# LECTURE NOTES ON NUMERICAL ANALYSIS

Complied by: DR. GABBY

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

UNIVERSITY OF CAPE COAST

2021 EDITION

# Contents

1 Introduction To Numerical Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1.1 Definition 5

1.2 Significant digits 6

1.3 Accuracy and Precision 6

1.4 Rounding and Chopping 7

1.5 Absolute and Relative Errors 7

1.6 Taylor Series and Maclaurin Series 8

2 Numerical Solutions To Transcendental Equations . . . . . . . . . . . . . . . 12

2.1 Zeros or Roots of An Equation 12

2.1.1 Direct Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.1.2 Iterative Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.2 Finding The Root of an Equation 15

2.2.1 Bisection Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2.2.2 Method of False-Position . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2.2.3 Newton-Raphson Method . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

2.2.4 Secant Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

3 Numerical Solutions To Systems of Equations . . . . . . . . . . . . . . . . . . . . 27
3.1 Direct Methods 27
3.1.1 Gaussian Elimination Method . . . . . . . . . . . . . . . . . . . . . . . . 28
3.1.2 Gauss-Jordan Elimination Method . . . . . . . . . . . . . . . . . . . . . 31
3.2 Iterative Methods 33
3.2.1 Gauss-Jacobi Method . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
3.2.2 Gauss-Seidel Method . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
4 Numerical Differentiation and Integration . . . . . . . . . . . . . . . . . . . . . . . . 42
4.1 Numerical Differentiation 42
4.1.1 Newton's Forward Difference . . . . . . . . . . . . . . . . . . . . . . . 42
4.1.2 Newton's Backward Difference . . . . . . . . . . . . . . . . . . . . . . 45
4.2 Numerical Integration 47
4.2.1 Simple Trapezium Rule . . . . . . . . . . . . . . . . . . . . . . . . . . 48
4.2.2 Composite Trapezium Rule . . . . . . . . . . . . . . . . . . . . . . . . 49
4.2.3 The Simple Simpson's 1/3 Rule . . . . . . . . . . . . . . . . . . . . . . 50
4.2.4 The Composite Simpson's 1/3 Rule . . . . . . . . . . . . . . . . . . . . 51
5 Numerical Solutions of Ordinary Differential Equations: Single-Step Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53
5.1 Single-Step Methods 53
5.2 Multi-Step Methods 54
5.3 Solution To IVP's Using Single-Step Methods 55
5.3.1 Taylor Series of Order 1 or Euler Method . . . . . . . . . . . . . . . . . 55
5.3.2 Backward Euler . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55
5.3.3 Modified Euler or Midpoint Method . . . . . . . . . . . . . . . . . . . . 56
5.3.4 Trapezium Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
5.3.5 Heun's Method or Euler-Cauchy Method . . . . . . . . . . . . . . . . . 56
5.3.6 Runge-Kutta Methods . . . . . . . . . . . . . . . . . . . . . . . . . . 60
6 Numerical Solutions of Ordinary Differential Equations: Multi-step methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
6.1 Multi-step Predictor Method 65
6.1.1 Adams-Bashforth Scheme . . . . . . . . . . . . . . . . . . . . . . . . 65

6.2 Multi-step Corrector Methods 69

6.2.1 Adams-Moulton Method . . . . . . . . . . . . . . . . . . . . . . . . . 69

6.2.2 Milne-Simpson Method . . . . . . . . . . . . . . . . . . . . . . . . . 70

6.3 Predictor-Corrector Methods 71

Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75

# 1. Introduction To Numerical Analysis

1.1 Definition 5

1.2 Significant digits 6

1.3 Accuracy and Precision 6

1.4 Rounding and Chopping 7

1.5 Absolute and Relative Errors 7

1.6 Taylor Series and Maclaurin Series 8

## 1.1 Definition

Numerical analysis is the area of mathematics and computer science that creates, analyzes, and implements algorithms for solving numerically the problems of continuous mathematics. Such problems originate generally from real-world applications of algebra, geometry, and calculus, and they involve variables which vary continuously. The overall goal of the field is the design and analysis of techniques to give approximate but not exact solutions to problems that do may/may not have analytical solutions.

When an analytical solution does not exist, numerical techniques are employed to solving hard problems.

There are times that an analytical solution to a problem may exist but might be cumbersome and time wasting to find such solution. In such instances, it is better to resort to numerical solutions.

Most numerical solutions are iterative. This implies that a sequence of approximate solution is obtained by repeating a given procedure. These solutions are implemented using computer programs.

**Example 1.1.1** Analytically we can find the roots of

$$x^2 - 1 = 0$$

as

$$x = \pm 1$$

Since the highest degree of $x$ is two, then we have two roots.

**Example 1.1.2** Find all the five roots of the problem

$$x^5 - 1 = 0$$

It will be very difficult to find analytical (exact) solution to this problem, hence it would require that we find the approximate solution to the polynomial function using numerical methods.

The techniques for solving such polynomial problems will be discussed later in this course, including problems related to system of equation, differentiation, integration, ODE. Check to the course outline for details.

Moreover, people who employ numerical methods for solving problems have the following concerns

1. The rate of convergence: that is how long it take for the method(iterations) to find an answer.
2. The completeness of the answer: Is the solution unique or do other solutions exist excluding your obtained approximate solution.

## 1.2 Significant digits

There are digits beginning with the leftmost nonzero digit and ending with the rightmost correct digit, including final zeros that are exact.

1. All non-zero digits are considered significant. For example 91 has two significant figures, likewise 123.45 has five significant figures.
2. Zeros appearing anywhere between two non-zero digits are significant. Example 101.1203 has seven significant figures.
3. Leading zeros are not significant. Example 0.00053 has two significant figures.
4. Trailing zeros in a number containing a decimal point are significant. Example 12.2300 has six significant figures.

   Again 0.0001200 has four significant figures (the zeros before 1 are not significant).

   In addition 120.00 has five significant figures since it has three trailing zeroes.

## 1.3 Accuracy and Precision

Accurate to $n$ decimal places means that you can trust $n$ digits to the right of the decimal place.

Accurate to $n$ significant digits means that you can trust a total of $n$ digits as being meaningful beginning with the leftmost nonzero digit.

**Example 1.3.1**
12.356 has three decimal places but five significant figures.

## 1.4 Rounding and Chopping

We say that a number $x$ is **chopped** to $n$ digits or figures when all digits that follow the *nth* digit are discarded and none of the remaining $n$ digits are changed.

Conversely, $x$ is **rounded** to $n$ digits or figures when $x$ is replaced by an $n$-digit number that approximates $x$ with minimum error.

**Example 1.4.1** The results of rounding some three-decimal numbers to two digits are

$$0.217 \approx 0.22, \qquad 0.365 \approx 0.36, \qquad 0.475 \approx 0.48, \qquad 0.592 \approx 0.59,$$

while chopping them gives

$$0.217 \approx 0.21, \qquad 0.365 \approx 0.36, \qquad 0.475 \approx 0.47, \qquad 0.592 \approx 0.59.$$

## 1.5 Absolute and Relative Errors

Suppose that $\alpha$ and $\beta$ are two numbers, of which one is regarded as an approximation to the other. The error of $\beta$ as an approximation to $\alpha$ is

$$\alpha - \beta$$

that is, the error equals the exact value minus the approximate value.

**The absolute error** of $\beta$ as an approximation to $\alpha$ is

$$|\alpha - \beta|$$

**The relative error** of $\beta$ as an approximation to $\alpha$ is

$$\frac{|\alpha - \beta|}{|\alpha|}$$

Generally

$$\text{absolute error} = |\text{exact value - approximate value}| \tag{1.1}$$

and

$$\text{relative error} = \frac{|\text{exact value - approximate value}|}{|\text{exact value}|} \tag{1.2}$$

For practical reasons, the relative error is usually more meaningful than the absolute error.

**Example 1.5.1** If $x = 0.00347$ is rounded to $\tilde{x} = 0.0035$, what is its number of significant digits. Again find the absolute error, and relative error. Interpret the results.

**Solution**
$\tilde{x} = 0.0035$ has two significant digits

$$\begin{aligned}\text{absolute error} &= |\text{exact value - approximate value}|\\ &= |0.00347 - 0.0035|\\ &= |-0.00003|\\ &= 0.00003\end{aligned}$$

$$\text{relative error} = \frac{|\text{exact value - approximate value}|}{|\text{exact value}|} = \frac{0.00003}{|0.00347|} = 0.008646 \qquad (1.3)$$

Clearly, the relative error is a better indication of the number of significant digits than the absolute error

## 1.6 Taylor Series and Maclaurin Series

Most students will have encountered infinite series (particularly Taylor series) in their study of calculus. Consequently, this section is particularly important for numerical analysis, and deserves careful study. Once students are well grounded with a basic understanding of Taylor series we can proceed to study the fundamentals of numerical methods with better comprehension.

**Definition 1.1** Taylor series is a representation of a function as an **infinite sum of terms** that are calculated from the values of the function's derivatives at a single point.

Some examples include

1. $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$
2. $\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$
3. $\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$
4. $\cosh x = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots$
5. $\sinh x = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots$
6. $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots, \qquad |x| < 1$
7. $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \qquad -1 < x <\leq 1$

**Example 1.6.1** Use five terms of the Taylor series to approximate $\ln(1.1)$.

**Solution**
Taking $x = 0.1$, then the first five terms of the series for ln(1 + x) gives us

$$\ln(1.1) \approx 0.1 - \frac{0.01}{2} + \frac{0.001}{3} - \frac{0.0001}{4} + \frac{0.00001}{5} = 0.095310333\ldots$$

You can punch ln(1.1) directly on your calculator and compare your answer.

This value is correct to six decimal places of accuracy.

Generally, the Taylor series for a function $f$ about the point $c$ is given as:

$$f(x) \approx f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 + \cdots \approx \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!}(x-c)^n \quad (1.4)$$

$f', f'', \cdots f^{(n)}$ are the derivatives of the function $f$.

Alternatively, the Taylor series of a function could be written as

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!}(x-c)^n + E_{k+1} \quad (1.5)$$

where $E_{k+1}$ is the reminder or the error term. Note that "=" is used here instead of "≈".

In the special case where $c = 0$, then we obtain the **Maclaurin series** given by:

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots \approx \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n \quad (1.6)$$

**Example 1.6.2** What is the Taylor series of the function

$$f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5$$

at the point $c = 2$?

**Solution**
To compute the coefficients in the series, we need the numerical values of $f^{(n)}(2)$ for $n \geq 0$. Here are the details of the computation:

$$f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5 \qquad f(2) = 207 \quad (1.7)$$

$$f'(x) = 15x^4 - 8x^3 + 45x^2 + 26x - 12 \qquad f'(2) = 396 \quad (1.8)$$

$$f''(x) = 60x^3 - 24x^2 + 90x + 26 \qquad f''(2) = 590 \quad (1.9)$$

$$f'''(x) = 180x^2 - 48x + 90 \qquad f'''(2) = 714 \quad (1.10)$$

$$f^{(4)}(x) = 360x - 48 \qquad f^{(4)}(2) = 672 \quad (1.11)$$

$$f^{(5)}(x) = 360 \qquad f^{(5)}(2) = 360 \quad (1.12)$$

$$f^{(6)}(x) = 0 \qquad f^{(6)}(2) = 0 \quad (1.13)$$

Therefore, we have

$$\begin{aligned} f(x) &\approx f(c)+f'(c)(x-c)+\frac{f''(c)}{2!}(x-c)^2+\frac{f'''(c)}{3!}(x-c)^3+ \\ &\frac{f^{(4)}(c)}{4!}(x-c)^4+\frac{f^{(5)}(c)}{5!}(x-c)^5+\frac{f^{(6)}(c)}{6!}(x-c)^6 \\ &\approx f(2)+f'(2)(x-2)+\frac{f''(2)}{2!}(x-2)^2+\frac{f'''(2)}{3!}(x-2)^3+ \\ &\frac{f^{(4)}(2)}{4!}(x-2)^4+\frac{f^{(5)}(2)}{5!}(x-2)^5+\frac{f^{(6)}(2)}{6!}(x-2)^6 \\ &\approx 207+396(x-2)+295(x-2)^2+119(x-2)^3+28(x-2)^4+3(x-2)^5 \end{aligned}$$

**Example 1.6.3** Find the Taylor series and the Maclaurin series of the following function using the first four terms of the series.

$$f(x)=e^{2x}, \qquad \text{where } c=1$$

**Solution**

i. For the Taylor series we obtain

$$\begin{aligned} f(x)&=e^{2x} & f(1)&=e^2 \\ f'(x)&=2e^{2x} & f'(1)&=2e^2 \\ f''(x)&=4e^{2x} & f''(1)&=4e^2 \\ f'''(x)&=8e^{2x} & f'''(1)&=8e^2 \end{aligned}$$

$$\begin{aligned} f(x) &\approx f(c)+f'(c)(x-c)+\frac{f''(c)}{2!}(x-c)^2+\frac{f'''(c)}{3!}(x-c)^3 \\ &\approx f(1)+f'(1)(x-1)+\frac{f''(1)}{2!}(x-1)^2+\frac{f'''(1)}{3!}(x-1)^3 \\ &= e^2+2e^2(x-1)+\frac{4e^2}{2}(x-1)^2+\frac{8e^2}{6}(x-1)^3 \\ &= e^2\left[1+2x-2+2x^2-4x+2+\frac{4}{3}(x^3-3x^2+3x-1)\right] \\ &= e^2\left(\frac{7}{3}+2x-2x^2+\frac{4}{3}x^3\right) \end{aligned}$$

ii. The Maclaurin series is obtained with $c=0$. Thus

$$\begin{aligned} f(x)&=e^{2x} & f(0)&=e^0=1 \\ f'(x)&=2e^{2x} & f'(0)&=2e^0=2 \\ f''(x)&=4e^{2x} & f''(0)&=4e^0=4 \\ f'''(x)&=8e^{2x} & f'''(0)&=8e^0=8 \end{aligned}$$

$$\begin{aligned} f(x) &\approx f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 \\ &\approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 \\ &= 1 + 2x + 2x^2 + \frac{4}{3}x^3 \end{aligned}$$

**Exercise 1.1**

Find the Taylor series and the Maclaurin series of the following function using the first four terms of the series.

1. $$f(x) = \sin(2x), \qquad \text{where } c = \pi$$

2. $$f(x) = \cosh(3x), \qquad \text{where } c = 2$$

# 2. Numerical Solutions To Transcendental Equations

2.1 Zeros or Roots of An Equation 12
2.1.1 Direct Methods
2.1.2 Iterative Methods
2.2 Finding The Root of an Equation 15
2.2.1 Bisection Method
2.2.2 Method of False-Position
2.2.3 Newton-Raphson Method
2.2.4 Secant Method

A problem of great importance in science is determining the roots or zeros of a function.

**Definition 2.1** A polynomial equation of the form

$$f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n \tag{2.1}$$

is called an algebraic equation

**Definition 2.2** An equation which contains polynomial term, exponential term, logarithm term, and trigonometric term are called transcendental equations.

Some examples of transcendental equations are

1. $2xe^{2x-1} + 1 = 0$

2. $\cosh(x) + \cos(2x) + x^2 = 0$

3. $x^2 + \ln(2x) + \dfrac{1}{e^{x^2}} = 0$

## 2.1 Zeros or Roots of An Equation

**Definition 2.3** A number $\alpha$ for which $f(x) = 0$ is called the roots/zero of the function $f$. Geometrically, the roots of an equation is the value of $x$ where the graph crosses the $x-$axis.

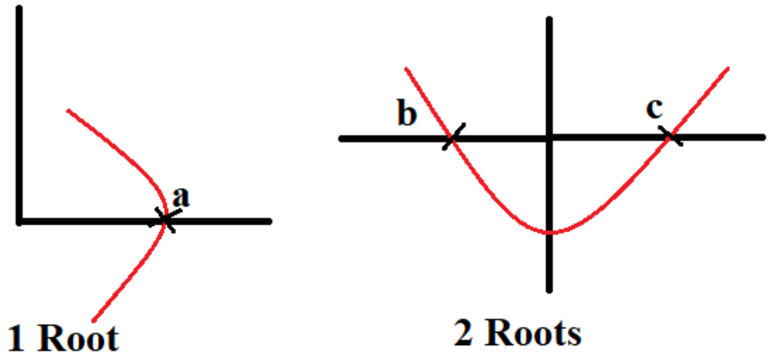


Figure 2.1: Roots of Equation

A polynomial equation of degree $n$ has exactly $n$ roots. These roots can either be real numbers, complex numbers or combination of real and complex numbers. Again, it can be a single root or multiple roots.

Some example $f(x) = 3x - 9$ has one root, and $f(x) = x^5 - 1$ will have five roots.

A transcendental equation may have one root, infinite number of roots, or no root.

The methods for finding the roots of an equation can be categorized under the following:

1. Direct methods
2. Iterative methods

### 2.1.1 Direct Methods

This gives the exact values of all the roots in a finite number of steps.

**Example 2.1.1** The roots of a quadratic equation: $ax^2 + bx + c = 0, \quad a \neq 0$ is

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

### 2.1.2 Iterative Methods

This is based on the idea of successive approximation. It starts with one or two initial approximations to the root in order to obtain the other sequences. This initial value are sometimes guessed. A sequence $x_k$ is said to converge to the exact root $\alpha$ if

$$\lim_{k\to\infty} x_k = \alpha \tag{2.2}$$

or

$$\lim_{k\to\infty} |x_k - \alpha| = 0 \tag{2.3}$$

Given an error tolerance $\epsilon$, an iterative procedure is terminated when

$$|x_{k+1} - x_k| \leq \epsilon \tag{2.4}$$

That is, a current solution value ($x_{k+1}$) minus the previous solution value ($x_k$) should be less or equal to a given threshold value ($\epsilon$).

