweden, BIT 6 (1966), 270)

Solution

Applying the Richardson's extrapolation formula

$$f'_{i,\, h} = \frac{4^i f'_{i-1,\, h} - f'_{i-1,\, 2h}}{4^i - 1}, \quad i = 1, 2, \ldots$$

where i denotes the ith iterate, we obtain

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
0.4	0.526010		
		0.540274	
0.2	0.536708		0.540299
		0.540297	
0.1	0.539400		

4.10 The formula

$$D_h = (2h)^{-1} (3\,f(a) - 4\,f(a - h) + f(a - 2h))$$

is suitable to approximation of $f'(a)$ where x is the last x-value in the table.

(a) State the truncation error $D_h - f'(a)$ as a power series in h.

(b) Calculate $f'(2.0)$ as accurately as possible from the table

x	$f(x)$	x	$f(x)$
1.2	0.550630	1.7	0.699730
1.3	0.577078	1.8	0.736559
1.4	0.604826	1.9	0.776685
1.5	0.634261	1.95	0.798129
1.6	0.665766	2.0	0.820576

(Royal Inst. Tech., Stockholm, Sweden, BIT 25 (1985), 300)

Solution

(a) Expanding each term in Taylor series about a in the given formula, we obtain

$$D_h - f'(a) = -\frac{h^2}{3} f'''(a) + \frac{h^3}{4} f^{iv}(a) - \frac{7h^4}{60} f^{v}(a) + \ldots$$

Hence, the error in $D_h - f'(a)$ of the form $c_1 h^2 + c_2 h^3 + \ldots$

(b) The extrapolation scheme for the given method can be obtained as

$$f'_{i,h} = \frac{2^{i+1} f'_{i-1,h} - f'_{i-1,2h}}{2^{i+1} - 1}, \quad i = 1, 2, \ldots$$

where i denotes the ith iterate. Using the values given in the table, we obtain

$$h = 0.4: \quad f'(2.0) = \frac{1}{2(0.4)} [3f(2.0) - 4f(1.6) + f(1.2)] = 0.436618.$$

$$h = 0.2: \quad f'(2.0) = \frac{1}{2(0.2)} [3f(2.0) - 4f(1.8) + f(1.6)] = 0.453145.$$

$$h = 0.1: \quad f'(2.0) = \frac{1}{2(0.1)} [3f(2.0) - 4f(1.9) + f(1.8)] = 0.457735.$$

$$h = 0.05: \quad f'(2.0) = \frac{1}{2(0.05)} [3f(2.0) - 4f(1.95) + f(1.9)] = 0.458970.$$

Using the extrapolation scheme, we obtain the following extrapolation table.

Extrapolation Table

h	$O(h^2)$	$O(h^3)$	$O(h^4)$	$O(h^5)$
0.4	0.436618			
		0.458654		
0.2	0.453145		0.459352	
		0.459265		0.459402
0.1	0.457735		0.459399	
		0.459382		
0.05	0.458970			

Hence, $f'(2.0) = 0.4594$ with the error 2.0×10^{-6}.

4.11 For the method

$$f'(x_0) = \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} + \frac{h^2}{3} f'''(\xi), \quad x_0 < \xi < x_2$$

determine the optimal value of h, using the criteria

(*i*) | RE | = | TE |,

(*ii*) | RE | +| TE | = minimum.

Using this method and the value of h obtained from the criterion | RE | = | TE |, determine an approximate value of $f'(2.0)$ from the following tabulated values of $f(x) = \log_e x$

x	2.0	2.01	2.02	2.06	2.12
$f(x)$	0.69315	0.69813	0.70310	0.72271	0.75142

given that the maximum roundoff error in the function evaluations is 5×10^{-6}.

Solution

If ε_0, ε_1 and ε_2 are the roundoff errors in the given function evaluations f_0, f_1 and f_2 respectively, then we have

$$f'(x_0) = \frac{-3f_0 + 4f_1 - f_2}{2h} + \frac{-3\varepsilon_0 + 4\varepsilon_1 - \varepsilon_2}{2h} + \frac{h^2}{3} f'''(\xi)$$

$$= \frac{-3f_0 + 4f_1 - f_2}{2h} + \text{RE} + \text{TE}.$$

Using $\varepsilon = \max (| \varepsilon_0 |, | \varepsilon_1 |, | \varepsilon_2 |)$,

and $$M_3 = \max_{x_0 \le x \le x_2} | f'''(x) |,$$

we obtain $$| \text{RE} | \le \frac{8\varepsilon}{2h}, \quad |\text{TE}| \le \frac{h^2 M_3}{3}.$$

If we use | RE | = | TE |, we get

$$\frac{8\varepsilon}{2h} = \frac{h^2 M_3}{3}$$

which gives $$h^3 = \frac{12\varepsilon}{M_3}, \quad \text{or} \quad h_{opt} = \left(\frac{12\varepsilon}{M_3}\right)^{1/3}$$

and $$| \text{RE} | = | \text{TE} | = \frac{4\varepsilon^{2/3} M_3^{1/3}}{(12)^{1/3}}.$$

If we use | RE | + | TE | = minimum, we get

$$\frac{4\varepsilon}{h} + \frac{M_3 h^2}{3} = \text{minimum}$$

which gives $$\frac{-4\varepsilon}{h^2} + \frac{2M_3 h}{3} = 0, \quad \text{or} \quad h_{opt} = \left(\frac{6\varepsilon}{M_3}\right)^{1/3}.$$

Minimum total error $= 6^{2/3} \varepsilon^{2/3} M_3^{1/3}$.

When, $f(x) = \log_e(x)$, we have

$$M_3 = \max_{2.0 \le x \le 2.12} |f'''(x)| = \max_{2.0 \le x \le 2.12} \left| \frac{2}{x^3} \right| = \frac{1}{4}.$$

Using the criterion, $|\text{RE}| = |\text{TE}|$ and $\varepsilon = 5 \times 10^{-6}$, we get

$$h_{opt} = (4 \times 12 \times 5 \times 10^{-6})^{1/3} \simeq 0.06.$$

For $h = 0.06$, we get

$$f'(2.0) = \frac{-3(0.69315) + 4(0.72271) - 0.75142}{0.12} = 0.49975.$$

If we take $h = 0.01$, we get

$$f'(2.0) = \frac{-3(0.69315) + 4(0.69813) - 0.70310}{0.02} = 0.49850.$$

The exact value of $f'(2.0) = 0.5$.

This verifies that for $h < h_{opt}$, the results deteriorate.

Newton-Cotes Methods

4.12 (*a*) Compute by using Taylor development

$$\int_{0.1}^{0.2} \frac{x^2}{\cos x}\, dx$$

with an error $< 10^{-6}$.

(*b*) If we use the trapezoidal formula instead, which step length (of the form 10^{-k}, 2×10^{-k} or 5×10^{-k}) would be largest giving the accuracy above ? How many decimals would be required in function values ?

(Royal Inst. Tech., Stockholm, Sweden, BIT 9(1969), 174)

Solution

(*a*)

$$\int_{0.1}^{0.2} \frac{x^2}{\cos x}\, dx = \int_{0.1}^{0.2} x^2 \left(1 - \frac{x^2}{2} + \frac{x^4}{24} - \ldots\right)^{-1} dx$$

$$= \int_{0.1}^{0.2} x^2 \left(1 + \frac{x^2}{2} + \frac{5x^4}{24} + \ldots\right) dx = \left[\frac{x^3}{3} + \frac{x^5}{10} + \frac{5x^7}{168} + \ldots\right]_{0.1}^{0.2}$$

$$= 0.00233333 + 0.000031 + 0.000000378 + \ldots = 0.002365.$$

(*b*) The error term in the composite trapezoidal rule is given by

$$|\text{TE}| \le \frac{h^2}{12}(b - a) \max_{0.1 \le x \le 0.2} |f''(x)|$$

$$= \frac{h^2}{120} \max_{0.1 \le x \le 0.2} |f''(x)|.$$

We have

$$f(x) = x^2 \sec x,$$

$$f'(x) = 2x \sec x + x^2 \sec x \tan x,$$

$$f''(x) = 2 \sec x + 4x \sec x \tan x + x^2 \sec x (\tan^2 x + \sec^2 x).$$

Since $f''(x)$ is an increasing function, we get

$$\max_{0.1 \le x \le 0.2} |f''(x)| = f''(0.2) = 2.2503.$$

We choose h such that

$$\frac{h^2}{120}(2.2503) \le 10^{-6}, \text{ or } h < 0.0073.$$

Therefore, choose $h = 5 \times 10^{-3} = 0.005$.

If the maximum roundoff error in computing f_i, $i = 0, 1, ..., n$ is ε, then the roundoff error in the trapezoidal rule is bounded by

$$|\text{ RE }| \le \frac{h}{2}\left[1 + \sum_{i=1}^{n-1} 2 + 1\right]\varepsilon = nh\varepsilon = (b-a)\varepsilon = 0.1\varepsilon.$$

To meet the given error criterion, 5 decimal accuracy will be required in the function values.

4.13 Compute

$$I_p = \int_0^1 \frac{x^p dx}{x^3 + 10} \quad \text{for } p = 0, 1$$

using trapezoidal and Simpson's rules with the number of points 3, 5 and 9. Improve the results using Romberg integration.

Solution

For 3, 5 and 9 points, we have $h = 1/2$, $1/4$ and $1/8$ respectively. Using the trapezoidal and Simpson's rules and Romberg integration we get the following

$p = 0$: **Trapezoidal Rule**

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
1 / 2	0.09710999		
		0.09763534	
1 / 4	0.09750400		0.09763357
		0.09763368	
1 / 8	0.09760126		

Simpson's Rule

h	$O(h^4)$	$O(h^6)$	$O(h^8)$
1 / 2	0.09766180		
		0.09763357	
1 / 4	0.09763533		0.09763357
		0.09763357	
1 / 8	0.09763368		

$p = 1$: **Trapezoidal Rule**

h	$O(h^2)$	$O(h^4)$	$O(h^6)$
1 / 2	0.04741863		
		0.04811455	
1 / 4	0.04794057		0.04811657
		0.04811645	
1 / 8	0.04807248		

Simpson's Rule

h	$O(h^4)$	$O(h^6)$	$O(h^8)$
1 / 2	0.04807333		
		0.04811730	
1 / 4	0.04811455		0.04811656
		0.04811658	
1 / 8	0.04811645		

4.14 The arc length L of an ellipse with half axes a and b is given by the formula $L = 4aE(m)$ where $m = (a^2 - b^2) / a^2$ and

$$E(m) = \int_0^{\pi/2} (1 - m \sin^2 \phi)^{1/2} \, d\phi .$$

The function $E(m)$ is an *elliptic integral,* some values of which are displayed in the table :

m	0	0.1	0.2	0.3	0.4	0.5
$E(m)$	1.57080	1.53076	1.48904	1.44536	1.39939	1.35064

We want to calculate L when $a = 5$ and $b = 4$.

(*a*) Calculate L using quadratic interpolation in the table.

(*b*) Calculate L applying Romberg's method to $E(m)$, so that a Romberg value is got with an error less than 5×10^{-5}. (Trondheim Univ., Sweden, BIT 24(1984), 258)

Solution

(*a*) For $a = 5$ and $b = 4$, we have $m = 9 / 25 = 0.36$.

Taking the points as $x_0 = 0.3$, $x_1 = 0.4$, $x_2 = 0.5$ we have the following difference table.

x	$f(x)$	Δf	$\Delta^2 f$
0.3	1.44536		
		– 0.04597	
0.4	1.39939		– 0.00278
		– 0.04875	
0.5	1.35064		

The Newton forward difference interpolation gives

$$P_2(x) = 1.44536 + (x - 0.3) \left(\frac{-0.04597}{0.1} \right) + (x - 0.3)(x - 0.4) \left(- \frac{0.00278}{2(0.01)} \right).$$

We obtain $\quad E(0.36) \approx P_2 (0.36) = 1.418112.$

Hence, $\quad L = 4aE(m) = 20E(0.36) = 28.36224.$

(*b*) Using the trapezoidal rule to evaluate

$$E(m) = \int_0^{\pi/2} (1 - m \sin^2 \phi)^{1/2} \, d\phi , \; m = 0.36$$

and applying Romberg integration, we get

h	$O(h^2)$ method	$O(h^4)$ method
$\pi / 4$	1.418067	
		1.418088
$\pi / 8$	1.418083	

Hence, using the trapezoidal rule with $h = \pi / 4$, $h = \pi / 8$ and with one extrapolation, we obtain $E(m)$ correct to four decimal places as

$$E(m) = 1.4181, \; m = 0.36.$$

Hence, $\quad L = 28.362.$

4.15 Calculate $\displaystyle\int_0^{1/2} \frac{x}{\sin x}\, dx$.

(a) Use Romberg integration with step size $h = 1 / 16$.

(b) Use 4 terms of the Taylor expansion of the integrand.

(Uppsala Univ., Sweden, BIT 26(1986), 135)

Solution

(a) Using trapezoidal rule we have with

$$h = \frac{1}{2}: \qquad I = \frac{h}{2}[f(a) + f(b)] = \frac{1}{4}\left[1 + \frac{1/2}{\sin 1/2}\right] = 0.510729$$

where we have used the fact that $\lim\limits_{x \to 0} (x / \sin x) = 1$.

$$h = \frac{1}{4}: \qquad I = \frac{1}{8}\left[1 + 2\left(\frac{1/4}{\sin 1/4}\right) + \left(\frac{1/2}{\sin 1/2}\right)\right] = 0.507988.$$

$$h = \frac{1}{8}: \qquad I = \frac{1}{16}\left[1 + 2\left\{\frac{1/8}{\sin 1/8} + \frac{2/8}{\sin 2/8} + \frac{3/8}{\sin 3/8}\right\} + \left(\frac{1/2}{\sin 1/2}\right)\right] = 0.507298.$$

$$h = \frac{1}{16}: \qquad I = \frac{1}{32}\left[1 + 2\left\{\frac{1/16}{\sin 1/16} + \frac{2/16}{\sin 2/16} + \frac{3/16}{\sin 3/16} + \frac{4/16}{\sin 4/16}\right.\right.$$
$$\left.\left. + \frac{5/16}{\sin 5/16} + \frac{6/16}{\sin 6/16} + \frac{7/16}{\sin 7/16}\right\} + \left(\frac{1/2}{\sin 1/2}\right)\right]$$
$$= 0.507126.$$

Using extrapolation, we obtain the following Romberg table :

Romberg Table

h	$O(h^2)$	$O(h^4)$	$O(h^6)$	$O(h^8)$
1 / 2	0.510729			
		0.507074		
1 / 4	0.507988		0.507068	
		0.507068		0.507069
1 / 8	0.507298		0.507069	
		0.507069		
1 / 16	0.507126			

(*b*) We write

$$I = \int_0^{1/2} \frac{x}{x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \ldots}\, dx$$

$$= \int_0^{1/2} \left[1 - \left(\frac{x^2}{6} - \frac{x^4}{120} + \frac{x^6}{5040} - \ldots\right)\right]^{-1} dx$$

$$= \int_0^{1/2} \left[1 + \frac{x^2}{6} + \frac{7}{360}x^4 + \frac{31}{15120}x^6 + \ldots\right] dx$$

$$= \left[x + \frac{x^3}{18} + \frac{7x^5}{1800} + \frac{31x^7}{105840} + \ldots\right]_0^{1/2} = 0.507068.$$

4.16 Compute the integral $\int_0^1 y\, dx$ where y is defined through $x = y\, e^y$, with an error $< 10^{-4}$.

(Uppsala Univ., Sweden, BIT 7(1967), 170)

Solution

We shall use the trapezoidal rule with Romberg integration to evaluate the integral. The solution of $y\, e^y - x = 0$ for various values of x, using Newton-Raphson method is given in the following table.

x	y	x	y
0	0	0.500	0.351734
0.125	0.111780	0.625	0.413381
0.250	0.203888	0.750	0.469150
0.375	0.282665	0.875	0.520135
		1.000	0.567143

Romberg integration gives

h	$O(h^2)$ method	$O(h^4)$ method	$O(h^6)$ method
0.5	0.317653		
		0.330230	
0.25	0.327086		0.330363
		0.330355	
0.125	0.329538		

The error of integration is 0.330363 – 0.330355 = 0.000008.

The result correct to five decimals is 0.33036.

4.17 The area A inside the closed curve $y^2 + x^2 = \cos x$ is given by

$$A = 4 \int_0^{\alpha} (\cos x - x^2)^{1/2}\, dx$$

where α is the positive root of the equation $\cos x = x^2$.

(*a*) Compute α to three correct decimals.

(*b*) Use Romberg's method to compute the area A with an absolute error less than 0.05.

(Linköping Univ., Sweden, BIT 28(1988), 904)

Solution

(*a*) Using Newton-Raphson method to find the root of equation

$$f(x) = \cos x - x^2 = 0$$

we obtain the iteration scheme

$$x_{k+1} = x_k + \frac{\cos x_k - x_k^2}{\sin x_k + 2x_k}, \quad k = 0, 1, \ldots$$

Starting with $x_0 = 0.5$, we get

$$x_1 = 0.5 + \frac{0.627583}{1.479426} = 0.924207.$$

$$x_2 = 0.924207 + \frac{-0.251691}{2.646557} = 0.829106.$$

$$x_3 = 0.829106 + \frac{-0.011882}{2.395540} = 0.824146.$$

$$x_4 = 0.824146 + \frac{-0.000033}{2.382260} = 0.824132.$$

Hence, the value of α correct to three decimals is 0.824.

The given integral becomes

$$A = 4\int_0^{0.824} (\cos x - x^2)^{1/2}\, dx.$$

(*b*) Using the trapezoidal rule with h = 0.824, 0.412 and 0.206 respectively, we obtain the approximation

$$A \approx \frac{4(0.824)}{2} [1 + 0.017753] = 1.677257$$

$$A \approx \frac{4(0.412)}{2} [1 + 2(0.864047) + 0.017753] = 2.262578.$$

$$A \approx \frac{4(0.206)}{2} [1 + 2(0.967688 + 0.864047 + 0.658115) + 0.017753]$$
$$= 2.470951.$$

Using Romberg integration, we obtain

h	$O(h^2)$ method	$O(h^4)$ method	$O(h^6)$ method
0.824	1.677257		
		2.457685	
0.412	2.262578		2.545924
		2.540409	
0.206	2.470951		

Hence, the area with an error less than 0.05 is 2.55.

4.18 (*a*) The natural logarithm function of a positive x is defined by

$$\ln x = -\int_x^1 \frac{dt}{t}.$$

We want to calculate ln (0.75) by estimating the integral by the trapezoidal rule $T(h)$. Give the maximal step size h to get the truncation error bound $0.5(10^{-3})$. Calculate $T(h)$ with $h = 0.125$ and $h = 0.0625$. Extrapolate to get a better value.

(*b*) Let $f_n(x)$ be the Taylor series of $\ln x$ at $x = 3/4$, truncated to $n + 1$ terms. Which is the smallest n satisfying

$$| f_n(x) - \ln x | \le 0.5(10^{-3}) \text{ for all } x \in [0.5, 1].$$

(Trondheim Univ., Sweden, BIT 24(1984), 130)

Solution

(*a*) The error in the composite trapezoidal rule is given as

$$| R | \le \frac{(b-a)h^2}{12} f''(\xi) = \frac{h^2}{48} f''(\xi),$$

where $f''(\xi) = \max_{0.75 \le x \le 1} | f''(x) |$.

Since $f(t) = -1/t$, we have $f'(t) = 1/t^2$, $f''(t) = -2/t^3$

and therefore $\max_{0.75 \le t \le 1} | f''(t) | = \max_{0.75 \le t \le 1} \left| \frac{2}{t^3} \right| = 4.740741.$

Hence, we find h such that

$$\frac{h^2}{48}(4.740741) < 0.0005$$

which gives $h < 0.0712$. Using the trapezoidal rule, we obtain

$h = 0.125 : t_0 = 0.75,\ t_1 = 0.875,\ t_2 = 1.0,$

$$I = -\frac{0.125}{2}\left[\frac{1}{t_0} + \frac{2}{t_1} + \frac{1}{t_2}\right] = -0.288690.$$

$h = 0.0625 : t_0 = 0.75,\ t_1 = 0.8125,\ t_2 = 0.875,\ t_3 = 0.9375,\ t_4 = 1.0,$

$$I = -\frac{0.0625}{2}\left[\frac{1}{t_0} + 2\left(\frac{1}{t_1} + \frac{1}{t_2} + \frac{1}{t_3}\right) + \frac{1}{t_4}\right] = -0.287935.$$

Using extrapolation, we obtain the extrapolated value as

$I = -0.287683.$

(*b*) Expanding $\ln x$ in Taylor series about the point $x = 3/4$, we get

$$\ln x = \ln(3/4) + \left(x - \frac{3}{4}\right)\left(\frac{4}{3}\right) - \left(x - \frac{3}{4}\right)^2 \cdot \frac{1}{2} \cdot \left(\frac{4}{3}\right)^2 + \ldots + \left(x - \frac{3}{4}\right)^n \frac{(n-1)!\,(-1)^{n-1}}{n!}\left(\frac{4}{3}\right)^n + R_n$$

with the error term

$$R_n = \frac{(x - 3/4)^{n+1}}{(n+1)!} \frac{n!\,(-1)^n}{\xi^{n+1}},\ 0.5 < \xi < 1.$$

We have

$$| R_n | \le \frac{1}{(n+1)} \max_{0.5 \le x \le 1} \left| \left(x - \frac{3}{4}\right)^{n+1} \right| \max_{0.5 \le x \le 1} \left| \frac{1}{x^{n+1}} \right| = \frac{1}{(n+1)\, 2^{n+1}}.$$

We find the smallest n such that

$$\frac{1}{(n+1)\,2^{n+1}} \le 0.0005$$

which gives $n = 7$.

4.19 Determine the coefficients a, b and c in the quadrature formula

$$\int_{x_0}^{x_1} y(x)dx = h(ay_0 + by_1 + cy_2) + R$$

where $x_i = x_0 + ih$, $y(x_i) = y_i$. Prove that the error term R has the form

$$R = ky^{(n)}(\xi), \quad x_0 \le \xi \le x_2$$

and determine k and n. (Bergen Univ., Sweden, BIT 4(1964), 261)

Solution

Making the method exact for $y(x) = 1$, x and x^2 we obtain the equations

$$x_1 - x_0 = h(a + b + c),$$
$$\frac{1}{2}(x_1^2 - x_0^2) = h(a\,x_0 + b\,x_1 + c\,x_2),$$
$$\frac{1}{3}(x_1^3 - x_0^3) = h(a\,x_0^2 + b\,x_1^2 + c\,x_2^2).$$

Simplifying the above equations, we get

$$a + b + c = 1,$$
$$b + 2c = 1/2,$$
$$b + 4c = 1/3.$$

which give $a = 5/12$, $b = 2/3$ and $c = -1/12$.

The error term R is given by

$$R = \frac{C}{3!}\,y'''(\xi),\ x_0 < \xi < x_2$$

where
$$C = \int_{x_0}^{x_1} x^3 dx - h\,[a\,x_0^3 + b\,x_1^3 + c\,x_2^3] = \frac{h^4}{4}.$$

Hence, we have the remainder as

$$R = \frac{h^4}{24}\,y'''(\xi).$$

Therefore, $\quad k = h^4/24 \quad$ and $\quad n = 3.$

4.20 Obtain a generalized trapezoidal rule of the form

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f_0 + f_1) + ph^2(f_0' - f_1').$$

Find the constant p and the error term. Deduce the composite rule for integrating

$$\int_a^b f(x)dx\,,\ a = x_0 < x_1 < x_2 \ldots < x_N = b.$$

Solution

The method is exact for $f(x) = 1$ and x. Making the method exact for $f(x) = x^2$, we get

$$\frac{1}{3}(x_1^3 - x_0^3) = \frac{h}{2}(x_0^2 + x_1^2) + 2ph^2(x_0 - x_1).$$

Since, $x_1 = x_0 + h$, we obtain on simplification $p = 1 / 12$.

The error term is given by

$$\text{Error} = \frac{C}{3!} f'''(\xi), \quad x_0 < \xi < x_1$$

where $$C = \int_{x_0}^{x_1} x^3 dx - \left[\frac{h}{2} (x_0^3 + x_1^3) + 3ph^2 (x_0^2 - x_1^2) \right] = 0.$$

Therefore, the error term becomes

$$\text{Error} = \frac{C}{4!} f^{iv}(\xi), \quad x_0 < \xi < x_1$$

where $$C = \int_{x_0}^{x_1} x^4 dx - \left[\frac{h}{2} (x_0^4 + x_1^4) + 4ph^2 (x_0^3 - x_1^3) \right] = \frac{h^5}{30}.$$

Hence, we have the remainder as

$$\text{Error} = \frac{h^5}{720} f^{iv}(\xi).$$

Writing the given integral as

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{N-1}}^{x_N} f(x)dx$$

where $x_0 = a$, $x_N = b$, $h = (b - a) / N$, and replacing each integral on the right side by the given formula, we obtain the composite rule

$$\int_a^b f(x)dx = \frac{h}{2} [f_0 + 2(f_1 + f_2 + \ldots + f_{N-1}) + f_N] + \frac{h^2}{12} (f_0' - f_N').$$

4.21 Determine α and β in the formula

$$\int_a^b f(x)\, dx = h \sum_{i=0}^{n-1} [f(x_i) + \alpha h f'(x_i) + \beta h^2 f''(x_i)] + O(h^p)$$

with the integer p as large as possible. (Uppsala Univ., Sweden, BIT 11(1971), 225)

Solution

First we determine the formula

$$\int_{x_0}^{x_1} f(x)dx = h[a\, f_0 + b\, f_0' + c\, f_0''].$$

Making the method exact for $f(x) = 1$, x and x^2, we get $a = 1$, $b = h / 2$ and $c = h^2 / 6$.

Hence, we have the formula

$$\int_{x_0}^{x_1} f(x)dx = h \left[f_0 + \frac{h}{2} f_0' + \frac{h^2}{6} f_0'' \right]$$

which has the error term

$$\text{TE} = \frac{C}{3!} f'''(\xi)$$

where $$C = \int_{x_0}^{x_1} x^3 dx - h \left[x_0^3 + \frac{3h}{2} x_0^2 + h^2 x_0 \right] = \frac{h^4}{4}$$

Using this formula, we obtain the composite rule as

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)\,dx$$

$$= h\sum_{i=0}^{n-1}\left(f_i + \frac{h}{2}f_i' + \frac{h^2}{6}f_i''\right)$$

The error term of the composite rule is obtained as

$$|\,TE\,| = \frac{h^4}{24}\,|\,f'''(\xi_1) + f'''(\xi_2) + \ldots + f'''(\xi_n)\,|$$

$$\le \frac{nh^4}{24}\,f'''(\xi) = \frac{(b-a)h^3}{24}\,f'''(\xi),$$

where $a < \xi < b$ and $f'''(\xi) = \max\,|\,f'''(x)\,|$, $a < x < b$.

4.22 Determine a, b and c such that the formula

$$\int_0^h f(x)dx = h\left\{af(0) + bf\left(\frac{h}{3}\right) + cf(h)\right\}$$

is exact for polynomials of as high degree as possible, and determine the order of the truncation error. (Uppsala Univ. Sweden, BIT 13(1973), 123)

Solution

Making the method exact for polynomials of degree upto 2, we obtain

$f(x) = 1$: $\quad h = h(a + b + c)$, or $\quad a + b + c = 1$.

$f(x) = x$: $\quad \dfrac{h^2}{2} = h\left(\dfrac{bh}{3} + ch\right)$, or $\quad \tfrac{1}{3}b + c = \tfrac{1}{2}$.

$f(x) = x^2$: $\quad \dfrac{h^3}{3} = h\left(\dfrac{bh^2}{9} + ch^2\right)$, or $\quad \tfrac{1}{9}b + c = \tfrac{1}{3}$.

Solving the above equations, we get $a = 0$, $b = 3/4$ and $c = 1/4$.

Hence, the required formula is

$$\int_0^h f(x)dx = \frac{h}{4}\,[3f(h/3) + f(h)].$$

The truncation error of the formula is given by

$$TE = \frac{C}{3!}\,f'''(\xi),\ 0 < \xi < h$$

where $$C = \int_0^h x^3 dx - h\left[\frac{bh^3}{27} + ch^3\right] = -\frac{h^4}{36}.$$

Hence, we have

$$TE = -\frac{h^4}{216}\,f'''(\xi) = O(h^4).$$

4.23 Find the values of a, b and c such that the truncation error in the formula

$$\int_{-h}^{h} f(x)dx = h[af(-h) + bf(0) + af(h)] + h^2 c\,[f'(-h) - f'(h)]$$

is minimized.

Suppose that the composite formula has been used with the step length h and $h / 2$, giving $I(h)$ and $I(h / 2)$. State the result of using Richardson extrapolation on these values.

(Lund Univ., Sweden, BIT 27(1987), 286)

Solution

Note that the abscissas are symmetrically placed. Making the method exact for $f(x) = 1$, x^2 and x^4, we obtain the system of equations

$$f(x) = 1 \;:\; 2a + b = 2,$$
$$f(x) = x^2 : 2a - 4c = 2 / 3,$$
$$f(x) = x^4 : 2a - 8c = 2 / 5,$$

which gives $a = 7 / 15$, $b = 16 / 15$, $c = 1 / 15$.

The required formula is

$$\int_{-h}^{h} f(x)dx = \frac{h}{15}\,[7f(-h) + 16f(0) + 7f(h)] + \frac{h^2}{15}\,[f'(-h) - f'(h)].$$

The error term is obtained as

$$R = \frac{C}{6!} f^{vi}(\xi), \quad -h < \xi < h$$

where $$C = \int_{-h}^{h} x^6 dx - \left[\frac{h}{15}(14h^6) - \frac{12}{15}h^7\right] = \frac{16}{105}\,h^7.$$

Hence, we get the error term as

$$R = \frac{h^7}{4725}\, f^{vi}(\xi), \; -h < \xi < h.$$

The composite integrating rule can be written as

$$\int_a^b f(x)dx = \frac{h}{15}\,[7(f_0 + f_{2n}) + 16(f_1 + f_3 + \ldots + f_{2n-1}) + 14(f_2 + f_4 + \ldots + f_{2n-2})]$$
$$+ \frac{h^2}{15}\,(f_0' - f_{2n}') + O(h^6).$$

The truncation error in the composite integration rule is obtained as

$$R = c_1 h^6 + c_2 h^8 + \ldots$$

If $I(h)$ and $I(h / 2)$ are the values obtained by using step sizes h and $h / 2$ respectively, then the extrapolated value is given

$$I = [64\, I(h / 2) - I(h)] / 63.$$

4.24 Consider the quadrature rule

$$\int_a^b f(x)dx = \sum_{i=0}^{n} w_i f(x_i)$$

where $w_i > 0$ and the rule is exact for $f(x) = 1$. If $f(x_i)$ are in error atmost by $(0.5)10^{-k}$, show that the error in the quadrature rule is not greater than $10^{-k}(b - a) / 2$.

Solution

We have $w_i > 0$. Since the quadrature rule is exact for $f(x) = 1$, we have

$$\sum_{i=0}^{n} w_i = b - a .$$

We also have

$$|\text{ Error }| = \left| \sum_{i=0}^{n} w_i \, [f(x_i) - f^*(x_i)] \right| \le \sum_{i=0}^{n} w_i \, |f(x_i) - f^*(x_i)|$$

$$\le (0.5)10^{-k} \sum_{i=0}^{n} w_i = \frac{1}{2} (b - a)10^{-k}.$$

Gaussian Integration Methods

4.25 Determine the weights and abscissas in the quadrature formula

$$\int_{-1}^{1} f(x) dx = \sum_{k=1}^{4} A_k f(x_k)$$

with $x_1 = -1$ and $x_4 = 1$ so that the formula becomes exact for polynomials of highest possible degree. (Gothenburg Univ., Sweden, BIT 7(1967), 338)

Solution

Making the method

$$\int_{-1}^{1} f(x) dx = A_1 f(-1) + A_2 f(x_2) + A_3 f(x_3) + A_4 f(1)$$

exact for $f(x) = x^i$, $i = 0, 1, ..., 5$, we obtain the equations

$$A_1 + A_2 + A_3 + A_4 = 2, \tag{4.102}$$

$$-A_1 + A_2 x_2 + A_3 x_3 + A_4 = 0, \tag{4.103}$$

$$A_1 + A_2 x_2^2 + A_3 x_3^2 + A_4 = \frac{2}{3}, \tag{4.104}$$

$$-A_1 + A_2 x_2^3 + A_3 x_3^3 + A_4 = 0, \tag{4.105}$$

$$A_1 + A_2 x_2^4 + A_3 x_3^4 + A_4 = \frac{2}{5}, \tag{4.106)]}$$

$$-A_1 + A_2 x_2^5 + A_3 x_3^5 + A_4 = 0. \tag{4.107}$$

Subtracting (4.104) from (4.102), (4.105) from (4.103), (4, 106) from (4.104) and (4.107) from (4.105), we get

$$\frac{4}{3} = A_2(1 - x_2^2) + A_3(1 - x_3^2),$$

$$0 = A_2 x_2 (1 - x_2^2) + A_3 x_3 (1 - x_3^2),$$

$$\frac{4}{15} = A_2 x_2^2 (1 - x_2^2) + A_3 x_3^2 (1 - x_3^2),$$

$$0 = A_2 x_2^3 (1 - x_2^2) + A_3 x_3^3 (1 - x_3^2).$$

Eliminating A_3 from the above equations, we get

$$\frac{4}{3}x_3 = A_2(1 - x_2^2)(x_3 - x_2),$$

$$-\frac{4}{15} = A_2 x_2 (1 - x_2^2)(x_3 - x_2),$$

$$\frac{4}{15}x_3 = A_2 x_2^2 (1 - x_2^2)(x_3 - x_2),$$

which give $x_2 x_3 = -1/5$, $x_2 = -x_3 = 1/\sqrt{5}$ and $A_1 = A_4 = 1/6$, $A_2 = A_3 = 5/6$.

The error term of the method is given by

$$\text{TE} = \frac{C}{6!} f^{vi}(\xi), \quad -1 < \xi < 1$$

where $$C = \int_{-1}^{1} x^6 - [A_1 + A_2 x_2^6 + A_3 x_3^6 + A_4] = \frac{2}{7} - \frac{26}{75} = -\frac{32}{525}.$$

Hence, we have $$\text{TE} = -\frac{2}{23625} f^{vi}(\xi).$$

4.26 Find the value of the integral

$$I = \int_2^3 \frac{\cos 2x}{1 + \sin x} dx$$

using Gauss-Legendre two and three point integration rules.

Solution

Substituting $x = (t + 5)/2$ in I, we get

$$I = \int_2^3 \frac{\cos 2x}{1 + \sin x} dx = \frac{1}{2}\int_{-1}^{1} \frac{\cos(t+5)}{1 + \sin((t+5)/2)} dt.$$

Using the Gauss-Legendre two-point formula

$$\int_{-1}^{1} f(x)dx = f\left(\frac{1}{\sqrt{3}}\right) + f\left(-\frac{1}{\sqrt{3}}\right)$$

we obtain $$I = \frac{1}{2}[0.56558356 - 0.15856672] = 0.20350842.$$

Using the Gauss-Legendre three-point formula

$$\int_{-1}^{1} f(x)dx = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

we obtain $$I = \frac{1}{18}[-1.26018516 + 1.41966658 + 3.48936887] = 0.20271391.$$

4.27 Determine the coefficients in the formula

$$\int_0^{2h} x^{-1/2} f(x)dx = (2h)^{1/2}[A_0 f(0) + A_1 f(h) + A_2 f(2h)] + R$$

and calculate the remainder R, when $f'''(x)$ is constant.

(Gothenburg Univ., Sweden, BIT 4(1964), 61)

Solution

Making the method exact for $f(x) = 1$, x and x^2, we get

$$f(x) = 1: \quad 2\sqrt{2h} = \sqrt{2h}\,(A_0 + A_1 + A_2)$$

$$\text{or} \quad A_0 + A_1 + A_2 = 2.$$

$$f(x) = x: \quad \frac{4h\sqrt{2h}}{3} = \sqrt{2h}\,(A_1 h + 2A_2 h)$$

$$\text{or} \quad A_1 + 2A_2 = \frac{4}{3}.$$

$$f(x) = x^2: \quad \frac{8h^2\sqrt{2h}}{5} = \sqrt{2h}\,(A_1 h^2 + 4A_2 h^2)$$

$$\text{or} \quad A_1 + 4A_2 = \frac{8}{5}.$$

Solving the above system of equations, we obtain

$$A_0 = 12 / 15,\ A_1 = 16 / 15 \text{ and } A_2 = 2 / 15.$$

The remainder R is given by

$$R = \frac{C}{3!} f'''(\xi), \quad 0 < \xi < 2h$$

where
$$C = \int_0^{2h} x^{-1/2}(x^3)dx - \sqrt{2h}\,[A_1 h^3 + 8A_2 h^3] = \frac{16\sqrt{2}}{105} h^{7/2}.$$

Hence, we have the remainder as

$$R = \frac{8\sqrt{2}}{315} h^{7/2} f'''(\xi).$$

4.28 In a quadrature formula

$$\int_{-1}^{1} (a - x) f(x) dx = A_{-1} f(-x_1) + A_0 f(0) + A_1 f(x_1) + R$$

the coefficients A_{-1}, A_0, A_1 are functions of the parameter a, x_1 is a constant and the error R is of the form $Cf^{(k)}(\xi)$. Determine A_{-1}, A_0, A_1 and x_1, so that the error R will be of highest possible order. Also investigate if the order of the error is influenced by different values of the parameter a. (Inst. Tech., Lund, Sweden, BIT 9(1969), 87)

Solution

Making the method exact for $f(x) = 1$, x, x^2 and x^3 we get the system of equations

$$A_{-1} + A_0 + A_1 = 2a,$$

$$x_1(-A_{-1} + A_1) = -\frac{2}{3},$$

$$x_1^2(A_{-1} + A_1) = \frac{2a}{3},$$

$$x_1^3(-A_{-1} + A_1) = -\frac{2}{5},$$

which has the solution

$$x_1 = \sqrt{\frac{3}{5}},\ A_{-1} = \frac{5}{9}\left[a + \sqrt{\frac{3}{5}}\right],$$

$$A_0 = \frac{8a}{9},\ A_1 = \frac{5}{9}\left[a - \sqrt{\frac{3}{5}}\right].$$

The error term in the method is given by

$$R = \frac{C}{4\,!}\, f^{iv}(\xi), \quad -1 < \xi < 1$$

where $$C = \int_{-1}^{1} (a - x)x^4 dx - [x_1^4\,(A_{-1} + A_1)] = 0$$

Therefore, the error term becomes

$$R = \frac{C}{5\,!}\, f^{v}(\xi), \quad -1 < \xi < 1$$

where $$C = \int_{-1}^{1} (a - x)x^5 dx - x_1^5(-A_{-1} + A_1) = -\frac{8}{175}.$$

Hence, we get $$R = -\frac{1}{2625}\, f^{v}(\xi).$$

The order of the method is four for arbitrary a. The error term is independent of a.

4.29 Determine x_i and A_i in the quadrature formula below so that σ, the order of approximation will be as high as possible

$$\int_{-1}^{1} (2x^2 + 1) f(x) dx = A_1 f(x_1) + A_2 f(x_2) + A_3 f(x_3) + R.$$

What is the value of σ ? Answer with 4 significant digits.

(Gothenburg Univ., Sweden, BIT 17 (1977), 369)

Solution

Making the method exact for $f(x) = x^i$, $i = 0, 1, 2, ..., 5$ we get the system of equations

$$A_1 + A_2 + A_3 = \frac{10}{3},$$
$$A_1x_1 + A_2x_2 + A_2x_3 = 0,$$
$$A_1x_1^2 + A_2x_2^2 + A_3x_3^2 = \frac{22}{15},$$
$$A_1x_1^3 + A_2x_2^3 + A_3x_3^3 = 0,$$
$$A_1x_1^4 + A_2x_2^4 + A_3x_3^4 = \frac{34}{35},$$
$$A_1x_1^5 + A_2x_2^5 + A_3x_3^5 = 0,$$

which simplifies to $$A_1(x_3 - x_1) + A_2(x_3 - x_2) = \frac{10}{3} x_3,$$

$$A_1(x_3 - x_1)x_1 + A_2(x_3 - x_2)x_2 = -\frac{22}{15},$$
$$A_1(x_3 - x_1)x_1^2 + A_2(x_3 - x_2)x_2^2 = \frac{22}{15} x_3,$$
$$A_1(x_3 - x_1)x_1^3 + A_2(x_3 - x_2)x_2^3 = -\frac{34}{35},$$
$$A_1(x_3 - x_1)x_1^4 + A_2(x_3 - x_2)x_2^4 = \frac{34}{35} x_3,$$

or
$$A_1(x_3 - x_1)(x_2 - x_1) = \frac{10}{3} x_2 x_3 + \frac{22}{15},$$
$$A_1(x_3 - x_1)(x_2 - x_1)x_1 = -\frac{22}{15}(x_2 + x_3),$$
$$A_1(x_3 - x_1)(x_2 - x_1)x_1^2 = \frac{22}{15} x_2 x_3 + \frac{34}{35},$$
$$A_1(x_3 - x_1)(x_2 - x_1)x_1^3 = -\frac{34}{35}(x_2 + x_3),$$

Solving this system, we have $x_1^2 = \frac{51}{77}$ or $x_1 = \pm 0.8138$ and $x_2 x_3 = 0$.

For $x_2 = 0$, we get $x_3 = -x_1$
$$A_1 = \frac{11}{15x_1^2} = 1.1072,$$
$$A_2 = \frac{10}{3} - 2A_1 = 1.1190,\ A_3 = 1.1072.$$

For $x_3 = 0$, we get the same method.

The error term is obtained as
$$R = \frac{C}{6!} f^{vi}(\xi), \quad -1 < \xi < 1$$

where
$$C = \int_{-1}^{1} (2x^2 + 1)\, x^6 dx - [A_1 x_1^6 + A_2 x_2^6 + A_3 x_3^6] = 0.0867.$$

The order σ, of approximation is 5.

4.30 Find a quadrature formula
$$\int_0^1 \frac{f(x)dx}{\sqrt{x(1-x)}} = \alpha_1 f(0) + \alpha_2 f\left(\frac{1}{2}\right) + \alpha_3 f(1)$$
which is exact for polynomials of highest possible degree. Then use the formula on
$$\int_0^1 \frac{dx}{\sqrt{x - x^3}}$$
and compare with the exact value. (Oslo Univ., Norway, BIT 7(1967), 170)

Solution

Making the method exact for polynomials of degree upto 2, we obtain
$$\text{for } f(x) = 1: \quad I_1 = \int_0^1 \frac{dx}{\sqrt{x(1-x)}} = \alpha_1 + \alpha_2 + \alpha_3,$$
$$\text{for } f(x) = x: \quad I_2 = \int_0^1 \frac{x\,dx}{\sqrt{x(1-x)}} = \frac{1}{2}\alpha_2 + \alpha_3,$$
$$\text{for } f(x) = x^2: \quad I_3 = \int_0^1 \frac{x^2 dx}{\sqrt{x(1-x)}} = \frac{1}{4}\alpha_2 + \alpha_3,$$

where
$$I_1 = \int_0^1 \frac{dx}{\sqrt{x(1-x)}} = 2\int_0^1 \frac{dx}{\sqrt{1-(2x-1)^2}} = \int_{-1}^1 \frac{dt}{\sqrt{1-t^2}} = \sin^{-1} t \Big|_{-1}^{1} = \pi,$$

$$I_2 = \int_0^1 \frac{xdx}{\sqrt{x(1-x)}} = 2\int_0^1 \frac{xdx}{\sqrt{1-(2x-1)^2}} = \int_{-1}^1 \frac{(t+1)}{2\sqrt{1-t^2}}dt$$

$$= \frac{1}{2}\int_{-1}^1 \frac{tdt}{\sqrt{1-t^2}} + \frac{1}{2}\int_{-1}^1 \frac{dt}{\sqrt{1-t^2}} dt = \frac{\pi}{2},$$

$$I_3 = \int_0^1 \frac{x^2dx}{\sqrt{x(1-x)}} = 2\int_0^1 \frac{x^2dx}{\sqrt{1-(2x-1)^2}} = \frac{1}{4}\int_{-1}^1 \frac{(t+1)^2}{\sqrt{1-t^2}}dt$$

$$= \frac{1}{4}\int_{-1}^1 \frac{t^2}{\sqrt{1-t^2}}dt + \frac{1}{2}\int_{-1}^1 \frac{t}{\sqrt{1-t^2}}dt + \frac{1}{4}\int_{-1}^1 \frac{dt}{\sqrt{1-t^2}} = \frac{3\pi}{8}.$$

Hence, we have the equations

$$\alpha_1 + \alpha_2 + \alpha_3 = \pi,$$
$$\frac{1}{2}\alpha_2 + \alpha_3 = \frac{\pi}{2},$$
$$\frac{1}{4}\alpha_2 + \alpha_3 = \frac{3\pi}{8},$$

which gives $\alpha_1 = \pi / 4$, $\alpha_2 = \pi / 2$, $\alpha_3 = \pi / 4$.

The quadrature formula is given by

$$\int_0^1 \frac{f(x)dx}{\sqrt{x(1-x)}} = \frac{\pi}{4}\left[f(0) + 2f\left(\frac{1}{2}\right) + f(1)\right].$$

We now use this formula to evaluate

$$I = \int_0^1 \frac{dx}{\sqrt{x - x^3}} = \int_0^1 \frac{dx}{\sqrt{1+x}\,\sqrt{x(1-x)}} = \int_0^1 \frac{f(x)dx}{\sqrt{x(1-x)}}$$

where $f(x) = 1/\sqrt{1+x}$.

We obtain

$$I = \frac{\pi}{4}\left[1 + \frac{2\sqrt{2}}{\sqrt{3}} + \frac{\sqrt{2}}{2}\right] \approx 2.62331.$$

The exact value is

$$I = 2.62205755.$$

4.31 There is a two-point quadrature formula of the form

$$I_2 = w_1 f(x_1) + w_2 f(x_2)$$

where $-1 \le x_1 < x_2 \le 1$ and $w_1 > 0$, $w_2 > 0$ to calculate the integral $\int_{-1}^1 f(x)dx$.

(*a*) Find w_1, w_2, x_1 and x_2 so that $I_2 = \int_{-1}^1 f(x)dx$ when $f(x) = 1, x, x^2$ and x^3.

(*b*) To get a quadrature formula I_n for the integral $\int_a^b f(x)dx$, let $x_i = a + ih$, $i = 0, 1, 2, ..., n$, where $h = (b - a) / n$, and approximate $\int_{x_{i-1}}^{x_i} f(x)dx$ by a suitable variant of the formula in (*a*). State I_n.

(Inst. Tech. Lyngby, Denmark, BIT 25(1985), 428)

Solution

(*a*) Making the method

$$\int_{-1}^{1} f(x)dx = w_1 f(x_1) + w_2 f(x_2)$$

exact for $f(x) = 1, x, x^2$ and x^3, we get the system of equations

$$\begin{aligned} w_1 + w_2 &= 2, \\ w_1 x_1 + w_2 x_2 &= 0, \\ w_1 x_1^2 + w_2 x_2^2 &= 2/3, \\ w_1 x_1^3 + w_2 x_2^3 &= 0, \end{aligned}$$

whose solution is $\quad x_2 = - x_1 = 1/\sqrt{3}, \; w_2 = w_1 = 1.$

Hence $\quad \int_{-1}^{1} f(x)dx = f(-1/\sqrt{3}) + f(1/\sqrt{3})$

is the required formula.

(*b*) We write

$$\begin{aligned} I_n &= \int_a^b f(x)dx \\ &= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{i-1}}^{x_i} f(x)dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)dx \end{aligned}$$

where $\quad x_0 = a, x_n = b, x_i = x_0 + ih, h = (b - a) / n.$

Using the transformation

$$x = \frac{1}{2}[(x_i - x_{i-1})t + (x_i + x_{i-1})] = \frac{h}{2} t + m_i$$

where $m_i = (x_i + x_{i-1}) / 2$, we obtain, on using the formula in (*a*),

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{h}{2}\left[f\left(m_i - \frac{h\sqrt{3}}{6}\right) + f\left(m_i + \frac{h\sqrt{3}}{6}\right)\right].$$

Hence, we get

$$I_n = \frac{h}{2}\sum_{i=1}^{n}\left[f\left(m_i - \frac{h\sqrt{3}}{6}\right) + f\left(m_i + \frac{h\sqrt{3}}{6}\right)\right].$$

4.32 Compute by Gaussian quadrature

$$I = \int_0^1 \frac{\ln(x+1)}{\sqrt{x(1-x)}} dx$$

The error must not exceed 5×10^{-5}. (Uppsala Univ., Sweden, BIT 5(1965), 294)

Solution

Using the transformation, $x = (t + 1) / 2$, we get

$$I = \int_0^1 \frac{\ln (x+1)}{\sqrt{x(1-x)}} dx = \int_{-1}^1 \frac{\ln \{(t+3)/2\}}{\sqrt{1-t^2}} dt$$

Using Gauss-Chebyshev integration method

$$\int_{-1}^1 \frac{f(t)}{\sqrt{1-t^2}} dt = \sum_{k=0}^{n} \lambda_k f(t_k)$$

where $$t_k = \cos\left(\frac{(2k+1)\pi}{2n+2}\right),\ k = 0, 1, ..., n,$$

$$\lambda_k = \pi / (n + 1),\ k = 0, 1, ..., n,$$

we get for $f(t) = \ln \{(t + 3) / 2\}$, and

$$n = 1: \quad I = \frac{\pi}{2}\left[f\left(-\frac{1}{\sqrt{2}}\right) + f\left(\frac{1}{\sqrt{2}}\right)\right] = 1.184022,$$

$$n = 2: \quad I = \frac{\pi}{3}\left[f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right)\right] = 1.182688,$$

$$n = 3: \quad I = \frac{\pi}{4}\left[f\left(\cos\left(\frac{\pi}{8}\right)\right) + f\left(\cos\left(\frac{3\pi}{8}\right)\right) + f\left(-\cos\left(\frac{3\pi}{8}\right)\right) + f\left(-\cos\left(\frac{\pi}{8}\right)\right)\right]$$
$$= 1.182662.$$

Hence, the result correct to five decimal places is $I = 1.18266$.

4.33 Calculate

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

correct to four decimal places. (Lund Univ., Sweden, BIT 20(1980), 389)

Solution

Since, the integrand is an even function, we write the integral as

$$I = \int_0^1 \frac{\cos(2x)}{\sqrt{1-x^2}} dx = \frac{1}{2}\int_{-1}^1 \frac{\cos(2x)}{\sqrt{1-x^2}} dx .$$

Using the Gauss-Chebyshev integration method, we get for $f(x) = (\cos(2x)) / 2$, (see problem 4.32)

$n = 1: \quad I = 0.244956,$

$n = 2: \quad I = 0.355464,$

$n = 3: \quad I = 0.351617,$

$$n = 4: \quad I = \frac{\pi}{5}\left[f\left(\cos\left(\frac{\pi}{10}\right)\right) + f\left(\cos\left(\frac{3\pi}{10}\right)\right) f(0) + f\left(-\cos\left(\frac{3\pi}{10}\right)\right) + f\left(-\cos\left(\frac{\pi}{10}\right)\right)\right]$$
$$= 0.351688.$$

Hence, the result correct to four decimal places is $I = 0.3517$.

4.34 Compute the value of the integral

$$\int_{0.5}^{1.5} \frac{2 - 2x + \sin(x - 1) + x^2}{1 + (x - 1)^2} dx$$

with an absolute error less than 10^{-4}. (Uppsala Univ., Sweden, BIT 27(1987), 130)

Solution

Using the trapezoidal rule, we get

$$h = 1.0 : \quad I = \frac{1}{2} [f(0.5) + f(1.5)] = 1.0.$$

$$h = 0.5 \quad I = \frac{1}{4} [f(0.5) + 2f(1) + f(1.5)] = 1.0.$$

Hence, the solution is $I = 1.0$.

4.35 Derive a suitable two point and three point quadrature formulas to evaluate

$$\int_0^{\pi/2} \left(\frac{1}{\sin x}\right)^{1/4} dx$$

Obtain the result correct to 3 decimal places. Assume that the given integral exists.

Solution

The integrand and its derivatives are all singular at $x = 0$. The open type formulas or a combination of open and closed type formulas discussed in the text converge very slowly. We write

$$\int_0^{\pi/2} \left(\frac{1}{\sin x}\right)^{1/4} dx = \int_0^{\pi/2} x^{-1/4} \left(\frac{x}{\sin x}\right)^{1/4} dx$$

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

We shall first construct quadrature rules for evaluating this integral.

We write

$$\int_0^{\pi/2} x^{-1/4} f(x) dx = \sum_{i=0}^{n} \lambda_i f(x_i).$$

Making the formula exact for $f(x) = 1, x, x^2, \ldots$, we obtain the following results for $n = 1$ and 2

	x_i	λ_i
$n = 1$	0.260479018	1.053852181
	1.205597553	0.816953346
$n = 2$	0.133831762	0.660235355
	0.739105922	0.779965743
	1.380816210	0.430604430

Using these methods with $f(x) = (x / \sin x)^{1/4}$, we obtain for

$n = 1 :$ $\quad I = 1.927616.$

$n = 2 :$ $\quad I = 1.927898.$

Hence, the result correct to 3 decimals is 1.928.

4.36 Compute

$$\int_0^{\pi/2} \frac{\cos x \log_e(\sin x)}{1+\sin^2 x}\, dx$$

to 2 correct decimal places. (Uppsala Univ., Sweden, BIT 11(1971), 455)

Solution

Substituting $\sin x = e^{-t}$, we get

$$I = -\int_0^{\infty} e^{-t} \left(\frac{t}{1+e^{-2t}}\right) dt.$$

We can now use the Gauss-Laguerre's integration methods (4.71) for evaluating the integral with $f(t) = t / (1 + e^{-2t})$. We get for

$n = 1:$ $I = -[0.3817 + 0.4995] = -0.8812.$

$n = 2:$ $I = -[0.2060 + 0.6326 + 0.0653] = -0.9039.$

$n = 3:$ $I = -[0.1276 + 0.6055 + 0.1764 + 0.0051] = -0.9146.$

$n = 4:$ $I = -[0.0865 + 0.5320 + 0.2729 + 0.0256 + .0003] = -0.9173.$

$n = 5:$ $I = -[0.0624 + 0.4537 + 0.3384 + 0.0601 + 0.0026 + 0.0000]$
$= -0.9172.$

Hence, the required value of the integral is – 0.917 or – 0.92.

4.37 Compute

$$\int_0^{0.8} \left(1+\frac{\sin x}{x}\right) dx$$

correct to five decimals. (Umea Univ., Sweden, BIT 20(1980), 261)

Solution

We have

$$I = 0.8 + \int_0^{0.8} \left(\frac{\sin x}{x}\right) dx.$$

The integral on the right hand side can be evaluated by the open type formulas. Using the methods (4.50) with $f(x) = \sin x / x$, we get for

$n = 2:$ $I = 0.8 + 0.8 f(0.4) = 1.578837.$

$n = 3:$ $I = 0.8 + \frac{3}{2}\left(\frac{0.8}{3}\right)\left[f\left(\frac{0.8}{3}\right) + f\left(\frac{1.6}{3}\right)\right] = 1.576581.$

$n = 4:$ $I = 0.8 + \left(\frac{0.8}{3}\right)[2f(0.2) - f(0.4) + 2f(0.6)] = 1.572077.$

$n = 5:$ $I = 0.8 + \left(\frac{0.8}{24}\right) \times [11f(0.16) + f(0.32) + f(0.48) + 11f(0.64)] = 1.572083.$

Hence, the solution correct to five decimals is 1.57208.

