SECOND EDITION

Elements of Numerical Analysis

Radhey S. Gupta

Elements of
Numerical Analysis

Second Edition

Radhey S. Gupta

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Cambridge House, 4381/4 Ansari Road, Daryaganj, Delhi 110002, India

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/9781107500495

Second edition first published 2015

Printed in India

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

Library of Congress Cataloging-in-Publication Data
Gupta, Radhey S.,
1937- Elements of numerical analysis. – Second edition / Radhey S. Gupta.
pages cm
Includes bibliographical references and index.
Summary: "Offers detailed discussion on difference equations, Fourier series, discrete Fourier transforms and finite element methods"– Provided by publisher.
ISBN 978-1-107-50049-5 (pbk.)
1. Numerical analysis–Textbooks. 2. Mathematics–Study and teaching (Higher)–Textbooks.
3. Mathematics–Study and teaching (Graduate)–Textbooks. I. Title.
QA297.G87 2015
518–dc23
2014038362

ISBN 978-1-107-50049-5 Paperback

Additional resources for this publication available at www.cambridge.org/9781107500495

To my supervisor
(Late) Professor John Crank

and
To my family
Wife: (Late) Pramila
Children: Virna, Neelam, Aditi and Mayank

Contents

Preface

As I think of writing about the present work I am pleasantly reminded of a few names which occupy very special place in my academic and professional career. In 1966, while I was working at the Post Office Research Station in London (now most probably at Martlesham) I got introduced to computers. I cannot forget, with how much patience and perseverence, B.E. Surtees had not only helped but had almost taught me Algol programming. Later I used the computer extensively for solving scientific problems. Further, the Research Station generously granted me day-release to attend MSc (Comp.Sc.) course at the City University, London. Although I could not complete the course, I developed a strong liking for Numerical Analysis. It would be my privilege to mention the name of Professor V.E. Price who taught the subject with full devotion and dedication. I must admit that I learnt the basics of Numerical Analysis from there and much of it makes part of Chapters 1 to 7 and 10 of the book. I was greatly impressed by the book *Modern Computing Methods* by E.T. Goodwin, and still am. My intense desire for working in Numerical Analysis was fulfilled when I joined PhD in 1969 at Brunel University under the guidance of Professor J. Crank who was known internationally in the field. Luckily, a very challenging problem came my way to work upon from Hammersmith Hospital, London. The problem required knowledge for solving partial differential equations numerically. The first book on p.d.e., I read was G.D. Smith's who was coincidently teaching in the same department. Therefore no wonder, my treatment for solving p.d.e.'s in Chapter 11 may be biased towards his book. I worked with Professor Crank for five years – three years for my PhD and two years as a postdoctoral research fellow. It was only his constant inspiration that kept me going and galloping. Those five years, I may call the most precious years of my life. I came to know with deep sense of sorrow and grief that Professor Crank passed away in October 2006. This book is a humble tribute to him. Same time when I was doing PhD, Professor J.R. Whiteman joined the department. He taught splines to the students of MSc (Numerical Analysis) and gave lectures on variational principle applied to Finite Element Method. I came to know about these techniques through him which have been provided in Chapters 8 and 12. Nick Papamichael was another good fellow in the department who taught Integral Equations to MSc (Numerical Analysis). I learnt initially about this topic from his notes which have been useful in writing Chapter 13.

After coming to India I joined I.I.T., Delhi as Pool Officer, with Professor M.P. Singh who was heading a Centre, concerned with problems in bio-mathematics and atmospheric science. I had a good interaction with the members of his team working on diffusion problems. Professor Singh was extremely helpful in providing me all facilities – academic and otherwise. He had been a source of constant encouragement and inspiration during that

