Numerical Methods

AIMA 203

Department of Artificial Intelligence,
Kathmandu University, Panchkhal, Kavre

Lecture notes by
Sandesh Thakuri
Sandesh.775509@cdmath.tu.edu.np

Text Book: S.S. Sastary, Introductory Methods of Numerical Analysis, Prentice Hall of India.

2025

Contents

Mathematical Preliminaries

In this chapter we state, without proof, certain mathematical results which would be useful in the sequel.

Theorem 1 (**Rolle's Theorem**)**.** If $f(x)$ is continuous in $[a, b]$, $f'(x)$ exists in (a, b) and $f(a) = f(b) = 0$, then, there exists at least one number $c \in (a, b)$ such that $f'(c) = 0$.

Theorem 2 (**Intermediate value theorem**)**.** Let $f(x)$ be continuous in $[a, b]$ and let k be any number between $f(a)$ and $f(b)$. Then there exists a number $c \in (a, b)$ such that$f(c) = k$.

1 Errors in Numerical Computing

1.1 Introduction

In practical applications, an engineer would finally obtain results in a numerical form. The aim of numerical analysis is to provide efficient methods for obtaining numerical answers to such problems.

1.1.1 Approximate Value

There are certain numbers whose exact value cannot be written. For the famous number π, we can only write value of π to certain degree of accuracy. For example, π is 3.1416 or 3.14159265. This values of π are called the approximate values of π. The exact value of π, we cannot write. Another such number is the *Euler's number 'e'*.

1.1.2 Significant Digits

(Significant Figures) The digits that are significant (important) in a number expressed as digits, are called significant digits.

Rules to identify significant figures in a number:

1. Non-zero digits within the given measurement or reporting resolution are significant.
2. Zeros between two significant non-zero digits are significant.
3. Zeros to the left of the first non-zero digit (leading zeros) are not significant.
4. Zeros to the right of the last non-zero digit (trailing zeros) in a number with the decimal point are significant.
5. Trailing zeros in an integer may or may not be significant.
6. Exact value of a number has an infinite number of significant digits.

1.1.3 Rounding off

In numerical computation, we come across numbers which have large number of digits and it will be necessary to bring them to the required number of significant figures. The first such process is Rounding off.

Rules for rounding off to a number of significant digits.

1. To round-off a number to n significant digits , discard all digits to the right of the nth digit, and if this discarded number is,
2. less than half a unit in the nth place, leave the nth digit unaltered;
3. greater than half a unit in the nth place, increase the nth digit by unity;
4. exactly half a unit in the nth place, increase the nth digit by unity if it is odd; otherwise, leave it unchanged.

1.1.4 Truncation

Second way of writing a large number of digits to a significant figure is by truncating the number to required number of digits.

1.2 Error

Both the rounding off and truncation causes error as they reduce a given number to an approximate value. Using approximate instead of exact value of number gives rise to the error.

1.2.1 Types of Error

Absolute Error

The absolute error E_A is the difference between the exact-value X and the approximate-value X_1 of a number.

$$E_A = X - X_1 = \delta X \tag{1.1}$$

Relative Error

The relative error E_R is the ratio of the absolute error to the exact-value.

$$E_R = \frac{E_A}{X} = \frac{X - X_1}{X} \tag{1.2}$$

Percentage Error

The percentage error E_P is

$$E_P = 100(E_R) = \frac{X - X_1}{X} \times 100 \tag{1.3}$$

1.2.2 Absolute errors of Sum, Product and Product

Suppose a_1 and a_2 are two approximate values of a number with their absolute errors E_A^1 and E_A^2 respectively.