This is often the **stopping criterion** for all iterative schemes.

### Initial Approximation of An Iterative Procedure

1. We count the number of changes of signs in the co-efficient of the given polynomial or function, the number of positive roots cannot exceed the number of changes in signs. If there are four changes in signs, then the function will have an even number of roots less or equal to four.

2. If a function is written as $f(-x) = 0$ and count the number of changes of signs in the co-efficient of the reduced function, the number of negative roots cannot exceed the number of changes in signs. If there are three changes in signs, then the function will have an odd number of roots less or equal to three.

**Theorem 2.1 (Intermediate Value Theorem)**
If $f(x)$ is continuous on the closed interval $[a,\ b]$ and $f(a) \times f(b) < 0$, then $f(x) = 0$ has at least one real root or an odd number of real roots in the open interval $(a,\ b)$.

**Example 2.1.2**
Determine the maximum number of positive and negative roots and the interval of length one unit in which the real roots lies in the following

$$8x^3 - 12x^2 - 2x + 3 = 0, \quad x \in [-2,\ 3]$$

**Solution**
Before we begin, let check the number of sign changes
The number of changes in the signs of the coefficients $(8, -12, -2, 3)$ is 2. Therefore, the equation has 2 or no positive roots.

Now, $f(-x) = -8x^3 - 12x^2 + 2x + 3$. The number of changes in signs in the coefficients $(-8, -12, 2, 3)$ is 1. Therefore, the equation has one negative root.

Note again that the equation will have a maximum of three roots.

So we begin by finding the functional values of the given problem within the given interval.

$$f(-2) = 8 * (-2)^3 - 12 * (-2)^2 - 2(-2) + 3 = -105$$

$$f(-1) = 8 * (-1)^3 - 12 * (-1)^2 - 2(-1) + 3 = -15$$

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

$$f(3) = 8 * 3^3 - 12 * 3^2 - 2(3) + 3 = 105$$

The other values are computed using this same procedure.

These values are presented in a tabular form as

| x | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| f(x) | -105 | -15 | 3 | -3 | 15 | 105 |

Now let determine the intervals that satisfy the intermediate value theorem.

1. $f(-2) \times f(-1) = -105 \times -15 \implies > 0$, hence, there is not root in the interval $(-2, -1)$
2. $f(-1) \times f(0) = -15 \times 3 \implies < 0$, hence, there is a root in the interval $(-1, 0)$
3. $f(0) \times f(1) = 3 \times -3 \implies < 0$, hence, there is a root in the interval $(0, 1)$
4. $f(1) \times f(2) = -3 \times 15 \implies < 0$, hence, there is a root in the interval $(1, 2)$
5. $f(2) \times f(3) = 15 \times 105 \implies > 0$, hence, there is no root in the interval $(2, 3)$

Considering the obtained results $(-1,\ 0),\ (0,\ 1),\ (1,\ 2)$, the function $f(x)$ will have one negative root and two positive roots.

**Exercise 2.1** Determine the maximum number of positive and negative roots and the interval of length one unit in which the real roots lies in the following

1. $3x^3 - 2x^2 - x + 3 = 0, \quad x \in [-3,\ 3]$
2. $xe^x - \cos(x) = 0, \quad x \in [-3,\ 3]$

## 2.2 Finding The Root of an Equation

There are several numerical methods for finding the root of the equation $f(x) = 0$. However, the following four methods will be considered in this course. The methods are

1. Bisection or interval halving method
2. Method of false position or chord method
3. Newton-Raphson method
4. Secant method

### 2.2.1 Bisection Method

The method is applicable to functions of the form $f(x) = 0$, where the function $f$ is continuous and defined on a closed interval $[a, b]$ and $f(a),\ f(b)$ have opposite signs. The function $f$ must have one root in the open interval $(a, b)$.

Suppose we need to find the root of $f(x) = 0$ given the error tolerance $\epsilon$, then the algorithm for the bisection method is as follows:

1. Find two numbers $a = x_0$ and $b = x_1$, for which $f$ has different signs. That is consider the interval $[a, b]$ or $[x_0, x_1]$.
2. Define $c$, such that $c = \dfrac{a+b}{2}$ or $c = \dfrac{x_0 + x_1}{2}$

3. If $b - c \leq \epsilon$, then accept $c$ as the root of the equation and stop the iteration, otherwise continue

4. If $f(a) \times f(c) \leq 0$, then set $c$ as the new $b$, otherwise set $c$ as the new $a$.

   Return to step two

□

The procedure is continued until the interval is sufficiently small. That is when $|x_{k+1} - x_k| \leq \epsilon$. The graphically representation of the method is illustrated with Figure 2.2.

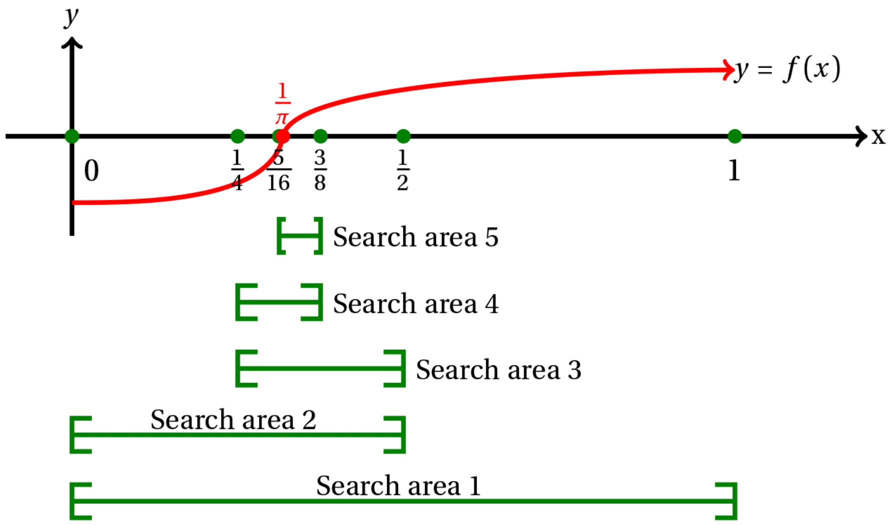


Figure 2.2: Bisection Method

### Advantage of the Bisection Method

1. The method is guaranteed to converge.

### Disadvantages of the Bisection Method

1. The method converges very slowly.
2. It cannot detect multiple roots.

**Example 2.2.1** Find the root of the equation

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

accurate to $\epsilon = 0.05$. Otherwise stop after the 5th iteration. Assume that the root lies in the interval $[0, 2]$.

**Solution**
**Iteration 1:**
Step 1: Considering the given interval, it implies that $a = 0$ and $b = 2$.

Step 2: $c = \dfrac{a+b}{2} = \dfrac{0+2}{2} = 1$

Step 3: Check stopping criterion: $b - c = 2 - 1 = 1 \implies \not< \epsilon$.
Hence we continue the iteration.

Step 4: $f(a) = f(0) = 0^2 + 2(0) - 3 = -3$

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

$f(a) \times f(c) = -3(0) = 0 \implies \leq 0$, hence set $c$ as new $b$

Therefore the new interval is $[0, 1]$

Return to step 2.

**Iteration 2:**
Considering the new interval, $a = 0$ and $b = 1$.

Step 2: $c = \dfrac{a+b}{2} = \dfrac{0+1}{2} = 0.5$

Step 3: Check stopping criterion: $b - c = 1 - 0.5 = 0.5 \implies \not< \epsilon$
Hence we continue the iteration.

Step 4: $f(a) = f(0) = 0^2 + 2(0) - 3 = -3$

$f(c) = f(0.5) = 0.5^2 + 2(0.5) - 3 = -1.75$

$f(a) \times f(c) = -3(-1.75) \implies > 0$, hence set $c$ as new $a$

Therefore the new interval is $[0.5, 1]$

Return to step 2.

**Iteration 3:**
Considering the new interval, $a = 0.5$ and $b = 1$.

Step 2: $c = \dfrac{a+b}{2} = \dfrac{0.5+1}{2} = 0.75$

Step 3: Check stopping criterion: $b - c = 1 - 0.75 = 0.75 \implies \not< \epsilon$
Hence we continue the iteration.

Step 4: $f(a) = f(0.5) = 0.5^2 + 2(0.5) - 3 = -1.75$

$f(c) = f(0.75) = 0.75^2 + 2(0.75) - 3 = -0.937$

$f(a) \times f(c) = -1.75(-0.936) \implies > 0$, hence set $c$ as new $a$

Therefore the new interval is $[0.75, 1]$

Return to step 2.

**Iteration 4:**
Considering the new interval, $a = 0.75$ and $b = 1$.

Step 2: $c = \dfrac{a+b}{2} = \dfrac{0.75+1}{2} = 0.875$

Step 3: Check stopping criterion: $b - c = 1 - 0.875 = 0.125 \implies \not< \epsilon$
Hence we continue the iteration.

Step 4: $f(a) = f(0.75) = 0.75^2 + 2(0.75) - 3 = -0.937$

$f(c) = f(0.875) = 0.875^2 + 2(0.875) - 3 = -0.47$

$f(a) \times f(c) = -0.937(-0.47) \implies > 0$, hence set $c$ as new $a$

Therefore the new interval is $[0.875, 1]$

Return to step 2.

**Iteration 5:**
Considering the new interval, $a = 0.875$ and $b = 1$.

Step 2: $c = \dfrac{a+b}{2} = \dfrac{0.875+1}{2} = 0.94$

Step 3: Check: $b - c = 1 - 0.94 = 0.06 \implies \not< \epsilon$

Though $b - c$ is not less then $\epsilon$, the otherwise statement in the question implies that the computations could be halted at the 5th iteration.

Hence $c = 0.94$ is root of the equation that lies in the interval $[0, 2]$

**Note 2.1** The function $f(x) = x^2 + 2x - 3$ have two roots, but the iterative scheme could find only one root at a time.

**Exercise 2.2** Using the interval halving or bisection method, find the root of the equation

$$x^6 - 1 = 0$$

that lies within the interval [1, 2] accurate to $\epsilon = 0.01$. Otherwise stop after the 6th iteration.

### 2.2.2 Method of False-Position

This method also requires the interval in which the root is expected to lie. The iterative formula is defined as

$$x_{k+1} = \frac{x_{k-1} \times f(x_k) - x_k \times f(x_{k-1})}{f(x_k) - f(x_{k-1})} \tag{2.5}$$

The iterative procedure begins to find $x_2$(that is $k = 1$), given that $x_0$ and $x_1$ are given.

That is, starting with the initial interval $[x_0, x_1]$ is which the root lies, then $x_2$ is computed as

$$x_2 = \frac{x_0 \times f(x_1) - x_1 \times f(x_0)}{f(x_1) - f(x_0)} \tag{2.6}$$

If $f(x_0) \times f(x_2) < 0$, then the root lies in the interval $(x_0, x_2)$, otherwise the root lies in the interval $(x_2, x_1)$.

To simplify the subsequent computations and iterations, we let $x_2 = x_1$ when the chosen interval is $(x_0, x_2)$ , likewise, we set $x_2 = x_0$ when the chosen interval is $(x_2, x_1)$. The could help us use the iterative formula (2.6) repeatedly without any complications.

The iteration is continued until the required accuracy criterion is satisfied. That is when $|x_{k+1} - x_k| \leq \epsilon$. $\epsilon$ is a given tolerance level.

The method of false-position is graphically illustrated using Figure 2.3

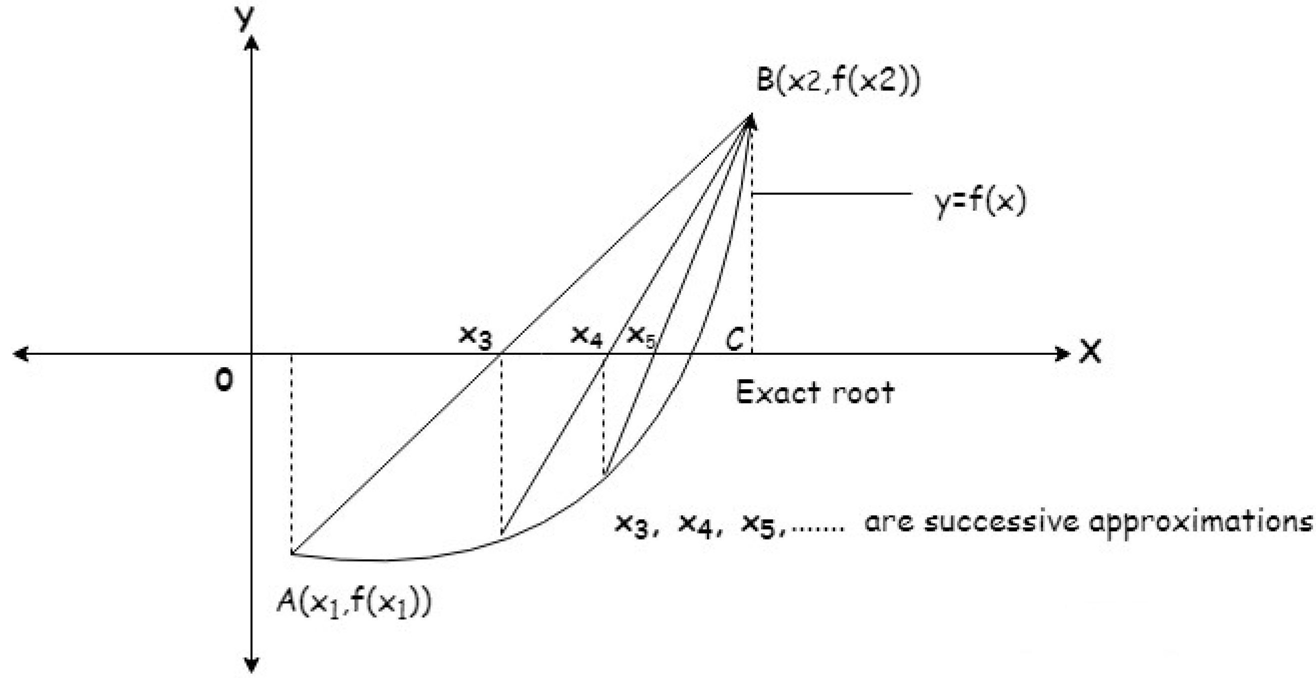


Figure 2.3: Method of False-Position

**Advantages of the Method False-Position**

1. The method is guaranteed to converge.
2. The method converges faster then the bisection method.

### Disadvantage of the Method False-Position

1. It cannot detect multiple roots.

**Example 2.2.2** Find the positive root of the function

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

accurate to $\epsilon = 0.05$ using the method of false-position. Otherwise stop after the 5th iteration. Take initial interval [0, 2]

**Solution**
**Iteration 1**
Step 1: We start with the initial interval $[x_0, x_1] = [0, 2]$. Then

$$x_2 = \frac{x_0 \times f(x_1) - x_1 \times f(x_0)}{f(x_1) - f(x_0)} \tag{2.7}$$

Since $x_0 = 0$ and $x_1 = 2$ are known, let compute their corresponding functional values. The details are not shown in this section, because it has been explained in the previous section. Hence

$$f(x_0) = f(0) = -3, \qquad f(x_1) = f(2) = 5$$

Thus

$$x_2 = \frac{x_0 \times f(x_1) - x_1 \times f(x_0)}{f(x_1) - f(x_0)} = \frac{0(5) - 2(-3)}{5 - (-3)} = \frac{6}{8} = 0.75$$

Step 2: Check stopping criterion: $|x_{k+1} - x_k| \leq \epsilon$.

Since this is the first iteration, we will skip this step. There is no previous iterative value to make such comparison.

Hence we continue the iteration.

Step 3: Check the new interval: $f(x_2) = f(0.75) = 0.75^2 + 2(0.75) - 3 = -0.9375$

Note that $f(x_0) = -3$

Therefore $f(x_0) \times f(x_2) = -0.9375(-3) = 2.8125 \implies > 0$.
Hence the root will lie in the interval $(x_2, x_1)$.

With proper substitution, the new interval is $[0.75, 2]$

**Iteration 2**
Step 1: $x_0 = 0.75$ and $x_1 = 2$

The corresponding functional values are

$$f(0.75) = -0.9375, \qquad f(2) = 5$$

Then

$$x_2 = \frac{x_0 \times f(x_1) - x_1 \times f(x_0)}{f(x_1) - f(x_0)} = \frac{0.75(5) - 2(-0.9375)}{5 - (-0.9375)} = 0.9494$$

Step 2: Check stopping criterion: $|x_{k+1} - x_k| = |0.9494 - 0.75| = 0.1994 \implies \not< \epsilon$
Hence we continue the iteration.

Step 3: Check the new interval: $f(x_2) = f(0.9494) = 0.9494^2 + 2(0.9494) - 3 = -0.2076$

Note that $f(x_0) = -0.9375$

Therefore $f(x_0) \times f(x_2) > 0$.
Hence the root will lie in the interval $(x_2, x_1)$.

With proper substitution the new interval is $[0.9494, 2]$

**Iteration 3**
Step 1: $x_0 = 0.9494$ and $x_1 = 2$

The corresponding functional values are

$$f(0.9494) = -0.2076, \qquad f(2) = 5$$

Then

$$x_2 = \frac{x_0 \times f(x_1) - x_1 \times f(x_0)}{f(x_1) - f(x_0)} = \frac{0.9494(5) - 2(-0.2076)}{5 - (-0.2076)} = 0.9894$$

Step 2: Check stopping criterion: $|x_{k+1} - x_k| = |0.9894 - 0.9494| = 0.04 \implies < \epsilon$

Since the stopping criterion is satisfied, we halt the iteration process here.