period and afterwards also, as I stayed there for less than $1^1/_2$ years only. I joined in 1976, the Department of Mathematics at University of Roorkee (now I.I.T., Roorkee). There I got an opportunity to hone and extend my knowledge of Numerical Analysis further through teaching and guiding research. Professor C. Prasad, Head of Department, wanted to see an all-round expansion in Numerical Analysis and Computer Science in the department. I was entrusted to carry out various activities in these areas. A postgraduate diploma course in Computer Science was started in the department in 1978. Same year, I also organised a short term course on Numerical Solution of Partial Differential Equations under Quality Improvement Program. Professor M.B. Kanchi (Civil Engineering Department) gave lectures on Finite Element Method (FEM) in this program. It inspired me to broaden my knowledge on FEM which has been included in Chapter 12. Further, I thought to provide the reader an exposure to a very important class of problems known as free and moving boundary problems. Such problems arise in almost all branches of engineering and applied sciences. A brief introduction to these problems is given in Chapter 15. I have included a list of my research papers on moving boundary problems so that the interested readers may search other papers through cross references. My stay of 21 years at University of Roorkee (I.I.T., Roorkee) had been extremely fruitful academically as well as in personal relations. I have always cherished its memories in my heart and will continue to do so all my life. As would have been clear, the present book is, in a way, direct or indirect contribution from various people — to whom I feel greatly indebted. Whatever faults are there, they are mine — criticism and suggestions would be most welcome. I do hope the book will be useful to students, to teachers and to those who want to use Numerical Analysis as a tool for solving practical problems.

Preface to Second Edition

It gave me a great sense of satisfaction and happiness to hear some good words about the book from my old colleagues and acquaintances working in the field of Numerical Analysis. In this edition I have included a new chapter dealing with Fourier Series, Fourier Transform and Fast Fourier Transform (FFT). In fact I wanted to include this topic in the first edition itself but I was not truly prepared then. Now I have also added first order hyperbolic equation as a new section in the chapter on partial differential equations. I thank Cambridge University Press for bringing out this revised edition.

1

Errors in Computation

1.1 Introduction

For solving a mathematical problem by numerical method, an input is provided in the form of some numerical data or it is generated/created as called for by the problem. The input is processed through arithmetic operations together with logical operations, which are performed in a systematic manner and the output is produced in the form of some numbers. Thus the whole exercise in Numerical Analysis is all about manipulation of numbers. Whether we are working by hand or on a computing machine, there is always a constraint in regard to physical size of the numbers, i.e., the number of digits a number can contain. Inside a computer the size of the number is dependent on its word-length (number of bits) which also puts a limit on the range of numbers that can be represented in a particular computer. Further, it may be noted that all numbers are not represented exactly inside the computer and that the input given in the decimal form is converted to binary in the computer. It should also be remembered that fractions cannot be stored in their natural form; they are converted to decimals, for example 2/5 is input as 0.4 and 1/3 as 0.333... up to a finite number of digits acceptable by a computer.

1.2 Floating Point Representation of Number

When a number x is expressed as,

$$x = p \times 10^q$$

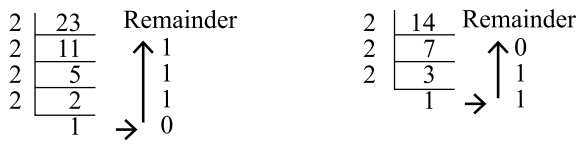
where $0.1 \leq |p| < 1.0$ and q is an integer (positive (+ve) or negative (−ve)), it is called 'floating point' representation of number x. A floating point form consists of two parts; the fractional part p (alongwith the sign) is known as mantissa and the other part q as exponent, a power raised to a radix (in the case of decimal system, 10). At some places it is referred to

as 'normalised floating point' and when $1 \leq p < 10$, the form is called 'scientific notation'. A few examples of floating point representation, $fl(x)$ of number x are given as follows:

x	$fl(x)$	*Mantissa*(p)	*Exponent*(q)
2.0456	0.20456×10^{1}	0.20456	1
-32.7652	-0.327652×10^{2}	-0.327652	2
0.00234	0.234×10^{-2}	0.234	-2
0.000000034	0.34×10^{-7}	0.34	-7
34000000	0.34×10^{8}	0.34	8