Sum

If E_A is the absolute error of $a_1 + a_2$ then,

$$E_A = (a_1 + E_A^1) + (a_2 + E_A^2) - (a_1 + a_2) = E_A^1 + E_A^2 \tag{1.4}$$

Product

If E_A is the absolute error of $a_1 a_2$ then,

$$E_A = (a_1 + E_A^1)(a_2 + E_A^2) - (a_1 a_2) = a_1 E_A^2 + a_2 E_A^1 + E_A^1 E_A^2 \approx a_1 E_A^2 + a_2 E_A^1 \tag{1.5}$$

Quotient

If E_A is the absolute error of a_1/a_2 then,

$$E_A = \frac{a_1 + E_A^1}{a_2 + E_A^2} - \frac{a_1}{a_2} = \frac{a_2 E_A^1 - a_1 E_A^2}{a_2(a_2 + E_A^2)} = \frac{a_2 E_A^1 - a_1 E_A^2}{a_2 a_2(1 + E_A^2/a_2)} \approx \frac{a_2 E_A^1 - a_1 E_A^2}{(a_2)^2}$$

This implies,

$$E_A = \frac{a_1}{a_2}\left[\frac{E_A^1}{a_1} - \frac{E_A^2}{a_2}\right] \tag{1.6}$$

1.2.3 Operations on numbers of different absolute accuracies

While dealing with several numbers of different number of significant digits, the following procedure may be adopted:

1. Isolate the number with the greatest absolute error,
2. Round-off all other numbers retaining in them one digit more than in the isolated number,
3. Add up, and
4. Round-off the sum by discarding one digit.

1.2.4 Upper limit of Absolute Error

The number ΔX such that

$$|X_1 - X| \leq \Delta X. \tag{1.7}$$

Then ΔX is an upper limit on the magnitude of the absolute error and is said to measure *absolute accuracy.*

Theorem 3. If the number X is rounded to N decimal places, then $\Delta X = \frac{1}{2}(10^{-N})$.

1.2.5 Exercise

1. Round off the following numbers to two decimal places.
   48.21461, 2.3742, 52.275

2. Round off the following numbers to four significant figures:
   38.46235, 0.70029, 0.0022218, 19.235101, 2.36425

3. Two numbers 3.1425 and 34.5851 are rounded to 2 decimal places. Find the error in their sum, product and quotient.

4. Find the absolute error in the sum of the numbers 105.6, 27.28, 5.63, 0.1467, 0.000523, 208.5, 0.0235, 0.432, 0.0467, where each number is correct to the digits given.

2 Root Finding

2.1 Introduction

Roots of an equation

$$f(x) = 0 \tag{2.1}$$

are the zeros, of f, which means, the values of x that makes the value of f zero. Basically equations are categorized into two. If f is a polynomial then the Equation-2.1 is a **polynomial equation** and if f is a non-polynomial then Equation-2.1 is a **transcendental equation**. For the polynomial equations following results hold:

1. Every polynomial equation of degree n has at most n real roots.

2. If n is odd then, the polynomial equation has at least one real root whose sign is opposite to that of the last term.

3. If n is even and the constant term is negative, then the equation has at-least one positive root and at-least one negative root.

4. The imaginary roots occurs in a pair (conjugate-pair). If the coefficients of f are rationals then, the irrational roots occurs in pairs (conjugate-pair).

5. **Descartes' Rule of Signs**

   a). A polynomial equation cannot have more number of positive real roots than the number of changes of signs in the coefficients of $f(x)$.

   b). A polynomial equation cannot have more number of negative real roots than the number of changes of signs in the coefficients of $f(-x)$.

2.2 Bisection Method

Theorem 4 (Bolzano's Theorem)**.** If f(x) is continuous in $[a, b]$, and if $f(a)$ and $f(b)$ are of opposite signs, then $f(c) = 0$ for at least one number $c \in (a, b)$.

The Bisection method is based on Theorem-4. The word "bisection" means "half". Using this method the root c of f is given by $c \approx \frac{a+b}{2}$. Let $x_1 = \frac{a+b}{2}$. If $f(x_1) \neq 0$ then, the root, c lies either in $[a, x_1]$ or in $[x_1, b]$. If $f(a)f(x_1) < 0$ then, c lies in $[a, x_1]$ else, it lies in $[x_1, b]$.

At each step of this method, the given interval is bisected, so the length of the interval is halfed. At nth step the length of the interval is $\frac{|b-a|}{2^n}$. If the tolerance of the given approximation is ϵ then we must have $\frac{|b-a|}{2^n} \leq \epsilon$. And the number of steps required to reach this accuracy is $n \geq log_2(|b-a|)$.