Hence the root of the equation is 0.9894.

**Exercise 2.3** Using the method of false position, find the root of the equation

$$x^6 - 1 = 0$$

that lies within the interval [1, 2] accurate to $\epsilon = 0.01$. Otherwise stop after the 6th iteration.

### 2.2.3 Newton-Raphson Method

This is also called the Newton's method. It approximates the curve near a root by a straight line. If $x_0$ is the initial approximation then $h(x_0, f(x_0))$ is a point on the curve. If we draw a tangent to the curve at a point $f$, the point of intersection of the tangent to the $x-$axis is taken as the next approximation to the root. The process is continued until the required accuracy criterion is obtained. That is when

$$|x_{k+1} - x_k| < \epsilon$$

.

The method is graphically illustrated with Figure 2.4.

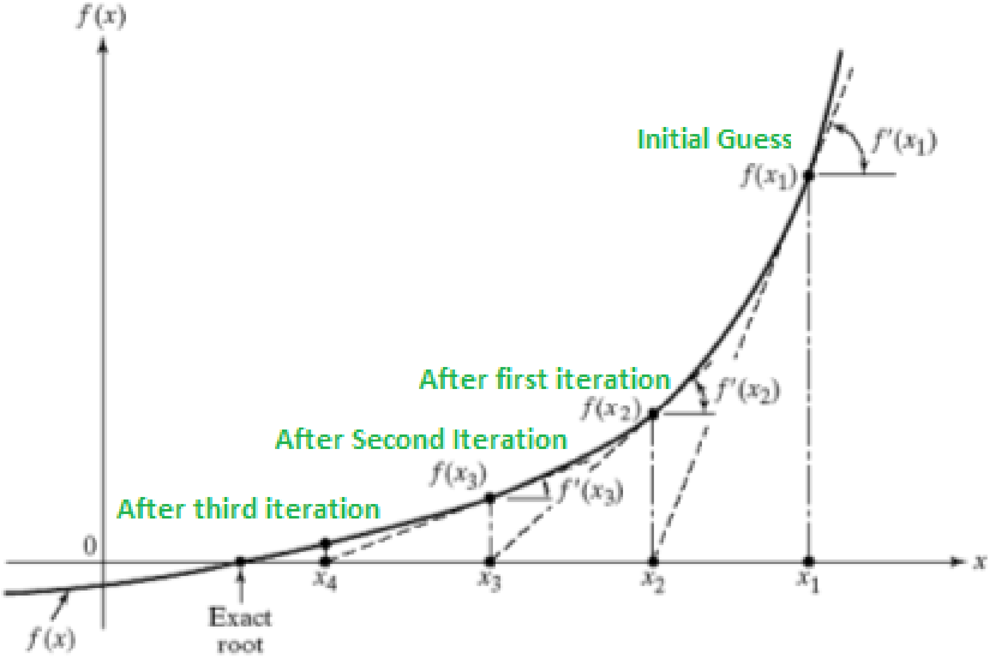


Figure 2.4: Newton-Raphson Method

The iterative procedure is explained as follows.

Given $x_0$, then $x_1, x_2, \cdots, x_{k+1}$ are obtained as:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \qquad \text{where } f'(x_0) \neq 0 \tag{2.8}$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}, \qquad \text{where } f'(x_1) \neq 0 \tag{2.9}$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}, \qquad \text{where } f'(x_2) \neq 0 \tag{2.10}$$

$$\vdots \qquad \vdots$$

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad \text{where } f'(x_k) \neq 0 \tag{2.11}$$

### Advantage of the Newton-Raphson Method

1. The method converges faster then the bisection and false position methods.

### Disadvantage of the Newton-Raphson Method

1. The method diverge if the initial approximation is far from the root.

**Example 2.2.3** Use the Newton's method to find the root of the function

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

that lies in the interval [0, 2]. Take $\epsilon = 0.05$ and $x_0 = 0$.

**Solution**
**Iteration 1: k = 0**
Given the formula

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad f'(x_0) \neq 0$$

Since we know $x_0 = 0$, we let $k = 0$, then

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \qquad f'(x_0) \neq 0 \tag{2.12}$$

We need to find $f(x_0)$ and $f'(x_0)$. Note that $f'$ is the derivative of the given function. Thus

$$f'(x) = 2x + 2$$

$$f'(x_0) = f'(0) = 2(0) + 2 = 2$$

$$f(x_0) = f(0) = 0^2 + 2(x) - 3 = -3$$

Substituting these values into equation (2.12).

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{(-3)}{2} = 1.5$$

Checking the stopping criterion:

Since this is the first iteration, we will skip this step. There is no previous iterative value to make such comparison.

Hence we continue to the next iteration and find $x_2$.

**Iteration 2: k = 1**
With $k = 1$, the formula reduces to

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

From the above, we know that $x_1 = 1.5$. Therefore the functional values are:

$$f'(x_1) = f'(1.5) = 2(1.5) + 2 = 5$$

$$f(x_1) = f(1.5) = 1.5^2 + 2(1.5) - 3 = 2.25$$

Substituting these values into the formula

$$x_2 = 1.5 - \frac{2.25}{5} = 1.05$$

Checking the stopping criterion: $|x_2 - x_1| = |1.05 - 1.5| = 0.45 \implies \not< \epsilon$.
Hence compute $x_3$.

**Iteration 3: k = 2**
With $k = 2$, the formula reduces to

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

From the above, we know that $x_2 = 1.05$. Therefore the functional values are:

$$f'(x_2) = f'(1.05) = 2(1.05) + 2 = 4.205$$

$$f(x_2) = f(1.05) = 1.05^2 + 2(1.05) - 3 = 0.2025$$

Substituting these values into the formula

$$x_3 = 1.05 - \frac{0.2025}{4.205} = 1.0018$$

Checking the stopping criterion: $|x_3 - x_2| = |1.0018 - 1.05| = 0.048 \implies < \epsilon$.
Hence stop the iteration.

Therefore the root of the equation is 1.0018

**Exercise 2.4** Using the Newton-Raphson method, find the root of the equation

$$x^6 - 1 = 0$$

with the initial approximation $x_0 = 1$ and accurate to $\epsilon = 0.01$. Otherwise stop after the 6th iteration.

### 2.2.4 Secant Method

Assuming we need to find the root of the equation $f(x) = 0$ which lies in the interval $[x_0, x_1]$, then the two points $(x_0, f(x_0))$ and $(x_1, f(x_1)$ form a straight line called the **secant line** which is viewed as the approximation to the graph of $f(x)$. The point where this secant line crosses the $x-$axis is considered as the root of the equation. The iteration is continued until the interval in which the root lies becomes significantly small. That is when

$$|x_{k+1} - x_k| < \epsilon$$

This method is similar to the method false position, but with different iterative procedure. The iterative procedure for the secant method is given by:

$$x_2 = x_1 - f(x_1)\left[\frac{x_1 - x_0}{f(x_1) - f(x_0)}\right] \tag{2.13}$$

$$x_3 = x_2 - f(x_2)\left[\frac{x_2 - x_1}{f(x_2) - f(x_1)}\right] \tag{2.14}$$

$$\vdots \qquad \vdots$$

$$x_{k+1} = x_k - f(x_k)\left[\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\right] \tag{2.15}$$

### Advantages of the Secant Method

1. The method is faster after few initial iterations.
2. Compared to the Newton's method, this does not require differentiation.

### Disadvantage of the Secant Method

1. It is slow compared to the Newton-Raphson method.

**Example 2.2.4** Use the Secant method to find the root of the function

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

that lies in the interval [0, 2]. Take $\epsilon = 0.06$.

**Solution**
**Iteration 1: k=1**
Given the general formula

$$x_{k+1} = x_k - f(x_k)\left[\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\right] \tag{2.16}$$

For $k = 1$, we have

$$x_2 = x_1 - f(x_1)\left[\frac{x_1 - x_0}{f(x_1) - f(x_0)}\right] \tag{2.17}$$

From the given interval $x_0 = 0$ and $x_1 = 2$. Therefore the functional values are

$$f(x_0) = f(0) = 0^2 + 2(0) - 3 = -3$$
$$f(x_1) = f(1) = 2^2 + 2(2) - 3 = 5$$

Substituting these value into the iterative formula (2.17), we obtain

$$x_2 = 2 - 5\left[\frac{2 - 0}{5 - (-3)}\right] = 2 - \frac{10}{8} = 0.75$$

Checking the stopping criterion:

Since this is the first iteration, we will skip this step. There is no previous iterative value to make such comparison.

Hence we continue the iteration.

**Iteration 2: k=2**
For $k = 2$, we have

$$x_3 = x_2 - f(x_2)\left[\frac{x_2 - x_1}{f(x_2) - f(x_1)}\right] \tag{2.18}$$

The functional values are

$$f(x_2) = f(0.75) = 0.75^2 + 2(0.75) - 3 = -0.9375$$
$$f(x_1) = f(2) = 2^2 + 2(2) - 3 = 5$$

Substituting these value into the iterative formula (2.18), we obtain

$$x_3 = 0.75 - (-0.9375)\left[\frac{0.75 - 2}{-0.9375 - 5}\right] = 0.75 + 0.197 = 0.947$$

Checking the stopping criterion: $|x_3 - x_2| = |0.947 - 0.75| = 0.197 \implies \not< \epsilon$.
Hence continue to find $x_4$.

**Iteration 3: k=3**
For $k = 3$, we have

$$x_4 = x_3 - f(x_3)\left[\frac{x_3 - x_2}{f(x_3) - f(x_2)}\right] \tag{2.19}$$

The functional values are

$$f(x_2) = f(0.75) = 0.75^2 + 2(0.75) - 3 = -0.9375$$
$$f(x_3) = f(0.947) = 0.947^2 + 2(0.947) - 3 = -0.21$$

Substituting these value into the iterative formula (2.18), we obtain

$$x_4 = 0.947 - (-0.21)\left[\frac{0.947 - 0.75}{-0.21 - (-0.9375)}\right] = 0.947 + 0.0568 = 1.0038$$

Checking the stopping criterion: $|x_4 - x_3| = |1.0038 - 0.947| = 0.0568 \implies < \epsilon$. Hence stop the iteration.

Therefore the root of the equation is 1.0038

**Exercise 2.5** Using the secant method, find the root of the equation

$$x^6 - 1 = 0$$

that lies within the interval [1, 2] accurate to $\epsilon = 0.01$. Otherwise stop after the 6th iteration.

# 3. Numerical Solutions To Systems of Equations

3.1 Direct Methods 27

3.1.1 Gaussian Elimination Method

3.1.2 Gauss-Jordan Elimination Method

3.2 Iterative Methods 33

3.2.1 Gauss-Jacobi Method

3.2.2 Gauss-Seidel Method

The method for solving system of equations can be classified into

1. Direct Method: This produce the exact solution after a finite number of steps. The direct methods considered in this course are
   (a) Gaussian Elimination Method
   (b) Gauss-Jordan Elimination Method
2. Indirect/Iterative Method: This is based on the method of successive approximations. It start with an initial approximation to the solution to obtain a sequence of approximate solutions. The direct methods considered in this course are
   (a) Gauss-Jacobi Method
   (b) Gauss-Seidel Method

## 3.1 Direct Methods

Consider a system of linear equations

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots a_{1n}x_n &= b_1\\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots a_{2n}x_n &= b_2\\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots a_{3n}x_n &= b_3\\
\vdots \qquad\qquad\qquad\qquad & \vdots \qquad \vdots\\
a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \cdots a_{mn}x_n &= b_m
\end{aligned}$$

Equation (3.1) can be recast as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \tag{3.1}$$

The $x_i's$ are the unknown to be determined. Thus, equation (3.1) is of the form

$$Ax = b \tag{3.2}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots a_{mn} \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and } b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix}$$

The matrix $A$ is appended by $b$ to form what we call the **Augmented Matrix**. This is denoted by

$$A|b = \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & \cdots a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots a_{mn} & b_m \end{array}\right] \tag{3.3}$$

The augmented matrix is solved using either a direct or an iterative numerical schemes. These schemes are aided by elementary row operations.

**Elementary Row Operation** are algebraic manipulations on the rows of a given matrix. This involves interchanging of any two rows, division and multiplication of any row by a non-zero constant. We begin by finding the solution to problem (3.3) using the direct methods.

### 3.1.1 Gaussian Elimination Method

This is based on the idea of reducing a given system of equation

$$Ax = b$$

to an **upper triangular matrix** of the form

$$Ux = z$$

using the technique of elementary row operations. $U$ is the upper triangular matrix, and $z$ is the new column vector on the right hand side.

With this technique, all the elements below the leading diagonal are reduced to zero. This is made possible by using the 'pivot element' to manipulate all values below it column. The reduced system $Ux = z$ is solved using back substitution.

For a given square matrix, the element on the leading diagonal are called the **pivot points.**

The Gaussian elimination method may fail when any one of the pivot points is zero or a very small number relative to the other values. To overcome such computational difficulty, we use a procedure called **Partial Pivoting** to solve the given problem. With this technique, first search through a given pivot column to find the largest number in magnitude. That number is used as the pivot by interchanging rows. The procedure is continued until an upper triangular matrix is obtained.

Let illustrate this method by considering an example.

**Example 3.1.1** Solve the following system of equations using the Gaussian elimination method

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 4x_1 + 3x_2 - x_3 &= 6 \\ 3x_1 + 5x_2 + 3x_3 &= 4 \end{aligned}$$

**Solution**
This problem is first recast into matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix} \tag{3.4}$$

The augmented matrix is deduced from (3.4) as

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 4 & 3 & -1 & 6 \\ 3 & 5 & 3 & 4 \end{array}\right] \tag{3.5}$$

Now, let reduce (3.5) to an upper triangular matrix using elementary row operations.

**Iteration 1**
The first pivot point is 1 in the first column (first term).

$$\left[\begin{array}{ccc|c} \boxed{1} & 1 & 1 & 1 \\ 4 & 3 & -1 & 6 \\ 3 & 5 & 3 & 4 \end{array}\right] \tag{3.6}$$

We are to reduce the values beneath 1, that is 4 and 3 to zeros using elementary row operations. The following manipulations are used here.
$NR_2 = 4R_1 - R_2, \qquad \Longrightarrow\ 4 \to 0$
$NR_3 = 3R_1 - R_3, \qquad \Longrightarrow\ 3 \to 0$

Note that these computations affect the entire row.

**Note 3.1**
NR is used to denote New Row, such that $NR_2$ is read ‘new row 2’.
R is used to denote Row, such that $R_1$ is read ‘row 1’.

Therefore the new matrix is

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 5 & -2 \\ 0 & -2 & 0 & -1 \end{array}\right] \tag{3.7}$$

**Iteration 2**

The second pivot point is 1 in the second column (diagonal value).

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & \boxed{1} & 5 & -2 \\ 0 & -2 & 0 & -1 \end{array}\right] \tag{3.8}$$

We are to reduce the value beneath 1, that is -2 to zero using elementary row operations. The following manipulations are used here.
$NR_3 = 2R_2 + R_3, \qquad \implies -2 \to 0$

Therefore the new matrix is

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 5 & -2 \\ 0 & 0 & 10 & -5 \end{array}\right] \tag{3.9}$$

Now we have an upper triangular matrix. So the solution could be finally obtained using back substitution. That is substitution and solving from the last row. We have

$$10x_3 = -5 \implies x_3 = -0.5$$

$$\begin{aligned} x_2 + 5x_3 &= -2, \qquad \text{but } x_3 = -0.5 \\ x_2 + 5(-0.5) &= -2 \\ x_2 &= 0.5 \end{aligned}$$

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + 0.5 - 0.5 &= 1 \\ x_1 &= 1 \end{aligned}$$

These are the solutions to the given problem.

**Exercise 3.1** Solving the following system of equations using the Gaussian elimination method. Apply partial pivoting when necessary.

1.

$$\begin{aligned} 10x + 4y - 2z &= 20 \\ 3x + 12y - z &= 28 \\ x + 4y + 7z &= 2 \end{aligned}$$

2.

$$\begin{aligned} 2a + b + c + d &= 2 \\ 4a + 2c + d &= 3 \\ 3a + 2b + 2c &= -1 \\ a + 3b + 2c + 6d &= 2 \end{aligned}$$

### 3.1.2 Gauss-Jordan Elimination Method

This method is based on the idea of reducing the given system of equation

$$Ax = b$$

to a **diagonal matrix** of the form

$$Dx = z$$

$D$ is the diagonal matrix, and $z$ is the new column vector on the right hand side.

All solution techniques of the Gaussian elimination method do apply here. While a given problem is reduced to an upper triangular matrix in the case of Gaussian elimination, it is reduced to a diagonal matrix in the case of Gauss-Jordan elimination.

Let illustrate this method by considering an example.

**Example 3.1.2** Solve the following system of equations using the Gauss-Jordan elimination method

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 4x_1 + 3x_2 - x_3 &= 6 \\ 3x_1 + 5x_2 + 3x_3 &= 4 \end{aligned}$$

**Solution**
This problem is first recast into matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix} \tag{3.10}$$

The augmented matrix is deduced from (3.10) as

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 4 & 3 & -1 & 6 \\ 3 & 5 & 3 & 4 \end{array}\right] \tag{3.11}$$

Note that we are to reduce value above and beneath the leading diagonals to zeros.

Now, let reduce (3.11) to an upper triangular matrix using elementary row operations.

