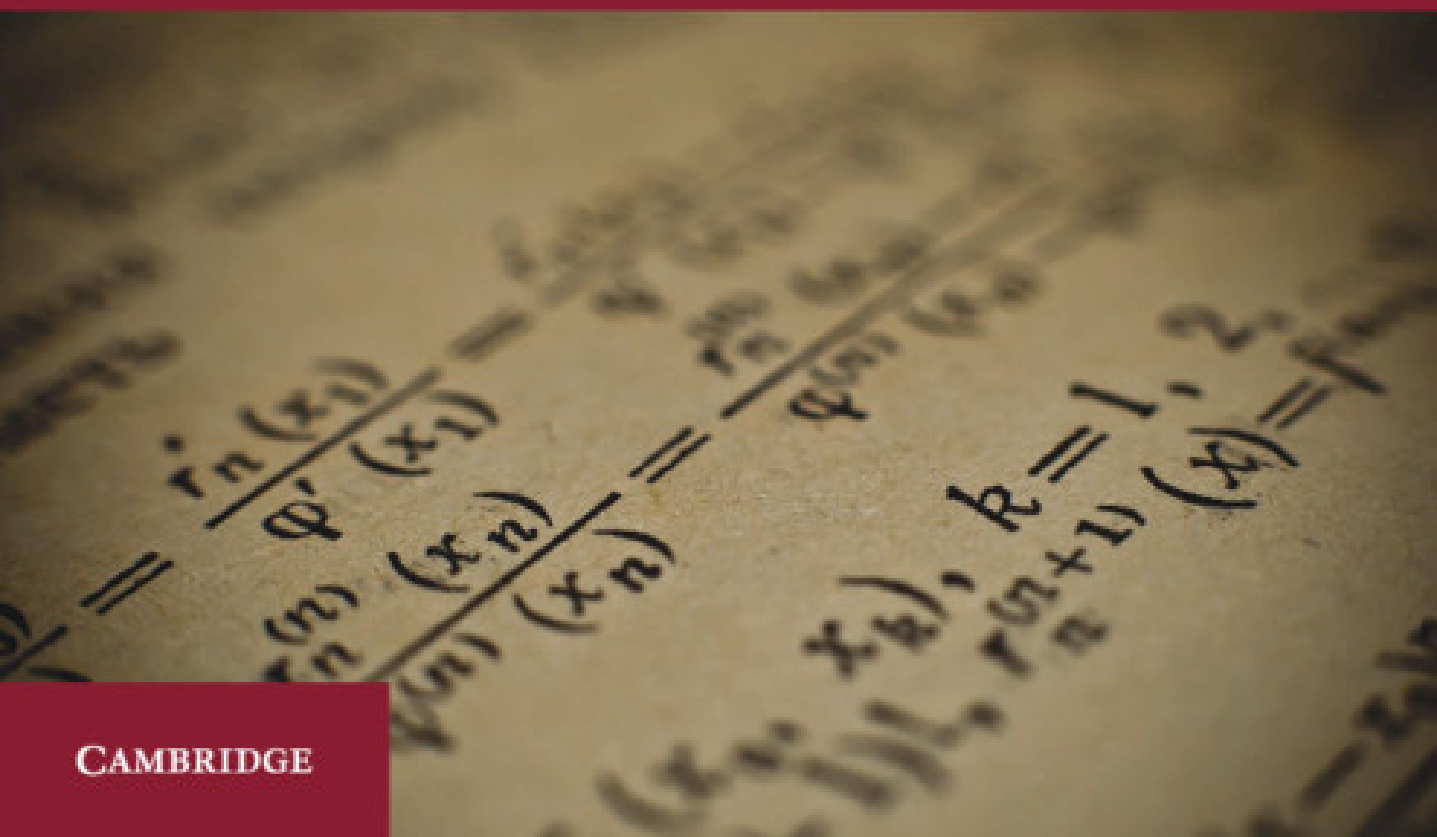
Numerical Methods

Fundamentals and Applications

Rajesh Kumar Gupta



CAMBRIDGE

Numerical Methods

Numerical methods play an important role in solving complex engineering and science problems. This textbook provides essential information on a wide range of numerical techniques, and it is suitable for undergraduate and postgraduate/research students from various engineering and science streams. It covers numerical methods and their analysis to solve nonlinear equations, linear and nonlinear systems of equations, eigenvalue problems, interpolation and curve-fitting problems, splines, numerical differentiation and integration, ordinary and partial differential equations with initial and boundary conditions. C-programs for various numerical methods are presented to enrich problem-solving capabilities. The concepts of error and divergence of numerical methods are described by using unique examples. The introductions to all chapters carry graphical representations of the problems so that readers can visualize and interpret the numerical approximations.

C-Programs are available at www.cambridge.org/9781108716000

Rajesh Kumar Gupta is an associate professor of mathematics at Central University of Haryana and Central University of Punjab (on lien), India. He has more than 13 years of teaching and research experience. He has published 65 research papers in reputed international journals on the applications of Lie symmetry analysis to nonlinear partial differential equations governing important physical phenomena and related fields.

Numerical Methods

Fundamentals and Applications

Rajesh Kumar Gupta

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, vic 3207, Australia

314 to 321, 3rd Floor, Plot No.3, Splendor Forum, Jasola District Centre, New Delhi 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108716000

First published 2019

Printed in India

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Gupta, Rajesh Kumar, 1979 author.
Title: Numerical methods: fundamentals and applications / Rajesh Kumar Gupta.
Description: Cambridge; New York, NY: Cambridge University Press, 2019. |
Includes bibliographical references and index.
Identifiers: LCCN 2019013359 | ISBN 9781108716000 (alk. paper)
Subjects: LCSH: Numerical analysis—Problems, exercises, etc. | Mathematical notation.
Classification: LCC QA297 .G8725 2019 | DDC 518—dc23 LC record available at https://lccn.loc.gov/2019013359

ISBN 978-1-108-71600-0 Paperback

Additional resources for this publication at www.cambridge.org/9781108716000

To My Parents

Sh. Murari Lal and Smt. Santosh Devi

To My Teacher

Professor Karanjeet Singh

To My Wife and Children

Dr Usha Rani Gupta and Aaradhya and Reyansh

Contents

Preface

There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world.

Nikolai Ivanovich Lobachevsky
(December 1, 1792–February 24, 1856)
His work is mainly on hyperbolic geometry, also known as Lobachevskian geometry.

The rapid growth of science and technology during the last few decades has made a tremendous change to the nature of various mathematical problems. It is not easy to solve these new problems for analytical solutions by conventional methods. In fact, the study of these mathematical problems for analytical solutions is not only regarded as a difficult endeavor, rather it is almost impossible to get analytical solutions in many cases. The tools for analysis and for obtaining the analytical solutions of complex and nonlinear mathematical systems are limited to very few special categories. Due to this reason, when confronted with such complex problems we usually simplify them by invoking certain restrictions on the problem and then solve it. But these solutions, however, fail to render much needed information about the system. These shortcomings of analytical solutions lead us to seek alternates, and various numerical techniques developed for different types of mathematical problems seem to be excellent options. During the last century, the numerical techniques have witnessed a veritable explosion in research, both in their application to complex mathematical systems and in the very development of these techniques. At many places in this book, we will compare numerical techniques with analytical techniques, and point out various problems which can not be solved through analytical techniques, and to which numerical techniques provide quite good approximate solutions.

Many researchers are using numerical techniques to investigate research problems. Numerical techniques are now widely used in a lot of engineering and science fields. Almost all universities now offer courses on introductory and advanced computer-oriented numerical methods to their engineering and science students, keeping in mind the utilization merits of these techniques. In addition, computer-oriented problems are part of various other courses of engineering/technology.

It gives me immense pleasure in presenting the book to our esteemed readers. This book is written keeping several goals in mind. It provides essential information on various numerical techniques to the students from various engineering and science streams. The aim of the book is to make the subject easy to understand, and to provide in-depth knowledge about various numerical tools in a simple and concise manner.

Students learn best when the course is problem-solution oriented, especially when studying mathematics and computing. This book contains many examples for almost all numerical techniques designed from a problem-solving perspective. In fact, theoretical and practical introductions to numerical techniques and worked examples make this book student-friendly.

While the main emphasis is on problem-solving, sufficient theory and examples are also included in this book to help students understand the basic concepts. The book includes theories related to errors and convergence, limitations of various methods, comparison of various methods for solving a specific type of problem and scope for further improvements, etc.

The practical knowledge of any subject is thought to be an essential part of the curriculum for an engineering student. Numerical methods require tedious and repetitive arithmetic operations, wherein for large-scale problems it is almost impossible to do such cumbersome arithmetic operations manually. Fortunately most numerical techniques are algorithmic in nature, so it is easy to implement them with the aid of a computer. To enrich problem-solving capabilities, we have presented the basic C-programs for a wide range of methods to solve algebraic and transcendental equations, linear and nonlinear systems of equations, eigenvalue problems, interpolation problems, curve fitting and splines, numerical integration, initial and boundary value problems, etc.

The section below provides an overview of the contents of the book. Each chapter contains a brief introduction and it also emphasis the need for numerical techniques for solving specific problems. We have provided exercises in all chapters with the aim of helping students check their capabilities and understanding, and also illustrate how various numerical methods are the better problem solvers.

Chapter-by-chapter Introduction to the Book

The book comprises sixteen chapters.

Chapter 1: Number Systems explains integral and fractional numbers in the binary, octal, decimal and hexadecimal number systems. It also includes the conversion from one number system to another number system.

Chapter 2: Error Analysis primarily presents various types of errors, and some standard remedies to trace and reduce these errors.

Except Chapters 1 and 2, all other chapters of this book have been devoted to numerical techniques which are used to solve some specific type of problems. In each chapter, various numerical methods will be introduced to solve these problems.

Chapter 3: Nonlinear Equations consists of various techniques to solve nonlinear equations in single variable. Primary aim is to determine the value of variable or parameter x, called root of the equation that satisfies the equation

$$f(x) = 0$$

Roots of simple equations like quadratic equation $x^2 - 3x + 2 = 0$ can be obtained easily. But in the case of higher order polynomial equations like $3x^5 + x^4 + 3x^3 - 2x^2 - 3x + 9 = 0$ and transcendental equations viz. $2e^x \cos x - x = 0$, we do not have any general method to compute the roots of these equations. Numerical techniques will be helpful for computing roots of such equations.

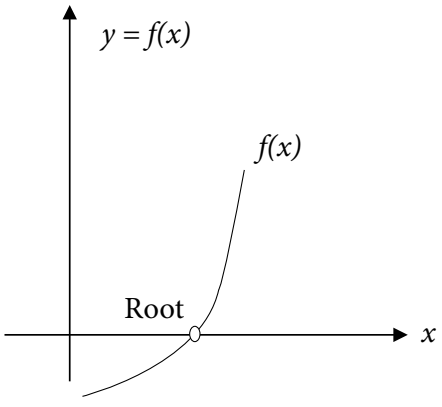


Fig. 1 Root of $f(x) = 0$

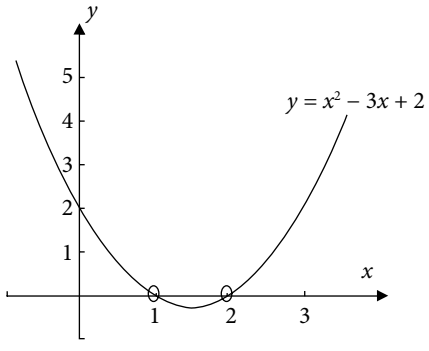


Fig. 2 Roots of $x^2 - 3x + 2 = 0$

These problems are especially valuable in engineering design contexts where due to the complexity of the design equations it is often impossible to solve these equations with analytical methods.

Chapter 4: Nonlinear Systems and Polynomial Equations deals with the numerical techniques to solve the systems of nonlinear equations, say, the system of two equations $\begin{matrix} f(x,\ y) = 0 \\ g(x,\ y) = 0 \end{matrix}$.

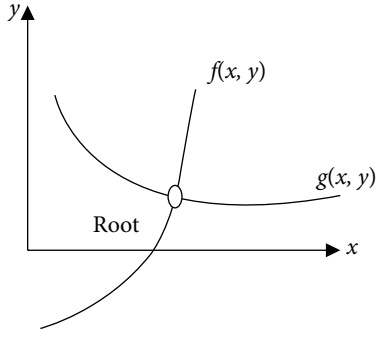


Fig. 3 Solution of $\begin{matrix} f(x,\ y) = 0 \\ g(x,\ y) = 0 \end{matrix}$

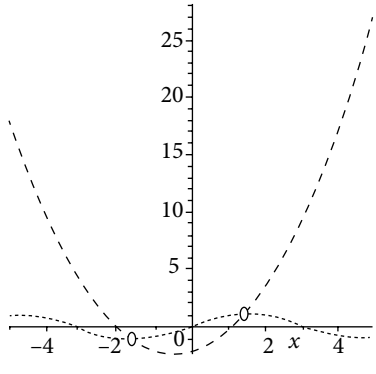


Fig. 4 Roots of $\begin{matrix} y - \sin(x) = 0 \\ x^2 + x - y - 2 = 0 \end{matrix}$

The aim is to find coordinate (x, y), which satisfies these two equations simultaneously. Since there is no general analytical method for the solution of such systems of nonlinear equations, therefore we will apply numerical methods to solve such kind of problems. This chapter also includes some numerical methods for the roots of polynomial equations.

Chapter 5: Systems of Linear Equations is devoted to obtain solution of the system of linear algebraic equations

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ &\vdots \\ &\vdots \\ &a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{aligned}$$

e.g.,
$$\begin{aligned} x_1 - 2x_2 + 3x_3 &= 15 \\ 2x_1 - x_2 + 3x_3 &= 15 \quad \text{with } n = 3. \\ x_1 + x_2 - 3x_3 &= -9 \end{aligned}$$

In case of system of two algebraic equations, we have two lines, and their point of intersection is the solution.

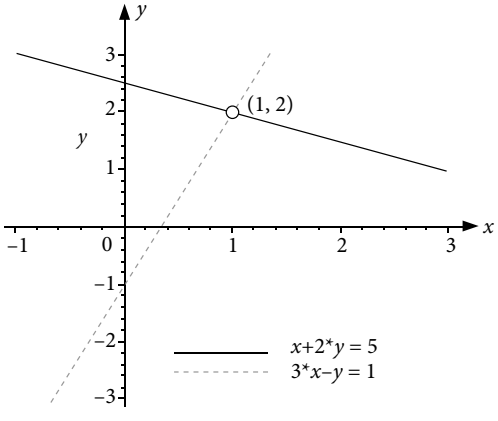


Fig. 5 Linear system in two variables (x, y)

Such equations have many important applications in science and engineering, specifically in the mathematical modeling of large systems of interconnected elements such as electrical circuits, structures, lattice and fluid networks, etc. In this chapter, we will discuss various direct and iterative methods to solve these systems of linear equations. Also, we will discuss problems that arise in applying these methods on the computer and some remedies for these problems.

Chapter 6: Eigenvalues and Eigenvectors is to deduce eigenvalues and eigenvectors for a square matrix A. A column vector X is an eigenvector corresponding to eigenvalue λ of a square matrix A, if

$$AX = \lambda X. \qquad \text{(or)} \qquad (A - \lambda \mathrm{I})X = 0$$

The nontrivial solutions of this homogeneous system exist, only if

$$p(\lambda) = \det(A - \lambda \mathrm{I}) = 0$$

$p(\lambda)$ is the polynomial of degree n for a square matrix of order n. There are only n eigenvalues of matrix A, including repetitions (eigenvalues may be complex). The polynomial $p(\lambda)$ is known as characteristic polynomial, and the equation $p(\lambda) = 0$ is called characteristic equation.

For example, the characteristic equation for the matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$ is given by

$$p(\lambda) = |A - \lambda \mathrm{I}| = \begin{vmatrix} 1-\lambda & 2 \\ 3 & 2-\lambda \end{vmatrix} = (\lambda - 4)(\lambda + 1) = 0$$

The roots of the characteristic equation give eigenvalues – 1 and 4.

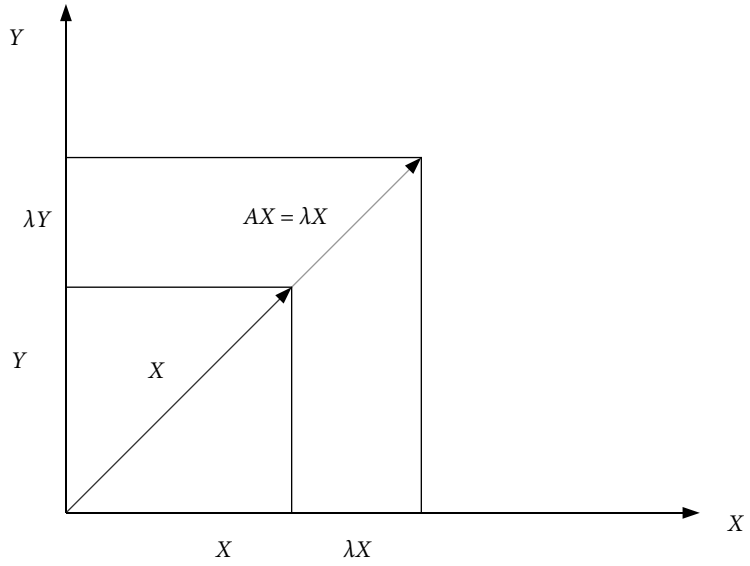


Fig. 6 Eigenvalue λ and eigenvector X of matrix A

These types of problems arise in different streams of science and engineering especially in the case of oscillatory systems like elasticity, vibrations, etc.