1.3 Binary Numbers

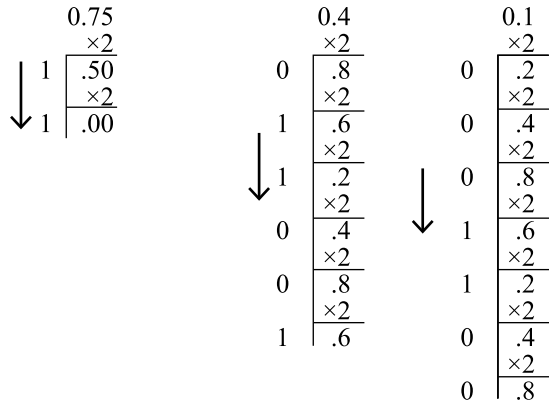
The decimal numbers (radix 10) are converted to binary form with digits 0 and 1 (radix 2) in the computer. An integer decimal number, may be converted to binary equivalent by following procedure:



Divide repeatedly by 2 until last quotient is 1, keeping the remainder against the quotient; read the binary digits in the direction of arrow. Thus we get,

$$23 = 10111;\ 14 = 1110.$$

The fractional decimal number is converted to binary form in the following manner:



Multiply by 2 until the decimal part is zero, saving digit 0 or 1 before the decimal point. Read the digits saved in a top-down manner. Thus the converted numbers are

$$0.75 = 0.11;\ 0.4 = 0.0110011\ (0011 \text{ recurring});\ 0.1 = 0.0001100\ (1100 \text{ recurring})$$

When a decimal number consists of both parts, integral as well as fractional, then both parts are converted to binary forms separately. For example, 23.75 will convert to 10111.11. It should be clear from the above examples that the integer numbers in the decimal system can be converted exactly in the binary system but most of the non-integers may be represented approximately due to non-terminating character of the converted numbers.

For conversion from binary to decimal, we simply multiply the binary digits by their respective place-value and add. For example, 10111.11 can be converted to decimal form as,

$$\begin{array}{cccccccc} 2^4 & 2^3 & 2^2 & 2^1 & 2^0 & 2^{-1} & 2^{-2} & \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & = 1\times 2^4 + 0\times 2^3 + 1\times 2^2 + 1\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 1\times 2^{-2} \\ & & & & & & & = 16+0+4+2+1+.5+.25 \\ & & & & & & & = 23.75 \end{array}$$

It may also be noted that largest k-digit binary integer will have the value $2^k - 1$ in decimal. For example, the largest 2-digit binary number will be $11 = 2^2 - 1 = 3$ and a 3-digit largest binary number will be $111 = 2^3 - 1 = 7$ and so on. Obviously all the k digits will be binary 1's. A k-digit binary number can represent 2^k decimal numbers from 0 to $2^k - 1$.

1.3.1 Binary number representation in computer

As stated earlier, all the input data is converted to binary inside the computer; while the decimal integers are represented exactly in the computer memory, the non-integers are represented in floating point form. We would like to explain very briefly as how the floating point numbers are stored in the computer memory. Consider the floating point representation of binary numbers given below:

Binary number	*Floating point form*	*Mantissa*	*Exponent*
0.0111	0.1110×10^{-01}	$+0.1110$	-01
-1.101	$-0.1101\times 10^{+01}$	-0.1101	$+01$
11.1	$0.1110\times 10^{+10}$	$+0.1110$	$+10$

It may be noted that all numbers are in binary so that 10 is equal to 2 in decimal. The other thing to be noted is that mantissa is expressed in four digits and exponent in two digits, in each case.

Let us now consider a hypothetical case of a computer having a word length of 8 bits only. Out of eight bits, the left-most bit is used for storing the sign of mantissa. Let 0 denote positive (+ve) and 1 denote negative (−ve) sign of mantissa. The next four bits are used for storing the binary digits of mantissa. The right-most 3 bits are used for storing the exponent part; the first bit for storing its sign and last two bits for its value, digit 0 showing positive (+ve) and digit 1 showing negative (−ve) exponent (See Fig. 1.1).

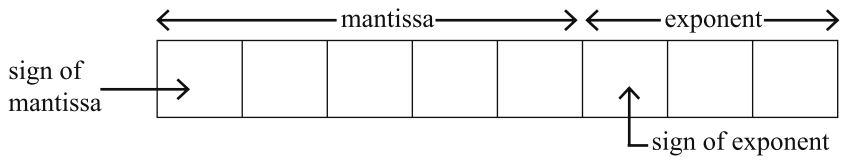


Figure 1.1 Floating point representation in 8-bit computer memory.