So simplicity we will pick it up from the Gaussian solution above. That is the iterations 1 and 2 for this question. The upper triangular matrix is given by

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 5 & -2 \\ 0 & 0 & 10 & -5 \end{array}\right] \tag{3.12}$$

**Iteration 3**

Since we need a diagonal matrix, we will start manipulating the other non-diagonal matrix to zero starting from the leading diagonal of the last column.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 5 & -2 \\ 0 & 0 & \boxed{10} & -5 \end{array}\right] \tag{3.13}$$

We are to reduce the values above 10, that is 5 and 1 to zeros elementary row operations. The following manipulations are used here.
$NR_2 = R_3 - 2R_2, \qquad \implies 5 \to 0$
$NR_1 = R_3 - 10R_1, \qquad \implies 1 \to 0$

The new matrix is

$$\left[\begin{array}{ccc|c} -10 & -10 & 0 & -15 \\ 0 & -2 & 0 & -1 \\ 0 & 0 & 10 & -5 \end{array}\right] \tag{3.14}$$

**Iteration 4**

The next pivot point is -2 in the second column (diagonal value).

$$\left[\begin{array}{ccc|c} -10 & -10 & 0 & -15 \\ 0 & \boxed{-2} & 0 & -1 \\ 0 & 0 & 10 & -5 \end{array}\right] \tag{3.15}$$

We are to reduce the value above $-2$, that is $-10$ to zero using elementary row operations. The following manipulations are used here.
$NR_1 = -5R_2 + R_1, \qquad \implies -10 \to 0$

The new matrix

$$\left[\begin{array}{ccc|c} -10 & 0 & 0 & -10 \\ 0 & -2 & 0 & -1 \\ 0 & 0 & 10 & -5 \end{array}\right] \tag{3.16}$$

Since we have reduced the system to a diagonal matrix, the values of $x$ can be obtained using direct substitution. That is

$$-10x_1 = -10 \implies x_1 = 1$$
$$-2x_2 = -1 \implies x_2 = 0.5$$
$$10x_3 = -5 \implies x_3 = -0.5$$

This solves the given problem.

**Exercise 3.2** Solving the following system of equations using the Gauss-Jordan elimination method. Apply partial pivoting when necessary.

1.

$$\begin{aligned} 10x + 4y - 2z &= 20 \\ 3x + 12y - z &= 28 \\ x + 4y + 7z &= 2 \end{aligned}$$

2.

$$\begin{aligned} 2a + b + c + d &= 2 \\ 4a + 2c + d &= 3 \\ 3a + 2b + 2c &= -1 \\ a + 3b + 2c + 6d &= 2 \end{aligned}$$

## 3.2 Iterative Methods

These are based on the idea of successive approximations. Given an initial solution $x_0$ you can solve the system of equation $Ax = b$ to obtain the approximate solution $x_1, x_2, x_3, \cdots, x_k$. We stop the iteration process when $|x_{k+1} - x_k| < \epsilon$, where $\epsilon$ is the error term. The following iterative methods are considered to solve system of equations.

### 3.2.1 Gauss-Jacobi Method

This method is sometimes called the Jacobi method. The method assumes that the entries of the leading diagonal of a matrix $A$ cannot be zero, that is $a_{ii} \neq 0$.

Given the following system of linear equations

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \tag{3.17}$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \tag{3.18}$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \tag{3.19}$$

The Jacobi method makes the coefficient of the leading diagonal of each equation the subject. Therefore equation (3.17) reduces to (3.20), (3.18) reduces to (3.21), and (3.19) reduces to (3.22).

$$x_1 = \frac{1}{a_{11}}\left[b_1 - a_{12}x_2 - a_{13}x_3\right] \tag{3.20}$$
$$x_2 = \frac{1}{a_{22}}\left[b_2 - a_{21}x_1 - a_{23}x_3\right] \tag{3.21}$$
$$x_3 = \frac{1}{a_{33}}\left[b_3 - a_{31}x_1 - a_{32}x_2\right] \tag{3.22}$$

Iteratively, the equations (3.20) to (3.22) becomes

$$x_1^{(k+1)} = \frac{1}{a_{11}}\left[b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right] \tag{3.23}$$
$$x_2^{(k+1)} = \frac{1}{a_{22}}\left[b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}\right] \tag{3.24}$$
$$x_3^{(k+1)} = \frac{1}{a_{33}}\left[b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}\right] \tag{3.25}$$

Equations (3.23) to (3.25) are the iterative formula for the Gauss-Jacobi method for solving three equation with three unknowns. The method can be generalized for higher order system of equations.

### Convergence and Diagonal Dominance

The method is guaranteed to converge if the coefficient of the matrix $A$ is diagonally dominant. That is

$$|a_{ii}| \geq \sum_{j=1}^{n} |a_{ij}|, \quad i \neq j \tag{3.26}$$

If the system is not diagonally dominant, we may exchange the equations if possible.

**Example 3.2.1** Solve the following system of equation using the Gauss-Jacobi method taking the origin as the initial solution.

$$2x + y = 4 \tag{3.27}$$
$$x - y = 5 \tag{3.28}$$

Take $\epsilon = 0.05$, otherwise stop on the 5th iteration

**Solution**

We first have to identify the leading diagonals based on the arrangement of the equations. Again, we inquire if the system is diagonal dominant. Yes, it is indeed diagonal dominant. So we a sure that the solution will converge.

The leading diagonals are $x$ in equation (3.27), and $y$ in equation (3.28).

$$2\boxed{x} + y = 4 \tag{3.29}$$
$$x - \boxed{y} = 5 \tag{3.30}$$

Making these variables the subject we have

$$x = \frac{1}{2}[4 - y] \tag{3.31}$$
$$y = x - 5 \tag{3.32}$$

Iteratively we have

$$x^{(k+1)} = \frac{1}{2}[4 - y^{(k)}] \tag{3.33}$$
$$y^{(k+1)} = x^{(k)} - 5 \tag{3.34}$$

**Iteration 1: when k =0**

$$x^{(1)} = \frac{1}{2}[4 - y^{(0)}]$$
$$y^{(1)} = x^{(0)} - 5$$

The initial approximations are the values at the origin, therefore

$$x^{(0)} = 0, \quad y^{(0)} = 0$$

Substituting the initial values, we have

$$x^{(1)} = \frac{1}{2}[4 - 0] = 2$$
$$y^{(1)} = 0 - 5 = -5$$

Check stopping criterion:

Since this is the first iteration, we will skip this step. There is no previous iterative value to make such comparison.

Hence we continue the iteration.

**Note 3.2** The following are not equivalent

$$x^0 \neq x^{(0)}, \quad x^1 \neq x^{(1)}, \quad \text{and } x^2 \neq x^{(2)}$$

The former are exponents, while the latter are iterative numbers.

**Iteration 2: when k =1**

$$x^{(2)} = \frac{1}{2}[4 - y^{(1)}]$$
$$y^{(2)} = x^{(1)} - 5$$

From iteration one

$$x^{(1)} = 2, \quad y^{(1)} = -5$$

Substituting these values, we have

$$x^{(2)} = \frac{1}{2}[4 - (-5)] = 4.5$$
$$y^{(2)} = 2 - 5 = -3$$

Check stopping criterion:
$|x^{(2)} - x^{(1)}| = |4.5 - 2| = 2.5 \implies \not< \epsilon$

$|y^{(2)} - y^{(1)}| = |-3 - (-5)| = 2 \implies \not< \epsilon$
Hence move to the next iteration.

**Iteration 3: when k =2**

$$\begin{aligned} x^{(3)} &= \frac{1}{2}[4 - y^{(2)}] = \frac{1}{2}[4 - (-3)] = 3.5 \\ y^{(3)} &= x^{(2)} - 5 = 4.5 - 5 = -0.5 \end{aligned}$$

Check stopping criterion:
$|x^{(3)} - x^{(2)}| = |3.5 - 4.5| = 1 \implies \not< \epsilon$
$|y^{(3)} - y^{(2)}| = |-0.5 - (-3)| = 2.5 \implies \not< \epsilon$
Hence move to the next iteration.

**Iteration 4: when k =3**

$$\begin{aligned} x^{(4)} &= \frac{1}{2}[4 - y^{(3)}] = \frac{1}{2}[4 - (-0.5)] = 2.25 \\ y^{(4)} &= x^{(3)} - 5 = 3.5 - 5 = -1.5 \end{aligned}$$

Check stopping criterion:
$|x^{(4)} - x^{(3)}| = |2.25 - 3.5| = 1.25 \implies \not< \epsilon$
$|y^{(4)} - y^{(3)}| = |-1.5 - (-0.5)| = 1 \implies \not< \epsilon$
Hence move to the next iteration.

**Iteration 5: when k =4**

$$\begin{aligned} x^{(5)} &= \frac{1}{2}[4 - y^{(4)}] = \frac{1}{2}[4 - (-1.5)] = 2.75 \\ y^{(5)} &= x^{(4)} - 5 = 2.25 - 5 = -2.75 \end{aligned}$$

Based on the otherwise condition we halt the iteration process here. Therefore the solution to the system is

$$x = 2.75, \quad y = -2.75$$

Let consider a 3 × 3 example.

**Example 3.2.2** Solve the system of equation

$$4x_1 + x_2 + x_3 = 2 \tag{3.35}$$
$$x_1 + 5x_2 + 2x_3 = -6 \tag{3.36}$$
$$x_1 + 2x_2 + 3x_3 = -4 \tag{3.37}$$

using the Jacobi method by performing five iterations. Use the initial approximations $[x_1,\ x_2,\ x_3] = [0,\ 0,\ 0]$

**Solution**
We first have to identify the leading diagonals based on the arrangement of the equations. Again, we inquire if the system is diagonal dominant. Yes it is indeed diagonal dominant.
The leading diagonals are $x_1$ in equation (3.35), $x_2$ in equation (3.36), and $x_3$ in equation (3.37).

$$4\boxed{x_1} + x_2 + x_3 = 2 \tag{3.38}$$
$$x_1 + 5\boxed{x_2} + 2x_3 = -6 \tag{3.39}$$
$$x_1 + 2x_2 + 3\boxed{x_3} = -4 \tag{3.40}$$

Making these variables the subject we have

$$x_1 = \frac{1}{4}[2 - (x_2 + x_3)] \tag{3.41}$$
$$x_2 = \frac{1}{5}[-6 - (x_1 + 2x_3)] \tag{3.42}$$
$$x_3 = \frac{1}{3}[-4 - (x_1 + 2x_2)] \tag{3.43}$$

Iteratively we have

$$x_1^{(k+1)} = 0.25[2 - (x_2^{(k)} + x_3^{(k)})] \tag{3.44}$$
$$x_2^{(k+1)} = 0.2[-6 - (x_1^{(k)} + 2x_3^{(k)})] \tag{3.45}$$
$$x_3^{(k+1)} = 0.33[-4 - (x_1^{(k)} + 2x_2^{(k)})] \tag{3.46}$$

**Iteration 1: when k= 0**
From the initial conditions we have $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0$.
Substituting

$$x_1^{(1)} = 0.25[2 - (x_2^{(0)} + x_3^{(0)})] = 0.25(2) = 0.5$$
$$x_2^{(1)} = 0.2[-6 - (x_1^{(0)} + 2x_3^{(0)})] = 0.2(-6) = -1.2$$
$$x_3^{(1)} = 0.3333[-4 - (x_1^{(0)} + 2x_2^{(0)})] = 0.33(-4) = -1.3333$$

Since we are required to perform five iterations there is no need checking the stopping criterion.

**Iteration 2: when k= 1**

$$x_1^{(2)} = 0.25[2 - (x_2^{(1)} + x_3^{(1)})] = 0.25[2 - (-1.2 - 1.33333)] = 1.13333$$
$$x_2^{(2)} = 0.2[-6 - (x_1^{(1)} + 2x_3^{(1)})] = 0.2[-6 - (0.5 + 2(-1.33333))] = -0.76668$$
$$x_3^{(2)} = 0.33[-4 - (x_1^{(1)} + 2x_2^{(1)})] = 0.33333[-4 - (0.5 + 2(-1.2))] = -0.7$$

**Iteration 3: when k= 2**

$$x_1^{(3)} = 0.25[2 - (x_2^{(2)} + x_3^{(2)})] = 0.25[2 - (-0.76668 - 0.7)] = 0.86667$$
$$x_2^{(3)} = 0.2[-6 - (x_1^{(2)} + 2x_3^{(2)})] = 0.2[-6 - (1.13333 + 2(-0.7))] = -1.14667$$
$$x_3^{(3)} = 0.33[-4 - (x_1^{(2)} + 2x_2^{(2)})] = 0.33333[-4 - (1.13333 + 2(-0.76668))] = -1.19998$$

**Iteration 4: when k= 3**

$$x_1^{(4)} = 0.25[2 - (x_2^{(3)} + x_3^{(3)})] = 0.25[2 - (-1.14667 - 1.19999)] = 1.08666$$
$$x_2^{(4)} = 0.2[-6 - (x_1^{(3)} + 2x_3^{(3)})] = 0.2[-6 - (0.86667 + 2(-1.19998))] = -0.89334$$
$$x_3^{(4)} = 0.33[-4 - (x_1^{(3)} + 2x_2^{(3)})] = 0.33333[-4 - (0.86667 + 2(-1.14667))] = -0.85777$$

**Iteration 5: when k= 4**

$$\begin{aligned}
x_1^{(5)} &= 0.25[2 - (x_2^{(4)} + x_3^{(4)})] = 0.25[2 - (-0.89334 - 0.85777)] = 0.93778\\
x_2^{(5)} &= 0.2[-6 - (x_1^{(4)} + 2x_3^{(4)})] = 0.2[-6 - (1.08666 + 2(-0.85777))] = -1.07422\\
x_3^{(5)} &= 0.33[-4 - (x_1^{(4)} + 2x_2^{(4)})] = 0.33333[-4 - (1.08666 + 2(-0.89334))] = -1.09998
\end{aligned}$$

Therefore the system has the solution

$$x_1 = 0.93778,\ x_2 = -1.07422,\ x_3 = -1.0998$$

**Exercise 3.3** Solving the following system of equations using the Gauss-Jacobi method with the origin as the initial solution. Take $\epsilon = 0.01$, otherwise stop on the 5th iteration

1.

$$\begin{aligned}
10x + 4y - 2z &= 20\\
3x + 12y - z &= 28\\
x + 4y + 7z &= 2
\end{aligned}$$

2.

$$\begin{aligned}
2a + b + c + d &= 2\\
4a + 2c + d &= 3\\
3a + 2b + 2c &= -1\\
a + 3b + 2c + 6d &= 2
\end{aligned}$$

### 3.2.2 Gauss-Seidel Method

This method is similar to the Gauss-Jacobi method. However, this uses a continual iterative technique. Given the system of equation below

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \tag{3.47}$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \tag{3.48}$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \tag{3.49}$$

The Gauss-Seidel method likewise makes the coefficient of the leading diagonal of each equation the subject. Therefore equation (3.47) reduces to (3.50), (3.48) reduces to (3.51), and (3.49) reduces to (3.52).

$$x_1 = \frac{1}{a_{11}}\left[b_1 - a_{12}x_2 - a_{13}x_3\right] \tag{3.50}$$
$$x_2 = \frac{1}{a_{22}}\left[b_2 - a_{21}x_1 - a_{23}x_3\right] \tag{3.51}$$
$$x_3 = \frac{1}{a_{33}}\left[b_3 - a_{31}x_1 - a_{32}x_2\right] \tag{3.52}$$

However the continual iterative method of the Gauss-Seidel approach is deduced from equations (3.50) to (3.52) as

$$x_1^{(k+1)} = \frac{1}{a_{11}}\left[b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right] \tag{3.53}$$

$$x_2^{(k+1)} = \frac{1}{a_{22}}\left[b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right] \tag{3.54}$$

$$x_3^{(k+1)} = \frac{1}{a_{33}}\left[b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right] \tag{3.55}$$

Equations (3.53) to (3.55) are the iterative formula for the Gauss-Seidel method for solving three equation with three unknowns. The method can be generalized for higher order system of equations.

### Convergence and Diagonal Dominance

The method is guaranteed to converge if the coefficient of the matrix $A$ is diagonally dominant. That is

$$|a_{ii}| \geq \sum_{j=1}^{n} |a_{ij}|, \quad i \neq j \tag{3.56}$$

If the system is not diagonally dominant, we may exchange the equations if possible.

If both methods converges, then the Gauss-Seidel method converges at least two times faster than the Gauss-Jacobi method.

**Example 3.2.3** Solve the following system of equation

$$20x + y - 2z = 17 \tag{3.57}$$

$$3x + 20y - z = -18 \tag{3.58}$$

$$2x - 3y + 20z = 25 \tag{3.59}$$

using the Gauss-Seidel method with the initial solution

$$x^{(0)} = y^{(0)} = z^{(0)} = 1$$

Take $\epsilon = 0.01$, otherwise stop on the 5th iteration

**Solution**

We first have to identify the leading diagonals based on the arrangement of the equations. Again, we inquire if the system is diagonal dominant. Yes, it is indeed diagonal dominant

The leading diagonals are $x$ in equation (3.57), $y$ in equation (3.58) and $z$ in equation (3.58).

$$20\boxed{x} + y - 2z = 17 \tag{3.60}$$

$$3x + 20\boxed{y} - z = -18 \tag{3.61}$$

$$2x - 3y + 20\boxed{z} = 25 \tag{3.62}$$

Making these variables the subject we have

$$x = \frac{1}{20}[17 - y + 2z] \tag{3.63}$$
$$y = \frac{1}{20}[-18 - 3x + z] \tag{3.64}$$
$$z = \frac{1}{20}[25 - 2x + 3y] \tag{3.65}$$

Iteratively we have

$$x^{(k+1)} = \frac{1}{20}[17 - y^{(k)} + 2z^{(k)}] \tag{3.66}$$
$$y^{(k+1)} = \frac{1}{20}[-18 - 3x^{(k+1)} + z^{(k)}] \tag{3.67}$$
$$z^{(k+1)} = \frac{1}{20}[25 - 2x^{(k+1)} + 3y^{(k+1)}] \tag{3.68}$$

**Iteration 1: when k =0**
The initial solution is

$$x^{(0)} = 1,\ y^{(0)} = 1,\ z^{(0)} = 1$$

$$\begin{aligned} x^{(1)} &= \frac{1}{20}[17 - y^{(0)} + 2z^{(0)}] = 0.05[17 - 1 - 2(1)] = 0.9 \\ y^{(1)} &= \frac{1}{20}[-18 - 3x^{(1)} + z^{(0)}] = 0.05[-18 - 3(0.9) + 1] = -0.9895 \\ z^{(1)} &= \frac{1}{20}[25 - 2x^{(1)} + 3y^{(1)}] = 0.05[25 - 2(0.9) + 3(-0.9895)] = 1.01225 \end{aligned}$$

**Note 3.3** Be meticulous on how this continual substitution is done.

Since this is the very first iteration, we may not worry ourselves checking the stopping criterion.