4.38 Integrate by Gaussian quadrature ($n = 3$)

$$\int_1^2 \frac{dx}{1+x^3}.$$

Solution

Using the transformation $x = (t + 3) / 2$, we get

$$I = \int_1^2 \frac{dx}{1+x^3} = \frac{1}{2}\int_{-1}^{1} \frac{dt}{1+[(t+3)/2]^3}.$$

Using the Gauss-Legendre four-point formula

$$\int_{-1}^{1} f(x)dx = 0.652145\,[f(0.339981) + f(-0.339981)] + 0.347855\,[f(0.861136) + f(-0.861136)]$$

we obtain $I = \frac{1}{2}$ [0.652145 (0.176760 + 0.298268) + 0.347855(0.122020 + 0.449824)]

$= 0.254353.$

4.39 Use Gauss-Laguerre or Gauss-Hermite formulas to evaluate

(*i*) $\displaystyle\int_0^\infty \frac{e^{-x}}{1+x}\,dx$, (*ii*) $\displaystyle\int_0^\infty \frac{e^{-x}}{\sin x}\,dx$,

(*iii*) $\displaystyle\int_{-\infty}^\infty \frac{e^{-x^2}}{1+x^2}\,dx$, (*iv*) $\displaystyle\int_{-\infty}^\infty e^{-x^2}\,dx$.

Use two-point and three-point formulas.

Solution

(*i, ii*) Using the Gauss-Laguerre two-point formula

$$\int_0^\infty e^{-x} f(x)dx = 0.853553\, f(0.585786) + 0.146447\, f(3.414214)$$

we obtain $I_1 = \displaystyle\int_0^\infty \frac{e^{-x}}{1+x}\,dx = 0.571429$, where $f(x) = \dfrac{1}{1+x}$.

$I_2 = \displaystyle\int_0^\infty e^{-x} \sin x\, dx = 0.432459$, where $f(x) = \sin x$.

Using the Gauss-Laguerre three-point formula

$$\int_0^\infty e^{-x} f(x)dx = 0.711093\, f(0.415775) + 0.278518\, f(2.294280) + 0.010389\, f(6.289945)$$

we obtain $I_1 = \displaystyle\int_0^\infty \frac{e^{-x}}{1+x}\,dx = 0.588235.$

$I_2 = \displaystyle\int_0^\infty e^{-x} \sin x\, dx = 0.496030.$

(*iii, iv*) Using Gauss-Hermite two-point formula

$$\int_{-\infty}^\infty e^{-x^2} f(x)dx = 0.886227\,[f(0.707107) + f(-0.707107)]$$

we get $I_3 = \displaystyle\int_{-\infty}^\infty \frac{e^{-x^2}}{1+x^2}\,dx = 1.181636$, where $f(x) = \dfrac{1}{1+x^2}$.

$I_4 = \displaystyle\int_{-\infty}^\infty e^{-x^2}\,dx = 1.772454$, where $f(x) = 1$.

Using Gauss-Hermite three-point formula

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx = 1.181636\, f(0) + 0.295409\, [f(1.224745) + f(-1.224745)]$$

we obtain
$$I_3 = \int_{-\infty}^{\infty} \frac{e^{-x^2}}{1+x^2} dx = 1.417963.$$

$$I_4 = \int_{-\infty}^{\infty} e^{-x^2}\, dx = 1.772454.$$

4.40 Obtain an approximate value of

$$I = \int_{-1}^{1} (1-x^2)^{1/2} \cos x\, dx$$

using

(a) Gauss-Legendre integration method for $n = 2, 3$.

(b) Gauss-Chebyshev integration method for $n = 2, 3$.

Solution

(a) Using Gauss-Legendre three-point formula

$$\int_{-1}^{1} f(x)dx = \frac{1}{9}\,[5f(-\sqrt{0.6}) + 8f(0) + 5f(\sqrt{0.6})]$$

we obtain
$$I = \frac{1}{9}\,[5\sqrt{0.4}\cos\sqrt{0.6} + 8 + 5\sqrt{0.4}\cos\sqrt{0.6}]$$
$$= 1.391131.$$

Using Gauss-Legendre four-point formula

$$\int_{-1}^{1} f(x)dx = 0.652145\, [f(0.339981) + f(-0.339981)] + 0.347855\, [f(0.861136) + f(-0.861136)]$$

we obtain
$$I = 2 \times 0.652145\, [\sqrt{1-(0.339981)^2}\, \cos(0.339981)] + 2 \times 0.347855\, [\sqrt{1-(0.861136)^2}\, \cos(0.861136)]$$
$$= 1.156387 + 0.230450 = 1.3868837.$$

(b) We write
$$I = \int_{-1}^{1} \sqrt{1-x^2}\, \cos x\, dx = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\, f(x)\, dx$$

where $f(x) = (1 - x^2) \cos x$.

Using Gauss-Chebyshev three-point formula

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)dx = \frac{\pi}{3}\left[f\left(\frac{\sqrt{3}}{2}\right) + f(0) + f\left(-\frac{\sqrt{3}}{2}\right)\right]$$

we obtain
$$I = \frac{\pi}{3}\left[\frac{1}{4}\cos\frac{\sqrt{3}}{2} + 1 + \frac{1}{4}\cos\frac{\sqrt{3}}{2}\right] = 1.386416.$$

Using Gauss-Chebyshev four-point formula

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)dx = \frac{\pi}{4} [f(0.923880) + f(0.382683) + f(-0.382683) + f(-0.923880)]$$

we obtain $$I = \frac{\pi}{4} [2(0.088267) + 2(0.791813)] = 1.382426.$$

4.41 The Radau quadrature formula is given by

$$\int_{-1}^{1} f(x)dx = B_1 f(-1) + \sum_{k=1}^{n} H_k f(x_k) + R$$

Determine x_k, H_k and R for $n = 1$.

Solution

Making the method

$$\int_{-1}^{1} f(x)dx = B_1 f(-1) + H_1 f(x_1) + R$$

exact for $f(x) = 1$, x and x^2, we obtain the system of equations

$$B_1 + H_1 = 2,$$
$$-B_1 + H_1 x_1 = 0,$$
$$B_1 + H_1 x_1^2 = 2/3,$$

which has the solution $x_1 = 1/3$, $H_1 = 3/2$, $B_1 = 1/2$.

Hence, we obtain the method

$$\int_{-1}^{1} f(x)dx = \frac{1}{2} f(-1) + \frac{3}{2} f\left(\frac{1}{3}\right).$$

The error term is given by

$$R = \frac{C}{3!} f'''(\xi), \quad -1 < \xi < 1$$

where $$C = \int_{-1}^{1} x^3 dx - [-B_1 + H_1 x_1^3] = \frac{4}{9}.$$

Hence, we have

$$R = \frac{2}{27} f'''(\xi), -1 < \xi < 1.$$

4.42 The Lobatto quadrature formula is given by

$$\int_{-1}^{1} f(x)dx = B_1 f(-1) + B_2 f(1) + \sum_{k=1}^{n-1} H_k f(x_k) + R$$

Determine x_k, H_k and R for $n = 3$.

Solution

Making the method

$$\int_{-1}^{1} f(x)dx = B_1 f(-1) + B_2 f(-1) + H_1 f(x_1) + H_2 f(x_2) + R$$

exact for $f(x) = x^i$, $i = 0, 1, ..., 5$, we obtain the system of equations

$$B_1 + B_2 + H_1 + H_2 = 2,$$
$$-B_1 + B_2 + H_1x_1 + H_2x_2 = 0,$$
$$B_1 + B_2 + H_1x_1^2 + H_2x_2^2 = \frac{2}{3},$$
$$-B_1 + B_2 + H_1x_1^3 + H_2x_2^3 = 0,$$
$$B_1 + B_2 + H_1x_1^4 + H_2x_2^4 = \frac{2}{5},$$
$$-B_1 + B_2 + H_1x_1^5 + H_2x_2^5 = 0,$$

or

$$H_1(1 - x_1^2) + H_2(1 - x_2^2) = \frac{4}{3},$$
$$H_1(1 - x_1^2)x_1 + H_2(1 - x_2^2)x_2 = 0,$$
$$H_1(1 - x_1^2)x_1^2 + H_2(1 - x_2^2)x_2^2 = \frac{4}{15},$$
$$H_1(1 - x_1^2)x_1^3 + H_2(1 - x_2^2)x_2^3 = 0,$$

or

$$H_1(1 - x_1^2)(x_2 - x_1) = \frac{4}{3}x_2,$$
$$H_1(1 - x_1^2)(x_2 - x_1)x_1 = -\frac{4}{15},$$
$$H_1(1 - x_1^2)(x_2 - x_1)x_1^2 = \frac{4}{15}x_2.$$

Solving the system, we get $x_1x_2 = -1/5$, and $x_1 = -x_2$.
The solution is obtained as

$$x_1 = 1/\sqrt{5},\ x_2 = -1/\sqrt{5},$$
$$H_1 = H_2 = 5/6,\ B_1 = B_2 = 1/6.$$

The method is given by

$$\int_{-1}^{1} f(x)dx = \frac{1}{6}\,[f(-1) + f(1)] + \frac{5}{6}\left[f\left(\frac{1}{\sqrt{5}}\right) + f\left(-\frac{1}{\sqrt{5}}\right)\right].$$

The error term is

$$R = \frac{C}{6!}\, f^{vi}(\xi), \quad -1 < \xi < 1$$

where $$C = \int_{-1}^{1} x^6 dx - [B_1 + B_2 + H_1x_1^6 + H_2x_2^6]$$

$$= \left[\frac{2}{7} - \left(\frac{1}{3} + \frac{1}{75}\right)\right] = -\frac{32}{525}.$$

Hence, we have

$$R = -\frac{2}{23625}\, f^{vi}(\xi),\ -1 < \xi < 1.$$

4.43 Obtain the approximate value of

$$I = \int_{-1}^{1} e^{-x^2} \cos x \, dx$$

using

(*a*) Gauss-Legendre integration method for $n = 2, 3$.

(*b*) Radau integration method for $n = 2, 3$.

(*c*) Lobatto integration method for $n = 2, 3$.

Solution

(*a*) Using Gauss-Legendre 3-point formula

$$\int_{-1}^{1} f(x)dx = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

we obtain $I = 1.324708.$

Using Gauss-Legendre 4-point formula

$$\int_{-1}^{1} f(x)dx = 0.652145\,[f(0.339981) + f(-0.339981)] + 0.347855\,[f(0.861136) + f(-0.861136)]$$

we obtain $I = 1.311354.$

(*b*) Using Radau 3-point formula

$$\int_{-1}^{1} f(x)\,dx = \frac{2}{9} f(-1) + \frac{16+\sqrt{6}}{18} f\left(\frac{1-\sqrt{6}}{5}\right) + \frac{16-\sqrt{6}}{18} f\left(\frac{1+\sqrt{6}}{5}\right)$$

we obtain $I = 1.307951.$

Using Radau 4-point formula

$$\int_{-1}^{1} f(x)dx = 0.125000\, f(-1) + 0.657689\, f(-0.575319) + 0.776387\, f(0.181066) + 0.440924\, f(0.822824)$$

we obtain $I = 1.312610.$

(*c*) Using Lobatto 3-point formula

$$\int_{-1}^{1} f(x)\,dx = \frac{1}{3}\,[f(-1) + 4f(0) + f(1)]$$

we obtain $I = 1.465844.$

Using Lobatto 4-point formula

$$\int_{-1}^{1} f(x)dx = 0.166667\,[f(-1) + f(1)] + 0.833333\,[f(0.447214) + f(-0.447214)]$$

we obtain $I = 1.296610.$

4.44 Evaluate

$$I = \int_{0}^{\infty} e^{-x} \log_{10}(1 + x)\, dx$$

correct to two decimal places, using the Gauss-Laguerre's integration methods.

Solution

Using the Gauss-Laguerre's integration methods (4.71) and the abscissas and weights given in Table 4.7, with $f(x) = \log_{10}(1 + x)$, we get for

$n = 1:$ $I = 0.2654.$

$n = 2:$ $I = 0.2605.$

$n = 3:$ $I = 0.2594.$

$n = 4:$ $I = 0.2592.$

Hence, the result correct to two decimals is 0.26.

4.45 Calculate the weights, abscissas and the remainder term in the Gaussian quadrature formula

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{\exp(-t) f(t)}{\sqrt{t}}\, dt = A_1 f(t_1) + A_2 f(t_2) + C f^{(n)}(\xi).$$

(Royal Inst. Tech., Stockholm, Sweden, BIT 20(1980), 529)

Solution

Making the method

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-t} f(t)}{\sqrt{t}}\, dt = A_1 f(t_1) + A_2 f(t_2)$$

exact for $f(t) = 1, t, t^2$ and t^3 we obtain

$$A_1 + A_2 = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-t}}{\sqrt{t}}\, dt \qquad \text{(substitute } \sqrt{t} = T)$$

$$= \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-T^2}\, dT = \frac{2}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = 1.$$

$$A_1 t_1 + A_2 t_2 = \frac{1}{\sqrt{\pi}} \int_0^\infty \sqrt{t} e^{-t}\, dt \qquad \text{(integrate by parts)}$$

$$= \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{e^{-t}}{\sqrt{t}}\, dt = \frac{1}{2}.$$

$$A_1 t_1^2 + A_2 t_2^2 = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{3/2} e^{-t}\, dt \qquad \text{(integrate by parts)}$$

$$= \frac{3}{2\sqrt{\pi}} \int_0^\infty \sqrt{t} e^{-t}\, dt = \frac{3}{4}.$$

$$A_1 t_1^3 + A_2 t_2^3 = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{5/2} e^{-t}\, dt \qquad \text{(integrate by parts)}$$

$$= \frac{5}{2\sqrt{\pi}} \int_0^\infty t^{3/2} e^{-t}\, dt = \frac{15}{8}.$$

Simplifying the above system of equations, we get

$$A_1(t_2 - t_1) = t_2 - \frac{1}{2},$$

$$A_1(t_2 - t_1) t_1 = \frac{1}{2} t_2 - \frac{3}{4},$$

$$A_1(t_2 - t_1)t_1^2 = \frac{3}{4}t_2 - \frac{15}{8},$$

which give $$t_1 = \frac{\frac{1}{2}t_2 - \frac{3}{4}}{t_2 - \frac{1}{2}} = \frac{\frac{3}{4}t_2 - \frac{15}{8}}{\frac{1}{2}t_2 - \frac{3}{4}}.$$

Simplifying, we get

$$4t_2^2 - 12t_2 + 3 = 0, \quad \text{or} \quad t_2 = \frac{3 \pm \sqrt{6}}{2}.$$

We also obtain

$$t_1 = \frac{3 \mp \sqrt{6}}{2}, A_1 = \frac{3+\sqrt{6}}{6}, A_2 = \frac{3-\sqrt{6}}{6}.$$

Hence, the required method is

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-t}}{\sqrt{t}} f(t)dt = \frac{3+\sqrt{6}}{6} f\left(\frac{3-\sqrt{6}}{2}\right) + \frac{3-\sqrt{6}}{6} f\left(\frac{3+\sqrt{6}}{2}\right)$$

The error term is given by

$$R = \frac{C}{4\,!} f^{iv}(\xi), \quad 0 < \xi < \infty$$

where $$C = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{7/2} e^{-t} dt - [A_1 t_1^4 + A_2 t_2^4]$$

$$= \frac{7}{2\sqrt{\pi}} \int_0^\infty t^{5/2} e^{-t} dt - [A_1 t_1^4 + A_2 t_2^4]$$

$$= \frac{105}{16} - \frac{3+\sqrt{6}}{6}\left(\frac{3-\sqrt{6}}{2}\right)^4 - \frac{3-\sqrt{6}}{6}\left(\frac{3+\sqrt{6}}{2}\right)^4 = \frac{105}{16} - \frac{81}{16} = \frac{3}{2}.$$

Hence, the error term is given by $f^{iv}(\xi) / 16$.

4.46 The total emission from an absolutely black body is given by the formula

$$E = \int_0^\infty E(\nu) d\nu = \frac{2\pi h}{c^3} \int_0^\infty \frac{\nu^3 d\nu}{e^{h\nu/kT} - 1}.$$

Defining $x = h\nu \,/\, kT$, we get

$$E = \frac{2\pi h}{c^3} \left(\frac{kT}{h}\right)^4 \int_0^\infty \frac{x^3 dx}{e^x - 1}.$$

Calculate the value of the integral correct to 3 decimal places.

(Royal Inst. Tech., Stockholm, Sweden, BIT 19(1979), 552)

Solution

We write

$$I = \int_0^\infty \frac{x^3 dx}{e^x - 1} = \int_0^\infty e^{-x} \left(\frac{x^3}{1 - e^{-x}}\right) dx$$

Applying the Gauss-Laguerre integration methods (4.71) with $f(x) = x^3 / (1 - e^{-x})$, we get for

$n = 1:$ $\quad I = 6.413727,$
$n = 2:$ $\quad I = 6.481130,$
$n = 3:$ $\quad I = 6.494531.$

Hence, the result correct to 3 decimal places is 6.494.

4.47 (*a*) Estimate $\int_0^{0.5}\int_0^{0.5} \frac{\sin xy}{1+xy}\,dx\,dy$ using Simpson's rule for double integrals with both step sizes equal to 0.25.

(*b*) Calculate the same integral correct to 5 decimals by series expansion of the integrand.

(Uppsala Univ., Sweden, BIT 26(1986), 399)

Solution

(*a*) Using Simpson's rule with $h = k = 0.25$, we have three nodal points each, in x and y directions. The nodal points are (0, 0), (0, 1 / 4), (0, 1 / 2), (1 / 4, 0), (1 / 4, 1 / 4), (1 / 4, 1 / 2), (1 / 2, 0), (1 / 2, 1 / 4) and (1 / 2, 1 / 2). Using the double Simpson's rule, we get

$$I = \frac{1}{144}\left[f(0,0) + 4f\left(\frac{1}{4},0\right) + f\left(\frac{1}{2},0\right) + 4\left\{f\left(0,\frac{1}{4}\right) + 4f\left(\frac{1}{4},\frac{1}{4}\right) + f\left(\frac{1}{2},\frac{1}{4}\right)\right\}\right.$$
$$\left. + f\left(0,\frac{1}{2}\right) + 4f\left(\frac{1}{4},\frac{1}{2}\right) + f\left(\frac{1}{2},\frac{1}{2}\right)\right]$$

$$= \frac{1}{144}\,[0 + 0 + 0 + 4\,(0 + 0.235141 + 0.110822) + 0 + 0.443288 + 0.197923]$$
$$= 0.014063.$$

(*b*) Using the series expansions, we get

$$I = \int_0^{1/2}\int_0^{1/2} (1+xy)^{-1} \sin xy\,dx\,dy$$

$$= \int_0^{1/2}\int_0^{1/2} (1 - xy + x^2y^2 - \ldots)\left(xy - \frac{x^3y^3}{6} + \ldots\right)dx\,dy$$

$$= \int_0^{1/2}\int_0^{1/2}\left[xy - x^2y^2 + \frac{5}{6}x^3y^3 - \frac{5}{6}x^4y^4 + \frac{101}{120}x^5y^5 - \frac{101}{120}x^6y^6 + \ldots\right]dx\,dy$$

$$= \int_0^{1/2}\left(\frac{x}{8} - \frac{x^2}{24} + \frac{5x^3}{384} - \frac{5x^4}{960} + \frac{101x^5}{46080} - \frac{101x^6}{107520} + \ldots\right)dx$$

$$= \frac{1}{64} - \frac{1}{576} + \frac{5}{24576} - \frac{5}{153600} + \frac{101}{17694720} - \frac{101}{96337920} + \ldots = 0.014064.$$

4.48 Evaluate the double integral

$$\int_0^1\left(\int_1^2 \frac{2xy}{(1+x^2)(1+y^2)}\,dy\right)dx$$

using

(*i*) the trapezoidal rule with $h = k = 0.25$.

(*ii*) the Simpson's rule with $h = k = 0.25$.

Compare the results obtained with the exact solution.

Solution

Exact solution is obtained as

$$I = \int_0^1 \frac{2x}{1+x^2}\, dx \,.\int_1^2 \frac{y}{1+y^2}\, dy \;=\; \frac{1}{2}\Big[\ln(1+x^2)\Big]_0^1 \Big[\ln(1+y^2)\Big]_1^2$$

$$= \frac{1}{2}(\ln 2)\ln(5/2) = 0.317562.$$

With $h = k = 1/4$, we have the nodal points

$$(x_i, y_i),\ i = 0, 1, 2, 3, 4,\ j = 0, 1, 2, 3, 4,$$

where $x_i = i/4$, $i = 0, 1, ..., 4$; $y_j = 1 + (j/4)$, $j = 0, 1, ..., 4$.

Using the trapezoidal rule, we obtain

$$I = \int_0^1 \int_1^2 f(x, y) dy\, dx$$

$$= \frac{k}{2}\int_0^1 [f(x, y_0) + 2f(x, y_1) + 2f(x, y_2) + 2f(x, y_3) + f(x, y_4)]\, dx$$

$$= \frac{hk}{4}[S_1 + 2S_2 + 4S_3] = \frac{1}{64}(S_1 + 2S_2 + 4S_3)$$

where $\quad S_1 = f(x_0, y_0) + f(x_4, y_0) + f(x_0, y_4) + f(x_4, y_4) = 0.9.$

$$S_2 = \sum_{i=1}^{3} [f(x_i, y_0) + f(x_i, y_4)] + \sum_{j=1}^{3} [f(x_0, y_j) + f(x_4, y_j)] = 3.387642.$$

$$S_3 = \sum_{i=1}^{3} [f(x_i, y_1) + f(x_i, y_2) + f(x_i, y_3)] = 3.078463.$$

Hence, we get $\quad I = 0.312330.$

Using Simpson's rule, we obtain

$$I = \frac{k}{3}\int_0^1 [f(x, y_0) + 4f(x, y_1) + 4f(x, y_3) + 2f(x, y_2) + f(x, y_4)]\, dx$$

$$= \frac{hk}{9}[T_1 + 2T_2 + 4T_3 + 8T_4 + 16T_5]$$

$$= \frac{1}{144}[T_1 + 2T_2 + 4T_3 + 8T_4 + 16T_5]$$

where $\quad T_1 = f(x_0, y_0) + f(x_4, y_0) + f(x_0, y_4) + f(x_4, y_4) = 0.9.$

$$T_2 = f(x_2, y_0) + f(x_2, y_4) + f(x_0, y_2) + f(x_4, y_2) = 1.181538.$$

$$T_3 = f(x_0, y_1) + f(x_4, y_1) + f(x_0, y_3) + f(x_4, y_3) + f(x_1, y_4) + f(x_3, y_4) + f(x_1, y_0) + f(x_3, y_0) + f(x_2, y_2)$$

$$= 2.575334.$$

$$T_4 = f(x_2, y_1) + f(x_2, y_3) + f(x_1, y_2) + f(x_3, y_2) = 1.395131.$$

$$T_5 = f(x_1, y_1) + f(x_3, y_1) + f(x_1, y_3) + f(x_3, y_3) = 1.314101.$$

Hence, we get $\quad I = 0.317716.$

4.49 Evaluate the double integral

$$\int_1^5 \left(\int_1^5 \frac{dx}{(x^2 + y^2)^{1/2}}\right) dy$$

using the trapezoidal rule with two and four subintervals and extrapolate.

Solution

With $h = k = 2$, the nodal point are

$(1, 1), (3, 1), (5, 1), (1, 3), (3, 3), (5, 3), (1, 5), (3, 5), (5, 5).$

Using the trapezoidal rule, we get

$$I = \frac{2 \times 2}{4} [f(1, 1) + 2f(1, 3) + f(1, 5) + 2\{f(3, 1) + 2f(3, 3) + f(3, 5)\}$$
$$+ f(5, 1) + 2f(5, 3) + f(5, 5)]$$
$$= 4.1345.$$

With $h = k = 1$, the nodal points are

$(i, j), i = 1, 2, ..., 5, j = 1, 2, ..., 5.$

Using the trapezoidal rule, we get

$$I = \frac{1}{4} [f(1, 1) + 2(f(1, 2) + f(1, 3) + f(1, 4)) + f(1, 5)$$
$$+ 2\{f(2, 1), + 2(f(2, 2) + f(2, 3) + f(2, 4)) + f(2, 5)\}$$
$$+ 2\{f(3, 1) + 2(f(3, 2) + f(3, 3) + f(3, 4)) + f(3, 5)\}$$
$$+ 2\{f(4, 1) + 2(f(4, 2) + f(4, 3) + f(4, 4)) + f(4, 5)\}$$
$$+ f(5, 1) + 2(f(5, 2) + f(5, 3) + f(5, 4)) + f(5, 5)]$$
$$= 3.9975.$$

Using extrapolation, we obtain the better approximation as

$$I = \frac{4(3.9975) - 4.1345}{3} = 3.9518.$$

4.50 A three dimensional Gaussian quadrature formula has the form

$$\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} f(x, y, z) dx\, dy\, dz = f(\alpha, \alpha, \alpha) + f(-\alpha, \alpha, \alpha) + f(\alpha, -\alpha, \alpha)$$
$$+ f(\alpha, \alpha, -\alpha) + f(-\alpha, -\alpha, \alpha) + f(-\alpha, \alpha, -\alpha)$$
$$+ f(\alpha, -\alpha, -\alpha) + f(-\alpha, -\alpha, -\alpha) + R$$

Determine α so that $R = 0$ for every f which is a polynomial of degree 3 in 3 variables *i.e.*

$$f = \sum_{i, j, k = 0}^{3} a_{ijk} x^i y^j z^k \qquad \text{(Lund Univ., Sweden, BIT 15(1975), 111)}$$

Solution

For $i = j = k = 0$, the method is exact.

The integral

$$\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} x^i y^j z^k \, dx\, dy\, dz = 0$$

when i and / or j and / or k is odd. In this case also, the method is exact.

For $f(x, y, z) = x^2y^2z^2$, we obtain

$$\frac{8}{27} = 8\alpha^6.$$

The value of α is therefore $\alpha = 1 / \sqrt{3}$.

Note that $\alpha = -1 / \sqrt{3}$ gives the same expression on the right hand side.

CHAPTER 5

Numerical Solution of Ordinary Differential Equations

5.1 INTRODUCTION

Many problems in science and engineering can be formulated either in terms of the initial value problems or in terms of the boundary value problems.

Initial Value Problems

An mth order initial value problem (IVP), in its canonical representation, can be written in the form

$$u^{(m)} = f(x, u, u', \ldots, u^{(m-1)}) \tag{5.1 i}$$

$$u(a) = \eta_1,\ u'(a) = \eta_2,\ \ldots,\ u^{(m-1)}(a) = \eta_m \tag{5.1 ii}$$

where m represents the highest order derivative.

The equation (5.1 i) can be expressed as an equivalent system of m first order equations

$$\begin{aligned} y_1 &= u, \\ y_1' &= y_2, \\ y_2' &= y_3, \\ &\ldots\ldots\ldots\ldots \\ y_{m-1}' &= y_m, \\ y_m' &= f(x, y_1, y_2, \ldots, y_m). \end{aligned} \tag{5.2 i}$$

The initial conditions become

$$y_1(a) = \eta_1,\ y_2(a) = \eta_2,\ \ldots,\ y_m(a) = \eta_m. \tag{5.2 ii}$$

The system of equations (5.2 i) and the initial conditions (5.2 ii), in vector form becomes

$$\begin{aligned} \mathbf{y}' &= \mathbf{f}(x, \mathbf{y}), \\ \mathbf{y}(a) &= \boldsymbol{\eta}, \end{aligned} \tag{5.3}$$

where $\mathbf{y} = [y_1 \quad y_2 \ldots y_m]^T, \quad \boldsymbol{\eta} = [\eta_1 \quad \eta_2 \ldots \eta_m]^T$

$$\mathbf{f}(x, \mathbf{y}) = [y_2 \quad y_3 \ldots f\,]^T.$$

Thus, the methods of solution for the first order initial value problem (IVP)

$$\begin{aligned} \frac{dy}{dx} &= f(x, y), \\ y(a) &= \eta, \end{aligned} \tag{5.4}$$

may be used to solve the system of first order initial value problems (5.3) and the mth order initial value problem (5.1).

The behaviour of the solution of (5.4) can be predicted by considering the homogeneous linearized form

$$\frac{dy}{dx} = \lambda y,$$
$$y(a) = \eta, \tag{5.5}$$

where λ may be regarded as a constant. The equation (5.5) is called a *test problem*.

We will assume the existence and uniqueness of the solution of (5.4) and also that $f(x, y)$ has continuous partial derivatives with respect to x and y of as high an order as required.

Boundary Value Problems

An mth order boundary value problem (BVP) can be represented symbolically as

$$L\,y = r(x),$$
$$U_\mu y = \gamma_\mu, \quad \mu = 1, 2, ..., m \tag{5.6}$$

where L is an mth order differential operator, $r(x)$ is a given function of x and U_μ are the m boundary conditions.

The simplest boundary value problem is given by a second order differential equation of the form

$$-y'' + p(x)y' + q(x)y = r(x), \tag{5.7}$$

where $p(x)$, $q(x)$ and $r(x)$ are continuous functions of x or constants, with one of the three boundary conditions

(*i*) *first kind :* $\quad y(a) = \gamma_1, y(b) = \gamma_2,$ (5.8)

(*ii*) *second kind :* $\quad y'(a) = \gamma_1, y'(b) = \gamma_2,$ (5.9)

(*iii*) *third kind :* $\quad a_0 y(a) - a_1 y'(a) = \gamma_1,$
$\quad b_0 y(b) + b_1 y'(b) = \gamma_2.$ (5.10)

A homogeneous boundary value problem possesses only a trivial solution $y(x) \equiv 0$. We, therefore, consider those boundary value problems in which a parameter λ occurs either in the differential equation or in the boundary conditions, and we determine values of λ, called *eigenvalues*, for which the boundary value problem has a nontrivial solution. Such a solution is called an *eigenfunction* and the entire problem is called an *eigenvalue* or *a characteristic value problem.*

In general, a boundary value problem does not always have a unique solution. However, we shall assume that the boundary value problem under consideration has a unique solution.

Difference Equations

A k-th order linear nonhomogeneous difference equation with constant coefficients may be written as

$$a_0 y_{m+k} + a_1 y_{m+k-1} + ... + a_k y_m = \phi_m \tag{5.11}$$

where m can take only the integer values and $a_0, a_1, a_2, ..., a_k$ are constants.

The general solution of (5.11) is of the form

$$y_m = y_m^{(H)} + y_m^{(P)} \tag{5.12}$$

where $y_m^{(H)}$ is the solution of the associated homogeneous difference equation

$$a_0 y_{m+k} + a_1 y_{m+k-1} + ... + a_k y_m = 0 \tag{5.13}$$

and $y_m^{(P)}$ is any particular solution of (5.11).

In order to obtain $y_m^{(H)}$, we attempt to determine a solution of the form

$$y_m = \xi^m. \tag{5.14}$$

Substituting (5.14) into (5.13), we get the polynomial equation

$$a_0 \xi^{m+k} + a_1 \xi^{m+k-1} + ... + a_k\, \xi^m = 0$$

or

$$a_0 \xi^k + a_1 \xi^{k-1} + ... + a_k = 0. \tag{5.15}$$

This equation is called the *characteristic equation* of (5.13). The form of the complementary solution $y_m^{(H)}$ depends upon the nature of the roots of (5.15).

Real and Unequal Roots

If the roots of (5.15) are real and unequal, then the solution of (5.13) is of the form

$$y_m^{(H)} = C_1\xi_1^m + C_2\xi_2^m + \ldots + C_k\xi_k^m \tag{5.16}$$

where C_i's are arbitrary constants and ξ_i, $i = 1(1)k$ are the real and unequal roots of (5.15).

Real and p, $(p \le k)$ Equal Roots

The form of the solution (5.16) gets modified to

$$y_m^{(H)} = (C_1 + C_2 m + \ldots + C_p m^{p-1})\, \xi^m + C_{p+1}\, \xi^m_{p+1} + \ldots + C_k\, \xi_k^m \tag{5.17}$$

where $\quad \xi_1 = \xi_2 = \ldots = \xi_p = \xi.$

Two Complex Roots and $k - 2$ Distinct and Real Roots

The form of the solution (5.16) becomes

$$y_m^{(H)} = r^m (C_1 \cos m\theta + C_2 \sin m\theta) + C_3\xi_3^m + \ldots + C_k\xi_k^m \tag{5.18}$$

where $\quad \xi_1 = \alpha + i\beta,\ \xi_2 = \alpha - i\beta \quad \text{and} \quad r^2 = (\alpha^2 + \beta^2),\ \theta = \tan^{-1}(\beta / \alpha).$

The particular solution $y_m^{(P)}$ will depend on the form of ϕ_m. If $\phi_m = \phi$, *a* constant, then we have

$$y_m^{(P)} = \phi / (a_0 + a_1 + \ldots + a_k). \tag{5.19}$$

From the form of the solutions (5.16)-(5.18), we conclude that the solution of the difference equation (5.13) will remain bounded as $m \to \infty$ if and only if the roots of the characteristic equation (5.15), ξ_i, lie inside the unit circle in the complex plane and are simple if they lie on the unit circle. This condition is called the *root condition*.

Routh-Hurwitz Criterion

To test the root condition when the degree of the characteristic equation is high, or the coefficients are functions of some parameters, we use the transformation

$$\xi = \frac{1+z}{1-z} \tag{5.20}$$

which maps the interior of the unit circle $|\,\xi\,| = 1$ onto the left half plane $z \le 0$, the unit circle $|\,\xi\,| = 1$ onto the imaginary axis, the point $\xi = 1$ onto $z = 0$, and the point $\xi = -1$ onto $z = -\infty$ as shown in Fig. 5.1.

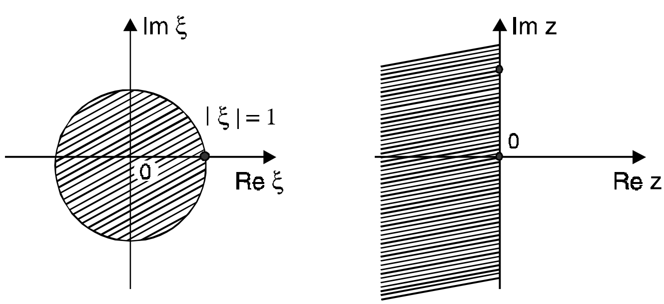


Fig. 5.1. Mapping of the interior of the unit circle onto the left half plane.

Substituting (5.20) into (5.15) and grouping the terms together, we get

$$b_0 z^k + b_1 z^{k-1} + \dots + b_k = 0 \tag{5.21}$$

This is called the *transformed characteristic equation.*

For $k = 1$, we get $b_0 = a_0 - a_1,\ b_1 = a_0 + a_1,$

For $k = 2$, we get $b_0 = a_0 - a_1 + a_2,\ b_1 = 2\,(a_0 - a_2),\ b_2 = a_0 + a_1 + a_2,$

For $k = 3$, we get $b_0 = a_0 - a_1 + a_2 - a_3,\ b_1 = 3a_0 - a_1 - a_2 + 3a_3,$
$b_2 = 3a_0 + a_1 - a_2 - 3a_3,\ b_3 = a_0 + a_1 + a_2 + a_3.$

Note that $\quad b_k = \sum_{i=0}^{k} a_i.$

Denote

$$D = \begin{vmatrix} b_1 & b_3 & b_5 & \cdots & b_{2k-1} \\ b_0 & b_2 & b_4 & \cdots & b_{2k-2} \\ 0 & b_1 & b_3 & \cdots & b_{2k-3} \\ 0 & b_0 & b_2 & \cdots & b_{2k-4} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & b_k \end{vmatrix}$$

where $b_j \geq 0$ for all j. *Routh-Hurwitz criterion* states that the real parts of the roots of (5.21) are negative if and only if the principal minors of D are positive.

Using the Routh-Hurwitz criterion, we obtain for

$$\begin{aligned} k = 1 :&\quad b_0 > 0,\ b_1 > 0, \\ k = 2 :&\quad b_0 > 0,\ b_1 > 0,\ b_2 > 0, \\ k = 3 :&\quad b_0 > 0,\ b_1 > 0,\ b_2 > 0,\ b_3 > 0,\ b_1 b_2 - b_3 b_0 > 0, \end{aligned} \tag{5.22}$$

as the necessary and sufficient conditions for the real parts of the roots of (5.21) to be negative.

If any one or more of the b_i's are equal to zero and other b_j's are positive, then it indicates that a root lies on the unit circle $|\ \xi\ | = 1$. If any one or more of the b_j's are negative, then there is atleast one root for which $|\ \xi_i\ | > 1$.

5.2 SINGLESTEP METHODS

We consider the numerical solution of the initial value problem (5.4)

$$\frac{dy}{dx} = f(x, y), \quad x \in [a, b]$$
$$y(a) = \eta.$$

Divide the interval $[a, b]$ into N equally spaced subintervals such that

$$x_n = a + nh, \quad n = 0, 1, 2, \dots, N,$$
$$h = (b - a) / N.$$

The parameter h is called the step size and x_n, $n = 0(1)N$ are called the mesh or step points.

A *single step method* for (5.4) is a related first order difference equation. A general single step method may be written as

$$y_{n+1} = y_n + h\,\phi\,(t_{n+1}, t_n, y_{n+1}, y_n, h),\; n = 0, 1, 2,... \tag{5.23}$$

where ϕ is a function of the arguments $t_n, t_{n+1}, y_n, y_{n+1}$, h and also depends on f. We often write it as $\phi\,(t, y, h)$. This function ϕ is called the *increment function*. If y_{n+1} can be obtained simply by evaluating the right hand side of (5.23), then the method is called an *explicit method*. In this case, the method is of the form

$$y_{n+1} = y_n + h\,\phi\,(t_n, y_n, h).$$

If the right hand side of (5.23) depends on y_{n+1} also, then it is called an *implicit method*. The general form in this case is as given in (5.23).

The local truncation error T_{n+1} at x_{n+1} is defined by

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h\,\phi\,(t_{n+1}, t_n, y\,(t_{n+1}), y(t_n), h). \tag{5.24}$$

The largest integer p such that

$$|\, h^{-1}\, T_{n+1} \,| = O(h^p) \tag{5.25}$$

is called the *order* of the single step method.

We now list a few single step methods.

Explicit Methods

Taylor Series Method

If the function $y(x)$ is expanded in the Taylor series in the neighbourhood of $x = x_n$, then we have

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2!}\, y_n'' + \ldots + \frac{h^p}{p!}\, y_n^{(p)}$$

$$n = 0, 1, 2, \ldots, N - 1 \tag{5.26}$$

with remainder

$$R_{p+1} = \frac{h^{p+1}}{(p+1)!}\, y^{(p+1)}\,(\xi_n),\; x_n < \xi_n < x_{n+1}.$$

The equation (5.26) is called the *Taylor series method* of order p. The value p is chosen so that $|\, R_{p+1} \,|$ is less than some preassigned accuracy. If this error is ε, then

$$h^{p+1}\,|\, y^{(p+1)}\,(\xi_n) \,| < (p + 1)\,!\;\varepsilon$$

or

$$h^{p+1}\,|\, f^{(p)}\,(\xi_n) \,| < (p + 1)\,!\;\varepsilon.$$

For a given h, this equation will determine p, and if p is specified then it will give an upper bound on h. Since ξ_n is not known, $|\, f^{(p)}\,(\xi_n)$ is replaced by its maximum value in $[x_0, b]$. As the exact solution may not be known, a way of determining this value is as follows. Write one more non-vanishing term in the series than is required and then differentiate this series p times. The maximum value of this quantity in $[x_0, b]$ gives the required bound.

The derivatives $y_n^{(p)}$, $p = 2, 3, \ldots$ are obtained by successively differentiating the differential equation and then evaluating at $x = x_n$. We have

$$y'(x) = f\,(x, y),$$

$$y''(x) = f_x + f f_y$$

$$y'''(x) = f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_y\,(f_x + ff_y).$$

Substituting $p = 1$ in (5.26), we get

$$y_{n+1} = y_n + h\,f_n,\; n = 0(1)N - 1 \tag{5.27}$$

which is known as the *Euler method*.

Runge-Kutta methods

The general Runge-Kutta method can be written as

$$y_{n+1} = y_n + \sum_{i=1}^{v} w_i K_i, \quad n = 0, 1, 2, ..., N-1 \tag{5.28}$$

where $$K_i = hf\left(x_n + c_i h, y_n + \sum_{m=1}^{i-1} a_{im} K_m\right)$$

with $c_1 = 0$.

For $v = 1$, $w_1 = 1$, the equation (5.28) becomes the Euler method with $p = 1$. This is the lowest order Runge-Kutta method. For higher order Runge-Kutta methods, the minimum number of function evaluations (v) for a given order p is as follows :

p	2	3	4	5	6	...
v	2	3	4	6	8	...

We now list a few Runge-Kutta methods.

Second Order Methods

Improved Tangent method :

$$y_{n+1} = y_n + K_2, \quad n = 0(1)N-1, \tag{5.29}$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right).$$

Euler-Cauchy method (Heun method)

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2), \quad n = 0(1)N-1, \tag{5.30}$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + h, y_n + K_1).$$

Third Order Methods

Nyström method

$$y_{n+1} = y_n + \frac{1}{8}(2K_1 + 3K_2 + 3K_3), \quad n = 0(1)N-1, \tag{5.31}$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_1\right),$$
$$K_3 = hf\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_2\right).$$

Heun method

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_3), \quad n = 0(1)N-1, \tag{5.32}$$
$$K_1 = hf(x_n, y_n),$$

$$K_2 = hf\left(x_n + \frac{1}{3}h, y_n + \frac{1}{3}K_1\right)$$

$$K_3 = hf\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_2\right).$$

Classical method

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 4K_2 + K_3), \quad n = 0(1)N-1, \tag{5.33}$$

$$K_1 = hf(x_n, y_n),$$

$$K_2 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1\right),$$

$$K_3 = hf(x_n + h, y_n - K_1 + 2K_2).$$

Fourth Order Methods

Kutta method

$$y_{n+1} = y_n + \frac{1}{8}(K_1 + 3K_2 + 3K_3 + K_4), \quad n = 0(1)N-1, \tag{5.34}$$

$$K_1 = hf(x_n, y_n),$$

$$K_2 = hf\left(x_n + \frac{1}{3}h, y_n + \frac{1}{3}K_1\right),$$

$$K_3 = hf\left(x_n + \frac{2}{3}h, y_n - \frac{1}{3}K_1 + K_2\right),$$

$$K_4 = hf(x_n + h, y_n + K_1 - K_2 + K_3).$$

Classical method

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4), \quad n = 0(1)N-1, \tag{5.35}$$

$$K_1 = hf(x_n, y_n),$$

$$K_2 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1\right)$$

$$K_3 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2\right)$$

$$K_4 = hf(x_n + h, y_n + K_3).$$

Implicit Runge-Kutta Methods

The Runge-Kutta method (5.28) is modified to

$$y_{n+1} = y_n + \sum_{i=1}^{v} w_i K_i, \quad n = 0(1)N-1, \tag{5.36}$$

where
$$K_i = hf\left(x_n + c_i h,\; y_n + \sum_{m=1}^{v} a_{im} K_m\right).$$

With v function evaluations, implicit Runge-Kutta methods of order $2v$ can be obtained. A few methods are listed.

Second Order Method

$$y_{n+1} = y_n + K_1, \quad n = 0(1)N - 1, \tag{5.37}$$

$$K_1 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1\right).$$

Fourth Order Method

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2), \quad n = 0(1)N - 1, \tag{538}$$

$$K_1 = hf\left(x_n + \left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)h,\ y_n + \frac{1}{4}K_1 + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)K_2\right),$$

$$K_2 = hf\left(x_n + \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)h,\ y_n + \left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)K_1 + \frac{1}{4}K_2\right).$$

5.3 MULTISTEP METHODS

The general k-step or multistep method for the solution of the IVP (5.4) is a related k-th order difference equation with constant coefficients of the form

$$\begin{aligned} y_{n+1} &= a_1 y_n + a_2 y_{n-1} + \ldots + a_k y_{n-k+1} \\ &\quad + h\,(b_0 y'_{n+1} + b_1 y'_n + \ldots + b_k y'_{n-k+1}) \end{aligned} \tag{5.39}$$

or symbolically, we write (5.39) as

$$\rho(E)\, y_{n-k+1} - h\sigma(E)\, y'_{n-k+1} = 0$$

where $\rho(\xi) = \xi^k - a_1 \xi^{k-1} - a_2 \xi^{k-2} \ldots - a_k,$

$$\sigma(\xi) = b_0 \xi^k + b_1 \xi^{k-1} + b_2 \xi^{k-2} + \ldots + b_k. \tag{5.40}$$

The formula (5.39) can be used if we know the solution $y(x)$ and $y'(x)$ at k successive points. The k-values will be assumed to be known. Further, if $b_0 = 0$, the method (5.39) is called an *explicit* or a *predictor* method. When $b_0 \neq 0$, it is called an *implicit* or a *corrector* method. The local truncation error of the method (5.39) is given by

$$T_{n+1} = y(x_{n+1}) - \sum_{i=1}^{k} a_i y(x_{n-i+1}) - h \sum_{i=0}^{k} b_i y'(x_{n-i+1}). \tag{5.41}$$

Expanding the terms on the right hand side of (5.41) in Taylor's series and rearranging them we obtain

$$\begin{aligned} T_{n+1} &= C_0 y(x_n) + C_1 h\, y'(x_n) + C_2 h^2 y''(x_n) + \ldots \\ &\quad + C_p h^p y^{(p)}(x_n) + C_{p+1} h^{p+1} y^{(p+1)}(x_n) + \ldots \end{aligned} \tag{5.42}$$

where $$C_0 = 1 - \sum_{m=1}^{k} a_m$$

$$\begin{aligned} C_q &= \frac{1}{q!}\left[1 - \sum_{m=1}^{k} a_m (1 - m)^q\right] \\ &\quad - \frac{1}{(q-1)!} \sum_{m=0}^{k} b_m (1 - m)^{q-1}, \quad q = 1(1)p. \end{aligned} \tag{5.43}$$

Definitions

Order : If $$C_0 = C_1 = \ldots = C_p = 0$$
and $C_{p+1} \neq 0$ in (5.42), then the multistep method (5.39) is said to be of order p.

Consistency : If $p \geq 1$, then the multistep method (5.39) is said to be consistent, *i.e.,* if $C_0 = C_1 = 0$ or

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1). \tag{5.44}$$

Convergence : If $\lim_{h \to 0} y_n = y(x_n),\ 0 \leq n \leq N$ (5.45)

and provided the rounding errors arising from all the initial conditions tend to zero, then the linear multistep method (5.39) is said to be convergent.

Attainable Order of Linear Multistep Methods

As the number of coefficients in (5.39) is $2k + 1$, we may expect that they can be determined so that $2k + 1$ relations of the type (5.43) are satisfied and the order would be equal to $2k$. However, the order of a k-step method satisfying the root condition cannot exceed $k + 2$. If k is odd, then it cannot exceed $k + 1$.

Linear Multistep Methods

We now determine a few wellknown methods which satisfy the root condition.

Putting $y(t) = e^t$ and $e^h = \xi$ into (5.39), we get

$$\rho(\xi) - \log \xi\ \sigma(\xi) = 0. \tag{5.46}$$

As $h \to 0$, $\xi \to 1$ and we may use (5.46) for determining $\rho(\xi)$ or $\sigma(\xi)$ of maximum order if $\sigma(\xi)$ or $\rho(\xi)$ is given.

If $\sigma(\xi)$ is specified, then (5.46) can be used to determine $\rho(\xi)$ of degree k. The term $(\log \xi)\sigma(\xi)$ can be expanded as a power series in $(\xi - 1)$ and terms upto $(\xi - 1)^k$ can be used to give $\rho(\xi)$. Similarly, if $\rho(\xi)$ is given, we can use the equation

$$\sigma(\xi) - \frac{\rho(\xi)}{\log \xi} = 0 \tag{5.47}$$

to determine $\sigma(\xi)$ of degree $\leq k$. The term $\rho(\xi) / (\log \xi)$ is expanded as a power series in $(\xi - 1)$, and terms upto $(\xi - 1)^k$ for implicit methods and $(\xi - 1)^{k-1}$ for explicit methods are used to get $\sigma(\xi)$.

Adams-Bashforth Methods

$$\rho(\xi) = \xi^{k-1} (\xi - 1),$$

$$\sigma(\xi) = \xi^{k-1} \sum_{m=0}^{k-1} \gamma_m (1 - \xi^{-1})^m.$$

$$\gamma_m + \frac{1}{2} \gamma_{m-1} + \ldots + \frac{1}{m+1} \gamma_0 = 1,\ m = 0, 1, 2, \ldots \tag{5.48}$$

We have the following methods.

(*i*) $k = 1 : \gamma_0 = 1$.

$$y_{n+1} = y_n + hy'_n$$

$$T_{n+1} = \frac{1}{2} h^2 y''(x_n) + O(h^3),\ \ p = 1.$$

(*ii*) $k = 2$; $\gamma_0 = 1$, $\gamma_1 = 1/2$.

$$y_{n+1} = y_n + \frac{h}{2} (3y'_n - y'_{n-1}),$$

$$T_{n+1} = \frac{5}{12} h^3 y'''(x_n) + O(h^4), \; p = 2.$$

(*iii*) $k = 3$: $\gamma_0 = 1, \gamma_1 = 1/2, \gamma_2 = 5/12.$

$$y_{n+1} = y_n + \frac{h}{12}(23y'_n - 16y'_{n-1} + 5y'_{n-2}),$$

$$T_{n+1} = \frac{3}{8} h^4 y^{(4)}(x_n) + O(h^5), \; p = 3.$$

Nyström methods

$$\rho(\xi) = \xi^{k-2}(\xi^2 - 1),$$

$$\sigma(\xi) = \xi^{k-1} \sum_{m=0}^{k-1} \gamma_m (1 - \xi^{-1})^m.$$

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \ldots + \frac{1}{m+1}\gamma_0 = \left| \begin{array}{l} 2, \; m = 0 \\ 1, \; m = 1, 2, \ldots \end{array} \right. \tag{5.49}$$

We have the following methods.

(*i*) $k = 2$: $\gamma_0 = 2, \gamma_1 = 0.$

$$y_{n+1} = y_{n-1} + 2hy'_n,$$

$$T_{n+1} = \frac{h^3}{3} y^{(3)}(x_n) + O(h^4), p = 2.$$

(*ii*) $k = 3$: $\gamma_0 = 2, \gamma_1 = 0 \; \gamma_2 = 1/3.$

$$y_{n+1} = y_{n-1} + \frac{h}{3}(7y'_n - 2y'_{n-1} + y'_{n-2}),$$

$$T_{n+1} = \frac{h^4}{3} y^{(4)}(x_n) + O(h^5), \; p = 3.$$

Adams-Moulton Methods

$$\rho(\xi) = \xi^{k-1}(\xi - 1),$$

$$\sigma(\xi) = \xi^k \sum_{m=0}^{k} \gamma_m (1 - \xi^{-1})^m.$$

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \ldots + \frac{1}{m+1}\gamma_0 = \left| \begin{array}{l} 1, \; m = 0 \\ 0, \; m = 1, 2, \ldots \end{array} \right. \tag{5.50}$$

We have the following methods.

(*i*) $k = 1$: $\gamma_0 = 1, \gamma_1 = -1/2.$

$$y_{n+1} = y_n + \frac{h}{2}(y'_{n+1} + y'_n)$$

$$T_{n+1} = -\frac{1}{12} h^3 y'''(x_n) + O(h^4), \; p = 2.$$

(*ii*) $k = 2$: $\gamma_0 = 1, \gamma_1 = -1/2, \gamma_2 = -1/12.$

$$y_{n+1} = y_n + \frac{h}{12}(5y'_{n+1} + 8y'_n - y'_{n-1}),$$

$$T_{n+1} = -\frac{1}{24} h^4 y^{(4)}(x_n) + O(h^5), \; p = 3.$$

(*iii*) $k = 3$: $\gamma_0 = 1, \gamma_1 = -\dfrac{1}{2}, \gamma_2 = -\dfrac{1}{12}, \gamma_3 = -\dfrac{1}{24}.$

$$y_{n+1} = y_n + \frac{h}{24}(9y'_{n+1} + 19y'_n - 5y'_{n-1} + y'_{n-2}),$$

$$T_{n+1} = -\frac{19}{720}h^5 y^{(5)}(x_n) + O(h^6), \ \ p = 4.$$

Milne Method

$$\rho(\xi) = \xi^{k-2}(\xi^2 - 1),$$

$$\sigma(\xi) = \xi^k \sum_{m=0}^{k} \gamma_m (1 - \xi^{-1})^m$$

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \ldots + \frac{1}{m+1}\gamma_0 = \left| \begin{array}{ll} 2, & m = 0 \\ -1, & m = 1 \\ 0, & m = 2, 3, \ldots\ldots \end{array} \right. \tag{5.51}$$

We have the following method.

$k = 2$: $\gamma_0 = 2, \gamma_1 = -2, \gamma_2 = 1/3.$

$$y_{n+1} = y_{n-1} + \frac{h}{3}(y'_{n+1} + 4y'_n + y'_{n-1}),$$

$$T_{n+1} = -\frac{1}{90}h^5 y^{(5)}(x_n) + O(h^6), \ p = 4.$$

Numerical Differentiation Methods

$$\sigma(\xi) = \xi^k, \ \rho(\xi) \text{ of degree } k. \tag{5.52}$$

We have the following methods.

(*i*) $k = 1$: $y_{n+1} = y_n + hy'_{n+1},$

$$T_{n+1} = -\frac{h^2}{2}y''(x_n) + O(h^3), \ \ p = 1.$$

(*ii*) $k = 2$: $y_{n+1} = \dfrac{4}{3}y_n - \dfrac{1}{3}y_{n-1} + \dfrac{2}{3}hy'_{n+1},$

$$T_{n+1} = -\frac{2h^3}{9}y^{(3)}(x_n) + O(h^4), \ \ p = 2.$$

(*iii*) $k = 3$: $y_{n+1} = \dfrac{18}{11}y_n - \dfrac{9}{11}y_{n-1} + \dfrac{2}{11}y_{n-2} + \dfrac{6}{11}hy'_{n+1},$

$$T_{n+1} = -\frac{3}{22}h^4 y^{(4)}(x_n) + O(h^5), \ \ p = 3.$$

5.4 PREDICTOR CORRECTOR METHODS

When $\rho(\xi)$ and $\sigma(\xi)$ are of the same degree, we produce an implicit or a corrector method. If the degree of $\sigma(\xi)$ is less than the degree of $\rho(\xi)$, then we have an explicit or a predictor method. Corrector method produces a non-linear equation for the solution at x_{n+1}. However, the predictor method can be used to predict a value of $y_{n+1}^{(0)}$ and this value can be taken as the starting approximation of the iteration for obtaining y_{n+1} using the corrector method. Such methods are called the *Predictor-Corrector* methods.

Suppose that we use the implicit method

$$y_{n+1} = hb_0 f_{n+1} + \sum_{m=1}^{k} (a_m y_{n-m+1} + hb_m f_{n-m+1})$$

to find y_{n+1}.

This equation may be written as

$$y = F(y) \tag{5.53}$$

where
$$y = y_{n+1},$$
$$F(y) = hb_0 f(x_{n+1}, y) + c,$$
$$c = \sum_{m=1}^{k} (a_m y_{n-m+1} + hb_m f_{n-m+1}).$$

An iterative method can be used to solve (5.53) with suitable first approximation $y^{(0)}$. The general iterative procedure can be written as

$$y^{(s+1)} = F(y^{(s)}),\ s = 0, 1, 2, \ldots.. \tag{5.54}$$

which converges if

$$\left| h \frac{\partial f(x_{n+1}, y)}{\partial y} b_0 \right| < 1. \tag{5.55}$$

P(EC)^m E method

We use the explicit (predictor) method for predicting $y_{n+1}^{(0)}$ and then use the implicit (corrector) method iteratively until the convergence is obtained. We write (5.54) as

P : Predict some value $y_{n+1}^{(0)}$,

E : Evaluate $f(x_{n+1}, y_{n+1}^{(0)})$,

C : Correct $y_{n+1}^{(1)} = hb_0 f(x_{n+1}, y_{n+1}^{(0)}) + c$,

E : Evaluate $f(x_{n+1}, y_{n+1}^{(1)})$,

C : Correct $y_{n+1}^{(2)} = hb_0 f(x_{n+1}, y_{n+1}^{(1)}) + c$.