According to the memory configuration of Fig. 1.1 the binary numbers given above will be represented in the floating point form as follows:

	Binary number with decimal equivalent	*Representation in 8-bit memory*
(*a*)	0.0111 (0.4375)	01110101
(*b*)	−1.101 (−1.875)	11101001
(*c*)	11.1 (3.5)	01110010

It may be stated that the positive exponent varies from 000 to 011, i.e., from 0 to 3 in decimal. The negative exponent should vary from 101 to 111, i.e., from −1 to −3. But 100 may be considered as −4, since 000 is already zero, hence negative exponent varies from −1 to −4 in decimal.

It may be noted that the largest positive number that can be stored under present configuration would be, $0.1111 \times 10^{11} = 111.1$ (binary) $= 7.5$ (decimal). The algebraically smallest number that can be stored would be -7.5 (in decimal). Thus the range of numbers that can be represented in the computer memory would be $-7.5 \leq x \leq 7.5$. The smallest positive non-zero number represented in the above memory configuration would be, $0.1000 \times 10^{100} = 0.00001$ (binary) $= 2^{-5} = 0.03125$ (decimal). However, it may also be mentioned that even the simplest computer has a memory of 32-bit word and two or more words can be adjoined to store a number in floating point. Thus the space (number of bits) occupied by the mantissa and the exponent would be manifolds that of shown in Fig. 1.1

but the logic remains same. When a fixed number of decimal digits are kept in all numbers, it is called 'Fixed Point' representation.

1.4 Significant Digits

All the digits from 0 to 9 in a number, except the zeros which are used for fixation of decimal point, are called significant digits (or figures). For example, in the number .003456, the first two zeros are not significant since we can also express the number as $.3456 \times 10^{-2}$, while the other four digits, namely, 3, 4, 5 and 6 are significant. But in the number 20.003456, all the eight digits are significant. In order to find the number of significant digits in a number, express it in floating point; the mantissa part gives the number of significant digits.Whether the last zeros in a number are significant or not may depend on the context. For example, in measuring the heights of the students, in 168.00 cm, the last zero may not be significant and we can express the height as 168.0 cm, showing that the height is being measured nearest to the $\frac{1}{10}$th part of the centimeter, so that zero in 168.0 cm is significant.

1.5 Rounding and Chopping a Number

In scientific computing we are encountered by numbers with too many digits. More often than not, we have to shorten/reduce them to a size which may not affect the end result within a desired accuracy. There are two ways of reducing the size of or truncating the number, viz., (*i*) rounding (*ii*) chopping. Let us first discuss the procedure for rounding off a number x in decimal system.

Let the number x be expressed in floating point form with s digits in mantissa and with exponent q, as

$$x = \cdot d_1 d_2 \ldots d_n d_{n+1} \ldots . d_s \times 10^q.$$

If the number x is to be rounded to n significant digits, following procedure would be adopted:

(*i*) if $d_{n+1} < 5$, then no change in any of the digits from d_1 to d_n and the rounded number would be,

$$x \simeq .d_1 d_2 \ldots . \, d_n \times 10^q.$$

(*ii*) if $d_{n+1} > 5$, then digit d_n is incremented by 1, i.e. d_n becomes $d_n + 1$; as a cosequence of this other digits may get affected and even the exponent may have to be adjusted accordingly.

(*iii*) if $d_{n+1} = 5$, then there will be two ways for rounding, depending upon d_n being an even digit $(0, 2, 4, 6, 8)$ or an odd digit (1, 3, 5, 7, 9). If d_n is even, then case (i) applies and if d_n is odd then case (ii) applies. Thus probability of both cases is $\frac{1}{2}$ when $d_{n+1} = 5$.

Given below are some examples of rounding the numbers to four places of decimal (four significant digits):

	Floating point number	*Rounded to four decimals*
(a)	0.245684×10^2	0.2457×10^2
(b)	0.245629×10^{-2}	0.2456×10^{-2}
(c)	0.245659×10^2	0.2456×10^2
(d)	0.245750×10^2	0.2458×10^2
(e)	0.999951×10^2	0.1000×10^3
(f)	0.999858×10^2	0.9998×10^2

The difference between examples (e) and (f) may be noted. It may also be observed that in example (e) all zeros in the rounded number are significant.