**Iteration 2: when k =1**

$$\begin{aligned} x^{(2)} &= \frac{1}{20}[17 - y^{(1)} + 2z^{(1)}] = 0.05[17 - 1(-0.9895) - 2(1.01225)] = 1.00475 \\ y^{(2)} &= \frac{1}{20}[-18 - 3x^{(2)} + z^{(1)}] = 0.05[-18 - 3(1.00475) + 1.01225] = -0.999 \\ z^{(2)} &= \frac{1}{20}[25 - 2x^{(2)} + 3y^{(2)}] = 0.05[25 - 2(1.00475) + 3(-0.999)] = 1.000 \end{aligned}$$

Check stopping criterion:
$|x^{(2)} - x^{(1)}| = |1.00475 - 0.9| = 0.10475 \implies \not< \epsilon$
$|y^{(2)} - y^{(1)}| = |-0.999 - (-0.9895)| = 0.0095 \implies < \epsilon$
$|z^{(2)} - z^{(1)}| = |1.01225 - 1.000| = 0.01225 \implies \not< \epsilon$
Since two of these variables do not satisfy the condition we move to the next iteration.

**Iteration 3: when k =2**

$$\begin{aligned}
x^{(3)} &= \frac{1}{20}[17 - y^{(2)} + 2z^{(2)}] = 0.05[17 - 1(-0.999) - 2(1.0)] = 1.00 \\
y^{(3)} &= \frac{1}{20}[-18 - 3x^{(3)} + z^{(2)}] = 0.05[-18 - 3(1.00) + 1.00] = -1 \\
z^{(3)} &= \frac{1}{20}[25 - 2x^{(3)} + 3y^{(3)}] = 0.05[25 - 2(1.00) + 3(-1)] = 1.000
\end{aligned}$$

Check stopping criterion:
$|x^{(3)} - x^{(2)}| = |1.00 - 1.00475| = 0.00475 \implies < \epsilon$
$|y^{(3)} - y^{(2)}| = |-1 - (-0.999)| = 0.001 \implies < \epsilon$
$|z^{(3)} - z^{(2)}| = |1.00 - 1.000| = 0.00 \implies < \epsilon$
Since this criterion is satisfied we stop the iteration here.

Therefore the solution to the system is

$$x = 1, \quad y = -1, \quad z = 1$$

**Exercise 3.4** Solving the following system of equations using the Gauss-Seidel method, taking the origin as the initial solution. Take $\epsilon = 0.01$, otherwise stop on the 5th iteration

1.

$$\begin{aligned}
10x + 4y - 2z &= 20 \\
3x + 12y - z &= 28 \\
x + 4y + 7z &= 2
\end{aligned}$$

2.

$$\begin{aligned}
2a + b + c + d &= 2 \\
4a + 2c + d &= 3 \\
3a + 2b + 2c &= -1 \\
a + 3b + 2c + 6d &= 2
\end{aligned}$$

# 4. Numerical Differentiation and Integration

4.1 Numerical Differentiation 42

4.1.1 Newton's Forward Difference

4.1.2 Newton's Backward Difference

4.2 Numerical Integration 47

4.2.1 Simple Trapezium Rule

4.2.2 Composite Trapezium Rule

4.2.3 The Simple Simpson's 1/3 Rule

4.2.4 The Composite Simpson's 1/3 Rule

The chapter considers numerical methods of differentiating and integrating functions.

## 4.1 Numerical Differentiation

Analytical techniques for differentiating functions were treated in first year calculus. Here, approximations to the derivatives of functions are obtained numerically using the method of finite differencing. Specifically, Newton's forward difference and Newton's backward difference.

### 4.1.1 Newton's Forward Difference

Consider the data $(x_i, f(x_i))$, given by equidistant points such that

$$x_i = [x_0 + ih], \tag{4.1}$$

where $i = 0, 1, 2, \cdots$ and $h$ is the step size or the mesh size. Then the Newton's forward difference at the beginning of the table

$$x = x_0$$

is given by the following approximations

**First Derivative**

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \cdots\right] \tag{4.2}$$

**Second Derivative**

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0 + \frac{137}{180}\Delta^6 f_0 + \cdots\right] \tag{4.3}$$

where $\Delta f$ is the change in $f$. That is

$$\Delta f = y_{i+1} - y_i \tag{4.4}$$

Moreover, the subscript denote the location of this difference. If $x \in [0, n]$, then $\Delta f_0$ is the difference at the beginning of the table, while $\Delta f_n$ is the difference at the end of the table.

We use the forward difference method when we need the value of the derivatives at the points near the beginning of the table of values.

**Example 4.1.1** If

$$y = x^3, \quad h = 1, \quad x_0 = 1, \quad i = 1:3$$

Find the first and second derivatives of the function using the Newton's forward difference scheme.
Hence compute the absolute error in each case.

**Solution**
From the question $i = 1 : 3$. We use these $i$ values to compute the respective $x_i$ values. Note that

$$x_i = [x_0 + hi] \tag{4.5}$$

Then

$$\begin{aligned}
\text{When } i = 1 &\implies x_1 = x_0 + ih = 1 + (1)1 = 2 \\
\text{When } i = 2 &\implies x_2 = x_0 + ih = 1 + (2)1 = 3 \\
\text{When } i = 3 &\implies x_3 = x_0 + ih = 1 + (3)1 = 4
\end{aligned}$$

Therefore the $x$ nodal points are

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_3 = 4$$

Again from the question

$$y = x^3$$

Iteratively we have

$$y_i = (x_i)^3 \tag{4.6}$$

Now let begin the iterations to find the functional values:

**Iteration 1: when i =0**
We have $x_0 = 1$.
Then the initial value for y is

$$y_0 = (x_0)^3 \quad \implies \quad y_0 = 1^3 = 1 \tag{4.7}$$

**Iteration 2: when i =1**
We have $x_1 = 2$

$$y_1 = (x_1)^3 \quad \implies \quad y_1 = 2^3 = 8 \tag{4.8}$$

**Iteration 3: when i =2**
We have $x_2 = 3$

$$y_2 = (x_2)^3 \quad \Longrightarrow \quad y_2 = 3^3 = 27 \tag{4.9}$$

**Iteration 4: when i =3**
We have $x_3 = 4$

$$y_3 = (x_3)^3 \quad \Longrightarrow \quad y_3 = 4^3 = 64$$

Let present the $y$ result in table form to estimate $\Delta f, \Delta^2 f, \cdots$. We use equation (4.4) to compute for all the $\Delta f, \cdots, \Delta^4 f$ values.

| $y = f$ | 1 | | 8 | | 27 | | 64 |
|---|---|---|---|---|---|---|---|
| $\Delta f$ | | 7 | | 19 | | 37 | |
| $\Delta^2 f$ | | | 12 | | 18 | | |
| $\Delta^3 f$ | | | | 6 | | | |
| $\Delta^4 f$ | | | | 0 | | | |

**Note 4.1** Note that $1, 7, 12, 6$ are the beginning value $(f_0)$, while $6, 18, 37, 64$ are the end values $(f_n)$

**First Derivative**

$$\begin{aligned} f'(x_0) &= \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \cdots\right] \\ f'(1) &= \frac{1}{1}\left[7 - \frac{1}{2}(12) + \frac{1}{3}(6) + 0\right] \\ &= 3 \end{aligned}$$

We will need the exact solution, since we have compute the absolute error.

**Exact solution**

$$y = f = x^3 \implies f'(x) = 3x^2 \implies f'(1) = 3$$

Recall that

$$\begin{aligned} \text{Absolute error} &= |\text{Exact value - Approximate value}| \\ &= |3 - 3| = 0 \end{aligned}$$

**Second Derivative**

$$\begin{aligned} f''(x_0) &= \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0 + \frac{137}{180}\Delta^6 f_0 + \cdots\right] \\ f''(1) &= \frac{1}{1^2}\left[12 - 6 + 0\right] \\ &= 6 \end{aligned}$$

**Exact solution**
Here $x = x_0$, that is the first value on the nodal table: $x_0 = x = 1$.

$$y = f = x^3 \implies f''(x) = 6x \implies f''(1) = 6$$

$$\begin{aligned}\text{Absolute error} &= |\text{Exact value - Approximate value}|\\ &= |6-6| = 0\end{aligned}$$

### 4.1.2 Newton's Backward Difference

Considering the same data points $(x_i, f(x_i))$, at equidistant points such that

$$x_i = [x_0 + ih], \tag{4.10}$$

where $i = 0, 1, 2, \cdots$ and $h$ is the step size or the mesh size. Then the Newton's backward difference at the end of the table of values

$$x = x_n$$

is given by the following approximations

**First Derivative**

$$f'(x_n) = \frac{1}{h}\left[\Delta f_n + \frac{1}{2}\Delta^2 f_n + \frac{1}{3}\Delta^3 f_n + \frac{1}{4}\Delta^4 f_n + \cdots\right] \tag{4.11}$$

**Second Derivative**

$$f''(x_n) = \frac{1}{h^2}\left[\Delta^2 f_n + \Delta^3 f_n + \frac{11}{12}\Delta^4 f_n + \frac{5}{6}\Delta^5 f_n + \frac{137}{180}\Delta^6 f_n + \cdots\right] \tag{4.12}$$

We use the backward difference method when we need the value of the derivatives at the points near the end of the table of values.

**Example 4.1.2** If

$$y = x^3, \quad h = 1, \quad x_0 = 1, \quad i = 1:3$$

Find the first and second derivatives of the function using the Newton's backward difference scheme.
Hence compute the absolute error in each case.

**Solution**
From the question $i = 1 : 3$. We use these $i$ values to compute the respective $x_i$ values. Note that

$$x_i = [x_0 + hi] \tag{4.13}$$

Then

$$\begin{aligned}\text{When } i = 1 &\implies x_1 = x_0 + h = 1 + 1 = 2\\ \text{When } i = 2 &\implies x_2 = x_1 + h = 2 + 1 = 3\\ \text{When } i = 3 &\implies x_3 = x_2 + h = 2 + 1 = 4\end{aligned}$$

Therefore the $x$ nodal points are

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_3 = 4$$

Again from the question

$$y = x^3$$

Iteratively we have

$$y_i = (x_i)^3 \tag{4.14}$$

Now let begin the iterations to find the functional values:

**Iteration 1: when i =0**
We have $x_0 = 1$.
Then the initial value for y is

$$y_0 = (x_0)^3 \quad \Longrightarrow \quad y_0 = 1^3 = 1 \tag{4.15}$$

**Iteration 2: when i =1**
We have $x_1 = 2$

$$y_1 = (x_1)^3 \quad \Longrightarrow \quad y_1 = 2^3 = 8 \tag{4.16}$$

**Iteration 3: when i =2**
We have $x_2 = 3$

$$y_2 = (x_2)^3 \quad \Longrightarrow \quad y_2 = 3^3 = 27 \tag{4.17}$$

**Iteration 4: when i =3**
We have $x_3 = 4$

$$y_3 = (x_3)^3 \quad \Longrightarrow \quad y_3 = 4^3 = 64$$

Let present the $y$ result in table form to estimate $\Delta f, \Delta^2 f, \cdots$. We use equation (4.4) to compute for all the $\Delta f, \cdots, \Delta^4 f$ values.

| $y = f$ | 1 | | 8 | | 27 | | 64 |
|---|---|---|---|---|---|---|---|
| $\Delta f$ | | 7 | | 19 | | 37 | |
| $\Delta^2 f$ | | | 12 | | 18 | | |
| $\Delta^3 f$ | | | | 6 | | | |
| $\Delta^4 f$ | | | | 0 | | | |

**Note 4.2** Note that $1, 7, 12, 6$ are the beginning value $(f_0)$, while $6, 18, 37, 64$ are the end values $(f_n)$

**First Derivative**

$$\begin{aligned}
f'(x_n) &= \frac{1}{h}\left[\Delta f_n + \frac{1}{2}\Delta^2 f_n + \frac{1}{3}\Delta^3 f_n + \frac{1}{4}\Delta^4 f_n + \cdots\right] \\
f'(4) &= \frac{1}{1}\left[37 + \frac{1}{2}(18) + \frac{1}{3}(6) + 0\right] \\
&= 48
\end{aligned}$$

We will need the exact solution, since we have compute the absolute error.

**Exact solution**
Here $x = x_n$, that is the last value on the nodal table: $x_n = x = 4$.

$$y = f = x^3 \implies f'(x) = 3x^2 \implies f'(4) = 48$$

$$\begin{aligned}\text{Absolute error} &= |\text{Exact value - Approximate value}| \\ &= |48 - 48| = 0\end{aligned}$$

**Second Derivative**

$$\begin{aligned}f''(x_n) &= \frac{1}{h^2}\left[\Delta^2 f_n + \Delta^3 f_n + \frac{11}{12}\Delta^4 f_n\right] \\ f''(4) &= \frac{1}{1^2}[18 + 6 + 0] \\ &= 24\end{aligned}$$

**Exact solution**

$$y = f = x^3 \implies f''(x) = 6x \implies f''(4) = 24$$

$$\begin{aligned}\text{Absolute error} &= |\text{Exact value - Approximate value}| \\ &= |24 - 24| = 0\end{aligned}$$

**Exercise 4.1** Find the first and second derivatives of the following functions using both forward and backward difference. Compute the relative error in each case

1. $$f(x) = e^{2x}, \quad x \in [0,\ 1.2], \quad h = 0.3, \quad x_0 = 0$$

2. $$y = 3x^3 + 2e^x, \quad h = 0.2, \quad x_0 = 0, \quad i = 0 : 0.6$$

## 4.2 Numerical Integration

This deals with the problem of finding an approximate value of the integral

$$I = \int_a^b w(x)f(x)dx \tag{4.18}$$

where $w(x) > 0$ the weight function lies in the open interval $(a, b)$, and $I$ a definite integral. $I$ reduces to an indefinite integral when the limits of integration are not specified.

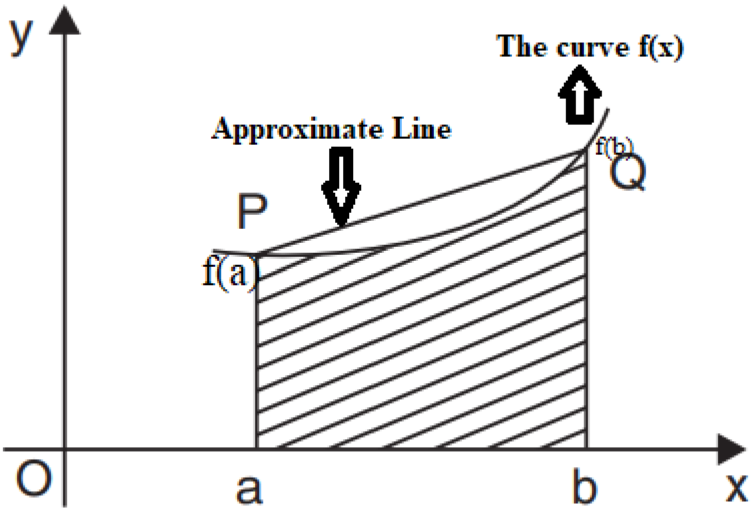
When $w(x) = 1$ and $x_k$'s are equidistant with $x_0 = a,\ x_n = b,\ h = \frac{b-a}{N}$, where $N$ is the number of subdivision, then the integral (4.18) reduces to

$$I = \int_{x_0}^{x_n} f(x)dx \tag{4.19}$$

This integral (4.19) defines the area under the curve above the $x-$axis within the interval $[x_0,\ x_n]$. The following integration techniques will be employed in finding the value of this area.