The sequence of operations

PECECE ...CE (5.56)

is denoted by $P(EC)^m E$ and is called a predictor-corrector method. Note that the predictor may be of the same order or of lower order than the corrector method.

If the predictor is of lower order, then the order of the method *PECE* is generally that of the predictor. Further application of the corrector raises the order of the combination by 1, until the order of the corrector is obtained. Further application may atmost reduce the magnitude of the error constant. Therefore, in practical applications, we may use only 2 or 3 corrector iterations.

***PM$_p$CM$_c$* Method**

For $m = 1$, the predictor-corrector method becomes *PECE*. If the predictor and the corrector methods are of the same order p then we can use the estimate of the truncation error to modify the predicted and the corrected values. Thus, we may write this procedure as $PM_p\, CM_c$. This is called the *modified predictor-corrector method.* We have

$$y(x_{n+1}) - y_{n+1}^{(p)} = C_{n+1}^{(p)}\, h^{p+1}\, y^{(p+1)}(x_n) + O(h^{p+2}),$$

$$y(x_{n+1}) - y_{n+1}^{(c)} = C_{p+1}^{(c)}\, h^{p+1}\, y^{(p+1)}(x_n) + O(h^{p+2}), \tag{5.57}$$

where $y_{n+1}^{(p)}$ and $y_{p+1}^{(c)}$ represent the solution values obtained by using the predictor and corrector methods respectively. Estimating $h^{p+1}\, y^{(p+1)}(x_n)$ from (5.57), we may obtain the modified predicted and corrected values m_{n+1} and y_{n+1} respectively and write the modified predictor-corrector method as follows :

Predicted value :

$$(i)\ p_{n+1} = \sum_{m=1}^{k} (a_m^{(0)} y_{n-m+1} + h b_m^{(0)} f_{n-m+1}).$$

Modified predicted value :

$$(ii)\ m_{n+1} = p_{n+1} + C_{p+1}^{(p)}\, (C_{p+1}^{(c)} - C_{p+1}^{(p)})^{-1}\, (p_n - c_n).$$

Corrected value :

$$(iii)\ c_{n+1} = hb_0 f(x_{n+1}, m_{n+1}) + \sum_{m=1}^{k} (a_m y_{n-m+1} + hb_m f_{n-m+1}).$$

Modified corrected value :

$$(iv)\ y_{n+1} = c_{n+1} + C_{p+1}^{(c)}\, (C_{p+1}^{(c)} - C_{p+1}^{(p)})^{-1} (p_{n+1} - c_{n+1}). \tag{5.58}$$

The quantity $(p_1 - c_1)$ in (5.58 (*ii*)) required for modification of the first step is generally put as zero.

5.5 STABILITY ANALYSIS

A numerical method is said to be *stable* if the cumulative effect of all the errors is bounded independent of the number of mesh points. We now examine the stability of the single step and multistep methods.

Single Step Methods

The application of the single step method (5.23) to the test problem (5.5) leads to a first order difference equation of the form

$$y_{n+1} = E(\lambda\, h)\, y_n \tag{5.59}$$

where $E(\lambda\, h)$ depends on the single-step method.

The analytical solution of the test problem (5.5) gives

$$y(x_{n+1}) = e^{\lambda\, h}\, y(x_n). \tag{5.60}$$

To find the error equation, we substitute $y_{n+1} = \varepsilon_{n+1} + y(x_{n+1})$ into (5.59) and use (5.60) to get

$$\begin{aligned}\varepsilon_{n+1} &= E(\lambda h)\, \varepsilon_n - [e^{\lambda h} - E(\lambda h)]\, y(x_n)\\ &= E(\lambda h)\, \varepsilon_n - T_{n+1}\end{aligned} \tag{5.61}$$

where T_{n+1} is the local truncation error and is independent of ε_n.

The first term on the right side of (5.61) represents the propagation of the error from the step x_n to x_{n+1} and will not grow if $|\, E(\lambda\, h)\, | \le 1$.

Definitions

Absolute Stability : If $| E(\lambda h) | \le 1$, $\lambda < 0$, then the single step method (5.23) is said to be absolutely stable.

Interval of Absolute Stability : If the method (5.23) is absolutely stable for all $\lambda h \in (\alpha, \beta)$ then the interval (α, β) on the real line is said to be the interval of absolute stability.

Relative Stability : If $E(\lambda h) \le e^{\lambda h}$, $\lambda > 0$, then the singlestep method (5.23) is said to be relatively stable.

Interval of Relative Stability : If the method (5.23) is relatively stable for all $\lambda h \in (\alpha, \beta)$, then the interval (α, β) on the real line is said to be the interval of relative stability.

Multistep Methods

Applying the method (5.39) to (5.5) and substituting $y_n = \varepsilon_n + y(x_n)$, we obtain the error equation

$$\varepsilon_{n+1} = \sum_{m=1}^{k} a_m \varepsilon_{n-m+1} + \lambda h \sum_{m=0}^{k} b_m \varepsilon_{n-m+1} - T_{n+1} \tag{5.62}$$

where T_{n+1} is the local truncation error and is independent of ε_n.

We assume that T_{n+1} is a constant and is equal to T. The characteristic equation of (5.62) is given by

$$\rho(\xi) - h\lambda\sigma(\xi) = 0. \tag{5.63}$$

The general solution of (5.62) may be written as

$$\varepsilon_n = A_1 \xi_{1h}^n + A_2 \xi_{2h}^n + \ldots + A_k \xi_{kh}^n + \frac{T}{h\lambda\sigma(1)} \tag{5.64}$$

where A_i are constants to be determined from the initial errors and $\xi_{1h}, \xi_{2h}, \ldots, \xi_{kh}$ are the distinct roots of the characteristic equation (5.63). For $h \to 0$, the roots of the characteristic equation (5.63) approach the roots of the equation

$$\rho(\xi) = 0. \tag{5.65}$$

The equation (5.65) is called the *reduced characteristic equation.* If $\xi_1, \xi_2, \ldots, \xi_k$ are the roots of $\rho(\xi) = 0$, then for sufficiently small $h\lambda$, we may write

$$\xi_{ih} = \xi_i (1 + h\lambda) \kappa_i + O(| \lambda h |^2),\ i = 1(1)^k \tag{5.66}$$

where κ_i are called the *growth parameters.* Substituting (5.66) into (5.63) and neglecting terms of $O(| \lambda h |^2)$, we obtain

$$\kappa_i = \frac{\sigma(\xi_i)}{\xi_i \rho'(\xi_i)},\quad i = 1(1)k. \tag{5.67}$$

From (5.66), we have

$$\xi_{ih}^n \approx \xi_i^n e^{\lambda h n k_i},\ i = 1(1)k. \tag{5.68}$$

For a consistent method we have, $\rho'(1) = \sigma(1)$. Hence, we get $\kappa_1 = 1$ and (5.66) becomes

$$\xi_{1h} = 1 + h\lambda + O(| \lambda h |^2). \tag{5.69}$$

Definitions

The multistep method (5.39) is said to be

stable if $| \xi_i | < 1$, $i \ne 1$.

unstable if $| \xi_i | > 1$ for some i or if there is a multiple root of $\rho(\xi) = 0$ of modulus unity.

weakly stable if ξ_i's are simple and if more than one of these roots have modulus unity.

absolutely stable if $| \xi_{ih} | \le 1, \lambda < 0, i = 1(1)\, k$.

relatively stable if $| \xi_{ih} | \le | \xi_{1h} |, \lambda > 0, i = 2(1)k$.

A-stable if the interval of absolute stability is $(-\infty, 0)$.

To obtain the interval of absolute stability of a multistep method, we often use the Routh-Hurwitz criterion.

5.6. SYSTEM OF DIFFERENTIAL EQUATIONS

The system of m equations in vector form may be written as

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}\,(x, y_1, y_2, ..., y_m),$$
$$\mathbf{y}(0) = \boldsymbol{\eta}, \tag{5.70}$$

where $\mathbf{y} = [y_1\ y_2\ ...\ y_m]^T$, $\boldsymbol{\eta} = [\eta_1\ \eta_2\ ...\ \eta_m]^T$,

$$\mathbf{f} = \begin{bmatrix} f_1\,(x, y_1, y_2, ... y_m) \\ f_2\,(x, y_1, y_2, ... y_m) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ f_m\,(x, y_1, y_2, ... y_m) \end{bmatrix}.$$

The singlestep and multistep methods developed in Sections 5.2 and 5.3 can be directly written for the system (5.70).

Taylor Series Method

We write (5.26) as

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}'_n + \frac{h^2}{2!}\,\mathbf{y}''_n + ... + \frac{h^p}{p!}\,\mathbf{y}_n^{(p)},$$
$$n = 0, 1, 2, ...\, N - 1, \tag{5.71}$$

where $\mathbf{y}_n^{(k)} = [y_{1,n}^{(k)}\quad y_{2,n}^{(k)}\quad ...\quad y_{m,n}^{(k)}]^T.$

Second order Runge-Kutta Method

The second order Runge-Kutta method (5.30) becomes

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}\,(\mathbf{K}_1 + \mathbf{K}_2),\ n = 0, 1, 2, ..., N - 1,$$

where $\mathbf{K}_j = [K_{1j}\ K_{2j}\ .....\ K_{mj}]^T,\ j = 1, 2.$

and $K_{i1} = hf_i\,(x_n, y_{1,n}, y_{2,n}\ ..., y_{m,n}),$

$K_{i2} = hf_i(x_n + h, y_{1,n} + K_{11}, y_{2,n} + K_{21}, ..., y_{m,n} + K_{m1}).$

$i = 1, 2, ..., m.$

Fourth Order Runge-Kutta Method

The fourth order Runge-Kutta method (5.35) becomes

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{6}\,(\mathbf{K}_1 + 2\mathbf{K}_2 + 2\mathbf{K}_3 + \mathbf{K}_4),$$
$$n = 0, 1, 2, ..., N - 1, \tag{5.72}$$

where $\mathbf{K}_j = [K_{1j}\ K_{2j}\ ...\ K_{mj}]^T,\ j = 1, 2, 3, 4,$

and
$$K_{i1} = hf_i\,(x_n, y_{1,n}\ y_{2,n}, \ldots, y_{m,n})$$
$$K_{i2} = hf_i\left(x_n + \frac{h}{2}, y_{1,n} + \frac{1}{2}K_{11}, y_{2,n} + \frac{1}{2}K_{21}, \ldots, y_{m,n} + \frac{1}{2}K_{m1}\right)$$
$$K_{i3} = hf_i\left(x_n + \frac{h}{2}, y_{1,n} + \frac{1}{2}K_{12}, y_{2,n} + \frac{1}{2}K_{22}, \ldots, y_{m,n} + \frac{1}{2}K_{m2}\right),$$
$$K_{i4} = h\,f_i\,(x_n + h, y_{1,n} + K_{13}, y_{2,n} + K_{23}, \ldots, y_{m,n} + K_{m3})$$
$$i = 1, 2, \ldots, m.$$

Stability Analysis

The stability of the numerical methods for the system of first order differential equations is discussed by applying the numerical methods to the homogeneous locally linearized form of the equation (5.70). Assuming that the functions f_i have continuous partial derivatives $(\partial f_i / \partial y_k) = a_{ik}$ and $\mathbf{A}$ denotes the $m \times m$ matrix (a_{ik}), we may, to terms of first order, write (5.70) as

$$\frac{d\mathbf{y}}{dx} = \mathbf{A}\mathbf{y} \tag{5.73}$$

where $\mathbf{A}$ is assumed to be a constant matrix with distinct eigenvalues λ_i, $i = 1(1)n$. The analytic solution $\mathbf{y}(x)$ of (5.73) satisfying the initial conditions $\mathbf{y}(0) = \boldsymbol{\eta}$ is given by

$$\mathbf{y}(x) = \exp\,(\mathbf{A}x)\,\boldsymbol{\eta} \tag{5.74}$$

where $\exp\,(\mathbf{A}x)$ is defined as the matrix function

$$\exp\,(\mathbf{A}x) = \mathbf{I} + \mathbf{A}x + \frac{(\mathbf{A}x)^2}{2!} + \ldots \tag{5.75}$$

and $\mathbf{I}$ is a unit matrix.

The transformation $\mathbf{y} = \mathbf{PZ}$, where $\mathbf{P}$ is the $m \times m$ non-singular matrix formed by the eigenvectors corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_m$, *i.e.*,

$$\mathbf{P} = [\mathbf{y}_1\ \mathbf{y}_2 \ldots \mathbf{y}_m],$$

transforms (5.73) into a decoupled system of equations

$$\frac{d\mathbf{Z}}{dx} = \mathbf{DZ} \tag{5.76}$$

where
$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_m \end{bmatrix}.$$

Application of the Taylor series method (5.71) to (5.76) leads to an equation of the form

$$\mathbf{v}_{n+1} = \mathbf{E}(\mathbf{D}h)\,\mathbf{v}_n \tag{5.77}$$

where $\mathbf{E}(\mathbf{D}h)$ represents an approximation to $\exp\,(\mathbf{D}h)$. The matrix $\mathbf{E}(\mathbf{D}h)$ is a diagonal matrix and each of its diagonal element $E_s(\lambda_s h)$, $s = 1(1)m$ is an approximation to the diagonal element $\exp\,(\lambda_s h)$, $s = 1(1)m$ respectively, of the matrix $\exp\,(\mathbf{D}h)$. We therefore, have the important result that the stability analysis of the Taylor series method (5.71) as applied to the differential system (5.73) can be discussed by applying the Taylor series method (5.71) to the scalar equation

$$y' = \lambda_s y \tag{5.78}$$

where λ_s, $s = 1(1)m$ are the eigenvalues of $\mathbf{A}$. Thus, the Taylor series method (5.71) is absolutely stable if $|E_s\,(\lambda_s h)| < 1$, $s = 1(1)m$, where $\mathrm{Re}(\lambda_s) < 0$ and Re is the real part of λ_s.

Multistep Methods

The multistep method (5.39) for (5.70) may be written in the form

$$\mathbf{y}_{n+1} = \sum_{m=1}^{k} a_m \mathbf{y}_{n-m+1} + h \sum_{m=0}^{k} b_m \mathbf{f}_{n-m+1} \tag{5.79}$$

where a_m's and b_m's have the same values as in the case of the method (5.39).

The stability analysis can be discussed by applying the method (5.79) to (5.73) or (5.78).

Boundary Value Problems

5.7 SHOOTING METHODS

Consider the numerical solution of the differential equation (5.7)

$$-y'' + p(x)y' + q(x)y = r(x), \quad a < x < b$$

subject to the boundary conditions (5.10)

$$\begin{aligned} a_0 y(a) - a_1 y'(a) &= \gamma_1, \\ b_0 y(b) + b_1 y'(b) &= \gamma_2, \end{aligned} \tag{5.80}$$

where a_0, a_1, b_0, b_1, γ_1 and γ_2 are constants such that

$$a_0 a_1 \geq 0, \; |\, a_0 \,| + |\, a_1 \,| \neq 0,$$
$$b_0 b_1 \geq 0, \; |\, b_0 \,| + |\, b_1 \,| \neq 0, \quad \text{and} \quad |\, a_0 \,| + |\, b_0 \,| \neq 0.$$

The boundary value problem (5.7) subject to the boundary conditions (5.80) will have a unique solution if the functions $p(x)$, $q(x)$ and $r(x)$ are continuous on $[a, b]$ and $q(x) > 0$.

To solve the differential equation (5.7) subject to (5.80) numerically, we first define the function $y(x)$ as

$$y(x) = \phi_0(x) + \mu_1 \phi_1(x) + \mu_2 \phi_2(x) \tag{5.81}$$

where μ_1 and μ_2 are arbitrary constants and ϕ's are the solutions on $[a, b]$ of the following IVPs :

$$\begin{aligned} -\phi_0'' + p(x)\phi_0' + q(x)\phi_0 &= r(x), \\ \phi_0(a) = 0, \; \phi_0'(a) &= 0. \end{aligned} \tag{5.82}$$

$$\begin{aligned} -\phi_1'' + p(x)\phi_1' + q(x)\phi_1 &= 0, \\ \phi_1(a) = 1, \; \phi_1'(a) &= 0. \end{aligned} \tag{5.83}$$

$$\begin{aligned} -\phi_2'' + p(x)\phi_2' + q(x)\phi_2 &= 0, \\ \phi_2(a) = 0, \; \phi_2'(a) &= 1. \end{aligned} \tag{5.84}$$

The first condition in (5.80) will be satisfied by (5.81) if

$$a_0 \mu_1 - a_1 \mu_2 = \gamma_1. \tag{5.85}$$

The *shooting method* requires the solution of the three initial value problems (5.82), (5.83) and (5.84).

Denoting $\phi_i(x) = w^{(i+1)}(x)$ and $\phi_i'(x) = v^{(i+1)}(x)$, $i = 0, 1, 2$, the IVPs (5.82)-(5.84) can be written as the following equivalent first order systems.

$$\begin{bmatrix} w^{(1)} \\ v^{(1)} \end{bmatrix}' = \begin{bmatrix} v^{(1)} \\ pv^{(1)} + qw^{(1)} - r \end{bmatrix}, \quad \begin{bmatrix} w^{(1)}(0) \\ v^{(1)}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{5.86}$$

$$\begin{bmatrix} w^{(2)} \\ v^{(2)} \end{bmatrix}' = \begin{bmatrix} v^{(2)} \\ pv^{(2)} + qw^{(2)} \end{bmatrix}, \quad \begin{bmatrix} w^{(2)}(0) \\ v^{(2)}(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{5.87}$$

$$\begin{bmatrix} w^{(3)} \\ v^{(3)} \end{bmatrix}' = \begin{bmatrix} v^{(3)} \\ pv^{(3)} + qw^{(3)} \end{bmatrix}, \quad \begin{bmatrix} w^{(3)}(0) \\ v^{(3)}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{5.88}$$

respectively.

Now, any of the numerical methods discussed in Sections 5.2 and 5.3 can be applied to solve (5.86), (5.87) and (5.88). We denote the numerical solutions of (5.86), (5.87) and (5.88) by

$$w_i^{(1)},\ v_i^{(1)}\ ;\ v_i^{(2)},\ w_i^{(2)}\ ;\ w_i^{(3)},\ v_i^{(3)}\ ;\ i = 0, 1, ..., N \tag{5.89}$$

respectively.

The solution (5.81) at $x = b$ gives

$$y^{(b)} = w_N^{(1)} + \mu_1 w_N^{(2)} + \mu_2\, w_N^{(3)}, \tag{5.90}$$

$$y'(b) = v_N^{(1)} + \mu_1 v_N^{(2)} + \mu_2\, v_N^{(3)}. \tag{5.91}$$

Substituting (5.90) and (5.91) into the second condition in (5.80) we obtain

$$(b_0 w_N^{(2)} + b_1 v_N^{(2)})\, \mu_1 + (b_0 w_N^{(3)} + b_1 v_N^{(3)})\, \mu_2 = \gamma_2 - (b_0 w_N^{(1)} + b_1 v_N^{(1)}). \tag{5.92}$$

We can determine μ_1 and μ_2 from (5.85) and (5.92).

Thus, the numerical solution of the boundary value problem is given by

$$y(x_i) = w_i^{(1)} + \mu_1 w_i^{(2)} + \mu_2\, w_i^{(3)},\ \ i = 1(1)N - 1 \tag{5.93}$$

Alternative

When the boundary value problem in nonhomogeneous, then it is sufficient to solve the two initial value problems

$$-\ \phi_1'' + p(x)\,\phi_1' + q(x)\,\phi_1 = r(x), \qquad \text{...[5.94 } (i)]$$

$$-\ \phi_2'' + p(x)\,\phi_2' + q(x)\,\phi_2 = r(x) \qquad \text{...[5.94 } (ii)]$$

with suitable initial conditions at $x = a$.

We write the general solution of the boundary value problem in the form

$$y(x) = \lambda\phi_1(x) + (1 - \lambda)\,\phi_2(x) \tag{5.95}$$

and determine λ so that the boundary condition at the other end, that is, at $x = b$ is satisfied.

We solve the inital value problems [5.94 (i)],[5.94 (ii)] upto $x = b$ using the initial conditions.

(i) *Boundary conditions of the first kind* :

$$\phi_1(a) = \gamma_1,\ \phi_1'(a) = 0,$$
$$\phi_2(a) = \gamma_1,\ \phi_2'(a) = 1.$$

From (5.95), we obtain

$$y(b) = \gamma_2 = \lambda\,\phi_1(b) + (1 - \lambda)\,\phi_2(b),$$

which gives
$$\lambda = \frac{\gamma_2 - \phi_2(b)}{\phi_1(b) - \phi_2(b)},\ \ \phi_1(b) \neq \phi_2(b). \tag{5.96}$$

(ii) *Boundary conditions of the second kind* :

$$\phi_1(a) = 0,\ \phi_1'(a) = \gamma_1$$
$$\phi_2(a) = 1,\ \phi_2'(a) = \gamma_1.$$

From (5.95), we obtain

$$y'(b) = \gamma_2 = \lambda\, \phi'_1(b) + (1 - \lambda)\, \phi'_2(b).$$

which gives
$$\lambda = \frac{\gamma_2 - \phi'_2(b)}{\phi'_1(b) - \phi'_2(b)}, \quad \phi'_1(b) \neq \phi'_2(b). \tag{5.97}$$

(*iii*) *Boundary conditions third kind* :

$$\phi_1(a) = 0,\ \phi'_1(a) = -\gamma_1 / a_1,$$
$$\phi_2(a) = 1,\ \phi'_2(a) = (a_0 - \gamma_1) / a_1.$$

From (5.95), we obtain

$$y(b) = \lambda\phi_1(b) + (1 - \lambda)\, \phi'_1(b),$$
$$y'(b) = \lambda\, \phi'_2(b) + (1 - \lambda)\, \phi'_2(b).$$

Substituting in the second condition, $b_0\, y(b) + b_1 y'(b) = \gamma_2$, in (5.10), we get

$$\gamma_2 = b_0[\lambda\, \phi_1(b) + (1 - \lambda)\, \phi_2(b)] + b_1\, [\lambda\, \phi'_1(b) + (1 - \lambda)\, \phi'_2(b)]$$

which gives

$$\lambda = \frac{\gamma_2 - [b_0\, \phi_2(b) + b_1\phi'_2(b)]}{[b_0\phi_1(b) + b_1\phi'_1(b)] - [b_0\phi_2(b) + b_1\phi'_2(b)]} \tag{5.98}$$

The results obtained are identical in both the approaches.

Nonlinear Second Order Differential Equations

We now consider the nonlinear differential equation

$$y'' = f(x, y, y'), \quad a < x < b$$

subject to one of the boundary conditions (5.8) to (5.10). Since the differential equation is non linear, we cannot write the solution in the form (5.81) or (5.95).

Depending on the boundary conditions, we proceed as follows :

Boundary condition of the first kind : We have the boundary conditions as

$$y(a) = \gamma_1 \quad \text{and} \quad y(b) = \gamma_2.$$

We assume $y'(a) = s$ and solve the initial value problem

$$y'' = f(x, y, y'),$$
$$y(a) = \gamma_1,\ y'(a) = s \tag{5.99}$$

upto $x = b$ using any numerical method. The solution, $y(b, s)$ of the initial value problem (5.99) should satisfy the boundary condition at $x = b$. Let

$$\phi(s) = y(b, s) - \gamma_2. \tag{5.100}$$

Hence, the problem is to find s, such that $\phi(s) = 0$.

Boundary condition of the second kind : We have the boundary conditions as

$$y'(a) = \gamma_1 \quad \text{and} \quad y'(b) = \gamma_2.$$

We assume $y(a) = s$ and solve the initial value problem

$$y'' = f(x, y, y'),$$
$$y(a) = s,\ y'(a) = \gamma_1, \tag{5.101}$$

upto $x = b$ using any numerical method. The solution $y(b, s)$ of the initial value problem (5.101) should satisfy the boundary condition at $x = b$. Let

$$\phi(s) = y'(b, s) - \gamma_2. \tag{5.102}$$

Hence, the problem is to find s, such that $\phi(s) = 0$.

Boundary condition of the third kind : We have the boundary conditions as

$$a_0 y(a) - a_1 y'(a) = \gamma_1$$
$$b_0 y(b) + b_1 y'(b) = \gamma_2.$$

Here, we can assume the value of $y(a)$ or $y'(a)$.

Let $y'(a) = s$. Then, from

$$a_0 y(a) - a_1 y'(a) = \gamma_1, \text{ we get } y(a) = (a_1 s + \gamma_1) / a_0.$$

We now solve the initial value problem

$$y'' = f(x, y, y'),$$
$$y(a) = (a_1 s + \gamma_1) / a_0, \; y'(a) = s, \tag{5.103}$$

upto $x = b$ using any numerical method. The solution $y(b, s)$ of the initial value problem (5.103) should satisfy the boundary condition at $x = b$. Let

$$\phi(s) = b_0 y(b, s) + b_1 y'(b, s) - \gamma_2. \tag{5.104}$$

Hence, the problem is to find s, such that $\phi(s) = 0$.

The function $\phi(s)$ in (5.100) or (5.102) or (5.104) is a nonlinear function in s. We solve the equation

$$\phi(s) = 0 \tag{5.105}$$

by using any iterative method discussed in Chapter 1.

Secant Method

The iteration method for solving $\phi(s) = 0$ is given by

$$s^{(k+1)} = s^{(k)} - \left[\frac{s^{(k)} - s^{(k-1)}}{\phi(s^{(k)}) - \phi(s^{(k-1)})}\right] \phi(s^{(k)}), \; k = 1, 2, \tag{5.106}$$

which $s^{(0)}$ and $s^{(1)}$ are two initial approximations to s. We solve the initial value problem (5.99) or (5.101) or (5.103) with two guess values of s and keep iterating till

$$|\, \phi(s^{(k+1)}) \,| < \text{(given error tolerance)}.$$

Newton-Raphson Method

The iteration method for solving $\phi(s) = 0$ is given by

$$s^{(k+1)} = s^{(k)} - \frac{\phi(s^{(k)})}{\phi'(s^{(k)})}, \; k = 0, 1, \ldots.. \tag{5.107}$$

where $s^{(0)}$ is some initial approximation to s.

To determine $\phi'(s^{(k)})$, we proceed as follows :

Denote $y_s = y(x, s)$, $y_s' = y'(x, s)$, $y_s'' = y''(x, s)$.

Then, we can write (5.103) as

$$y_s'' = f(x, y_s, y'_s), \qquad \ldots[5.108\,(i)]$$
$$y_s(a) = (a_1 s + \gamma_1) / a_0, \; y_s'(a) = s. \qquad \ldots[(5.108\,(ii)]$$

Differentiating [5.108 (i)] partially with respect to s, we get

$$\frac{\partial}{\partial s}(y_s'') = \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y_s}\frac{\partial y_s}{\partial s} + \frac{\partial f}{\partial y_s'}\frac{\partial y_s'}{\partial s}$$
$$= \frac{\partial f}{\partial y_s}\frac{\partial y_s}{\partial s} + \frac{\partial f}{\partial y_s'}\frac{\partial y_s'}{\partial s} \tag{5.109}$$

since x is independent of s.

Differentiating [5.108 (*ii*)] partially with respect to s, we get

$$\frac{\partial}{\partial s}[y_s(a)] = \frac{a_1}{a_0}, \frac{\partial}{\partial s}[y'_s(a)] = 1.$$

Let $v = \frac{\partial y_s}{\partial s}$. Then,we have

$$v' = \frac{\partial v}{\partial x} = \frac{\partial}{\partial x}\left[\frac{\partial y_s}{\partial s}\right] = \frac{\partial}{\partial s}\left[\frac{\partial y_s}{\partial x}\right] = \frac{\partial}{\partial s}(y'_s)$$

$$v'' = \frac{\partial v'}{\partial x} = \frac{\partial}{\partial x}\left[\frac{\partial}{\partial s}\left(\frac{\partial y_s}{\partial x}\right)\right] = \frac{\partial}{\partial s}\left[\frac{\partial^2 y_s}{\partial x^2}\right] = \frac{\partial}{\partial s}(y''_s).$$

From (5.108) and (5.109), we obtain

$$v'' = \frac{\partial f}{\partial y_s}(x, y_s, y'_s)\, v + \frac{\partial f}{\partial y'_s}(x, y_s, y'_s)\, v' \qquad ...[5.111\ (i)]$$

$$v(a) = a_1 / a_0, v'(a) = 1. \qquad ...[5.111\ (ii)]$$

The differential equation [5.111 (*i*)] is called the first variational equation. It can be solved step by step along with (5.108), that is, (5.108) and (5.111) can be solved to gether as a single system. When the computation of one cycle is completed, $v(b)$ and $v'(b)$ are available.

Now, from (5.104), at $x = b$, we have

$$\frac{d\phi}{ds} = b_0 \frac{\partial y_s}{\partial s} + b_1 \frac{\partial y'_s}{\partial s} = b_0\, v(b) + b_1 v'(b). \qquad (5.112)$$

Thus, we have the value of $\phi'(s^{(k)})$ to be used in (5.107).

If the boundary conditions of the first kind are given, then we have

$$a_0 = 1, a_1 = 0, b_0 = 1, b_1 = 0 \quad \text{and} \quad \phi(s) = y_s(b) = \gamma_2.$$

The initial conditions (5.110), on v become

$$v(a) = 0, v'(a) = 1.$$

Then, we have from (5.112)

$$\frac{d\phi}{ds} = v(b). \qquad (5.113)$$

5.8 FINITE DIFFERENCE METHODS

Let the interval $[a, b]$ be divided into $N + 1$ subintervals, such that

$$x_j = a + jh, j = 0, 1, ..., N + 1,$$

where $x_0 = a$, $x_{N+1} = b$ and $h = (b - a) / (N + 1)$.

Linear Second Order Differential Equations

We consider the linear second order differential equation

$$-y'' + p(x)y' + q(x)y = r(x) \qquad (5.114)$$

subject to the boundary conditions of the first kind

$$y(a) = \gamma_1, y(b) = \gamma_2. \qquad (5.115)$$

Using the second order finite difference approximations

$$y'(x_j) \approx \frac{1}{2h}[y_{j+1} - y_{j-1}],$$

$$y''(x_j) \approx \frac{1}{h^2}[y_{j+1} - 2y_j + y_{j-1}],$$

at $x = x_j$, we obtain the difference equation

$$-\frac{1}{h^2}(y_{j+1}-2y_j+y_{j-1})+\frac{1}{2h}(y_{j+1}-y_{j-1})\,p(x_j)+q(x_j)\,y_j=r(x_j), \tag{5.116}$$

$j = 1, 2, ..., N.$

The boundary conditions (5.115) become

$$y_0 = \gamma_1,\ y_{N+1} = \gamma_2.$$

Multiplying (5.116) by $h^2 / 2$, we obtain

$$A_j y_{j-1} + B_j y_j + C_j y_{j+1} = \frac{h^2}{2}\, r(x_j), \quad j = 1, 2, ..., N \tag{5.117}$$

where $A_j = -\frac{1}{2}\left(1+\frac{h}{2}\,p(x_j)\right)$, $B_j = \left(1+\frac{h^2}{2}\,q(x_j)\right)$, $C_j = -\frac{1}{2}\left(1-\frac{h}{2}\,p(x_j)\right)$.

The system (5.117) in matrix notation, after incorporating the boundary conditions, becomes

$$\mathbf{A\,y = b}$$

where
$$\mathbf{y} = [y_1, y_2, ..., y_N]^T$$

$$\mathbf{b} = \frac{h^2}{2}\left[r(x_1)-\frac{2A_1\gamma_1}{h^2}, r(x_2), ..., r(x_{N-1}), r(x_n)-\frac{2C_n\gamma_2}{h^2}\right]^T,$$

$$\mathbf{A} = \begin{bmatrix} B_1 & C_1 & & & \mathbf{0} \\ A_2 & B_2 & C_2 & & \\ \cdots & \cdots & .. & & .. \\ & & A_{N-1} & B_{N-1} & C_{N-1} \\ \mathbf{0} & & & A_N & B_N \end{bmatrix}$$

The solution of this system of linear equations gives the finite difference solution of the differential equation (5.114) satisfying the boundary conditions (5.115).

Local Truncation Error

The local truncation error of (5.117) is defined by

$$T_j = A_j y(x_{j-1}) + B_j y(x_j) + C_j y(x_{j+1}) - \frac{h^2}{2}\, r(x_j)$$

Expanding each term on the right hand side Taylor's series about x_j, we get

$$T_j = -\frac{h^4}{24}[y^{(4)}(\xi_1) - 2p(x_j)\, y^{(3)}(\xi_2)], \quad j = 1, 2, ..., N$$

where $x_{j-1} < \xi_1 < x_{j+1}$ and $x_{j-1} < \xi_2 < x_{j+1}$.

The largest value of p for which the relation

$$T_j = 0\ (h^{p+2})$$

holds is called the order of the difference method. (5.118)

Therefore, the method (5.116) is of second order.

Derivative Boundary Conditons

We now consider the boundary conditions

$$a_0 y(a) - a_1 y'(a) = \gamma_1,$$
$$b_0 y(b) + b_1 y'(b) = \gamma_2. \tag{5.119}$$

The difference equation (5.116) at the internal nodes, $j = 1, 2, ..., N$ gives N equations in $N + 2$ unknowns. We obtain two more equations using the boundary conditions (5.119). We obtain the second order approximations for the boundary conditons as follows.

(*i*) *At* $x = x_0$: $a_0 y_0 - \frac{a_1}{2h}[y_1 - y_{-1}] = \gamma_1$,

or $$y_{-1} = -\frac{2ha_0}{a_1} y_0 + y_1 + \frac{2h}{a_1}\gamma_1. \quad (5.120)$$

At $x = x_{N+1}$: $b_0 y_{N+1} + \frac{b_1}{2h}[y_{N+2} - y_N] = \gamma_2$,

or $$y_{N+2} = y_N - \frac{2hb_0}{b_1} y_{N+1} + \frac{2h}{b_1}\gamma_2. \quad (5.121)$$

The values y_{-1} and y_{N+2} can be eliminated by assuming that the difference equation (5.116) holds also for $j = 0$ and $N + 1$, that is, at the boundary points x_0 and x_{N+1}. Substituting the values of y_{-1} and y_{N+1} from (5.120) and (5.121) into the equations (5.116) for $j = 0$ and $j = N + 1$, we obtain two more equations.

(*ii*) *At* $x = x_0$: $a_0 y_0 - \frac{a_1}{2h}(-3y_0 + 4y_1 - y_2) = \gamma_1$,

or $$(2ha_0 + 3a_1) y_0 - 4a_1 y_1 + a_1 y_2 = 2h\gamma_1. \quad (5.122)$$

At $x = x_{N+1}$: $b_0 y_{N+1} + \frac{b_1}{2h}(3y_{N+1} - 4y_N + y_{N-1}) = \gamma_2$

or $$b_1 y_{N-1} - 4b_1 y_N + (2hb_0 + 3b_1) y_{N+1} = 2h\,\gamma_2. \quad (5.123)$$

Fourth Order Method when y' is Absent in (5.114)

Consider the differential equation
$$-y'' + q(x)y = r(x),\ a < x < b \quad (5.124)$$
subject to the boundary conditions of the first kind
$$y(a) = \gamma_1,\ y(b) = \gamma_2. \quad (5.125)$$
We write the differential equation as
$$y'' = q(x)y - r(x) = f(x, y). \quad (5.126)$$
A fourth order difference approximation for (5.126) is obtained as
$$y_{j-1} - 2y_j + y_{j+1} = \frac{h^2}{12}(y''_{j-1} + 10y''_j + y''_{j+1}),\ j = 1, 2, ..., N, \quad (5.127)$$
which is also called the *Numeröv method*.

We can also write the method as
$$\left[1 - \frac{h^2}{12} q_{j-1}\right] y_{j-1} - \left[2 + \frac{5h^2}{6} q_j\right] y_j + \left[1 - \frac{h^2}{12} q_{j+1}\right] y_{j+1} = -\frac{h^2}{12}[r_{j-1} + 10r_j + r_{j+1}],$$
(5128)

where $r_i = r(x_i),\ q_i = q(x_i),\ i = j - 1, j, j + 1$.

The truncation error associated with (5.127) is given by
$$T_j = -\frac{h^6}{240} y^{(6)}(\xi),\ x_{j-1} < \xi < x_{j+1}.$$

Derivative boundary conditions (5.119)

Fourth order approximations to the boundary conditions

$$a_0 y(a) - a_1 y'(a) = \gamma_1, \qquad \text{...[5.129(}i\text{)]}$$

$$b_0 y(b) + b_1 y'(b) = \gamma_2 \qquad \text{...[5.129(}ii\text{)]}$$

are given as follows.

$$\text{At } x = x_0: \quad y_1 = y_0 + h y_0' + \frac{h^2}{6}[y_0'' + 2y_{1/2}''] \tag{5.130}$$

where $$y_{1/2} = y_0 + \frac{h}{2} y_0' + \frac{h^2}{8} y_0''.$$

Solving (5.130) for y_0', we get

$$y_0' = \frac{1}{h}(y_1 - y_0) - \frac{h}{6}(y_0'' + 2y_{1/2}'').$$

Substituting in [5.129(i)], we get an $O(h^4)$ approximation valid at $x = a$ as

$$\frac{1}{h}(y_1 - y_0) - \frac{h}{6}[y_0'' + 2y_{1/2}''] = \frac{1}{a_1}(a_0 y_0 - \gamma_1). \tag{5.131}$$

$$\text{At } x = x_N: \quad y_N = y_{N+1} - h y_{N+1}' + \frac{h^2}{6}[2y_{N+1/2}'' + y_{N+1}''] \tag{5.132}$$

where $$y_{N+1/2} = y_N + \frac{h}{2} y_N' + \frac{h^2}{8} y_N''.$$

Solving (5.132) for y_{N+1}', we obtain

$$y_{N+1}' = \frac{1}{h}(y_{N+1} - y_N) + \frac{h}{6}[2y_{N+1/2}'' + y_{N+1}''].$$

Substituting in [5.129(ii)], we get an $O(h^4)$ aproximation valid at $x = b$ as

$$\frac{1}{h}(y_{N+1} - y_N) + \frac{h}{6}(2y_{N+1/2}'' + y_{N+1}'') = \frac{1}{b_1}(\gamma_2 - b_0 y_{N+1}). \tag{5.133}$$

Nonlinear Second Order Differential Equation y″ = f(x, y)

We consider the nonlinear second order differential equation

$$y'' = f(x, y) \tag{5.134}$$

subject to the boundary conditions (5.119).

Substituting

$$y_j'' = \frac{1}{h^2}(y_{j+1} - 2y_j + y_{j-1})$$

in (5.134), we obtain

$$y_{j+1} - 2y_j + y_{j-1} = h^2 f(x_j, y_j), \; j = 1, 2, ..., N \tag{5.135}$$

with the truncation error

$$\text{TE} = \frac{h^4}{12} y^{iv}(\xi), \; x_{j-1} < \xi < x_{j+1}.$$

The system of equations (5.135) contains N equations in $N + 2$ unknowns. Two more equations are obtained by using suitable difference approximations to the boundary conditions.

The difference approximations (5.120), (5.121) or (5.122), (5.123) at $x = a$ and $x = b$ can be used to obtain the two required equations. The totality of equations (5.135), (5.120), (5.121) or (5.135), (5.122), (5.123) are of second arder.

We write the Numeröv method for the solution of (5.134) as

$$y_{j+1} - 2y_j + y_{j-1} = \frac{h^2}{12} [f_{j+1} + 10 f_j + f_{j-1}] \tag{5.136}$$

which is of fourth order, that is, TE = $O(h^6)$.

Suitable approximations to the boundary conditions can be written as follows :

$$\text{At} \quad x = x_0 : \qquad hy_0' = y_1 - y_0 - \frac{h^2}{6} [2f(x_0, y_0) + f(x_1, y_1)] \tag{5.137}$$

$$\text{At} \quad x = x_{N+1} : \quad hy_{n+1}' = y_{N+1} - y_N + \frac{h^2}{6} [f(x_N, y_N) + 2f(x_{N+1}, y_{N+1})] \tag{5.138}$$

The truncation error in (5.137) and (5.138) is $O(h^4)$.

Substituting in (5.119), we obtain the difference equations corresponding to the boundary conditions at $x = x_0$ and $x = x_{N+1}$ respectively as

$$(ha_0 + a_1) y_0 - a_1 y_1 + \frac{a_1 h^2}{6} (2f_0 + f_1) = h\gamma_1 \tag{5.139}$$

$$hb_0 y_{N+1} + b_1(y_{N+1} - y_N) + \frac{b_1 h^2}{6} (f_N + 2f_{N+1}) = h\gamma_2. \tag{5.140}$$

Alternately, we can use the $O(h^4)$ difference approximations (5.131) and (5.133) at the boundary points $x = a$ and $x = b$.

The difference approximations discussed above produce a system of $(N + 2)$ nonlinear equations in $(N + 2)$ unknowns.

This system of nonlinear equations can be solved by using any iteration method.

5.9 PROBLEMS AND SOLUTIONS

Difference Equations

5.1 Solve the difference equation

$$y_{n+1} - 2 \sin x \, y_n + y_{n-1} = 0$$

when $y_0 = 0$ and $y_1 = \cos x$. (Lund Univ., Sweden, BIT 9(1969), 294)

Solution

Substituting $y_n = A\, \xi^n$, we obtain the characteristic equation as

$$\xi^2 - 2(\sin x)\xi + 1 = 0$$

whose roots are $\xi_1 = -ie^{ix} = \sin x - i \cos x$, $\xi_2 = ie^{-ix} = \sin x + i \cos x$.

The general solution is given by

$$y_n = C_1 i^n e^{-inx} + C_2 (-1)^n i^n e^{inx}.$$

The initial conditions give

$$C_1 + C_2 = 0,$$

$$C_1 ie^{-ix} - C_2 ie^{ix} = \cos x,$$

or $\quad i(C_1 - C_2) \cos x - (C_1 + C_2)\, i \sin x] = \cos x.$

We get $C_1 - C_2 = 1 / i$ and $C_1 + C_2 = 0$, whose solution is

$$C_1 = 1/(2i),\ C_2 = -1/(2i).$$

The general solution becomes

$$y_n = \frac{(i)^{n+1}}{2}\,[(-1)^n e^{inx} - e^{-inx}].$$

5.2 Find y_n from the difference equation

$$\Delta^2 y_{n+1} + \frac{1}{2}\Delta^2 y_n = 0,\ n = 0, 1, 2, \ldots$$

when $\quad y_0 = 0,\ y_1 = 1/2,\ y_2 = 1/4.$

Is this method numerically stable ? (Gothenburg Univ., Sweden, BIT 7(1967), 81)

Solution

The difference equation may be written in the form

$$y_{n+3} - \frac{3}{2} y_{n+2} + \frac{1}{2} y_n = 0.$$

The characteristic polynomial

$$\xi^3 - \frac{3}{2}\xi^2 + \frac{1}{2} = 0$$

has the roots 1, 1, $-1/2$. The general solution becomes

$$y_n = C_1 + C_2 n + C_3 \left(-\frac{1}{2}\right)^n.$$

The initial conditions lead to the following equations

$$C_1 + C_3 = 0,$$

$$C_1 + C_2 - \frac{1}{2}C_3 = \frac{1}{2},$$

$$C_1 + 2C_2 + \frac{1}{4}C_3 = \frac{1}{4},$$

which give $C_1 = 1/3,\ C_2 = 0,\ C_3 = -1/3.$

Hence, the solution is

$$y_n = \frac{1}{3}\left[1 + (-1)^{n+1}\frac{1}{2^n}\right].$$

The characteristic equation does not satisfy the root condition (since 1 is a double root) and hence the difference equation is unstable.

5.3 Show that all solutions of the difference equation

$$y_{n+1} - 2\lambda y_n + y_{n-1} = 0$$

are bounded, when $n \to \infty$ if $-1 < \lambda < 1$, while for all complex values of λ there is atleast one unbounded solution. (Stockholm Univ., Sweden, BIT 4(1964), 261)

Solution

The characteristic equation

$$\xi^2 - 2\lambda\,\xi + 1 = 0$$

has the roots $\quad \xi = \lambda \pm \sqrt{\lambda^2 - 1}.$

The product of roots satisfies the equation $\xi_1 \xi_2 = 1$.

The general solution is given by

$$y_n = C_1 \xi_1^n + C_2 \xi_2^n .$$

Now, $| y_n |$ is bounded if $| \xi_1 | \le 1$ and $| \xi_2 | \le 1$.

For λ real and $| \lambda | < 1$, ξ is a complex pair given by

$$\xi_{1, 2} = \lambda \pm i \sqrt{1 - \lambda^2}$$

and $| \xi_{1, 2} | = 1$. Hence, both the solutions are bounded.

For complex values of λ, ξ is also complex, but they do not form a complex pair. However,

$$| \xi_1 | \, | \xi_2 | = 1.$$

Hence, either $| \xi_1 | > 1$, $| \xi_2 | < 1$ or $| \xi_1 | < 1$, $| \xi_2 | > 1$ while satisfying the above equation. Hence, there is one unbounded solution.

5.4 (*i*) Each term in the sequence 0, 1, 1 / 2, 3 / 4, 5 / 8, ..., is equal to the arithmetic mean of the two preceeding terms. Find the general term.

(*ii*) Find the general solution of the recurrence relation

$$y_{n+2} + 2b \, y_{n+1} + c \, y_n = 0$$

where b and c are real constants.

Show that solutions tend to zero as $n \to \infty$, if and only if, the point (b, c) lies in the interior of a certain region in the b-c plane, and determine this region.

Solution

(*i*) If $y_0, y_1, ..., y_n$ is the sequence, then we have

$$y_{n+2} = \frac{1}{2}(y_{n+1} + y_n), \quad \text{or} \quad 2y_{n+2} = y_{n+1} + y_n$$

which is a second order difference equation with initial conditions

$$y_0 = 0, y_1 = 1.$$

The characteristic equation is

$$2\xi^2 - \xi - 1 = 0$$

whose roots are 1, – 1 / 2.

The general solution becomes

$$y_n = C_1 + C_2 \left(- \frac{1}{2} \right)^n$$

Using the initial conditions, we obtain $C_1 = 2 / 3$, $C_2 = - 2 / 3$.

Hence, the general term becomes

$$y_n = \frac{2}{3} \left[1 - \left(- \frac{1}{2} \right)^n \right].$$

(*ii*) The characteristic equation of the given difference equation is

$$\xi^2 + 2b\xi + c = 0$$

whose roots are $\quad \xi = - b \pm \sqrt{b^2 - c}$.

The general solution of the difference equation is given by

$$y_n = C_1\xi_1^n + C_2\xi_2^n.$$

Now, $y_n \to 0$ as $n \to \infty$ if and only if $|\xi_1| < 1$ and $|\xi_2| < 1$.

Substituting $\xi = (1+z)/(1-z)$, we get the transformed characteristic equation as

$$(1 - 2b + c)z^2 + 2(1 - c)z + 1 + 2b + c = 0.$$

The Routh-Hurwitz criterion requires that

$$1 - 2b + c \geq 0,\ 1 - c \geq 0, \text{ and } 1 + 2b + c \geq 0.$$

Therefore, $|\xi_i| < 1$ and hence $y_n \to 0$ as $n \to \infty$ if the point (b, c) lies in the interior of the triangular region of the b-c plane bounded by the straight lines

$$c = 1,\ 2b - 1 = c,\ 1 + 2b + c = 0$$

as shown in Fig. 5.2.

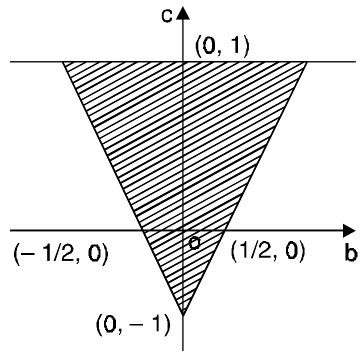


Fig. 5.2. Stability region

5.5 Solve the difference equation

$$\Delta^2 y_n + 3\Delta y_n - 4y_n = n^2$$

with the initial conditions $y_0 = 0,\ y_2 = 2$. (Stockholm Univ., Sweden, BIT 7(1967), 247)

Solution

Substituting for the forward differences in the difference equation we obtain

$$y_{n+2} + y_{n+1} - 6y_n = n^2.$$

Substituting $y_n = A\,\xi^n$ in the homogeneous equation, we obtain the characteristic equation as

$$\xi^2 + \xi - 6 = 0$$

whose roots are $\xi_1 = -3$ and $\xi_2 = 2$.

The complementary solution may be written as

$$y_n^{(H)} = C_1(-3)^n + C_2 2^n.$$

To obtain the particular solution, we set

$$y_n^{(P)} = an^2 + bn + c$$

where a, b and c are constants to be determined.

Substituting in the difference equation, we get

$$[a(n + 2)^2 + b(n + 2) + c] + [a(n + 1)^2 + b(n + 1) + c] - 6\,[an^2 + bn + c] = n^2$$

or $$-4an^2 + (6a - 4b)n + (5a + 3b - 4c) = n^2.$$

Comparing the coefficients of like powers of n, we obtain

$$-4a = 1,\ 6a - 4b = 0,\ 5a + 3b - 4c = 0.$$

The solution is $a = -1/4,\ b = -3/8$ and $c = -19/32$.

The particular solution becomes

$$y_n^{(P)} = -\frac{1}{4}n^2 - \frac{3}{8}n - \frac{19}{32}.$$

Thus, the general solution of the difference equation takes the form

$$y_n = C_1(-3)^n + C_2 2^n - \frac{1}{32}(8n^2 + 12n + 19).$$

The initial conditions $y_0 = 0$, $y_2 = 2$ yield the equations

$$C_1 + C_2 = 19 / 32,$$
$$9C_1 + 4C_2 = 139 / 32,$$

whose solution is $C_1 = 63 / 160$, $C_2 = 32 / 160$.

The general solution becomes

$$y_n = \frac{1}{160}[63(-3)^n + 32(2)^n - 40n^2 - 60n - 95].$$

5.6 Find the general solution of the difference equation

$$y_{n+1} - 2y_n = \frac{n}{2^n}.$$ [Linköping Univ., Sweden, BIT 27 (1987), 438]

Solution

The characteristic equations is $\xi - 2 = 0$, which gives $\xi = 2$.

The complemetary solution is given by

$$y_n = A\, 2^n.$$

We assume the particular solution in the form

$$y_n = \frac{n}{2^n} A_1 + \frac{B_1}{2^n}.$$

Substituting in the difference equation we obtain

$$\frac{n+1}{2^{n+1}} A_1 + \frac{B_1}{2^{n+1}} - 2\frac{n}{2^n} A_1 - \frac{2B_1}{2^n} = \frac{n}{2^n},$$

or $$\frac{n}{2^n}\left(-\frac{3}{2}A_1\right) + \frac{1}{2^n}\left(\frac{1}{2}A_1 - \frac{3}{2}B_1\right) = \frac{n}{2^n}.$$

Comparing the coefficients of $n / 2^n$ and $1 / 2^n$, we get

$$-\frac{3}{2}A_1 = 1, \quad \frac{1}{2}A_1 - \frac{3}{2}B_1 = 0,$$

which gives $$A_1 = -2/3, \quad B_1 = -2/9.$$

The particular solution becomes

$$y_n = -\left(\frac{2}{3}n + \frac{2}{9}\right)2^{-n}.$$

Therefore, the general solution is given by

$$y_n = A2^n - \left(\frac{2}{3}n + \frac{2}{9}\right)2^{-n}.$$

5.7 A sequence of functions $f_n(x)$, $n = 0, 1, \ldots$ defines a recursion formula

$$f_{n+1}(x) = 2x f_n(x) - f_{n-1}(x), \quad |\, x \,| < 1$$
$$f_0(x) = 0, \; f_1(x) = 1.$$

(*a*) Show that $f_n(x)$ is a polynomial and give the degree and leading coefficient.

(*b*) Show that

$$\begin{bmatrix} f_{n+1}(x) \\ T_{n+1}(x) \end{bmatrix} = \begin{bmatrix} x & 1 \\ x^2 - 1 & 1 \end{bmatrix} \begin{bmatrix} f_n(x) \\ T_n(x) \end{bmatrix}$$

where $T_n(x) = \cos(n \cos^{-1} x)$. (Univ. Stockholm, Sweden, BIT 24(1984), 716)

Solution

The characteristic equation of the given recurrence formula is

$$\xi^2 - 2x\xi + 1 = 0$$

with $| x | < 1$ and having the roots $\xi = x \pm i\sqrt{1-x^2}$.

We may write $x = \cos\theta$. Then, we obtain $\xi = e^{\pm i\theta}$.

The general solution becomes

$$f_n(x) = A \cos (n\theta) + B \sin (n\theta).$$

Using the conditions $f_0(x) = 0$ and $f_1(x) = 1$, we obtain

$$A = 0,\ B = 1 / \sin\theta.$$

Therefore, the general solution is given by

$$f_n(x) = \frac{\sin (n\theta)}{\sin\theta}.$$

(*a*) We have

$$f_{n+1}(x) = \frac{\sin (n+1)\theta}{\sin\theta} = \frac{\sin (n\theta)\cos\theta}{\sin\theta} + \frac{\cos (n\theta)\sin\theta}{\sin\theta} = x\, f_n(x) + T_n(x)$$

where $\quad T_n(x) = \cos (n\theta) = \cos (n \cos^{-1} x).$

Hence,

$$f_1(x) = x\, f_0(x) + T_0(x) = 1$$
$$f_2(x) = x\, f_1(x) + T_1(x) = x + x = 2x$$
$$f_3(x) = x\, f_2(x) + T_2(x) = x(2x) + (2x^2 - 1) = 2^2 x^2 - 1$$
$$f_4(x) = x\, f_3(x) + T_3(x) = x(2^2 x^2 - 1) + (4x^3 - 3x) = 2^3 x^3 - 4x$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

Thus, $f_n(x)$ is a polynomial of degree $n - 1$ and its leading coefficient is 2^{n-1}.

(*b*) We have

$$\cos (n + 1)\theta = \cos (n\theta) \cos\theta - \sin (n\theta) \sin\theta,$$

or $\quad T_{n+1}(x) = xT_n(x) - (1 - x^2)\, f_n(x).$

We may now write

$$\begin{bmatrix} f_{n+1}(x) \\ T_{n+1}(x) \end{bmatrix} = \begin{bmatrix} x & 1 \\ x^2 - 1 & x \end{bmatrix} \begin{bmatrix} f_n(x) \\ T_n(x) \end{bmatrix}.$$

5.8 Consider the recursion formula

$$y_{n+1} = y_{n-1} + 2hy_n,$$

$$y_0 = 1,\ y_1 = 1 + h + h^2 \left(\frac{1}{2} + \frac{h}{6} + \frac{h^2}{24} \right).$$

Show that $y_n - e^{nh} = O(h^2)$ as $h \to 0$, for nh = constant.

(Uppsala Univ., Sweden, BIT 14(1974), 482)

Solution

The characteristic equation is

$$\xi^2 - 2h\xi - 1 = 0,$$

whose roots are

$$\xi_{1h} = h + (1 + h^2)^{1/2}$$
$$= 1 + h + \frac{1}{2}h^2 - \frac{1}{8}h^4 + O(h^6)$$
$$= e^h - \frac{1}{6}h^3 + O(h^4) = e^h \left(1 - \frac{1}{6}h^3 + O(h^4) \right).$$

$$\xi_{2h} = h - (1 + h^2)^{1/2}$$

$$= -\left(1 - h + \frac{1}{2}h^2 - \frac{1}{8}h^4 + O(h^6)\right)$$

$$= -\left(e^{-h} + \frac{1}{6}h^3 + O(h^4)\right) = -e^{-h}\left(1 + \frac{1}{6}h^3 + O(h^4)\right).$$

The general solution is given by

$$y_n = C_1\xi_{1h}^n + C_2\xi_{2h}^n$$

where C_1 and C_2 are arbitrary constants to be determined using the initial conditions. Satisfying the initial conditions and solving for C_1 and C_2, we get

$$C_1 = \frac{y_1 - \xi_{2h}}{\xi_{1h} - \xi_{2h}} = 1 + \frac{1}{12}h^3 + O(h^4).$$

$$C_2 = \frac{\xi_{1h} - y_1}{\xi_{1h} - \xi_{2h}} = -\frac{1}{12}h^3 + O(h^4).$$

Substituting for C_1 and C_2 into the general solution, we have

$$y_n = \left(1 + \frac{1}{12}h^3 + O(h^4)\right)e^{x_n}\left(1 - \frac{1}{6}x_n h^2 + O(h^4)\right) + \frac{1}{12}(-1)^{n-1}h^3 e^{-x_n} + O(h^4)$$

$$= e^{x_n} - \frac{1}{6}x_n h^2 e^{x_n} + O(h^3)$$

where $x_n = nh$.