However, a more conventional way for rounding, is straight in that if $d_{n+1} < 5$, then all the digits from d_1 to d_n remain unaltered [case (i)] while if $d_{n+1} \geq 5$, then d_n is incremented by 1 and necessary changes are made in the digits d_1 to d_n and also in the exponent, if necessary [case (ii)].

When all the digits after d_n are ignored, irrespective of whatever value d_{n+1} has, the procedure for truncating the number is known as 'chopping off' the number or simply 'chopping'. If there are sufficient number of significant digits in a number, like in a computer, the process of chopping may not affect the result in normal circumstances.

1.6 Errors due to Rounding/Chopping

Suppose a number x is rounded to x^*, then the modulus of the difference between x and x^*, i.e. $|x - x^*|$ is known as rounding error or error due to rounding in x^*.

Let x be a number which has been rounded to 4 decimals, say $x^* = 0.4387$. Then lower and upper bounds for the actual number x would be,

$$0.43865 \leq x < 0.43875$$

or $\quad 0.43865 - 0.4387 \leq x - x^* < 0.43875 - 0.4387$

or $$-0.00005 \leq x - x^* < 0.00005$$

or $$|x - x^*| \leq 0.00005 = \frac{1}{2} \times 10^{-4}$$

$$= \frac{1}{2} \times \text{ unit at } 4^{\text{th}} \text{ decimal place.}$$

The above result can be generalised for a number x represented in the floating point form as,

$$x = \cdot d_1 d_2 \ldots . d_n d_{n+1} \ldots . d_s \times 10^q.$$

If x is rounded to n decimals, then the maximum rounding error would be,

$$|x - x^*| \leq \frac{1}{2} \times 10^{-n} \times 10^q$$

$$= \frac{1}{2} \times 10^{q-n}. \tag{1.1}$$

If the number x is chopped off to n decimals, then it is easy to see that the maximum error due to chopping would be,

$$|x - x^*| \leq 10^{q-n}. \tag{1.2}$$

That is, the error in chopping a number is twice that in the rounding.

1.7 Measures of Error in Approximate Numbers

Let x^* be an approximation of exact number x, then we can measure the magnitude of error in three different forms:

(*i*) absolute error (a.e.) $= |x - x^*|$ (1.3a)

(*ii*) relative error (r.e.) $= \left|\frac{x - x^*}{x}\right|$ or $\left|\frac{x - x^*}{x^*}\right|$ (1.3b)

(*iii*) percentage error (p.e.) $=$ r.e. $\times 100$ (1.3c)

1.8 Errors in Arithmetic Operations

Let x_1 and x_2 be two numbers which are rounded to x_1^* and x_2^* respectively and let ε_1 and ε_2 be the corresponding rounding errors in them, such that $x_1 = x_1^* + \varepsilon_1$ and $x_2 = x_2^* + \varepsilon_2$. We are going to study below, the effects of rounding errors on arithmetic operations, viz., addition, subtraction, multiplication and division.

(*i*) Addition

$$x_1 + x_2 = x_1^* + \varepsilon_1 + x_2^* + \varepsilon_2$$

$$= x_1^* + x_2^* + \varepsilon_1 + \varepsilon_2$$

or $\quad |(x_1 + x_2) - (x_1^* + x_2^*)| = |(x_1 - x_1^*) + (x_2 - x_2^*)| = |\varepsilon_1 + \varepsilon_2| \leq |\varepsilon_1| + |\varepsilon_2|.$

This can be generalised to n numbers as,

$$\left| \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^* \right| = \left| \sum_{i=1}^{n} (x_i - x_i^*) \right| = \left| \sum_{i=1}^{n} \varepsilon_i \right| \leq \sum_{i=1}^{n} |\varepsilon_i| \tag{1.4}$$

That is, the total absolute error in the sum of n numbers will be less than or equal to the sum of the absolute errors in each of them. Although it gives the upper bound for the absolute error in the sum, the actual error will be much smaller since some of the errors may be positive and some negative so that cumulative effect would be much reduced.