Chapter 7: Eigenvalues and Eigenvectors of Real Symmetric Matrices deals with the eigenvalues and eigenvectors of real symmetric matrices. Some methods are applicable only to real symmetric matrices. Since these methods are easy to implement and provide all the eigenvalues and eigenvectors at a time, hence need more exploration.

Chapter 8: Interpolation is most important part of numerical methods, as it deals with the approximation of the data sets with the polynomials. This chapter deals with the task of constructing a polynomial function $P(x)$ of minimum degree, which passes through a given set of discrete data points (x_i, y_i), $i = 0, 1, ..., n$. This polynomial is known as interpolating polynomial. It estimates the value of the dependent variable y for any intermediate value of the independent variable, x.

For example: consider the data set (0, –1), (1, 1), (2, 9), (3, 29), (5, 129). The aim is to construct a polynomial of minimum degree which passes through all these points. We will discuss methods to construct such polynomial. The polynomial $P(x) = x^3 + x - 1$ is the required polynomial and it passes through all these points.

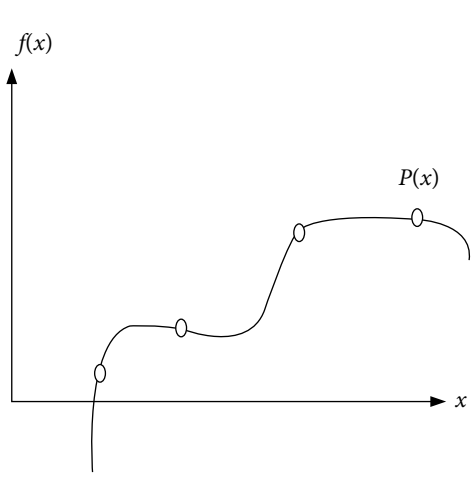


Fig. 7 Interpolation

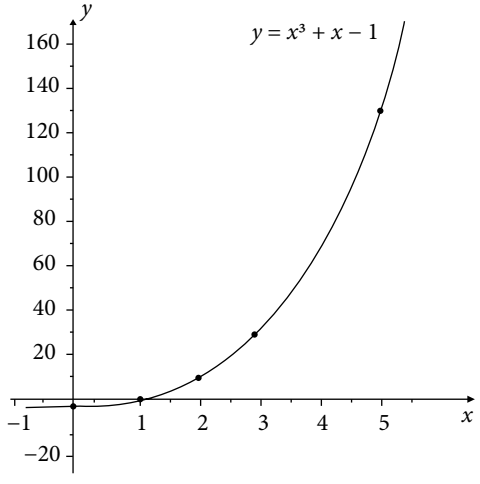


Fig. 8 Interpolating polynomial for data set (0, –1), (1, 1), (2, 9), (3, 29), (5, 129)

A data set is either the table of values of well-defined functions or the table of data points from observations during an experiment. These types of problems are most common in various experiments where only inputs and corresponding outputs are known. In most of the experimental cases, we have data points, i.e., inputs (x) and correspondingly outputs (y). Also, many practical problems involve data points instead of the mathematical model for the problem. For example, Indian government carries out national census after a gap of 10 years to speculate about the development in population of country. Hence, we have populations in these years as follows:

Years	Population (in crores)
1961	43.9235
1971	54.8160
1981	68.3329
1991	84.6421
2001	102.8737
2011	121.0193

This population data is exact up to four decimal digits. But, in intermediate years such as 1977, 2010, etc., we do not have exact population. The numerical techniques can be used to compute approximate populations in these years.

Except for data points, sometimes, we also require approximating different functions with polynomials due to the simple structure of the polynomials. The polynomials are also easy for analysis like differentiation and integration etc.

This chapter is devoted to various techniques for the polynomial approximations of functions and data points. The chapter also includes the piecewise interpolation.

Chapter 9: Finite Operators introduces various finite operators including finite difference operators (forward, backward and central difference operators) and other operators like average or mean operator, shift operator, and differential operator. The chapter contains the relations between these operators. This chapter also presents construction of finite difference tables and the error propagation in these tables.

These finite difference operators are helpful in constructing solutions of difference equations and also used to construct interpolating polynomials for equally spaced points, as discussed in Chapter 10.

Chapter 10: Interpolation for Equal Intervals and Bivariate Interpolation contains some interpolation methods for equally spaced points. The methods discussed in Chapter 8 are applicable for both unequally as well as equally spaced points. Rather, the interpolating polynomial obtained from any formula is unique, but for equally spaced points, the calculations for interpolation become simpler and hence need more exploration.

We will also discuss the extension of interpolation from one independent variable to two independent variables known as bivariate interpolation.

Chapter 11: Splines, Curve Fitting, and Other Approximating Curves discusses approximations of data set other than interpolation. In interpolation, we fit a polynomial of the degree $\leq n$ to $(n + 1)$ data points. But if the data set is large, say 50 data points, then it is impractical to fit a polynomial of degree 49 to the data set. In this case, other approximation techniques like least squares curve fitting, spline fitting, etc., can be used. In this chapter, we will discuss different approximation techniques which have certain advantages over interpolation in some real time problems.

Curve fitting is to construct an approximate function $f(x)$ (like exponential, polynomial, logistic curve, etc.) for a table of data points.

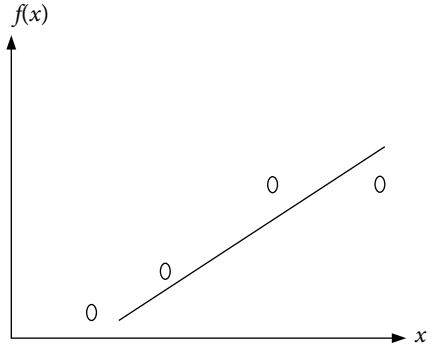


Fig. 9 Straight line fitting

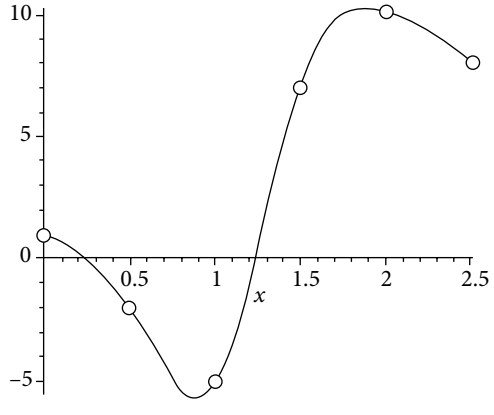


Fig. 10 Cubic spline curve

Interpolating polynomials have global effect, i.e., if we change a point in the data set, then complete polynomial will change. Also if we change the order of data points, the interpolating polynomial remain same, which is not recommended for certain applications like computer graphics and designing, etc. In these cases, we can apply Bězier and B-Spline curves.

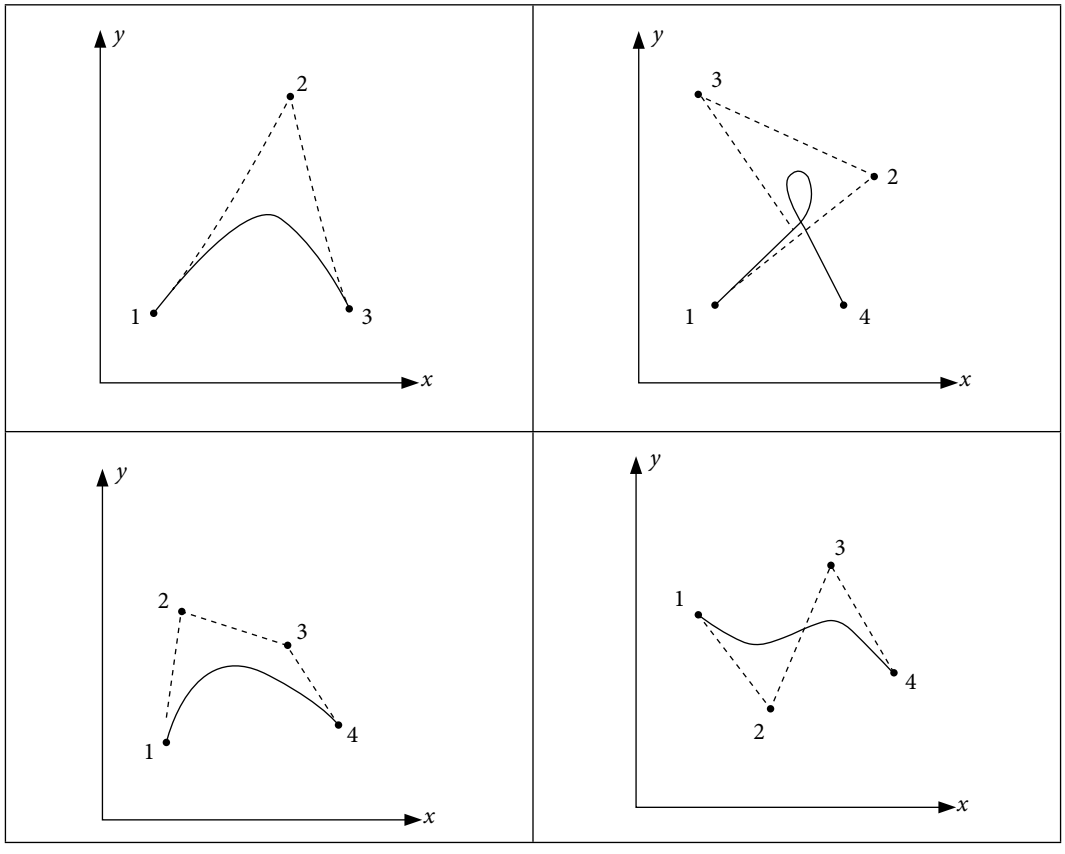


Fig. 11 Bězier curves

In approximations of any polynomial by lower order polynomial, the maximum absolute error can be minimized by Chebyshev polynomials. We can deduce best lower order approximation to a given polynomial by using Chebyshev polynomials.

The polynomial approximations are best approximations for smooth functions and experiments (data set). But if function/experiment behaves in chaos or singular manner (i.e. tends to infinity at some points), then we have to approximate with some other function. One of the functions is a rational function of polynomials, and the approximation is known as Padé approximation.

Chapter 12: Numerical Differentiation is devoted to obtaining numerical differentiation from discrete data points. This chapter elaborates some numerical differentiation techniques based on interpolation.

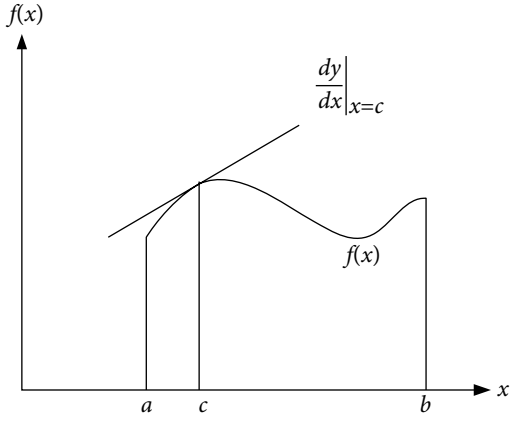


Fig. 12 Differentiation

Chapter 13: Numerical Integration deals with approximating the finite integral of the functions, which are complicated enough to integrate analytically. For example, we don't have exact closed form solutions of integrations like $\int_0^{\pi} \sqrt{1+\cos^2 x}\, dx$, $\int_1^2 \frac{\sin x}{x} dx$, $\int_0^2 e^{-x^2} dx$ etc. In these cases, we can simply apply numerical methods for the approximate solutions. Sometimes we have to find the integration from a set of discrete data points $\{(x_i, y_i),\ i = 0,1,...,n\}$. It is not possible to integrate data points analytically, so it is imperative to approximate these integrations by numerical methods. For example, the value of integral $\int_0^5 y(x)dx$ for the given data set (0, –1), (1, 1), (2, 9), (3, 29), (5, 129) can be obtained only through numerical methods.

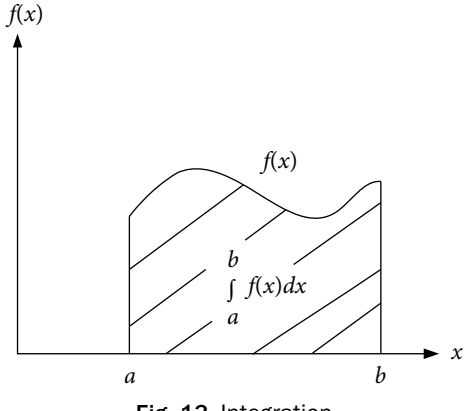


Fig. 13 Integration

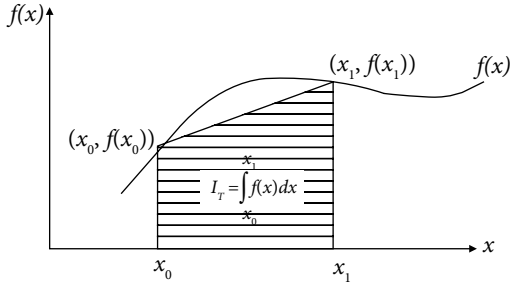


Fig. 14 Numerical Integration

Chapter 14: First Order Ordinary Differential Equations: Initial Value Problems provides a detailed description of standard numerical techniques for the solution of first order ordinary differential equation (ODE) with the initial condition

$$\frac{dy}{dx} = f(x, y),\ y(x_0) = y_0$$

The ODE with initial conditions is known as initial value problem (IVP). Most of the physical laws have a rate of change of quantity rather than the magnitude of the quantity itself; e.g., velocity of any fluid (rate of change of distance), radioactive decay (rate of change of radioactive material), etc. Differential equations govern all these physical phenomena. This chapter contains some basic definitions on differential equations.

The main aim of this chapter in to study numerical methods for the solutions of first order IVP. Differential equations, especially nonlinear, are not easy to solve analytically, as very few analytical methods exist in the literature for a limited class of differential equations. Hence, numerical methods play an important role in the theories of the differential equations.

Consider the following examples

i) $\frac{dy}{dx} = x + y^2,\ y(1) = 2$

ii) $\frac{d^2y}{dx^2} = x\frac{dy}{dx} + \sin y; \quad y(0) = 1,\ y'(0) = 1,$ etc.

These examples are difficult to solve analytically, but we can use numerical techniques for approximate solutions of such ODEs.

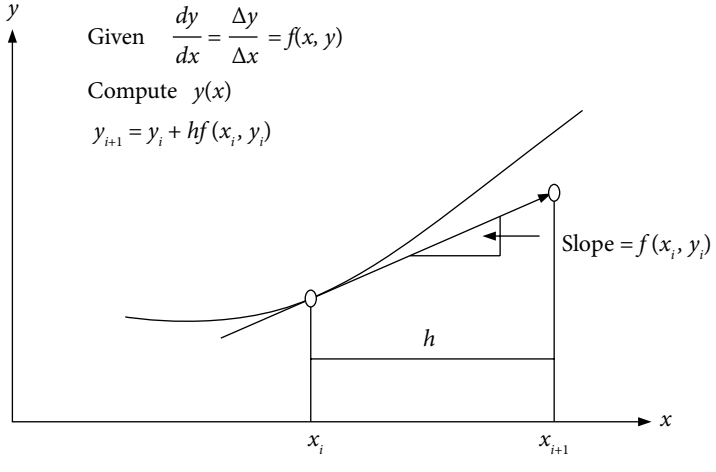


Fig. 15 First order ODE

Chapter 15: Systems of First Order ODEs and Higher Order ODEs: Initial and Boundary Value Problems elucidates the steps involved for finding numerical solutions of a system of first order ODEs and higher order ODEs with initial and boundary conditions, for examples

Systems of First Order ODEs:

i)
$$\frac{dy}{dx} = x + y - z^2$$
$$\frac{dz}{dx} = z - \sin(xy)$$
$$y(0) = 1,\ z(0) = -1$$

ii)
$$\frac{dy}{dx} = w + \sin(x)y - z^2$$
$$\frac{dz}{dx} = z^2 - \sin(xy)$$
$$\frac{dw}{dx} = x + w - 2y$$
$$y(1) = 1,\ z(1) = -1,\ w(1) = 1.3$$

Second and Higher Order Initial Value Problems

i) $\frac{d^2y}{dx^2} + x\frac{dy}{dx} + y = 3; \quad y(0) = 1,\ y'(0) = 2$

ii) $\frac{d^3y}{dx^3} + \sin x \frac{d^2y}{dx^2} + xy = \cos x; \quad y(0) = 1,\ y'(0) = 2,\ y''(0) = 2$

Second and Higher Order Boundary Value Problems

i) $x^2\frac{d^2y}{dx^2} + (x-1)\frac{dy}{dx} + y = 3; \quad y(0) + 2y'(0) = 1,\ y(1) = 3$

ii) $\frac{d^3y}{dx^3} + \sin x \frac{d^2y}{dx^2} + xy = \cos x; \quad y(0) = 1,\ y'(1) = 2,\ y(3) + y''(3) = -4$

In last chapter, we have described various numerical methods for the solutions of the first order ODE $\frac{dy}{dx} = f(x, y);\ y(x_0) = y_0$. In this chapter, we will generalize these methods to find the numerical solutions of system of first order ODEs.

The chapter deals with the conversion of higher order ODEs to the systems of first order ODEs. This chapter also includes the finite difference approximations of derivatives and further solutions of boundary value problems using these finite differences.

Chapter 16: Partial Differential Equations: Finite Difference Methods presents various finite difference methods for the solutions of some standard linear partial differential equations (PDEs). The finite difference method is a simple and most commonly used method to solve PDEs. In this method, we select some node points in the domain of the PDE. Various derivative terms in the PDE and the derivate boundary conditions are replaced by their finite difference approximations at these node points. The PDE is converted to a set of linear algebraic equations at node points. This system of linear algebraic equations can be solved by any direct/iterative procedure discussed in Chapter 5. The solution of this system of linear equations leads to the solution of PDE at node points. An important advantage of this method is that the procedure is algorithmic, and the calculations can be carried out on the computer. So, the solutions can be obtained in a systematic and easy way.