Hence, we obtain

$$y_n - e^{x_n} = O(h^2).$$

5.9 The linear triangular system of equations

$$\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \mathbf{0} \\ & -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ \mathbf{0} & & & -1 & 2 & \end{bmatrix}\mathbf{x} = \begin{bmatrix} 0.001 \\ 0.002 \\ 0.003 \\ \vdots \\ 0.998 \\ 0.999 \end{bmatrix}$$

can be associated with the difference equations

$$-x_{n-1} + 2x_n - x_{n+1} = \frac{n}{1000}, \; n = 1(1)999,$$

$$x_0 = 0, \; x_{1000} = 0.$$

Solve the system by solving the difference equation.

(Lund Univ., Sweden, BIT 20(1980), 529)

Solution

The characteristic equation of the difference equation is

$$-\xi^2 + 2\xi - 1 = 0$$

whose roots are 1, 1.

The complementary solution is given by

$$x_n^{(H)} = C_1 + C_2 n.$$

Since, $\xi = 1$ is a double root of the characteristic equation, we assume

$$x_n^{(P)} = C_3 n^2 + C_4 n^3$$

where the constants C_3 and C_4 are to be determined.

Substituting in the difference equation, we get

$$C_3 [-(n-1)^2 + 2n^2 - (n+1)^2] + C_4 [-(n-1)^3 + 2n^3 - (n+1)^3] = \frac{n}{1000}$$

or
$$-2C_3 - 6nC_4 = \frac{n}{1000}.$$

Comparing the coefficients of like powers of n, we obtain

$$C_3 = 0,\ C_4 = -1/6000.$$

The general solution becomes

$$x_n = C_1 + C_2 n - \frac{n^3}{6000}.$$

The constants C_1 and C_2 are determined by satisfying the boundary conditions. We get

$$C_1 = 0,\ C_2 = 1000/6.$$

Hence, we have the solution

$$x_n = -\frac{1}{6000}(n^3 - 10^6 n),\ n = 1(1)999.$$

5.10 We want to solve the tridiagonal system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is $(N-1) \times (N-1)$ and

$$\mathbf{A} = \begin{bmatrix} -3 & 1 & & & & \mathbf{0} \\ 2 & -3 & 1 & & & \\ & 2 & -3 & 1 & & \\ & \ddots & \ddots & \ddots & & \\ & & & 2 & -3 & 1 \\ \mathbf{0} & & & & 2 & -3 \end{bmatrix},\ \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

State the difference equation which replaces the matrix formulation of the problem, and find the solution. (Umea Univ., Sweden, BIT 24(1984), 398)

Solution

Assuming that $\mathbf{x} = [x_1\ x_2\ \ldots\ x_{N-1}]^T$, we get the difference equation

$$2x_{n-1} - 3x_n + x_{n+1} = 0,\ \ n = 1, 2, \ldots, N-1.$$

For $n = 1$, the difference equation is $2x_0 - 3x_1 + x_2 = 0$.

To match the first equation $-3x_1 + x_2 = 1$ in $\mathbf{Ax} = \mathbf{b}$, we set $x_0 = -1/2$. Similarly, comparing the difference equation for $n = N-1$ and the last equation in $\mathbf{Ax} = \mathbf{b}$, we set $x_N = 0$.

Hence, The boundary conditions are $x_0 = -1/2$, and $x_N = 0$.

The characteristic equation of the difference equation is

$$\xi^2 - 3\xi + 2 = 0$$

whose roots are $\xi = 1$, $\xi = 2$.

The general solution becomes

$$x_n = C_1 + C_2 2^n.$$

The boundary conditions give

$$C_1 + C_2 = -1/2,$$
$$C_1 + C_2 2^N = 0.$$

The solution of this system is

$$C_1 = -\frac{2^{N-1}}{2^N - 1},\ C_2 = \frac{1}{2(2^N - 1)}.$$

The general solution is given by

$$x_n = \frac{2^{n-1} - 2^{N-1}}{2^N - 1}.$$

5.11 Consider the recursion formula for vectors

$$\mathbf{T}\,\mathbf{y}^{(j+1)} = \mathbf{y}^{(j)} + \mathbf{c},$$
$$\mathbf{y}(0) = \mathbf{a}$$

where
$$\mathbf{T} = \begin{bmatrix} 1+2s & -s & & & \mathbf{0} \\ -s & 1+2s & -s & & \\ & \cdots & & & \\ & -s & 1+2s & -s & \\ \mathbf{0} & & -s & 1+2s \end{bmatrix}$$

Is the formula stable, *i.e.* is there any constant k such that $|\ \mathbf{y}^{(n)}\ | < k$ for all $n \ge 0$?

(Royal Inst. Tech., Stockholm, Sweden, BIT 19(1979), 425)

Solution

The matrix may be written as

$$\mathbf{T} = \mathbf{I} + s\,\mathbf{J}$$

where
$$\mathbf{J} = \begin{bmatrix} 2 & -1 & & & \mathbf{0} \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ \mathbf{0} & & & -1 & 2 \end{bmatrix}$$

The recursion formula may be written as

$$\mathbf{y}^{(j+1)} = \mathbf{A}\mathbf{y}^{(j)} + \mathbf{A}\mathbf{c},\ j = 0, 1, 2, \ldots$$

where $\mathbf{A} = (\mathbf{I} + s\mathbf{J})^{-1}$.

Setting $j = 0, 1, 2, \ldots, n$, we get

$$\mathbf{y}^{(1)} = \mathbf{A}\mathbf{y}^{(0)} + \mathbf{A}\mathbf{c},$$
$$\mathbf{y}^{(2)} = \mathbf{A}\mathbf{y}^{(1)} + \mathbf{A}\mathbf{c}$$
$$= \mathbf{A}^2\mathbf{y}^{(0)} + (\mathbf{A} + \mathbf{I})\mathbf{A}\mathbf{c},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\mathbf{y}^{(n)} = \mathbf{A}^n\mathbf{y}^{(0)} + (\mathbf{A}^{n-1} + \mathbf{A}^{n-2} + \ldots + \mathbf{I})\mathbf{A}\mathbf{c}$$
$$= \mathbf{A}^n\mathbf{y}^{(0)} + (\mathbf{I} - \mathbf{A}^n)(\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{A}\mathbf{c}$$

since, $(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^{n-1})(\mathbf{I} - \mathbf{A}) = \mathbf{I} - \mathbf{A}^n$ and $(\mathbf{I} - \mathbf{A})^{-1}$ exists.

The method is stable if $\|\ \mathbf{A}\ \| < 1$. We know that if λ_i is the eigenvalue of $\mathbf{J}$, then the eigenvalue μ_i of $\mathbf{A}$ is $(1 + s\lambda_i)^{-1}$, where

$$\lambda_i = 4\sin^2\left(\frac{i\pi}{2M}\right),\quad i\ (1)\ (M - 1).$$

Thus, we have

$$\mu_i = \frac{1}{1 + 4s\sin^2(i\pi/(2M))},\ i = 1(1)(M - 1).$$

Hence, $0 < \mu_i < 1$, for $s > 0$.

For $s < 0$, we may also have $1 + 4s < -1$ and hence $| \mu_i | < 1$.

This condition gives $s < -1/2$. Hence, the method is stable for $s < -1/2$ or $s > 0$.

Initial Value Problems

Taylor Series Method

5.12 The following IVP is given

$$y' = 2x + 3y, \; y(0) = 1.$$

(*a*) If the error in $y(x)$ obtained from the first four terms of the Taylor series is to be less than 5×10^{-5} after rounding, find x.

(*b*) Determine the number of terms in the Taylor series required to obtain results correct to 5×10^{-6} for $x \le 0.4$

(*c*) Use Taylor's series second order method to get $y(0.4)$ with step length $h = 0.1$.

Solution

(*a*) The analytic solution of the IVP is given by

$$y(x) = \frac{11}{9} e^{3x} - \frac{2}{9}(3x + 1).$$

We have from the differential equation and the initial condition

$$y(0) = 1, \; y'(0) = 3, \; y''(0) = 11,$$
$$y'''(0) = 33, \; y^{iv}(0) = 99.$$

Hence, the Taylor series method with the first four terms becomes

$$y(x) = 1 + 3x + \frac{11}{2}x^2 + \frac{11}{2}x^3.$$

The remainder term is given by

$$R_4 = \frac{x^4}{24} y^{(4)}(\xi).$$

Now $| R_4 | < 5 \times 10^{-5}$ may be approximated by

$$\left| \frac{x^4}{24} 99e^{3x} \right| < 5 \times 10^{-5},$$

or $$x^4 e^{3x} < 0.00001212, \quad \text{or} \quad x \le 0.056.$$

(This value of x can be obtained by applying the Newton-Raphson method on

$$f(x) = x^4 e^{3x} - 0.00001212 = 0).$$

Alternately, we may not use the exact solution. Writing one more term in the Taylor series, we get

$$y(x) = 1 + 3x + \frac{11}{2}x^2 + \frac{11}{2}x^3 + \frac{33}{8}x^4.$$

Differentiating four times, we get $y^{(4)}(x) = 99$. We approximate $\max | y^{(4)}(\xi) | = 99$. Hence, we get

$$| R_4 | = \left| \frac{x^4}{24} y^{(4)}(\xi) \right| \le \frac{99}{24} x^4$$

Now, $\dfrac{99}{24} x^4 < 5 \times 10^{-5}$, gives $x \le 0.059$.

(*b*) If we use the first p terms in the Taylor series method then we have

$$\max_{0 \le x \le 0.4} \left| \frac{x^p}{p!} \right| \max_{\xi \in [0, 0.4]} | y^{(p)}(\xi) | \le 5 \times 10^{-6}.$$

Substituting from the analytic solution we get

$$\frac{(0.4)^p}{p!} (11)\, 3^{p-2}\, e^{1.2} \le 5 \times 10^{-6} \quad \text{or} \quad p > 10.$$

Alternately, we may use the procedure as given in (*a*).

Writing $p + 1$ terms in the Taylor series, we get

$$y(x) = 1 + 3x + \ldots + \frac{(11)\, 3^{p-2}}{p!} x^p$$

Differentiating p times, we get $y^{(p)}(x) = (11)3^{p-2}$. We approximate

$$\max | y^{(p)}(\xi) | = (11)3^{p-2}.$$

Hence, we get

$$\left[\max_{0 \le x \le 0.4} \frac{|x|^p}{p!} \right] (11)3^{p-2} \le 5 \times 10^{-6}$$

or $$\frac{(0.4)^p}{p!} (11)3^{p-2} \le 5 \times 10^{-6}, \quad \text{which gives } p \ge 10.$$

(*c*) The second order Taylor series method is given by

$$y_{n+1} = y_n + h y_n' + \frac{h^2}{2} y_n'', \; n = 0, 1, 2, 3$$

We have
$$y_n' = 2x_n + 2y_n$$
$$y_n'' = 2 + 3y_n' = 2 + 3(2x_n + 3y_n) = 2 + 6x_n + 9y_n.$$

With $h = 0.1$, the solution is obtained as follows :

$n = 0, x_0 = 0 :$
$$y_0 = 1$$
$$y_0' = 2 \times 0 + 3y_0 = 3$$
$$y_0'' = 2 + 6 \times 0 + 9 \times 1 = 11,$$
$$y_1 = 1 + 0.1(3) + \frac{(0.1)^2}{2} \times 11 = 1.355.$$

$n = 1, x_1 = 0.1 :$
$$y_1' = 2 \times 0.1 + 3(1.355) = 4.265.$$
$$y_1'' = 2 + 6 \times 0.1 + 9(1.355) = 14.795.$$
$$y_2 = y_1 + h y_1' + \frac{1}{2} h^2 y_1''$$
$$= 1.355 + 0.1(4.265) + \frac{(0.1)^2}{2} (14.795) = 1.855475.$$

$n = 2, x_2 = 0.2 :$
$$y_2' = 2 \times 0.2 + 3(1.855475) = 5.966425.$$
$$y_2'' = 2 + 6 \times 0.2 + 9(1.855475) = 19.899275.$$
$$y_3 = y_2 + h y_2' + \frac{h^2}{2} y_2''$$
$$= 1.855475 + 0.1(5.966425) + \frac{(0.1)^2}{2} (19.899275) = 2.5516138.$$

$$n = 3,\ x_3 = 0.3: \quad y_3' = 2 \times 0.3 + 3(2.5516138) = 8.2548414$$

$$y_3'' = 26.764524$$

$$y_4 = 2.5516138 + 0.1(8.2548414) + \frac{(0.1)^2}{2}(26.764524)$$

$$= 3.5109205.$$

Hence, the solution values are

$$y(0.1) \approx 1.355,\ y(0.2) \approx 1.85548,$$

$$y(0.3) \approx 2.55161,\ y(0.4) \approx 3.51092.$$

5.13 Compute an approximation to $y(1)$, $y'(1)$ and $y''(1)$ with Taylor's algorithm of order two and steplength $h = 1$ when $y(x)$ is the solution to the initial value problem

$$y''' + 2y'' + y' - y = \cos x,\ 0 \le x \le 1,\ y(0) = 0,\ y'(0) = 1,\ y''(0) = 2.$$

(Uppsala Univ., Sweden, BIT 27(1987), 628)

Solution

The Taylor series method of order two is given by

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{h^2}{2} y''(x_0).$$

$$y'(x_0 + h) = y'(x_0) + hy''(x_0) + \frac{h^2}{2} y'''(x_0).$$

$$y''(x_0 + h) = y''(x_0) + hy'''(x_0) + \frac{h^2}{2} y^{(4)}(x_0).$$

We have

$$y(0) = 0,\ y'(0) = 1,\ y''(0) = 2,$$

$$y'''(0) = -2\,y''(0) - y'(0) + y(0) + 1 = -4,$$

$$y^{(4)}(0) = -2\,y'''(0) - y''(0) + y'(0) = 7.$$

For $h = 1$, $x_0 = 0$, we obtain

$$y(1) \approx 2,\ y'(1) \approx 1,\ y''(1) \approx 3/2.$$

Alternately, we can use the vector form of the Taylor series method. Setting $y = v_1$, we write the given IVP as

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}' = \begin{bmatrix} v_2 \\ v_3 \\ \cos x + v_1 - v_2 - 2v_3 \end{bmatrix},\ \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}(0) = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix},$$

or

$$\mathbf{v}' = \mathbf{f}(x, \mathbf{v}),\ \mathbf{v}(0) = [0 \quad 1 \quad 2]^T.$$

where $\mathbf{v} = [v_1 \quad v_2 \quad v_3]^T$.

The Taylor series method of second order gives

$$\mathbf{v}(1) = \mathbf{v}(0) + h\,\mathbf{v}'(0) + \frac{h^2}{2}\mathbf{v}''(0) = \mathbf{v}(0) + \mathbf{v}'(0) + 0.5\,\mathbf{v}''(0)$$

We have $\mathbf{v}(0) = [0 \quad 1 \quad 2]^T$

$$\mathbf{v}'(0) = \begin{bmatrix} v_2 \\ v_3 \\ \cos x + v_1 - v_2 - 2v_3 \end{bmatrix}(0) = \begin{bmatrix} 1 \\ 2 \\ -4 \end{bmatrix}$$

$$\mathbf{v}''(0) = \begin{bmatrix} v_2' \\ v_3' \\ -\sin x + v_1' - v_2' - 2v_3' \end{bmatrix}(0) = \begin{bmatrix} 2 \\ -4 \\ 7 \end{bmatrix}$$

Hence, we obtain

$$\begin{bmatrix} v_1(1) \\ v_2(1) \\ v_3(1) \end{bmatrix} = \begin{bmatrix} y(1) \\ y'(1) \\ y''(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ -4 \end{bmatrix} + 0.5 \begin{bmatrix} 2 \\ -4 \\ 7 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1.5 \end{bmatrix}$$

5.14 Apply Taylor series method of order p to the problem $y' = y$, $y(0) = 1$ to show that

$$| y_n - y(x_n) | \le \frac{h^p}{(p+1)!} x_n e^{x_n}.$$

Solution

The p-th order Taylor series method for $y' = y$ is given by

$$y_{n+1} = \left(1 + h + \frac{h^2}{2!} + \ldots + \frac{h^p}{p!}\right) y_n = A\, y_n,\ n = 0, 1, 2, \ldots$$

where
$$A = 1 + h + \frac{h^2}{2!} + \ldots + \frac{h^p}{p!}.$$

Setting $n = 0, 1, 2, \ldots$, we obtain the solution of this first order difference equation which satisfies the initial condition, $y(0) = y_0 = 1$, as

$$y_n = A^n = \left(1 + h + \frac{h^2}{2!} + \ldots + \frac{h^p}{p!}\right)^n.$$

The analytic solution of the initial value problem gives

$$y(x_n) = e^{x_n}.$$

Hence, we have

$$y(x_n) - y_n = e^{nh} - \left(1 + h + \frac{h^2}{2} + \ldots + \frac{h^p}{p!}\right)^n \le n \frac{h^{p+1}}{(p+1)!} e^{\theta h} e^{(n-1)h}.$$

Since, $nh = x_n$ and $0 < \theta < 1$, we get

$$| y_n - y(x_n) | \le \frac{h^p}{(p+1)!} x_n e^{x_n}.$$

Runge-Kutta Methods

5.15 Given the equation

$$y' = x + \sin y$$

with $y(0) = 1$, show that it is sufficient to use Euler's method with the step $h = 0.2$ to compute $y(0.2)$ with an error less than 0.05.

(Uppsala Univ., Sweden, BIT 11(1971), 125)

Solution

The value $y(0.2)$ with step length $h = 0.2$ is the first value to be computed with the help of the Euler method and so there is no question of propagation error contributing to the numerical solution. The error involved will only be the local truncation error given by

$$| T_1 | = \frac{1}{2} h^2 \, | y''(\xi) |, \ 0 < \xi < h.$$

Using the differential equation, we find

$$y''(\xi) = 1 + \cos(y(\xi))\, y'(\xi) = 1 + \cos(y(\xi))\, [\xi + \sin (y(\xi))].$$

We obtain $\max_{\xi \in [0, 0.2]} | y''(\xi) | \leq 2.2.$

Hence, we have

$$| T_1 | \leq \frac{1}{2} (0.2)^2 \, 2.2 = 0.044 < 0.05.$$

5.16 Consider the initial value problem

$$y' = x(y + x) - 2, \; y(0) = 2.$$

(*a*) Use Euler's method with step sizes $h = 0.3$, $h = 0.2$ and $h = 0.15$ to compute approximations to $y(0.6)$ (5 decimals).

(*b*) Improve the approximation in (*a*) to $O(h^3)$ by Richardson extrapolation.

(Linköping Inst. Tech., Sweden, BIT 27(1987), 438)

Solution

(*a*) The Euler method applied to the given problem gives

$$y_{n+1} = y_n + h[x_n(y_n + x_n) - 2], \; n = 0, 1, 2, \ldots$$

We have the following results.

$h = 0.3$:

$n = 0, x_0 = 0$: $\quad y_1 = y_0 + 0.3[-2] = 2 - 0.6 = 1.4.$

$n = 1, x_1 = 0.3$: $\quad y_2 = y_1 + 0.3[0.3(y_1 + 0.3) - 2] = 1.4 - 0.447 = 0.953.$

$h = 0.2$:

$n = 0, x_0 = 0$: $\quad y_1 = y_0 + 0.2[-2] = 2 - 0.4 = 1.6.$

$n = 1, x_1 = 0.2$: $\quad y_2 = y_1 + 0.2[0.2(y_1 + 0.2) - 2] = 1.6 + 0.04(1.6 + 0.2) - 0.4 = 1.272.$

$n = 2, x_2 = 0.4$: $\quad y_3 = y_2 + 0.2[0.4(y_2 + 0.4) - 2]$

$\qquad = 1.272 + 0.08(1.272 + 0.4) - 0.4 = 1.00576.$

$h = 0.15$:

$n = 0, x_0 = 0$: $\quad y_1 = y_0 + 0.15[-2] = 2 - 0.3 = 1.7.$

$n = 1, x_1 = 0.15$: $\quad y_2 = y_1 + 0.15[0.15(y_1 + 0.15) - 2]$

$\qquad = 1.7 + 0.0225(1.7 + 0.15) - 0.3 = 1.441625.$

$n = 2, x_2 = 0.30$: $\quad y_3 = y_2 + 0.15[0.3(y_2 + 0.3) - 2]$

$\qquad = 1.441625 + 0.045(1.441625 + 0.3) - 0.3 = 1.219998.$

$n = 3, x_3 = 0.45$: $\quad y_4 = y_3 + 0.15[0.45(y_3 + 0.45) - 2]$

$\qquad = 1.2199981 + 0.0675(1.6699988) - 0.3 = 1.032723.$

(*b*) Since the Euler method is of first order, we may write the error expression in the form

$$y(x, h) = y(x) + c_1 h + c_2 h^2 + c_3 h^3 + \ldots$$

We now have

$$y(0.6, 0.3) = y(0.6) + 0.3c_1 + 0.09c_2 + O(h^3).$$
$$y(0.6, 0.2) = y(0.6) + 0.2c_1 + 0.04c_2 + O(h^3).$$
$$y(0.6, 0.15) = y(0.6) + 0.15c_1 + 0.0225c_2 + O(h^3).$$

Eliminating c_1, we get

$$p = 0.2y(0.6, 0.3) - 0.3y(0.6, 0.2)$$
$$= -0.1y(0.6) + 0.006c_2 + O(h^3).$$
$$q = 0.15y(0.6, 0.2) - 0.2y(0.6, 0.15)$$
$$= -0.05y(0.6) + 0.0015c_2 + O(h^3).$$

Eliminating c_2, we have

$$0.0015\,p - 0.006\,q = 0.00015y(0.6) + O(h^3).$$

Hence, the $O(h^3)$ result is obtained from

$$y(0.6) \approx \frac{0.0015\,p - 0.006\,q}{0.00015} = 10\,p - 40\,q.$$

From (*a*) we have

$$y(0.6, 0.3) = 0.953\ ;\ y(0.6, 0.2) = 1.00576\ ;\quad \text{and}\ \ y(0.6, 0.15) = 1.03272.$$

Substituting these values, we get

$$p = -\,0.111128,\ q = -\,0.05568, \text{ and}$$
$$y(0.6) = 1.11592.$$

5.17 (*a*) Show that Euler's method applied to $y' = \lambda\, y, y(0) = 1, \lambda < 0$ is stable for step-sizes $-\,2 < \lambda\, h < 0$ (stability means that $y_n \to 0$ as $n \to \infty$).

(*b*) Consider the following Euler method for $y' = f(x, y)$,

$$y_{n+1} = y_n + p_1 hf(x_n, y_n)$$
$$y_{n+2} = y_{n+1} + p_2 hf(x_{n+1}, y_{n+1}),\ n = 0, 2, 4, \ldots$$

where $p_1, p_2 > 0$ and $p_1 + p_2 = 2$. Apply this method to the problem given in (*a*) and show that this method is stable for

$$-\,\frac{2}{p_1 p_2} < \lambda h < 0, \text{ if } 1 - \frac{1}{\sqrt{2}} < p_1, p_2 < 1 + \frac{1}{\sqrt{2}}.$$

(Linköping Univ., Sweden, BIT 14(1974), 366)

Solution

(*a*) Applying the Euler method on $y' = \lambda\, y$, we obtain

$$y_{n+1} = (1 + \lambda h)\, y_n,\ \ n = 0, 1, 2, \ldots$$

Setting $n = 0, 1, 2, \ldots$, we get

$$y_n = (1 + \lambda h)^n\, y_0.$$

The solution which satisfies the initial condition $y_0 = 1$, is given by

$$y_n = (1 + \lambda h)^n,\ n = 0, 1, 2, \ldots$$

The Euler method will be stable (in the sense $y_n \to 0$ as $n \to \infty$) if

$$|\ 1 + \lambda h\ | < 1, \lambda < 0, \quad \text{or} \quad -\,2 < \lambda h < 0.$$

(*b*) The application of Euler's method gives

$$y_{n+1} = (1 + p_1 h\lambda) y_n$$
$$y_{n+2} = (1 + p_2 h\lambda) y_{n+1}$$

or
$$y_{n+2} = (1 + p_1 h\lambda)(1 + p_2 h\lambda) y_n.$$

The characteristic equation of this difference equation is

$$\xi^2 = (1 + p_1 h\lambda)(1 + p_2 h\lambda).$$

The stability condition $|\ \xi\ | < 1$ is satisfied if (using the Routh-Hurwitz criterion)

$$(i)\ \ 1 - (1 + p_1 h\lambda)(1 + p_2 h\lambda) > 0,$$

and
$$(ii)\ \ 1 + (1 + p_1 h\lambda)(1 + p_2 h\lambda) > 0.$$

The condition (*i*) is satisfied if

$$-\,2h\lambda - p_1 p_2 h^2 \lambda^2 > 0,$$

or
$$-\,h\lambda\, p_1 p_2 \left(\frac{2}{p_1 p_2} + h\lambda \right) > 0, \quad \text{or} \quad -\,\frac{2}{p_1 p_2} < h\lambda < 0.$$

The condition (*ii*) is satisfied if

$$2 + 2h\lambda + p_1 p_2 h^2\lambda^2 > 0 \quad \text{or} \quad \left(\sqrt{p_1 p_2}\, h\lambda + \frac{1}{\sqrt{p_1 p_2}}\right)^2 + 2 - \frac{1}{p_1 p_2} > 0$$

A sufficient condition is

$$2 - \frac{1}{p_1 p_2} > 0, \quad \text{or} \quad 2p_1 p_2 - 1 > 0.$$

Substituting $p_2 = 2 - p_1$, we have

$$2p_1^2 - 4p_1 + 1 < 0, \quad \text{or} \quad (p_1 - 1)^2 - \frac{1}{2} < 0.$$

Similarly, we obtain

$$(p_2 - 1)^2 - \frac{1}{2} < 0.$$

Hence, if follows

$$1 - \frac{1}{\sqrt{2}} < p_1, p_2 < 1 + \frac{1}{\sqrt{2}}.$$

5.18 (*a*) Give the exact solution of the IVP $y' = xy$, $y(0) = 1$.

(*b*) Estimate the error at $x = 1$, when Euler's method is used, with step size $h = 0.01$. Use the error formula

$$|\, y(x_n) - y(x_n\,;\, h)\,| \le \frac{hM}{2L}\, [\exp\,(x_n - a)\, L - 1]$$

when Euler's method is applied to the problem $y' = f(x, y)$; $y(x) = A$, in $a \le x \le b$ and $h = (b - a) / N$, $x_n = a + nh$ and $|\, \partial f / \partial y \,| \le L$; $|\, y''(x)\,| \le M$.

(Uppsala Univ., Sweden, BIT 25(1985), 428)

Solution

(*a*) Integrating the differential equation

$$\frac{1}{y}\frac{dy}{dx} = x$$

we obtain $y = ce^{x^2/2}$.

The initial condition gives $\quad y(0) = c = 1.$

The exact solution becomes $\quad y(x) = e^{x^2/2}.$

(*b*) We have at $x = 1$,

$$\left|\frac{\partial f}{\partial y}\right| = |\, x \,| \le L = 1,$$

$$|\, y''(x)\,| = (1 + x^2)e^{x^2/2} \le M = 3.297442,$$

$$|\, y(x_n) - y(x_n; h)\,| \le \frac{1}{2}\, [(0.01)\, 3.297442]\, (e - 1) = 0.0283297.$$

Hence, we obtain

$$|\, y(x_n) - y(x_n; h)\,| \le 0.03.$$

5.19 Apply the Euler-Cauchy method with step length h to the problem

$$y' = -\,y,\ y(0) = 1.$$

(*a*) Determine an explicit expression for y_n.

(*b*) For which values of h is the sequence $\{y_n\}_0^{\infty}$ bounded ?

(*c*) Compute $\lim_{h \to 0} \{(y(x\,;\,h) - e^{-x}) \,/\, h^2\}$.

Solution

Applying the Euler-Cauchy method

$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + h, y_n + K_1),$$
$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2),$$

to $y' = -y$, we obtain

$$K_1 = -hy_n,$$
$$K_2 = -h(y_n - hy_n) = -h(1-h)y_n,$$
$$y_{n+1} = \left(1 - h + \frac{1}{2}h^2\right) y_n.$$

(*a*) The solution of the first order difference equation satisfying the initial condition $y(0) = 1$ is given by

$$y_n = \left(1 - h + \frac{1}{2}h^2\right)^n, \ n = 0, 1, 2, \ldots\ldots$$

(*b*) The sequence $\{y_n\}_0^{\infty}$ will remain bounded if and only if

$$\left| 1 - h + \frac{1}{2}h^2 \right| \le 1, \quad \text{or} \quad 0 < h \le 2.$$

(*c*) The analytic solution of the IVP gives $y(x_n) = e^{-x_n}$.

We also have

$$e^{-h} = 1 - h + \frac{h^2}{2} - \frac{h^3}{6} e^{-\theta h}, \ \ 0 < \theta < 1.$$

The solution in (*a*) can be written as

$$y_n = \left(e^{-h} + \frac{h^3}{6} + O(h^4)\right)^n = \left[e^{-h}\left(1 + \frac{h^3}{6} + O(h^4)\right)\right]^n$$
$$= e^{-nh}\left(1 + \frac{nh^3}{6} + O(h^4)\right) = e^{-nh} + \frac{1}{6} x_n e^{-nh} h^2 + O(h^4).$$

Hence, at a fixed point $x_n = x$, and $h \to 0$, we obtain

$$\lim_{h \to 0} \frac{(y(x\,;\,h) - e^{-x})}{h^2} = \frac{1}{6} x\, e^{-x}.$$

5.20 Heun's method with step size h for solving the differential equation

$$y' = f(x, y), \ y(0) = c$$

can be written as

$$K_1 = h\,f(x_n, y_n),$$
$$K_2 = h\,f(x_n + h, y_n + K_1),$$
$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2).$$

(*a*) Apply Heun's method to the differential equation $y' = \lambda y$, $y(0) = 1$. Show that

$$y_n = [H(\lambda h)]^n$$

and state the function H. Give the asymptotic expression for $y_n - y(x_n)$ when $h \to 0$.

(*b*) Apply Heun's method to the differential equation $y' = f(x)$, $y(0) = 0$ and find y_n.

(Royal Inst. Tech., Stockholm, Sweden, BIT 26(1986), 540)

Solution

(*a*) We have

$$K_1 = \lambda h\, y_n,$$
$$K_2 = \lambda h\,(y_n + \lambda h\, y_n) = \lambda h(1 + \lambda h) y_n,$$
$$y_{n+1} = y_n + \frac{1}{2}\,[\lambda h y_n + \lambda h\,(1 + \lambda h)\, y_n]$$
$$= \left(1 + \lambda h + \frac{1}{2}(\lambda h)^2\right) y_n.$$

This is a first order difference equation. The general solution satisfying the initial condition, $y(0) = 1$, is given by

$$y_n = \left(1 + \lambda h + \frac{1}{2}(\lambda h)^2\right)^n.$$

Therefore, we have

$$H(\lambda h) = 1 + \lambda h + \frac{1}{2}\,(\lambda h)^2.$$

The analytic solution of the test equation gives

$$y(x_n) = (e^{\lambda h})^n .$$

Hence, we may write y_n in the form

$$y_n = \left[e^{\lambda h} - \frac{1}{6}(\lambda h)^3 + O(h^4)\right]^n = e^{\lambda nh}\left[1 - \frac{1}{6} n(\lambda h)^3 + O(h^4)\right]$$
$$= y(x_n) - \frac{1}{6} x_n\, \lambda^3 h^2\, e^{\lambda x_n} + O(h^4)$$

Therefore, the asymptotic expression for $y_n - y(x_n)$ is given by

$$\lim_{h \to 0}\left[\left(\frac{y_n - y(x_n)}{h^2}\right)\right] = -\frac{1}{6} x_n\, \lambda^3\, e^{\lambda x_n}.$$

(*b*) The Heun method for $y' = f(x)$, becomes

$$y_n = \int_0^{x_n} f(x)\, dx = T(h).$$

where $T(h)$ is the expression for the trapezoidal rule of integration.

5.21 Consider the following Runge-Kutta method for the differential equation $y' = f(x, y)$

$$y_{n+1} = y_n + \frac{1}{6}\,(K_1 + 4K_2 + K_3),$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right),$$
$$K_3 = hf(x_n + h, y_n - K_1 + 2K_2).$$

(*a*) Compute $y(0.4)$ when

$$y' = \frac{y + x}{y - x},\ y(0) = 1$$

and $h = 0.2$. Round to five decimals.

(*b*) What is the result after one step of length h when $y' = -y$, $y(0) = 1$.

(Lund Univ., Sweden, BIT 27(1987), 285)

Solution

(*a*) Using $y_0 = 1$, $h = 0.2$, we obtain

$$K_1 = 0.2\left[\frac{y_0 + x_0}{y_0 - x_0}\right] = 0.2.$$

$$K_2 = 0.2\left[\frac{y_0 + 0.1 + 0.1}{y_0 + 0.1 - 0.1}\right] = 0.24.$$

$$K_3 = 0.2\left[\frac{y_0 - 0.2 + 0.48 + 0.2}{y_0 - 0.2 + 0.48 - 0.2}\right] = 0.27407.$$

$$y_1 = 1 + \frac{1}{6}(0.2 + 0.96 + 0.27407) = 1.23901.$$

Now, using $y_1 = 1.23901$, $x_1 = 0.2$, we obtain

$$K_1 = 0.2\left(\frac{y_1 + 0.2}{y_1 - 0.2}\right) = 0.277.$$

$$K_2 = 0.2\left(\frac{y_1 + 0.13850 + 0.3}{y_1 + 0.13850 - 0.3}\right) = 0.31137.$$

$$K_3 = 0.2\left(\frac{y_1 - 0.277 + 0.62274 + 0.4}{y_1 - 0.277 + 0.62274 - 0.4}\right) = 0.33505.$$

$$y_2 = y_1 + \frac{1}{6}(0.277 + 4 \times 0.31137 + 0.33505) = 1.54860.$$

(*b*) For $f(x, y) = -y$, we get

$$K_1 = -hy_0,$$

$$K_2 = -h\left(y_0 - \frac{1}{2}hy_0\right) = \left(-h + \frac{1}{2}h^2\right)y_0,$$

$$K_3 = -h\left(y_0 + hy_0 + 2\left(-h + \frac{1}{2}h^2\right)y_0\right) = (-h + h^2 - h^3)\,y_0,$$

$$y_1 = y_0 + \frac{1}{6}\left(-hy_0 + 4\left(-h + \frac{1}{2}h^2\right)y_0 + (-h + h^2 - h^3)y_0\right)$$

$$= \left(1 - h + \frac{1}{2}h^2 - \frac{1}{6}h^3\right)y_0.$$

Therefore, $$y_1 = \left(1 - h + \frac{1}{2}h^2 - \frac{1}{6}h^3\right).$$

5.22 Use the classical Runge-Kutta formula of fourth order to find the numerical solution at $x = 0.8$ for

$$\frac{dy}{dx} = \sqrt{x + y}\,,\ y(0.4) = 0.41.$$

Assume the step length $h = 0.2$.

Solution

For $n = 0$ and $h = 0.2$, we have

$$x_0 = 0.4,\ y_0 = 0.41,$$

$$K_1 = hf(x_0, y_0) = 0.2\,(0.4 + 0.41)^{1/2} = 0.18,$$

$$K_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}K_1\right) = 0.2\left[0.4 + 0.1 + 0.41 + \frac{1}{2}(0.18)\right]^{1/2} = 0.2,$$

$$K_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{1}{2}K_2\right) = 0.2\left[0.4 + 0.1 + 0.41 + \frac{1}{2}(0.2)\right]^{1/2}$$

$$= 0.2009975,$$

$$K_4 = hf(x_0 + h, y_0 + K_3) = 0.2\,[0.4 + 0.2 + 0.41 + 0.2009975]^{1/2}$$

$$= 0.2200907.$$

$$y_1 = y_0 + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

$$= 0.41 + 0.2003476 = 0.6103476.$$

For $n = 1$, $x_1 = 0.6$, and $y_1 = 0.6103476$, we obtain

$$K_1 = 0.2200316,\ K_2 = 0.2383580,\ K_3 = 0.2391256,\ K_4 = 0.2568636,$$

$$y_2 = 0.6103476 + 0.2386436 = 0.8489913.$$

Hence, we have $y(0.6) \approx 0.61035$, $y(0.8) \approx 0.84899$.

5.23 Find the implicit Runge-Kutta method of the form

$$y_{n+1} = y_n + W_1 K_1 + W_2 K_2,$$

$$K_1 = hf(y_n),$$

$$K_2 = hf(y_n + a(K_1 + K_2)),$$

for the initial value problem $y' = f(y)$, $y(t_0) = y_0$.

Obtain the interval of absolute stability for $y' = \lambda y$, $\lambda < 0$.

Solution

Expanding K_2 in Taylor series, we get

$$K_2 = hf_n + ha(K_1 + K_2)f_y + \frac{1}{2}ha^2(K_1 + K_2)^2 f_{yy}$$

$$+ \frac{1}{6}ha^3(K_1 + K_2)^3 f_{yyy} + \ldots\ldots$$

where $f_y = \partial f(x_n) / \partial y$.

We assume the expression for K_2 in the form

$$K_2 = hA_1 + h^2A_2 + h^3A_3 + \ldots.$$

Substituting for K_2 and equating coefficients of like powers of h, we obtain

$$A_1 = f_n,$$

$$A_2 = 2a\,f_n f_y,$$

$$A_3 = aA_2 f_y + 2a^2 f_{yy} f_n^2 = 2a^2 f_n f_y^2 + 2a^2 f_n^2 f_{yy},$$

We also have

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(x_n) + \frac{h^3}{6} y'''(x_n) + ...$$

where $\quad y' = f,$

$$y'' = f_y f,$$

$$y''' = f_{yy} f^2 + f_y^2 f, ...$$

The truncation error in the Runge-Kutta method is given by

$$\begin{aligned} T_{n+1} &= y_{n+1} - y(x_{n+1}) \\ &= y_n + W_1 h f_n + W_2 [h f_n + h^2 \, 2a f_n f_y \\ &\quad + h^3 (2a^2 f_n f_y^2 + 2a^2 f_n^2 f_{yy})] - [y(x_n) + h f_n \\ &\quad + \frac{h^2}{2} f_y f_n + \frac{h^3}{6} (f_{yy} f_n^2 + f_n f_y^2)] + O(h^4). \end{aligned}$$

To determine the three arbitrary constants a, W_1 and W_2, the necessary equations are

$$W_1 + W_2 = 1,$$

$$2W_2 a = 1/2,$$

$$2W_2 a^2 = 1/6,$$

whose solution is $a = 1/3$, $W_2 = 3/4$, $W_1 = 1/4$.

The implicit Runge-Kutta method becomes

$$K_1 = hf(y_n),$$

$$K_2 = hf\left(y_n + \frac{1}{3}(K_1 + K_2)\right),$$

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_2).$$

The truncation error is of the form $O(h^4)$ and hence, the order of the method is three.

Applying the method to $y' = \lambda y$, $\lambda < 0$, we get

$$K_1 = \lambda h \, y_n = \bar{h} y_n,$$

$$K_2 = h\lambda\left(y_n + \frac{1}{3}(K_1 + K_2)\right) = \bar{h}\left(y_n + \frac{1}{3}\bar{h} y_n + \frac{1}{3}K_2\right).$$

Solving for K_2, we get

$$K_2 = \frac{\bar{h}\,[1 + (\bar{h}/3)]}{1 - (\bar{h}/3)} y_n$$

where $\bar{h} = h\lambda$.

Therefore, $$y_{n+1} = y_n + \frac{1}{4}\bar{h} y_n + \frac{3\bar{h}}{4}\left[\frac{1 + (\bar{h}/3)}{1 - (\bar{h}/3)}\right] y_n = \left[\frac{1 + (2\bar{h}/3) + (\bar{h}^2/6)}{1 - (\bar{h}/3)}\right] y_n$$

The characteristic equation is

$$\xi = \frac{1 + (2\bar{h}/3) + (\bar{h}^2/6)}{1 - (\bar{h}/3)}.$$

For absolute stability we require $|\xi| < 1$. Hence, we have

$$-\left(1-\frac{\bar{h}}{3}\right) < 1 + \frac{2}{3}\bar{h} + \frac{\bar{h}^2}{6} < 1 - \frac{\bar{h}}{3}.$$

The right inequality gives

$$\bar{h} + \frac{\bar{h}^2}{6} < 0, \quad \text{or} \quad \frac{\bar{h}}{6}(6 + \bar{h}) < 0.$$

Since, $\bar{h} < 0$, we require $6 + \bar{h} > 0$, which gives $\bar{h} > -6$.

The left inequality gives

$$2 + \frac{\bar{h}}{3} + \frac{\bar{h}^2}{6} > 0$$

which is always satisfied for $\bar{h} > -6$.

Hence, the interval of absolute stability is $\bar{h} \in (-6, 0)$.

5.24 Determine the interval of absolute stability of the implicit Runge-Kutta method

$$y_{n+1} = y_n + \frac{1}{4}(3K_1 + K_2),$$

$$K_1 = hf\left(x_n + \frac{h}{3}, y_n + \frac{1}{3}K_1\right)$$

$$K_2 = hf(x_n + h, y_n + K_1),$$

when applied to the test equation $y' = \lambda y$, $\lambda < 0$.

Solution

Applying the method to the test equation we have

$$K_1 = h\lambda\left(y_n + \frac{1}{3}K_1\right),$$

or

$$K_1 = \left[\frac{\bar{h}}{1-(\bar{h}/3)}\right]y_n, \text{ where } \bar{h} = \lambda h.$$

$$K_2 = h\lambda(y_n + K_1) = \bar{h}\left[y_n + \frac{\bar{h}}{1-(\bar{h}/3)}y_n\right] = \left[\frac{\bar{h} + (2\bar{h}^2/3)}{1-(\bar{h}/3)}\right]y_n$$

Therefore,

$$y_{n+1} = \left[\frac{1 + (2\bar{h}/3) + (\bar{h}^2/6)}{1-(\bar{h}/3)}\right]y_n.$$

The characteristic equation of themethod is same as in example 5.23. The interval of absolute stability is $(-6, 0)$.

5.25 Using the implicit method

$$y_{n+1} = y_n + K_1,$$

$$K_1 = hf\left(t_n + \frac{h}{2}, y_n + \frac{1}{2}K_1\right),$$

find the solution of the initial value problem

$$y' = t^2 + y^2,\ y(1) = 2,\ 1 \le t \le 1.2 \text{ with } h = 0.1.$$

Solution

We have $f(t, y) = t^2 + y^2$. Therefore,

$$K_1 = \left[\left(t_n + \frac{h}{2}\right)^2 + \left(y_n + \frac{K_1}{2}\right)^2\right].$$

We obtain the following results.

$n = 0$: $h = 0.1$, $t_0 = 1$, $y_0 = 2$.

$$K_1 = 0.1\ [(1.05)^2 + (2 + 0.5\ K_1)^2] = 0.51025 + 0.2\ K_1 + 0.025\ K_1^2.$$

This is an implicit equation in K_1 and can be solved by using the Newton-Raphson method. We have

$$F(K_1) = 0.51025 - 0.8\ K_1 + 0.025\ K_1^2,$$
$$F'(K_1) = -0.8 + 0.05K_1.$$

We assume $K_1^{(0)} = h\, f(t_0, y_0) = 0.5$. Using the Newton-Raphson method

$$K_1^{(s+1)} = K_1^{(s)} - \frac{F(K_1^{(s)})}{F'(K_1^{(s)})},\ s = 0, 1, \ldots\ldots$$

We obtain $K_1^{(1)} = 0.650322$, $K_1^{(2)} = 0.651059$, $K_1^{(3)} = 0.651059$.

Therefore, $K_1 \approx K_1^{(3)} = 0.651059$ and $y(1.1) \approx y_1 = 2.651059$.

$n = 1$: $h = 0.1$, $t_1 = 1.1$, $y_1 = 2.65106$

$$K_1 = 0.1[(1.15)^2 + (2.65106 + 0.5\ K_1)^2]$$
$$= 0.835062 + 0.265106\ K_1 + 0.025\ K_1^2$$
$$F(K_1) = 0.835062 - 0.734894\ K_1 + 0.025\ K_1^2$$
$$F'(K_1) = -0.734894 + 0.05\ K_1.$$

We assume $K_1^{(0)} = hf(t_1, y_1) = 0.823811$. Using the Newton-Raphson method, we get

$$K_1^{(1)} = 1.17932,\ K_1^{(2)} = 1.18399,\ K_1^{(3)} = 1.18399.$$

Therefore, $K_1 \approx K_1^{(3)} = 1.18399$ and $y(1.2) \approx y_2 = 3.83505$.

5.26 Solve the initial value problem

$$u' = -2tu^2,\ u(0) = 1$$

with $h = 0.2$ on the interval [0, 0.4]. Use the second order implicit Runge-Kutta method.

Solution

The second order implicit Runge-Kutta method is given by

$$u_{j+1} = u_j + K_1,\ j = 0, 1$$
$$K_1 = hf\left(t_j + \frac{h}{2}, u_j + \frac{1}{2}K_1\right)$$

which gives $$K_1 = -h\ (2t_j + h)\left(u_j + \frac{1}{2}K_1\right)^2.$$

This is an implicit equation in K_1 and can be solved by using an iterative method. We generally use the Newton-Raphson method. We write

$$F(K_1) = K_1 + h(2t_j + h)\left(u_j + \frac{1}{2}K_1\right)^2 = K_1 + 0.2\ (2t_j + 0.2)\left(u_j + \frac{1}{2}K_1\right)^2$$

We have $F'(K_1) = 1 + h\,(2t_j + h)\,(u_j + \frac{1}{2}K_1) = 1 + 0.2\,(2t_j + 0.2)\,(u_j + \frac{1}{2}K_1)$.

The Newton-Raphson method gives

$$K_1^{(s+1)} = K_1^{(s)} - \frac{F\,(K_1^{(s)})}{F'\,(K_1^{(s)})}, \; s = 0, 1, \ldots\ldots$$

We assume $K_1^{(0)} = hf\,(t_j,\; u_j),\; j = 0, 1.$

We obtain the following results.

$j = 0$; $t_0 = 0,\, u_0 = 1,\; K_1^{(0)} = -\,h(2t_0 u_0^2) = 0,$

$F\,(K_1^{(0)}) = 0.04,\; F'(K_1^{(0)}) = 1.04,\; K_1^{(1)} = -\,0.03846150,$

$F\,(K_1^{(1)}) = 0.00001483,\; F'(K_1^{(1)}) = 1.03923077,\; K_1^{(2)} = -\,0.03847567$

$F\,(K_1^{(2)}) = 0.30 \times 10^{-8}.$

Therefore, $K_1 \approx K_1^{(2)} = -\,0.03847567,$

and $u(0.2) \approx u_1 = u_0 + K_1 = 0.96152433.$

$j = 1$; $t_1 = 0.2,\, u_1 = 0.96152433,\; K_1^{(0)} = -\,h(2t_1 u_1^2) = -\,0.07396231,$

$F\,(K_1^{(0)}) = 0.02861128,\; F'(K_1^{(0)}) = 1.11094517,\; K_1^{(1)} = -\,0.09971631,$

$F\,(K_1^{(1)}) = 0.00001989,\; F'(K_1^{(1)}) = 1.10939993,\; K_1^{(2)} = -\,0.09973423,$

$F\,(K_1^{(2)}) = 0.35 \times 10^{-7},\; F'(K_1^{(2)}) = 1.10939885,\; K_1^{(3)} = -\,0.099773420.$

Therefore, $K_1 \approx K_1^{(3)} = -\,0.09973420,$

and $u(0.4) \approx u_2 = u_1 + K_1 = 0.86179013.$

Multistep Methods

5.27 Find the solution at $x = 0.3$ for the differential equation

$$y' = x - y^2,\; y(0) = 1$$

by the Adams-Bashforth method of order two with $h = 0.1$. Determine the starting values using a second order Runge-Kutta method.

Solution

The second order Adams-Bashforth method is

$$y_{n+1} = y_n + \frac{h}{2}\,(3y_n' - y_{n-1}'),\; n = 1, 2, \ldots..$$

We need the value of $y(x)$ at $x = x_1$ for starting the computation. This value is determined with the help of the second order Runge-Kutta method

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2),$$

$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + h, y_n + K_1).$$

We have $y' = x - y^2$, $y_0 = 1$, $x_0 = 0$,

$$K_1 = 0.1(0 - 1) = -0.1,$$
$$K_2 = 0.1(0.1 - (1 - 0.1)^2) = -0.071,$$
$$y_1 = 1 + \frac{1}{2}(-0.1 - 0.071) = 0.9145.$$
$$y_0' = 0 - 1 = -1,$$
$$y_1' = 0.1 - (0.9145)^2 = -0.73631.$$

Using the Adams-Bashforth method, we now obtain

$$y_2 = y_1 + \frac{0.1}{2}(3y_1' - y_0')$$
$$= 0.9145 + \frac{0.1}{2}(-3 \times 0.73631 + 1) = 0.85405.$$
$$y_2' = 0.2 - (0.85405)^2 = -0.52940,$$
$$y_3 = y_2 + \frac{0.1}{2}(3y_2' - y_1')$$
$$= 0.85405 + \frac{0.1}{2}(3 \times (-0.52940) + 0.73631) = 0.81146.$$

Therefore, we have

$$y_1 = 0.9145,\ y_2 = 0.85405,\ y_3 = 0.81146.$$

5.28 Derive a fourth order method of the form

$$y_{n+1} = ay_{n-2} + h(by_n' + cy_{n-1}' + dy_{n-2}' + ey_{n-3}')$$

for the solution of $y' = f(x, y)$. Find the truncation error.

Solution

The truncation error of the method is written as

$$T_{n+1} = y(x_{n+1}) - ay(x_{n-2}) - h[by'(x_n)$$
$$+ cy'(x_{n-1}) + dy'(x_{n-2}) + ey'(x_{n-3})]$$
$$= C_0 y(x_n) + C_1 hy'(x_n) + C_2 h^2 y''(x_n) + C_3 h^3 y'''(x_n)$$
$$+ C_4 h^4 y^{(4)}(x_n) + C_5 h^5 y^{(5)}(x_n) + \ldots\ldots.$$

To determine *a, b, c, d* and *e* we have the following equations

$$C_0 = 1 - a = 0,$$
$$C_1 = 1 + 2a - (b + c + d + e) = 0,$$
$$C_2 = \frac{1}{2}(1 - 4a) + (c + 2d + 3e) = 0,$$
$$C_3 = \frac{1}{6}(1 + 8a) - \frac{1}{2}(c + 4d + 9e) = 0,$$

$$C_4 = \frac{1}{24}(1 - 16a) + \frac{1}{6}(c + 8d + 27e) = 0,$$

whose solution is, $a = 1$, $b = 21 / 8$, $c = -9 / 8$, $d = 15 / 8$, $e = -3 / 8$.

Thus, we obtain the method

$$y_{n+1} = y_{n-2} + \frac{h}{8}(21y'_n - 9y'_{n-1} + 15y'_{n-2} - 3y'_{n-3});$$

with the truncation error

$$T_{n+1} = \left[\frac{1}{120}(1 + 32a) - \frac{1}{24}(c + 16d + 81e)\right] h^5 y^{(5)}(\xi) = \frac{27}{80} h^5 y^{(5)}(\xi)$$

where $x_{n-3} < \xi < x_{n+1}$.

5.29 If $\rho(\xi) = (\xi - 1)(\xi - \lambda)$ where λ is a real and $-1 \le \lambda < 1$, find $\sigma(\xi)$, such that the resulting method is an implicit method. Find the order of the method for $\lambda = -1$.

Solution

We have
$$\sigma(\xi) = \frac{\rho(\xi)}{\log \xi} = \frac{(\xi - 1)(1 - \lambda + \xi - 1)}{\log(1 + \xi - 1)}$$

$$= \frac{(\xi - 1)[(1 - \lambda) + (\xi - 1)]}{[(\xi - 1) - \frac{1}{2}(\xi - 1)^2 + \frac{1}{3}(\xi - 1)^3 - .....]}$$

$$= [(1 - \lambda) + \{\xi - 1)]\left[1 - \left\{\frac{1}{2}(\xi - 1) - \frac{1}{3}(\xi - 1)^2 + ...\right\}\right]^{-1}$$

$$= 1 - \lambda + \frac{3 - \lambda}{2}(\xi - 1) + \frac{5 + \lambda}{12}(\xi - 1)^2 - \frac{1 + \lambda}{24}(\xi - 1)^3 + O((\xi - 1)^4)$$

Hence,
$$\sigma(\xi) = 1 - \lambda + \frac{3 - \lambda}{2}(\xi - 1) + \frac{5 + \lambda}{12}(\xi - 1)^2$$

Thus, we find that for $\lambda \ne -1$, the order is 3 while for $\lambda = -1$, the order is 4.

5.30 One method for the solution of the differential equation $y' = f(y)$ with $y(0) = y_0$ is the implicit mid-point method

$$y_{n+1} = y_n + hf\left(\frac{1}{2}(y_n + y_{n+1})\right).$$

Find the local error of this method. (Lund Univ., Sweden, BIT 29(1989), 375)

Solution

The truncation error of the method is given by

$$T_{n+1} = y(x_{n+1}) - y(x_n) - hf\left(\frac{1}{2}(y(x_n) + y(x_{n+1}))\right)$$

$$= y(x_n) + hy'(x_n) + \frac{1}{2}h^2 y''(x_n)$$

$$+ \frac{1}{6}h^3 y'''(x_n) + ... - y(x_n) - hf\left(y(x_n) + \frac{1}{2}hy'(x_n) + \frac{1}{4}h^2 y''(x_n) + ...\right)$$

$$= hy'(x_n) + \frac{1}{2}h^2y''(x_n) + \frac{1}{6}h^3y'''(x_n) - h\left[f_n + \left(\frac{1}{2}hy'_n + \frac{1}{4}h^2y''_n + \ldots\right)f_y\right.$$
$$\left. + \frac{1}{2}\left(\frac{1}{2}hy'_n + \frac{1}{4}h^2y''_n + \ldots\right)^2 f_{yy} + \ldots\right]$$

We have $\quad y' = f,\ y'' = ff_y,\ y''' = ff_y^2 + f^2f_{yy}.$

On simplification, we obtain

$$T_{n+1} = -\frac{1}{24}h^3 f_n\,(2f_y^2 - ff_{yy})_{x_n} + O(h^4).$$

5.31 Consider an implicit two-step method

$$y_{n+1} - (1+a)\,y_n + a\,y_{n-1} = \frac{h}{12}[(5+a)y'_{n+1} + 8(1-a)y'_n - (1+5a)y'_{n-1}]$$

where $-1 \le a < 1$, for the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$.

(*i*) Show that the order of the two-step method is 3 if $a \ne -1$ and is 4 if $a = -1$.

(*ii*) Prove that the interval of absolute stability is $(-6(a+1)/(a-1), 0)$ and that the interval of relative stability is $(3(a+1)/(2(a-1)), \infty)$.