### 4.2.1 Simple Trapezium Rule

This is also called the trapezoidal method. Let the curve $y = f(x),\ a \leq x \leq b$ be approximated by the line joining the points $P(a, f(a))$ and $Q(b, f(b))$ on the curve as illustrated below



Then the trapezium method uses the area under the approximate line to obtain the formula

$$I = \frac{1}{2}h\big[f(a) + f(b)\big] \tag{4.20}$$

$$= \frac{1}{2}(b-a)\Big[f(a) + f(b)\Big], \qquad N = 1 \tag{4.21}$$

**Example 4.2.1** Approximate the following integrals using the simple trapezium rule

1. $\displaystyle\int_3^5 2x^2 dx$
2. $\displaystyle\int_0^{\pi/4} \sin(x) dx$

Hence determine the absolute error (AE)

**Solution**

1. Given $\displaystyle\int_3^5 2x^2 dx$, then $a = 3$ and $b = 5$. The functional values are

$f(a) = f(3) = 2(3)^2 = 18$
$f(b) = f(5) = 2(5)^2 = 50$

Therefore

$$I = \frac{1}{2}(b-a)\big[f(a)+f(b)\big] \tag{4.22}$$
$$= \frac{1}{2}(5-3)[18+50] \tag{4.23}$$
$$= 68 \tag{4.24}$$

For absolute error, the exact solution is required

**Exact solution**

$$\int_3^5 2x^2dx = \left.\frac{2x^3}{3}\right|_3^5 = 65.33$$

$$AE = |ES - AS| = |65.33 - 68| = 2.66$$

2. Given $\int_0^{\pi/4} \sin(x)dx$, then $a = 0$ and $b = \frac{\pi}{4}$. The functional values are

$f(a) = f(0) = \sin(0) = 0$
$f(b) = f(\pi/4) = \sin(\pi/4) = 0.707$

Therefore

$$I = \frac{1}{2}(b-a)\big[f(a)+f(b)\big] \tag{4.25}$$
$$= \frac{1}{2}(\pi/4-0)[0+0.707] \tag{4.26}$$
$$= \frac{\pi}{8}(0.707) \tag{4.27}$$
$$= 15.9075 \tag{4.28}$$

**Exact solution**

$$\int_0^{\pi/4} \sin(x)dx = -\cos(x)\Big|_0^{\pi/4} = -\cos(\pi/4) + \cos(0) = 0.293$$

$$AE = |ES - AS| = |0.293 - 15.9075| = 15.614$$

### 4.2.2 Composite Trapezium Rule

Here, the idea is to split the interval $(a,\ b)$ into a sequence of $N$ smaller sub-interval with width $h = \frac{b-a}{N}$. These yields the composite trapezium formula

$$\int_a^b f(x)dx = \frac{h}{2}\left\{f(x_0) + 2\Big[f(x_1) + f(x_2) + \cdots + f(x_{N-1})\Big] + f(x_N)\right\} \tag{4.29}$$

**Example 4.2.2** Find the approximate solution for the following using the composite trapezium rule with 4 equal sub-intervals.

1. $\displaystyle\int_3^5 2x^2\ dx$

2. $\displaystyle\int_0^2 \ln(x)\ dx$

**Solution**

For 4 sub-intervals $\implies h = \dfrac{b-a}{N} = \dfrac{5-3}{4} = 0.5.$

Therefore the $x$ points are

$$x_0 = 3, \quad x_1 = 3.5, \quad x_2 = 4, \quad x_3 = 4.5, \quad x_4 = 5$$

The functional value are

$$\begin{aligned} f(x_0) &= f(3) = 18 \\ f(x_1) &= f(3.5) = 24.5 \\ f(x_2) &= f(4) = 32 \\ f(x_3) &= f(4.5) = 40.5 \\ f(x_4) &= f(5) = 50 \end{aligned}$$

Substituting into the formula

$$\begin{aligned} \int_a^b f(x)dx &= \frac{h}{2}\Big\{f(x_0) + 2\Big[f(x_1) + f(x_2) + \cdots + \ f(x_{N-1})\Big] + f(x_N)\Big\} \\ \int_3^5 2x\ dx &= \frac{0.5}{2}\Big\{f(x_0) + 2\Big[f(x_1) + f(x_2) + \ f(x_3)\Big] + f(x_4)\Big\} \\ &= \frac{0.5}{2}\{18 + 2[24.5 + \ 32 + 40.5] + 50\} \\ &= 65.5 \end{aligned}$$

From above the exact solution is 65.33, therefore the absolute error is

$$AE = |ES - AS| = |65.33 - 65.5| = 0.23$$

By comparison, we realize that the composite trapezium is more accurate than the simple trapezium method.

### 4.2.3 The Simple Simpson's 1/3 Rule

This is based on a quadratic curve through equally spaced point rather than a line as is the case of the simple trapezium rule. The interval $(a, b)$ is subdivided into two equal parts with the step length $h = \dfrac{b-a}{2}$.

We approximate the integral curve by the parabola joining these points. The formula of the Simpson's 1/3 rule is deduced from interpolation techniques (check the reference books for details), and its given by

$$\int_a^b f(x)dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{4.30}$$

**Example 4.2.3** Find the derivative of the following functions using the simple Simpson's 1/3 rule

1. $\int_3^5 2x^2 dx$

2. $\int_1^3 \frac{1}{1+x}dx$

**Solution**

For 2 sub-intervals $\implies h = \frac{b-a}{2} = \frac{5-3}{2} = 1.$

Therefore the $x$ points are

$$3,\ 4,\ 5$$

The functional value are

$$\begin{aligned} f(x_0) &= f(3) = 18 = f(a) \\ f(x_1) &= f(4) = 32 = f\left(\frac{a+b}{2}\right) \\ f(x_2) &= f(5) = 50 = f(b) \end{aligned}$$

Substituting into the formula

$$\begin{aligned} \int_a^b f(x)dx &= \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \\ \int_3^5 2x^2 dx &= \frac{5-3}{6}[18 + 4(32) + 50] \\ &= \frac{1}{3}(196) \\ &= 65.33 \end{aligned}$$

### 4.2.4 The Composite Simpson's 1/3 Rule

This is an extension of the simple Simpson's rule. Here, the given interval $a, b$ can be divided into any finite sub-intervals of equal length say $h$. For $N$ sub-intervals the formula is given as

$$\begin{aligned} \int_a^b f(x)dx = \frac{h}{3}\Big\{ f(x_0) &+ 4\Big[f(x_1) + f(x_3) + f(x_5) + \cdots + f(x_{N-1})\Big] \\ &+ 2\Big[f(x_2) + f(x_4) + f(x_6) + \cdots + f(x_{N-2})\Big] + f(x_N)\Big\} \end{aligned}$$

**Example 4.2.4** Find the derivative of the following functions using the composite Simpson's 1/3 rule using 4 sub-intervals

1. $\displaystyle\int_3^5 2x^2 dx$
2. $\displaystyle\int_1^3 \frac{1}{1+x} dx$

**Solution**

For 4 sub-intervals $\implies h = \dfrac{b-a}{4} = \dfrac{5-3}{4} = 0.5.$

Therefore the $x$ points are

$$3,\ 3.5,\ 4,\ 4.5,\ 5$$

The functional value are

$$\begin{aligned} f(x_0) &= f(3) = 18 \\ f(x_1) &= f(3.5) = 24.5 \\ f(x_2) &= f(4) = 32 \\ f(x_3) &= f(4.5) = 40.5 \\ f(x_4) &= f(5) = 50 \end{aligned}$$

Substituting into the formula

$$\begin{aligned} \int_a^b f(x)dx = \frac{h}{3}\Big\{ & f(x_0) + 4\Big[f(x_1) + f(x_3) + f(x_5) + \cdots + f(x_{N-1})\Big] \\ & + 2\Big[f(x_2) + f(x_4) + f(x_6) + \cdots + f(x_{N-2})\Big] + f(x_N)\Big\} \end{aligned}$$

$$\begin{aligned} \int_3^5 2x^2 dx &= \frac{h}{3}\Big\{f(x_0) + 4\Big[f(x_1) + f(x_3)\Big] + 2\Big[f(x_2)\Big] + f(x_4)\Big\} \\ &= \frac{0.5}{3}\Big\{18 + 4(24.5 + 40.5) + 2(32) + 50\Big\} \\ &= 0.167(392) \\ &= 65.474 \end{aligned}$$

**Exercise 4.2** Evaluate the function

$$\int_0^6 \frac{1}{1+x^2}$$

using

1. Simple trapezium rule
2. Composite trapezium rule with 6 sub-intervals
3. Simple Simpson's 1/3 rule
4. Composite Simpson's 1/3 rule with 8 sub-intervals

# 5. Numerical Solutions of Ordinary Differential Equations: Single-Step Methods

5.1 Single-Step Methods 53

5.2 Multi-Step Methods 54

5.3 Solution To IVP's Using Single-Step Methods 55

5.3.1 Taylor Series of Order 1 or Euler Method

5.3.2 Backward Euler

5.3.3 Modified Euler or Midpoint Method

5.3.4 Trapezium Method

5.3.5 Heun's Method or Euler-Cauchy Method

5.3.6 Runge-Kutta Methods

The general form of an nth order ODE is

$$f(x,\ y',\ y'',\ y''',\ \cdots, y^{(n)}) \tag{5.1}$$

The order of an ODE is the order of its highest derivative. A differential equation together with some initial conditions is called an **Initial Value Problem(IVP)**. A first order IVP can be written as

$$y' = f(x, y); \qquad y(x_0) = y_0 \tag{5.2}$$

Given an interval $[x_0,\ b]$ in which a solution is desired. The interval is divided into finite number of sub-intervals by points

$$x_0 < x_1 < x_2 < \cdots < x_n; \qquad x_n = b \tag{5.3}$$

These points are called mesh points or grid points. The spacing between the points are given by

$$h_i = x_i - x_{i-1}, \qquad i = 1, 2, 3, \cdots, n \tag{5.4}$$

The methods for solving this IVP's can be classified into single-step methods and multi-step methods.

## 5.1 Single-Step Methods

With this method, the solution at any point $y_{i+1}$ is obtained by using the solution at only the previous point $y_i$. A general single step method can be written as

$$y_{i+1} = y_i + hf(x_{i+1},\ x_i,\ y_{i+1},\ y_i,\ h) \tag{5.5}$$

The function $f$ is called the increment function. Since the right hand side of the equation depend on $y_{i+1}$, equation (5.5) is referred to as an **Implicit scheme**.

In the explicit case, the right hand side does not not depend on $y_{i+1}$. For the **Explicit scheme** equation (5.5) reduces to

$$y_{i+1} = y_i + hf(x_i,\ y_i,\ h) \tag{5.6}$$

Given an initial value $y_0$, then the other values of $y$ are computed successively as

$$\begin{aligned}
y_1 &= y_0 + hf(x_0,\ y_0,\ h), && \text{when } i = 0 && (5.7)\\
y_2 &= y_1 + hf(x_1,\ y_1,\ h), && \text{when } i = 1 && (5.8)\\
y_3 &= y_2 + hf(x_2,\ y_2,\ h), && \text{when } i = 2 && (5.9)\\
\vdots \quad & \qquad\quad \vdots && \quad\vdots &&\\
y_n &= y_{n-1} + hf(x_{n-1},\ y_{n-1},\ h), && \text{when } i = n-1 && (5.10)
\end{aligned}$$

The solution of $y_1$ requires only one previous point $y_0$.
The solution of $y_2$ requires only one previous point $y_1$.
The solution of $y_3$ requires only one previous point $y_2$.
The solution of $y_n$ requires only one previous point $y_{n-1}$.

Hence this schemes are called single-step methods.

> **Note 5.1** All single-step methods are self starting, that is, they do not require values of $y$ or it's derivatives beyond the immediate previous point.

## 5.2 Multi-Step Methods

With this method, the solution at point $y_{i+1}$ is obtained using the solution at a number of previous points, $y_i,\ y_{i-1},\ y_{i-2},\ y_{i-3}, \cdots$.

Two-step implicit depends on $y_{i+1}, y_i, y_{i-1}$
Two-step explicit depends on $y_i, y_{i-1}$

Four-step implicit depends on $y_{i+1}, y_i, y_{i-1}, y_{i-2}, y_{i-3}$
Four-step explicit depends on $y_i, y_{i-1}, y_{i-2}, y_{i-3}$

A classical example of a two-step implicit method can be written as

$$y_{i+1} = y_i + hf(x_{i+1},\ x_i,\ x_{i-1},\ y_{i+1},\ y_i,\ y_{i-1},\ h) \tag{5.11}$$

A classical example of a three-step explicit method can be written as

$$y_{i+1} = y_i + hf(\ x_i,\ x_{i-1}, x_{i-2},\ y_i,\ y_{i-1},\ y_{i-2},\ h) \tag{5.12}$$

A **general $k$-step explicit** method can be written as

$$y_{i+1} = y_i + hf(x_{i-k+1},\ \cdots,\ x_{i-1},\ x_i,\ y_{i-k+1},\ \cdots,\ y_{i-1},\ y_i,\ h) \tag{5.13}$$

and the **implicit** case as

$$y_{i+1} = y_i + hf(x_{i-k+1},\ \cdots,\ x_{i-1},\ x_i,\ x_{i+1},\ y_{i-k+1},\ \cdots,\ y_{i-1},\ y_i,\ y_{i+1},\ h) \tag{5.14}$$

## 5.3 Solution To IVP's Using Single-Step Methods

Some numerical techniques used for solving IVP's include:

1. Euler or Taylor series Method
2. Backward Euler
3. Modified Euler or Midpoint Method
4. Trapezium Method
5. Heun's Method or Euler-Cauchy Method
6. Runge-Kutta (RK) Methods
   (a) Second-order RK Method
   (b) Fourth-order RK Method

All these methods are derived using Taylor series. Given the Taylor series

$$y(x_{i+1}) = y(x_i) + hf[(x_i + \theta h), y(x_i + \theta h)]; \qquad 0 \leq \theta \leq 1 \tag{5.15}$$

Several numerical schemes can be deduced from the above depending on the value of $\theta$.

### 5.3.1 Taylor Series of Order 1 or Euler Method

This is obtained from equation (5.15) be letting

$$\theta = 0$$

The scheme is given by the formula

$$y(x_{i+1}) = y(x_i) + hf[x_i,\ y(x_i)] \tag{5.16}$$

This is an explicit scheme

### 5.3.2 Backward Euler

This is obtained from equation (5.15) be letting

$$\theta = 1$$

The scheme is given by the formula

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + hf[(x_i + h),\ y(x_i + h)] && (5.17) \\ &= y(x_i) + hf[x_{i+1},\ y(x_{i+1})] && (5.18) \end{aligned}$$

This is an implicit scheme.

### 5.3.3 Modified Euler or Midpoint Method

This is obtained from equation (5.15) be letting

$$\theta = \frac{1}{2}$$

The scheme is given by the formula

$$y(x_{i+1}) = y(x_i) + hf\left[\left(x_i + \frac{h}{2}\right),\ y\left(x_i + \frac{h}{2}\right)\right] \tag{5.19}$$

However, $x_i + \frac{h}{2}$ is not a nodal point, hence we approximate $y\left(x_i + \frac{h}{2}\right)$ using the Euler method with spacing $\frac{h}{2}$. The Euler approximation is given by equation (5.20).

$$y\left(x_i + \frac{h}{2}\right) = y_i + \frac{h}{2}f(x_i,\ y_i) \tag{5.20}$$

Substituting equation (5.20) into equation (5.19), we obtain the Modified Euler as

$$y(x_{i+1}) = y(x_i) + hf\left[\left(x_i + \frac{h}{2}\right),\ y_i + \frac{h}{2}f(x_i,\ y_i)\right] \tag{5.21}$$

### 5.3.4 Trapezium Method

Let the continuously varied slope in $x_i$ and $x_{i+1}$ be approximated by the mean of the slope, then the trapezium method is deduced as

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + \frac{h}{2}\left\{f[x_i,\ y(x_i)] + f[x_{i+1},\ y(x_{i+1})]\right\} && (5.22)\\ &= y_i + \frac{h}{2}[f_i + f_{i+1}] && (5.23) \end{aligned}$$

This is an implicit scheme. When this is converted to an explicit scheme we obtain the Heun's method.

### 5.3.5 Heun's Method or Euler-Cauchy Method

This is the explicit form of the trapezium method. This conversion is made possible by using the approximation

$$y(x_{i+1}) = y(x_i) + hf[x_i,\ y(x_i)] \tag{5.24}$$

Substituting equation (5.24) into equation (5.22) we obtain the Euler-Cauchy iterative scheme as

$$y(x_{i+1}) = y(x_i) + \frac{h}{2}\left\{f[x_i,\ y(x_i)] + f[x_{i+1},\ y(x_i) + hf[x_i,\ y(x_i)]]\right\} \tag{5.25}$$

**Note 5.2** The following are equivalent

$$y_{i+1} = y(x_{i+1}), \qquad y_i = y(x_i), \qquad \cdots$$

**Note 5.3** The following are equivalent

$$f_i = f[x_i,\ y(x_i)], \qquad f_{i+1} = f[x_{i+1},\ y(x_{i+1})], \quad \cdots$$

Let look at some examples.