PDEs are of great significance in describing the systems in which the behavior of any physical quantity depends on two or more independent variables. Laplace and Poisson equations (steady-state flow, fluid mechanics, electromagnetic theory and torsion problems), heat conduction equation (temperature distribution) and wave equation (vibrations, fluid dynamics, etc.) are some important examples of second order linear PDEs. Numerical techniques for the solution

of PDEs include finite difference methods (FDMs), finite volume methods (FVMs) and finite element methods (FEMs). This chapter contains only a few finite difference techniques for the solutions of following PDEs governing some important physical phenomena.

Parabolic Equation (Heat Conduction or Diffusion Equation)

$$\frac{\partial u}{\partial t} = c\frac{\partial^2 u}{\partial x^2} \qquad \text{(1-Dimensional heat conduction equation)}$$

$$\frac{\partial u}{\partial t} = c\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = c\nabla^2 u \qquad \text{(2-Dimensional heat conduction equation)}$$

Elliptic Equation (Laplace and Poisson Equations)

$$\nabla^2 u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad \text{(Laplace equation in 2-dimensions)}$$

$$\nabla^2 u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \qquad \text{(Poisson equation in 2-dimensions)}$$

Hyperbolic Equation (Wave Equation)

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} \qquad \text{(1-Dimensional wave equation)}$$

The primary focus is on the preliminary material and the basic concepts of the finite difference techniques used in the book along with their application procedures to derive the numerical solutions of the PDEs.

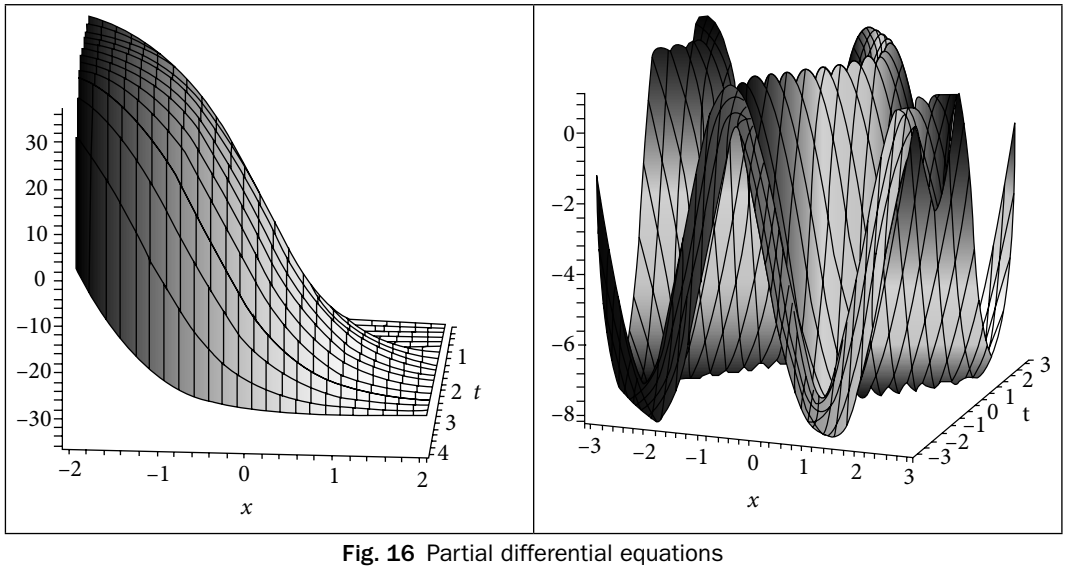


Fig. 16 Partial differential equations

Any Information concerning corrections/errors in this book will be gratefully received.

Rajesh Kumar Gupta
rajeshateli@gmail.com

Acknowledgments

I owe this work to the grace of Almighty, whose divine light provided me strength to complete this book.

It is a great pleasure to thank my mathematics teachers for their expert guidance, support, and encouragement.

I take this opportunity to thank the authorities, my colleagues and students at Thapar University, Patiala, and at the Central University of Punjab, Bathinda, for their support, suggestions and constructive criticism.

I want to thank reviewers and staff at Cambridge University Press who worked to ensure the quality publication of this book.

I am also grateful to my parents, Sh. Murari Lal and Smt. Santosh Devi, brother, Shiv Shanker, sisters, Hemlata and Poonam, brother's wife, Suman, my wife, Usha, and children, Aastha, Akshit, Yashvi, Aadhya, Aaradhya and Reyansh, for providing me a lovely environment in our home.

My friends Dr Harsh and Himani, Dr Amit Kumar, Dr Anoop and Kamal, Dr Amit Bhardwaj, Dr Sunil Singla, Dr Khusneet Jindal, Dr Aklank, Dr Rajendra and Geeta, Dr Phool Singh, Yashpal, Nardeep, Gandhi, Sanjay and Aaditya, PhD scholars and all my well-wishers deserve my heartfelt gratitude for their love and constant support.

Number Systems

Chapter 1

All the mathematical sciences are founded on relations between physical laws and laws of numbers, so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers.

In a few years, all great physical constants will have been approximately estimated, and that the only occupation which will be left to men of science will be to carry these measurements to another place of decimals.

James Clerk Maxwell
(June 13, 1831–November 5, 1879)
He pioneered the classical theory of "Electromagnetism".

1.1 Introduction

In everyday life, we are habituated to doing arithmetic using numbers based on the decimal system. Any number in the decimal number system, say for example, 349.15, can be expressed as a polynomial in the base or radix 10 with integral coefficients 0 to 9.

$$(349.15)_{10} = 3 \times 10^2 + 4 \times 10^1 + 9 \times 10^0 + 1 \times 10^{-1} + 5 \times 10^{-2}$$

In number 349.15, 349 is an integral part and .15 is a fractional part. Note that the subscript (10) in the number $(349.15)_{10}$ denotes the base of the number system.

There is no intrinsic reason to use the decimal number system. Computers read electrical signals, and the state of an electrical impulse is either on or off. Hence, binary system, with base 2 and with integer coefficients 0 and 1, is convenient for computers. However, most computer users prefer to work with the familiar decimal system. It is cumbersome to work with the binary number system, as a large number of binary digits are required to represent even a moderate-sized decimal number. Hence the octal and hexadecimal number systems are also used for this purpose. If the base is two, eight or sixteen, the number is called as the binary, octal or hexadecimal number, respectively. Any number $x = (a_n a_{n-1} \ldots a_1 a_0 . b_1 b_2 \ldots)_\beta$ with base β can be represented as follows

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} \ldots \quad (1.1)$$

The number system with base β contains numbers from 0 to β–1. For examples, decimal number system, with base 10, contains numbers from 0 to 9. Similarly, binary system, with base 2, contains numbers 0 and 1.

Table 1.1 Binary, Octal, Decimal and Hexadecimal Numbers

Binary Base: 2 Digits: 0, 1	Octal Base: 8 Digits: 0, 1, 2, 3, 4, 5, 6, 7	Decimal Base: 10 Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9	Hexadecimal Base: 16 Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F
0000	00	00	0
0001	01	01	1
0010	02	02	2
0011	03	03	3
0100	04	04	4
0101	05	05	5
0110	06	06	6
0111	07	07	7
1000	10	08	8
1001	11	09	9
1010	12	10	A
1011	13	11	B
1100	14	12	C
1101	15	13	D
1110	16	14	E
1111	17	15	F

To work with the computer-preferred binary and the people-preferred decimal, and also with the octal and hexadecimal number systems, it is imperative to have algorithms for conversion from one number system to another number system. In the next two sections, some algorithms are discussed to convert the integral and fractional parts of a number from one number system to another number system.

1.2 Representation of Integers

The arithmetic for various number systems with some examples has been discussed in this section. We will use this for conversion of integers in different number systems.

Example 1.1

Explore the addition and multiplication in the decimal, binary, octal and hexadecimal number systems with some examples.

Decimal Arithmetic (For base 10, digits are 0 … 9)
$(1295)_{10} + (357)_{10} = (1652)_{10}$
$(734)_{10} \times (46)_{10} = (33764)_{10}$

Binary Arithmetic (For base 2, digits are 0 and 1)
$(101011)_2 + (11011)_2 = (1000110)_2$
$(11101)_2 \times (1001)_2 = (100000101)_2$

Octal Arithmetic (For base 8, digits are 0 … 7)
$(1635)_8 + (274)_8 = (2131)_8$
$(752)_8 \times (23)_8 = (22136)_8$

Hexadecimal Arithmetic (For base 16, digits are 0 … 9, A, B, C, D, E, F)
$(5AB7)_{16} + (F63)_{16} = (6A1A)_{16}$
$(A4B)_{16} \times (7A)_{16} = (4E7BE)_{16}$

Note: Arithmetic for numbers with base β:
Consider the addition of two numbers $(1635)_8$ and $(274)_8$ in the octal number system with the base $\beta = 8$. Note that, the addition of numbers 5 and 4 will produce number 9. For $\beta = 8$, we have 1 carry, and the remaining number is 1. Similarly, other calculations give the following result

$$\begin{array}{r} 1\ 1\ 1 \quad \text{Carry} \\ (1\ 6\ 3\ 5)_8 \\ +\ (2\ 7\ 4)_8 \\ \hline (2\ 1\ 3\ 1)_8 \end{array}$$

$\Rightarrow (1635)_8 + (274)_8 = (2131)_8$

Similarly, consider the multiplication of two numbers. For example, multiplication of numbers 7 and 5 will produce number 35. In octal system (base $\beta = 8$), for number 32, we have 4 carry; and remaining is 3. So, final result is $(7)_8 \times (5)_8 = (43)_8$.

1.2.1 Conversion from Any Number System to the Decimal Number System

Conversion from any number system to the decimal form may be obtained directly from the definition (1.1)

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} \ldots$$

Some of the examples are as follows

Example 1.2

$(1101.101)2 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} = (13.625)_{10}$

$(347.623)8 = 3 \times 8^2 + 4 \times 8^1 + 7 \times 8^0 + 6 \times 8^{-1} + 2 \times 8^{-2} + 3 \times 8^{-3} = (231.787109375)_{10}$

$(A5F.B42)16 = 10 \times 16^2 + 5 \times 16^1 + 15 \times 16^0 + 11 \times 16^{-1} + 4 \times 16^{-2} + 2 \times 16^{-3}$
$= (2655.70361328125)_{10}$

1.2.2 Conversion between Binary, Octal and Hexadecimal Number Systems

Conversion in the binary, octal and hexadecimal can be accomplished easily since four/three binary digits make one hexadecimal/octal digit, respectively. To convert from the binary to the octal/hexadecimal, we have to partition the binary digits in groups of three/four (starting from right in an integral part and from left in fractional part) and then replace each group by its equivalent octal/hexadecimal digit. To convert from octal and hexadecimal, we have to replace all digits by their binary equivalents.

Example 1.3

$(1101.101)_2 = (001\ 101.\ 101) = (\underbrace{001}_{1}\ \underbrace{101}_{5}.\ \underbrace{101}_{5}) = (15.5)_8$

$(1101.101)_2 = (1101.\ 1010) = (\underbrace{1101}_{D}.\ \underbrace{1010}_{A}) = (D.A)_{16}$

$(347.623)_8 = (\underbrace{011}_{3}\ \underbrace{100}_{4}\ \underbrace{111}_{7}.\ \underbrace{110}_{6}\ \underbrace{010}_{2}\ \underbrace{011}_{3}) = (11100111.110010011)_2$

$(A5F.B42)_{16} = (\underbrace{1010}_{A}\ \underbrace{0101}_{5}\ \underbrace{1111}_{F}.\ \underbrace{1011}_{B}\ \underbrace{0100}_{4}\ \underbrace{0010}_{2}) = (101001011111.101101000010)_2$

1.2.3 Conversion from Decimal Number System to Any Other Number System

The conversion of the integer N in decimal number system to another number system can be easily obtained in a systematic manner described as follows. Let there be a number N with base β

$$N = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0$$

Division by the base β will give

$$\frac{N}{\beta} = a_n\beta^{n-1} + a_{n-1}\beta^{n-2} + \ldots + a_1 + \frac{a_0}{\beta}$$

The digit a_0 is the remainder after the base β divides the number N. Let us consider the above equation in the form

$$\frac{N}{\beta} = N_0 + \frac{a_0}{\beta}, \text{ where } N_0 = a_n\beta^{n-1} + a_{n-1}\beta^{n-2} + \ldots + a_1$$

On dividing N_0 by base β, we get

$$\frac{N_0}{\beta} = a_n\beta^{n-2} + a_{n-1}\beta^{n-3} + \ldots + \frac{a_1}{\beta}$$

The number a_1 is the remainder. We can continue the process till the quotient is 0. Apparently, the conversion from decimal number system to a number system with base β can be achieved by the following algorithm.

$N = \beta N_0 + a_0$

$N_0 = \beta N_1 + a_1$

$N_1 = \beta N_2 + a_2$

$\vdots$

till the quotient is 0.

Example 1.4

Convert the decimal number $(231)_{10}$ into its binary equivalent.

Ans.

$231 = 115 \times 2 + 1$	$N_0 = 115$	$a_0 = 1$
$115 = 57 \times 2 + 1$	$N_1 = 57$	$a_1 = 1$
$57 = 28 \times 2 + 1$	$N_2 = 28$	$a_2 = 1$
$28 = 14 \times 2 + 0$	$N_3 = 14$	$a_3 = 0$
$14 = 7 \times 2 + 0$	$N_4 = 7$	$a_4 = 0$
$7 = 3 \times 2 + 1$	$N_5 = 3$	$a_5 = 1$
$3 = 1 \times 2 + 1$	$N_6 = 1$	$a_6 = 1$
$1 = 0 \times 2 + 1$	$N_7 = 0$	$a_7 = 1$

Thus the binary representation of the decimal number $(231)_{10}$ is $(11100111)_2$. It can be computed from the expression $\left(a_n a_{n-1} \ldots a_1 a_0\right)_2$.

Example 1.5

Compute the hexadecimal equivalent of the decimal number $(2655)_{10}$.

Ans.

$2655 = 165 \times 16 + 15$ $\quad N_0 = 165$ $\quad a_0 = (15)_{10} = (F)_{16}$

$165 = 10 \times 16 + 5$ $\quad N_1 = 10$ $\quad a_1 = (5)_{10} = (5)_{16}$

$10 = 0 \times 16 + 10$ $\quad N_2 = 0$ $\quad a_2 = (10)_{10} = (A)_{16}$

So, $(A5F)_{16}$ is hexadecimal equivalent of $(2655)_{10}$.

1.2.4 Conversion from One Number System to Any Other Number System

So far, we have discussed the algorithms for conversion of integers in some number systems. The following recursive algorithm can be utilized for conversion of integers in any general number systems.

Algorithm 1.1

Consider a number N with the coefficients $a_n, a_{n-1}, \ldots, a_0$

$$N = a_n \beta^n + a_{n-1} \beta^{n-1} + \ldots + a_1 \beta + a_0 \beta^0$$

Calculate the following numbers $b_n, b_{n-1}, \ldots, b_0$ recursively using

$$b_n = a_n$$

$$b_i = a_i + b_{i+1} \beta, \; i = n-1, \; n-2, \ldots, 0$$

Then $b_0 = N$.

Example 1.6

Convert the binary number $(110111)_2$ into its decimal equivalent.

Ans.

$(110111)_2 = 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$

Since the conversion is from binary to decimal, we will use decimal arithmetic for this conversion. Note that each digit in the following calculation is in decimal number system.

$b_5 = a_5 = 1$

$b_4 = a_4 + b_5\beta = 1 + 1 \times 2 = 3$

$b_3 = 0 + 3 \times 2 = 6$

$b_2 = 1 + 6 \times 2 = 13$

$b_1 = 1 + 13 \times 2 = 27$

$b_0 = 1 + 27 \times 2 = 55$

Example 1.7

Compute the binary equivalent of the decimal number $(231)_{10}$ using recursive algorithm 1.1.

Ans.

$(231)_{10} = 2 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 = (10)_2 \times (1010)_2^{\,2} + (11)_2 \times (1010)_2 + (1)_2$

This conversion uses binary arithmetic as follows

$b_2 = a_2 = (10)_2$

$b_1 = a_1 + b_2\beta = (11)_2 + (10)_2 \times (1010)_2 = (10111)_2$

$b_0 = a_0 + b_1\beta = (1)_2 + (10111)_2 \times (1010)_2 = (11100111)_2$

Example 1.8

Compute the octal equivalent of the decimal number $(231)_{10}$.

Ans.

$(231)_{10} = 2 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 = (2)_8 \times (12)_8^{\,2} + (3)_8 \times (12)_8 + (1)_8$

On using octal arithmetic in the Algorithm 1.1, we have

$b_2 = a_2 = (2)_8$

$b_1 = a_1 + b_2\beta = (3)_8 + (2)_8 \times (12)_8 = (27)_8$

$b_0 = a_0 + b_1\beta = (1)_8 + (27)_8 \times (12)_8 = (1)_8 + (346)_8 = (347)_8$

Example 1.9

Convert the decimal number $(2655)_{10}$ into hexadecimal number.

Ans.