Solution

(*i*) The truncation error of the two-step method is given by

$$T_{n+1} = y(x_{n+1}) - (1+a)\,y(x_n) + ay(x_{n-1})$$
$$-\frac{h}{12}[(5+a)\,y'(x_{n+1}) + 8(1-a)\,y'(x_n) - (1+5a)\,y'(x_{n-1})]$$
$$= C_0y(x_n) + C_1h\,y'(x_n) + C_2h^2y''(x_n) + C_3h^3y'''(x_n) + C_4h^4y^{(4)}(x_n) + \ldots$$

where $\quad C_0 = 0,\ C_1 = 0,\ C_2 = 0,\ C_3 = 0,\quad C_4 = -(1+a)/24.$

Hence, the truncation error is

$$T_{n+1} = -\frac{1}{24}(1+a)h^4\,y^{(4)}(x_n) + \left(\frac{a-1}{90}\right)h^5\,y^{(5)}(x_n) + O(h^6).$$

Therefore, the two-step method has order 3 if $a \ne -1$ and order 4 if $a = -1$.

(*ii*) The characteristic equation of the method is given by

$$\left(1 - \frac{\bar{h}}{12}(5+a)\right)\xi^2 - \left((1+a) + \frac{2}{3}\bar{h}(1-a)\right)\xi + \left(a + \frac{\bar{h}}{12}(1+5a)\right) = 0.$$

Absolute Stability : Setting $\xi = (1+z)/(1-z)$, the transformed characteristic equation is obtained as

$$\left(2(1+a) + \frac{\bar{h}}{3}(1-a)\right)z^2 + 2\left((1-a) - \frac{\bar{h}}{2}(1+a)\right)z - \bar{h}\,(1-a) = 0.$$

The Routh-Hurwitz criterion is satisfied if

$$2(1+a) + \frac{\bar{h}}{3}(1-a) > 0,$$
$$(1-a) - \frac{\bar{h}}{2}\,(1+a) > 0,$$
$$-\bar{h}\,(1-a) > 0.$$

For $\bar{h} < 0$ and $-1 \le a < 1$, the conditions will be satisfied if $\bar{h} \in (-6(1+a)/(1-a), 0)$. Hence, the interval of absolute stability is $\bar{h} \in (-6(1+a)/(1-a), 0)$.

Relative Stability : It is easily verified that the roots of the characteristic equation are real and distinct for all $\bar{h}$ and for all a. The end points of the interval of relative stability are given by $\xi_{1h} = \xi_{2h}$ and $\xi_{1h} = -\xi_{2h}$. The first condition is never satisfied that is, the interval extends to $+\infty$, ; the second condition gives

$$(1+a) + \frac{2}{3}\bar{h}(1-a) = 0 \quad \text{or} \quad \bar{h} = \frac{3(a+1)}{2(a-1)}.$$

Hence, the interval of relative stability is

$$\bar{h} \in \left(\frac{3(a+1)}{2(a-1)},\ \infty\right).$$

5.32 Determine the constants α, β and γ so that the difference approximation

$$y_{n+2} - y_{n-2} + \alpha(y_{n+1} - y_{n-1}) = h\,(\beta\,(f_{n+1} + f_{n-1}) + \gamma f_n]$$

for $y' = f(x, y)$ will have the order of approximation 6. Is the difference equation stable for $h = 0$? (Uppsala Univ., Sweden, BIT 9(1969), 87)

Solution

The truncation error of the method is given by

$$\begin{aligned} T_{n+1} &= y(x_{n+2}) - y(x_{n-2}) + \alpha\,(y(x_{n+1}) - y(x_{n-1})) \\ &\quad - h\,[\beta\,(y'(x_{n+1}) + y'(x_{n-1})) + \gamma y'(x_n)] \\ &= C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + C_3 h^3 y'''(x_n) \\ &\quad + C_4 h^4 y^{(4)}(x_n) + C_5 h^5 y^{(5)}(x_n) \\ &\quad + C_6 h^6 y^{(6)}(x_n) + C_7 h^7 y^{(7)}(x_n) + \ldots\ldots \end{aligned}$$

where $\quad C_0 = 0, \quad C_1 = 4 + 2\alpha - 2\beta - \gamma,\ C_2 = 0, \quad C_3 = \frac{1}{6}(16 + 2\alpha) - \beta,$

$$C_4 = 0, \quad C_5 = \frac{1}{120}(64 + 2\alpha) - \frac{\beta}{12},\ C_6 = 0, \quad C_7 = \frac{1}{5040}(256 + 2\alpha) - \frac{\beta}{360}.$$

Setting $C_i = 0$, $i = 1, 3, 5$, we obtain

$$\alpha = 28,\ \beta = 12,\ \gamma = 36 \quad \text{and} \quad C_7 = 1/35.$$

The sixth order method is

$$y_{n+2} + 28y_{n+1} - 28y_{n-1} - y_{n-2} = h(12 f_{n+1} + 36 f_n + 12 f_{n-1})$$

with the truncation error

$$T_{n+1} = \frac{1}{35} h^7 y^{(7)}(x_n) + O(h^8).$$

The reduced characteristic equation ($h = 0$) is

$$\xi^4 + 28\xi^3 - 28\xi - 1 = 0$$

whose roots are $\xi = 1, -1, -0.03576, -27.96424$.

Since the root condition is not satisfied, the method is unstable.

5.33 The difference equation

$$\frac{1}{(1+a)}(y_{n+1}-y_n)+\frac{a}{(1+a)}(y_n-y_{n-1})=-hy_n,\ h>0,\ a>0$$

which approximates the differential equation $y' = -y$ is called strongly stable, if for sufficiently small values of h $\lim_{n\to\infty} y_n = 0$ for all solutions y_n. Find the values of a for which strong stability holds. (Royal Inst. Tech., Stockholm, Sweden, BIT 8(1968), 138)

Solution

The reduced characteristic equation ($h = 0$) is

$$\xi^2-(1-a)\,\xi-a=0.$$

whose roots are $1, -a$. The root condition is satisfied if $|a| < 1$. The characteristic equation is given by

$$\xi^2-[1-a-h(1+a)]\,\xi-a=0.$$

Setting $\xi = (1+z)/(1-z)$, the transformed characteristic equation is obtained as

$$[2(1-a)-h(1+a)]\,z^2+2(1+a)z+h(1+a)=0.$$

Since, $|a| < 1$, we get $1 + a > 0$.

The Routh-Hurwitz criterion is satisfied if

$$0<h<2(1-a)/(1+a).$$

5.34 To solve the differential equation $y' = f(x, y)$, $y(x_0) = y_0$, the method

$$y_{n+1}=\frac{18}{19}(y_n-y_{n-2})+y_{n-3}+\frac{4h}{19}(f_{n+1}+4f_n+4f_{n-2}+f_{n-3})$$

is suggested.

(*a*) What is the local truncation error of the method ?

(*b*) Is the method stable ? (Lund Univ., Sweden, BIT 20(1980), 261)

Solution

The truncation error of the method may be written as

$$T_{n+1}=y(x_{n+1})-\frac{18}{19}(y(x_n)-y(x_{n-2}))-y(x_{n-3})$$
$$-\frac{4h}{19}(y'(x_{n+1})+4y'(x_n)+4y'(x_{n-2})+y'(x_{n-3}))$$
$$=C_0y(x_n)+C_1h\,y'(x_n)+C_2h^2y''(x_n)+C_3h^3y'''(x_n)+\ldots..$$

where $C_0 = 0 = C_1 = C_2$, $C_3 = 2/3$.

Hence, the truncation error becomes

$$T_{n+1}=\frac{2}{3}h^3y'''(x_n)+O(h^4).$$

The reduced characteristic equation is (set $h = 0$)

$$\xi^4-\frac{18}{19}(\xi^3-\xi)-1=0$$

whose roots are ± 1, $(9 \pm i\sqrt{280})/19$.

The roots have modulus one and hence the root condition is satisfied.

The characteristic equation is given by

$$(19 - 4\bar{h})\,\xi^4 - (18 + 16\bar{h})\,\xi^3 + (18 - 16\bar{h})\,\xi - (19 + 4\bar{h}) = 0$$

where $\bar{h} = \lambda h < 0$.

Let the characteristic equation be written as

$$a\xi^4 + b\xi^3 + c\xi^2 + d\xi + e = 0.$$

Substituting $\xi = (1 + z) / (1 - z)$, we obtain the transformed characteristic equation as

$$v_0 z^4 + v_1 z^3 + v_2 z^2 + v_3 z + v_4 = 0$$

where $v_0 = a - b + c - d + e,\ v_1 = 4a - 2b + 2d - 4e,$

$v_2 = 6a - 2c + 6e,\ \ v_3 = 4a + 2b - 2d - 4e,$

$v_4 = a + b + c + d + e.$

Substituting $a = 19 - 4\bar{h}$, $b = -(18 + 16\bar{h})$, $c = 0$, $d = 18 - 16\bar{h}$, and $e = -(19 + 4\bar{h})$, we obtain the transformed equation as

$$6\bar{h}\ z^4 + 56z^3 - 12\bar{h}\,z^2 + 20z - 10\bar{h} = 0.$$

The necessary condition for the application of the Routh-Hurwitz criterion is $v_i \ge 0$. Since $\bar{h} < 0$, this condition is violated. Hence, there is atleast one root which lies in the right half-plane of the z-plane. Hence, the method is unstable for $h > 0$. It is stable for $h = 0$ and so we may conclude that the method is weakly stable.

5.35 For the corrector formula

$$y_{n+1} - \alpha y_{n-1} = Ay_n + By_{n-2} + h(Cy'_{n+1} + Dy'_n + Ey'_{n-1}) + T$$

we have $T = O(h^5)$.

(*a*) Show that $A = 9(1 - \alpha) / 8$, $B = -(1 - \alpha) / 8$, and determine C, D and E.

(*b*) Find the zero-stability conditions.

Solution

(*a*) Expanding each term in the truncation error in Taylor's series, we have

$$\begin{aligned} T &= y(x_{n+1}) - \alpha y(x_{n-1}) - Ay(x_n) - By(x_{n-2}) \\ &\quad - h(Cy'(x_{n+1}) + Dy'(x_n) + Ey'(x_{n-1})) \\ &= C_0 y(x_n) + C_1 h\, y'(x_n) + C_2 h^2 y''(x_n) \\ &\quad + C_3 h^3 y'''(x_n) + C_4 h^4 y^{(4)}(x_n) + O(h^5) \end{aligned}$$

where $C_0 = 1 - \alpha - A - B,$

$C_1 = 1 + \alpha + 2B - (C + D + E),$

$$C_2 = \frac{1}{2}(1 - \alpha - 4B) - (C - E),$$

$$C_3 = \frac{1}{6}(1 + \alpha + 8B) - \frac{1}{2}(C + E),$$

$$C_4 = \frac{1}{24}(1 - \alpha - 16B) - \frac{1}{6}(C - E).$$

Setting $C_i = 0$, $i = 0, 1, \ldots, 4$, we obtain

$$A = \frac{9}{8}(1 - \alpha),\ B = -\frac{1}{8}(1 - \alpha),\ C = -\frac{1}{24}(\alpha - 9),$$

$$D = \frac{1}{12}(9 + 7\alpha),\ E = \frac{1}{24}(17\alpha - 9).$$

(*b*) The reduced characteristic equation is (set $h = 0$)

$$\xi^3 - \frac{9}{8}(1 - \alpha)\,\xi^2 - \alpha\xi + \frac{1}{8}(1 - \alpha) = 0,$$

with one root $\xi = 1$ and the other two roots are the roots of

$$\xi^2 + \frac{1}{8}(9\alpha - 1)\xi + \frac{1}{8}(\alpha - 1) = 0.$$

Setting $\xi = (1 + z) / (1 - z)$, we get the transformed equation as

$$(1 - \alpha)\, z^2 + \frac{1}{4}(9 - \alpha)\, z + \frac{1}{4}(3 + 5\alpha) = 0.$$

Routh-Hurwitz criterion gives the conditions $1 - \alpha > 0$, $9 - \alpha > 0$, and $3 + 5\alpha > 0$ which give $\alpha \in (-0.6, 1)$. Hence, the root condition is satisfied if $-0.6 < \alpha < 1$. Therefore, the method is stable for $-0.6 < \alpha < 1$.

5.36 Use the two-step formula

$$y_{n+1} = y_{n-1} + \frac{h}{3}(y'_{n+1} + 4y'_n + y'_{n-1})$$

to solve the test problem $y' = \lambda y$, $y(0) = y_0$, where $\lambda < 0$.

Determine $\lim_{n \to \infty} |y_n|$ and $\lim_{n \to \infty} y(x_n)$ where $x_n = nh$, h fixed, and $y(x)$ is the exact solution of the test problem. (Uppsala Univ., Sweden, BIT 12(1972), 272)

Solution

We apply the method to the test equation $y' = \lambda y$, $\lambda < 0$, and obtain

$$\left(1 - \frac{\bar{h}}{3}\right) y_{n+1} - \frac{4\bar{h}}{3}\, y_n - \left(1 + \frac{\bar{h}}{3}\right) y_{n-1} = 0.$$

The characteristic equation is given by

$$\left(1 - \frac{\bar{h}}{3}\right)\xi^2 - \frac{4\bar{h}}{3}\xi - \left(1 + \frac{\bar{h}}{3}\right) = 0,$$

whose roots are

$$\xi_{1h} = \left[\frac{2\bar{h}}{3} + \left(1 + \frac{\bar{h}^2}{3}\right)^{1/2}\right] \Big/ \left(1 - \frac{\bar{h}}{3}\right)$$

$$= 1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{\bar{h}^3}{6} + \frac{\bar{h}^4}{24} + \frac{\bar{h}^5}{72} + \ldots \simeq e^{\bar{h}} + c_1 h^5,$$

$$\xi_{2h} = \left[\frac{2\bar{h}}{3} - \left(1 + \frac{\bar{h}^2}{3}\right)^{1/2}\right] \Big/ \left(1 - \frac{\bar{h}}{3}\right)$$

$$= -\left(1 - \frac{\bar{h}}{3} + \frac{\bar{h}^2}{18} + \frac{\bar{h}^3}{54} + \ldots\right) \simeq -(e^{-\bar{h}/3} + c_2 h^3),$$

where $c_1 = \lambda^5 / 180$ and $c_2 = 2\lambda^3 / 81$.

The general solution of the difference equation is

$$y_n = A\xi_{1h}^n + B\xi_{2h}^n$$

We have $$\xi_{1h}^n \simeq (e^{\lambda h} + c_1 h^5)^n = e^{\lambda nh}(1 + c_1 h^5 e^{-\lambda h})^n = e^{\lambda nh}(1 + nc_1 h^5 e^{-\lambda h} + \ldots)$$
$$\simeq e^{\lambda nh}\left(1 + \frac{n\lambda^5 h^5}{180}\right).$$
We also find $$\xi_{2h} = (-1)^n (e^{-\lambda h/3} + c_2 h^3)^n = (-1)^n e^{-n\lambda h/3}(1 + c_2 h^3 e^{n\lambda h/3})^n$$
$$\simeq (-1)^n e^{-\lambda nh/3}\left(1 + \frac{2}{81} n\lambda^3 h^3\right).$$
Hence, the general solution is
$$y_n \simeq Ae^{\lambda nh}\left(1 + \frac{n\lambda^5}{180} h^5\right) + B(-1)^n e^{-\lambda nh/3}\left(1 + \frac{2}{81} n\lambda^3 h^3\right)$$
For $\lambda < 0$, the limiting value as $n \to \infty$ is given by
$$\lim_{n\to\infty} y_n = A \lim_{n\to\infty} e^{\lambda nh}\left(1 + \frac{n\lambda^5}{180} h^5\right) + B \lim_{n\to\infty} (-1)^n e^{-\lambda nh/3}\left(1 + \frac{2}{81} n\lambda^3 h^3\right)$$
The first term on the right tends to zero whereas the second term oscillates and tends to infinity.

Therefore, we obtain $\lim_{n\to\infty} |\, y_n \,| = \infty.$

In the limit, the analytic solution tends to zero
$$\lim_{n\to\infty} y(x_n) = \lim_{n\to\infty} e^{\lambda x_n} = 0.$$

5.37 The formula
$$y_{n+3} = y_n + \frac{3h}{8}(y_n' + 3y_{n+1}' + 3y_{n+2}' + y_{n+3}')$$
with a small step length h is used for solving the equation $y' = -y$. Investigate the convergence properties of the method. (Lund Univ., Sweden, BIT 7(1967), 247)

Solution

The formula may be written as $\rho(E)y_n - h\sigma(E)y_n' = 0$, where
$$\rho(\xi) = \xi^3 - 1 \quad \text{and} \quad \sigma(\xi) = \frac{3}{8}(\xi^3 + 3\xi^2 + 3\xi + 1) = \frac{3}{8}(\xi + 1)^3.$$
The roots of the reduced characteristic equation are 1, ω, ω^2 where ω is the cube root of unity. The growth parameters are given by
$$\kappa_j = \frac{\sigma(\xi_j)}{\xi_j\, \rho'(\xi_j)}, \quad j = 1, 2, 3.$$
We have $$\kappa_j = \frac{(\xi_j + 1)^3}{8\xi_j^3}.$$
We obtain the following values for the growth parameters.
$$j = 1: \quad \xi_1 = 1, \quad \kappa_1 = \frac{1}{8} \cdot \frac{8}{1} = 1.$$
$$j = 2: \quad \xi_2 = \omega, \quad \kappa_2 = \frac{1}{8}\frac{(1+\omega)^3}{\omega^3} = -\frac{1}{8}.$$
$$j = 3: \quad \xi_3 = \omega^2, \quad \kappa_3 = \frac{1}{8}\frac{(1+\omega^2)^3}{\omega^6} = -\frac{1}{8}.$$

The difference equation has the following approximate roots.

$$\xi_{jh} = \xi_j (1 - h\kappa_j + O(h^2)), \quad j = 1, 2, 3.$$

$$\xi_{1h} = 1 - h + O(h^2).$$

$$\xi_{2h} = \left(1 + \frac{h}{8} + O(h^2)\right)\left(-\frac{1}{2} + \frac{\sqrt{3}}{2} i\right).$$

$$\xi_{3h} = \left(1 + \frac{h}{8} + O(h^2)\right)\left(-\frac{1}{2} - \frac{\sqrt{3}}{2} i\right).$$

The solution of the difference equation becomes

$$\begin{aligned} y_n &= C_1(1 - h + O(h)^2)^n \\ &\quad + C_2\left(1 + \frac{h}{8} + O(h^2)\right)^n \left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right)^n \\ &\quad + C_3\left(1 + \frac{h}{8} + O(h^2)\right)^n \left(-\frac{1}{2} - i\frac{\sqrt{3}}{2}\right)^n \\ &\approx C_1 e^{-nh} + (-1)^n e^{nh/8}\left(a_1 \cos\frac{n\pi}{3} + a_2 \sin\frac{n\pi}{3}\right). \end{aligned}$$

The first term is the desired solution of the differential equation. The second term arises because the first order differential equation is discretized with the help of the third order difference equation. The behaviour of the extraneous solution is exactly opposite to that of the analytic solution. This term oscillates and grows at an exponential rate and no matter how small initially, it over shadows the first term if $x_n = nh$ is sufficiently large. The roots of the reduced characteristic equation satisfy the root condition. Hence, the method is weakly stable.

5.38 (*a*) Show that if the trapezoidal rule is applied to the equation $y' = \lambda y$, where λ is an arbitrary complex constant with negative real part, then for all $h > 0$

$$|\, y_n \,| < |\, y_0 \,|, \; n = 1, 2, 3 \ldots$$

(*b*) Show that if **A** is a negative definite symmetric matrix, then a similar conclusion holds for the application of the trapezoidal rule to the system $\mathbf{y}' = \mathbf{A}\mathbf{y}$, $\mathbf{y}(0)$ given, $h > 0$.

(Stockholm Univ., Sweden, BIT 5(1965), 68)

Solution

(*a*) The trapezoidal rule is

$$y_{n+2} = y_n + \frac{h}{2}(y'_{n+1} + y'_n), \quad n = 0, 1, 2 \ldots$$

Substituting $y' = \lambda y$, we obtain

$$y_{n+1} = \left[\frac{1 + \lambda h/2}{1 - \lambda h/2}\right] y_n, \quad n = 0, 1, 2 \ldots$$

Hence, the growth factor is given by

$$[1 + (\lambda h / 2)] / [1 - (\lambda h / 2)].$$

Setting $(\lambda h / 2) = u + iv$ where u and v are real with $u < 0$, we get

$$| y_{n+1} | = \left| \frac{(1+u)+iv}{(1-u)-iv} \right| | y_n | = \left[\frac{1+u^2+v^2+2u}{1+u^2+v^2-2u} \right]^{1/2} | y_n |$$

Since $u < 0$, the growth factor is always less than one. Hence

$$| y_n | < | y_0 |, \; n = 1, 2, \ldots$$

(*b*) Since **A** is a negative definite symmetric matrix, the eigenvalues of **A** are real, negative and distinct. We define the matrix

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]$$

formed by the eigenvectors of **A**.

Using the transformation $\mathbf{y} = \mathbf{Yz}$, we get for the system

$$\mathbf{y}' = \mathbf{Ay}$$

$$\mathbf{z}' = \mathbf{Dz}$$

where $\mathbf{D} = \mathbf{Y}^{-1}\mathbf{AY}$ is the diagonal matrix with the eigenvalues located on the diagonal.

We get similar conclusion as in (*a*), since λ is an eigenvalue of **A**.

Predictor-Corrector Methods

5.39 Let a linear multistep method for the initial value problem

$$y' = f(x, y), \; y(0) = y_0$$

be applied to the test equation $y' = -y$. If the resulting difference equation has at least one characteristic root $\alpha(h)$ such that $| \alpha(h) | > 1$ for arbitrarily small values of h, then the method is called weakly stable. Which of the following methods are weakly stable ?

(*a*) $y_{n+1} = y_{n-1} + 2hf(x_n, y_n)$.

(*b*) $\bar{y}_n = -y_n + 2y_{n-1} + 2hf(x_n, y_n)$

$y_{n+1} = y_{n-1} + 2hf(x_n, \bar{y}_n)$.

(*c*) $\bar{y}_{n+1} = -4y_n + 5y_{n-1} + 2h(2f_n + f_{n-1})$

$y_{n+1} = y_{n-1} + \frac{1}{3}h[f(x_{n+1}, \bar{y}_{n+1}) + 4f_n + f_{n-1}]$

$f_i = f(x_i, y_i)$. (Gothenburg Univ., Sweden, BIT 8(1968), 343)

Solution

Apply the given methods to $y' = -y$.

(*a*) The difference equation is obtained as

$$y_{n+1} = y_{n-1} - 2hy_n.$$

The characteristic equation is given by

$$\xi^2 + 2h\xi - 1 = 0.$$

Setting $\xi = (1 + z) / (1 - z)$, the transformed characteristic equation is obtained as

$$-hz^2 + 2z + h = 0.$$

Applying the Routh-Hurwitz criterion, we find that there is at least one root which lies in the right half plane of the z-plane or there is atleast one root of the characteristic equation which is greater than one.

The reduced characteristic equation is $\xi^2 - 1 = 0$, whose roots are 1, –1. Hence, the method is weakly stable.

(*b*) The difference equation is obtained as

$$\bar{y}_n = -y_n + 2y_{n-1} - 2hy_n,$$
$$y_{n+1} = y_{n-1} - 2h\bar{y}_n.$$

The composite scheme is given by

$$y_{n+1} = y_{n-1} - 2h[-y_n + 2y_{n-1} - 2hy_n]$$

or
$$y_{n+1} - (2h + 4h^2)y_n - (1 - 4h)y_{n-1} = 0.$$

The characteristic equation is given by

$$\xi^2 - (2h + 4h^2)\xi - (1 - 4h) = 0.$$

The reduced characteristic equation is $\xi^2 - 1 = 0$, whose roots are 1, – 1.

Setting $\xi = (1 + z) / (1 - z)$, the transformed characteristic equation is obtained as

$$(3h + 2h^2)z^2 + 2(1 - 2h)z + h(1 - 2h) = 0.$$

The Routh-Hurwitz criterion requires

$$3h + 2h^2 > 0,\ 1 - 2h > 0 \quad \text{and} \quad h(1 - 2h) > 0.$$

We obtain the condition as $h < 1 / 2$. The method is absolutely stable for $h < 1 / 2$.

Hence, the method is not weakly stable.

(*c*) The difference equation is obtained as

$$\bar{y}_{n+1} = -4y_n + 5y_{n-1} - 2h(2y_n + y_{n-1})$$
$$y_{n+1} = y_{n-1} - \frac{h}{3}(\bar{y}_{n+1} + 4y_n + y_{n-1})$$

Eliminating $\bar{y}_{n+1}$, we obtain the difference equation

$$y_{n+1} = y_{n-1} + \frac{2h^2}{3} y_n - \frac{h}{3}(6 - 2h)y_{n-1}.$$

The characteristic equation is given by

$$\xi^2 - \frac{2}{3} h^2 \xi - \left(1 - 2h + \frac{2}{3} h^2\right) = 0$$

Setting $\xi = (1 + z) / (1 - z)$, the transformed characteristic equation is obtained as

$$h z^2 + 2\left(1 - h + \frac{1}{3} h^2\right) z + \left(h - \frac{2}{3} h^2\right) = 0.$$

The Routh-Hurwitz criterion requires

$$h > 0,\ 1 - h + \frac{1}{3} h^2 > 0 \quad \text{and} \quad h\left(1 - \frac{2}{3} h\right) > 0.$$

The third inequality gives $h < 3 / 2$, which also satisfies the second equation. The method is absolutely stable for $h < 3 / 2$.

Hence, the method is not weakly stable.

The method in (*a*) is the only weakly stable method.

5.40 Find the characteristic equations for the PECE and P(EC)^{2}E methods of the P-C set

$$y^*_{n+1} = y_n + \frac{h}{2}(3y_n' - y_{n-1}'),$$
$$y_{n+1} = y_n + \frac{h}{2}(y^{*\prime}_{n+1} + y_n').$$

when applied to the test equation $y' = \lambda y,\ \lambda < 0$.

Solution

We apply the P-C set to the test equation $y' = \lambda y$ and get

$$P: \quad y^*_{n+1} = y_n + \frac{\lambda h}{2}(3y_n - y_{n-1}),$$

$$E: \quad y^{*\prime}_{n+1} = \lambda y^*_{n+1},$$

$$C: \quad y_{n+1} = y_n + \frac{\lambda h}{2}(y^*_{n+1} + y_n) = \left(1 + \bar{h} + \frac{3}{4}\bar{h}^2\right) y_n - \frac{\bar{h}^2}{4} y_{n-1},$$

$$E: \quad y'_{n+1} = \lambda y_{n+1},$$

where $\bar{h} = \lambda h$.

The characteristic equation of the *PECE* method is obtained as

$$\xi^2 - \left(1 + \bar{h} + \frac{3}{4}\bar{h}^2\right)\xi + \frac{\bar{h}^2}{4} = 0.$$

The next corrector iteration is given by

$$y_{n+1} = y_n + \frac{\bar{h}}{2}\left[\left(1 + \bar{h} + \frac{3}{4}\bar{h}^2\right) y_n - \frac{\bar{h}^2}{4} y_{n-1} + y_n\right]$$

$$= \left(1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{3}{8}\bar{h}^3\right) y_n - \frac{\bar{h}^3}{8} y_{n-1}.$$

The characteristic equation is given by

$$\xi^2 - \left(1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{3}{8}\bar{h}^3\right)\xi + \frac{\bar{h}^3}{8} = 0.$$

5.41 Apply the *P-C* set

$$P: \quad y_{n+1} = y_n + h f_n,$$

$$C: \quad y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$$

to the test problem $y' = -y$, $y(0) = 1$.

(*a*) Determine an explicit expression for y_n obtained using P(EC)^{m}E algorithm.

(*b*) For which values of h, when the corrector is iterated to converge, is the sequence $\{y_n\}_0^\infty$ bounded ?

(*c*) Show also that the application of the corrector more than twice does not improve the results.

Solution

(*a*) The P(EC)^{m}E method can be written as

$$y^{(0)}_{n+1} = y_n + h f_n,$$

$$y^{(s)}_{n+1} = y_n + \frac{h}{2}\left(f^{(s-1)}_{n+1} + f_n\right), \quad s = 1(1)m,$$

$$y_{n+1} = y^{(m)}_{n+1},$$

where $\quad f^{(s-1)}_{n+1} = f(x_{n+1}, y^{(s-1)}_{n+1})$.

When applied to the test problem $y' = -y$, the above P(EC)^{m}E scheme becomes

$$y_{n+1}^{(0)} = (1 - h)y_n,$$

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}[-(1-h)y_n - y_n] = \left(1 - h + \frac{h^2}{2}\right)y_n,$$

$$y_{n+1}^{(2)} = y_n + \frac{h}{2}\left[-\left(1 - h + \frac{h^2}{2}\right)y_n - y_n\right]$$

$$= \left(1 - h + \frac{h^2}{2} - \frac{h^2}{2^2}\right)y_n,$$

....

$$y_{n+1}^{(m)} = \left(1 - h + \frac{h^2}{2} - \frac{h^2}{4} + \ldots + \frac{(-h)^{m+1}}{2^m}\right)y_n$$

$$= [1 - h\{1 + (-p) + (-p)^2 + \ldots + (-p)^m\}]y_n$$

$$= \left[1 - 2p\left\{\frac{1-(-p)^{m+1}}{1-(-p)}\right\}\right]y_n = \frac{1}{1+p}[1 - p - 2(-p)^{m+2}]y_n$$

where $p = h / 2$.

Therefore, we have

$$y_{n+1} = \left(\frac{1 - p - 2(-p)^{m+2}}{1+p}\right)y_n.$$

The solution of the first order difference scheme satisfying the initial condition becomes

$$y_n = \left(\frac{1 - p - 2(-p)^{m+2}}{1+p}\right)^n.$$

(*b*) If the corrector is iterated to converge, *i.e.*, $m \to \infty$, the last equation in (*a*) will converge if $p < 1$, or $0 < h < 2$, which is the required condition.

(*c*) The analytic solution is $y(x) = e^{-x}$,

so that $y(x_{n+1}) = e^{-h}\, y(x_n)$.

The local truncation error is given by

$$T_{n+1} = y_{n+1} - y(x_{n+1})$$

$$= \left[\frac{1 - (h/2) - 2(-h/2)^{m+2}}{1 + (h/2)} - e^{-h}\right]y(x_n)$$

We find that

$$\frac{1 - (h/2) - 2(-h/2)^{m+2}}{1 + (h/2)} - e^{-h}$$

becomes $-\dfrac{1}{2}h^2 + O(h^3)$ for 0 corrector,

$-\dfrac{1}{6}h^3 + O(h^4)$ for 1 corrector,

$$\frac{1}{12}\, h^3 + O(h^4) \quad \text{for 2 correctors,}$$

$$\frac{1}{12}\, h^3 + O(h^4) \quad \text{for 3 correctors.}$$

We thus notice that the application of the corrector more than twice does not improve the result because the minimum local truncation error is obtained at this stage.

5.42 Obtain the PM_pCM_c algorithm for the P-C set

$$y_{n+1} = y_n + \frac{h}{2}\,(3y_n' - y_{n-1}')$$

$$y_{n+1} = y_n + \frac{h}{2}\,(y_{n+1}' + y_n')$$

Find the interval of absolute stability when applied to $y' = \lambda y,\ \lambda < 0$.

Solution

First, we obtain the truncation error of the P-C set.

$$T_{n+1}^{(P)} = y(x_{n+1}) - y(x_n) - \frac{h}{2}\,[3y'(x_n) - y'(x_{n-1})] = \frac{5}{12}\, h^3 y'''(x_n) + \ldots$$

or $$y(x_{n+1}) - y_{n+1}^{(P)} = \frac{5}{12}\, h^3 y'''(x_n) + \ldots$$

and $$T_{n+1}^{(C)} = y(x_{n+1}) - y(x_n) - \frac{h}{2}\,[y'(x_{n+1}) + y'(x_n)] = -\frac{1}{12}\, h^3 y'''(x_n) + \ldots$$

or $$y(x_{n+1}) - y_{n+1}^{(C)} = -\frac{1}{12} h^3 y'''(x_n) + \ldots$$

Comparing with (5.57), we get $C_{p+1}^{(P)} = 5/12$, $C_{p+1}^{(C)} = -1/12$,

and $$[C_{p+1}^{(c)} - C_{p+1}^{(P)}]^{-1} = -2.$$

From (5.58), we now write $PM_p\ CM_c$ algorithm as

$$p_{n+1} = y_n + \frac{h}{2}(3y_n' - y_{n-1}'),$$

$$m_{n+1} = p_{n+1} - \frac{5}{6}(p_n - c_n),$$

$$c_{n+1} = y_n + \frac{h}{2}(m_{n+1}' + y_n'),$$

$$y_{n+1} = c_{n+1} + \frac{1}{6}(p_{n+1} - c_{n+1}).$$

Applying the method on $y' = \lambda y$, and substituting $p_n = b_1\xi^n$, $c_n = b_2\xi^n$, $m_n = b_3\xi^n$ and $y_n = b_4\xi^n$ into the PM_pCM_c algorithm we obtain

$$b_1\xi^2 = \left(\xi + \frac{\bar{h}}{2}\,(3\xi - 1)\right) b_4,$$

$$b_3\xi = \left(\xi - \frac{5}{6}\right) b_1 + \frac{5}{6} b_2,$$

$$b_2\xi = \left(1 + \frac{\bar{h}}{2}\right) b_4 + \frac{\bar{h}}{2}\,\xi\, b_3,$$

$$b_4 = b_2 + \frac{1}{6}b_1 - \frac{1}{6}b_2 = \frac{5}{6}b_2 + \frac{1}{6}\, b_1,$$

where b_i, $i = 1(1)4$ are arbitrary parameters and $\overline{h} = \lambda h$.

Eliminating b_i's we get the characteristic equation as

$$\begin{vmatrix} \xi^2 & 0 & 0 & \theta - \xi(1+3\theta) \\ 6\xi - 5 & 5 & -6\xi & 0 \\ 0 & \xi & -\theta\xi & -(1+\theta) \\ -1 & -5 & 0 & 6 \end{vmatrix} = 0,$$

where $\theta = \overline{h}\ / 2$. Expanding the determinant, we obtain

$$6\xi^3 - 3\,(5\theta^2 + 6\theta + 2)\xi^2 + 2(3\theta + 10\theta^2)\,\xi - 5\theta^2 = 0.$$

Setting $\xi = (1 + z) / (1 - z)$, the transformed equation is obtained as

$$v_0 z^3 + v_1 z^2 + v_2 z + v_3 = 0,$$

where $\quad v_0 = 4(3 + 6\theta + 10\theta^2)$, $v_1 = 4(6 + 3\theta - 5\theta^2)$,

$v_2 = 4(3 - 6\theta - 5\theta^2)$, $v_3 = -12\theta$.

The Routh-Hurwitz criterion requires $v_i > 0$ and $v_1 v_2 - v_0 v_3 > 0$.

Since $\theta < 0$, we find $v_0 > 0_1, v_4 > 0$ for all θ. We use Newton-Raphson method to find $\theta < 0$ satisfying $v_1 v_2 - v_0 v_3 > 0$. We obtain $-0.6884 \le \theta < 0$. For these values of θ, we find that $v_1 > 0$, $v_2 > 0$.

Hence, the second order modified Adams predictor-corrector (PM_pCM_c) method is absolutely stable for all θ in the interval $[-0.6884, 0)$.

5.43 The formulas

$$y^*_{n+1} = y_n + \frac{h}{24}(55y'_n - 59y'_{n-1} + 37y'_{n-2} - 9y'_{n-3}) + T_1$$

$$y_{n+1} = y_n + \frac{h}{24}(9y^{*\prime}{}_{n+1} + 19y'_n - 5y'_{n-1} + y'_{n-2}) + T_2$$

may be used as a P-C set to solve $y' = f(x, y)$. Find T_1 and T_2 and an estimate of the truncation error of the P-C set. Construct the corresponding modified P-C set.

Solution

We have

$$T_1 = y(x_{n+1}) - y(x_n) - \frac{h}{24}\,[55y'(x_n) - 59y'(x_{n-1}) + 37y'(x_{n-2}) - 9y'(x_{n-3})]$$

$$= C_0 y(x_n) + C_1 h\, y'(x_n) + C_2 h^2\, y''(x_n) + C_3 h^3\, y'''(x_n) + C_4 h^4\, y^{(4)}(x_n) + C_5 h^5\, y^{(5)}(x_n) + \ldots$$

$$= \frac{251}{720} h^5 y^{(5)}(\xi_1)$$

where $x_{n-3} < \xi_1 < x_{n+1}$.

Similarly, we obtain

$$T_2 = y(x_{n+1}) - y(x_n) - \frac{h}{24}\,[9y'(x_{n+1}) + 19y'(x_n) - 5y'(x_{n-1}) + y'(x_{n-2})]$$

$$= C_0 y(x_n) + C_1 h\, y'(x_n) + C_2 h^2\, y''(x_n) + C_3 h^3\, y'''(x_n) + C_4 h^4\, y^{(4)}(x_n) + C_5 h^5\, y^{(5)}(x_n) + \ldots$$

$$= -\frac{19}{720} h^5 y^{(5)}(\xi_2).$$

where $x_{n-2} < \xi_2 < x_{n+1}$. The estimate of the truncation error is obtained as follows :

$$y(x_{n+1}) - y^*_{n+1} = \frac{251}{720} h^5 y^{(5)}(x_n) + O(h^6),$$

$$y(x_{n+1}) - y_{n+1} = -\frac{19}{720} h^5 y^{(5)}(x_n) + O(h^6),$$

$$y_{n+1} - y^*_{n+1} = \frac{3}{8} h^5 y^{(5)}(x_n) + O(h^6).$$

Therefore,
$$y(x_{n+1}) = y^*_{n+1} + \frac{251}{270}(y_{n+1} - y^*_{n+1}),$$

$$y(x_{n+1}) = y_{n+1} - \frac{19}{270}(y_{n+1} - y^*_{n+1}).$$

The modified P-C method may be written as

$$p_{n+1} = y_n + \frac{h}{24}(55y'_n - 59y'_{n-1} + 37y'_{n-2} - 9y'_{n-3}),$$

$$m_{n+1} = p_{n+1} - \frac{251}{270}(p_n - c_n),$$

$$c_{n+1} = y_n + \frac{h}{24}[9m'_{n+1} - 19y'_n - 5y'_{n-1} + y'_{n-2}],$$

$$y_{n+1} = c_{n+1} + \frac{19}{270}(p_{n+1} - c_{n+1}).$$

5.44 Which of the following difference methods are applicable for solving the initial value problem.

$$y' + \lambda y = 0,\ y(0) = 1,\ \lambda > 0.$$

For what values of λ are the methods stable ?

(*a*) $y_{n+1} = \frac{1}{2} y_n - \frac{1}{4} y_{n-1} + \frac{h}{3}(2y'_n + y'_{n-1})$

(*b*) $\begin{cases} y^*_{n+1} = y_n + h(2y'_n - y'_{n-1}) & \text{(predictor)} \\ y_{n+1} = y_n + \dfrac{h}{2}(y^{*\prime}_{n+1} + y'_n) & \text{(corrector)} \end{cases}$

using the corrector just once. (Gothenburg Univ., Sweden, BIT 6(1966), 83)

Solution

(*a*) Substituting $y' = -\lambda y$, $\lambda > 0$, in the method we get the difference equation as

$$y_{n+1} - \left(\frac{1}{2} - \frac{2}{3}\bar{h}\right) y_n + \left(\frac{1}{4} + \frac{\bar{h}}{3}\right) y_{n-1} = 0$$

where $\bar{h} = \lambda h$.

The reduced characteristic equation is given by (set $h = 0$)

$$\xi^2 - \frac{1}{2}\xi + \frac{1}{4} = 0$$

whose roots are $(1 \pm i\sqrt{3})/4$.

We have | ξ | = 1 / 2 and the root condition is satisfied. The characteristic equation of the method is given by

$$\xi^2 - \left(\frac{1}{2} - \frac{2}{3}\bar{h}\right)\xi + \left(\frac{1}{4} + \frac{\bar{h}}{3}\right) = 0.$$

Setting $\xi = (1 + z) / (1 - z)$, we get the transformed characteristic equation as

$$v_0 z^2 + v_1 z + v_2 = 0$$

where $$v_0 = \frac{7}{4} - \frac{\bar{h}}{3}, v_1 = 2\left(\frac{3}{4} - \frac{\bar{h}}{3}\right), \; v_2 = \frac{3}{4} + \bar{h}, \bar{h} > 0$$

Routh-Hurwitz criterion is satisfied if $0 < \bar{h} < 9 / 4$. Hence, the method is absolutely stable in this interval.

(*b*) We have

$$P: \quad y^*_{n+1} = (1 - 2\bar{h})y_n + \bar{h}y_{n-1}$$

$$E: \quad y^{*\prime}_{n+1} = -\bar{h}y^*_{n+1}$$

$$C: \quad y_{n+1} = y_n - \frac{\bar{h}}{2}[(1 - 2\bar{h})y_n + \bar{h}y_{n-1} + y_n] = (1 - \bar{h} + \bar{h}^2)y_n - \frac{\bar{h}^2}{2}y_{n-1}$$

$$E: \quad y'_{n+1} = -\bar{h}y_{n+1}.$$

where $\bar{h} > 0$.

The characteristic equation of the PECE method is obtained as

$$\xi^2 - (1 - \bar{h} + \bar{h}^2)\xi + \frac{\bar{h}^2}{2} = 0.$$

Setting $\xi = (1 + z) / (1 - z)$, the transformed characteristic polynomial is obtained as

$$\left(2 - \bar{h} + \frac{3}{2}\bar{h}^2\right) z^2 + (2 - \bar{h}^2)\, z + \frac{\bar{h}}{2}(2 - \bar{h}) = 0.$$

The Routh-Hurwitz criterion is satisfied if

$$4 - 2\bar{h} + 3\bar{h}^2 > 0, \; 2 - \bar{h}^2 > 0, \; \bar{h}(2 - \bar{h}) > 0.$$

We obtain $\bar{h}^2 < 2$, or $\bar{h} < \sqrt{2}$ as the required condition.

System of differential equations

5.45 Use the Taylor series method of order two, for step by step integration of the initial value problem

$$y' = xz + 1, \; y(0) = 0,$$
$$z' = -xy, \; z(0) = 1,$$

with $h = 0.1$ and $0 \le x \le 0.2$.

Solution

The second order Taylor series method for the IVP can be written as

$$y_{n+1} = y_n + hy'_n + \frac{h^2}{2}y''_n,$$

$$z_{n+1} = z_n + hz'_n + \frac{h^2}{2}z''_n.$$

Using the differential equations, the second order Taylor series method becomes

$$y_{n+1} = \left(1 - \frac{h^2 x_n^2}{2}\right) y_n + \left(hx_n + \frac{h^2}{2}\right) z_n + h,$$

$$z_{n+1} = \left(-hx_n - \frac{h^2}{2}\right) y_n + \left(1 - \frac{h^2 x_n^2}{2}\right) z_n - \frac{h^2}{2} x_n.$$

With $h = 0.1$, we obtain

$$n = 0, x_0 = 0: \quad y_1 = 0 + \frac{(0.1)^2}{2} + 0.1 = 0.105, \ z_1 = 1.$$

$$n = 1, x_1 = 0.1: \ y_2 = \left(1 - \frac{(0.1)^2 (0.1)^2}{2}\right) 0.105 + \left(0.1 \times 0.1 + \frac{(0.1)^2}{2}\right) 1 + 0.1 = 0.219995.$$

$$z_2 = \left(-(0.1)^2 - \frac{(0.1)^2}{2}\right) 0.105 + \left(1 - \frac{(0.1)^4}{2}\right) - \frac{(0.1)^2}{2} \times 0.1 = 0.997875.$$

Therefore, the required values are

$$y_1 = 0.105, y_2 = 0.219995, z_1 = 1.0, z_2 = 0.997875.$$

5.46 The system

$$y' = z$$
$$z' = -by - az$$

where $0 < a < 2\sqrt{b}$, $b > 0$, is to be integrated by Euler's method with known values. What is the largest step length h for which all solutions of the corresponding difference equation are bounded ? (Royal Inst. Tech., Stockholm, Sweden, BIT 7(1967), 247)

Solution

The application of the Euler method to the system yields

$$y_{n+1} = y_n + hz_n$$
$$z_{n+1} = z_n + h(-by_n - az_n)$$
$$n = 0, 1, 2, \ldots$$

We write the system in the matrix form as

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & h \\ -bh & 1-ah \end{bmatrix} \begin{bmatrix} y_n \\ z_n \end{bmatrix} = \mathbf{Ay}$$

The characteristic equation of $\mathbf{A}$ is given by

$$\xi^2 - (2 - ah)\xi + 1 - ah + bh^2 = 0.$$

Using the transformation $\xi = (1 + z) / (1 - z)$, we obtain the transformed characteristic equation as

$$(4 - 2ah + bh^2)z^2 + 2(a - bh)hz + bh^2 = 0.$$

The Routh-Hurwitz criterion requires

$$4 - 2ah + bh^2 \geq 0, \ a - bh \geq 0, \ bh^2 \geq 0.$$

As $b > 0$, we require

$$(2 - \sqrt{b}\,h)^2 + 2(2\sqrt{b} - a)h \geq 0,$$
$$a - bh \geq 0.$$

Since $0 < a < 2\sqrt{b}$, the conditions will be satisfied if $0 < h \leq a / b$.

5.47 Euler method and Euler-Cauchy (Heun) methods are used for solving the system

$$y' = -kz, \; y(x_0) = y_0$$
$$z' = ky, \; z(x_0) = z_0, \; k > 0.$$

If the numerical method is written as

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_n \\ y_n \end{bmatrix}$$

determine **A** for both the methods. Does there exist a value of h for which the solutions do not grow exponentially as n increases.

Solution

Let $\quad f_1(t, y, z) = -k\,z, f_2(t, y, z) = k\,y.$

Euler method gives

$$y_{n+1} = y_n + h\,f_1(t_n, y_n, z_n) = y_n - hk\,z_n,$$
$$z_{n+1} = z_n + h\,f_2(t_n, y_n, z_n) = z_n + hk\,y_n.$$

In matrix notation, we can write

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_n \\ z_n \end{bmatrix}, \quad \text{where } \mathbf{A} = \begin{bmatrix} 1 & -hk \\ hk & 1 \end{bmatrix}.$$

The eigenvalues of **A** are $\lambda_{1,2} = 1 \pm ihk$.

Since, $|\,\lambda\,| = \sqrt{1 + h^2k^2} > 1$, $y_n \to \infty$ as $n \to \infty$. Hence, Euler method diverges.

Euler-Cauchy method gives

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} y_n \\ z_n \end{bmatrix} + \frac{1}{2}\begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}$$

where
$$K_{11} = hf_1(t_n, y_n, z_n) = -hkz_n.$$
$$K_{21} = hf_2(t_n, y_n, z_n) = hky_n,$$
$$K_{12} = hf_1(t_n + h, y_n + K_{11}, z_n + K_{21}) = -kh(z_n + khy_n),$$
$$K_{22} = hf_2(t_n + h, y_n + K_{11}, z_n + K_{21}) = kh(y_n - kh\,z_n)$$

In matrix notation, we write the system as

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_n \\ z_n \end{bmatrix}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} 1 - (k^2h^2)/2 & -kh \\ kh & 1 - (k^2h^2)/2 \end{bmatrix}.$$

The eigenvalues of **A** are $\lambda_{1,2} = [1 - (k^2h^2) / 2] \pm ikh$.

Since, $|\,\lambda\,| = \sqrt{1 + (k^2h^2)/4} > 1$, $y_n \to \infty$ as $h \to \infty$. Hence, Heun's method also diverges.

Therefore, for both the methods, there does not exist any value of h for which solutions do not grow exponentially as n increases.

5.48 The classical Runge-Kutta method is used for solving the system

$$y' = -kz, \; y(x_0) = y_0$$
$$z' = ky, \; z(x_0) = z_0$$

where $k > 0$ and x, x_0, y, y_0, z and z_0 are real. The step length h is supposed to be > 0. Putting $y_n \approx y(x_0 + nh)$ and $z_n \approx z(x_0 + nh)$, prove that

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_n \\ z_n \end{bmatrix}$$

where **A** is a real 2×2 matrix. Find under what conditions the solutions do not grow exponentially for increasing values of n. (Bergen Univ., Norway, BIT 6(1966), 359)

Solution

We apply the classical fourth order Runge-Kutta method to the system of equations

$$y' = f(x, y, z) = -kz,$$
$$z' = g(x, y, z) = ky.$$

We have for $\alpha = kh$,

$$K_1 = hf(x_n, y_n, z_n) = -khz_n = -\alpha z_n,$$
$$l_1 = hg(x_n, y_n, z_n) = khy_n = \alpha y_n,$$
$$K_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}, z_n + \frac{l_1}{2}\right) = -kh\left(z_n + \frac{1}{2}\alpha y_n\right) = -\alpha z_n - \frac{1}{2}\alpha^2 y_n,$$
$$l_2 = hg\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_1, z_n + \frac{1}{2}l_1\right) = kh\left(y_n + \frac{1}{2}(-\alpha z_n)\right)$$
$$= \alpha y_n - \frac{1}{2}\alpha^2 z_n,$$
$$K_3 = hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_2, z_n + \frac{1}{2}l_2\right) = -kh\left(z_n + \frac{1}{2}\left(\alpha y_n - \frac{1}{2}\alpha^2 z_n\right)\right)$$
$$= -\frac{1}{2}\alpha^2 y_n + \left(-\alpha + \frac{1}{4}\alpha^3\right)z_n,$$
$$l_3 = hg\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_2, z_n + \frac{1}{2}l_2\right) = kh\left(y_n - \frac{1}{2}\alpha z_n - \frac{1}{4}\alpha^2 y_n\right)$$
$$= \left(\alpha - \frac{1}{4}\alpha^3\right)y_n - \frac{1}{2}\alpha^2 z_n,$$
$$K_4 = hf(x_n + h, y_n + K_3, z_n + l_3) = -kh\left(z_n + \left(\alpha - \frac{1}{4}\alpha^3\right)y_n - \frac{1}{2}\alpha^2 z_n\right)$$
$$= \left(-\alpha^2 + \frac{1}{4}\alpha^4\right)y_n + \left(-\alpha + \frac{1}{2}\alpha^3\right)z_n,$$
$$l_4 = hg(x_n + h, y_n + K_3, z_n + l_3) = kh\left(y_n - \frac{1}{2}\alpha^2 y_n + \left(-\alpha + \frac{1}{4}\alpha^3\right)z_n\right)$$
$$= \left(\alpha - \frac{1}{2}\alpha^3\right)y_n + \left(-\alpha^2 + \frac{1}{4}\alpha^4\right)z_n,$$
$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$
$$= \left(1 - \frac{\alpha^2}{2} + \frac{\alpha^4}{24}\right)y_n + \left(-\alpha + \frac{\alpha^3}{6}\right)z_n$$
$$z_{n+1} = z_n + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$$
$$= \left(\alpha - \frac{\alpha^3}{6}\right)y_n + \left(1 - \frac{\alpha^2}{2} + \frac{\alpha^4}{24}\right)z_n$$

Thus, we have the system of difference equations

$$\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 1 - \dfrac{\alpha^2}{2} + \dfrac{\alpha^4}{24} & -\alpha + \dfrac{\alpha^3}{6} \\ \alpha - \dfrac{\alpha^3}{6} & 1 - \dfrac{\alpha^2}{2} + \dfrac{\alpha^4}{24} \end{bmatrix} \begin{bmatrix} y_n \\ z_n \end{bmatrix}$$

The characteristic equation is given by

$$\begin{vmatrix} \left(1 - \dfrac{\alpha^2}{2} + \dfrac{\alpha^4}{24}\right) - \xi & -\alpha + \dfrac{\alpha^3}{6} \\ \alpha - \dfrac{\alpha^3}{6} & \left(1 - \dfrac{\alpha^2}{2} + \dfrac{\alpha^4}{24}\right) - \xi \end{vmatrix} = 0.$$

We obtain

$$\xi = \left(1 - \frac{\alpha^2}{2} + \frac{\alpha^4}{24}\right) \pm i \left(\alpha - \frac{\alpha^3}{6}\right).$$

We have

$$|\ \xi\ |^2 = \left(1 - \frac{\alpha^2}{2} + \frac{\alpha^4}{24}\right)^2 + \left(\alpha - \frac{\alpha^3}{6}\right)^2 = \frac{1}{576}\ (576 - 8\alpha^6 + \alpha^8).$$

Now, $|\ \xi\ |^2 \le 1$ gives $|\ 576 - 8\alpha^6 + \alpha^8\ | \le 576$,

or $$-576 \le 576 - 8\alpha^6 + \alpha^8 \le 576.$$

The right inequality gives $\alpha^2 \le 8$. The left inequality is satisfied for these values of α. Hence, for $0 < \alpha^2 \le 8$, the solutions do not grow exponentially for increasing values of n.

5.49 The solution of the system of equations

$$y' = u,\ \ y(0) = 1,$$
$$u' = -4y - 2u, \quad u(0) = 1,$$

is to be obtained by the Runge-Kutta fourth order method. Can a step length $h = 0.1$ be used for integration. If so find the approximate values of $y(0.2)$ and $u(0.2)$.

Solution

We have

$$K_1 = hu_n,$$
$$l_1 = h(-4y_n - 2u_n),$$
$$K_2 = h\left[u_n + \frac{1}{2}(-4hy_n - 2hu_n)\right] = -2h^2 y_n + (h - h^2)u_n,$$
$$l_2 = (-4h + 4h^2)\,y_n - 2hu_n,$$
$$K_3 = (-2h^2 + 2h^3)\,y_n + (h - h^2)u_n,$$
$$l_3 = (-4h + 4h^2)\,y_n + (-2h + 2h^3)u_n,$$
$$K_4 = (-4h^2 + 4h^3)\,y_n + (h - 2h^2 + 2h^4)u_n,$$
$$l_4 = (-4h + 8h^2 - 8h^4)\,y_n + (-2h + 4h^3 - 4h^4)u_n,$$

$$y_{n+1} = \left(1 - 2h^2 + \frac{4}{3}h^3\right) y_n + \left(h - h^2 + \frac{1}{3}h^4\right) u_n$$

$$u_{n+1} = \left(-4h + 4h^2 - \frac{4}{3}h^4\right) y_n + \left(1 - 2h + \frac{4}{3}h^3 - \frac{2}{3}h^4\right) u_n$$

or
$$\begin{bmatrix} y_{n+1} \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} 1-2h^2+(4/3)h^3 & h-h^2+(1/3)h^4 \\ -4h+4h^2-(4/3)h^4 & 1-2h+(4/3)h^3-(2/3)h^4 \end{bmatrix} \begin{bmatrix} y_n \\ u_n \end{bmatrix}$$

For $h = 0.1$, we obtain
$$\begin{bmatrix} y_{n+1} \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} 0.98133 & 0.09003 \\ -0.36013 & 0.80127 \end{bmatrix} \begin{bmatrix} y_n \\ u_n \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_n \\ u_n \end{bmatrix}.$$

We find that the roots of the characteristic equation $\xi^2 - 1.7826\xi + 0.818733 = 0$, are complex with modulus $|\ \xi\ | \le 0.9048$. Hence, $h = 0.1$ is a suitable step length for the Runge-Kutta method.

We have the following results.

$n = 0,\ x_0 = 0$: $\begin{bmatrix} y_1 \\ u_1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_0 \\ u_0 \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.07136 \\ 0.44114 \end{bmatrix}.$

$n = 1,\ x_1 = 0.1$: $\begin{bmatrix} y_2 \\ u_2 \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_1 \\ u_1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1.07136 \\ 0.44114 \end{bmatrix} = \begin{bmatrix} 1.09107 \\ -0.03236 \end{bmatrix}.$

5.50 Find the values of $y(0.4)$ and $u(0.4)$ for the system of equations
$$y' = 2y + u,\ y(0) = 1,$$
$$u' = 3y + 4u,\ u(0) = 1,$$
using the fourth order Taylor series method with $h = 0.2$.