2.2.1 Procedure

1. Choose two real numbers a and b such that $f(a)f(b) < 0$.
2. Set $x_0 = 0$ and $x_1 = \frac{a+b}{2}$.
3. Do
   $\epsilon_r = \left|\frac{x_0 - x_1}{x_0}\right|$
   If $\epsilon_r < tolerance$ then $root = x_1$,
   else
   $x_0 = x_1$ and if $f(a)f(x_1) < 0$ then $x_1 = \frac{a+x_1}{2}$
   if $f(x_1)f(b) < 0$ then $x_1 = \frac{x_1+b}{2}$.

2.2.2 Exercise

Using Bisection Method:

1. Find a root of $f(x) = x^3 - x - 1 = 0$, correct to 4 decimal places

2.3 Iteration Method

Steps:

1. Re-write the given equation $f(x) = 0$ in the form $x = \phi(x)$. This equation is of **iterative-type**. Meaning we can substitute a value of x in $\phi(x)$ to get another value of x, and continue this process to get the desired value of x if the iteration is of convergent one.
2. Choose an initial root of f, x_0.
3. $x_1 = \phi(x_0)$, $x_2 = \phi(x_1)$ and so on.

The sequence $x_0, x_1, x_2, ...$ may not converge to a definite number. But if the sequence converges to a definite number ζ, then ζ is a root of the given equation.

2.3.1 Exercise

Using Iteration Method:

1. Find a root of $2x - 3 - cosx = 0$, correct to 3 decimal places

2.4 Newton-Rapshon's Method

Steps:

1. Choose an initial guess solution of the given equation $f(x) = 0$, x_0.
2. Let x_1 be a solution, which is more close to the exact solution of $f(x) = 0$. Then Using Taylor's expansion of f about x_0:

$$f(x_1) = f(x_0) + (x_1 - x_0)f'(x_0) + (x_1 + x_0)^2 f''(x_0) + ... = 0$$

Neglecting the second and higher order derivatives, we get

$$f(x_0) + (x_1 - x_0)f'(x_0) = 0 \tag{2.2}$$

The equation-2.8 is a linear equation, so this is an linear approximation. This equation is infact the tangent to the curve of the function $f(x)$ at $(x_0, f(x_0)$. And it is the point x_1 where the tangent meets the $x-axis$. So, the next approximation after x_0 by Newton-Rapshon's method is the point on $x - axis$, where the tangent to the f at x_0 meets the $x - axis$. This point can be solved as follows:

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

3. Successive approximation are given by $x_2, x_3, x_4, ...$, where $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

2.4.1 Exercise

Using Newton-Rapshon's Method:

1. Find a root of $f(x) = xe^x - 1 = 0$, correct to 4 decimal places

2.5 Secant Method

In Newton-Rapshon's method we use a tangent to the curve to get close to the root of the function. So, Newton-Rapshon's method requires the evaluation of derivatives of the function, which may not always exit. So we replace the tangent, with a secant to approximate the root of the function.

Steps:

1. Choose two initial guess solutions of the given equation $f(x) = 0$, x_{-1} and x_0.

2. The slope of the secant is $\dfrac{f(x_0)-f(x_{-1})}{x_0-x_{-1}}$.

3. Then equation of the line passing through the points of given by the two initial guesses is $f(x)-f(x_0)=\dfrac{f(x_0)-f(x_{-1})}{x_0-x_{-1}}(x-x_0)$.

4. x_1 is the point where the secant meets the $x-axis$ so, $f(x_1)=0$. This gives,

$$\begin{aligned} 0-f(x_0) &= \frac{f(x_0)-f(x_{-1})}{x_0-x_{-1}}(x_1-x_0) \\ x_1-x_0 &= -\frac{x_0-x_{-1}}{f(x_0)-f(x_{-1})}f(x_0) \\ x_1 &= x_0-\frac{x_0-x_{-1}}{f(x_0)-f(x_{-1})}f(x_0) \end{aligned}$$

5. This generalizes to

$$x_{n+1}=x_n-\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}f(x_n) \tag{2.3}$$

You can get this relation-2.3 just by plugging $f'(x_n)=\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$ as the slope of the tangent in Newton Rapshon's method is just approximated by the slope of the secant in Secant method.