$(2655)_{16} = 2 \times 10^3 + 6 \times 10^2 + 5 \times 10^1 + 5 \times 10^0$

$$= (2)_{16} \times (A)_{16}^3 + (6)_{16} \times (A)_{16}^2 + (5)_{16} \times (A)_{16} + (5)$$

$$b_3 = a_3 = (2)_{16}$$

$$b_2 = a_2 + b_3\beta = (6)_{16} + (2)_{16} \times (A)_{16} = (6)_{16} + (14)_{16} = (1A)_{16}$$

$$b_1 = a_1 + b_2\beta = (5)_{16} + (1A)_{16} \times (A)_{16} = (5)_{16} + (104)_{16} = (109)_{16}$$

$$b_0 = a_0 + b_1\beta = (5)_{16} + (109)_{16} \times (A)_{16} = (5)_{16} + (A5A)_{16} = (A5F)_{16}$$

1.3 Representation of Fractions

In a number system with base β, the fractional part can always be written as follows

$$x_F = \sum_{k=1}^{\infty} b_k \beta^{-k}$$

where b_k is a non-negative integer less than the number β. If $b_k = 0$ for all k greater than a positive integer, then the fractional part is said to be terminating otherwise non-terminating. For example $\frac{1}{4} = 0.25$ is terminating, while $\frac{1}{6} = 0.166666...$ is non-terminating. Conversion of the fractional part from one number system to another number system can be achieved with the help of the following algorithm.

Algorithm 1.2

On multiplying the fraction $x_F = \sum_{k=1}^{\infty} b_k \beta^{-k} = .b_1 b_2 b_3 \ldots$ with base β, we get

$$\beta x_F = \sum_{k=1}^{\infty} b_k \beta^{-k+1} = b_1 + \sum_{k=1}^{\infty} b_{k+1} \beta^{-k}$$

Thus the number b_1 is an integral part of the product βx_F. On repeating the process, we find that b_2 is an integral part of $\beta(\beta x_F)_F$, b_3 is an integral part of $\beta\left(\beta(\beta x_F)_F\right)_F$ and so on. One can easily conclude the following algorithm for a general base β from the procedure above.

Let $c_0 = x_F$

$b_1 = (\beta c_0)_I$, $c_1 = (\beta c_0)_F$

$b_2 = (\beta c_1)_I$, $c_2 = (\beta c_1)_F$

$\vdots$

where subscript *I* denotes an integral part, while subscript *F* denotes the fractional part.

Example 1.10

Calculate the binary equivalent of the decimal number $(.3125)_{10}$ using the recursive algorithm 1.2.

Ans.

Let $c_0 = (.3125)_{10}$

$2(.3125)_{10} = (.6250)_{10}$ $b_1 = 0$ $c_1 = (.6250)_{10}$

$2(.6250)_{10} = (1.250)_{10}$ $b_2 = 1$ $c_2 = (.250)_{10}$

$2(.250)_{10} = (.50)_{10}$ $b_3 = 0$ $c_3 = (.50)_{10}$

$2(.50)_{10} = (1.00)_{10}$ $b_4 = 1$ $c_4 = (0)_{10}$

The binary equivalent of $(.3125)_{10}$ is $(.b_1b_2b_3b_4)_2 = (.0101)_2$. This example has a terminating binary fraction, but not each terminating decimal fraction will give a terminating binary fraction, and this is true for other number systems also.

Example 1.11

Find the binary equivalent of the decimal number $(0.3)_{10}$.

Ans.

Let $c_0 = (.3)_{10}$

$2(.3)_{10} = (.6)_{10}$ $b_1 = 0$ $c_1 = (.6)_{10}$

$2(.6)_{10} = (1.2)_{10}$ $b_2 = 1$ $c_2 = (.2)_{10}$

$2(.2)_{10} = (.4)_{10}$ $b_3 = 0$ $c_3 = (.4)_{10}$

$2(.4)_{10} = (.8)_{10}$ $b_4 = 0$ $c_4 = (.8)_{10}$

$2(.8)_{10} = (1.6)_{10}$ $b_5 = 1$ $c_5 = (.6)_{10}$

$\vdots$

Since the digits are repeating, we can conclude that the binary equivalent of $(.3)_{10}$ is a non-terminating fraction $(.0\ 1001\ 1001\ 1001\ \ldots)_2$ (or) $(.01001)$

Example 1.12

Find the decimal representation of the binary number $(.0101)_2$.

Ans.

Using the algorithm 1.2 and binary arithmetic, we get

$c_0 = (.0101)_2$

$(1010)_2 (.0101)_2 = (11.0010)_2$ $\quad b_1 = (11)_2 = (3)_{10}$ $\quad c_1 = (.0010)_2$

$(1010)_2 (.0010)_2 = (1.010)_2$ $\quad b_2 = (1)_2 = (1)_{10}$ $\quad c_2 = (.010)_2$

$(1010)_2 (.010)_2 = (10.10)_2$ $\quad b_3 = (10)_2 = (2)_{10}$ $\quad c_3 = (.10)_2$

$(1010)_2 (.10)_2 = (101.0)_2$ $\quad b_4 = (101)_2 = (5)_{10}$ $\quad c_4 = (0)_2$

Hence $(.3125)_{10}$ is decimal equivalent of the binary fraction $(.0101)_2$.

Example 1.13

Convert the octal fraction $(.71)_8$ to its equivalent decimal representation.

Ans.

Let $c_0 = (.71)_8$

$(12)_8 (.71)_8 = (10.72)_8$ $\quad b_1 = (10)_8 = (8)_{10}$ $\quad c_1 = (.72)_8$

$(12)_8 (.72)_8 = (11.04)_8$ $\quad b_2 = (11)_8 = (9)_{10}$ $\quad c_2 = (.04)_8$

$(12)_8 (.04)_8 = (0.5)_8$ $\quad b_3 = (0)_8 = (0)_{10}$ $\quad c_3 = (.5)_8$

$(12)_8 (.5)_8 = (6.2)_8$ $\quad b_4 = (6)_8 = (6)_{10}$ $\quad c_4 = (.2)_8$

$(12)_8 (.2)_8 = (2.4)_8$ $\quad b_5 = (2)_8 = (2)_{10}$ $\quad c_5 = (.4)_8$

$(12)_8 (.4)_8 = (5.0)_8$ $\quad b_6 = (5)_8 = (5)_{10}$ $\quad c_6 = (0)_8$

The decimal representation is $(.890625)_{10}$.

Example 1.14

Convert the hexadecimal fraction $(.B4)_{16}$ to its equivalent decimal representation.

Ans.

Let $c_0 = (.B4)_{16}$

$(A)_{16} (.B4)_{16} = (7.08)_{16}$ $\quad b_1 = (7)_{16} = (7)_{10}$ $\quad c_1 = (.08)_{16}$

$(A)_{16} (.08)_{16} = (0.5)_{16}$ $\quad b_2 = (0)_{10}$ $\quad c_2 = (.5)_{16}$

$(A)_{16}(.5)_{16} = (3.2)_{16}$ $b_3 = (3)_{10}$ $c_3 = (.2)_{16}$

$(A)_{16}(.2)_{16} = (1.4)_{16}$ $b_4 = (1)_{10}$ $c_4 = (.4)_{16}$

$(A)_{16}(.4)_{16} = (2.8)_{16}$ $b_5 = (2)_{10}$ $c_5 = (.8)_{16}$

$(A)_{16}(.8)_{16} = (5.0)_{16}$ $b_6 = (5)_{10}$ $c_6 = (0)_{16}$

The decimal representation is $(.703125)_{10}$.

For conversion from one number system to another number system, one can separately convert the integral and fractional part and then combine them. For example, the decimal equivalent of the number $(.B4)_{16}$ is $(.703125)_{10}$ and decimal equivalent of the number $(A5F)_{16}$ is $(2655)_{10}$. Therefore decimal equivalent of the number $(A5F.B4)_{16}$ is $(2655.703125)_{10}$.

Exercise 1

1. Perform the given arithmetic in the following examples, where the subscript in the number represents the base of the number system:

a) $(583)_{10} + (3057)_{10}$ b) $(312)_{10} \times (281)_{10}$

c) $(10110111)_2 + (101011)_2$ d) $(10101)_2 \times (1101)_2$

e) $(6047)_8 + (165)_8$ f) $(536)_8 \times (37)_8$

g) $(3A73)_{16} + (E84)_{16}$ h) $(85D)_{16} \times (23)_{16}$

Ans. a) $(3640)_{10}$ b) $(87672)_{10}$ c) $(11100010)_2$ d) $(100010001)_2$

e) $(6234)_8$ f) $(25142)_8$ g) $(48F7)_{16}$ h) $(124B7)_{16}$

2. Convert the following numbers into their decimal equivalents:

a) $(11011.110)_2$ b) $(67.243)_8$ c) $(2A7.3F)_{16}$

Ans. a) $(27.75)_{10}$ b) $(55.31835938)_{10}$ c) $(679.2460938)_{10}$

3. Find the binary, octal and hexadecimal forms of the following numbers:

a) $(101101.110)_2$ b) $(573.42)_8$ c) $(A05.9A)_{16}$

Ans. a) $[(55.6)_8, (2D.C)_{16}]$ b) $[(101111011.10001)_2, (17B.88)_{16}]$

c) $[(101000000101.10011010)_2, (5005.464)_8]$

4. Compute the binary, octal and hexadecimal equivalents of the decimal number $(5680)_{10}$.

Ans. $(1011000110000)_2, (13060)_8, (1630)_{16}$

5. Use the algorithm 1.1 for the following conversions:

a) $(1101101)_2$ in decimal b) $(5691)_{10}$ in octal

c) $(237)_8$ in decimal d) $(110111)_2$ in hexadecimal

e) $(2AD3)_{16}$ in decimal f) $(4529)_{10}$ in hexadecimal

g) $(438)_{10}$ in binary h) $(110111)_2$ in octal

Ans. a) $(109)_{10}$ b) $(13070)_8$ c) $(159)_{10}$ d) $(37)_{16}$

e) $(10963)_{10}$ f) $(11B1)_{16}$ g) $(110110110)_2$ h) $(67)_8$

6. Obtain the following conversions for the fractional numbers with the aid of recursive algorithm 1.2

a) $(.1101101)_2$ in decimal
b) $(.50)_{10}$ in octal
c) $(.237)_8$ in decimal
d) $(.A3)_{16}$ in decimal
e) $(.45)_{10}$ in hexadecimal
f) $(.325)_{10}$ in binary

Ans. a) $(.8515625000)_{10}$ b) $(.4\overline{0})_8$ c) $(.310\overline{5468}750)_{10}$
d) $(.1367187500)_{10}$ e) $(.7\overline{3})_{16}$ f) $(.0\overline{101001})_2$

7. Obtain the decimal equivalents of the numbers $(A23.4D)_{16}$, $(126.54)_8$, $(10101.11)_2$.

Ans. $(2595.300781)_{10}$, $(86.6875)_{10}$, $(21.750000)_{10}$

8. Compute the binary, octal and hexadecimal equivalents of the decimal number $(238.40)_{10}$.

Ans. $(11101110.0\overline{1100})_2$, $(356.\overline{3146})_8$, $(EE.\overline{6})_{16}$

9. Calculate the decimal equivalent of the octal number $(.647)_8$ with the aid of the recursive algorithm.

Ans. $(.8261718750)_{10}$

Error Analysis

Chapter 2

I claim to be a simple individual liable to err like any other fellow mortal. I own, however, that I have humility enough to confess my errors and to retrace my steps.

Mohandas Karamchand Gandhi (Mahatma Gandhi**)**
(October 2, 1869–January 30, 1948)
He embraced non-violent civil disobedience and led India to independence from British rule.

Numerical methods use arithmetic operations to solve complex mathematical problems. The numerical processes are algorithmic, so these methods can be applied easily with the advent of high-speed computers. In fact, the development of more efficient computers has played a vital role in a veritable explosion in the usage of numerical techniques for engineering and scientific problems. The common characteristic of numerical techniques is that all these involve cumbersome arithmetic operations. During the implementation of the numerical techniques on a computer, we often come across various types of errors. The precisions (number of digits in the representation of a number) of a numerical solution can be diminished by these several possible errors. This chapter deals with various types of errors, and some standard remedies to trace and reduce these errors.

In Section 2.1, measurement of the error will be discussed. Section 2.2 presents the various sources of errors in mathematical modeling of a real world problem. The study of errors during the implementation of numerical methods for the solution of a mathematical model is the primary objective of Section 2.3. The last section is about some interesting discussion on error.

2.1 Absolute, Relative and Percentage Errors

The difference between the exact value and an approximate value of a quantity is called error in the measurement. Its absolute value is called absolute error. Let x be the exact value and $\tilde{x}$ be an approximate value of a given quantity; then the absolute error is given by

Absolute error = $E_a = |x - \tilde{x}|$

Absolute error is not a complete measurement of the error. For example, let absolute error in any quantity be 0.1 m. This information is not complete until we define the quantity for the 0.1 m error. If the 0.1 m error is in 10000 m, then it is small enough to be ignored. But, we cannot neglect 0.1 m error if it is in 1 m. In fact, the error in any quantity depends on the size of that quantity, so relative error and percentage error are the best measurements of the error.

The relative and percentage errors are defined as follows

$$\text{Relative error} = E_r = \left|\frac{x - \tilde{x}}{x}\right|$$

$$\text{Percentage error} = E_p = 100E_r = 100\left|\frac{x - \tilde{x}}{x}\right|$$

Let there exist a number $\varepsilon > 0$, such that $|x - \tilde{x}| \leq \varepsilon$. Then ε is an upper limit of the absolute error and measures the absolute accuracy.

The relative and percentage errors are independent of the units of the quantities used while the absolute error is expressed in terms of these units.

Example 2.1

An approximation to the value of π is given by $\frac{22}{7}$, while its true value in 8 decimal digits is 3.1415926. Calculate the absolute, relative and percentage errors in the approximation.

Ans. Exact value = x = 3.1415926

Approximate value = $\tilde{x} = \frac{22}{7} = 3.1428571$

$$E_a = |x - \tilde{x}| = |-0.0012645| = 0.0012645$$

$$E_r = \left|\frac{x - \tilde{x}}{x}\right| = \frac{0.0012645}{3.1415926} = 0.000402502$$

$$E_p = 100E_r = 100\left|\frac{x - \tilde{x}}{x}\right| = .0402502\%$$

To recognize the major sources of errors and then how to quantify or minimize these errors in the numerical computations are the primary objectives of this chapter.

Analysing any real world problem involves the following three major steps: the first step is to convert the real world problem into a mathematical model; the second step is to solve that model analytically or numerically; and the last step is to analyze the obtained solution for its physical and real-time significance.

After the analysis part is complete, we implement the model for its application.

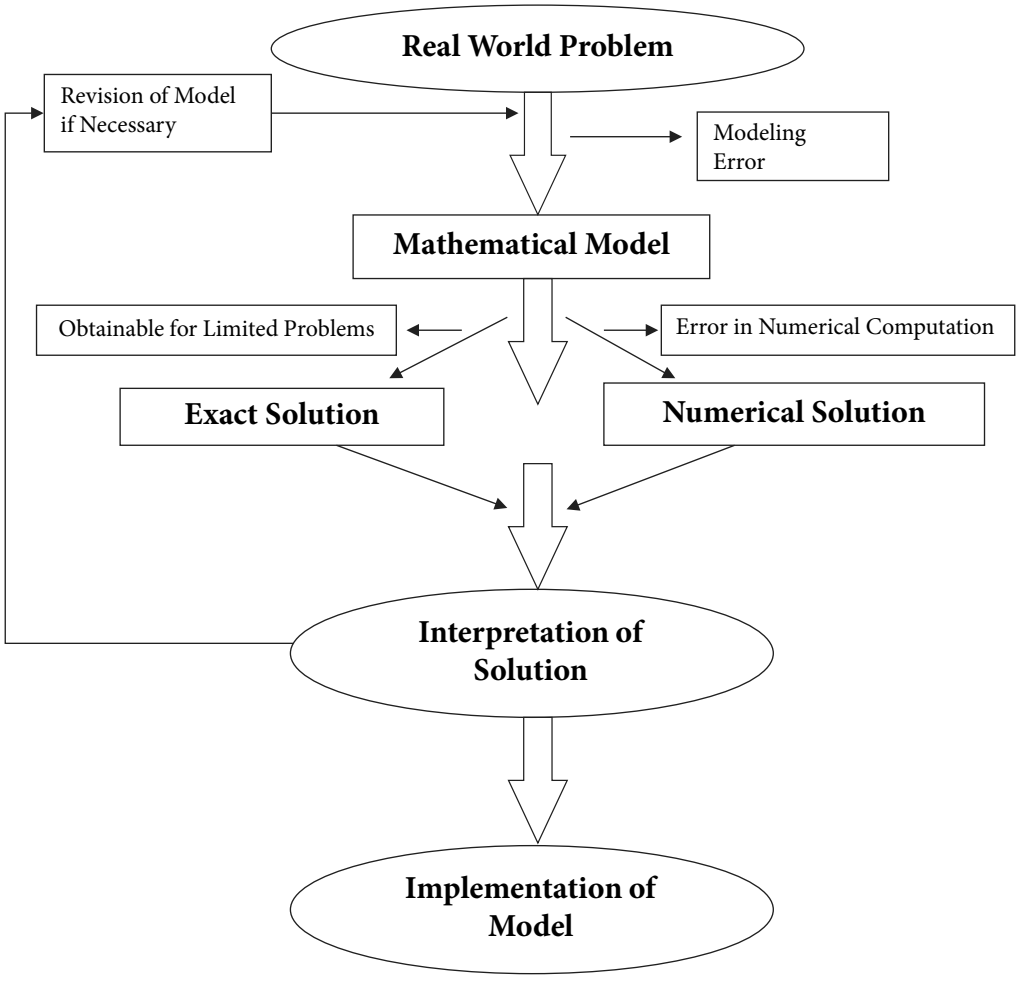


Fig. 2.1 Steps in solving a real world problem

In fact, the error is a multifaceted problem, but it mainly arises during two stages: error occurred during the mathematical modeling of the real world problem, and error when we solve the mathematical model numerically.

In this chapter, we will discuss different types of errors: those encountered during the first step (modeling) and the second step (mathematical model to the solution).