Solution

We have
$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \text{ where } \mathbf{y} = \begin{bmatrix} y \\ u \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}.$$

Differentiating the given system of equations, we get
$$\mathbf{y}'' = \mathbf{A}\mathbf{y}' = \mathbf{A}^2\mathbf{y} = \begin{bmatrix} 7 & 6 \\ 18 & 19 \end{bmatrix}\mathbf{y},$$
$$\mathbf{y}''' = \mathbf{A}\mathbf{y}'' = \mathbf{A}^3\mathbf{y} = \begin{bmatrix} 32 & 31 \\ 93 & 94 \end{bmatrix}\mathbf{y},$$
$$\mathbf{y}^{iv} = \mathbf{A}\mathbf{y}''' = \mathbf{A}^4\mathbf{y} = \begin{bmatrix} 157 & 156 \\ 468 & 469 \end{bmatrix}\mathbf{y}.$$

Substituting in the fourth order Taylor series method
$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}_n' + \frac{h^2}{2}\mathbf{y}_n'' + \frac{h^3}{6}\mathbf{y}_n''' + \frac{h^4}{24}\mathbf{y}_n^{iv}$$
$$= \left[\mathrm{I} + h\mathbf{A} + \frac{h^2}{2}\mathbf{A}^2 + \frac{h^3}{6}\mathbf{A}^3 + \frac{h^4}{24}\mathbf{A}^4 \right]\mathbf{y}_n .$$

we get
$$\begin{bmatrix} y_{n+1} \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} 1+2h+\frac{7}{2}h^2+\frac{19}{6}h^3+\frac{193}{24}h^4 & h+\frac{4h^2}{2}+\frac{31}{6}h^3+\frac{156}{24}h^4 \\ 3h+\frac{18}{2}h^2+\frac{93}{6}h^3+\frac{468}{24}h^4 & 1+4h+\frac{19}{2}h^2+\frac{94}{6}h^3+\frac{469}{24}h^4 \end{bmatrix} \begin{bmatrix} y_n \\ u_n \end{bmatrix}$$

For $h = 0.2$, we have
$$\begin{bmatrix} y_{n+1} \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} 1.593133 & 0.371733 \\ 1.1152 & 2.3366 \end{bmatrix} \begin{bmatrix} y_n \\ u_n \end{bmatrix},\ n = 0, 1, \ldots$$

Therefore, we obtain

$$n = 0: \quad \begin{bmatrix} y_0 \\ u_0 \end{bmatrix}_{x=0} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \quad \begin{bmatrix} y_1 \\ u_1 \end{bmatrix}_{x=0.2} = \begin{bmatrix} 1.964866 \\ 3.4518 \end{bmatrix}$$

$$n = 1: \quad \begin{bmatrix} y_1 \\ u_1 \end{bmatrix}_{x=0.2} = \begin{bmatrix} 1.964866 \\ 3.4518 \end{bmatrix}; \quad \begin{bmatrix} y_2 \\ u_2 \end{bmatrix}_{x=0.4} = \begin{bmatrix} 4.313444 \\ 10.256694 \end{bmatrix}.$$

5.51 To integrate a system of differential equations

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \ \ \mathbf{y}_0 \text{ is given,}$$

one can use Euler's method as predictor and apply the trapezoidal rule once as corrector, *i.e.*

$$\mathbf{y}^*_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}[\mathbf{f}(x_n, \mathbf{y}_n) + \mathbf{f}(x_{n+1}, \mathbf{y}^*_{n+1})]$$

(also known as Heun's method).

(*a*) If this method is used on $\mathbf{y}' = \mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is a constant matrix, then $\mathbf{y}_{n+1} = \mathbf{B}(h)\mathbf{y}_n$. Find the matrix $\mathbf{B}(h)$.

(*b*) Assume that $\mathbf{A}$ has real eigenvalues λ satisfying $\lambda_i \in [a, b]$, $a < b < 0$. For what values of h is it true that $\lim_{n \to \infty} \mathbf{y}_n = \mathbf{0}$?

(*c*) If the scalar equation $y' = \lambda y$ is integrated as above, which is the largest value of p for which

$$\lim_{h \to 0} \frac{y_n - e^{\lambda x} y_0}{h^p}, \quad x = nh,$$

x fixed, has finite limit ? (Royal Inst. Tech., Stockholm, Sweden, BIT 8(1968), 138)

Solution

(*a*) Applying the Heun method to $\mathbf{y}' = \mathbf{A}\mathbf{y}$, we have

$$\mathbf{y}^*_{n+1} = \mathbf{y}_n + h\mathbf{A}\mathbf{y}_n = (\mathbf{I} + h\mathbf{A})\mathbf{y}_n,$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}[\mathbf{A}\mathbf{y}_n + \mathbf{A}\mathbf{y}^*_{n+1}] = \mathbf{y}_n + \frac{h}{2}[\mathbf{A}\mathbf{y}_n + \mathbf{A}(\mathbf{I} + h\mathbf{A})\mathbf{y}_n]$$

$$= \left(\mathbf{I} + h\mathbf{A} + \frac{h^2}{2}\mathbf{A}^2\right)\mathbf{y}_n.$$

Hence, we have $\mathbf{B}(h) = \mathbf{I} + h\mathbf{A} + \dfrac{h^2}{2}\mathbf{A}^2$.

(*b*) Since $\mathbf{A}$ has real eigenvalues λ_i, $\mathbf{A}^2$ will have real eigenvalues λ_i^2. The stability requirement is $\rho(\mathbf{B}(h)) \le 1$. Hence we require

$$\left| 1 + h\lambda_i + \frac{h^2}{2}\lambda_i^2 \right| \le 1.$$

This condition is satisfied for $h\lambda_i \in (-2, 0)$. Since $\lambda_i \in [a, b]$, $a < b < 0$; $\lim_{n \to \infty} \mathbf{y}_n = \mathbf{0}$ if $0 < h < -2 / a$.

(*c*) Here, we have

$$y_{n+1} = \left(1 + \lambda h + \frac{h^2\lambda^2}{2}\right) y_n, \quad n = 0, 1, 2, \ldots$$

$$y_n = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2}\right)^n y_0.$$

Hence,
$$y_n = \left[e^{\lambda h} - \frac{1}{6}\lambda^3 h^3 + O(h^4)\right]^n y_0 = e^{\lambda n h}\left[1 - \frac{1}{6}\lambda^3 n h^3 + O(h^4)\right] y_0$$
$$= e^{\lambda x_n} y_0 - \frac{1}{6} h^2 \lambda^3 x_n e^{\lambda x_n} y_0 + O(h^4).$$

We find
$$\lim_{h \to 0} \frac{y_n - e^{\lambda x_n} y_0}{h^2} = -\frac{1}{6} y_0 \lambda^3 x_n e^{\lambda x_n}.$$

Therefore, we obtain $p = 2$.

5.52 Consider the problem

$$\mathbf{y}' = \mathbf{A}\mathbf{y}$$
$$\mathbf{y}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -2 & 1 \\ 1 & -20 \end{bmatrix}$$

(*a*) Show that the system is asymptotically stable.

(*b*) Examine the method
$$y_{n+1} = y_n + \frac{h}{2}(3F_{n+1} - F_n)$$
for the equation $y' = F(x, y)$. What is its order of approximation ? Is it stable ? Is it *A*-stable ?

(*c*) Choose step sizes $h = 0.2$ and $h = 0.1$ and compute approximations to $y(0.2)$ using the method in (*b*). Finally, make a suitable extrapolation to $h = 0$. The exact solution is $\mathbf{y}(0.2) = [0.68 \quad 0.036]^T$ with 2 significant digits.

(Gothenburg Univ., Sweden, BIT 15(1975), 335)

Solution

(*a*) The eigenvalues of **A** are given by
$$\begin{vmatrix} -2-\lambda & 1 \\ 1 & -20-\lambda \end{vmatrix} = 0.$$

We obtain $\lambda_1 = -20.055$, $\lambda_2 = -1.945$.

The eigenvectors corresponding to the eigenvalues are
$$\lambda_1 = -20.055 : \quad [-1 \quad 18.055]^T.$$
$$\lambda_2 = -1.945 : \quad [18.055 \quad 1]^T.$$

The analytic solution is given by
$$\mathbf{y} = k_1 \begin{bmatrix} -1 \\ 18.055 \end{bmatrix} e^{-20.055x} + k_2 \begin{bmatrix} 18.055 \\ 1 \end{bmatrix} e^{-1.945x}$$

Satisfying the initial conditions we obtain
$$k_1 = -3.025 \times 10^{-3}, \; k_2 = 0.055.$$

The system is asymptotically stable, since as $x \to \infty$, $y(x) \to 0$.

(*b*) The truncation error may be written as
$$T_{n+1} = y(x_{n+1}) - y(x_n) - \frac{h}{2}[3y'(x_{n+1}) - y'(x_n)] = -h^2 y''(\xi)$$
where $x_n < \xi < x_{n+1}$. Therefore, the method is of order one.

Applying the method to the test equation $y' = \lambda y$, $\lambda < 0$, we get

$$y_{n+1} = [(1 - \bar{h} / 2) / (1 - 3\bar{h} / 2)] y_n$$

where $\bar{h} = \lambda h$.

The characteristic equation is

$$\xi = \frac{1 - \bar{h}/2}{1 - 3\bar{h}/2}.$$

For $\lambda < 0$, we have $\bar{h} < 0$ and the stability condition $| \xi | \le 1$ is always satisfied. Hence, the method is absolutely stable for $\bar{h} \in (-\infty, 0)$. The method is also A-stable as $\lim_{n \to \infty} y_n = 0$ for all $\bar{h} < 0$.

(*c*) The method $\quad y_{n+1} = y_n + \dfrac{h}{2} (3F_{n+1} - F_n)$

when applied to the given system, leads to the equations

$$y_{1,\, n+1} = y_{1,\, n} + \frac{h}{2} [3(-2y_{1,\, n+1} + y_{2,\, n+1}) - (-2y_{1,\, n} + y_{2,\, n})]$$

$$y_{2,\, n+1} = y_{2,\, n} + \frac{h}{2} [3(y_{1,\, n+1} - 20y_{2,\, n+1}) - (y_{1,\, n} - 20y_{2,\, n})]$$

or
$$\begin{bmatrix} 1+3h & -3h/2 \\ -3h/2 & 1+30h \end{bmatrix} \begin{bmatrix} y_{1,\, n+1} \\ y_{2,\, n+1} \end{bmatrix} = \begin{bmatrix} 1+h & -h/2 \\ -h/2 & 1+10h \end{bmatrix} \begin{bmatrix} y_{1,\, n} \\ y_{2,\, n} \end{bmatrix}$$

Inverting the coefficient matrix, we obtain

$$\begin{bmatrix} y_{1,\, n+1} \\ y_{2,\, n+1} \end{bmatrix} = \frac{1}{D(h)} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,\, n} \\ y_{2,\, n} \end{bmatrix}$$

where $\quad D(h) = 1 + 33h + (351 / 4)h^2$, $a_{11} = 1 + 31h + (117 / 4)h^2$,

$a_{12} = h$, $a_{21} = h$, $a_{22} = 1 + 13h + (117 / 4)h^2$.

We have the following results.

$h = 0.2$: $\quad \mathbf{y}(0) = [1 \quad 0]^T$, $\mathbf{y}(0.2) = [0.753 \quad 0.018]^T$.

$h = 0.1$: $\quad \mathbf{y}(0) = [1 \quad 0]^T$, $\mathbf{y}(0.1) = [0.8484 \quad 0.01931]^T$,

$\mathbf{y}(0.2) = [0.720 \quad 0.026]^T$.

The extrapolated values are given by

$$\mathbf{y}(0.2) = 2\mathbf{y}(0.2 \,;\, 0.1) - \mathbf{y}(0.2 \,;\, 0.2)$$

$$= 2 \begin{bmatrix} 0.720 \\ 0.026 \end{bmatrix} - \begin{bmatrix} 0.753 \\ 0.018 \end{bmatrix} = \begin{bmatrix} 0.687 \\ 0.034 \end{bmatrix}$$

Shooting Methods

5.53 Find the solution of the boundary value problem

$$y'' = y + x, \quad x \in [0, 1],$$

$$y(0) = 0, \; y(1) = 0$$

with the shooting method. Use the Runge-Kutta method of second order to solve the initial value problems with $h = 0.2$.

Solution

We assume the solution of the differential equation in the form

$$y(x) = \phi_0(x) + \mu_1\phi_1(x) + \mu_2\phi_2(x)$$

where μ_1, μ_2 are parameters to be determined.

The related initial value problems are given by

$$I: \quad \phi_0'' = \phi_0 + x, \quad \phi_0(0) = 0, \; \phi_0'(0) = 0.$$

$$II: \quad \phi_1'' = \phi_1, \quad \phi_1(0) = 0, \; \phi_1'(0) = 1.$$

$$III: \quad \phi_2'' = \phi_2, \quad \phi_2(0) = 1, \; \phi_2'(0) = 0.$$

The solution satisfies the boundary condition at $x = 0$. We get

$$y(0) = 0 = \phi_0(0) + \mu_1\phi_1(0) + \mu_2\phi_2(0) = 0 + \mu_1(0) + \mu_2$$

which gives $\mu_2 = 0$. Hence, we have

$$y(x) = \phi_0(x) + \mu_1\phi_1(x)$$

and it is sufficient to solve the initial value problems *I* and *II*.

We write these IVP as the following equivalent first order systems

$$I: \begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}' = \begin{bmatrix} V^{(1)} \\ W^{(1)} + x \end{bmatrix}$$

where $W^{(1)} = \phi_0$. The initial conditions are

$$W^{(1)}(0) = 0, \; V^{(1)}(0) = 0$$

$$II: \begin{bmatrix} W^{(2)} \\ V^{(2)} \end{bmatrix}' = \begin{bmatrix} V^{(2)} \\ W^{(2)} \end{bmatrix}$$

where $W^{(2)} = \phi_1$. The initial conditions are

$$W^{(2)}(0) = 0, \; V^{(2)}(0) = 1$$

With $h = 0.2$ and $[W_0^{(1)} \;\; V_0^{(1)}]^T = [0 \;\; 0]^T$,

we have

$$\begin{bmatrix} W_{n+1}^{(1)} \\ V_{n+1}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 + \frac{1}{2}h^2 & h \\ h & 1 + \frac{1}{2}h^2 \end{bmatrix} \begin{bmatrix} W_n^{(1)} \\ V_n^{(1)} \end{bmatrix} + \begin{bmatrix} \frac{1}{2}h^2 \\ h \end{bmatrix} x_n + \begin{bmatrix} 0 \\ \frac{1}{2}h^2 \end{bmatrix}$$

or

$$\begin{bmatrix} W_{n+1}^{(1)} \\ V_{n+1}^{(1)} \end{bmatrix} = \begin{bmatrix} 1.02 & 0.2 \\ 0.2 & 1.02 \end{bmatrix} \begin{bmatrix} W_n^{(1)} \\ V_n^{(1)} \end{bmatrix} + \begin{bmatrix} 0.02 \\ 0.2 \end{bmatrix} x_n + \begin{bmatrix} 0 \\ 0.02 \end{bmatrix}.$$

For $n = 0: \begin{bmatrix} W_1^{(1)} \\ V_1^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.02 \end{bmatrix}$, $\quad n = 1: \begin{bmatrix} W_2^{(1)} \\ V_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.008 \\ 0.0804 \end{bmatrix}$

$n = 2: \begin{bmatrix} W_3^{(1)} \\ V_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0.03224 \\ 0.183608 \end{bmatrix}$, $\quad n = 3: \begin{bmatrix} W_4^{(1)} \\ V_4^{(1)} \end{bmatrix} = \begin{bmatrix} 0.0816064 \\ 0.3337281 \end{bmatrix}$

$n = 4: \begin{bmatrix} W_5^{(1)} \\ V_5^{(1)} \end{bmatrix} = \begin{bmatrix} 0.1659841 \\ 0.5367238 \end{bmatrix}.$

Similarly, for $[W_0^{(2)} \ \ V_0^{(2)}]^T = [0 \quad 1]^T$, we have from

$$\begin{bmatrix} W_{n+1}^{(2)} \\ V_{n+1}^{(2)} \end{bmatrix} = \begin{bmatrix} 1.02 & 0.2 \\ 0.2 & 1.02 \end{bmatrix} \begin{bmatrix} W_n^{(2)} \\ V_n^{(2)} \end{bmatrix},$$

For $n = 0:$ $\begin{bmatrix} W_1^{(2)} \\ V_1^{(2)} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 1.02 \end{bmatrix}$, $n = 1:$ $\begin{bmatrix} W_2^{(2)} \\ V_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.408 \\ 0.0804 \end{bmatrix}$

$n = 2:$ $\begin{bmatrix} W_3^{(2)} \\ V_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0.63224 \\ 1.183608 \end{bmatrix}$, $n = 3:$ $\begin{bmatrix} W_4^{(2)} \\ V_4^{(2)} \end{bmatrix} = \begin{bmatrix} 0.8816064 \\ 1.3337202 \end{bmatrix}$

$n = 4:$ $\begin{bmatrix} W_5^{(2)} \\ V_5^{(2)} \end{bmatrix} = \begin{bmatrix} 1.1659826 \\ 1.5236074 \end{bmatrix}.$

The boundary conditions at $x = 1$ will be satisfied if

$$y(1) = \phi_0(1) + \mu_1\phi_1(1) = 0$$

or

$$\mu_1 = -\frac{\phi_0(1)}{\phi_1(1)} = -\frac{W_5^{(1)}}{W_5^{(2)}} = -0.142355.$$

The solution is given as $y(x) = \phi_0(x) - 0.142355\ \phi_1(x)$

yielding the numerical solution

$$y(0.2) \approx -0.28471 \times 10^{-1},\ y(0.4) \approx -0.500808 \times 10^{-1},$$
$$y(0.6) \approx -0.577625 \times 10^{-1},\ y(0.8) \approx -0.438946 \times 10^{-1},$$
$$y(1.0) \approx 0.$$

Alternative

We write the general solution of the boundary value problem as

$$y(x) = \lambda\phi_0(x) + (1 - \lambda)\ \phi_1(x)$$

and determine λ so that the boundary condition at $x = b = 1$ is satisfied.

We solve the two initial value problems

$$\phi_0'' = \phi_0 + x,\ \phi_0(0) = 0,\ \phi_0'(0) = 0,$$

$$\phi_1'' = \phi_1 + x,\ \phi_1(0) = 0,\ \phi_1'(0) = 1.$$

Using the second order Runge-Kutta method with $h = 0.2$, we obtain the equations (see equations of system 1)

$$\begin{bmatrix} W_{n+1}^{(i)} \\ V_{n+1}^{(i)} \end{bmatrix} = \begin{bmatrix} 1.02 & 0.2 \\ 0.2 & 1.02 \end{bmatrix} \begin{bmatrix} W_n^{(i)} \\ V_n^{(i)} \end{bmatrix} + \begin{bmatrix} 0.02 \\ 0.2 \end{bmatrix} x_n + \begin{bmatrix} 0 \\ 0.02 \end{bmatrix},\ n = 0, 1, 2, 3, 4\ ;\ i = 1, 2.$$

where $W^{(1)} = \phi_0,\ V^{(1)} = \phi_0',\ W^{(2)} = \phi_1,\ V^{(2)} = \phi_1'.$

Using the conditions $i = 1,\ W_0^{(1)} = 0,\ V_0^{(0)} = 0$, we get

$$\begin{bmatrix} W_1^{(1)} \\ V_1^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.02 \end{bmatrix}, \begin{bmatrix} W_2^{(1)} \\ V_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0.008 \\ 0.0804 \end{bmatrix}, \begin{bmatrix} W_3^{(1)} \\ V_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0.03224 \\ 0.183608 \end{bmatrix},$$

$$\begin{bmatrix} W_4^{(1)} \\ V_4^{(1)} \end{bmatrix} = \begin{bmatrix} 0.081606 \\ 0.333728 \end{bmatrix}, \begin{bmatrix} W_5^{(1)} \\ V_5^{(1)} \end{bmatrix} = \begin{bmatrix} 0.165984 \\ 0.536724 \end{bmatrix}.$$

Using the conditions $i = 2$, $W_0^{(2)} = 0$, $V_0^{(2)} = 1$, we get

$$\begin{bmatrix} W_1^{(2)} \\ V_1^{(2)} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 1.04 \end{bmatrix} \begin{bmatrix} W_2^{(2)} \\ V_2^{(2)} \end{bmatrix} = \begin{bmatrix} 0.416 \\ 1.1608 \end{bmatrix}, \begin{bmatrix} W_3^{(2)} \\ V_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0.66448 \\ 1.367216 \end{bmatrix},$$

$$\begin{bmatrix} W_4^{(2)} \\ V_4^{(2)} \end{bmatrix} = \begin{bmatrix} 0.963213 \\ 1.667456 \end{bmatrix} \begin{bmatrix} W_5^{(2)} \\ V_5^{(2)} \end{bmatrix} = \begin{bmatrix} 1.331968 \\ 2.073448 \end{bmatrix}.$$

From (5.96), we obtain

$$\lambda = -\frac{W_5^{(2)}}{W_5^{(1)} - W_5^{(2)}} = 1.142355 \quad (\text{since } \gamma_2 = 0).$$

Hence, we get $\quad y(x) = 1.142355\, W^{(1)}(x) - 0.142355\, W^{(2)}(x)$.

Substituting $x = 0.2, 0.4, 0.6, 0.8$ and 1.0 we get

$$y(0.2) \approx -0.028471,\ y(0.4) \approx -0.0500808,$$
$$y(0.6) \approx -0.0577625,\ y(0.8) \approx -0.0438952,$$
$$y(1.0) \approx 0.$$

5.54 Find the solution of the boundary value problem

$$x^2 y'' - 2y + x = 0,\ x \in [2, 3]$$
$$y(2) = y(3) = 0$$

with the shooting method. Use the fourth order Taylor series method with $h = 0.25$ to solve the initial value problems. Compare with the exact solution

$$y(x) = (19x^2 - 5x^3 - 36) / 38x.$$

Solution

We assume the solution of the boundary value problem as

$$y(x) = \phi_0(x) + \mu_1\phi_1(x) + \mu_2\phi_2(x)$$

where μ_1, μ_2 are parameters to be determined.

The boundary value problem is replaced by the following three initial value problems.

$$I: \quad x^2\phi_0'' - 2\phi_0 + x = 0,\ \phi_0(2) = 0,\ \phi_0'(2) = 0.$$

$$II: \quad x^2\phi_1'' - 2\phi_1 = 0,\ \phi_1(2) = 0,\ \phi_1'(2) = 1.$$

$$III: \quad x^2\phi_2 - 2\phi_2 = 0,\ \phi_2(2) = 1,\ \phi'_2(2) = 0.$$

Using the boundary conditions, we get

$$y(2) = 0 = \phi_0(2) + \mu_1\phi_1(2) + \mu_2\phi_2(2) = 0 + 0 + \mu_2(1)$$

which gives $\mu_2 = 0$.

$$y(3) = 0 = \phi_0(3) + \mu_1\phi_1(3)$$

which gives $\mu_1 = -\phi_0(3) / \phi_1(3)$.

Hence, it is sufficient to solve the systems I and II.

The equivalent systems of first order initial value problems are

$$I: \quad \phi_0(x) = W^{(1)}$$
$$\phi'_0(x) = W^{(1)\prime} = V^{(1)},$$

$$\phi_0'' = V^{(1)\prime} = \frac{2}{x^2} W^{(1)} - \frac{1}{x},$$

and $$\begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}' = \begin{bmatrix} V^{(1)} \\ \dfrac{2}{x^2} W^{(1)} - \dfrac{1}{x} \end{bmatrix}, \text{ with } \begin{bmatrix} W^{(1)}(2) \\ V^{(1)}(2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

II : $$\phi_1(x) = W^{(2)}$$
$$\phi'(x) = W^{(2)\prime} = V^{(2)}$$
$$\phi_1''(x) = V^{(2)\prime} = (2 / x^2)\, W^{(2)},$$

and $$\begin{bmatrix} W^{(2)} \\ V^{(2)} \end{bmatrix}' = \begin{bmatrix} V^{(2)} \\ (2/x^2)\, W^{(2)} \end{bmatrix}, \text{ with } \begin{bmatrix} W^{(2)}(2) \\ V^{(2)}(2) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Denote $\mathbf{y} = [W^{(i)}\ V^{(i)}]^T$, $i = 1, 2$. Then, the Taylor series method of fourth order gives

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}_n' + \frac{h^2}{2}\mathbf{y}_n'' + \frac{h^2}{6}\mathbf{y}_n''' + \frac{h^4}{24}\mathbf{y}_n^{(4)}.$$

I : For the first system, we get

$$h\mathbf{y}' = \begin{bmatrix} hV^{(1)} \\ t\,\{(2/x)\, W^{(1)} - 1\} \end{bmatrix},\ h^2\mathbf{y}'' = \begin{bmatrix} 2t^2\, W^{(1)} - ht \\ 2t^2\, V^{(1)} - 4(t^2/x)\, W^{(1)} + t^2 \end{bmatrix},$$

$$h^3\mathbf{y}''' = \begin{bmatrix} 2\,ht^2\, V^{(1)} - 4t^3\, W^{(1)} + ht^2 \\ (16/x)\, t^3\, W^{(1)} - 8t^3 V^{(1)} - 4t^3 \end{bmatrix},$$

$$h^4\,\mathbf{y}^{(4)} = \begin{bmatrix} 16t^4\, W^{(1)} - 8t^3\, V^{(1)} - 4ht^3 \\ -(80/x)\, t^4\, W^{(1)} + 40t^4 V^{(1)} + 20t^4 \end{bmatrix},\ t = h / x.$$

The Taylor series method becomes

$$\begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}_{n+1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}_n + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where $a_{11} = 1 + t^2 - (2/3)\, t^3 + (2/3)\, t^4$, $a_{12} = h\,[1 + (t^2/3) - (t^3/3)]$,

$$a_{21} = \frac{1}{x}\left[2t - 2t^2 + \frac{8}{3}t^3 - \frac{10}{3}t^4\right],\ a_{22} = 1 + t^2 - \frac{4}{3}t^3 + \frac{5}{3}t^4,$$

$$b_1 = h\left[-\frac{t}{2} + \frac{t^2}{6} - \frac{t^3}{6}\right],\ b_2 = -t + \frac{t^2}{2} - \frac{2}{3}t^3 + \frac{5}{6}t^4.$$

We obtain the following results, with $h = 0.25$.

$x_0 = 2$, $(t / x) = 1/8$, $W_0^{(1)} = 0$, $V_0^{(1)} = 0$.

$a_{11} = 1.014486$, $a_{12} = 0.251139$, $b_1 = -0.015055$,
$a_{21} = 0.111572$, $a_{22} = 1.013428$, $b_2 = -0.118286$.

$W_1^{(1)} = -0.015055$, $V_1^{(1)} = -0.118286$.

$x_1 = 2.25$, $(t / x) = 1/9$.

$a_{11} = 1.011533$, $a_{12} = 0.250914$, $b_1 = -0.013432$,
$a_{21} = 0.089191$, $a_{22} = 1.010771$, $b_2 = -0.105726$,

$W_2^{(1)} = -0.058340$, $V_2^{(1)} = -0.226629$.

$x_2 = 2.5$, $(t / x) = 0.1$.

$$a_{11} = 1.0094,\ a_{12} = 0.25075,\ b_1 = -\ 0.012125,$$
$$a_{21} = 0.072933,\ a_{22} = 1.008833,\ b_2 = -\ 0.095583,$$
$$W_3^{(1)} = -\ 0.127841,\ V_3^{(1)} = -\ 0.328469.$$

$x_3 = 2.75,\ (t\ /\ x) = 1\ /\ 11.$

$$a_{11} = 1.007809,\ a_{12} = 0.250626,\ b_1 = -\ 0.011051,$$
$$a_{21} = 0.060751,\ a_{22} = 1.007377,\ b_2 = -\ 0.087221,$$
$$W_4^{(1)} = -\ 0.222213,\ V_4^{(1)} = -\ 0.425880.$$

II : For the second system, we obtain the system

$$\begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}_{n+1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} W^{(1)} \\ V^{(1)} \end{bmatrix}_n$$

where a_{11}, a_{12}, a_{21} and a_{22} are same as defined in system *I*.

We obtain the following results.

$x_0 = 2,\ (t\ /\ x) = 1\ /\ 8:\qquad W_1^{(2)} = 0.251139,\ V_1^{(2)} = 1.013428.$

$x_1 = 2.25,\ (t\ /\ x) = 1\ /\ 9:\ W_2^{(2)} = 0.508319,\ V_2^{(2)} = 1.046743.$

$x_2 = 2.5,\ (t\ /\ x) = 0.1:\quad W_3^{(2)} = 0.775568,\ V_3^{(2)} = 1.093062.$

$x_3 = 2.75,\ (t\ /\ x) = 1\ /\ 11:\qquad W_4^{(2)} = 1.055574,\ V_4^{(2)} = 1.148242.$

Hence, we obtain $\quad \mu_1 = -\dfrac{W_4^{(1)}}{W_4^{(2)}} = \dfrac{0.222213}{1.055574} = 0.210514.$

We get $y(x) = \phi_0(x) + 0.210514\ \phi_1(x)$.

Setting $x = 2.25, 2.5, 2.75, 3.0$, we obtain

$$y(2.25) = 0.037813,\ y(2.5) = 0.048668,$$
$$y(2.75) = 0.035427,\ y(3.0) = 0.0000001.$$

The error in satistying the boundary condition $y(3) = 0$ is 1×10^{-7}.

5.55 Use the shooting method to solve the mixed boundary value problem

$$u'' = u - 4x\,e^x,\ 0 < x < 1,$$
$$u(0) - u'(0) = -\ 1,\ u(1) + u'(1) = -\ e.$$

Use the Taylor series method

$$u_{j+1} = u_j + hu_j' + \frac{h^2}{2}\ u_j'' + \frac{h^3}{6} u_j'''.$$

$$u_{j+1}' = u_j' + hu_j'' + \frac{h^2}{2} u_j'''$$

to solve the initial value problems. Assume $h = 0.25$. Compare with the exact solution $u(x) = x(1 - x)\,e^x$.

Solution

We assume the solution in the form

$$u(x) = u_0(x) + \mu_1 u_1(x) + \mu_2 u_2(x)$$

where $u_0(x)$, $u_1(x)$ and $u_2(x)$ satisfy the differential equations

$$u_0'' - u_0 = -\,4\,x\,e^x,\ u_1'' - u_1 = 0,$$
$$u_2'' - u_2 = 0.$$

The initial conditions may be assumed as

$$u_0(0) = 0, \quad u_0'(0) = 0,$$
$$u_1(0) = 1, \quad u_1'(0) = 0,$$
$$u_2(0) = 0, \quad u_2'(0) = 1.$$

We solve the three, second order initial value problems

$$u_0'' = u_0 - 4\,x\,e^x, \qquad u_0(0) = 0,\ u_0'(0) = 0,$$
$$u_1'' = u_1. \qquad u_1(0) = 1,\ u_1'(0) = 0,$$
$$u_2'' = u_2, \qquad u_2(0) = 0,\ u_2'(0) = 1$$

by using the given Taylor series method with $h = 0.25$. We have the following results.

(*i*) $i = 0$, $u_{0,\,0} = 0$, $u_{0,0}' = 0$.

Hence, $\quad u_{0,j}'' = u_{0,j} - 4x_j\, e^{x_j},\ \ u_{0,j}''' = u_{0,j}' - 4(x_j + 1)\, e^{x_j},\ \ j = 0, 1, 2, 3.$

$$u_{0,j+1} = u_{0,j} + hu_{0,j}' + \frac{h^2}{2}(u_{0,j} - 4x_j\, e^{x_j}) + \frac{h^3}{6}[u_{0,j}' - 4(x_j + 1)e^{x_j}]$$

$$= \left(1 + \frac{h^2}{2}\right)u_{0,j} + \left(h + \frac{h^3}{6}\right)u_{0,j}' - \left[\frac{2}{3}h^3\,(1 + x_j) + 2h^2 x_j\right]e^{x_j}$$

$$= 1.03125\; u_{0,j} + 0.25260\; u_{0,j}' - (0.13542\, x_j + 0.01042)e^{x_j}.$$

$$u_{0,j+1}' = u_{0,j}' + h\,[u_{0,j} - 4x_j\; e^{x_j}] + \frac{h^2}{2}\,[u_{0,j}' - 4(x_j + 1)e^{x_j}]$$

$$= hu_{0,j} + \left(1 + \frac{h^2}{2}\right)u_{0,j}' - 2[2hx_j + h^2(1 + x_j)]\; e^{x_j}$$

$$= 0.25\; u_{0,j} + 1.03125\; u_{0,j}' - 2(0.5625x_j + 0.0625)e^{x_j}.$$

Therefore, we obtain

$$u_0(0.25) \approx u_{0,\,1} = -\,0.01042, \quad u_0'(0.25) \approx u_{0,\,1}' = -\,0.12500,$$
$$u_0(0.50) \approx u_{0,\,2} = -\,0.09917, \quad u_0'(0.50) \approx u_{0,\,2}' = -\,0.65315,$$
$$u_0(0.75) \approx u_{0,\,3} = -\,0.39606, \quad u_0'(0.75) \approx u_{0,\,3}' = -\,1.83185,$$
$$u_0(1.00) \approx u_{0,\,4} = -\,1.10823, \quad u_0'(1.00) \approx u_{0,\,4}' = -\,4.03895.$$

(*ii*) $i = 1$, $u_{1,\,0} = 1$, $u_{1,0}' = 0$.

$$u_{1,j}'' = u_{1,j},\ u_{1,j}''' = u_{1,j}',\ \ j = 0, 1, 2, 3.$$

$$u_{1,j+1} = u_{1,j} + hu_{1,j}' + \frac{h^2}{2}u_{1,j} + \frac{h^3}{6}u_{1,j}'$$

$$= \left(1 + \frac{h^2}{2}\right)u_{1,j} + \left(h + \frac{h^3}{6}\right)u_{1,j}' = 1.03125\; u_{1,j} + 0.25260\; u_{1,j}'$$

$$u_{1,j+1}' = u_{1,j}' + h\; u_{1,j} + \frac{h^2}{2}\; u_{1,j}'$$

$$= hu_{1,j} + \left(1 + \frac{h^2}{2}\right)u_{1,j}' = 0.25\; u_{1,j} + 1.03125u_{1,j}'$$

Hence,

$$u_1(0.25) \approx u_{1,\,1} = 1.03125, \qquad u_1'(0.25) \approx u_{1,1}' = 0.25,$$
$$u_1(0.50) \approx u_{1,\,2} = 1.12663, \qquad u_1'(0.50) \approx u_{1,2}' = 0.51563,$$

$$u_1(0.75) \approx u_{1,3} = 1.29209, \quad u_1'(0.75) \approx u_{1,3}' = 0.81340,$$
$$u_1(1.00) \approx u_{1,4} = 1.53794, \quad u_1'(1.00) \approx u_{1,4}' = 1.16184.$$

(*iii*) $i = 2$, $u_{2,0} = 0$, $u'_{2,0} = 1$.

$$u''_{2,j} = u_{2,j}, u'''_{2,j} = u'_{2,j}, j = 0, 1, 2, 3.$$

Since the differential equaton is same as for u_1, we get

$$u_{2,j+1} = 1.03125\, u_{2,j} + 0.25260 u'_{2,j}$$
$$u'_{2,j+1} = 0.25\, u_{2,j} + 1.03125 u'_{2,j}$$

Hence, $\quad u_2(0.25) \approx u_{2,1} = 0.25260, \quad u_2'(0.25) \approx u'_{2,1} = 1.03125,$

$$u_2(0.50) \approx u_{2,2} = 0.52099, \quad u_2'(0.50) \approx u'_{2,2} = 1.12663,$$
$$u_2(0.75) \approx u_{2,3} = 0.82186, \quad u_2'(0.75) \approx u'_{2,3} = 1.29208,$$
$$u_2(1.00) \approx u_{2,4} = 1.17393, \quad u_2'(1.00) \approx u'_{2,4} = 1.53792.$$

From the given boundary conditions, we have

$$a_0 = a_1 = 1, b_0 = b_1 = 1, \gamma_1 = -1, \gamma_2 = -e.$$
$$\mu_1 - \mu_2 = -1$$
$$[u_1(1) + u_1'(1)]\,\mu_1 + [u_2(1) + u_2'(1)]\,\mu_2 = -e - [u_0(1) + u_0'(1)]$$

or $\quad 2.69978\mu_1 + 2.71185\,\mu_2 = 2.42890.$

Solving these equations, we obtain $\mu_1 = -0.05229$, $\mu_2 = 0.94771$.

We obtain the solution of the boundary value problem from

$$u(x) = u_0(x) - 0.05229\, u_1(x) + 0.94771\, u_2(x).$$

The solutions at the nodal points are given in the Table 5.1. The maximum absolute error which occurs at $x = 0.75$, is given by

max. abs. error = 0.08168.

Table 5.1. Solution of Problem 5.55.

x_j	*Exact* : $u(x_j)$	u_j
0.25	0.24075	0.17505
0.50	0.41218	0.33567
0.75	0.39694	0.31526
1.00	0.0	– 0.07610

Alternative

Here, we solve the initial value problems

$$u_1'' - u_1 = -4x\, e^x,\ u_1(0) = 0,\ u_1'(0) = -(\gamma_1 / a_1) = 1,$$
$$u_2'' - u_2 = -4x\, e^x,\ u_2(0) = 1,\ u_2'(0) = [(a_0 - \gamma_1)) / a_1] = 2.$$

Using the given Taylor's series method with $h = 0.25$, we obtain (as done earlier)

$$u_{i,j+1} = 1.03125\, u_{i,j} + 0.25260\, u'_{i,j} - (0.13542x_j + 0.01042)e^{x_j}$$
$$u'_{i,j+1} = 0.25\, u_{i,j} + 1.03125\, u'_{i,j} - 2(0.5625x_j + 0.0625)e^{x_j}$$
$$i = 1, 2 \quad \text{and} \quad j = 0, 1, 2, 3.$$

Using the initial conditions, we obtain

$$u_1(0.25) \approx u_{1,1} = 0.24218, \qquad u_1'(0.25) \approx u_{1,1}' = 0.90625,$$
$$u_1(0.50) \approx u_{1,2} = 0.42182, \qquad u_1'(0.50) \approx u_{1,2}' = 0.47348,$$
$$u_1(0.75) \approx u_{1,3} = 0.42579, \qquad u_1'(0.75) \approx u_{1,3}' = -0.53976,$$
$$u_1(1.00) \approx u_{1,4} = 0.06568, \qquad u_1'(1.00) \approx u_{1,4}' = -2.50102.$$
$$u_2(0.25) \approx u_{2,1} = 1.52603, \qquad u_2'(0.25) \approx u_{2,1}' = 2.18750,$$
$$u_2(0.50) \approx u_{2,2} = 2.06943, \qquad u_2'(0.50) \approx u_{2,2}' = 2.11573,$$
$$u_2(0.75) \approx u_{2,3} = 2.53972, \qquad u_2'(0.75) \approx u_{2,3}' = 1.56571,$$
$$u_2(1.00) \approx u_{2,4} = 2.77751, \qquad u_2'(1.00) \approx u_{2,4}' = 0.19872.$$

Using the boundary condition at $x = 1$, we obtain, on using (5.98),

$$\lambda = \frac{-e-[u_2(1)+u'_2(1)]}{[u_1(1)+u_1'(1)]-[u_2(1)+u_2'(1)]} = \frac{-5.69451}{-2.43534-2.97623} = 1.05228.$$

Hence, we have

$$u(x) = \lambda u_1(x) + (1-\lambda)u_2(x) = 1.05228\ u_1(x) - 0.05228\ u_2(x)$$

Substituting $x = 0.25$, 0.5, 0.75 and 1.0, we get

$$u(0.25) \approx 0.17506,\ u(0.50) \approx 0.33568,$$
$$u(0.75) \approx 0.31527,\ u(1.00) \approx -0.07609.$$

These values are same as given in the Table except for the round off error in the last digit.

5.56 Use the shooting method to find the solution of the boundary value problem

$$y'' = 6y^2,$$
$$y(0) = 1,\ y(0.5) = 4/9.$$

Assume the initial approximations

$$y'(0) = \alpha_0 = -1.8,\ \ y'(0) = \alpha_1 = -1.9,$$

and find the solution of the initial value problem using the fourth order Runge-Kutta method with $h = 0.1$. Improve the value of $y'(0)$ using the secant method once. Compare with the exact solution $y(x) = 1/(1+x)^2$.

Solution

We use the fourth order Runge-Kutta method to solve the initial value problems :

$$I: \qquad y'' = 6y^2,$$
$$y(0) = 1,\ y'(0) = -1.8,$$
$$II: \qquad y'' = 6y^2,$$
$$y(0) = 1,\ y'(0) = -1.9,$$

and obtain the solution values at $x = 0.5$. We then have

$$g(\alpha_0) = y(\alpha_0\,;b) - \frac{4}{9},\ \ g(\alpha_1) = y(\alpha_1\,;b) - \frac{4}{9}.$$

The secant method gives

$$\alpha_{n+1} = \alpha_n - \left[\frac{\alpha_n - \alpha_{n-1}}{g(\alpha_n) - g(\alpha_{n-1})}\right] g(\alpha_n),\ n = 1, 2, \ldots$$

The solution values are given by

x	$y(0) = 1$, $\alpha_0 = -1.8$	$y(0) = 1$, $\alpha_0 = -1.9$	$y(0) = 1$, $\alpha_{sc} = 1.998853$	$y(0) = 1$, $y'(0) = -2$
0.1	0.8468373	0.8366544	0.8265893	0.8264724
0.2	0.7372285	0.7158495	0.6947327	0.6944878
0.3	0.6605514	0.6261161	0.5921678	0.5917743
0.4	0.6102643	0.5601089	0.5108485	0.5102787
0.5	0.5824725	0.5130607	0.4453193	0.4445383

Difference Methods

5.57 Use the Numerov method with $h = 0.2$, to determine $y(0.6)$, where $y(x)$ denotes the solution of the initial value problem

$$y'' + xy = 0, \quad y(0) = 1, y'(0) = 0.$$

Solution

The Numerov method is given by

$$y_{n+1} - 2y_n + y_{n-1} = \frac{h^2}{12}(y''_{n+1} + 10y''_n + y''_{n-1}), \ n \ge 1.$$

$$= -\frac{h^2}{12}[x_{n+1}\, y_{n+1} + 10x_n\, y_n + x_{n-1}\, y_{n-1}]$$

Solving for y_{n+1}, we get

$$\left[1 + \frac{h^2}{12} x_{n+1}\right] y_{n+1} = 2y_n - y_{n-1} - \frac{h^2}{12}[10x_n y_n + x_{n-1} y_{n-1}].$$

Here, we require the values y_0 and y_1 to start the computation. The Numerov method has order four and we use a fourth order single step method to determine the value y_1. The Taylor series method gives

$$y(h) = y(0) + hy'(0) + \frac{h^2}{2} y''(0) + \frac{h^3}{6} y'''(0) + \frac{h^4}{24} y^{(4)}(0).$$

We have $\quad y(0) = 1, \ y'(0) = 0, \ y''(0) = 0, \ y'''(0) = -1, \ y^{(4)}(0) = 0,$

$$y^{(5)}(0) = 0, \ y^{(6)}(0) = 4.$$

Hence, we obtain

$$y(h) = 1 - \frac{h^3}{6} + \frac{h^6}{180} + \ldots$$

For $h = 0.2$, we get

$$y(0.2) \approx y_1 = 1 - \frac{(0.2)^3}{6} + \ldots \approx 0.9986667.$$

We have the following results, using the Numerov method.

$$n = 1: \quad \left[1 + \frac{h^2}{12}(0.4)\right] y_2 = 2y_1 - y_0 - \frac{h^2}{12}[10(0.2)y_1 + 0]$$

or $$y_2 = \frac{1}{1.0013333}\left[2(0.9986667) - 1 - \frac{0.04}{12}\{2(0.9986667)\}\right] = 0.9893565.$$

$$n = 2 : \quad \left[1 + \frac{h^2}{12}(0.6)\right] y_3 = 2y_2 - y_1 - \frac{h^2}{12}\,[10(0.4)\,y_2 + 0.2\,y_1]$$

or $$y_3 = \frac{1}{1.002}\left[2(0.9893565) - 0.9986667 - \frac{0.04}{12}\{10(0.4)\,0.9893565 + (0.2)\,0.9986667\}\right]$$

$$= 0.9642606.$$

5.58 Solve the boundary value problem

$$y'' + (1 + x^2)\,y + 1 = 0, \; y\,(\pm\,1) = 0$$

with step lengths $h = 0.5, 0.25$ and extrapolate. Use a second order method.

Solution

Replacing x by $-x$, the boundary value problem remains unchanged. Thus, the solution of the problem is symmetrical about the y-axis. It is sufficient to solve the problem in the interval [0, 1]. The nodal points are given by

$$x_n = nh, \; n = 0, 1, 2, \ldots\ldots., N$$

where $Nh = 1$.

The second order method gives the difference equation

$$\frac{1}{h^2}\,[y_{n-1} - 2y_n + y_{n+1}] + (1 + x_n^2)y_n + 1 = 0,$$

or $$-y_{n-1} + [2 - (1 + x_n^2)h^2]\,y_n - y_{n+1} = h^2, \; n = 0, 1, 2, \ldots\ldots., N.$$

The boundary condition gives $y_N = 0$.

For $h = 1/2$, $N = 2$, we have

$$n = 0 : \qquad -y_{-1} + (7/4)\,y_0 - y_1 = 1/4\,,$$

$$n = 1 : \qquad -y_0 + (27/16)\,y_1 - y_2 = 1/4.$$

Due to symmetry $y_{-1} = y_1$ and the boundary condition gives $y_2 = 0$.

The system of linear equations is given by

$$\begin{bmatrix} 7/4 & -2 \\ -1 & 27/16 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

whose solution is $y_0 = 0.967213$, $y_1 = 0.721311$.

For $h = 1/4$, $N = 4$, we have the system of equations

$$\begin{bmatrix} 31/16 & -2 & 0 & 0 \\ -1 & 495/256 & -1 & 0 \\ 0 & -1 & 123/64 & -1 \\ 0 & 0 & -1 & 487/256 \end{bmatrix}\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Using the Gauss-elimination method to solve the system of equations, we obtain

$$y_0 = 0.941518, \; y_1 = 0.880845, \; y_2 = 0.699180, \; y_3 = 0.400390.$$

Using the extrapolation formula

$$y(x) = \frac{1}{3}(4y_{h/2} - y_h),$$

the extrapolated values at $x = 0, 0.5$ are obtained as

$$y_0 = 0.932953, \; y_1 = 0.691803.$$

5.59 Use a second order method for the solution of the boundary value problem

$$y'' = xy + 1, \; x \in [0, 1],$$
$$y'(0) + y(0) = 1, y(1) = 1,$$

with the step length $h = 0.25$.

Solution

The nodal points are $x_n = nh$, $n = 0(1)4$, $h = 1 / 4$, $Nh = 1$. The discretizations of the differential equation at $x = x_n$ and that of the boundary conditions at $x = 0$ and $x = x_N = 1$ lead to

$$-\frac{1}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + x_n y_n + 1 = 0, \; n = 0(1)\,3,$$
$$\frac{y_1 - y_{-1}}{2h} + y_0 = 1, y_4 = 1.$$

Simplifying we get

$$-y_{n-1} + (2 + x_n h^2)y_n - y_{n+1} = -h^2, \; n = 0(1)3$$
$$y_{-1} = 2hy_0 + y_1 - 2h, \quad y_4 = 1.$$

We have the following results.

$$n = 0, x_0 = 0: \quad -y_{-1} + 2y_0 - y_1 = -\frac{1}{16},$$
$$n = 1, x_1 = 0.25: \quad -y_0 + \frac{129}{64}y_1 - y_2 = -\frac{1}{16},$$
$$n = 2, x_2 = 0.5: \quad -y_1 + \frac{65}{32}y_2 - y_3 = -\frac{1}{16};$$
$$n = 3, x_3 = 0.75: \quad -y_2 + \frac{131}{64}y_3 - y_4 = -\frac{1}{16},$$

and $\quad y_{-1} = \frac{1}{2}y_0 + y_1 - \frac{1}{2}, y_4 = 1.$

Substituting for y_{-1} and y_4, we get the following system of equations

$$\begin{bmatrix} 3/2 & -2 & 0 & 0 \\ -1 & 129/64 & -1 & 0 \\ 0 & -1 & 65/32 & -1 \\ 0 & 0 & -1 & 131/64 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = -\frac{1}{16}\begin{bmatrix} 9 \\ 1 \\ 1 \\ -15 \end{bmatrix}$$

Using the Gauss elimination method, we find

$$y_0 = -7.4615, y_1 = -5.3149, \; y_2 = -3.1888, y_3 = -1.0999.$$

5.60 A table of the function $y = f(x)$ is given

x	4	5	6	7
y	0.15024	0.56563	1.54068	3.25434

x	8	9	10
y	5.51438	7.56171	8.22108

It is known that f satisfies the differential equation

$$y'' + \left(1 - \frac{4}{x} - \frac{n(n+1)}{x^2}\right)y = 0$$

where n is a positive integer.

(*a*) Find n.

(*b*) Compute $f(12)$ using Numeröv's method with step size 1.

(Uppsala Univ., Sweden, BIT 8(1968), 343)

Solution

(*a*) The application of the Numeröv method at $x = 5$ gives, with $h = 1$,

$$y_4 - 2y_5 + y_6 = \frac{1}{12}(y_4'' + 10y_5'' + y_6'')$$

or $$12y_4 - 24y_5 + 12y_6 = y_4'' + 10y_5'' + y_6''.$$

We now use the differential equation and the given data to find y_4'', y_5'' and y_6''. We have, $x_4 = 4$, $x_5 = 5$, $x_6 = 6$, and

$$y_4'' = -\left(1 - 1 - \frac{n(n+1)}{16}\right)y_4 = \frac{n(n+1)}{16} y_4 .$$

$$y_5'' = -\left(1 - \frac{4}{5} - \frac{n(n+1)}{25}\right)y_5 = -0.113126 + \frac{1}{25} n(n+1)y_5,$$

$$y_6'' = -\left(1 - \frac{4}{6} - \frac{n(n+1)}{36}\right)y_6 = -0.51356 + \frac{1}{36} n(n+1)y_6.$$

Substituting into the Numeröv method, we get

$$8.36074 = 0.278439\, n(n+1),$$

or $$n(n+1) - 30.027187 = 0.$$

Solving for n, we get

$$n = \frac{-1 + 11.0049}{2} = 5.0025 \approx 5.$$

Hence, we obtain $n = 5$.

(*b*) We take $n = 5$ and apply Numeröv's method at $x = 10$ and 11.

We have at $x = 10$,

$$12y_{11} - 24y_{10} + 12y_9 = y_9'' + 10y_{10}'' + y_{11}''$$

where $$y_{10}'' = -\left(1 - \frac{4}{10} - \frac{30}{100}\right) y_{10} = -2.466324,$$

$$y_9'' = -\left(1 - \frac{4}{9} - \frac{30}{81}\right)y_9 = -1.400317.$$

$$y_{11}'' = -\left(1 - \frac{4}{11} - \frac{30}{121}\right)y_{11} = -0.388430y_{11}.$$

Substituting, we get $12.388430\, y_{11} = 24y_{10} - 12y_9 - 26.063557$.

Simplifying, we get $y_{11} = 6.498147$.

We have at $x = 11$,

$$12y_{12} - 24y_{11} + 12y_{10} = y_{12}'' + 10y_{11}'' + y_{10}''.$$

where $$y_{12}'' = -\left(1 - \frac{1}{3} - \frac{5}{24}\right)y_{12} = -\frac{11}{24} y_{12}.$$

Substituting, we get $$12.458333\, y_{12} = 24y_{11} - 12y_{10} + 10y_{11}'' + y_{10}''$$
$$= 20.1157y_{11} - 12y_{10} + y_{10}''$$

Simplifying, we get $y_{12} = 2.375558$.

5.61 Find difference approximations of the solution $y(x)$ of the boundary value problem

$$y'' + 8(\sin^2 \pi x)y = 0, \; 0 \le x \le 1,$$
$$y(0) = y(1) = 1$$

taking step-lengths $h = 1/4$ and $h = 1/6$. Also find an approximate value for $y'(0)$.

(Chalmer's Inst. Tech., Gothenburg, Sweden, BIT 8(1968), 246)

Solution

The nodal points are given by $x_n = nh$, $n = 0(1)\, N$,

$$Nh = 1.$$

We apply the second order method at $x = x_n$ and obtain

$$y_{n-1} - 2y_n + y_{n+1} + 8h^2 \sin^2 (\pi x_n)\, y_n = 0$$

or
$$-y_{n-1} + [2 - 8h^2 \sin^2(\pi x_n)]\, y_n - y_{n+1} = 0.$$

The boundary conditions become $y_0 = y_N = 1$.

The solution is symmetrical about the point $x = 1/2$. It is sufficient to consider the interval $[0, 1/2]$.

For $h = 1/4$, we have the mesh points as 0, 1 / 4 and 1 / 2.

We have the following difference equations.

$$n = 1: \quad -y_0 + \left(2 - 8 \cdot \frac{1}{16} \cdot \frac{1}{2}\right) y_1 - y_2 = 0, \quad \text{or} \quad \frac{7}{4} y_1 - y_2 = 1.$$

$$n = 2: \quad -y_1 + \left(2 - 8 \cdot \frac{1}{16} \cdot 1\right) y_2 - y_3 = 0, \quad \text{or} \quad -2y_1 + \frac{3}{2} y_2 = 0,$$

since $y_1 \approx y_3$.

Solving, we get $y_1 = 2.4$, $y_2 = 3.2$.

We also find

$$y_0' = \frac{y_1 - y_0}{h} = \frac{2.4 - 1.0}{0.25} = 5.6,$$

which is a first order approximation.

A second order approximation is given by

$$y_0' = \frac{1}{2h} [-3y_0 + 4y_1 - y_2] = 6.8.$$

For $h = 1/6$, we have the mesh points as 0, 1 / 6, 1 / 3 and 1 / 2.

We have the following system of equations

$$\begin{bmatrix} 35/18 & -1 & 0 \\ -1 & 11/6 & -1 \\ 0 & -2 & 16/9 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

where we have incorporated the boundary condition $y_0 = 1$, and have used the symmetric condition $y_2 = y_4$.

The solution is obtained as $y_1 = 1.8773$, $y_2 = 2.6503$, $y_3 = 2.9816$.

A first order approximation to y_0' is given by

$$y_0' = \frac{1}{h} [y_1 - y_0] = 5.2638.$$

A second order approximation to y_0' is given by

$$y_0' = \frac{1}{2h}[-3y_0 + 4y_1 - y_2] = 5.5767.$$

5.62 Determine a difference approximation of the problem

$$\frac{d}{dx}\left[(1+x^2)\frac{dy}{dx}\right] - y = x^2 + 1,$$
$$y(-1) = y(1) = 0.$$

Find approximate value of $y(0)$ using the steps $h = 1$ and $h = 0.5$ and perform Richardson extrapolation. (Royal Inst. Tech., Stockholm, Sweden, BIT 7(1967), 338)

Solution

We write the differential equation as

$$(1 + x^2)\, y'' + 2xy' - y = x^2 + 1.$$

The boundary value problem is symmetric about at $x = 0$. Therefore, it is sufficient to consider the interval [0, 1].

A second order difference approximation is given by

$$\frac{1}{h^2}(1 + x_n^2)(y_{n-1} - 2y_n + y_{n+1}) + \frac{2x_n}{2h}(y_{n+1} - y_{n-1}) - y_n = x_n^2 + 1,$$

or $$(1 + x_n^2 - hx_n)\, y_{n-1} - [2(1 + x_n^2) + h^2]\, y_n + (1 + x_n^2 + hx_n)\, y_{n+1} = h^2(x_n^2 + 1).$$

For $h = 1$, we have only one mesh point as 0. We have the difference approximation as $y_{-1} - 3y_0 + y_1 = 1$, which gives $y_0 = -1/3$, since $y_{-1} = 0 = y_1$.