2.6 System of Non-linear equations

For now we consider only a system of two equations. Let a system of two equations be

$$f(x,y)=0, \qquad g(x,y)=0 \tag{2.4}$$

2.6.1 Method of Iteration

First we assume that the sytem of equations 2.4 may be written in the form

$$x=F(x,y), \qquad y=G(x,y) \tag{2.5}$$

where the function F and G satisty the following conditions in a **closed** neighborhood of R of the root (α,β):

i) F and G and their firt partial derivatives are continuous in R, and

ii) $\left|\dfrac{\partial F}{\partial x}\right|+\left|\dfrac{\partial F}{\partial y}\right|<1$ and $\left|\dfrac{\partial G}{\partial x}\right|+\left|\dfrac{\partial G}{\partial y}\right|<1$, for all (x,y) in R.

If (x_0,y_0) is an initial approximation to the root (α,β), then Equations 2.5 give the sequence

$$\begin{aligned} x_1 &= F(x_0,y_0), & y_1 &= G(x_0,y_0) \\ x_2 &= F(x_1,y_1), & y_2 &= G(x_1,y_1) \\ &\dots \\ x_{n+1} &= F(x_n,y_n), & y_{n+1} &= G(x_n,y_n) \end{aligned} \tag{2.6}$$

For faster convergence, recently computed values of x_i may be used in the evaluation of y_i in Equations. Above conditions are sufficient for convergence and in the limit we obtain,

$$\alpha = F(\alpha, \beta) \quad and \quad \beta = G(\alpha, \beta) \tag{2.7}$$

Hence (α, β) is the root of the system 2.4.

2.6.2 Newton-Raphson Method

Let $(x_0, y_0$ be an initial approximation to the root of the system of equations in two variables 2.4. If $(x_0 + h, y_0 + k$ is the root of the system, then we must have

$$f(x_0 + h, y_0 + k) = 0 \qquad g(x_0 + h, y_0 + k) = 0$$

Assuming that f and g are sufficiently differentiable, we expand both of these functions by Taylor's series to obtain

$$f_0 + h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} \ldots = 0$$
$$g_0 + h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} \ldots = 0$$

where, $\dfrac{\partial f}{\partial x_0} = \left[\dfrac{\partial f}{\partial x}\right]_{x=x_0}$, $f_0 = f(x_0, y_0)$, etc

Neglating the second and higher-order derivatives terms, we get,

$$h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} \ldots = -f_0$$
$$h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} \ldots = g_0 \tag{2.8}$$

The system of equations 2.8 possesses a unique solution if

$$D = \begin{vmatrix} \frac{\partial f}{\partial x_0} & \frac{\partial f}{\partial y_0} \\ \frac{\partial g}{\partial x_0} & \frac{\partial g}{\partial y_0} \end{vmatrix} \neq 0$$

By Cramer's rule

$$h = \frac{1}{D}\begin{vmatrix} -f_0 & \frac{\partial f}{\partial y_0} \\ -g_0 & \frac{\partial g}{\partial y_0} \end{vmatrix} \quad and \quad k = \frac{1}{D}\begin{vmatrix} \frac{\partial f}{\partial y_0} & -f_0 \\ \frac{\partial g}{\partial y_0} & -g_0 \end{vmatrix} \tag{2.9}$$

The new approximations are, therefore

$$x_1 = x_0 + h \quad and \quad y_1 = y_0 + k \tag{2.10}$$

2.6.3 Exercise

1. Find a real root of the system: $y^2 - 5y + 4 = 0$ and $3x^2y - 10x + 7 = 0$ correct to 4 decimal places using initial approximation $(0, 0)$.
2. Solve the system: $x^2 + y = 11$, $x + y^2 = 7$.
3. Solve the system: $x^2 - y^2 = 4$, $x^2 + y^2 = 16$.