2.2 Errors in Modeling of Real World Problems

The errors in modeling are not directly connected with numerical techniques, but they have a profound impact on the success of a model. Thus, before implementation of a numerical method to the mathematical model, we must have knowledge of the following errors

1. Modeling Error
2. Error in Original Data (Inherent Error)
3. Blunder

2.2.1 Modeling Error

Most of the physical phenomena in nature are inherently nonlinear, and the mathematical models governing these physical systems must be nonlinear. In real world phenomena, it is not possible to include nonlinearity and all other parameters in the mathematical model which govern the situation. For example, while introducing a model for calculating the force acting on the free falling body, it is not always possible to include the air resistance coefficient (drag coefficient) properly. It is not possible to exactly measure the magnitude and direction of the wind force acting on a free-falling body. We simplify the problem by assuming that wind force acting on the body is directly proportional to the velocity of the body. There are many simplifications in each mathematical model, and certainly, these simplifications produce errors in the mathematical model. Further, the model itself is not perfectly governing the situation itself. To check the validity of such models in the real world problem, we may need to perform sensitivity analysis. If our mathematical model is inadequate or inaccurate, then, in that case, obtained results are erroneous.

To reduce such errors, we must ensure that the mathematical model must be formulated and refined by incorporating more features of the situation, which are essential to reduce the error. Simply, the error can be reduced by working with the best model.

2.2.2 Error in Original Data (Inherent Error)

The mathematical model of any physical situation always has associated quantities which are imperfectly known. The reason is that the modeled problem often depends on some instruments whose measurements are of doubtful accuracy. For example, if we want to compute the area of a circular disk, then the radius of the disk is required. But, we cannot measure the radius of the disk with perfect accuracy as very high precision machines can measure up to the accuracy of maximum 5 to 6 decimal digits. Inherent errors can be minimized using high precision computing systems and by taking better data.

2.2.3 Blunder

There is extensive use of the computer in applications of various numerical techniques; chances that the computers make mistakes are very less. But, during the implementation of algorithms on computers, we can make mistakes at various steps, like problem formulations, selection of numerical procedures, programming, and result interpretations, etc. These lead to blunders or gross errors. Some frequent and common types of errors are as follows

i) Inaccurate or inadequate knowledge of the nature of the problem.
ii) Avoiding certain important features of the problem during formulation of the problem.
iii) Some wrong assumptions during the formulation of the problem.
iv) Error in selecting the mathematical equation, which describes a part of the problem.
v) Errors in input data.
vi) Selection of an inappropriate numerical process to determine a solution of the mathematical model.
vii) Implementing a wrong algorithm or avoiding certain important features of a mathematical model in the algorithm.
viii) Starting with a wrong initial guess.
ix) Other simple mistakes like misprints, wrong subscripts in variables, forgetting unit conversion, forgetting negative sign, etc.
x) Implementing infinite series without having knowledge of the convergence.

These errors can be reduced to a large extent by acquiring a hold over various intricacies of the real world phenomena, mathematical modeling of the phenomena and the numerical methods for the solutions of these mathematical models. We must carefully examine the results to avoid such blunders, and a test run with known results is worthwhile in this regard. Test problems more often reveal the mistake and permit its correction.

2.3 Errors in Implementation of Numerical Methods

In this section, we will discuss those errors, which are due to the way that computers store numbers and do arithmetic. In any numerical computation, we come across following types of errors

i) Round-off Error
ii) Overflow and Underflow
iii) Floating Point Arithmetic and Propagated Error
iv) Truncation Error
v) Machine eps (Epsilon)
vi) Epilogue
vii) Loss of Significance: Condition and Stability

There are several potential sources of errors in numerical computation. But, *round-off* and *truncation* errors can occur in any numerical computation.

2.3.1 Round-off Error

During the implementation of a numerical algorithm with computing devices mainly calculator and computer, we have to work with a finite number of digits in representing a number. The number of digits depends on the word length of the computing device and software. The scientific calculations are carried out in floating point arithmetic. It is

necessary to have knowledge of floating point representations of numbers and the basic arithmetic operations performed by the computer (+, -, *, /) in these representations.

Floating Point Representation of Numbers

To understand the major sources of error during the implementation of numerical algorithms, it is necessary to discuss how the computer stores the numbers.

An m-digits floating point number in the base β is of the following form

$$x = \pm \left(.d_1 d_2 d_3 \cdots d_m\right)_\beta \beta^n$$

where $\left(.d_1 d_2 d_3 \cdots d_m\right)_\beta$ is called as a mantissa and the integer n is called the exponent. A non-zero number is said to be normalized if $d_1 \neq 0$.

All the real numbers are stored in normalized form in the computer to avoid wastage of computer memory on storing useless non-significant zeroes. For example, 0.002345 can be represented in a wasteful manner as $(0.002345)10^0$ which is wasting two important decimal points. However, the normalized form is $(0.2345)10^{-2}$, which eliminates these useless zeroes; also known as spurious zeroes.

If we want to enter the number 234.1205, then this number stored in the computer in normalized form, i.e., $(0.2341205)10^3$. Similarly, the number 0.00008671213 stored in the computer in normalized form $(0.8671213)10^{-4}$.

The digits used in mantissa to express a number are called as significant digits or significant figures. More precisely, *digits in the normalized form mantissa of a number are significant digits.*

a) All non-zero digits are significant. For examples, the numbers 3.1416, 4.7894 and 34.211 have five significant digits each.
b) All zeroes between non-zero digits are significant. For examples, the numbers 3.0156 and 7.5608 have five significant digits each.
c) Trailing zeroes following a decimal point are significant. So, the numbers 3.5070 and 76.500 have five significant digits each.
   (Why the number 5.1 has two significant digits, and number 5.10 has three significant digits? To explain this, let us assume we are reading Chapter 5 of a book, and it contains 12 sections. The number 5.1 represents first section of Chapter 5, while the number 5.10 represents tenth section of Chapter 5.)
d) Zeroes between the decimal point and preceding a non-zero digit are not significant. i.e., the numbers 0.0023401 and 0.00023401 have five significant digits each.
e) Trailing zeroes are significant if the decimal point is not present, i.e., the numbers 45067000 and 45000 have eight and five significant digits, respectively.

To compute the significant digits in a number, simply convert the number in the normalized form and then compute the significant digits.

There is a limit on the mantissa (m) and exponent (n) as the storage capacity of any machine is finite. The precision or length m of the floating point numbers usually depends on the word length of the computer and software, and it may vary widely. For example, in

single precision (float variable, 32 bits), the C-programming allows 23 bits for mantissa (m), 8 bits for exponent (n), and 1 bit for sign (±). Similarly, double variable gives 52 bits for mantissa, 11 bits for exponent, and 1 bit for sign. Note that the calculations in double precision require more storage and more running time as compared to single precision.

To understand the limit on storing capacity of the computer, consider the number 10/3. Since the computer can enter the number only in normalized form, hence the computer first solves 10/3 = 3.333333…, and then stores the number. There are infinite numbers of 3's in the expression, but computer will store the number up to its capacity. Let the capacity of the computer be ten digits (i.e., mantissa limit m <= 10), then the number will store as $(0.3333333333)10^1$. All computing devices represent such numbers with some imprecision. For examples, 5/3 =1.6666666…, $\sqrt{2} = 1.414213...$ and $\pi = 3.141592...$ cannot be expressed by a finite number of digits, since the computer cannot store 50/3, $\sqrt{2}$, etc. These numbers may be approximated by rounding off the last precision to m-digits floating point number. For example, let m = 6, then we can approximate 50/3, $\sqrt{2}$ and π by numbers 16.6667, 1.41421 and 3.14159, respectively.

This process of rounding off the numbers during the computation will give rise to round off errors.

Rounding and Chopping

Rounding and chopping are two commonly used ways of converting a given real number x into its m-digits floating point representation $fl(x)$. In the case of chopping, the number x is retained up to m-digits, and remaining digits are simply chopped off. For example, consider 6-digits floating point representation, then

$$x_1 = \frac{2}{3} \qquad fl(x_1) = 0.666666$$

$$x_2 = 3456789 \qquad fl(x_2) = (.345678)10^7$$

$$x_3 = -0.0011223344 \qquad fl(x_3) = -(.112233)10^{-2}$$

In rounding, the normalized floating point number $fl(x)$ is chosen such that it is nearest to the number x. In the case of a tie, some special rules such as symmetric rounding can be used. Rules to round off a number to m significant figures are as follows

i) Discard all digits to the right of m-th digit.
ii) If the last discarded number is

   a) less than half of base β in the $(m + 1)$th place, leave the m-th digit unchanged;
   b) greater than half of base β in the $(m + 1)$th place, increase the m-th digit by unity;
   c) exactly half of base β in the $(m + 1)$th place, increase the m-th digit by unity if it is odd, otherwise leave the m-th digit unchanged. It is known as symmetric rounding around even number. Similarly, we can have symmetric rounding about odd number.

Consider the following numbers with 6-digits floating point representation

$$x_1 = \frac{2}{3} \qquad fl(x_1) = 0.666667$$
$$x_2 = 3456789 \qquad fl(x_2) = (.345679)10^7$$
$$x_3 = -0.0011223344 \qquad fl(x_3) = -(.112233)10^{-2}$$

The difference between x and $fl(x)$ is called the round-off error. If the number is correct up to p decimal points, then the maximum absolute error in chopping and rounding is given by

$$\text{Absolute error} = E_a = |x - fl(x)| \le \begin{cases} \frac{1}{2}\beta^{-p} & \text{in rounding} \\ \beta^{-p} & \text{in chopping} \end{cases}$$

For example, if the number 12.345 (β = 10) is correct up to digits mentioned, then the maximum absolute error in this number is .001 in the case of chopping, and it is .0005 in the case of rounding.

The relative error in the floating point representation of x is as follows

$$\text{Relative Error} = |\delta| = \left|\frac{x - fl(x)}{x}\right|$$

Let the number be correct up to m significant digits in normalized form. Then the maximum relative error is given by the following expression

$$|\delta| \le \begin{cases} \frac{1}{2}\beta^{1-m} & \text{in rounding} \\ \beta^{1-m} & \text{in chopping} \end{cases}$$

For example, let us assume that the number 123.45 (β = 10) is correct up to digits mentioned. It contains five significant digits, so the maximum relative error in this number is .0001 in the case of chopping, and it is .00005 in the case of rounding.

Note: *It is worth mentioning here that generally we use rounding. Until it is not mentioned to use chopping specifically, we will use rounding for computational work throughout the book.*

Example 2.2

Consider the irrational number π = 3.14159265358979... It has an infinite number of digits. So, computer representation of the number π will produce the round-off error depending on the number of significant digits in arithmetic. In Table 2.1, we are presenting the absolute and percentage errors for 1,2,…,6 significant digits, while considering the exact value of π = 3.141593.

Table 2.1

Number of digits	Approximation for π (Rounding)	Absolute error	Percentage error
1	3	0.141593	4.507045%
2	3.1	0.041593	1.323946%
3	3.14	0.001593	0.050707%
4	3.142	0.000407	0.012955%
5	3.1416	0.000007	0.000234%
6	3.14159	0.000003	0.000095%

Example 2.3

Compute the absolute and relative errors in the four significant digits approximations of the numbers 124678 and 345.635.

Ans. Four significant digits approximations of the numbers 124678 and 345.635 are as follows

$$fl(x)=\begin{cases}(.1247)10^{6} & x=124678\\ (.3456)10^{3} & x=345.635\end{cases}\qquad \text{rounding}$$

$$fl(x)=\begin{cases}(.1246)10^{6} & x=124678\\ (.3456)10^{3} & x=345.635\end{cases}\qquad \text{chopping}$$

$$\text{Absolute error} = E_a=|x-fl(x)|=\begin{cases}|124678-124700|=22\\ |345.635-345.6|=.035\end{cases}\qquad \text{rounding}$$

$$E_a=|x-fl(x)|=\begin{cases}|124678-124600|=78\\ |345.635-345.6|=.035\end{cases}\qquad \text{chopping}$$

$$\text{Relative error} = E_r=\left|\frac{x-fl(x)}{x}\right|=\begin{cases}(.1764545)10^{-3}\\ (.1012628)10^{-3}\end{cases}\qquad \text{rounding}$$

$$E_r=\left|\frac{x-fl(x)}{x}\right|=\begin{cases}(.6256116)10^{-3}\\ (.1012628)10^{-3}\end{cases}\qquad \text{chopping}$$

Example 2.4

The true value of π correct to 8-significant digits is 3.1415926. Calculate the absolute and relative error.

Ans. The value is correct up to 7-decimal digits, so the maximum absolute errors in case of rounding and chopping are $\frac{1}{2}10^{-7}$ and 10^{-7}, respectively.

$$\text{Relative error in rounding} = E_r = \left|\frac{X - X'}{X}\right| = \frac{0.5 \times 10^{-7}}{3.1415926} = 0.15915494 \times 10^{-7}$$

$$\text{Relative error in chopping} = E_r = \frac{10^{-7}}{3.1415926} = 0.31830989 \times 10^{-7}$$

It is easy to see that, in between every two numbers, there are infinitely many numbers, which we cannot represent exactly using the computer. Let us consider the machine with 6–digits floating point arithmetic. Consider any two numbers, say 2.12346 and 2.12347. Then, it is easy to see that we cannot represent the in-between numbers like 2.1234652, 2.12346521, 2.1234603112, etc. and these are infinitely many numbers.

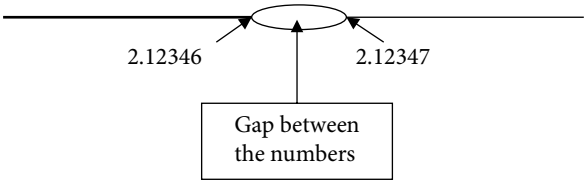


Fig. 2.2 Gaps between floating point numbers

In fact, using computer, we can only represent finite numbers of real numbers and in between every two such numbers, we have infinite numbers, which cannot be represented by the computer.

2.3.2 Overflow and Underflow

The normalized form for an m-digits non-zero floating point number in the base β is given by

$$x = \pm\left(.d_1 d_2 d_3 \cdots d_m\right)_\beta \beta^n, \qquad d_1 \neq 0$$

where $\left(.d_1 d_2 d_3 \cdots d_m\right)_\beta$ is called as mantissa and the integer n is called as exponent.

The exponent n is restricted to a range $l < n < L$, for integers l and L; generally $l = -L$. This limit varies widely and depends on the computational device used. If in the floating point representation of a number x, the exponent n exceeds the limit, i.e., either $|x| \geq \beta^{L}$ (overflow) or $0 \leq |x| \leq \beta^{l-1}$ (underflow), it results either in a stop or else $fl(x)$ is represented by a special number (either 0 or infinity). These special numbers are not subject to the usual rules of arithmetic when combined with ordinary floating point numbers.

Let a hypothetical computer with maximum ten digits mantissa and exponent range (–20, 20) in the decimal number system, then the overflow and underflow can be structured in the following figure

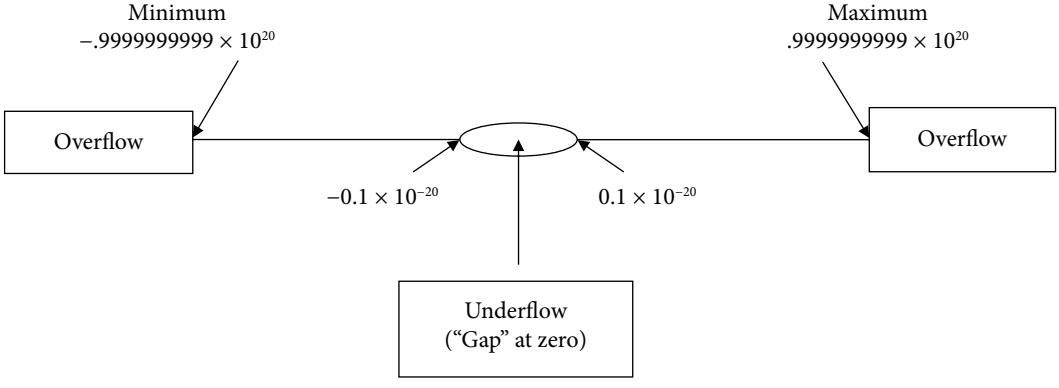


Fig. 2.3 Overflow and underflow

Rather the limit is quite awesome, but it is not able to represent physical quantities like Avogadro's number (6.022×10^{23}) and Plank's constant (6.626×10^{-34} *J.s.*), etc.

2.3.3 Floating Point Arithmetic and Error Propagation

In the last section, we have discussed the errors in number representations. These errors further propagate while performing basic arithmetic operations using a computer. The result of an arithmetic operation is usually not accurate to the same length as the numbers used for the operations. The floating point numbers are first converted into the normalized forms as soon as they enter in the computer.

Here we will explain the arithmetic operations with 6-significant digits numbers. For example, let us take numbers $x = 123.456$ and $y = 12.3456$ with six significant digits. The various arithmetic operations (+, – , *, /) on these two numbers are as follows

$x + y = (.123456)10^3 + (.123456)10^2$ (Normalized form)
$= (.123456)10^3 + (.012346)10^3$ (Equal exponent using symmetric rounding)
$= (.135802)10^3$

$x - y = (.123456)10^3 - (.123456)10^2$
$= (.123456)10^3 - (.012346)10^3$ (Equal exponent using symmetric rounding)
$= (.111110)10^3$

$x * y = (.123456)10^3 * (.123456)10^2$
$= (.123456) * (.123456)\ 10^{3+2}$ (Add the exponents)
$= (.015241)10^5$
$= (.152410)10^4$

$x / y = (.123456)10^3 / (.123456)10^2$
$= (.123456) / (.123456)\ 10^{3-2}$ (Subtract the exponents)
$= (1.00000)10^1$
$= (0.100000)10^2$

Note: If two floating point numbers are added or subtracted, first they are converted into the numbers with equal exponents. The results in various arithmetic operations are not correct up to six significant digits due to rounding errors.