For $h = 1/2$, we have two mesh points as 0 and 1 / 2. We have the following difference approximations.

$$n = 0,\ x_0 = 0: \quad y_{-1} - \frac{9}{4}y_0 + y_1 = \frac{1}{4}, \qquad \text{or} \qquad -9y_0 + 8y_1 = 1.$$

$$n = 1,\ x_1 = \frac{1}{2}: \quad y_0 - \frac{11}{4}y_1 + \frac{3}{2}y_2 = \frac{5}{16}, \qquad \text{or} \quad 4y_0 - 11y_1 = \frac{5}{4}.$$

Solving, we get $y_0 = -0.3134$, $y_1 = -0.2276$.

The extrapolated value at $x = 0$ is given by

$$y_0 = \frac{1}{3}[4y_{h/2}(0) - y_h(0)] = -0.3068.$$

5.63 Given the boundary value problem

$$(1 + x^2)\, y'' + \left(5x + \frac{3}{x}\right) y' + \frac{4}{3} y + 1 = 0,$$
$$y(-2) = y(2) = 0.6.$$

(*a*) Show that the solution is symmetric, assuming that it is unique.

(*b*) Show that when $x = 0$, the differential equation is replaced by a central condition

$$4y'' + \frac{4}{3}y + 1 = 0$$

(*c*) Discretize the differential equation and the central condition at $x_n = nh$, $n = \pm N, \pm N - 1, \ldots, \pm 1, 0$ and formulate the resulting three point numerical problem. Choose $h = 1$ and find approximate values of $y(0)$, $y(1)$ and $y(-1)$.

(Royal Inst. Tech., Stockholm, Sweden, BIT 18 (1978), 236)

Solution

(*a*) Replacing x by $-x$ in the boundary value problem, we get

$$(1 + x^2)\, y''(-x) + \left(5x + \frac{3}{x}\right) y'(-x) + \frac{4}{3} y(-x) + 1 = 0,$$

$$y(2) = y(-2) = 0.6.$$

The function $y(-x)$ satisfies the same boundary value problem.

Hence, we deduce that the solution is symmetric about $x = 0$.

(*b*) Taking the limits as $x \to 0$, we get

$$\lim_{x \to 0} \left[(1 + x^2)\, y'' + \left(5x + \frac{3}{x}\right) y' + \frac{4}{3} y + 1 \right] = 4y'' + \frac{4}{3} y + 1 = 0.$$

(*c*) The discretization of the differential equation at $x = x_n$ may be written as

$$\frac{1}{h^2} (1 + x_n^2)(y_{n-1} - 2y_n + y_{n+1})$$

$$+ \frac{1}{2h} \left(5x_n + \frac{3}{x_n}\right)(y_{n+1} - y_{n-1}) + \frac{4}{3} y_n + 1 = 0,$$

$$n \neq 0,\ n = \pm 1, \pm 2, \ldots\ldots., \pm(N - 1).$$

At $n = 0$, we get from the central condition

$$\frac{4}{h^2} (y_{-1} - 2y_0 + y_1) + \frac{4}{3} y_0 + 1 = 0.$$

Due to symmetry, we need to consider the discretization of the boundary value problem at $n = 0(1)N - 1$, with the boundary condition $y_N = 0.6$.

For $h = 1$, we have the following difference equations.

$$n = 0,\ x_0 = 0: \quad 4(-2y_0 + 2y_1) + \frac{4}{3} y_0 + 1 = 0,$$

$$n = 1,\ x_1 = 1: \quad (1 + x_1^2)(y_0 - 2y_1 + y_2) + \frac{1}{2}\left(5x_1 + \frac{3}{x_1}\right)(y_2 - y_0) + \frac{4}{3} y_1 + 1 = 0,$$

and $y_2 = 0.6$.

Simplifying, we get

$$\begin{bmatrix} -20/3 & 8 \\ 2 & 8/3 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4.6 \end{bmatrix}$$

The solution is obtained as $y_0 = 1.1684$, $y_1 = 0.8487$.

5.64 We solve the boundary value problem

$$(1 + x^2)\, y'' - y = 1,$$

$$y'(0) = 0,\ y(1) = 0$$

with the band matrix method. The interval (0, 1) is divided into N subintervals of lengths $h = 1 / N$. In order to get the truncation error $O(h^2)$ one has to discretize the equation as well as boundary conditions by central differences. To approximate the boundary condition at $x = 0$, introduce a fictive $x_1 = -h$ and replace $y'(0)$ by a central difference approximation. x_1 is eliminated by using this equation together with the main difference equation at $x = 0$.

(*a*) State the system arisen.

(*b*) Solve the system with $h = 1/3$. Use 5 decimals in calculations.

(*c*) If the problem is solved with $h = 1/4$ we get $y(0) \approx -0.31980$.

Use this result and the one of (*b*) to get a better estimate of $y(0)$.

(Inst. Tech., Linköping, Sweden, BIT 24(1984), 129)

Solution

The nodal points are $x_n = nh,\ n = 0(1)N$.

(*a*) The second order discretization of the boundary condition $y'(0)$ and the differential equation is given by

$$\frac{1}{h}(y_1 - y_{-1}) = 0,$$

$$\frac{1}{h^2}(1 + x_n^2)(y_{n-1} - 2y_n + y_{n+1}) - y_n = 1,$$

and $y_N = 0$.

After simplification we obtain, for

$$n = 0: \qquad (2 + h^2)\, y_0 - 2y_1 = -h^2,$$

$$1 \le n \le N-1: \qquad -(1 + n^2 h^2)\, y_{n-1} + [2 + (2n^2 + 1)h^2] y_n - (1 + n^2h^2) y_{n+1} = -h^2,$$

and $y_N = 0$, with $y_{-1} = y_1$.

(*b*) For $h = 1/3$, we have the system of equations

$$\begin{bmatrix} 19/9 & -2 & 0 \\ -10/9 & 21/9 & -10/9 \\ 0 & -13/9 & 3 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} = -\frac{1}{9}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

whose solution is

$$y_0 = -0.32036,\ y_1 = -0.28260,\ y_2 = -0.17310.$$

(*c*) We have $\qquad y(0, h) = y_0 + c_1 h^2 + O(h^3).$

Therefore,
$$y\left(0, \frac{1}{4}\right) = y(0) + \frac{C_1}{16} + O(h^3),$$

$$y\left(0, \frac{1}{3}\right) = y(0) + \frac{C_1}{9} + O(h^3).$$

Richardson extrapolation (eliminating C_1) gives

$$y(0) \approx [16 y_{1/4}(0) - 9 y_{1/3}(0)]/7 = -0.31908.$$

5.65 (*a*) Find the coefficients a and b in the operator formula

$$\delta^2 + a\delta^4 = h^2 D^2 (1 + b\delta^2) + O(h^8)$$

(*d*) Show that this formula defines an explicit multistep method for the integration of the special second order differential equation $y'' = f(x, y)$.

Prove by considering the case $f(x, y) = 0$ that the proposed method is unstable.

(Stockholm Univ., Sweden, BIT 8(1968), 138)

Solution

(a) We assume that the function $y(x) \in C^{P+1}[a, b]$ for $p \geq 1$. Applying the difference operator on $y(x_n)$, the truncation error T_n is written as

$$T_n = \delta^2 y(x_n) + a\delta^4 y(x_n) - h^2(1 + b\delta^2)\, y''(x_n)$$

We know that

$$\delta^2 y(x_n) = h^2 y''(x_n) + \frac{h^4}{12} y^{(4)}(x_n) + \frac{h^6}{360} y^{(6)}(x_n) + \frac{h^8}{20160} y^{(8)}(x_n) + \ldots$$

$$\delta^4 y(x_n) = h^4 y^{(4)}(x_n) + \frac{h^6}{6} y^{(6)}(x_n) + \frac{h^8}{80} y^{(8)}(x_n) + \ldots$$

$$\delta^2 y''(x_n) = h^2 y^{(4)}(x_n) + \frac{h^4}{12} y^{(6)}(x_n) + \frac{h^8}{360} y^{(8)}(x_n) + \ldots$$

Substituting the expansions in the truncation error, we obtain

$$T_n = C_2 h^2 y''(x_n) + C_4 h^4 y^{(4)}(x_n) + C_6 h^6 y^{(6)}(x_n) + C_8 h^8 y^{(8)}(x_n) + \ldots.$$

where $\quad C_2 = 0,\ C_4 = \dfrac{1}{12} + a - b,\ \ C_6 = \dfrac{1}{360} + \dfrac{a}{6} - \dfrac{b}{12}.$

Setting $\quad C_4 = 0,\ C_6 = 0$ we get $a = 1/20$, $b = 2/15$.

The truncation error is given by

$$T_n = \left(\frac{1}{20160} + \frac{a}{80} - \frac{b}{360}\right) h^8 y^{(8)}(x_n) + O(h^4)$$

$$= \frac{23}{75600} h^8 y^{(8)}(x_n) + O(h^{10})$$

(b) The characteristic equation when $f(x, y) = 0$, is

$$\xi(\xi - 1)^2 + \frac{1}{20}(\xi - 1)^4 = 0$$

whose roots are 1, 1, $-9 \pm 4\sqrt{5}$.

The root condition is not satisfied. Hence, the method is unstable.

5.66 (a) Determine the constants in the following relations :

$$h^{-4}\delta^4 = D^4(1 + a\delta^2 + b\delta^4) + O(h^6),$$

$$hD = \mu\delta + a_1\Delta^3 E^{-1} + (hD)^4(a_2 + a_3\mu\delta + a_4\delta^2) + O(h^7).$$

(b) Use the relations in (a) to construct a difference method for the boundary value problem

$$y^{iv}(x) = p(x)y(x) + q(x)$$

$y(0)$, $y(1)$, $y'(0)$ and $y'(1)$ are given.

The step size is $h = 1/N$, where N is a natural number. The boundary conditions should not be approximated with substantially lower accuracy than the difference equation. Show that the number of equations and the number of unknowns agree.

(Uppsala Univ., Sweden, BIT 8(1968), 59)

Solution

(*a*) Applying the difference operators on $y(x_n)$, we obtain the truncation error at $x = x_n$ as

$$\begin{aligned} T_n^{(2)} &= \delta^4 y(x_n) - h^4(1 + a\delta^2 + b\delta^4) y^{(4)}(x_n) + O(h^{10}) \\ &= C_6 h^6 y^{(6)}(x_n) + C_8 h^8 y^{(8)}(x_n) + O(h^{10}) \end{aligned}$$

where $\quad C_6 = \dfrac{1}{6} - a,\ C_8 = \dfrac{1}{80} - \dfrac{a}{12} - b.$

Setting $C_6 = 0$, $C_8 = 0$, we obtain $a = 1/6$, $b = -1/720$.

Next, we apply the first derivative operator hD on $y(x_n)$ and write as

$$\begin{aligned} T_n^{(1)} &= h\, y'(x_n) - \mu\delta\, y(x_n) - a_1 \Delta^3 y(x_n - h) \\ &\quad - h^4(a_2 + a_3\, \mu\delta + a_4 \delta^2)\, y^{(4)}(x_n) + O(h^7) \\ &= hy'(x_n) - \frac{1}{2}\, [y(x_{n+1}) - y(x_{n-1})] - a_1 [y(x_{n+2}) \\ &\quad - 3y(x_{n+1}) + 3y(x_n) - y(x_{n-1})] - h^4 (a_2 + a_4\delta^2) y^{(4)}(x_n) \\ &\quad - \frac{1}{2}\, h^4 a_3\, [y^{(4)}(x_{n+1}) - y^{(4)}(x_{n-1})] \\ &= C_3 h^3 y^{(3)}(x_n) + C_4 h^4 y^{(4)}(x_n) + C_5 h^5 y^{(5)}(x_n) + C_6 h^6 y^{(6)}(x_n) + O(h^7) \end{aligned}$$

where $\quad C_3 = -\dfrac{1}{6} - a_1,\ C_4 = -\dfrac{a_1}{2} - a_2,$

$$C_5 = -\frac{1}{120} - \frac{a_1}{4} - a_3,\ C_6 = -\frac{a_1}{12} - a_4.$$

Setting $C_3 = C_4 = C_5 = C_6 = 0$, we obtain

$$a_1 = -1/6,\ a_2 = 1/12,\ a_3 = 1/30,\ a_4 = 1/72.$$

(*b*) The difference scheme at $x = x_n$, can be written as

$$\delta^4 y_n = h^4 \left(1 + \frac{1}{6}\delta^2 - \frac{1}{720}\delta^4\right) [p(x_n) y_n + q(x_n)],$$

$$n = 1\,(1)\,N - 1,$$

$$y'(x_n) = h^{-1}\left[\mu\delta y_n - \frac{1}{6}\Delta^3 E^{-1} y_n\right] + h^3\left(\frac{1}{12} + \frac{1}{30}\mu\delta + \frac{1}{72}\delta^2\right) [p(x_n) y_n + q(x_n)],$$

$$n = 0,\ N.$$

When $n = 1$, the first equation contains the unknown y_{-1} outside [0, 1]. This unknown can be eliminated using the second equation at $n = 0$. Similarly, when $n = N - 1$, the first equation contains, y_{N+1} outside [0, 1] which can be eliminated using the second equation at $n = N$. Further, $y(0)$, $y(1)$ are prescribed. Hence, we finally have $(N - 1)$ equations in $N - 1$ unknowns.

5.67 The differential equation $y'' + y = 0$, with initial conditions $y(0) = 0$, $y(h) = K$, is solved by the Numeröv method.

(*a*) For which values of h is the sequence $\{y_n\}_0^\infty$ bounded ?

(*b*) Determine an explicit expression for y_n. Then, compute y_6 when $h = \pi/6$ and $K = 1/2$.

Solution

The Numeröv method

$$y_{n+1} - 2y_n + y_{n-1} = \frac{h^2}{12}(y''_{n+1} + 10y''_n + y''_{n-1})$$

is applied to the equation $y'' = -y$ yielding

$$y_{n+1} - 2By_n + y_{n-1} = 0$$

where $$B = \left(1 - \frac{5}{12}h^2\right)\Big/\left(1 + \frac{1}{12}h^2\right).$$

The characteristic equation is

$$\xi^2 - 2B\xi + 1 = 0$$

whose roots are $$\xi = B \pm \sqrt{B^2 - 1}.$$

(*a*) The solution y_n will remain bounded if

$$B^2 \le 1, \quad \text{or} \quad \left(1 - \frac{5}{12}h^2\right)^2 \le \left(1 + \frac{h^2}{12}\right)^2 \quad \text{or} \quad -\frac{h^2}{6}(6 - h^2) \le 0.$$

Hence, we obtain $0 < h^2 \le 6$.

(*b*) Since, $|B| \le 1$, let $B = \cos\theta$. The roots of the characteristic equation are given by $\xi = \cos\theta \pm i\sin\theta$, and the solution can be written as

$$y_n = C_1 \cos n\theta + C_2 \sin n\theta.$$

Satisfying the initial conditions, we obtain

$$y_0 = C_1 = 0,$$
$$y_1 = K = C_2 \sin\theta, \quad \text{or} \quad C_2 = K/\sin\theta.$$

We have $$y_n = K\frac{\sin n\theta}{\sin\theta}.$$

For $n = 6$, $h = \pi/6$ and $K = 1/2$, we have

$$B = \cos\theta = \left[1 - \frac{5}{12}\cdot\frac{\pi^2}{36}\right]\Big/\left[1 + \frac{1}{12}\cdot\frac{\pi^2}{36}\right] = 0.865984,$$

and $\theta = 0.523682$.

Hence, $$y_6 = \frac{1}{2}\frac{\sin 6\theta}{\sin\theta} \approx -0.0005.$$

5.68 A diffusion-transport problem is described by the differential equation for $x > 0$,

$py'' + Vy' = 0$, $p > 0$, $V > 0$, $p/V << 1$ (and starting conditions at $x = 0$).

We wish to solve the problem numerically by a difference method with stepsize h.

(*a*) Show that the difference equation which arises when central differences are used for y'' and y' is stable for any $h > 0$ but that when p/h is too small the numerical solution contains slowly damped oscillations with no physical meaning.

(*b*) Show that when forward-difference approximation is used for y' then there are no oscillations. (This technique is called upstream differencing and is very much in use in the solution of streaming problems by difference methods).

(*c*) Give the order of accuracy of the method in (*b*).

[Stockholm Univ., Sweden, BIT 19(1979), 552]

Solution

(a) Replacing the derivatives y'' and y' in the differential equation by their central difference approximations, we obtain

$$\frac{p}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + \frac{V}{2h}(y_{n+1} - y_{n-1}) + O(h^2) = 0.$$

Neglecting the truncation error, we get

$$\left(1 + \frac{Vh}{2p}\right) y_{n+1} - 2y_n + \left(1 - \frac{Vh}{2p}\right) y_{n-1} = 0.$$

The characteristic equation is given by

$$\left(1 + \frac{Vh}{2p}\right) \xi^2 - 2\xi + \left(1 - \frac{Vh}{2p}\right) = 0,$$

or
$$(1 + Re)\, \xi^2 - 2\xi + (1 - Re) = 0$$

where $Re = Vh / (2p)$ is called the cell Reynold number.

The roots are given by $\xi = 1$ and $\xi = (1 - Re) / (1 + Re)$.

The solution of the difference equation is given by

$$y_n = C_1 + C_2 \left(\frac{1 - Re}{1 + Re}\right)^n$$

when $(p \; / \; h)$ is too small, that is

Hence, move if $Re >> 1$, then the solution will contain slowly damped oscillations.

(b) Let now the forward difference approximation to y' be used. Neglecting the truncation error we get the difference equation

$$\left(1 + \frac{Vh}{p}\right) y_{n+1} - 2\left(1 + \frac{Vh}{2p}\right) y_n + y_{n-1} = 0,$$

or
$$(1 + 2\,Re) y_{n+1} - 2(1 + Re) y_n + y_{n-1} = 0.$$

The characteristic equation is given by

$$(1 + 2Re)\, \xi^2 - 2\,(1 + Re)\, \xi + 1 = 0,$$

whose roots are $\xi = 1$ and $1 / (1 + 2Re)$. The solution of the difference equation is

$$y_n = A + \frac{B}{(1 + 2Re)^n}\ .$$

Hence, for $Re > 1$, the solution does not have any oscillations.

(c) The truncation error of the difference scheme in (b) is defined by

$$T_n = \left(1 + \frac{Vh}{p}\right) y(x_{n+1}) - 2\left(1 + \frac{Vh}{2p}\right) y(x_n) + y(x_{n-1}).$$

Expanding each term in Taylor's series, we get

$$T_n = \left[\frac{V}{p}\, y'(x_n) + y''(x_n)\right] h^2 + O(h^3)$$

where $x_{n-1} < \xi < x_{n+1}$.

The order of the difference scheme in (b) is one.

5.69 In order to illustrate the significance of the fact that even the boundary conditions for a differential equation are to be accurately approximated when difference methods are used, we examine the differential equation

$$y'' = y,$$

with boundary conditions $y'(0) = 0$, $y(1) = 1$, which has the solution $y(x) = \dfrac{\cosh x}{\cosh (1)}$.

We put $x_n = nh$, assume that $1 / h$ is an integer and use the difference approximation

$$y_n'' \approx (y_{n+1} - 2y_n + y_{n-1}) / h^2.$$

Two different representations for the boundary conditions are

(1) symmetric case : $y_{-1} = y_1$; $y_N = 1$, $N = 1 / h$,

(2) non-symmetric case

$$y_0 = y_1, \; y_N = 1.$$

(*a*) Show that the error $y(0) - y_0$ asymptotically approaches ah^2 in the first case, and bh in the second case, where a and b are constants to be determined.

(*b*) Show that the truncation error in the first case is $O(h^2)$ in the closed interval [0, 1].

[Stockholm Univ., Sweden, BIT 5(1965) 294]

Solution

(*a*) Substituting the second order difference approximation into the differential equation, we get the difference equation

$$y_{n+1} - 2\left(1 + \frac{h^2}{2}\right) y_n + y_{n-1} = 0.$$

The characteristic equation is given by

$$\xi^2 - 2\left(1 + \frac{h^2}{2}\right)\xi + 1 = 0$$

with roots
$$\xi_{1h} = 1 + \frac{h^2}{2} + \left[\left(1 + \frac{h^2}{2}\right)^2 - 1\right]^{1/2}$$

$$= 1 + \frac{h^2}{2} + h\left[1 + \frac{h^2}{4}\right]^{1/2} = 1 + h + \frac{h^2}{2} + \frac{h^3}{8} + \ldots$$

$$= e^h\left(1 - \frac{1}{24}h^3 + O(h^4)\right),$$

$$\xi_{2h} = 1 + \frac{h^2}{2} - \left[\left(1 + \frac{h^2}{2}\right)^2 - 1\right]^{1/2} = 1 - h + \frac{h^2}{2} - \frac{h^3}{8} + \ldots$$

$$= e^{-h}\left(1 + \frac{1}{24}h^3 + O(h^4)\right).$$

The solution of the difference equation may be written as

$$y_n = C_1 e^{nh} \left(1 - \frac{n}{24} h^3 + O(h^4)\right) + C_2 e^{-nh} \left(1 + \frac{n}{24} h^3 + O(h^4)\right)$$

where C_1 and C_2 are arbitrary parameters to be determined with the help of the discretization of the boundary conditions.

(1) Symmetric case : We have $y_{-1} = y_1$. Hence, we obtain

$$C_1 e^h \left(1 - \frac{1}{24} h^3 + O(h^4)\right) + C_2 e^{-h} \left(1 + \frac{1}{24} h^3 + O(h^4)\right)$$

$$= C_1 e^{-h} \left(1 + \frac{1}{24} h^3 + O(h^4)\right) + C_2 e^{h} \left(1 - \frac{1}{24} h^3 + O(h^4)\right)$$

We get $C_1 = C_2$.

Next, we satisfy $y_N = 1$, $Nh = 1$.

$$y_N = C_1 e \left(1 - \frac{1}{24} h^2 + O(h^3)\right) + C_2 e^{-1} \left(1 + \frac{1}{24} h^2 + O(h^3)\right) = 1.$$

Since $C_1 = C_2$, we obtain

$$C_1 = \frac{1}{[2 \cosh (1) - (h^2 / 12) \sinh (1) + O(h^3)]}$$

$$= \frac{1}{2 \cosh (1)} \left[1 - \frac{h^2 \sinh (1)}{24 \cosh (1)} + O(h^3)\right]^{-1}$$

$$= \frac{1}{2 \cosh (1)} \left[1 + \frac{h^2 \sinh (1)}{24 \cosh (1)} + O(h^3)\right]$$

The solution of the difference equation becomes

$$y_n = C_1 \left[2 \cosh x_n - \frac{x_n h^2}{12} \sinh x_n\right].$$

$$y_0 = 2C_1 = \frac{1}{\cosh (1)} + \frac{h^2}{24} \frac{\sinh (1)}{\cosh^2 (1)} + O(h^3).$$

We get from the analytic solution $y(0) = 1 / \cosh (1)$.

Hence, we have $y(0) - y_0 = ah^2$,

where $$a = - \frac{1}{24} \frac{\sinh (1)}{\cosh^2 (1)} = - 0.020565.$$

(2) Non-symmetric case : $y_0 = y_1$, $y_N = 1$.

Satifying the boundary conditions, we obtain

$$C_1 + C_2 = C_1 e^h + C_2 e^{-h} + O(h^3),$$

or $$(e^h - 1)\, C_1 = (1 - e^{-h}) = C_2 + O(h^3),$$

or $$\left[h + \frac{h^2}{2} + O(h^3)\right] C_1 = \left[h - \frac{h^2}{2} + O(h^3)\right] C_2$$

or $$C_1 = \left\{\left(1 + \frac{h}{2}\right)^{-1} \left(1 - \frac{h}{2}\right) + O(h^2)\right\} C_2 = [1 - h + O(h^2)]\, C_2.$$

$$y_N = C_1 e + C_2 e^{-1} + O(h^3),$$

or
$$1 = [(1 - h)\, e + e^{-1} + O(h^2)]\, C_2.$$

Neglecting the error term, we obtain

$$C_2 = \frac{1}{2 \cosh (1) - he} = \frac{1}{2 \cosh (1)} \left[1 - \frac{he}{2 \cosh (1)} \right]^{-1}$$

$$= \frac{1}{2 \cosh (1)} \left[1 + \frac{he}{2 \cosh (1)} \right],$$

where the $O(h^2)$ term is neglected.

$$C_1 = (1 - h) C_2 = \frac{1}{2 \cosh (1)} \left[1 - h + \frac{he}{2 \cosh (1)} \right]$$

where the $O(h^2)$ term is neglected.

Thus, we have
$$y_0 = C_1 + C_2$$

$$= \frac{1}{2 \cosh (1)} \left[2 - h + \frac{he}{\cosh (1)} \right] = \frac{1}{2 \cosh (1)} \left[2 + \frac{h \sinh (1)}{\cosh (1)} \right]$$

$$= \left[\frac{1}{\cosh (1)} + \frac{h \sinh (1)}{2 \cosh^2 (1)} \right]$$

$$y(0) - y_0 = \frac{1}{\cosh (1)} - y_0 = - \frac{h}{2} \left(\frac{\sinh (1)}{2 \cosh^2 (1)} \right) = 12\, ah.$$

We have $\quad bh = 12\, ah, \quad$ or $\quad b = 12a = - 0.24678.$

(*b*) Hence, from (*a*), in the symmetric case the truncation error is $O(h^2)$ while in the nonsymmetric case it is $O(h)$.

5.70 A finite difference approximation to the solution of the two-point boundary value problem

$$y'' = f(x) y + g(x), \; x \in [a, b]$$
$$y(a) = A, \; y(b) = B$$

is defined by

$$- h^{-2} (y_{n-1} - 2y_n + y_{n+1}) + f(x_n) y_n = - g(x_n), \; 1 \le n \le N - 1,$$

and
$$y_0 = A, \; y_N = B,$$

where N is an integer greater than 1, $h = (b - a) / N$, $x_n = a + nh$, and y_n denotes the approximation to $y(x_n)$.

(*i*) Prove that if $f(x) \ge 0$, $x \in [a, b]$ and $y(x) \in C^4 [a, b]$, then

$$| y(x_n) - y_n | \le \frac{h^2}{24} M_4 (x_n - a)(b - x_n)$$

where
$$M_4 = \max_{x \in [a, b]} | y^{(4)} (x) |.$$

(*ii*) Show that with $N = 3$, the difference scheme gives an approximation to the solution of

$$y'' - y = 1, \; x \in [0, 1]$$
$$y(0) = 0, \; y(1) = e - 1,$$

for which $| y(x_n) - y_n | \le \dfrac{e}{864}, \; 0 \le n \le 3.$

Solution

(*i*) The difference equation at $x = x_n$ is defined by

$$-y_{n-1} + 2y_n - y_{n+1} + h^2 f_n y_n = -g_n h^2,\ n = 1(1)\ N-1.$$

Incorporating the boundary conditions $y_0 = A$ and $y_N = B$ into the difference equations, we write the system of equations in matrix notation as

$$\mathbf{Jy} + h^2\ \mathbf{Fy} = \mathbf{D}$$

where
$$\mathbf{J} = \begin{bmatrix} 2 & -1 & & & \mathbf{0} \\ -1 & 2 & -1 & & \\ & \cdots & & & \\ & -1 & 2 & -1 & \\ \mathbf{0} & & -1 & 2 \end{bmatrix},\quad \mathbf{F} = \begin{bmatrix} f_1 & & & \mathbf{0} \\ & f_2 & & \\ \mathbf{0} & & \ddots & \\ & & & f_{N-1} \end{bmatrix}$$

$$\mathbf{y} = [y_1\ y_2\ \ldots\ y_{N-1}]^T,\ \mathbf{D} = [\text{A} - h^2 g_1 \quad - h^2 g_2\ \ldots\ B - h^2 g_{N-1}]^T.$$

Exact solution satisfies the equation

$$\mathbf{Jy}(x_n) + h^2\ \mathbf{F}\ \mathbf{y}(x_n) = \mathbf{D} - \mathbf{T}$$

where $\mathbf{T} = [T_1\ T_2\ \ldots\ T_{N-1}]^T$ is the truncation error.

In order to find the error equation, we put $y_n = y(x_n) + \varepsilon_n$ in the difference equation and obtain

$$\mathbf{J}\boldsymbol{\varepsilon} + h^2\ \mathbf{F}\boldsymbol{\varepsilon} = \mathbf{T}$$

where $\boldsymbol{\varepsilon} = [\varepsilon_1\ \varepsilon_2\ \ldots\ \varepsilon_{N-1}]^T.$

The truncation error is given by

$$T_n = y(x_{n+1}) - 2y(x_n) + y(x_{n-1}) - h^2 f(x_n)\, y(x_n) - h^2 g(x_n)$$

$$= \frac{h^4}{12}\ y^{(4)}\,(\xi),\ \ x_{n-1} < \xi < x_{n+1}.$$

Hence,
$$|\ T_n\ | \le \frac{h^4}{12}\ M_4,\ \ M_4 = \max_{x \in [a, b]}\ |\ y^{(4)}\,(x)\ |.$$

Since $f(x) \ge 0$, $x \in [a, b]$ we have

$$\mathbf{J} + h^2\ \mathbf{F} > \mathbf{J}.$$

The matrices $\mathbf{J}$ and $\mathbf{J} + h^2\,\mathbf{F}$ are irreducibly diagonal dominent with non-positive off diagonal elements and positive diagonal elements. Hence, $\mathbf{J}$ and $\mathbf{J} + h^2\ \mathbf{F}$ are monotone matices.

If follows that $(\mathbf{J} + h^2\mathbf{F})^{-1} < \mathbf{J}^{-1}$.

Hence, we get $\quad \boldsymbol{\varepsilon} = (\mathbf{J} + h^2\ \mathbf{F})^{-1}\ \mathbf{T} \le \mathbf{J}^{-1}\mathbf{T}.$

We now determine $\mathbf{J}^{-1} = (j_{i,\ j})$ explicitly. On multiplying the rows of $\mathbf{J}$ by the jth column of $\mathbf{J}^{-1}$, we have the following difference equations.

(*i*) $-2\,j_{1,j} - j_{2,j} = 0,$

(*ii*) $-j_{i-1,j} + 2\,j_{i,j} - j_{i+1,j} = 0,\quad 2 \le i \le j-1,$

(*iii*) $-j_{j-1,j} + 2\,j_{j,j} - j_{j+1,j} = 1,$

(*iv*) $-j_{i-1,j} + 2\,j_{i,j} - j_{i+1,j} = 0,\quad j+1 \le i \le N-2,$

(*v*) $-j_{N-2,j} + 2\,j_{N-1,j} = 0.$

On solving the difference equations, we get

$$j_{i,j} = \begin{cases} \dfrac{i(N-j)}{N}, & i \le j, \\ \dfrac{j(N-i)}{N}, & i \ge j, \end{cases}$$

Note that the matrix $\mathbf{J}^{-1}$ is symmetric. The row sum of the nth row of $\mathbf{J}^{-1}$ is

$$\sum_{j=1}^{N-1} j_{n,j} = \frac{n(N-n)}{2} = \frac{(x_n - a)(b - x_n)}{2h^2}.$$

Thus, we have

$$|\,\varepsilon_n\,| \le \frac{h^4}{12} M_4 \frac{(x_n - a)(b - x_n)}{2h^2}$$

or $$|\,y(x_n) - y_n\,| \le \frac{h^2}{24} M_4\,(x_n - a)(b - x_n).$$

(*ii*) We are given that

$$N = 3, f(x) = 1, g(x) = 1, A = 0,$$
$$B = e - 1, a = 0, b = 1, h = 1/3.$$

We have $$y^{(4)}(x) = y''(x) = y(x) + 1.$$

Therefore, $$M_4 = \max_{x \in [0,1]} |y^{(4)}(x)\,| = \max_{x \in [0,1]} |\,y(x) + 1\,| = e - 1 + 1 = e.$$

Maximum of $(x_n - a)(b - x_n)$ occurs for $x_n = (a + b)/2$ and its maximum magnitude is $(b - a)^2 / 4 = 1/4$. Hence

$$|\,y(x_n) - y_n\,| \le \frac{e}{864}, \; 0 \le n \le 3.$$

5.71 Consider the homogeneous boundary value problem

$$y'' + \Lambda y = 0,$$
$$y(0) = y(1) = 0.$$

(*a*) Show that the application of the fourth order Numerov method leads to the system

$$\left[\mathbf{J} - \frac{\lambda}{1 + \lambda/12}\mathbf{I}\right]\mathbf{y} = \mathbf{0}$$

where $\lambda = h^2 \Lambda$.

(*b*) Show that the approximation to the eigenvalues by the second and fourth order methods are given by $2(1 - \cos n\pi h) / h^2$ and $12(1 - \cos n\pi h) / [h^2(5 + \cos n\pi h)]$, $1 \le n \le N - 1$, respectively, where $h = 1 / N$.

(*c*) Noticing that $\Lambda_n = n^2 \pi^2$, show that the relative errors

$$\frac{\Lambda_n - h^{-2}\lambda_n}{\Lambda_n}$$

for the second and the fourth order methods are given by $\Lambda_n h^2 / 12$ and $\Lambda_n^2 h^4 / 240$, respectively, when terms of higher order in h are neglected.

Solution

(a) The application of the Numeröv method leads to the system

$$y_{n+1} - 2y_n + y_{n-1} + \frac{\lambda}{12}(y_{n+1} + 10y_n + y_{n-1}) = 0,$$
$$n = 1, 2, \ldots N-1,$$

or
$$-y_{n+1} + 2y_n - y_{n-1} - \frac{\lambda}{1+\lambda/12} y_n = 0,$$
$$n = 1, 2, \ldots, N-1,$$

where $\lambda = h^2\Lambda$.

Incorporating the boundary conditions we obtain

$$\left(\mathbf{J} - \frac{\lambda}{1+\lambda/12}\mathbf{I}\right)\mathbf{y} = \mathbf{0}.$$

(b) For the second order method, we have

$$y_{n+1} - 2[1-(\lambda/2]y_n + y_{n-1} = 0,$$
$$y_0 = y_N = 0.$$

The characteristic equation of the difference equation is given by

$$\xi^2 - 2[1-(\lambda/2)]\,\xi + 1 = 0.$$

Substitute $\cos\theta = 1-(\lambda/2)$. Then the roots of the characteristic equation are given by $\xi = \cos\theta \pm i\sin\theta = e^{\pm i\theta}$. The solution of the difference scheme becomes

$$y_n = C_1 \cos n\theta + C_2 \sin n\theta.$$

Boundary conditions $y(0) = 0$, $y(1) = 0$ lead to $C_1 = 0$, $C_2 \sin(N\theta) = 0$, or $\theta = n\pi/N$. Since $h = 1/N$, we have

$$1 - \frac{1}{2}\lambda_n = \cos\theta = \cos(n\pi h), \quad \text{or} \quad \lambda_n = 2[1-\cos(n\pi h)],$$

or
$$\Lambda_n = \frac{2}{h^2}[1-\cos(n\pi h)].$$

Similarly, for the Numeröv method, we substitute

$$\frac{1-5\lambda/12}{1+\lambda/12} = \cos\theta$$

and find that $\theta = n\pi/N = n\pi h$.

The eigenvalue is given by

$$\left\{\left[1-\frac{5}{12}\lambda_n\right]\Big/\left[1+\frac{1}{12}\lambda_n\right]\right\} = \cos(n\pi h), \text{ or } \quad \lambda_n = \frac{12[1-\cos(n\pi h)]}{5+\cos(n\pi h)}$$

or
$$\Lambda_n = \frac{12}{h^2}\left[\frac{1-\cos(n\pi h)}{5+\cos(n\pi h)}\right].$$

(c) The analytical solution of the eigenvalue problem gives $\Lambda_n = n^2\pi^2$.

We have
$$1-\cos(n\pi h) = 1 - 1 + \frac{1}{2}n^2\pi^2h^2 - \frac{1}{24}n^4\pi^4h^4 + O(h^6)$$
$$= \frac{1}{2}n^2\pi^2h^2\left[1-\frac{1}{12}n^2\pi^2h^2 + O(h^4)\right]$$

For the second order method, we obtain

$$\frac{1}{h^2}\lambda_n = \frac{2}{h^2}[1 - \cos(n\pi h)]$$
$$= n^2\pi^2\left[1 - \frac{1}{12}n^2\pi^2h^2 + O(h^4)\right],$$

Thus, the relative error in the eigenvalue is given by

$$\frac{\Lambda_n - h^{-2}\lambda_n}{\Lambda_n} = \frac{\Lambda_n}{12}h^2 + O(h^4).$$

We have

$$[5 + \cos(n\pi h)]^{-1} = \left[6 - \frac{1}{2}n^2\pi^2h^2 + \frac{1}{24}n^4\pi^4h^4 + O(h^6)\right]^{-1}$$
$$= \frac{1}{6}\left[1 - \left\{\frac{1}{12}n^2\pi^2h^2 - \frac{1}{144}n^4\pi^4h^4 + O(h^6)\right\}\right]^{-1}$$
$$= \frac{1}{6}\left[1 + \frac{1}{12}n^2\pi^2h^2 + O(h^6)\right]$$

$$\frac{[1 - \cos(n\pi h)]}{[5 + \cos(n\pi h)]} = \left(\frac{1}{6}\right)\left(\frac{1}{2}n^2\pi^2h^2\right)\left[1 - \frac{1}{12}n^2\pi^2h^2 + \frac{1}{360}n^4\pi^4h^4 + O(h^6)\right]$$
$$\times\left[1 + \frac{1}{12}n^2\pi^2h^2 + O(h^6)\right] = \frac{1}{12}n^2\pi^2h^2\left[1 - \frac{1}{240}n^4\pi^4h^4 + O(h^6)\right]$$

$$\frac{1}{h^2}\lambda_n = \frac{12}{h^2}\frac{[1 - \cos(n\pi h)]}{[5 + \cos(n\pi h)]} = n^2\pi^2\left[1 - \frac{1}{240}n^4\pi^4h^4 + O(h^6)\right]$$

Therefore, for the Numeröv method, the relative error is given by

$$\frac{\Lambda_n - h^{-2}\lambda_n}{\Lambda_n} = \frac{1}{240}n^2\pi^2h^4 + O(h^6) = \frac{1}{240}\Lambda_n^2h^4 + O(h^6).$$

5.72 Solving the differential equation $y'' = y$, $0 \le x \le 1$, with boundary conditions $y(0) = y(1) = 1$, is associated with minimizing the integral

$$I = \int_0^1 (y'^2 + y^2)\,dx.$$

Find I_{min} using the approximate solution $y = 1 + ax(1 + x)$.

[Lund Univ., Sweden, BIT 29(1989), 158]

Solution

We have
$$I = \int_0^1 [a^2(1 - 2x)^2 + 1 + 2a(x - x^2) + a^2(x - x^2)^2]\,dx$$
$$= \frac{11}{30}a^2 + 1 + \frac{1}{3}a.$$

Setting $dI / da = 0$, we find that the minimum is obtained for $a = -5/11$.

Hence, we get $$I_{min} = \frac{61}{66} = 0.9242.$$

5.73 Solve the boundary value problem

$$y'' + y^2 = 0,$$
$$y(0) = 0,\ y(1) = 1,$$

by minimizing the integral

$$I = \int_0^1 (3y'^2 - 2y^3)\, dx.$$

Use the trial function $y = ax + bx^2$. Compute a and b as well as the minimum value.

[Lund Univ., Sweden, BIT 29(1989), 376]

Solution

The boundary condition $y(1) = 1$, gives $a + b = 1$. Substituting $y(x) = ax + bx^2$, in the integral and simplifying we obtain

$$I = \int_0^1 [3(a + 2bx)^2 - 2\,(ax + bx^2)^3]\, dx$$

$$= 3\left(a^2 + 2ab + \frac{4}{3} b^2\right) - 2\left(\frac{1}{4} a^3 + \frac{3}{5} a^2 b + \frac{1}{2} ab^2 + \frac{1}{7} b^3\right)$$

$$= \frac{1}{70} [a^3 - 66a^2 + 150a - 260],$$

For minimum value of I, we require $dI \,/\, da = 0$, which gives

$$a^2 - 44a + 50 = 0.$$

Solving, we get

$$a = 22 - \sqrt{434} = 1.1673 \text{ and}$$
$$b = 1 - a = -\,0.1673.$$

The other value of a is rejected.

We find $\quad I_{min} = 2.47493.$

5.74 In order to determine the smallest value of λ for which the differential equation

$$y'' = \frac{1}{3 + x} y' - \lambda(3 + x)y,$$
$$y(-\,1) = y(1) = 0,$$

has non-trivial solutions, Runge-Kutta method was used to integrate two first order differential equations equivalent to this equation, but with starting values $y(-\,1) = 0$, $y'(-\,1) = 1$.

Three step lengths h, and three values of λ were tried with the following results for $y(1)$

h \ λ	0.84500	0.84625	0.84750
2 / 10	0.0032252	0.0010348	– 0.0011504
4 / 30	0.0030792	0.0008882	– 0.0012980
1 / 10	0.0030522	0.0008608	– 0.0013254

(a) Use the above table to calculate λ, with an error less than 10^{-5}.

(*b*) Rewrite the differential equation so that classical Runga-Kutta method can be used.

[Inst. Tech., Stockholm, Sweden, BIT 5(1965), 214]

Solution

(*a*) Note from the computed results that $y(1)$ is a function of λ. Denote the dependence as $y(1, \lambda)$.

We now use the Müller method to find the improved value of λ.

The parameter values for various values of h are as follows :

$$\lambda_{k-2} = 0.84500,\ \lambda_{k-1} = 0.84625,\ \lambda_k = 0.84750$$

$$h_k = \lambda_k - \lambda_{k-1} = 0.84750 - 0.84625 = 0.00125.$$

$$h_{k-1} = \lambda_{k-1} - \lambda_{k-2} = 0.84625 - 0.84500 = 0.00125.$$

$$\mu_k = \frac{h_k}{h_{k-1}} = 1.$$

$$\delta_k = 1 + \mu_k = 2.$$

$$g_k = \mu_k^2\, y_{k-2} - \delta_k^2\, y_{k-1} + (\mu_k + \delta_k)\, y_k$$

$$= y_{k-2} - 4y_{k-1} + 3y_k.$$

$$C_k = \mu_k\, (\mu_k y_{k-2} - \delta_k\, y_{k-1} + y_k) = y_{k-2} - 2y_{k-1} + y_k.$$

$$\mu_{k+1} = -\frac{2\,\delta_k\, y_k}{g_k \pm \sqrt{g_k^2 - 4\delta_k\, C_k\, y_k}}$$

The sign in the denominator is chosen as that of g_k.

$$\lambda_{k+1} = \lambda_k + (\lambda_k - \lambda_{k-1})\, \mu_{k+1}.$$

We obtain $h = 2 / 10$: $g_k = -\,0.0043652$, $C_k = 0.0000052$,

$\mu_{k+1} = -\,0.5267473$, $\lambda_{k+1} = 0.8468416$.

$h = 4 / 30$: $g_k = -\,0.0043676$, $C_k = 0.0000048$,

$\mu_{k+1} = -\,0.593989$, $\lambda_{k+1} = 0.8467575$.

$h = 1 / 10$: $g_k = -\,0.0043672$, $C_k = 0.0000052$,

$\mu_{k+1} = -\,0.6065413$, $\lambda_{k+1} = 0.8467418$.

Hence, the eigenvalue is obtained as 0.84674.

(*b*) Substituting $y' = z$ we have two first order differential equations

$$y' = z,$$

$$z' = \frac{1}{(3 + x)}\, z - \lambda(3 + x)y,$$

with initial conditions $y(-\,1) = 0$, $z(-\,1) = 1$.

This system can be used with prescribed λ and h to find $y(1)$.

5.75 Obtain the numerical solution of the nonlinear boundary value problem

$$u'' = \frac{1}{2}\,(1 + x + u)^3$$

$$u'(0) - u(0) = -\,1 / 2,\ u'(1) + u(1) = 1$$

with $h = 1 / 2$. Use a second order finite difference method.

Solution

The nodal points are $x_0 = 0$, $x_1 = 1 / 2$, $x_2 = 1$. We have

$$a_0 = -1,\ a_1 = -1,\ \gamma_1 = -1/2,\ b_0 = b_1 = \gamma_2 = 1.$$

The system of nonlinear equations, using (5.135), (5.137), (5.138), becomes

$$(1 + h)u_0 - u_1 + \frac{h^2}{2}\left[\frac{1}{3}(1 + x_0 + u_0)^3 + \frac{1}{6}(1 + x_1 + u_1)^3\right] - \frac{h}{2} = 0$$

$$-u_0 + 2u_1 - u_2 + \frac{h^2}{2}(1 + x_1 + u_1)^3 = 0$$

$$-u_1 + (1 + h)u_2 + \frac{h^2}{2}\left[\frac{1}{6}(1 + x_1 + u_1)^3 + \frac{1}{3}(1 + x_2 + u_2)^3\right] - h = 0.$$

The Newton-Raphson method gives the following linear equations

$$\begin{bmatrix} \frac{3}{2} + \frac{1}{8}(1 + u_0^{(s)})^2 & -1 + \frac{1}{16}\left(\frac{3}{2} + u_1^{(s)}\right)^2 & 0 \\ -1 & 2 + \frac{3}{8}\left(\frac{3}{2} + u_1^{(s)}\right)^2 & -1 \\ 0 & -1 + \frac{1}{16}\left(\frac{3}{2} + u_1^{(s)}\right)^2 & \frac{3}{2} + \frac{1}{8}(2 + u_2^{(s)})^2 \end{bmatrix} \begin{bmatrix} \Delta u_0^{(s)} \\ \Delta u_1^{(s)} \\ \Delta u_2^{(s)} \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{3}{2}u_0^{(s)} - u_1^{(s)} + \frac{1}{8}\left[\frac{1}{3}(1 + u_0^{(s)})^3 + \frac{1}{6}\left(\frac{3}{2} + u_1^{(s)}\right)^3\right] - \frac{1}{4} \\ -u_0^{(s)} + 2u_1^{(s)} - u_2^{(s)} + \frac{1}{8}\left(\frac{3}{2} + u_1^{(s)}\right)^3 \\ -u_1^{(s)} + \frac{3}{2}u_2^{(s)} + \frac{1}{8}\left[\frac{1}{6}\left(\frac{3}{2} + u_1^{(s)}\right)^3 + \frac{1}{3}(2 + u_2^{(s)})^3 - \frac{1}{2}\right] \end{bmatrix}$$

and $$u_0^{(s+1)} = u_0^{(s)} + \Delta u_0^{(s)},\ u_1^{(s+1)} = u_1^{(s)} + \Delta u_1^{(s)},\ u_2^{(s+1)} = u_2^{(s)} + \Delta u_2^{(s)}.$$

Using $u_0^{(o)} = 0.001$, $u_1^{(0)} = -0.1$, $u_2^{(0)} = 0.001$, we get after three iterations

$$u_0^{(3)} = -0.0023,\ u_1^{(3)} = -0.1622,\ u_2^{(3)} = -0.0228.$$

The analytical solution of the boundary value problem is

$$u(x) = \frac{2}{2 - x} - x - 1.$$

$$u(0) = 0,\ u(1/2) = -0.1667,\ u(1) = 0.$$

Sample Programs In C

PROGRAM 1

```
/*PROGRAM BISECTION
Findings simple root of f(x) = 0 using bisection method. Read the end points of the interval
(a, b) in which the root lies, maximum number of iterations n and error tolerance eps.*/

#include <stdio.h>
#include <math.h>

float f();

main()
{
    float     a, b, x, eps, fa, fx, ff, s;
    int    i, n;
    FILE       *fp;

    fp = fopen(“result”,“w”);

    printf(“Please input end points of interval (a, b),\n”);
    printf(“in which the root lies\n”);
    printf(“n: number of iterations\n”);
    printf(“eps: error tolerance\n”);
    scanf(“%f %f %d %E”,&a, &b, &n, &eps);
    fprintf(fp,“A = %f, B = %f, N = %d,”, a, b, n);
    fprintf(fp,“EPS = %e\n”, eps);

/*Compute the bisection point of a and b */

    x = (a + b)/2.0;
    for(i = 1; i <= n; i++)
    {
        fa = f(a);
        fx = f(x);
```

```
          if(fabs(fx) <=eps)
              goto l1;
          ff = fa * fx;
          if(ff < 0.0)
              x = (a + x)/2.0;
          else
          {
              a = x;
              x = (x + b)/2.0;
          }
    }
    printf("No. of iterations not sufficient\n");
    goto l2;
l1: fprintf(fp,"ITERATIONS = %d,",i);
    fprintf(fp," ROOT = %10.7f, F(X) = %E\n", x, fx);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
    fclose(fp);
12: return 0;
    }
/*********************************************************/
float f(x)
    float x;
       { float   fun;
              fun = cos(x) - x * exp(x);
              return(fun);
       }
/*********************************************************/
A = 0.000000, B = 1.000000, N = 40, EPS = 1.000000e-04
ITERATIONS = 25, ROOT = 0.5177526, F(X) = 1.434746E-05
/*********************************************************/
```

PROGRAM 2

```
/*PROGRAM REGULA-FALSI
Finding a simple root of f(x)=0 using Regula-Falsi method. Read the end points of the interval
(a, b) in which the root lies, maximum number of iterations n and the error tolerance eps. */

#include <stdio.h>
#include <math.h>
```

```
float f();

main()
   {
   float        a, b, x, eps, fa, fb, fx;
   int      i, n;
   FILE       *fp;

   fp = fopen("result","w");
   printf("Input the end points of the interval (a, b) in");
   printf("which the root lies");
   printf("n: number of iterations\n");
   printf("eps: error tolerance\n");
   scanf("%f %f %d %E", &a, &b, &n, &eps);
   fprintf(fp,"a = %f b = %f n = %d", a, b, n);
   fprintf(fp," eps = %e\n\n", eps);

/*Compute the value of f(x) at a & b and calculate the new
   approximation x and value of f(x) at x.            */
   for (i = 1; i <= n; i++)
         { fa = f(a);
            fb = f(b);
            x = a - (a - b) * fa / (fa - fb);
            fx = f(x);
      if(fabs(fx) <= eps)

/*Iteration is stopped when abs(f(x)) is less than or equal to eps.
   Alternate conditions can also be used.             */

               goto l1;
            if((fa * fx) < 0.0)
               b = x;
            else
               a = x;
         }
   printf("\nITERATIONS ARE NOT SUFFICIENT");
   goto l2;
11: fprintf(fp,"Number of iterations = %d\n", i);
   fprintf(fp,"Root = %10.7f, f(x) = %e\n", x, fx);
   printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
12: return 0;
   }
/*****************************************************/
```

```
float f(x)
   float x;
      { float fun;
            fun = cos(x) - x * exp(x);
            return(fun);
      }
/********************************************************/
a = 0.000000 b = 1.000000 n = 20 eps = 1.000000e-04
Number of iterations = 9
Root = 0.5177283, f(x) = 8.832585e-05
/********************************************************/
```

PROGRAM 3

```
/*PROGRAM SECANT METHOD
Finding a simple root of f(x) = 0 using Secant method. Read any two approximations to the
root, say, a, b; maximum number of iterations n and the error tolerance eps. The method
diverges if the approximations are far away from the exact value of the root.      */

#include <stdio.h>
#include <math.h>

float f();

main()
   {
   float       a, b, x, eps, fa, fb, fx;
   int      i, n;
   FILE       *fp;

   fp = fopen("result","w");
   printf("Input any two approximations to the root ");
   printf("n: number of iterations\n");
   printf("eps: error tolerance\n");
   scanf("%f %f %d %E", &a, &b, &n, &eps);
   fprintf(fp,"a = %f b = %f n = %d", a, b, n);
   fprintf(fp," eps = %e\n\n", eps);

/*Compute the value of f(x) at a & b and calculate the new
   approximation x and value of f(x) at x.             */
```

```
    for (i = 1; i <= n; i++)
            { fa = f(a);
                fb = f(b);
                x = a - (a - b) * fa / (fa - fb);
            fx = f(x);
            if(fabs(fx) <= eps)

/* Iteration is stopped when abs(f(x)) is less than or equal to eps.
    Alternate conditions can also be used.              */
                    goto l1;
                    a = b;
                    b = x;
                }
    printf("\nITERATIONS ARE NOT SUFFICIENT");
    goto 12;
11: fprintf(fp,"Number of iterations = %d\n", i);
    fprintf(fp,"Root = %10.7f, f(x) = %e\n", x, fx);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
12: return 0;
    }
/**********************************************************/
float f(x)
    float   x;
        { float fun;
                fun = cos(x) - x * exp(x);
                return(fun);
        }
/**********************************************************/
a = 0.100000 b = 0.200000 n = 40 eps = 1.000000e-04
Number of iterations = 5
Root = 0.5177556, f(x) = 5.281272e-06
/**********************************************************/
```

PROGRAM 4

```
/* PROGRAM NEWTON-RAPHSON METHOD
    Finding a simple root of f(x) = 0 using Newton-Raphson method. Read initial approximation
xold. Maximum number of iterations n and error tolerance eps.          */

#include <stdio.h>
#include <math.h>
```

```
float f();
float df();

main()
   {
   float        xold, eps, fx, dfx, xnew;
   int          i, n;
   FILE            *fp;

   fp = fopen("result","w");
   printf("Input value initial approximation xold\n");
   printf("n: number of iterations\n");
   printf("eps: error tolerance\n");
   scanf(%f %d %E", &xold, &n, &eps);
   fprintf("fp,"Input value initial approximation xold\n");
   fprintf(fp,"number of iterations n,");
   fprintf(fp," error tolerance eps\n");
   fprintf(fp,"xold = %f n = %d eps = %e\n\n", xold, n, eps);

/*Calculate f and its first derivative at xold  */
   for(i = 1; i <= n; i++)
         {   fx = f(xold);
             dfx = df(xold);
             xnew = xold - fx / dfx;
             fx = f(xnew);
             if(fabs(fx) <= eps) goto l10;

/* Iteration is stopped when abs(f(x)) is less than or equal to eps.
   Alternate conditions can also be used.      */
             xold = xnew;
         }
   printf("\nITERATIONS ARE NOT SUFFICIENT");
   goto 120;
l10:
   fprintf(fp,"Iterations = %d",i);
   fprintf(fp," Root = %10.7f, f(x) = %e\n", xnew, fx);
   printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
l20: return 0;
         }
/*************************************************************/
float f(x)
   float   x;
      { float   fun;
```

```
            fun = cos(x) - x * exp(x);
            return(fun);
      }
/*******************************************************/
float df(x)
   float   x;
         {  float     dfun;
            dfun = - sin(x) - (x + 1.0) * exp(x);
            return(dfun);
         }
/*******************************************************/
Input value initial approximation xold
number of iterations n, error tolerance eps
xold = 1.000000 n = 15 eps = 1.000000e-04
Iterations = 4 Root = 0.5177574, f(x) = 2.286344e-08
/*******************************************************/
```

PROGRAM 5

```
/* PROGRAM MULLER METHOD
Finding a root of f(x) = 0 using Muller method. Read three initial approximations x0, x1 and
x1, maximum number of iterations n and error tolerance eps.          */

#include <stdio.h>
#include <math.h>

float f();

main()
   {
   float        x, x0, x1, x2, fx, fx0, fx1, fx2;
   float        al, dl, c, g, p, q, eps;
   int       i, n;
   FILE          *fp;

   fp = fopen("result","w");
   printf("Input three initial approximations : x0, x1, x2\n");
   printf("number of iterations : n, \n");
   printf("error tolerance : eps\n");
   scanf("%f %f %f %d %E", &x0, &x1, &x2, &n, &eps);
   fprintf(fp,"Input three initial approximations x0, x1, x2\n");
```

```
    fprintf(fp,"Number of iterations n and error tolerance eps\n");
    fprintf(fp,"x0 = %f, x1 = %f, x2 = %f\n", x0, x1, x2);
    fprintf(fp,"n = %d, eps%e\n", n, eps);

/*Compute f(x) at x0, x1 and x2 */
    for(i = 1; i <= n; i++)
        {   fx0 = f(x0);
            fx1 = f(x1);
            fx2 = f(x2);

/*Calculate the next approximation x */
            al = (x2 - x1) / (x1 - x0);
            dl = 1.0 + al;
            g = al * al * fx0 - dl * dl * fx1 + (al + dl) * fx2;
            c = al * (al * fx0 - dl * fx1 + fx2);
            q = g * g - 4.0 * dl * c * fx2;
            if(q < 0.0)
               q = 0.0;
            p = sqrt(q);
            if(g < 0.0)
               p = - p;
             al = - 2.0 * dl * fx2 / (g + p);
               x = x2 + (x2 - x1) * al;
             fx = f(x);

            if(fabs(fx) <= eps) goto l10;

/* Iteration is stopped when abs(f(x)) is less than or equal to eps.
    Alternate conditions can also be used.          */
            x0 = x1;
            x1 = x2;
            x2 = x;
        }
     printf("ITERATIONS ARE NOT SUFFICIENT\n");
     fprintf(fp,"\nITERATIONS ARE NOT SUFFICIENT");
     goto l20;
l10:
     fprintf(fp,"ITERATIONS = %d ROOT = %10.7f", i, x);
     fprintf(fp," F(x) = %e\n", fx);
     printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
    fclose(fp);
```

```
l20: return 0;
  }
/*********************************************************/
float f(x)
  float   x;
    {  float   fun;
          fun = cos(x) - x * exp(x);
          return(fun);
    }
/*********************************************************/
Input three initial approximations x0, x1, x2
Number of iterations n and error tolerance eps
x0 = -1.000000, x1 = 0.000000, x2 = 1.000000
n = 10, eps1.000000e-06
ITERATIONS = 4 ROOT =   0.5177574 F(x) =   2.286344e-08
/*********************************************************/
```