**Example 5.3.1** Solve the following IVP

$$yy' = x, \quad y(0) = 1, \quad 0 \le x \le 0.6, \quad h = 0.2$$

using

1. Euler method
2. Modified Euler method
3. Euler-Cauchy method

In each case compute the absolute error at $x = 0.6$

**Solution**
**1. Euler**
Given the step size $h = 0.2$. Then the $x$ values are given by the **interval table**

$$x_0 = 0, \quad x_1 = 0.2, \quad x_2 = 0.4, \quad x_3 = 0.6 \tag{5.26}$$

The Euler formula is given as

$$y(x_{i+1}) = y(x_i) + hf[x_i,\ y(x_i)] \tag{5.27}$$

$$y_{i+1} = y_i + hf(x_i, y_i) \tag{5.28}$$

From equation (5.2), $y' = f(x, y)$. Therefore making $y'$ the subject from the question we have

$$y' = f(x, y) = \frac{x}{y} \tag{5.29}$$

Iteratively,

$$f(x_i, y_i) = \frac{x_i}{y_i} \tag{5.30}$$

**Iteration 1: when i=0**
The formula reduces to

$$y(x_1) = y(x_0) + hf[x_0,\ y(x_0)] \tag{5.31}$$

$$y_1 = y_0 + hf(x_0, y_0) \tag{5.32}$$

From the question, that is $y(0) = 1$, the initial conditions can be deduced as

$$x_0 = 0, \quad \text{and } y_0 = 1$$

Therefore

$$f(x_0, y_0) = \frac{x_0}{y_0} = \frac{0}{1} = 0$$

Hence

$$\begin{aligned} y_1 &= y_0 + hf(x_0, y_0) \\ &= 1 + 0.2(0) \\ &= 1 \end{aligned}$$

**Iteration 2: when i=1**
The formula reduces to

$$y_2 = y_1 + hf(x_1, y_1)$$

From the interval table (5.26), $x_1 = 0.2$, and from the previous solution $y_1 = 1$

Therefore

$$f(x_1, y_1) = \frac{x_1}{y_1} = \frac{0.2}{1} = 0.2$$

Hence

$$\begin{aligned} y_2 &= y_1 + hf(x_1, y_1) \\ &= 1 + 0.2(0.2) \\ &= 1.04 \end{aligned}$$

**Iteration 3: when i=2**
The formula reduces to

$$y_3 = y_2 + hf(x_2, y_2)$$

From the interval table (5.26), $x_2 = 0.4$, and from the previous solution $y_2 = 1.04$

Therefore

$$\begin{aligned} y_3 &= y_2 + hf(x_2, y_2) \\ &= 1.04 + 0.2\left(\frac{0.4}{1.04}\right) \\ &= 1.117 \end{aligned}$$

Therefore the nodal points are

$$(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1); \quad (x_2, y_2) = (0.4, 1.04); \quad (x_3, y_3) = (0.6, 1.117)$$

**Analytical Solution**
The differential equation is solved using separation of variables

$$y' = \frac{dy}{dx} = \frac{x}{y} \implies \int dyy = \int xdx \implies y^2 = x^2 + c$$

Implementing the initial condition to find $c$

$$1^2 = 0^2 + c \implies c = 1$$

Therefore the analytical solution is

$$y = \sqrt{x^2 + 1}$$

Hence at point $x_3 = 0.6$

$$y_3 = y(0.6) = \sqrt{0.6^2 + 1} = 1.166$$

The absolute error

$$AE = |ES - AS| = |1.166 - 1.117| = 0.049$$

2. **Modified Euler**

The formula is given as

$$y(x_{i+1}) = y(x_i) + hf\left[\left(x_i + \frac{h}{2}\right),\ y_i + \frac{h}{2}f(x_i,\ y_i)\right] \tag{5.33}$$

$$y_{i+1} = y_i + hf\left[x_i + \frac{h}{2},\ y_i + \frac{h}{2}f(x_i, y_i)\right] \tag{5.34}$$

Again

$$f(x_i, y_i) = \frac{x_i}{y_i} \tag{5.35}$$

The initial conditions can be deduced as

$$x_0 = 0, \quad \text{and } y_0 = 1$$

**Iteration 1: when i=0**
The formula reduces to

$$y_1 = y_0 + hf\left[x_0 + \frac{h}{2},\ y_0 + \frac{h}{2}f(x_0, y_0)\right] \tag{5.36}$$

$$= 1 + 0.2f\left[0 + \frac{0.2}{2},\ 1 + \frac{0.2}{2}\left(\frac{0}{1}\right)\right] \tag{5.37}$$

$$= 1 + 0.2f(0.1,\ 1) \tag{5.38}$$

$$= 1 + 0.2\left(\frac{0.1}{1}\right) \tag{5.39}$$

$$= 1.02 \tag{5.40}$$

**Iteration 2: when i=1**
From the interval table (5.26), $x_1 = 0.2$, and from the previous solution $y_1 = 1.02$

The solution is as follows

$$y_2 = y_1 + hf\left[x_1 + \frac{h}{2},\ y_1 + \frac{h}{2}f(x_1, y_1)\right] \tag{5.41}$$

$$= 1.02 + 0.2f\left[0.2 + \frac{0.2}{2},\ 1.02 + \frac{0.2}{2}\left(\frac{0.2}{1.02}\right)\right] \tag{5.42}$$

$$= 1.02 + 0.2f(0.3,\ 1.04) \tag{5.43}$$

$$= 1.02 + 0.2\left(\frac{0.3}{1.04}\right) \tag{5.44}$$

$$= 1.03 \tag{5.45}$$

**Iteration 3: when i=2**
From the interval table (5.26), $x_2 = 0.4$, and from the previous solution $y_2 = 1.03$

The solution is as follows

$$\begin{aligned} y_3 &= y_2 + hf\left[x_2 + \frac{h}{2},\ y_2 + \frac{h}{2}f(x_2, y_2)\right] &(5.46)\\ &= 1.03 + 0.2f\left[0.4 + \frac{0.2}{2},\ 1.03 + \frac{0.2}{2}\left(\frac{0.4}{1.03}\right)\right] &(5.47)\\ &= 1.03 + 0.2f(0.5,\ 1.07) &(5.48)\\ &= 1.03 + 0.2\left(\frac{0.5}{1.07}\right) &(5.49)\\ &= 1.123 &(5.50)\end{aligned}$$

Therefore the nodal points are

$$(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1.02); \quad (x_2, y_2) = (0.4, 1.03); \quad (x_3, y_3) = (0.6, 1.123)$$

Hence

$$AE = |ES - AS| = |1.167 - 1.123| = 0.043$$

### 5.3.6 Runge-Kutta Methods

This is also a single-step method used for solving IVPs. A Runge-Kutta method of second-order uses two slopes, that is $k_1$ and $k_2$, whereas the fourth-order Runge-Kutta uses four slopes; $k_1, k_2, k_3,$ and $k_4$.

**Second-Order Runge-Kutta**

A general second-order Runge-Kutta (RK2) is of the form

$$y_{i+1} = y_i + \left(1 - \frac{1}{2\theta}\right)k_1 + \frac{k_2}{2\theta} \tag{5.51}$$

where $k_1 = hf(x_i, y_i)$ and
$k_2 = hf(x_i + \theta h,\ y_i + \theta k_1)$

The value of $\theta$ is arbitrary such that $0 \leq \theta \leq 1$. This results to a myriad number of solution schemes

**When $\theta = 1$**

$$y_{i+1} = y_i + \frac{1}{2}k_1 + \frac{1}{2}k_2 \tag{5.52}$$

where $k_1 = hf(x_i, y_i)$ and
$k_2 = hf(x_i + h,\ y_i + k_1)$
This is the same as the Heun's method.

**When $\theta = 1/2$**

$$y_{i+1} = y_i + k_2 \tag{5.53}$$

where $k_1 = hf(x_i, y_i)$ and
$k_2 = hf(x_i + h/2,\ y_i + k_1/2)$
This is the same as the Modified Euler

**Fourth-Order Runge-Kutta**

In the case of RK4, the iterative scheme is given by

$$y_{i+1} = y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{5.54}$$

where $k_1 = hf(x_i, y_i)$,
$k_2 = hf\left(x_i + \frac{h}{2},\ y_i + \frac{k_1}{2}\right)$
$k_3 = hf\left(x_i + \frac{h}{2},\ y_i + \frac{k_2}{2}\right)$
$k_4 = hf\left(x_i + h,\ y_i + k_3\right)$

Let look at an example involving RK4

**Example 5.3.2** Solve the following IVP

$$yy' = x, \quad y(0) = 1, \quad 0 \le x \le 0.6, \quad h = 0.2$$

using Runge-Kutta fourth-order scheme.
Hence determine the absolute error at $x = 0.6$

**Solution**
Given the step size $h = 0.2$. Then the $x$ values are given by the **interval table**

$$x_0 = 0, \quad x_1 = 0.2, \quad x_2 = 0.4, \quad x_3 = 0.6 \tag{5.55}$$

We know that

$$y' = f(x, y) = \frac{x}{y} \tag{5.56}$$

Iteratively,

$$f(x_i, y_i) = \frac{x_i}{y_i} \tag{5.57}$$

The initial conditions can be deduced as

$$x_0 = 0, \qquad \text{and } y_0 = 1$$

**Iteration 1: when i=0**
The formula reduces to

$$y_1 = y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{5.58}$$

where

$$k_1 = hf(x_0, y_0) = 0.2\left(\frac{0}{1}\right) = 0 \tag{5.59}$$

$$\begin{aligned} k_2 = hf\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_1}{2}\right) &= 0.2f\left(0 + \frac{0.2}{2},\ 1 + \frac{0}{2}\right) \\ &= 0.2f(0.1,\ 1) \\ &= 0.2\left(\frac{0.1}{1}\right) \\ &= 0.02 \end{aligned} \tag{5.60}$$

$$\begin{aligned} k_3 = hf\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_2}{2}\right) &= 0.2f\left(0 + \frac{0.2}{2},\ 1 + \frac{0.02}{2}\right) \\ &= 0.2f(0.1,\ 1.01) \\ &= 0.2\left(\frac{0.1}{1.01}\right) \\ &= 0.02 \end{aligned} \tag{5.61}$$

$$\begin{aligned} k_4 = hf\left(x_0 + h,\ y_0 + k_3\right) &= 0.2f(0 + 0.2,\ 1 + 0.02) \\ &= 0.2f(0.2,\ 1.02) \\ &= 0.2\left(\frac{0.2}{1.02}\right) \\ &= 0.04 \end{aligned} \tag{5.62}$$

Now let substitute equations (5.59) to (5.62) into the main formula equation (5.58). Hence

$$\begin{aligned} y_1 &= y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ &= 1 + \frac{1}{6}[0 + 2(0.02) + 2(0.02) + 0.04] \\ &= 1.02 \end{aligned}$$

**Iteration 2: when i=1**
From the interval table (5.55), $x_1 = 0.2$, and from the previous solution $y_1 = 1.02$

The formula reduces to

$$y_2 = y_1 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{5.63}$$

where

$$k_1 = hf(x_1, y_1) = 0.2\left(\frac{0.2}{1.02}\right) = 0.04 \tag{5.64}$$

$$\begin{aligned} k_2 = hf\left(x_1 + \frac{h}{2},\ y_1 + \frac{k_1}{2}\right) &= 0.2f\left(0.2 + \frac{0.2}{2},\ 1.02 + \frac{0.04}{2}\right) \\ &= 0.2f(0.3,\ 1.04) \\ &= 0.2\left(\frac{0.3}{1.04}\right) \\ &= 0.06 \end{aligned} \tag{5.65}$$

$$\begin{aligned} k_3 = hf\left(x_1 + \frac{h}{2},\ y_1 + \frac{k_2}{2}\right) &= 0.2f\left(0.2 + \frac{0.2}{2},\ 1.02 + \frac{0.06}{2}\right) \\ &= 0.2f(0.3,\ 1.05) \\ &= 0.2\left(\frac{0.3}{1.05}\right) \\ &= 0.06 \end{aligned} \tag{5.66}$$

$$\begin{aligned} k_4 = hf\left(x_1 + h,\ y_1 + k_3\right) &= 0.2f(0.2 + 0.2,\ 1.02 + 0.06) \\ &= 0.2f(0.4,\ 1.08) \\ &= 0.2\left(\frac{0.4}{1.08}\right) \\ &= 0.079 \end{aligned} \tag{5.67}$$

Now let substitute equations (5.64) to (5.67) into the main formula equation (5.63). Hence

$$\begin{aligned} y_2 &= y_1 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ &= 1.02 + \frac{1}{6}[0.04 + 2(0.06) + 2(0.06) + 0.079] \\ &= 1.079 \end{aligned}$$

**Iteration 3: when i=2**
From the interval table (5.55), $x_2 = 0.4$, and from the previous solution $y_2 = 1.079$

The formula reduces to

$$y_3 = y_2 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{5.68}$$

where

$$k_1 = hf(x_2, y_2) = 0.2\left(\frac{0.4}{1.079}\right) = 0.074 \tag{5.69}$$

$$\begin{aligned} k_2 = hf\left(x_2 + \frac{h}{2},\ y_2 + \frac{k_1}{2}\right) &= 0.2f\left(0.4 + \frac{0.2}{2},\ 1.079 + \frac{0.074}{2}\right) \\ &= 0.2f(0.5,\ 1.116) \\ &= 0.2\left(\frac{0.5}{1.116}\right) \\ &= 0.09 \end{aligned} \tag{5.70}$$

$$\begin{aligned} k_3 = hf\left(x_2 + \frac{h}{2},\ y_2 + \frac{k_2}{2}\right) &= 0.2f\left(0.4 + \frac{0.2}{2},\ 1.079 + \frac{0.09}{2}\right) \\ &= 0.2f(0.5,\ 1.125) \\ &= 0.2\left(\frac{0.5}{1.125}\right) \\ &= 0.089 \end{aligned} \tag{5.71}$$

$$\begin{aligned} k_4 = hf\left(x_2 + h,\ y_2 + k_3\right) &= 0.2f(0.4 + 0.2,\ 1.079 + 0.089) \\ &= 0.2f(0.6,\ 1.168) \\ &= 0.2\left(\frac{0.6}{1.168}\right) \\ &= 0.103 \end{aligned} \tag{5.72}$$

Now let substitute equations (5.69) to (5.72) into the main formula equation (5.68). Hence

$$\begin{aligned} y_3 &= y_2 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ &= 1.079 + \frac{1}{6}[0.07 + 2(0.09) + 2(0.089) + 0.103] \\ &= 1.168 \end{aligned}$$

Therefore the nodal points are

$$(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1.02); \quad (x_2, y_2) = (0.4, 1.079); \quad (x_3, y_3) = (0.6, 1.168)$$

Hence

$$AE = |ES - AS| = |1.167 - 1.168| = 0.001$$

Comparing the absolute error, we may conclude that, the fourth-order Runge-Kutta scheme gives a more approximate solution than the other single-step schemes.

**Exercise 5.1** Solve the initial value problem

$$y' = 2x + 3y, \quad y(0) = 1, \quad x \in [0,\ 0.4], \quad h = 0.1$$

using

1. Euler method, hence determine the relative error at $x = 0.1$
2. Modified Euler method, hence determine the relative error at $x = 0.2$
3. Euler-Cauchy method, hence determine the relative error at $x = 0.3$
4. Fourth-Order Runge-Kutta, hence determine the relative error at $x = 0.4$

# 6. Numerical Solutions of Ordinary Differential Equations: Multi-step methods

**6.1 Multi-step Predictor Method** 65

6.1.1 Adams-Bashforth Scheme

**6.2 Multi-step Corrector Methods** 69

6.2.1 Adams-Moulton Method

6.2.2 Milne-Simpson Method

**6.3 Predictor-Corrector Methods** 71

A k-step multi-step method requires a previous $k$ number of values to start the iteration process. The $k$ values that are required for starting the iteration are obtained using some single-step schemes. The chosen single-step scheme should be of the same or lower order than the order of the multi-step method. A classical example of the explicit multi-step method is the **Adams-Bashforth scheme**. For the implicit schemes, we have the **Adams-Moulton scheme** and the **Milne-Simpson scheme**.

Explicit schemes are **predictor method**, while implicit schemes are **corrector methods**. However, multi-step schemes (explicit and implicit) are not self starting. They always require the assistance of some single-step methods to start the iteration.

Therefore, Euler, modified Euler, Euler-Cauchy and Runge-Kutta methods are single-step predictor methods, while backward Euler, and the trapezium methods are single-step corrector methods.

Moreover, the Adams-Bashforth scheme is a multi-step predictor method, while Adams-Moulton and Milne-Simpson are multi-step corrector methods.

## 6.1 Multi-step Predictor Method

### 6.1.1 Adams-Bashforth Scheme

A general Adams-Bashforth (AB) method is given by the equation

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\Delta f_i + \frac{5}{12}\Delta^2 f_i + \frac{3}{8}\Delta^3 f_i + \frac{251}{720}\Delta^4 f_i + \cdots\right] \tag{6.1}$$

A k-step AM method is of order $k$. By choosing different values for $k$, we obtain different schemes.

**Note 6.1**

$$\Delta f_i = f_i - f_{i-1} \tag{6.2}$$

again

$$\begin{aligned}\Delta^2 f_i &= \Delta f_i - \Delta f_{i-1}\\ &= (f_i - f_{i-1}) - (f_{i-1} - f_{i-2})\\ &= f_i - 2f_{i-1} + f_{i-2}\end{aligned} \tag{6.3}$$

Higher-order changes in $f_i$ can be deduced by the same continuous iterations.

Since equation (6.1) is an infinite series, each truncation yields a different iterative scheme.

**Case 1**
When $k = 1$, we have the first-order AB method or simply AB1 method. This is obtained by chopping equation (6.1) after the first term; considering the terms in the square bracket. This yields:

$$y_{i+1} = y_i + h(f_i) \tag{6.4}$$

$$= y_i + hf(x_i, y_i) \tag{6.5}$$

This is the same as the Euler method.

**Case 2**
For $k = 2$, we obtain the second-order AB method or simply AB2 method. This is obtained by chopping equation (6.1) after the second term; considering the terms in the square bracket. This yields:

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\Delta f_i\right] \tag{6.6}$$

$$= y_i + h\left[f_i + \frac{1}{2}(f_i - f_{i-1})\right] \tag{6.7}$$

$$= y_i + \frac{h}{2}[3f_i - f_{i-1}] \tag{6.8}$$

**Case 3**
For $k = 3$, we obtain the third-order AB method or simply AB3 method. This is obtained by chopping equation (6.1) after the third term; considering the terms in the square bracket. This yields:

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\Delta f_i + \frac{5}{12}\Delta^2 f_i\right] \tag{6.9}$$

Thus, (6.9) can be simplified as

$$y_{i+1} = y_i + \frac{h}{12}[23f_i - 16f_{i-1} + 5f_{i-2}] \tag{6.10}$$

**Case 4**
For $k = 4$, we obtain the fourth-order AB method or simply AB4 method. This is obtained

by chopping equation (6.1) after the fourth term ; considering the terms in the square bracket. This yields:

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\Delta f_i + \frac{5}{12}\Delta^2 f_i + \frac{3}{8}\Delta^3 f_i\right] \tag{6.11}$$

Thus, (6.11) can be simplified as

$$y_{i+1} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}] \tag{6.12}$$

**Note 6.2** The following are synonymous
$f_i = (x_i,\ y_i)$
$f_{i-1} = (x_{i-1},\ y_{i-1})$
$f_{i-2} = (x_{i-2},\ y_{i-2})$

The method is elaborated using the following example.