It is worth mentioning here that the result of subtraction of two nearly equal numbers leads to a very serious problem, i.e., loss of significant digits. For example, consider six significant digits numbers $x = 123.456$ and $y = 123.432$, then

$x - y = (.123456)10^3 - (.123432)10^3$ (Normalized form)
$= (.000024)10^3$ (Result containing only two significant digits, four non-significant zeroes are appended)

This subtraction of two nearly equal numbers is called as subtractive cancellation or loss of significance. It is a classical example of computer handling mathematics can create a numerical problem. We will discuss it, in detail, in Section 2.3.7.

2.3.3.1 Propagated Error in Arithmetic Operations

Propagated errors are the errors in the succeeding steps of a process due to an earlier error in the input. For example, error in the division of two numbers due to local errors in the numbers. In this section, we will see how errors in numbers may propagate through basic mathematical operations viz. addition, subtraction, multiplication, and division of two numbers.

Consider any two numbers x_1 and x_2. Let the errors in the numbers x_1 and x_2 be δx_1 and δx_2, respectively. Then errors in the addition, subtraction, multiplication, and division of these two numbers are as follows

i) **Addition**

Let $X = x_1 + x_2$ and the error in X is δX.

$$X + \delta X = x_1 + \delta x_1 + x_2 + \delta x_2$$

$$\delta X = \delta x_1 + \delta x_2$$

$$\text{Absolute Error} = |\delta X| \le |\delta x_1| + |\delta x_2|$$

$$\text{Relative Error} = \frac{|\delta X|}{|X|} \le \frac{|\delta x_1|}{|X|} + \frac{|\delta x_2|}{|X|} \tag{2.1}$$

ii) **Subtraction**

Similarly, the error in subtraction $X = x_1 - x_2$ is given by

$$\delta X = \delta x_1 - \delta x_2$$

$$\text{Absolute Error} = |\delta X| \le |\delta x_1| + |\delta x_2|$$

$$\text{Relative Error} = \frac{|\delta X|}{|X|} \le \frac{|\delta x_1|}{|X|} + \frac{|\delta x_2|}{|X|} \tag{2.2}$$

iii) **Multiplication**

Let $X = x_1 x_2$, then

$$\begin{aligned} X + \delta X &= (x_1 + \delta x_1)(x_2 + \delta x_2) \\ &= x_1 x_2 + x_2 \delta x_1 + x_1 \delta x_2 + \delta x_1 \delta x_2 \end{aligned}$$

Neglecting second order term ($\delta x_1 \delta x_2$), the error in the multiplication of two numbers is as follows

$$\delta X = x_2 \delta x_1 + x_1 \delta x_2$$

$$\text{Absolute Error} = |\delta X| \le |x_2 \delta x_1| + |x_1 \delta x_2|$$

$$\text{Relative Error} = \frac{|\delta X|}{|X|} \le \frac{|\delta x_1|}{|x_1|} + \frac{|\delta x_2|}{|x_2|} \tag{2.3}$$

iv) **Division**

Let $X = \dfrac{x_1}{x_2}$, then

$$X+\delta X=\frac{x_1+\delta x_1}{x_2+\delta x_2}=\left(\frac{x_1+\delta x_1}{x_2+\delta x_2}\right)\left(\frac{x_2-\delta x_2}{x_2-\delta x_2}\right)$$
$$=\frac{x_1x_2+x_2\delta x_1-x_1\delta x_2-\delta x_1\delta x_2}{x_2^2-\delta x_2^2}$$

On neglecting the second order terms ($\delta x_1\delta x_2$ and δx_2^2), the error is given by

$$\delta X=\frac{x_2\delta x_1-x_1\delta x_2}{x_2^2}$$

$$\text{Absolute Error} = |\delta X| \le \frac{|\delta x_1|}{|x_2|}+\frac{|x_1\delta x_2|}{x_2^2}$$

$$\text{Relative error} = \frac{|\delta X|}{|X|} \le \frac{|\delta x_1|}{|x_1|}+\frac{|\delta x_2|}{|x_2|} \qquad (2.4)$$

Example 2.5

The numbers $x_1 = 0.123$ and $x_2 = 12.37$ are correct up to the significant digits in the numbers. Compute the relative errors in the addition, subtraction, multiplication and division of these two numbers. Consider symmetric rounding.

Ans. Absolute errors in the numbers $x_1 = 0.123$ and $x_2 = 12.37$ are $\delta x_1 = .0005$ and $\delta x_2 = .005$, respectively.
Using the formulae (2.1 – 2.4) for various error terms, we have

$$\text{Relative error in the addition} = \left|\frac{\delta X}{X}\right| \le \left|\frac{\delta x_1}{X}\right|+\left|\frac{\delta x_2}{X}\right|, \text{ where } X=x_1+x_2$$

$$=\frac{.0005}{12.493}+\frac{.005}{12.493}=.000440246538$$

$$\text{Relative error in subtraction} = \left|\frac{\delta X}{X}\right| \le \left|\frac{\delta x_1}{X}\right|+\left|\frac{\delta x_2}{X}\right|, \text{ where } X=x_1-x_2$$

$$=\frac{.0005}{12.247}+\frac{.005}{12.247}=.000449089573$$

$$\text{Relative error in multiplication and division} = \left|\frac{\delta x_1}{x_1}\right|+\left|\frac{\delta x_2}{x_2}\right|$$

$$=\frac{.0005}{.123}+\frac{.005}{12.37}=.004469244369$$

Example 2.6

Calculate the absolute and relative errors in the expression $a+\frac{5b}{c}-3bc$, if the measurements of $a = 3.5435$, $b = .2588$ and $c = 1.0150$ are possibly correct up to four decimal points.

Ans. Let $x = a+\frac{5b}{c}-3bc = A+5B-3C$, where $A = a$, $B = \frac{b}{c}$ and $C = bc$.

Value of $x = a+\frac{5b}{c}-3bc = 4.03033$

Error in a, b and c is $\delta a = \delta b = \delta c = .00005$

Absolute error in $A = |\delta A| = .00005$

Absolute error in $B = |\delta B| = \frac{|c\delta b|+|b\delta c|}{c^2} = \frac{(1.015+0.2588)\times .00005}{(1.015)^2} = .00006182$

Absolute error in $C = |\delta C| = |c\delta b|+|b\delta c| = (1.015+0.2588)\times .00005 = .00006369$

Absolute error in $x = |\delta x| \le |\delta A|+5|\delta B|+3|\delta C|$

$= .00005+5(.00006182)+3(.00006369)$

$= .0005502$

Relative error in $x = \left|\frac{\delta x}{x}\right| = \frac{.0005502}{4.03033} = .0001365$

Percentage error in $x = 0.01365\%$

2.3.3.2 Error Propagation in Function of Single Variable

Let us consider a function $f(x)$ of a single variable, x. Assume that the variable x has some error and its approximate value is $\tilde{x}$. The effect of error in the value of x on the value of function $f(x)$ is given by

$$\Delta f(x) = |f(x)-f(\tilde{x})|$$

Evaluating $\Delta f(x)$ is difficult as the exact value of x is unknown and hence exact $f(x)$ is unknown. But if $\tilde{x}$ is close to x and the function $f(x)$ is infinitely differentiable in some interval containing the points $\tilde{x}$ and x, then Taylor series can be employed as follows

$$f(x) = f(\tilde{x})+(x-\tilde{x})f'(\tilde{x})+\frac{(x-\tilde{x})^2}{2!}f''(\tilde{x})+\ldots$$

Since the difference $(x-\tilde{x})$ is very small, hence neglecting the second and higher order terms of $(x-\tilde{x})$ will give following relation

$$f(x)-f(\tilde{x}) \simeq (x-\tilde{x}) f'(\tilde{x})$$

or
$$\begin{aligned} |\Delta f(x)| &\simeq |x-\tilde{x}||f'(\tilde{x})| \\ &\simeq \Delta x |f'(\tilde{x})| \end{aligned} \tag{2.5}$$

where $\Delta f(x) = |f(x)-f(\tilde{x})|$ represents the estimated error in the function value and $\Delta x = |x-\tilde{x}|$ is the estimated error of *x*.

Example 2.7

Let $\tilde{x} = 3.42$ be an approximate value of the variable *x* with an error bound $\Delta x = 0.003$. Compute the resulting error bound in the function value $f(x) = x^3$.

Ans. From Eq. (2.5), the resulting error in the function *f*(*x*) is given by

$$\Delta f(x) = (0.003)3(3.42)^2 = 0.1052676$$

Note that the approximate function value is $f(3.42) = 40.001688$. Therefore, the predicted value of *f*(*x*) must be in the range

$$f(3.42) = 40.001688 \pm 0.1052676$$

Equivalent Statement for Example 2.7: Let us assume that we want to compute the volume of a cube. We measure its length with a machine and find out that it is 3.42m. Let us also assume that the machine can measure with maximum error 0.003m. Find the volume of the cube.

2.3.3.3 Error Propagation in Function of More than One Variable

General Error Formula

The approach above can be generalized to the function of more than one independent variable. Let $y = f(x_1, x_2, ..., x_n)$ be a function of *n*-independent variables $x_1, x_2, ..., x_n$. Let $\delta x_1, \delta x_2, ..., \delta x_n$ be the errors in calculating the variables $x_1, x_2, ..., x_n$, respectively. Let error in *y* be δy, i.e.,

$$y + \delta y = f(x_1 + \delta x_1, x_2 + \delta x_2, ..., x_n + \delta x_n)$$

When the required partial derivatives exist, then Taylor's series expansion is given by

$$y+\delta y=f(x_1,x_2,...,x_n)+\left(\frac{\partial f}{\partial x_1}\delta x_1+\frac{\partial f}{\partial x_2}\delta x_2+...+\frac{\partial f}{\partial x_n}\delta x_n\right)+$$

$$+\text{ terms involving second and higher powers of } \delta x_1,\delta x_2,...,\delta x_n \qquad (2.6)$$

The errors in the numbers $x_1,x_2,...,x_n$ are small enough to neglect the second and higher degree terms of the numbers $\delta x_1,\delta x_2,...,\delta x_n$. We can obtain the following result from Eq. (2.6)

$$\delta y\approx\frac{\partial f}{\partial x_1}\delta x_1+\frac{\partial f}{\partial x_2}\delta x_2+...+\frac{\partial f}{\partial x_n}\delta x_n \qquad (2.7)$$

Equation (2.7) is known as the general error formula. Since the error term may be of any sign, (+)ve or (–)ve, we can take absolute values of the terms in the expression.

$$|\delta y|\approx\left|\frac{\partial f}{\partial x_1}\right||\delta x_1|+\left|\frac{\partial f}{\partial x_2}\right||\delta x_2|+...+\left|\frac{\partial f}{\partial x_n}\right||\delta x_n|$$

Example 2.8

Compute the absolute and relative errors in the function $f(x,y,z)=\frac{y^2\sin(x)}{4z^3}$ at $x = 1$ and $y = z = 5$, if the errors in the values of x, y and z are 0.05.

Ans. On using general error formula (2.7), the error $\delta f(x,y,z)$ in $f(x,y,z)$ is given by

$$\delta f(x,y,z)=\frac{\delta f}{\delta x}\delta x+\frac{\delta f}{\delta y}\delta y+\frac{\delta f}{\delta z}\delta z$$

$$=\frac{y^2\cos(x)}{4z^3}\delta x+\frac{y\sin(x)}{2z^3}\delta y-\frac{3y^2\sin(x)}{4z^4}\delta z$$

$$\text{Absolute error}=\left|\frac{y^2\cos(x)}{4z^3}\delta x\right|+\left|\frac{y\sin(x)}{2z^3}\delta y\right|+\left|\frac{3y^2\sin(x)}{4z^4}\delta z\right|$$

$$=.001350756+.000841471+.001262206$$

$$=.003454433$$

$$\text{Relative error}=\frac{\delta f(x,y,z)}{f(x,y,z)}=\frac{.003454433}{.04207355}=.082104624$$

Example 2.9

The radius r and height h of a right circular cylinder are measured as .25 m and 2.4 m, respectively, with a maximum error of 5%. Compute the resulting percentage error in the volume of the cylinder. Assume the value of π is exact for calculation.

Ans. The value of π is exact for calculation, so the volume $V = \pi r^2 h$ is dependent only on radius r and height h of the cylinder i.e., $V = V(r,h)$. Therefore, the error $\delta V(r,h)$ in the volume is given by

$$\delta V(r,h) = \frac{\partial V}{\partial r}\delta r + \frac{\partial V}{\partial h}\delta h = \left(2\pi rh\right)\delta r + \left(\pi r^2\right)\delta h$$

The radius r and height h of the cylinder are measured with a maximum error of 5% i.e.

$$\frac{\delta r}{r} = \frac{\delta h}{h} = 0.05$$

The relative error in volume $V(r,h)$ is given by

$$\begin{aligned} R.E. &= \frac{\delta V(r,h)}{V} \\ &= \frac{1}{\pi r^2 h}\left(\left(2\pi rh\right)\delta r + \left(\pi r^2\right)\delta h\right) \\ &= 2\frac{\delta r}{r} + \frac{\delta h}{h} \\ &= 2(0.05) + 0.05 = 0.15 \end{aligned}$$

Percentage error in the volume of cylinder = *R.E.* × 100 = 15%

2.3.4 Truncation Error

An infinite power series (generally Taylor series) represents the local behavior of a given function $f(x)$ near a given point $x = a$. Approximation of an infinite power series with its finite number of terms, while neglecting remaining terms, leads to the *truncation error*. If we approximate the power series by the n-th order polynomial, then truncation error is of order $n + 1$.

Taylor series for the function $f(x)$ at the point $x = a$ is given by

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{(n)!}f^{(n)}(a) + \cdots$$

(Or) $$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{(n)!}f^{(n)}(a) + R_n(x)$$

where $R_n(x) = \frac{(x-a)^{n+1}}{(n+1)!}f^{(n+1)}(\xi)$ for some ξ between a and x.

On replacing $x = a + h$, we get following form of the Taylor series

$$f(a+h) = f(a) + (h)f'(a) + \frac{(h)^2}{2!}f''(a) + \cdots\cdots + \frac{(h)^n}{(n)!}f^{(n)}(a) + R_n(x)$$

where $R_n(x) = \frac{(h)^{n+1}}{(n+1)!}f^{(n+1)}(\xi);\ a < \xi < a + h.$

For a convergent series, $R_n(x) \to 0$ as $n \to \infty$. Since it is not possible to compute an infinite number of terms, we approximate the function $f(x)$ by first n-terms, and neglecting higher order terms. Then the error is given by remainder term $R_n(x)$. The exact value of ξ is not known, so the value of ξ is such that the error term considered is maximum.

Example 2.10

Use the following Taylor series expansion to compute the value of irrational number *e*.

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Create a table for absolute and percentage errors with numbers of terms n = 1, 2, … 6 of Taylor series approximations. For the exact value of *e*, use e = 2.718282.

Ans. Computation of exact value of e^x requires an infinitely long series. Approximating e^x with the Taylor series to n terms gives an inexact answer. Table 2.2 contains Taylor series approximations of *e* of order n = 1, 2, …6. It also contains absolute and percentage errors in these approximations

Table 2.2

Number of terms	Taylor Series of e^x	Approximation for the function e	Absolute error	Percentage error
1	$e^x = 1$	1	1.718282	63.212058%
2	$e^x = 1 + x$	2	0.718282	26.424116%
3	$e^x = 1 + x + \frac{x^2}{2!}$	2.500000	0.218282	8.030146%
4	$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$	2.666667	0.051615	1.898810%
5	$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$	2.708333	0.009948	0.365966%
6	$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}$	2.716666	0.001616	0.059449%

Example 2.11

Calculate the number of terms required in Taylor series approximation of $\sin(x)$ to compute the value of $\sin\left(\frac{\pi}{12}\right)$ correct up to 4-decimal places.

Ans. Using Taylor series of $\sin(x)$ at point $x = 0$, we have

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{(x)^{2n-1}}{(2n-1)!}(-1)^{n-1} + R_{2n-1}(x)$$

If we retain only first $2n-1$ terms in this expression, then the error term is given by

$$R_{2n-1}(x) = \frac{(x)^{2n}}{(2n)!} f^{(2n)}(\xi); \quad 0 < \xi < x. \text{ at } x = \frac{\pi}{12} = 0.2618$$

The maximum value of $f^{(2n)}(\xi)$ is 1. The error term must be less than .00005 for 4-decimal points accuracy

$$R_{2n-1}(x) = \frac{(0.2618)^{2n}}{(2n)!} \leq .00005$$

$$\Rightarrow 2n \geq 5$$

Hence, 4-decimal points accuracy can be achieved by computing more than five terms of Taylor series.

Example 2.12

The Gauss error function $erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is used widely in probability theory (e.g., normal distribution), statistics and partial differential equations. But the exact integral is not available for a finite value of x, so we use approximations. For example, one way is to use Taylor polynomial for the function e^{-t^2} and compute the resulting integration.

Compute the approximate value of the error function $erf(0.1)=\frac{2}{\sqrt{\pi}}\int_0^{0.1} e^{-t^2}dt$ by using first four terms of the Taylor series.