PROGRAM 6

```
/*PROGRAM BAIRSTOW METHOD
Extraction of a quadratic factor from a polynomial
    {x**n + a[1] * x**(n - 1) + ..... + a[n-1] * x + a[n] = 0}
of degree greater than two using Bairstow method. n gives the degree of the polynomial. a[i]
represents coefficients of polynomial in decreasing powers of x. p & q are initial approxima-
tions. m is the number of iterations and eps is the desired accuracy.          */

#include <stdio.h>
#include <math.h>

main()
  {
  float       a[10], b[10], c[10], p, q, cc, den, delp;
  float             delq, eps;
  int      i, n, m, j, k, l;
  FILE          *fp;

  fp = fopen("result","w");

  printf("Input initial approximations of p & q:\n");
  printf("Degree of polynomial : n,\n");
  printf("Number of iterations :m,\n");
```

```
    printf("Desired accuracy :eps,\n");
    scanf("%f %f %d %d %E", &p, &q, &n, &m, &eps);
    fprintf(fp," Input initial approximations of p & q:\n");
    fprintf(fp,"Degree of polynomial: n,\n");
    fprintf(fp,"Number of iterations :m,\n");
    fprintf(fp,"Desired accuracy :eps,\n");
    fprintf(fp,"p = %f, q = %f, n = %d", p, q, n);
    fprintf(fp,"m = %d, eps = %e\n", m, eps);

/* Read coefficients of polynomial in decreasing order */
    printf("Input coefficients of polynomial in decreasing");
    printf(" order\n");
    fprintf(fp,"Coefficients of polynomial are\n");
    for (i = 1; i <= n; i++)
          {   scanf("%f", &a[i]);
              fprintf(fp," %.4f",a[i]);
          }
    fprintf(fp,"\n");

/* generate b[k] & c[k] */
    for (j = 1; j <= m;j++)
          {   b[1] = a[1] - p;
              b[2] = a[2] - p * b[1] - q;
              for (k = 3; k <= n; k++)
                 b[k] = a[k] - p * b[k-1] - q * b[k-2];
              c[1] = b[1] - p;
              c[2] = b[2] - p * c[1] - q;
              l = n - 1;
              for (k = 3; k <=l; k++)
                 c[k] = b[k] - p * c[k-1] - q * c[k-2];
              cc = c[n-1] - b[n-1];
              den = c[n-2] * c[n-2] - cc * c[n-3];
              if(fabs(den) == 0.0)
                 {   fprintf(fp,"WRONG INITIAL APPROXIMATION\n");
                     printf("\n WRONG INITIAL APPROXIMATION\n");
                     got l2;
                 }
              delp = -(b[n] * c[n-3] - b[n-1] * c[n-2]) / den;
              delq = -(b[n-1] * cc - b[n] * c[n-2]) / den;
              p = p + delp;
              q = q + delq;
              if((fabs(delp) <= eps) && (fabs(delq) <= eps))
```

```
            goto l2;
        }
    printf("ITERATIONS NOT SUFFICIENT\n");
    fprintf(fp,"ITERATIONS NOT SUFFICIENT\n");
    goto l3;
l2: fprintf(fp,"ITERATIONS = %d, P = %11.7e, ", j, p);
    fprintf(fp, "Q = %11.7e\n", q);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
    fclose(fp);
l3: return 0;
    }
/**********************************************************/
Input initial approximations of p & q:
Degree of polynomial: n,
Number of iterations :m,
Desired accuracy :eps,
p = 0.500000, q = 0.500000,n = 4 m = 10, eps = 1.000000e-06
Coefficients of polynomial are
    1.0000  2.0000  1.0000  1.0000
ITERATIONS = 7, P = 9.9999994e-01, Q = 1.0000000e+00
/**********************************************************/
```

PROGRAM 7

```
/*PROGRAM GAUSS ELIMINATION METHOD
Solution of a system of nxn linear equations using Gauss elimination method with partial
piviting. The program is for a 10×10 system. Change the dimension if higher order system is to
be solved.                                        */

#include <stdio.h>
#include <math.h>

main()
    {
    float     a[10][11], x[10], big, ab, t, quot, sum;
    int       n, m, l, i, j, k, jj, kp1, nn, ip1;
    FILE      *fp;

    fp = fopen("result","w");
    printf("Input number of equations : n\n");
    scanf("%d", &n);
```

```
fprintf(fp,"Order of the system = %d\n", n);
m = n + 1;
l = n - 1;
printf("Input the augmented matrix row-wise\n");
fprintf(fp,"Elements of the augmented matrix :\n");

for (i = 1; i <= n; i++)
    {   for (j = 1; j <= m; j++)
            {   scanf("%f", &a[i][j]);
                fprintf(fp," %.6f", a[i][j]);
            }
        fprintf(fp,"\n");
    }
for (k = 1; k <= l; k++)
    {   big = fabs(a[k][k]);
        jj = k;
        kp1 = k + 1;
        for(i = kp1; i <= n; i++)
            {   ab = fabs(a[i][k]);
                    if((big - ab) < 0.0)
                        {   big = ab;
                            jj = i;
                        }
            }
        if((jj - k) > 0)
            {   for (j = k; j <= m; j++)
                        {   t = a[jj][j];
                            a[jj][j] = a[k][j];
                            a[k][j] = t;
                        }
            }
        for (i = kp1; i <= n; i++)
            {   quot = a[i][k]/a[k][k];
                    for (j = kp1; j <= m; j++)
                        a[i][j] = a[i][j] - quot*a[k][j];
            }
        for (i = kp1; i <= n; i++)
            a[i][k] = 0.0;
    }
x[n] = a[n][m]/a[n][n];
for (nn = 1; nn <= 1; nn++)
```

```
        {   sum = 0.0;
            i = n - nn;
            ip1 = i + 1;
            for(j = ip1; j <= n; j++)
                sum = sum + a[i][j]*x[j];
            x[i] = (a[i][m] - sum)/a[i][i];
        }
    fprintf(fp,"SOLUTION VECTOR\n");
    for (i = 1; i <= n; i++)
        fprintf(fp," %8.5f", x[i]);
    fprintf(fp,"\n");
    printf("PLEASE SEE FILE 'result' FOR RESULTS\n");
    return 0;
  }
/*********************************************************/
Order of the system = 3
Elements of the augmented matrix :
   1.000000  1.000000  1.000000  6.000000
   3.000000  3.000000  4.000000  20.000000
   2.000000  1.000000  3.000000  13.000000
SOLUTION VECTOR
      3.00000    1.00000    2.00000
/*********************************************************/
```

PROGRAM 8

```
/*PROGRAM JORDAN METHOD
Matrix inversion and solution of NXN system of equations using Gauss Jordan method. If the
system of equations is larger than 15x15, change the dimensions if the float statement.     */

#include <stdio.h>
#include <math.h>

main()
  {
  float       a[15][15], ai[15][15], b[15], x[15];
  float    aa[15][30], big, ab, t, p, sum;
  int        n,m, m2, i, j, lj, k, kp1, jj, lk, li, l3;
  FILE        *fp;
```

```
    fp = fopen("result","w");
    printf("Input order of matrix : n\n");
    scanf("%d", &n);
    printf("Input augmented matrix row-wise\n");
    for (i = 1; i <= n; i++)
        {   for (j = 1; j <= n; j++)
                scanf("%f", &a[i][j]);
                scanf("%f", &b[i]);
        }
    fprintf(fp,"Order of the system = %d\n", n);
    fprintf(fp,"Elements of the augmented matrix :\n");
    for (i = 1; i <= n; i++)
        {   for (j = 1; j <= n; j++)
                fprintf(fp," %8.4f", a[i][j]);
            fprintf(fp,"   %8.4f\n", b[i]);
        }
    m = n + n;
    m2 = n + 1;
/* Generate the augmented matrix aa.     */
    for (i = 1; i <= n; i++)
        {   for (j = 1; j <= n; j++)
                aa[i][j] = a[i][j];
        }
    for (i = 1; i <= n; i++)
        {   for (j = m2; j <= m; j++)
                aa[i][j] = 0.0;
        }
    for (i = 1; i <= n; i++)
        {   j = i + n;
            aa[i][j] = 1.0;
        }

/*Generate elements of b matrix.     */
    for (lj = 1; lj <= n; lj++)
        {       /*Search for the largest pivot.  */
            k = lj;
            if(k < n)
                {   jj = k;
                    big = fabs(aa[k][k]);
                    kp1 = k + 1;
                    for(i = kp1; i <= n; i++)
```

```
                { ab = fabs(aa[i][k]);
                  if((big - ab) < 0.0)
                        { big = ab;
                          jj = i;
                        }
                }
/*Interchange rows if required.  */
            if((jj - k) != 0)
                { for (j = k; j <= m; j++)
                        { t = aa[jj][j];
                          aa[jj][j] = aa[k][j];
                          aa[k][j] = t;
                        }
                }
          }
      p = aa[lj][lj];
      for (i = lj; i <= m; i++)
            aa[lj][i] = aa[lj][i] / p;
      for (lk = 1; lk <= n; lk++)
            { t = aa[lk][lj];
              for (li = lj; li <= m; li++)
                  { if((lk - lj) != 0)
                        aa[lk][li] = aa[lk][li] - aa[lj][li] * t;
                  }
            }
    }
  for (i = 1; i <= n; i++)
    { for (j = m2; j <= m; j++)
          { l3 = j - n;
            ai[i][l3] = aa[i][j];
          }
    }
  fprintf(fp,"\n INVERSE MATRIX\n");
  for (i = 1; i <= n; i++)
    { for (j = 1; j <= n; j++)
          fprintf(fp," %11.5f", ai[i][j]);
      fprintf(fp,"\n");
    }
  for (i = 1; i <= n; i++)
    { sum = 0.0;
      for (k = 1; k <= n; k++)
```

```
                sum = sum + ai[i][k] * b[k];
            x[i] = sum;
        }
  fprintf(fp,"\n SOLUTION VECTOR\n");
  for (i = 1; i <= n; i++)
        fprintf(fp," %11.5f", x[i]);
  fprintf(fp,"\n");
  printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
  fclose(fp);
  return 0;
  }
/**********************************************************/
```

Order of the system = 4

Elements of the augmented matrix :

3.0000	4.0000	2.0000	2.0000	6.0000
2.0000	5.0000	3.0000	1.0000	4.0000
2.0000	2.0000	6.0000	3.0000	3.0000
1.0000	2.0000	4.0000	6.0000	6.0000

INVERSE MATRIX

0.59756	– 0.46341	0.17073	– 0.20732
– 0.14024	0.35366	– 0.18293	0.07927
– 0.18902	0.08537	0.23171	– 0.06707
0.07317	– 0.09756	– 0.12195	0.21951

SOLUTION VECTOR

1.00000	0.50000	– 0.50000	1.00000

/**/

PROGRAM 9

```
/*PROGRAM GAUSS-SEIDEL
Program to solve a system of equations using Gauss-Seidel iteration method. Order of the matrix is n, maximum number of iterations is niter, error tolerance is eps and the initial approximation to the solution vector x is oldx. If the system of equations is larger than 10x10, change the dimensions in float.                  */

#include <stdio.h>
#include <math.h>

main()
  {
  float      a[10][10], b[10], x[10], oldx[10], sum, big, c;
```

```
    float        eps;
    int      n, niter, i, j, ii, jj, k, l;
    FILE         *fp;

    fp = fopen("result","w");

    printf("Input the order of matrix : n\n");
    printf("Input the number of iterations : niter\n");
    printf("Input error tolerance : eps\n");
    scanf("%d %d %e", &n, &niter, &eps);
    fprintf(fp,"n = %d, niter = %d, eps = %e\n", n, niter, eps);
    printf("Input augmented matrix row-wise\n");
    fprintf(fp,"Elements of the augmented matrix\n");
    for (i = 1; i <= n; i++)
         {  for (j = 1; j <= n; j++)
              {  scanf("%f", &a[i][j]);
                 fprintf(fp,"%f ", a[i][j]);
              }
            scanf("%f", &b[i]);
            fprintf(fp,"   %f\n", b[i]);
         }
    printf("Input initial approx. to the solution vector\n");
    fprintf(fp,"Initial approx. to solution vector :\n");
    for (i = 1; i <= n; i++)
         {  scanf("%f", &oldx[i]);
            fprintf(fp,"%f   ", oldx[i]);
         }
    fprintf(fp,"\n")
    for (i = 1; i <= n; i++)
            x[i] = oldx[i];

/*Compute the new values for x[i] */
    for (ii = 1; ii <= niter; ii++)
         {  for (i = 1; i <= n; i++)
              {  sum = 0.0;
                 for (j = 1; j <= n; j++)
                      {  if((i - j) ! = 0)
                         sum = sum + a[i][j] * x[j];
                      }
                 x[i] = (b[i] - sum) / a[i][i];
              }
            big = fabs(x[1] - oldx[1]);
            for (k = 2; k <= n; k++)
```

```
            {   c = fabs(x[k] - oldx[k]);
                if(c > big)
                    big = c;
            }
        if(big <= eps)
            goto l10;
        for (l = 1; l <= n; l++)
            oldx[l] = x[l];
    }
  printf("ITERATIONS NOT SUFFICIENT\n");
  fprintf(fp,"ITERATIONS NOT SUFFICIENT\n");
  goto l20;
l10: fprintf(fp,"Number of iterations = %d\n", ii);
  fprintf(fp,"Solution vector\n");
  for(i = 1; i <= n; i++)
        fprintf(fp," %f", x[i]);
  fprintf(fp,"\n");
  printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
l20: return 0;
  }
/*********************************************************/
n = 4, niter = 30, eps = 1.000000e-06
Elements of the augmented matrix
3.000000  4.000000  2.000000  2.000000  6.000000
2.000000  5.000000  3.000000  1.000000  4.000000
2.000000  2.000000  6.000000  3.000000  3.000000
1.000000  2.000000  4.000000  6.000000  6.000000
Initial approx. to solution vector :
0.100000  0.100000  0.100000  0.100000
Number of iterations = 28
Solution vector
  1.000000  0.500000  -0.500000 1.000000
/*********************************************************/
```

PROGRAM 10

```
/* PROGRAM POWER METHOD
Program to find the largest eigen value in magnitude and the corresponding eigen vector of a
square matrix A of order n using power method. If the order of the matrix is greater than 10,
change the dimensions in float.                    */
```

```
#include <stdio.h>
#include <math.h>

main()
   {
   float lambda[10], a[10][10], v[10], y[10], max, sum, eps;
   float big, c;
   int i, j, n, ii, niter, k, l;
   FILE   *fp;

   fp = fopen("result","w");

/* Read the order of matrix A, number of iterations, coefficients of matrix A and the initial
vector c.*/

   printf("Input the order of matrix :n\n");
   printf("Input number of iterations : niter\n");
   printf("Input error tolerance : eps\n");
   scanf("%d %d %e", &n, &niter, &eps);
   fprintf(fp,"Order of the matrix = %d\n",n);
   fprintf(fp,"Number of iterations = %d\n", niter);
   fprintf(fp,"Error tolerance = %e\n", eps);
   printf("Input the coefficients of matrix row-wise\n");
   fprintf(fp,"Elements of the matrix\n");
   for (i = 1; i <= n; i++)
      {  for (j = 1; j <= n; j++)
               {  scanf("%f", &a[i][j]);
                  fprintf(fp," %f", a[i][j]);
               }
            fprintf(fp,"\n");
      }
   printf("Input the elements of the approx. eigen vector\n");
   fprintf(fp,"Approx. eigen vector\n");
   for (i = 1; i <=n; i++)
      {  scanf("%f", &v[i]);
            fprintf(fp," %f", v[i]);
      }
   fprintf(fp, "\n");
   for (ii = 1; ii <= niter; ii++)
      {  for (i = 1; i <= n; i++)
            {  sum = 0.0;
               for (k = 1; k <= n; k++)
                     sum = sum + a[i][k] * v[k];
```

```
                y[i] = sum;
            }
        for (i = 1; i <= n; i++)
            lambda[i] = fabs(y[i] / v[i]);

/* Normalise the vector y.    */
        max = fabs(y[1]);
        for (i = 2; i <= n; i++)
            {   if(fabs(y[i] > max)
                    max = fabs(y[i]);
            }
        for (i = 1; i <= n; i++)
            v[i] = y[i] / max;
        big = 0.0;
        for (j = 1; i <= n - 1; j++)
        {
            for (i = j + 1; i <= n; i++)
            {
                c = fabs(lambda[j] - lambda[i]);
                if(big < c)
                    big = c;
            }
        }
        if(big <= eps)
            goto l1;
    }
    printf("NUMBER OF ITERATIONS NOT SUFFICIENT\n");
    fprintf(fp,"NUMBER OF ITERATIONS NOT SUFFICIENT\n");
    goto l2;
l1: fprintf(fp,"Number of iterations = %d\n", ii);
    fprintf(fp,"Approx. to Eigen value = ");
    for (l = 1; l <= n; l++)
        fprintf(fp," %f", lambda[1]);
    fprintf(fp,"\n");
    fprintf(fp,"Eigen-vector = ");
    for (l = 1; l <= n; l++)
        fprintf(fp," %f", v[1]);
    fprintf(fp,"\n");
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
l2: return 0;
    }
/*********************************************************/
```

```
Order of the matrix = 3
Number of iterations = 20
Error tolerance = 1.000000e-04
Elements of the matrix
   - 15.000000      4.000000      3.000000
     10.000000   - 12.000000      6.000000
     20.000000    - 4.000000      2.000000
Approx. eigen vector
      1.000000      1.000000      1.000000
Number of iterations = 19
Approx. to Eigen value =   19.999981   20.000076   19.999981
Eigen-vector = - 1.000000   0.499998   1.000000
/*******************************************************/
```

PROGRAM 11

```
/* PROGRAM : LAGRANGE METHOD
   Programme for Lagrange interpolation.  */

#include <stdio.h>
#include <math.h>

main()
   {
   float     x[10], y[10], xin, yout, sum;
   int       n, i, j;
   FILE      *fp;

   fp = fopen("result","w");

/* Read in data.   */

   printf("Input number of points : n\n");
   scanf("%d", &n);
   fprintf(fp,"Number of points = %d\n", n);
   printf("Input the abscissas \n");
   fprintf(fp,"The abscissas are :\n");
   for (i = 1; i <= n; i++)
         {  scanf("%f", &x[i]);
            fprintf(fp,"%8.4f", x[i]);
         }
```

```
    fprintf(fp, "\n");
    printf("Input the ordinates\n");
    fprintf(fp,"The ordinates are :\n");
    for (i = 1; i <= n; i++)
        {   scanf("%f", &y[i]);
            fprintf(fp,"%8.4f", y[i]);
        }
    fprintf(fp,"\n");

/* Read in x value for which y is desired.      */

    printf("Input value of x for which y is required\n");
    scanf("%f", &xin);
    fprintf(fp,"The value of x for which y is required is ");
    fprintf(fp,"%5.3f\n", xin);
/* Compute the value of y.    */

    yout = 0.0;
    for (i = 1; i <= n; i++)
        {   sum = y[i];
            for (j = 1; j <= n; j++)
                {   if(i != j)
                    sum = sum * (xin - x[j]) / (x[i] - x[j]);
                }
            yout = yout + sum;
        }
    fprintf(fp,"Lagrange interpolation based on %d points\n",n);
    fprintf)fp,"At x = %5.3f, y = %8.5f\n", xin, yout);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n");
    fclose(fp);
    return 0;
    }
/************************************************************/
```

```
Number of points = 6
The abscissas are :
   0.0000  1.0000  2.0000  4.0000  5.0000  6.0000
The ordinates are :
   1.0000  14.0000    15.0000    5.0000  6.0000  19.0000
The value of x for which y is required is 3.000
Lagrange interpolation based on 6 points
At x = 3.000, y = 10.00000
/************************************************************/
```

PROGRAM 12

```
/* NEWTON-GREGORY INTERPOLATION
Program for interpolation in a uniformly spaced table using Newton-Gregory formula.      */

#include <stdio.h>
#include <math.h>

main()
   {
   float        y[10], d[10], xi, xf, x, h, fm, fj;
   float        yout, fnum, fden, x0, y0,u, ffx, ffxx;
   int       n, m, i, j, k;
   FILE           *fp;

   fp = fopen("result","w");

/* Read in starting value and last value of x, the step size and the y values. n gives the total
number of nodal points.   */

   printf("Input the number of abscissas,\n");
   printf("starting value of x,\n");
   printf("last value of x and\n");
   printf("the step size\n");
   scanf("%d %f %f %f", &n, &xi, &xf, &h);
   fprintf(fp,"The number of abscissas = %d\n", n);
   fprintf(fp,"The starting value of x = %f\n", xi);
   fprintf(fp,"The last value of x = %f\n", xf);
   fprintf(fp,"The step size = %f\n", h);
   printf("Input the ordinates\n");
   fprintf(fp,"The ordinates are :\n");
   for (i = 1; i <= n; i++)
         {  scanf("%f", &y[i]);
            fprintf(fp,"%f", y[i]);
         }
   fprintf(fp,"\n");

/* Read in value of x for which y is desired and m the degree of the polynomial to be used.
Maximum value of m is 15.  */

   printf("Input x for which interpolation is required\n");
   printf("and the degree of polynomial\n");
   scanf("%f %d", &x, &m);
   fprintf(fp,"The value of x for which interpolation is ");
```

```
    fprintf(fp,"required is %f\n", x);
    fprintf(fp,"The degree of polynomial = %d\n", m);
    fm = m + 1;
    ffx = x - xi - fm * h / 2.0;
    ffxx = xf - x - fm * h / 2.0;
    if(ffx > 0.0)
      {   if(ffxx <= 0.0)
                j = n - m;
            else
                   j = (x - xi) / h - fm / 2.0 + 2.0;
      }
    else
            j = 1;
    fj = j;
    x0 = xi + (fj - 1.0) * h;
    y0 = y[j];

/* Calculate required differences d[i] and y. */
    for (i = 1; i <= m; i++)
          {   d[i] = y[j+1] - y[j];
              j = j + 1;
          }
    for (j = 2; j < = m; j++)
          {   for (i = j; i <= m; i++)
                {   k = m - i + j;
                    d[k] = d[k] - d[k-1];
                }
          }
    u = (x - x0) / h;
    yout = y0;
    fnum = u;
    fden = 1.0;
    for (i = 1; i <= m; i++)
          {   yout = yout + fnum/fden * d[i];
              fnum = fnum * (u - i);
              fden = fden * (i + 1);
          }
    fprintf(fp,"Newton-Gregory interpolation of degree %d\n", m);
    fprintf(fp,"At x = %7.5f, y = %7.5f\n", x, yout);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
    fclose(fp);
```

```
    return 0;
    }
/**********************************************************/
The number of abscissas = 5
The starting valueof x = 0.100000
The last value of x = 0.500000
The step size = 0.100000
The ordinates are :
1.400000  1.560000  1.760000  2.000000  2.280000
The value of x for which interpolation is required is 0.250000
The degree of polynomial = 4
Newton-Gregory interpolation of degree 4
At x = 0.25000, y = 1.65500
/**********************************************************/
```

PROGRAM 13

```
/* CUBIC SPLINE INTERPOLATION
Program for cubic spline interpolation for arbitrary set of points. The second derivatives at the
end points are assumed as zeros (natural spline).            */

#include <stdio.h>
#include <math.h>

main()
    {
    float       x[20], y[20], sdr[20], a[20], b[20], c[20], r[20];
    float       t, xx, dxm, dxp, del, f;
    int     n, i, j, nm1, nm2, k;
    FILE            *fp;

    fp = fopen("result","w");

/* Read n the number of points, x and y values.   */
    printf("Input number of points\n");
    scanf("%d", &n);
    fprintf(fp,"Number of points = %d\n", n);
    printf("Input abscissas\n");
    fprintf(fp,"The abscissas are :\n");
    for (i = 1; i <= n; i++)
        {   scanf("%f", &x[i]);
```

```
            fprintf(fp,"%f", x[i]);
        }
    fprintf(fp,"\n");
    printf("Input ordinates\n");
    fprintf(fp,"The ordinates are :\n");
    for (i = 1; i <= n; i++)
        {   scanf("%f", &y[i]);
            fprintf(fp,"%f", y[i]);
        }
    fprintf(fp,"\n");

/* Read the value of x for which y is required. */
    printf("Input x for which interpolation is required\n");
    scanf("%f", &xx);
    fprintf(fp,"The value of x for which interpolation ");
    fprintf(fp,"is required is %f\n", xx);

/* Calculate second order derivatives needed in cubic spline interpolation. a, b and c are the
three diagonals of the tridiagonal system. r is the right hand side.      */

    nm2 = n - 2;
    nm1 = n - 1;
    sdr[1] = 0.0;
    sdr[n] = 0.0;
    c[1] = x[2] - x[1];
    for (i = 2; i <= nm1; i++)
        {   c[i] = x[i+1] - x[i];
            a[i] = c[i-1];
            b[i] = 2.0 * (a[i] + c[i]);
            r[i] = 6.0*((y[i+1]-y[i]/c[i]-(y[i]-y[i-1]/c[i-1]);
        }

/* Solve the tridiagonal system.   */
    for (i = 3; i <= nm1; i++)
        {   t = a[i] / b[i-1];
            b[i] = b[i] - t * c[i-1];
            r[i] = r[i] - t * r[i-1];
        }
    sdr[nm1] = r[nm1] / b[nm1];
    for (i = 2; i <= nm2; i++)
        {   k = n - i;
            sdr[k] = (r[k] - c[k] * sdr[ k + 1]) / b[k];
        }
```

```
/* Calculate the corresponding value of y. Find the proper interval.     */
   for (i = 1; i <= nm1; i++)
         {  j = i;
            if(xx <= x[i + 1])
               goto l1;
         }
l1: dxm = xx - x[j];
   dxp = x[j + 1] - xx;
   del = x[j + 1] - x[j];
   f = sdr[j] * dxp * (dxp * dxp / del - del)/6.0;
   f = f + sdr[j + 1] * dxm * (dxm * dxm / del - del) / 6.0;
   f = f + y[j] * dxp / del + y[j + 1] * dxm / del;
   fprintf(fp,"At x = %6.4f, interpolated value using", xx);
   fprintf(fp,"%d points is y = %8.4f\n", n, f);
   printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
   fclose (fp);
   return 0;
   }
/**********************************************************/
Number of points = 5
The abscissas are :
0.000000  1.000000  2.000000  3.000000  4.000000
The ordinates are :
1.000000  2.000000  33.000000    244.000000  1025.000000
The value of x for which interpolation is required is 1.750000
At x = 1.7500, interpolated value using 5 points is y = 21.1819
/**********************************************************/
```

PROGRAM 14

```
/* TRAPEZOIDAL RULE OF INTEGRATION
Program to evaluate the integral of f(x) between the limits a to be using Trapezoidal rule of
integration based on n subintervals or n+1 nodal points. The values of a, b and n are to be read
and the integrand is written as a function subprogram. The program is tested for
f(x) = 1 / (1 + x).                                                                          */

#include <stdio.h>
#include <math.h>
```

```
float f();

main()
   {
   float        a, b, h, sum, x, trap;
   int       n, i, m;
   FILE           *fp;

   fp = fopen("result","w");

   printf("Input limits a & b and no. of subintervals n\n");
   scanf("%f %f %d", &a, &b, &n);
   fprintf(fp,"Limits are a = %f, b = %f\n", a, b);
   fprintf(fp,"Number of subintervals = %d\n", n);
   h = (b - a) / n;
   sum = 0.0;
   m = n - 1;
   for (i = 1; i <= m; i++)
         {   x = a + i * h;
             sum = sum + f(x);
         }
   trap = h * (f(a) + 2.0 * sum + f(b)) / 2.0;
   fprintf(fp,"Value of integral with %d ", n);
   fprintf(fp,"Subintervals = %14.6e\n", trap);
   printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
   return 0;
   }
/********************************************************/
float     f(x)
   float     x;
         {   float    fun;
                fun = 1.0 / (1.0 + x);
                return(fun);
         }
/********************************************************/
Limits are a = 0.000000, b = 1.000000
Number of subintervals = 8
Value of integral with 8 subintervals = 6.941218e-01
/********************************************************/
```

PROGRAM 15

```
/* SIMPSON RULE OF INTEGRATION
Program to evaluate the integral of f(x) between the limits a to b using Simpsons rule of inte-
gration based on 2n subintervals or 2n+1 nodal points. The values of a, b and n are to be read
and the integrand is written as a function subprogram. The program is tested for
f(x) = 1 / (1 + x).                                                              */

#include <stdio.h>
#include <math.h>

float f();

main()
  {
  float     a, b, h, x, sum, sum1, sum2, simp;
  int       n, i, n1, n2;
  FILE      *fp;

  fp = fopen("result","w");

  printf("Input limits a & b and half the no. of ");
  printf("subintervals n\n");
  scanf("%f %f %d", &a, &b, &n);
  fprintf(fp,"The limits are a = %f, b = %f\n", a, b);
  h = (b - a) / (2.0 * n);
  sum = f(a) + f(b);
  sum1 = 0.0;
  n1 = 2 * n - 1;
  for (i = 1; i <= n1; i = i+2)
      {  x = a + i * h;
         sum1 = sum1 + f(x);
      }
  n2 = 2 * n - 2;
  sum2 = 0.0;
  for (i = 2; i <= n2; i = i+2)
      {  x = a + i * h;
         sum2 = sum2 + f(x);
      }
  simp = h * (sum + 4.0 * sum1 + 2.0 * sum2) / 3.0;
  fprintf(fp,"Value of integral with ");
  fprintf(fp,"%d Subintervals = %14.6e\n", 2 * n, simp);
  printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
```

```
    return 0;
    }
/*****************************************************/
float     f(x)
    float     x;
        {   float    fun;
                fun = 1.0 / (1.0 + x);
                return(fun);
        }
/*****************************************************/
The limits are a = 0.000000, b = 1.000000
Value of integral with 8 subintervals = 6.931545e-01
/*****************************************************/
```

PROGRAM 16

```
/* ROMBERG INTEGRATION
Program to evaluate the integral of f(x) between the limits a and b using Romberg integration
based on Trapezoidal rule. Values of a, b and desired accuracy are to be read and the integrand
is written as a function subprogram. Array r gives Romberg table. n gives number of extrapo-
lations. The program is tested for f(x) = 1 / (1 + x).                    */

#include <stdio.h>
#include <math.h>

float f();

main()
    {
    float         r[15][15], a, b, h, jj, kk, x, diff, eps;
    int        n, i, j, k, m, l, ii;
    int        x1, x2;
    FILE            *fp;

    fp = fopen("result","w");

    printf("Input limits a & b, \n");
    printf("the maximum no. of extrapolations n and\n");
    printf("the error tolerance eps\n");
    scanf("%f %f %d %E", &a, &b, &n, &eps);
    fprintf(fp,"The limits are : a = %f, b = %f\n", a, b);
```

```
fprintf(fp,"The maximum number of extrapolations = %d\n", n);
fprintf(fp,"The error tolerance = %11.4e\n",eps);
i = 1;
h = b - a;
r[1][1] = 0.5 * h * (f(a) + f(b));
for (ii = 1; ii <= n; ii++)
      {   h = h/2.0;
   x2 = 1;
            for (x1 = 1; x1 <= (ii - 1); x1++)
        x2 = x2 * 2;
   j = x2;
            i = i + 1;
            r[i][1] = 0.5 * r[I - 1][1];
            for (k = 1; k <= j; k++)
               {   x = a + (2.0 * k - 1.0) * h;
                   r[i][1] = r[i][1] + h * f(x);
               }
            for (k = 2; k <= i; k++)
               {
            x2 = 1;
                      for (x1 = 1; x1 <= (k - 1); x1++)
                          x2 = x2 * 4;
                      jj = x2 * r[i][k - 1] - r[i - 1][k - 1];
                      kk = x2 - 1;
                      r[i][k] = jj/kk;
               }
               diff = fabs(r[i][i] - r[i][i - 1]);
               if(diff <= eps)
               {   fprintf(fp,"Romberg table after %d ", i - 1);
                      fprintf(fp,"extrapolations\n");
                      for (l = 1; l <= i; l++)
                      {   for (m = 1; m <= l; m++)
                                fprintf(fp,"%10.6f ", r[l][m]);
                          fprintf(fp,"\n");
                      }
                      goto l2;
               }
      }
printf("Number of extrapolations are not sufficient\n");
fprintf(fp,"Number of extrapolations are not sufficient\n");
goto l1;
```

```
l2: printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
l1: return 0;
   }
/**********************************************************/
float    f(x)
   float    x;
      {  float    fun;
            fun = 1.0 / (1.0 + x);
            return(fun);
      }
/**********************************************************/
The limits are : a = 0.000000, b = 1.000000
The maximum number of extrapolations = 5
The error tolerance = 1.0000e-06
ROMBERG TABLE AFTER 3 EXTRAPOLATIONS
   0.750000
   0.708333  0.694444
   0.697024  0.693254  0.693175
   0.694122  0.693155  0.693148  0.693147
/**********************************************************/
```

PROGRAM 17

```
/* EULER METHOD FOR SOLVING FIRST ORDER INITIAL VALUE PROBLEM
Program to solve an IVP, dy/dx = f(x,y), y(x0) = y0, using Euler method. The initial values x0,
y0, the final value xf and the step size are to be read. f(x,y) is written as a function subprogram.
                                                                                          */

#include <stdio.h>
#include <math.h>

float f();

main()
   {
   float       x0, y0, h, xf, x, y;
   int      i, iter;
   FILE        *fp;

   fp = fopen("result","w");
```

```
    printf("Input initial point x0, initial value y0,\n");
    printf("step size h and final value xf\n");
    scanf("%f %f %%f", &x0, &y0, &h, &xf);
    fprintf(fp, "Initial point x0 = %f, initial ", x0);
    fprintf(fp),"value y0 = %f\n", y0);
    fprintf(fp,"Step size = %f\n", h);
    fprintf(fp,"Final value = %f\n",xf);
    iter = (xf - x0) / h + 1;
    for (i = 1; i <= iter; i++)
        {   y = y0 + h * f(x0,y0);
            x = x0 + h;
            if(x < xf)
                {   x0 = x;
                    y0 = y;
                }
        }
    fprintf(fp,"At x = %6.4f, y = %12.6e\n",x, y);
    printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
    fclose(fp);
    return 0;
    }
/*********************************************************/
float f(x, y)
    float     x, y;
        {   float    fun;
                fun = -2.0 * x * y * y;
                return(fun);
        }
/*********************************************************/
Initial point x0 = 0.000000, initial value y0 = 1.000000
Step size = 0.100000
Final value = 1.000000
At x = 1.0000, y = 5.036419e-01
/*********************************************************/
```

PROGRAM 18

/* RUNGE-KUTTA CLASSICAL FOURTH ORDER METHOD

Program to solve the IVP, dy/dx = f(x,y), y(x0) = y0 using the classical Runge-Kutta fourth order method with steps h and h/2 and also computes the estimate of the truncation error. Input parameters are: initial point, initial value, number of intervals and the step length h: Solutions with h, h/2 and the estimate of truncation error are available as output. The right hand side f(x,y) is computed as a function subprogram. */

```
#include <stdio.h>
#include <math.h>

float f();

main()
  {
  float      u[20], v[40], x0, y0, h, k1, k2, k3, k4;
  float      h1, v1, te, x1, u1;
  int     n, i, m, nn, ij;
  FILE        *fp;

  fp = fopen("result","w");

  printf("Input initial point x0, initial value y0,\n");
  printf("number of intervals n and step size h\n");
  scanf("%f %f %d %f", &x0, &y0, &n, &h);
  fprintf(fp,"Initial point x0 = %f, initial ", x0);
  fprintf(fp,"value y0 = %f\n", y0);
  fprintf(fp,"Number of intervals = %d,\n", n);
  x1 = x0;
  for (m = 1; m <= 2; m++)
        {  if(m == 1)
        {  nn = n;
           u(0) = y0;
        }
     else
           {  nn = 2 * n;
                 h = h / 2.0;
                 v[0] = y0;
           }
     for (i = 1; i <= nn; i++)
        {  if(m == 1)
              {  u1 = u[i-1];
                 h1 = h / 2.0;
                 k1 = h * f(x0, u1);
                 k2 = h * f(x0 + h1, u1 + 0.5 * k1);
                 k3 = h * f(x0 + h1, u1 + 0.5 * k2);
```

```
                k4 = h * f(x0 + h, u1 + k3);
                u[i] = u1 + (k1 + 2.0 * (k2 + k3) + k4)/6.0;
                x0 = x0 + h;
            }
      else
                {   v1 = v[i-1];
                    h1 = h / 2.0;
                    k1 = h * f(x1, v1);
                    k2 = h * f(x1 + h1, v1 + 0.5 * k1);
                    k3 = h * f(x1 + h1, v1 + 0.5 * k2);
                    k4 = h * f(x1 + h, v1 + k3);
                    v[i] = v1 + (k1 + 2.0 * (k2 + k3) + k4)/6.0;
                    x1 = x1 + h;
                }
      }
  }
te = 16.0 * (v[nn] - u[n]) / 15.0;
fprintf(fp,"Step = %4.2f\n", 2.0*h);
fprintf(fp,"Solution at nodal points\n");
for (i = 1; i <= n; i++)
        fprintf(fp,"%11.7f", u[i]);
fprintf(fp,"\n");
fprintf(fp,"Step = %4.2f\n", h);
fprintf(fp,"Solution at nodal points\n");
for (i = 1; i <= 2 * n; i++)
  {
        if(i == n + 1)
            fprintf(fp,"\n");
        fprintf(fp,"%11.7f", v[i]);
  }
  fprintf(fp,"\n");
  fprintf(fp,"Estimate of truncation error at ");
  fprintf(fp,"xf = %12.5e\n",te);
  printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
  return 0;
  }
/*******************************************************/
float   f(x, y)
  float   x, y;
    {   float   fun;
        fun = - 2.0 * x * y * y;
```

```
            return(fun);
      }
/******************************************************/
Initial point x0 = 0.000000, initial value y0 = 1.000000
Number of intervals = 5,
Step = 0.10
Solution at nodal points
   0.9900990   0.9615382   0.9174306   0.8620682   0.7999992
Step = 0.05
Solution at nodal points
      0.9975063   0.9900990   0.9779951   0.9615384   0.9411764
      0.9174311   0.8908685   0.8620689   0.8316008   0.8000000
Estimate of truncation error at xf = 7.62939e-07
/******************************************************/
```

PROGRAM 19

```
/* MILNE'S METHOD FOR SOLVING FIRST ORDER IVP
Program to solve an IVP, dy/dx = f(x,y), y(x0) = y0, using Milne-Simpson method. The initial
values x0, y0, the final value xf and the step size h are to be read. Starting values are calcu-
lated using classical fourth order Runge-Kutta method. Adams-Bashforth method of third or-
der is used as a predictor and Milne-Simpson method is iterated till
abs(yold - ynew) <= err where err is error tolerance.       */

#include <stdio.h>
#include <math.h>

float f();

main()
   {
   float        x[21], y[21], k1, k2, k3, k4, x0, y0;
   float        h, f0, f1, f2, f3, x1, y1, p, yold, eps;
   int      n, i, miter, iter, niter, m;
   FILE          *fp;

   fp = fopen("result","w");

   printf("Input initial point x0, initial value y0\n");
   printf("number of steps m, step size h,\n");
   printf("error tolerance eps\n");
```

```
    scanf("%f %f %d %f %E", &x0, &y0, &m, &h, &eps);
    fprintf(fp,"Initial point = %f\n", x0);
    fprintf(fp,"Initial value = %f\n", y0);
    fprintf(fp, "Error tolerance = %e\n", eps);
    printf("Input maximum number of iterations per step\n");
    scanf("%d", &niter);
    fprintf(fp,"Maximum number of Milne iterations = ");
    fprintf(fp,"%d\n", niter);
    x[0] = x0;
    y[0] = y0;
    for (i = 1; i <= 2; i++)
        {   x1 = x[i - 1];
            y1 = y[i - 1];
            k1 = h * f(x1 , y1);
            k2 = h * f(x1 + 0.5 * h, y1 + 0.5 * k1);
            k3 = h * f(x1 + 0.5 * h, y1 + 0.5 * k2);
            k4 = h * f(x1 + h, y1 + k3);
            y[i] = y1 + (k1 + 2.0 * k2 + 2.0 * k3 + k4) / 6.0;
            x[i] = x1 + h;
        }
    miter = 0;
    for (i = 3; i <= m; i++)
        {   iter = 0;
            x1 = x[i - 1];
            y1 = y[i - 1];
            f0 = f(x[i - 3], y[i - 3]);
            f1 = f(x[i - 2], y[i - 2]);
            f2 = f(x1, y1);
            y[i] = y1 + h * (23.0 * f2 - 16.0 * f1 + 5.0 * f0) / 12.0;
            x[i] = x1 + h;
            p = y[i - 2] + h * (4.0 * f2 + f1) / 3.0;
l2:     yold = y[i];
            iter = iter + 1;
            miter = miter + 1;
            f3 = f(x[i], yold);
            y[i] = p + h * f3 / 3.0;
            if((fabs(yold - y[i]) - eps) <= 0)
              goto l3;
            if (iter > niter)
              {
              printf("Iteration bound is not sufficient");
```

```
                fprintf(fp,"Iteration bound is not sufficient");
                goto l1;
                }
            goto l2;
l3:     printf(" ");
            }
    fprintf(fp,"Step = %6.4f\n", h);
    fprintf(fp,"Total number of Milne correctors used = ");
    fprintf(fp,"%d\n", miter);
    fprintf(fp,"Solution at nodal points\n");
    for (i = 1; i <= m; i++)
        {
            fprintf(fp,"%11.7f", y[i]);
            if(i == 5)
                fprintf(fp,"\n");
        }
        printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
l1: fclose(fp);
    return 0;
    }
/******************************************************/
float   f(x, y)
    float   x, y;
        {  float   fun;
            fun = - 2.0 *x *y *y;
            return(fun);
        }
/******************************************************/
Initial point = 0.000000
Initial value = 1.000000
Error tolerance = 1.000000e-06
Maximum number of Milne iterations = 5
Step = 0.1000
Total number of Milne correctors used = 28
Solution at nodal points
    0.9900990   0.9615382   0.9174208   0.8620606   0.7999858
    0.7352871   0.6711303   0.6097542   0.5524795   0.5000020
/******************************************************/
```

PROGRAM 20

```
/* SHOOTING METHOD FOR SOLVING SECOND ORDER LINEAR BVP
Program to solve the linear two point boundary value problem
u” = p[x](du/dx) + q[x]u + r[x] = G(x, u, du/dx), u[a] = s1, u[b] = s2, by shooting method using the
super-position principle. The intial value problem is solved by the fourth order Runge-Kutta
method for 2x2 system. It requires two approximations of the slope of the solution curve at the
starting point of integration. The linear function G is given as a function subprogram.      */

#include <stdio.h>
#include <math.h>

float f();
float g();

main()
   {
   float      u[50], v[50], w[50], k[3][5], h, a, b, ya, yb, va;
   float      x0, x1, x2, u1, v1, c1, c2, app1, app2;
   int        n, i, j, ij;
   FILE       *fp;

   fp = fopen(“result”,“w”);

   printf(“Input end points of interval of integration ”);
   printf(“a & b,\nvalues at boundary points ya & yb,\n”);
   printf(“two approximations to the slope app1 & app2,\n”);
   printf(“number of intervals n\n”);
   scanf(“%f %f %f %f %f %f”, &a, &b, &ya, &yb, &app1, &app2);
   scanf(“%d”, &n);
   fprintf(fp,“End points are a = %4.2f, b = %4.2f\n”, a, b);
   fprintf(fp,“Values at boundary points are ya = %4.2f”, ya);
   fprintf(fp,“,yb = %4.2f\n”,yb);
   fprintf(fp,“Two approximations to the slope are:\n”);
   fprintf(fp,“app1 = %f, app2 = %f\n”, app1, app2);
   fprintf(fp,“Number of intervals = %d\n”, n);
   h = (b - a) / n;
   u[0] = ya;
   v[0] = app1;
   x0 = a;
   for (j = 1; j <= n; j++)
         {  x1 = x0 + h / 2.0;
            x2 = x0 + h;
```

```
            u1 = u[j-1];
            v1 = v[j-1];
            for (i = 1; i <= 2; i++)
               k[i][1] = h * f(i, x0, u1, v1);
            for (i = 1; i <= 2; i++)
               k[i][2] = h * f(i, x1, u1 + 0.5 * k[1][1], v1 + 0.5 * k[2][1]);
            for (i = 1; i <= 2; i++)
               k[i][3] = h * f(i, x1, u1 + 0.5 * k[1][2], v1 + 0.5 * k[2][2]);
            for (i = 1; i <= 2; i++)
               k[i][4] = h * f(i, x2, u1 + k[1][3], v1 + k[2][3]);
            u[j] = u1 + (k[1][1] + 2.0 * (k[1][2] + k[1][3]) + k[1][4]) / 6.0;
            v[j] = v1 + (k[2][1] + 2.0 * (k[2][2] + k[2][3] + k[2][4]) / 6.0;
            x0 = x0 + h;
         }
   w(0) = ya;
   v[0] = app2;
   x0 = a;
   for (j = 1; j <= n; j++)
         {  x1 = x0 + h / 2.0;
            x2 = x0 + h;
            u1 = w[j - 1];
            v1 = v[j - 1];
            for (i = 1; i <= 2; i++)
               k[i][1] = h * f(i, x0, u1, v1);
            for (i = 1; i <= 2; i++)
               k[i][2] = h * f(i, x1, u1 + 0.5 * k[1][1], v1 + 0.5 * k[2][1]);
            for (i = 1; i <= 2; i++)
               k[i][3] = h * f(i, x1, u1 + 0.5 * k[1][2], v1 + 0.5 * k[2][2]);
            for (i = 1; i <= 2; i++)
               k[i][4] = h * f(i, x2, u1 + k[1][3], v1 + k[2][3]);
            w[j] = u1 + (k[1][1] + 2.0 * (k[1][2] + k[1][3]) + k[1][4]) / 6.0;
            v[j] = v1 + (k[2][1] + 2.0 * (k[2][2] + k[2][3]) + k[2][4]) / 6.0;
            x0 = x0 + h;
         }
   c2 = (yb - u[n])/(w[n] - u[n]);
   c1 = 1.0 - c2;
   for (i = 1; i <= n; i++)
         u[i] = c1 * u[i] + c2 * w[i];
   fprintf(fp,"Step h = %6.2f\n", h);
   fprintf(fp,"Solution at nodal points\n");
   for (i = 1; i <= n - 1; i++)
```

```
        fprintf(fp,"%12.5e ", u[i]);
  fprintf(fp,"\n");
  printf("\nPLEASE SEE FILE 'result' FOR RESULTS\n\n");
  return 0;
  }
/*********************************************************/
float   f(i, x, z1, z2)
  float   x, z1, z2;
  int     i;
    {  float   fun;
         if(i == 1)
           fun = z2;
         else
             fun = g(x, z1, z2);
         return(fun);
    }
/*********************************************************/
float   g(xx, zz1, zz2)
  float   xx, zz1, zz2;
    {  float fung;
         fung = zz1 + xx;
       return(fung);
    }
/*********************************************************/
```

End points are a = 0.00, b = 1.00

Values at boundary points are ya = 0.00, yb = 0.00

Two approximations to the slope are:

app1 = 0.100000, app2 = 0.200000

Number of intervals = 5

Step h =   0.20

Solution at nodal points

-2.86791e-02   -5.04826e-02   -5.82589e-02   -4.42937e-02

/***/

Bibliography

The following is a brief list of texts on numerical methods. There are various other texts which are not reported here.

Ahlberg, J.H., E.N. Nilson, and J.L. Walsh, *Theory of Splines and Their Application*, Academic Press, New York, 1967.

Aiken, R.C. (ed), *Stiff Computation,* Oxford University Press ; New York, 1985.

Allgower, E.L., K. Glasshoff, and H.O. Peitgen (eds), *Numerical Solution of Nonlinear Equations,* Lecture Notes in Mathematics, 878, Springer Verlag, New York, 1981.

Atkinson, K, *Elementary Numerical Analysis,* Wiley, New York, 1985.

Aziz, A.K., (ed), *Numerical Solution of Differential Equations,*
Van Nostrand, New York, 1969.

Aziz, A.K. (ed), *Numerical Solution* of *BVP for Ordinary Differential Equations,*
Academic Press, New York, 1974.

Bartels, R., J. Beatty and B. Barsky, *An Introduction to Splines for Use in Computer Graphics and Geometric Modelling,* Morgan Kaufmann, Los Altos, Calif. 1987.

Burden, R.L., and J.D. Faires, *Numerical Analysis,* 4th ed, PWS-kent, 1989.

Butcher, J.C., *The Numerical Analysis of Ordinary Differential Equations : Runge-Kutta and General Linear Methods,* Wiley, New York, 1987.

Byrne, G.D., and C. A. Hall (eds), *Numerical Solution of Systems of Non-linear Algebraic Equations,* Academic Press, New York, 1973.

Collatz., L., *Numerical Treatment of Differential Equations*, 3rd ed.,
Springer Verlag, Berlin, 1966.

Conte, S.D., and C. deBoor, *Elementary Numerical Analysis : An Algorithmic Approach,* 3rd ed., McGraw-Hill, New York, 1980.

Dahlquist, G., and A. Björck, *Numerical Methods*,
Prentice-Hall, Engle-wood Cliffs, N.J., 1974.

David Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole, Calif., 1991.

Davis, P.J., *Interpolation* and *Approximation*, Blaisdell, New York, 1963.

Davis, P.J., and P. Rabinowitz, *Methods of Numerical Integration,* 2nd ed.,
Academic Press, New York, 1984.

Ferziger, J.H., *Numerical Methods for Engineering Application,*
John Wiley, New York, 1981.

Forsythe, G.E., and C.B. Moler, *Computer Solution of Linear Algebraic Equations,* Prentice-Hall, Englewood Cliffs, N.J., 1967.

Forsythe, G.E., M.A. Malcolm, and C.B. Moler, *Computer Methods for Mathematical Computations,* Prentice-Hall, Englewood Cliffs, N.J., 1977.

Fox, L., *Numerical Solution of Ordinary and Partial Differential Equations,*
Pergamon, London, 1962.

Fröberg, C.E., *Introduction to Numerical Analysis,*
Addison-Wesley, Reading, Mass., 1969.

Gear, C.W., *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

Gerald, C.F., and P.O. Wheatley, *Applied Numerical Analysis,* 4th ed.,
Addison Wesley, Reading, Mass., 1989.

Golub, G.H., and C.F. van Loan, *Matrix Computations,* 2nd ed.,
Johns Hopkins University Press, Baltimore 1989.

Gourlay, A. R., and G.A. Watson, *Computational, Methods for Matrix Eigen-problems,* Wiley, London, 1973.

Henrici, P., *Discrete Variable Methods in Ordinary Differential Equations,*
John Wiley, New York, 1962.

—, *Elements of Numerical Analysis,* John Wiley, New York, 1964.

Hildebrand, F.B., *Introduction to Numerical Analysis,*
McGraw-Hill, New York, London, 1956.

Householder, A.S., *Principles of Numerical Analysis,* McGraw-Hill, New York, 1953.

Isaacson, E., and H.B. Keller, *Analysis of Numerical Methods,*
John Wiley, NewYork, 1966.

Jain, M.K., *Numerical Solution of Differential Equations,* 2nd ed.,
Wiley Eastern Ltd., New Delhi, 1984.

Jain, M.K., S.R.K. Iyengar and R.K.Jain, *Numerical methods for Scientific and Engineering Computation*, 4th ed., New Age International Publishers, New Delhi, 2003.

James, M.L., G.M. Smith and J.C. Wolford, *Applied Numerical Methods for Digital Computation with FORTRAN AND CSMP,* 2nd ed.,
IEP-A Dun-Donnelley Publ., New York, 1977.

Johnson L.W. and R.D. Riess, *Numerical Analysis,* 2nd ed.
Addison-Wesley, Reading, Mass., 1982.

Johnston, R.L., *Numerical Methods : A Software Approach,* John Wiley, New York, 1982.

Keller, H.B., *Numerical Methods for Two-point Boundary Value Problems*,
Blaisdell, Waltham, Mass., 1968.

Kubiček, M. and V. Hlaváček, *Numerical Solution of Nonlinear Boundary Value Problems with Applications,* Prentice-Hall, Englewood Cliffs, N.J., 1983.

Kuo, S.S., *Computer Applications of Numerical Methods,*
Addison-Wesley, Reading, Mass., 1972.

Lambert, J.D., *Computational Methods in Ordinary Differential Equations.,*
John Wiley, New York, 1973.

Lapidus L., and J. Seinfeld, *Numerical Solution of Ordinary Differential Equations,* Academic Press, New York, 1971.

Marchuk, G., *Methods of Numerical Mathematics,* Springer, New York, 1975.

Na, T.Y., *Computational Methods in Engineering Boundary Value Problems,*
Academic Press, 1979.

Prager, W.H., *Applied Numerical Linear Algebra,* Prentice-Hall, N..J., 1988.

Ralston, A., and P. Rabinowitz, *A First Course in Numerical Analysis,* 2nd ed., McGraw-Hill, New York, 1978.

Rice, J.R., *Numerical Methods, Software and Analysis,* McGraw-Hill, New York, 1983.

Roberts, S.M. and J.S. Shipman, *Two Point Boundary Value Problems, Shooting Methods*, Elsevier, New York, 1972.

Scheid, F., *Numerical Analysis,* McGraw-Hill, New York, 1988.

Shlomo Breuer., and Zwas. G., *Numerical Mathematics — A Laboratory Approach,* Cambridge University Press, 1993.

Stewart, G.W., *Introduction to Matrix Computations,* Academic Press, New York, 1973.

Stroud, A.H. and D. Secrist, *Gaussian Quadrature Formulas,* Prentice-Hall, Englewood Cliffs, N.J., 1966.

Todd, J., *Survey of Numerical Analysis,* McGraw-Hill, New York, 1962.

Traub, J.F., *Iterative Methods for the Solutions of Nonlinear Equations,* Prentice-Hall, Englewood Cliffs, 1964.

Varga, R.S., *Matrix Iterative Analysis,* Prentice-Hall, Englewood Cliffs, N.J., 1962.

Wait, R., *The Numerical Solution of Algebraic Equations,* John Wiley, New York, 1979.

Wendroff, B., *Theoretical Numerical Analysis,* Academic Press, New York, 1966.

Yakowitz Sidney and Ferenc Szidarovszky, *An Introduction to Numerical Computations,* Macmillan Publishing Company, 1986.

Young, D.M., *Iterative Solution of Large Linear Systems,* Academic Press, New York, 1971.

Young, D.M. and R.T. Gregory, *A Survey of Numerical Mathematics,* Vol. 1, 2, Addison-Wesley, Reading, Mass., 1972.

Index

H

I

J

K

L

M

N