**Example 6.1.1** Solve the following IVP

$$yy' = x, \quad x \in [0,\ 1], \quad y(0) = 1, \quad h = 0.2$$

using the Adams-Bashforth method of third-order.
Compute all previous values of $y$ using Runge-Kutta of fourth-order.

**Solution**
Given the step size $h = 0.2$. Then the $x$ values are given by the **interval table**

$$x_0 = 0, \quad x_1 = 0.2, \quad x_2 = 0.4, \quad x_3 = 0.6, \quad x_4 = 0.8, \quad x_5 = 1 \tag{6.13}$$

We know that

$$y' = f(x, y) = \frac{x}{y} \tag{6.14}$$

Iteratively,

$$f(x_i, y_i) = \frac{x_i}{y_i} \tag{6.15}$$

AB3 is given by the formula

$$y_{i+1} = y_i + \frac{h}{12}[23f_i - 16f_{i-1} + 5f_{i-2}] \tag{6.16}$$

To begin the iteration, we identify an $i$th value when substituted into equation (6.16) the first $f$ value will be $f_0$. So we begin the iteration with $i = 2$.

**Iteration 1: when i=2**
Equation (6.16) reduces

$$y_3 = y_2 + \frac{h}{12}[23f_2 - 16f_1 + 5f_0] \tag{6.17}$$

We are required to find the values of $f_0, f_1$ and $f_2$ before a solution could be obtained. From the question, these values are to obtained using the Runge-Kutta scheme. We will skip the details of the computation here, since it has been solved under the single-step scheme in the previous chapter. We can recall that, the nodal points of the RK4 solution are:

$$(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1.02); \quad (x_2, y_2) = (0.4, 1.079); \quad (x_3, y_3) = (0.6, 1.168)$$

Only the first three point is needed for this computation.

Thus

$$\begin{aligned} f_0 &= f(x_0, y_0) = \frac{x_0}{y_0} = \frac{0}{1} = 0 \\ f_1 &= f(x_1, y_1) = \frac{x_1}{y_1} = \frac{0.2}{1.02} = 1.923 \\ f_2 &= f(x_2, y_2) = \frac{x_2}{y_2} = \frac{0.4}{1.079} = 0.3707 \end{aligned} \tag{6.18}$$

Now we can substitute equation (6.18) into equation (6.17):

$$\begin{aligned} y_3 &= y_2 + \frac{h}{12}[23f_2 - 16f_1 + 5f_0] \\ &= 1.079 + \frac{0.2}{12}[23(0.3707) - 16(1.923) + 5(0)] \\ &= 1.079 + 0.167(5.47) \\ &= 1.99 \end{aligned}$$

**Iteration 2: when i=3**
Equation (6.16) reduces

$$y_4 = y_3 + \frac{h}{12}[23f_3 - 16f_2 + 5f_1] \tag{6.19}$$

The new $f$ value that is to estimated is $f_3$, since the other values $f_2$ and $f_1$ are known. Therefore

$$\begin{aligned} f_1 &= f(x_1, y_1) = \frac{x_1}{y_1} = \frac{0.2}{1.02} = 1.923 \\ f_2 &= f(x_2, y_2) = \frac{x_2}{y_2} = \frac{0.4}{1.079} = 0.3707 \\ f_3 &= f(x_3, y_3) = \frac{x_3}{y_3} = \frac{0.6}{1.99} = 0.3 \end{aligned} \tag{6.20}$$

Note: The values for $f_3$ are from the interval table (6.13), $x_3 = 0.6$, and the previous solution $y_3 = 1.99$.

Now we can substitute equation (6.20) into equation (6.19)

$$\begin{aligned} y_4 &= y_3 + \frac{h}{12}[23f_3 - 16f_2 + 5f_1] \\ &= 1.99 + \frac{0.2}{12}[23(0.3) - 16(0.37) + 5(0.19)] \\ &= 1.99 + 0.167(1.95) \\ &= 2.31 \end{aligned}$$

**Iteration 3: when i=4**
Equation (6.16) reduces

$$y_5 = y_4 + \frac{h}{12}[23f_4 - 16f_3 + 5f_2] \tag{6.21}$$

The new $f$ value that is to estimated is $f_4$, since the other values $f_3$ and $f_2$ are known. Therefore

$$\begin{aligned} f_2 &= f(x_2, y_2) = \frac{x_2}{y_2} = \frac{0.4}{1.079} = 0.3707 \\ f_3 &= f(x_3, y_3) = \frac{x_3}{y_3} = \frac{0.6}{1.99} = 0.3 \\ f_4 &= f(x_4, y_4) = \frac{x_4}{y_4} = \frac{0.8}{2.31} = 0.346 \end{aligned} \tag{6.22}$$

Now we can substitute equation (6.22) into equation (6.21)

$$\begin{aligned} y_5 &= y_4 + \frac{h}{12}[23f_4 - 16f_3 + 5f_2] \\ &= 2.31 + \frac{0.2}{12}[23(0.346) - 16(0.3) + 5(0.3707)] \\ &= 2.31 + 0.167(5.008) \\ &= 3.146 \end{aligned}$$

Thus, the nodal points are

$$\begin{aligned} &(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1.02); \quad (x_2, y_2) = (0.4, 1.079); \\ &(x_3, y_3) = (0.6, 1.99), \quad (x_4, y_4) = (0.8, 2.31), \quad (x_5, y_5) = (1, 3.146) \end{aligned}$$

The first three $y$ points were computed using RK4, whiles the last three value were computed using AB3.

Thus, the multi-step method (AB3) was started with a single-step method (RK4).

## 6.2 Multi-step Corrector Methods

The two corrector methods considered here are, Adams-Moulton method and Milne-Simpson method.

### 6.2.1 Adams-Moulton Method

The general formula for the Adams-Moulton (AM) method is given by

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\Delta f_{i+1} - \frac{1}{12}\Delta^2 f_{i+1} - \frac{1}{24}\Delta^3 f_{i+1} - \frac{19}{720}\Delta^4 f_{i+1} - \cdots\right] \tag{6.23}$$

A k-step AM method is of order $k+1$. By choosing different k values, we arrive at different methods.

**Case 1**
When $k = 0$, we get the first-order AM method or AM1. This is obtained by truncating equation (6.23) after the first term; considering the terms in the square bracket. Thus

$$\begin{aligned} y_{i+1} &= y_i + hf_{i+1} && (6.24) \\ &= y_i + hf(x_{i+1}, y_{i+1}) && (6.25) \end{aligned}$$

This is the same as the **backward Euler method**.

**Case 2**
When $k = 1$, we get the second-order AM method or AM2. This is obtained by truncating equation (6.23) after the second term; considering the terms in the square bracket. Thus

$$\begin{aligned} y_{i+1} &= y_i + h\left[f_{i+1} - \frac{1}{2}\Delta f_{i+1}\right] && (6.26) \\ &= y_i + h\left[f_{i+1} - \frac{1}{2}(f_{i+1} - f_i)\right] && (6.27) \\ &= y_i + \frac{h}{2}\left[f_{i+1} - f_i\right] && (6.28) \end{aligned}$$

This is also the same as the **trapezium method**.

**Case 3**
When $k = 2$, we get the third-order AM method or AM3. This is obtained by truncating equation (6.23) after the third term; considering the terms in the square bracket. Thus

$$\begin{aligned} y_{i+1} &= y_i + h\left[f_{i+1} - \frac{1}{2}\Delta f_{i+1} - \frac{1}{12}\Delta^2 f_{i+1}\right] && (6.29) \\ &= y_i + h\left[f_{i+1} - \frac{1}{2}(f_{i+1} - f_i) - \frac{1}{12}(f_{i+1} - 2f_i + f_{i-1})\right] && (6.30) \\ &= y_i + \frac{h}{12}\left[5f_{i+1} + 8f_i - f_{i-1}\right] && (6.31) \end{aligned}$$

**Case 4**
When $k = 3$, we get the fourth-order AM method or AM4. This is obtained by truncating equation (6.23) after the fourth term; considering the terms in the square bracket. Thus

$$\begin{aligned} y_{i+1} &= y_i + h\left[f_{i+1} - \frac{1}{2}\Delta f_{i+1} - \frac{1}{12}\Delta^2 f_{i+1} - \frac{1}{24}\Delta^3 f_{i+1}\right] && (6.32) \\ &= y_i + \frac{h}{24}\left[9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}\right] && (6.33) \end{aligned}$$

### 6.2.2 Milne-Simpson Method

The general formula for the Milne-Simpson (MS) method is given by

$$y_{i+1} = y_{i-1} + h\left[2f_{i+1} - 2\Delta f_{i+1} + \frac{1}{3}\Delta^2 f_{i+1} - 0\times\Delta^3 f_{i+1} - \frac{1}{90}\Delta^4 f_{i+1} - \cdots\right] \tag{6.34}$$

We will skip the derivation of MS1, MS2 and MS3. For $k = 3$, we obtain the simplified form of MS4 or fourth-order Milne-Simpson method as

$$y_{i+1} = y_{i-1} + h\left[2f_{i+1} - 2\Delta f_{i+1} + \frac{1}{3}\Delta^2 f_{i+1} - 0\times\Delta^3 f_{i+1}\right] \tag{6.35}$$

Simplified as

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left[f_{i+1} + 4f_i + f_{i-1}\right] \tag{6.36}$$

## 6.3 Predictor-Corrector Methods

We have derived explicit single-step methods (Euler, Modified Euler, Euler-Cauchy, Runge-Kutta), explicit multi-step method (Adams-Bashforth), implicit single-step methods (Backward Euler, Trapezium) and implicit multi-step methods (Adams-Moulton, Milne-Simpson) for solving initial value problems of the form

$$y' = f(x,y), \qquad y(x_0) = y_0$$

If we perform analysis for numerical stability of these methods, we find that all explicit methods require very small step lengths to be used for convergence. If the solution of the problem is required over a large interval, we may need to use the method thousands or even millions of steps, which is computationally very expensive.

However, most implicit methods have strong stability properties, that is, we can use sufficiently large step lengths for computations and we can obtain convergence. But, we need to solve a nonlinear algebraic equation for the solution at each nodal point. This procedure may also be computationally expensive as convergence is to be obtained for the solution of the nonlinear equation at each nodal point.

Therefore, we combine the explicit methods (which have weak stability properties) and implicit methods (which have strong stability properties) to obtain new methods. Such methods are called **Predictor-Corrector methods** or PC methods.

The order of the predictor should be less than or equal to the order of the corrector. If the orders of the predictor and corrector are same, then we may require only one or two corrector iterations. For example, if the predictor and corrector are both of fourth order, then the combination (PC method) is also of fourth order and we may require one or two corrector iterations.

If the order of the predictor is less than the order of the corrector, then we require more iterations of the corrector. For example, if we use a first-order predictor and a second-order corrector, then one application of the combination gives a result of first order. If corrector is iterated once more, then the order of the combination increases by one, that is the result is now of second-order. If we iterate a third time, then the truncation error of the combination reduces, that is, we may get a better result. Further iterations may not change the results.

In the computations we will denote $P$ for predictor and $C$ for corrector.

Some possible combinations of the predictor-corrector methods are as fellow:

1. Predictor: Euler

$$y_{i+1}^{(p)} = y_i + hf(x_i, y_i)$$

Corrector: Backward Euler

$$y_{i+1}^{(c)} = y_i + hf(x_{i+1}, y_{i+1}^{(p)})$$

Both are of first-order.

2. Predictor: Adams-Bashforth of fourth-order

$$y_{i+1}^{(p)} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}]$$

Corrector: Adams-Moulton of fourth-order

$$y_{i+1}^{(c)} = y_i + \frac{h}{24}\left[9f(x_{i+1}, y_{i+1}^{(p)}) + 19f_i - 5f_{i-1} + f_{i-2}\right]$$

Below is an example of the predictor-corrector methods.

**Example 6.3.1** Using the fourth-order Adams-Bashforth-Moulton (ABM) predictor-corrector method evaluate $y(0.8)$ if

$$yy' = x, \quad x \in [0,\ 0.8], \quad y(0) = 1, \quad h = 0.2, \quad \epsilon = 0.001$$

Compute the necessary previous values using the Euler method.

**Solution**

Given the step size $h = 0.2$. Then the $x$ values are given by the **interval table**

$$x_0 = 0, \quad x_1 = 0.2, \quad x_2 = 0.4, \quad x_3 = 0.6, \quad x_4 = 0.8 \tag{6.37}$$

We know that

$$y' = f(x, y) = \frac{x}{y} \tag{6.38}$$

Iteratively,

$$f(x_i, y_i) = \frac{x_i}{y_i} \tag{6.39}$$

AB4 is given by the formula

$$y_{i+1}^{(p)} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}] \tag{6.40}$$

To begin the iteration, we identify an $i$th value when substituted into equation (6.40) the first $f$ value will be $f_0$. So we begin with $i = 3$

When $i = 3$, equation (6.40) reduces

$$y_4^{(p)} = y_3 + \frac{h}{24}[55f_3 - 59f_2 + 37f_1 - 9f_0] \tag{6.41}$$

We are required to find the values of $f_0, f_1, f_2$ and $f_3$ before a solution would be obtained. From the question, these values are to obtained using the Euler method. We will skip the details of the computation here, since it has been solved under the single-step scheme in the previous chapter. We recall the that, nodal points of the Euler solution are

$$(x_0, y_0) = (0, 1); \quad (x_1, y_1) = (0.2, 1); \quad (x_2, y_2) = (0.4, 1.04); \quad (x_3, y_3) = (0.6, 1.117)$$

All these points are needed for this computation.

Thus

$$\begin{aligned} f_0 &= f(x_0, y_0) = \frac{x_0}{y_0} = \frac{0}{1} = 0 \\ f_1 &= f(x_1, y_1) = \frac{x_1}{y_1} = \frac{0.2}{1} = 0.2 \\ f_2 &= f(x_2, y_2) = \frac{x_2}{y_2} = \frac{0.4}{1.04} = 0.385 \\ f_3 &= f(x_3, y_3) = \frac{x_3}{y_3} = \frac{0.6}{1.117} = 0.537 \end{aligned} \tag{6.42}$$

Now we can substitute equation (6.42) into equation (6.41)

**Predictor**

$$\begin{aligned} y_4^{(p)} &= y_3 + \frac{h}{24}[55f_3 - 59f_2 + 37f_1 - 9f_0] \\ &= 1.117 + \frac{0.2}{24}[55(0.537) - 59(0.385) + 37(0.2) - 9(0)] \\ &= 1.23611 \end{aligned}$$

**Corrector**

The Adams-Moulton of fourth-order is

$$y_{i+1}^{(c)} = y_i + \frac{h}{24}\left[9f(x_{i+1}, y_{i+1}^{(p)}) + 19f_i - 5f_{i-1} + f_{i-2}\right]$$

From here, we can begin the iterations. The $y_{i+1}^{(p)}$ values are coming from the predictor solution.

**Iteration 1**

When $i = 3$, the AM4 is given us

$$\begin{aligned} y_4^{(c_1)} &= y_3 + \frac{h}{24}\left[9f(x_4, y_4^{(p)}) + 19f_3 - 5f_2 + f_1\right] \\ &= 1.117 + \frac{0.2}{24}\left[9\left(\frac{0.8}{1.23611}\right) + 19(0.537) - 5(0.385) + 0.2\right] \\ &= 1.236 \end{aligned}$$

We will stop the iteration procedure when the stopping criterion is satisfied. Since is the very first iteration, we will need the next iteration to make comparison.

In the next step, we replace $y_{i+1}^{(p)}$ with $y_{i+1}^{(c_1)}$ to obtain an iterative scheme in terms of a corrector function.

**Iteration 2**

$$\begin{aligned} y_4^{(c_2)} &= y_3 + \frac{h}{24}\left[9f(x_4, y_4^{(c_1)}) + 19f_3 - 5f_2 + f_1\right] \\ &= 1.117 + \frac{0.2}{24}\left[9\left(\frac{0.8}{1.236}\right) + 19(0.537) - 5(0.385) + 0.2\right] \\ &= 1.236 \end{aligned}$$

Check stopping criterion: $|y_4^{(c_2)} - y_4^{(c_1)}| = |1.236 - 1.236| = 0 \implies < \epsilon$. Hence stop iterations.

Therefore

$$y(0.8) = 1.236$$

**Exercise 6.1**

1. Using the Runge-Kutta method of order 4, find the $y's$ for $x = 0.1, 0.2, 0.3$ given that
$$\frac{dy}{dx} = xy + y^2, \qquad y(0) = 1$$
Hence find the solution at $x = 0.4$ using the Milne-Simpson method with $\epsilon = 0.05$

2. Given that
$$\frac{dy}{dx} = x^2(1+y), \quad y(1) = 1,\ y(1.1) = 1.233,\ y(1.2) = 1.548,\ y(1.3) = 1.979$$
Evaluate $y(1.4)$ using Adams-Bashforth-Moulton of fourth-order with $\epsilon = 0.05$.

# Bibliography

[1] Ward Cheney and David Kincaid. *Numerical Mathematics and Computing*. Brooks/Cole, 6th edition, 2008.

[2] Douglas J. Faires and Richard L. Burden. *Numerical Methods*. Brooks/Cole, 3rd edition, 2002.

[3] Joe E. Hoffman. *Numerical Methods for Engineers and Scientist*. Marcel Dekker, Inc, 2nd edition, 2001.

[4] S. R. K. Iyengar and R. K. Jain. *Numerical Methods*. New Age International, 2009.