Ans. Taylor series of e^{-t^2} at $t = 0$ is given by

$$e^{-t^2}=1-t^2+\frac{t^4}{2!}-\frac{t^6}{3!}+\cdots$$

Using the Taylor polynomial of first four terms, we have

$$\begin{aligned} erf(x) &= \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \\ &= \frac{2}{\sqrt{\pi}}\int_0^x\left(1-t^2+\frac{t^4}{2!}-\frac{t^6}{3!}\right)dt \\ &= \frac{2}{\sqrt{\pi}}\left(x-\frac{x^3}{3}+\frac{x^5}{10}-\frac{x^7}{42}\right) \end{aligned}$$

At $x = 0.1$, we have

$$erf(0.1)=\frac{2}{\sqrt{\pi}}\int_0^{0.1} e^{-t^2}dt=0.112463$$

2.3.5 Machine eps (Epsilon)

Machine epsilon for a given machine, for example a computer, is defined as the smallest positive number which, when added to 1, gives a number different from 1. In fact, the machine epsilon defines the lowest floating point number, which can take part in the arithmetic for a given machine. Machine epsilon depends on round-off of the floating point numbers. Since rounding is machine dependent, so machine epsilon also varies with the machine. Machine epsilon characterizes computer arithmetic in numerical analysis. The quantity is also called as macheps or unit round-off, and it has the symbol epsilon ε.

2.3.6 Epilogue

In following chapters, we will see that several alternative numerical methods are available for the solution of any problem. In the selection of any method, we have to keep in mind all aspects of the problems and the method itself. Only from experience can we develop the skill for right selection and this skill has a prominent role in effective implementation of the method. Following are the deciding factors for selection of a numerical method and its implementation to the problem.

1. Type of mathematical problem
2. Computer available
3. Development cost
4. Characteristics of the numerical method
5. Mathematical behavior of the problem
6. Ease of application
7. Maintenance

2.3.7 Loss of Significance: Condition and Stability

In this section, we will study the two related concepts of condition and stability for function and process, respectively. The condition is used to describe the sensitivity of the function and stability is used to describe the sensitivity of the process.

Condition:
The sensitivity of the function $f(x)$ with the change in the argument x is described by the condition number (CN). It is a relative change in the function $f(x)$ for per unit relative change in x. CN of the function $f(x)$ at any point x is given by

$$\text{CN} = \frac{\left|\dfrac{f(x)-f(\tilde{x})}{f(x)}\right|}{\left|\dfrac{x-\tilde{x}}{x}\right|} = \left|\frac{f(x)-f(\tilde{x})}{x-\tilde{x}}\right|\left|\frac{x}{f(x)}\right|$$

For small change in x, Lagrange mean value theorem gives

$$\frac{f(x)-f(\tilde{x})}{x-\tilde{x}} \approx f'(x)$$

So, CN is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| \tag{2.8}$$

If CN ≤1, then the function $f(x)$ is said to be well-conditioned. Otherwise, it is said to be ill-conditioned. The function with large CN is more ill-conditioned as compared to the function with small CN.

Note: *Let us consider a mathematical model of any system, in which variable x gives input, and output is the function f(x). If a small relative change in x (input) produces a large relative change in output f(x), then the system is said to be a sensitive system as fluctuation in input may break the system. Mathematically, if CN is large, then the function is more sensitive to changes and function is ill-conditioned.*

Example 2.13

Find the CNs of the functions $f(x)=\sqrt{x}$ and x^3.

Ans. Using Eq. (2.8), we have

CN of the function $\sqrt{x} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(\frac{1}{2}x^{\frac{-1}{2}}\right)}{\sqrt{x}}\right| = \frac{1}{2}$

CN of the function $x^3 = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(3x^2\right)}{x^3}\right| = 3$

CN of the function $\sqrt{x}$ is less than 1, so the function $\sqrt{x}$ is well conditioned. The function x^3 is an ill-conditioned function as CN > 1.

Example 2.14

Check the condition of the function $f(x)=\frac{1}{1-2x+x^2}$ at $x = 1.01$.

Ans.

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

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right|_{x=1.01} = \left|\frac{x\left(\frac{-2}{(1-x)^3}\right)}{\left(\frac{1}{(1-x)^2}\right)}\right|_{x=1.01} = 202$$

The function $f(x)=\frac{1}{1-2x+x^2}$ at $x = 1.01$ is highly ill-conditioned function. The function has a singular point $x = 1$, so near this point, there are sharp changes in the function value, which make the function highly ill-conditioned.

Example **2.15**

Find the CN of the function $f(x)=\sqrt{x+1}-\sqrt{x}$ at point $x = 11111$.

Ans.

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(\frac{1}{2\sqrt{x+1}}-\frac{1}{2\sqrt{x}}\right)}{\sqrt{x+1}-\sqrt{x}}\right|_{x=11111} \approx \frac{1}{2}$$

Example **2.16**

Compute the function $f(x)=\sqrt{x+1}-\sqrt{x}=\frac{1}{\sqrt{x+1}+\sqrt{x}}$ by using both the formulae at point $x = 11111$. Use six significant digits floating point rounding arithmetic.

Ans. We have two formulas $f(x)=\sqrt{x+1}-\sqrt{x}$ and $f(x)=\frac{1}{\sqrt{x+1}+\sqrt{x}}$ to compute the function $f(x)$ at point $x = 11111$. We will use both the formulas with six significant digits arithmetic, and see that both the processes will produce different results for the same function.

Process-I: $f(x)=\sqrt{x+1}-\sqrt{x}$

$$\begin{aligned} f(x) &= \sqrt{11112}-\sqrt{11111} \\ &= 105.413-105.409 \\ &= .004 \end{aligned}$$

Process-II: $f(x)=\frac{1}{\sqrt{x+1}+\sqrt{x}}$

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{11112}+\sqrt{11111}} \\ &= \frac{1}{105.413+105.409} \\ &= \frac{1}{210.822} \\ &= 0.00474334 \end{aligned}$$

Note that, the exact result up to 6 significant digits is .00474330.

Note: Here, it is candidly seen that if we compute the function $f(x)=\sqrt{x+1}-\sqrt{x}$ directly, then it is error-prone. This is due to the fact that if we subtract two approximately equal numbers, then there is a loss of significant digits. For example in Process-I, when we subtract 105.413 and 105.409, then these two numbers are correct up to six significant digits, but the result .004 contains only one significant digit. Since there is a loss of five significant digits, so the result obtained is highly erroneous. This step can be avoided by rationalizing the function $f(x)$. The result obtained in Process-II after rationalization is correct up to five significant digits.

Stability of the Process:

It is clear from Example 2.16 that computation of the same function from two different processes can produce different results. There are following two major phases for computation of the function value $f(x)$:

i) First phase is to check the condition of the function by computing the CN of the function.

ii) Second phase is to check the stability of the process involved in the computation of the function. The stability of process can be checked by calculating the condition of each step in the process.

The function $f(x)=1/(1-x^2)$ is ill-conditioned (CN $\gg 1$) near $x=\pm 1$. If the function is ill-conditioned then whatever process we will use, it tends to error. So every process will produce an error in computation of the function value $f(x)=1/(1-x^2)$ near $x=\pm 1$.

The function $f(x)=\sqrt{x+1}-\sqrt{x}$ at x = 11111 is well conditioned (CN ≈ 1/2**,** Example 2.15). If the function is well conditioned, then we have to compute the function value by the stable process. If even a single step of the process is ill-conditioned, then the whole process is an unstable process, and we have to switch over to any other alternate stable process.

Example 2.17

Discuss the stability of the Processes-I and II in Example 2.16. Hence, validate the results that the Processes-I yields erroneous result and Process-II produces a more accurate result for the same function $f(x)$.

Ans.

We will calculate the CN of each step involved in both the Processes-I and II.

Process-I: $f(x)=\sqrt{x+1}-\sqrt{x}$

$$\begin{aligned} f(x) &= \sqrt{11112}-\sqrt{11111} \\ &= 105.413-105.409 \\ &= .004 \end{aligned}$$

Various computational steps in the process are as follows

$x_1 = 11111$ ($f(x)$ = Constant, CN = 0)

$x_2 = x_1 + 1 = 11112$ ($f(x) = x + 1$, CN = 1)

$x_3 = \sqrt{x_2} = 105.413$ ($f(x) = \sqrt{x}$, CN = ½)

$x_4 = \sqrt{x_1} = 105.409$ ($f(x) = \sqrt{x}$, CN = ½)

$x_5 = x_4 - x_3 = .004$ ($f(x) = x - x_3$ and $f(x) = x_4 - x$, CN $\simeq$ 26352)

In the last step $x_5 = x_4 - x_3$, we can assume the function $f(x)$ of variable x_3 or x_4. Let $f(x) = x_4 - x$, so condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x(-1)}{x_4 - x}\right| = \left|\frac{105.409}{.004}\right| \approx 26352$$

This step is not a stable step as CN is very large. So the whole process is an unstable process due to this step. That's why the result obtained from this process is highly erroneous.

Process-II: $f(x) = \dfrac{1}{\sqrt{x+1} + \sqrt{x}}$

We will check the conditions of each step in Process-II, and conclude that each step in this process is well conditioned.

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{11112} + \sqrt{11111}} \\ &= \frac{1}{105.413 + 105.409} \\ &= \frac{1}{210.822} \\ &= 0.00474334 \end{aligned}$$

Various steps involved in this process are as follows

$x_1 = 11111$

$x_2 = x_1 + 1 = 11112$

$x_3 = \sqrt{x_2} = 105.413$

$x_4 = \sqrt{x_1} = 105.409$

$x_5 = x_4 + x_3 = 210.822$

$x_6 = \dfrac{1}{x_5} = 0.00474334$

The first four steps in the process are well conditioned as discussed in Process-I. For the fifth step, let $f(x) = x_4 + x$. The condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x(1)}{x_4 + x}\right| = \left|\frac{105.409}{222.822}\right| \approx \frac{1}{2}$$

The last step is $f(x) = \frac{1}{x}$, and the condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(\frac{-1}{x^2}\right)}{\frac{1}{x}}\right| = 1$$

From above discussion, it is clear that all the steps in Process-II are well conditioned, and hence this process is a stable process. Since the process is stable, so the result obtained is accurate to five significant digits.

Note: *Even a single step in the process can make the whole process unstable. So we have to be extra careful during a large process, and must avoid the steps (if possible) with the loss of significant digits. We can use any alternate approach like rationalization, Taylor series expansion, etc. to avoid loss of significant digits.*

Example 2.18

Discuss the stability of the function $f(x) = 1 - \cos(x)$, when x is nearly equal to zero. Find a stable way to compute the function $f(x)$.

Ans. If we directly compute the function $f(x) = 1 - \cos(x)$ at $x \approx 0$, then it will lead to subtraction of two nearly equal numbers and produce loss of significance. To avoid this loss, we can use any of the following three alternates

i) $f(x) = 1 - \cos(x)$

$$= 1 - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right)$$

$$= \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} - \cdots$$

ii) $f(x)=1-\cos(x)$

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

$$=\frac{2\sin^2(x)}{1+\cos(x)}$$

iii) $f(x)=1-\cos(x)$

$$=2\sin^2\frac{x}{2}$$

Example 2.19

Calculate the roots of the equation $x^2+123x+0.5=0$ using five digits floating point chopping arithmetic.

Ans. The roots of the quadratic equation $ax^2+bx+c=0$ are given by

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

The roots of the equation $x^2+123x+0.5=0$ using five digits floating point chopping arithmetic are given by

Root 1. $x_1=\dfrac{-b+\sqrt{b^2-4ac}}{2a}$

$b^2=15129$

$b^2-4ac=15127$

$\sqrt{b^2-4ac}=122.99$

$x_1=\dfrac{-123+122.99}{2}=-0.0005$

Root 2. $x_2=\dfrac{-b-\sqrt{b^2-4ac}}{2a}=\dfrac{-123-122.99}{2}=-122.995=-122.99$

The roots of the equation correct up to some significant digits are $x_1=-0.004065175$ and $x_2=-122.995934825$. The root $x_2=-122.99$ is correctly calculated up to five significant

digits. But the root $x_1 = -0.0005$ is not correct even up to one significant digit. This error is due to loss of significant digits which occurs due to subtraction of two nearly equal numbers (123 and 122.99).

To avoid the loss of significant digits, we will rationalize the formula for x_1, and then compute the root.

$$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} \times \frac{b+\sqrt{b^2-4ac}}{b+\sqrt{b^2-4ac}}$$

$$= \frac{-2c}{b+\sqrt{b^2-4ac}}$$

$$= \frac{-1}{123+122.99} = -.0040652$$

This value is correct up to five significant digits.

Note: There are two ways to produce the results with desired accuracy. One way is to use stable processes for the computation and another way is to use the computing device with very high precisions. For example, we want to compute the roots of the equation $x^2 + 123x + 0.5 = 0$ correct up to five significant digits. In that case, we can use the computing device with more than ten digits floating point arithmetic, such that the results can be obtained up to desired accuracy even after the loss of significance.

2.4 Some Interesting Facts about Error

a) Let us assume we are doing six significant digits arithmetic on a hypothetical computer. If we want to add a small number $x = 0.000123$ to a large number $y = 123.456$ using this computer, then

$x + y = (.123456)10^3 + (.123000)10^{-3}$ (Normalized form)
$= (.123456)10^3 + (.000000)10^3$ (Equal exponent using symmetric rounding)
$= (.123456)10^3$ (Result, we missed the addition!)

This type of situations occurred commonly during the computations of infinite series. In these series, the initial terms are comparatively large. So, usually after adding some terms of the series, we are in a situation of adding a small term to a very large term. It may produce high rounding error in the computation. To avoid this kind of error, we can use backward sum of the series instead of forward sum, such that the each new term is compatible with the magnitude of accumulated sum.

b) In the case of series with mixed signs (like Taylor series of sin(x)), sometimes individual terms are larger than the summation itself. For example, in Taylor series of sin(2.13), the first term is 2.13. It is called as smearing, and we should use these kinds of series with extra care.

c) While performing arithmetic computations in a numerical method, the steps involving large number of arithmetic operations must be computed in double precisions. Such operations are error-prone to round-off error. For example, in Gauss-Seidel method for the solution of system of linear equations, the inner product

$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$ is a common operation, and such computations must be made in double precisions.

The accumulated rounding error can create the disastrous results; the following two examples of rounding errors are picked from the internet.

(*http://mathworld.wolfram.com/RoundoffError.html*)

1. An index was started with initial value 1000.000 for Vancouver stock exchange (McCullough and Vinod 1999) in 1982. Three decimal digits chopping arithmetic has been used to compute the index for each change in market value for next 22 months. The computed value was 524.881, while its correct value up to three decimal points is 1009.811.
2. The Ariane rocket was launched on June 4, 1996 (European Space Agency 1996). In the 37th second of flight, a 64-bits floating point number was converted to a 16-bits number by the inertial reference system of the rocket. It was an overflow error, but the guidance system interpreted it as flight data, which led the rocket to getting destroyed.

Exercise 2

1. Define normalized form and hence the number of significant digits for floating point numbers with examples.

2. Find out the number of significant digits in the numbers 788500, 0.4785, .003523, 0.2300, and 7.880.

Ans. 6, 4, 4, 4, 4

3. Compute the absolute errors (A.E.) and relative errors (R.E.) in the four significant digits chopping approximations of the numbers 234168 and 64.2685.

Ans. A.E. = 68, 0.0085; R.E. = 0.000290389, 0.000132257

4. If β based real number $x=(0.d_1d_2d_3....d_nd_{n+1}....)_\beta \times \beta^e$ is chopped to n digits and $fl(x)$ is its representation, then show that

$$0 \le \frac{x - fl(x)}{x} \le \beta^{1-n}$$

5. If x is any number in decimal number system and $fl(x)$ is its machine representation up to n digits, then for rounding, show that

$$\left|\frac{x - fl(x)}{x}\right| \le \frac{1}{2} \times 10^{1-n}$$

6. The true value of e(exponential) correct to 10-significant digits is 2.718281828. Calculate absolute and relative errors, if we approximate this value by 2.718.

Ans. A.E. = .000281828, R.E. = 0.000103678

7. Find the relative errors for the following cases. Also, determine the number of significant digits in the approximations:

$x = 2.71828182$ and $\tilde{x} = 2.7182$

Ans. R.E. = 0.0000301, 5 Significant Digits

$y = 28350$ and $\tilde{y} = 28000$

Ans. R.E. = 0.0123457, At least 2 Significant Digits

$z = 0.000067$ and $\tilde{z} = 0.00006$.

Ans. R.E. = 0.104478, 1 Significant Digits

8. Define the terms error, absolute error, relative error and significant digits. The numbers $x = 1.28$ and $y = 0.786$ are correct to the digits specified. Find estimates to the relative errors in $x + y$, $x - y$, $x\,y$, and x/y.

Ans. R.E. in $x + y = 0.00266215$, $x - y = 0.0111336$, $x\,y$ and $x/y = 0.00454238$

9. Consider the following decimal numbers with a four digits normalized mantissas, $a = 0.2473*10^4$, $b = 0.8125*10^3$, $c = 0.1523*10^1$
Perform the following operations in four significant digits symmetric rounding and indicate the errors in the results.

i) $a + b - c$ **Ans.** $0.3283*10^4$, Error = 0.0000977
ii) b / c **Ans.** $0.5335*10^3$, Error = – 0.0001346
iii) $a - b$ **Ans.** $0.1661*10^4$, Error = – 0.00005
iv) $b / (a + c)$ **Ans.** 0.3283, Error = 0.00004611

10. The numbers $x_1 = 0.643$ and $x_2 = 1.631$ are correct to the significant digits in the numbers. Compute the relative errors in the addition, subtraction, multiplication and division of these two numbers.

Ans. R.E. in $x_1 + x_2 = 0.00043975$, R.E. in $x_1 - x_2 = 0.001012145$,
R.E. in x_1x_2 and $x_1 / x_2 = 0.001084164$

11. Calculate the absolute and relative errors in the expression $3a-2bc+\frac{b}{a}$, if the measurement of $a = 3.5435$, $b = .2588$ and $c = 1.0150$ are possible only to correct up to four decimal points.

Ans. Absolute Error = 0.0002925, Relative Error = 0.00002874

12. Estimate the error in evaluating the function $f(x)=e^{2x^2}\sin(x)$ near the point $x = 1$, if the absolute error in value of x is 10^{-4}.

Ans. Absolute Error = 0.0028863

13. The maximum error tolerance in the measurement of the area of a given circle is 0.1%. What is maximum relative error allowed in the measurement of the diameter?

Ans. 0.05

14. Compute the resulting error in the function $f(x)=x^3$ for value of $\tilde{x}=2.38$ with an error $\Delta\tilde{x}=0.005$

Ans. Absolute Error = 0.084966

15. Find the maximum possible error in the computed value of the hyperbolic sine function $\sinh(x)=\frac{e^x-e^{-x}}{2}$ at the point $x = 1$, if the maximum possible error in the value of x is $|dx|=0.01$.

Ans. 0.01543

16. Let the function $u=\frac{4x^2y^3}{z^4}$ and errors in the values of variables x, y, z are 0.001. Find the relative error in the function u at $x=y=z=1$.

Ans. 0.009

17. The radius r and height h of a right circular cylinder are measured as 2.5 m and 1.6 m, respectively, with a maximum error of 2%. Compute the resulting percentage error measured in the volume of the cylinder by the formula $V=\pi r^2h$. Assume the value of π is exact for calculation.

Ans. 0.06

18. Consider a function $u=e^x\sin(y)+x\ln(z)$. Let the variables x, y and z be measured with maximum possible errors of ±0.01, $\pm\left(2^o=\frac{\pi}{90}\right)$ and ±0.5, respectively. Estimate the maximum possible error in computing the function u for $x=0.1$, $y=\frac{\pi}{4}$ and $z=50$.

Ans. 0.4976

19. The voltage V in an electrical circuit satisfies the law $V=IR$, where I is the current and R is the resistance and their starting values are $I = 5$ amp, $R = 600$ ohms, respectively. Let us assume that after a certain time, resistance is changed 0.15% due to heating, and we changed the current I with 5%. Compute the percentage change in the voltage V.

Ans. 5.15%

20. The length of a simple pendulum measured is $l = 0.362$ m, while the constants $\pi=3.1416$ and g = 9.8 m/sec^2 are correct to the specified digits. Compute the relative error in the time-period $T=2\pi\sqrt{\frac{l}{g}}$.

Ans. 0.0032575

21. Compute the absolute and relative errors in the function $f(x,y,z)=y^2e^z\cos(x)$ at $x=1.5$ $y=2.3$ and $z=5$, if the error in the values of *x*, *y* and *z* are 0.05.

Ans. Absolute Error = 44.3483, Relative Error = 0.7985

22. Calculate the number of terms required in Taylor series approximation of the function $\cos(x)$ to compute the value of $\cos\left(\frac{\pi}{12}\right)$ correct up to 4-decimal places.

Ans. 5

23. Find the number of terms of the Taylor series expansion of the function e^x required to compute the value of *e* correct to six decimal places.

Ans. 10

24. Discuss CN and stability of the function $y=\sec(x)$ in the interval $\left[0,\frac{\pi}{2}\right]$.

Ans. CN = $x\tan(x)$; as we move from 0 to $\frac{\pi}{2}$ in the interval $\left[0,\frac{\pi}{2}\right]$, the CN increase and hence function $y=\sec(x)$ become ill-conditioned.

25. Calculate the function $f(x)=\cos(x)-\sin(x)$ at the point *x* = 0.785398 using 6-decimal digits floating-point round-off arithmetic. Discuss the condition and stability of process involved.

Ans. $f(x)=\cos(x)-\sin(x)=0$ at *x* = 0.785398 with 6-decimal digits floating-point round-off arithmetic.
CN of $f(x)=\cos(x)-\sin(x)$ at $x=0.785398$ is approximately 0, hence function is well conditioned.
But the process is not a stable process.
We can use any of the following stable processes for computation purpose

i) $f(x)=\dfrac{\cos(2x)}{\cos(x)+\sin(x)}$

ii) $f(x)=1-x-\dfrac{x^2}{2!}+\dfrac{x^3}{3!}+\dfrac{x^4}{4!}-\dfrac{x^5}{5!}-\dfrac{x^6}{6!}+\cdots$

26. Discuss the condition and stability of the function $f(x)=x-\sqrt{x^2-1}$ at $x=11111$, using six significant digits floating point rounding arithmetic. Find a stable way to compute the function.

Ans. $f(x)=\dfrac{1}{x+\sqrt{x^2-1}}$

27. Evaluate roots of the quadratic equation $x^2+234.56x+1.2345=0$, with the minimum loss of significant digits. Use five significant digits chopping arithmetic.

Ans. – 234.55, – 0.0052632

28. Avoiding loss of significance, find the smallest root of the quadratic equation $x^2-500x+2=0$ by using five significant digits rounding arithmetic.

Ans. 0.004, 500

29. Discuss the condition and stability of the function $f(x)=x-\sin(x)$, when *x* is nearly equal to zero. Find a stable way to compute the function *f*(*x*).

Ans. $\dfrac{x^3}{3!}-\dfrac{x^5}{5!}+\dfrac{x^7}{7!}-\dfrac{x^9}{9!}+\cdots$

30. Subtraction of nearly equal numbers leads to loss of significant digits. Obtain equivalent formulas for the following functions to avoid loss of significance.

a) $\cos^2(x)-\sin^2(x)$ for $x \approx \frac{\pi}{4}$

b) $x-\sin(x)$ for $x \approx 0$

c) $x-\sqrt{x^2-1}$ for large x

d) $1-\cos^2(x)$ for $x \approx 0$

e) $\sqrt{1+\cos(x)}$ for $x \approx \pi$

Ans. a) $\cos(2x)$, b) $\frac{x^3}{3!}-\frac{x^5}{5!}+\frac{x^7}{7!}-\frac{x^9}{9!}+\cdots$, c) $\frac{1}{x+\sqrt{x^2-1}}$, d) $\sin^2(x)$, e) $\sqrt{\frac{\sin^2(x)}{1-\cos(x)}}$

Nonlinear Equations

Chapter 3

Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.

Sir Isaac Newton
(December 25, 1642–March 20, 1726)
He was a great mathematician and physicist. He pioneered 'classical mechanics'.

3.1 Introduction

Mathematical models for many problems in different branches of science and engineering are formulated as

$$f(x) = 0 \tag{3.1}$$

where the variables x and $f(x)$ may be real or complex, and scalar or vector quantities. In this chapter, the variables x and $f(x)$ are real and scalar quantities. The value of x, which satisfies the Eq. (3.1), is called the root of the equation. It is also known as the zero of the function $f(x)$. For example, the quadratic equation

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

has roots 1 and 2.

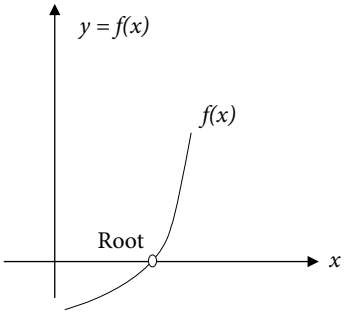


Fig. 3.1 Root of $f(x) = 0$

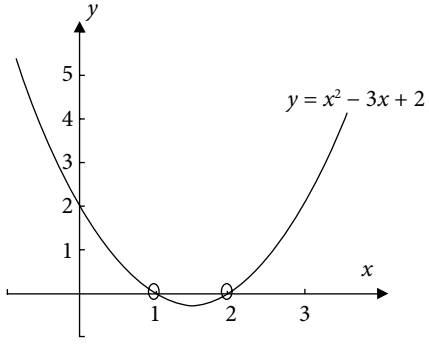


Fig. 3.2 Root of $x^2 - 3x + 2 = 0$

Root finding is also essential in many branches of mathematics. For example, the critical points of the function $f(x)$ are the roots of the equation $f'(x) = 0$. Eigenvalues of a square matrix A are the roots of the characteristic equation.

$$p(\lambda) = \det\left(A - \lambda \mathrm{I}\right) = 0$$

where $p(\lambda)$ is a polynomial of degree n (order of matrix A).

The nonlinear equations can be categorized broadly as polynomial equations and transcendental equations as follows

3.1.1 Polynomial Equations

The polynomial equations are given by

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

This equation is an nth degree polynomial equation, and has exactly n roots. These roots may be real or complex. Some examples of polynomial equations are

i) $3x^3 + x^2 - 9 = 0$
ii) $x^2 - 4x - 5 = 0$

3.1.2 Transcendental Equations

An equation which is not a polynomial equation is a transcendental equation. These equations involve trigonometric, exponential and logarithmic functions, etc. A few examples of the transcendental equations are as follows

i) $3\sin(x) - e^{-x} = 0$
ii) $3x^2 - 2\cos x = 0$
iii) $2e^x \sin x - \ln(x) = 0$

Transcendental equations may have finite or infinite numbers of real roots or may not have real roots at all.

The roots of simple equations are easy to compute by the direct methods. But in the case of higher order equations and transcendental equations, there is no general analytical method to compute the exact roots. So for this purpose, numerical techniques can be used to find approximate roots of the equation. The main objective of this chapter is to present and discuss the various numerical techniques which are useful for finding the approximate roots of the nonlinear equation, Eq. (3.1).

3.2 Methods for Solutions of the Equation $f(x) = 0$

So far, various methods have been developed for the solution of Eq. (3.1). All these methods have their advantages and disadvantages, and broadly categories as follows

i) Direct analytical methods
ii) Graphical methods
iii) Trial and error methods
iv) Iterative methods

In this section, we shall have a brief idea of these methods and conclude that iterative methods for finding numerical approximations for the roots of Eq. (3.1) are the best methods for the complex and complicated equations.

3.2.1 Direct Analytical Methods

We can solve the nonlinear equation by direct analytical methods in certain simple cases. For example, the roots of the quadratic equation $ax^2+bx+c=0$ are given by $x=\dfrac{-b\pm\sqrt{b^2-4ac}}{2a}$. Similarly, roots of cubic and quartic equations can be obtained using Cardano and Ferrari methods, respectively. The roots obtained by direct methods are exact roots of the equations. But these methods can be applied to some very special categories of the equations. The roots of higher order polynomial equations/ transcendental equations (like $x^5+2x^3+3x^2+5x+6=0$ and $2e^{-x}+x^3\sin x=0$) cannot be obtained from direct analytical methods. We don't have direct methods even for the solutions of simple transcendental equations.

3.2.2 Graphical Methods

Plotting of the function $f(x)$ with x-axis gives the root of the equation $f(x) = 0$. The points where the curve $f(x)$ crosses the x-axis, are the roots of the equation.

Solutions obtained using graphical methods are not accurate. But graphs of some standard curves are helpful in tracing the interval in which the root of the equation lies and are also important for an initial guess about the roots, etc. Let us discuss a few examples.

Case 1. Equations $x^2+x+1=0$, $x-e^x=0$, $x-\ln(x)=0$ with no real roots;

The graph (Fig. 3.3) of $y=x^2+x+1$ has no point of intersection with x – axis, so the equation $x^2+x+1=0$ has no real root.

Any equation $f(x) = 0$ can be rewritten as $f_1(x) = f_2(x)$, and points of intersections of the curves $y=f_1(x)$ and $y=f_2(x)$ provide the roots of the equation $f(x) = 0$. For example, consider the graphs of $y = e^x$ and $y = x$, then the points of intersection of these two curves are the roots of equation $x - e^x = 0$ ($x = e^x$). It is easy to see that there is no point of intersection (Fig. 3.4), so the equation $x - e^x = 0$ has no real root. Similarly, we can easily find that equation $x-\ln(x)=0$ also has no real root (Fig. 3.5).

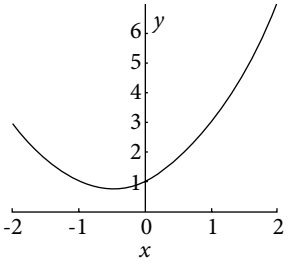


Fig. 3.3 $x^2 + x + 1 = 0$

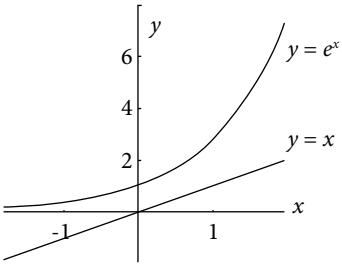


Fig. 3.4 $x = e^x$

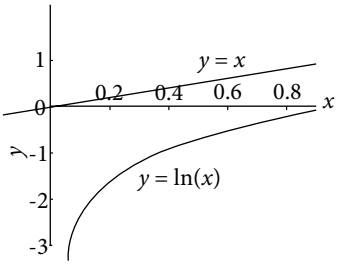


Fig. 3.5 $x - \ln(x)$

Case 2. Equations with finite numbers of real roots like $x^3 - 3x^2 + 2x = 0$, $x - \cos x = 0$, etc.

The points of intersection of two curves $y = \cos(x)$ and $y = x$ are the roots of equation $x - \cos x = 0$ (Fig. 3.6). There is only one point of intersection, so the equation $x - \cos x = 0$ has only one real root. Similarly, the graph of the function $y = x^3 - 3x^2 + 2x$ provides that the equation $x^3 - 3x^2 + 2x = 0$ has three real roots (Fig. 3.7).

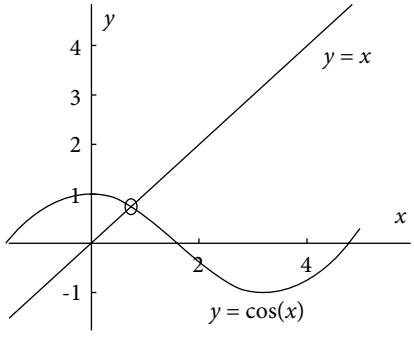


Fig. 3.6 $x = \cos x$, Root ≃ 0.7390851322

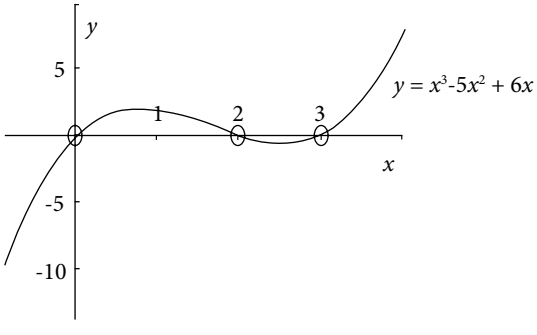


Fig. 3.7 $x^3 - 5x^2 + 6x = 0$, Roots = 0, 2, 3

Case 3. Equations with infinite numbers of real roots like $e^x - \cos x = 0$, $x - \tan x = 0$, $e^{-x} - \sin x = 0$, etc. The following graphs show that these equations have infinitely many real roots.

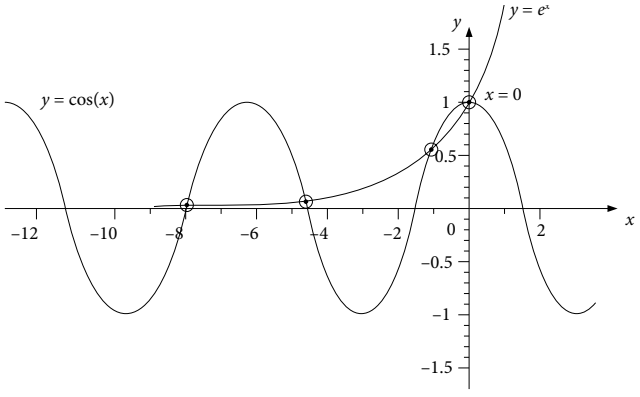


Fig. 3.8 $e^x = \cos(x)$, Roots ≃ 0, –1.2927, –4.7213, –7.8536, ...

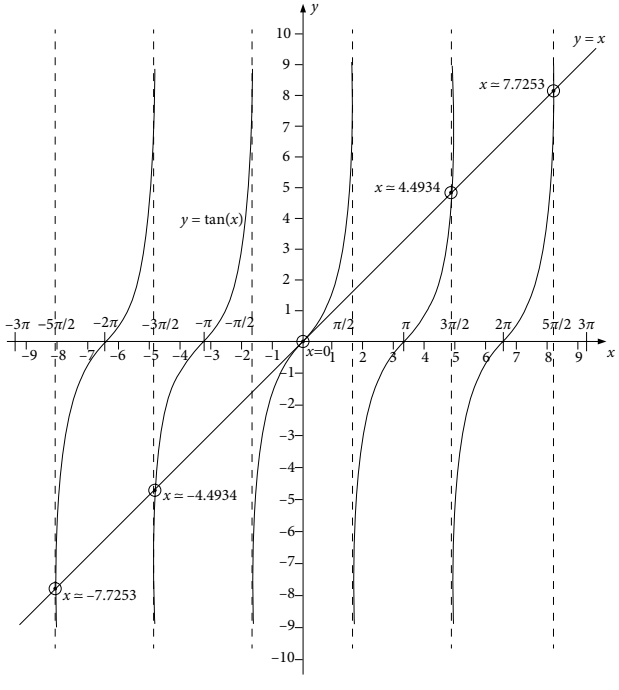


Fig. 3.9 $x = \tan x$, Roots ≃ 0, ±4.4934, ±7.7253, ...

3.2.3 Trial and Error Methods

Other approaches to obtain the approximate solutions are the trial and error techniques. These methods involve a series of guesses for the root of the equation (3.1) and check whether the value of the function is close to zero. The value of x, where function is close to zero, is the approximate root of the equation.

The trial and error methods are cumbersome and time-consuming. These methods are not algorithmic, so programming is not possible. Also, approaches in these methods vary from problem to problem. So, these methods are no longer in use.