(*ii*) Subtraction

$$x_1 - x_2 = x_1^* + \varepsilon_1 - x_2^* - \varepsilon_2$$

or $\quad |(x_1 - x_2) - (x_1^* - x_2^*)| = |(x_1 - x_1^*) - (x_2 - x_2^*)| = |\varepsilon_1 - \varepsilon_2| \leq |\varepsilon_1| + |\varepsilon_2|.$ (1.5)

Thus the absolute error in subtraction of two approximate numbers can be as great as the sum of their individual absolute errors; since the errors can be positive or negative and if ε_1 and ε_2 are of opposite signs, then under subtraction, they will be added up.

(*iii*) Multiplication

$$x_1 \cdot x_2 = (x_1^* + \varepsilon_1) \cdot (x_2^* + \varepsilon_2)$$

or $\quad x_1 \cdot x_2 - x_1^* \cdot x_2^* = x_1^* \varepsilon_2 + x_2^* \varepsilon_1$, neglecting $\varepsilon_1 \varepsilon_2$

$$\text{a.e.} = |x_1 x_2 - x_1^* \cdot x_2^*| \leq |x_1^* \varepsilon_2| + |x_2^* \varepsilon_1| \tag{1.6a}$$

$$\text{r.e.} = \left|\frac{x_1x_2 - x_1^*x_2^*}{x_1^*x_2^*}\right| \leq \left|\frac{\varepsilon_1}{x_1^*}\right| + \left|\frac{\varepsilon_2}{x_2^*}\right|. \tag{1.6b}$$

Thus the maximum relative error in the product of two approximate numbers will be less than or equal to the sum their individual relative errors. This can be generalised to n numbers.

(*iv*) Division

$$\frac{x_1}{x_2} = \frac{x_1^* + \varepsilon_1}{x_2^* + \varepsilon_2} = \frac{x_1^*\left(1 + \dfrac{\varepsilon_1}{x_1^*}\right)}{x_2^*\left(1 + \dfrac{\varepsilon_2}{x_2^*}\right)} = \frac{x_1^*}{x_2^*}\left(1 + \frac{\varepsilon_1}{x_1^*}\right)\left(1 + \frac{\varepsilon_2}{x_2^*}\right)^{-1}$$

$$= \frac{x_1^*}{x_2^*}\left(1 + \frac{\varepsilon_1}{x_1^*}\right)\left(1 - \frac{\varepsilon_2}{x_2^*}\right), \text{ neglecting } {\varepsilon_2}^2 \text{ and higher powers}$$

$$= \frac{x_1^*}{x_2^*}\left(1 + \frac{\varepsilon_1}{x_1^*} - \frac{\varepsilon_2}{x_2^*}\right), \text{ neglecting } \varepsilon_1\varepsilon_2 \text{ term.}$$

$$\text{a.e.} = \left|\frac{x_1}{x_2} - \frac{x_1^*}{x_2^*}\right| \leq \left|\frac{\varepsilon_1}{x_2^*}\right| + \left|\frac{\varepsilon_2 x_1^*}{{x_2^*}^2}\right|. \tag{1.7a}$$

$$\text{r.e.} = \left|\left(\frac{x_1}{x_2} - \frac{x_1^*}{x_2^*}\right) \div \frac{x_1^*}{x_2^*}\right| \leq \left|\frac{\varepsilon_1}{x_1^*}\right| + \left|\frac{\varepsilon_2}{x_2^*}\right|. \tag{1.7b}$$

Like multiplication, the relative error in the division of a number by another number cannot exceed the sum of their individual relative errors.

1.9 Computation of Errors Using Differentials

Let z be a function of two variables x and y defined as $z = f(x,\ y)$. If increments δx and δy are given to x and y respectively, then the corresponding increment δz in z is given by,

$$\delta z = f(x + \delta x,\ y + \delta y) - f(x,\ y).$$

Expanding the first term by Taylor's series (see Appendix A),

$$\delta z = \left[f(x,\ y) + \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}\delta x^2 + 2\frac{\partial^2 f}{\partial x \partial y}\delta x \cdot \delta y + \frac{\partial^2 f}{\partial y^2}\delta y^2\right) + \cdots\right] - f(x,\ y).$$

Neglecting higher powers of δx and δy and their products, assuming they are small, above may be written as,

$$\delta z \simeq \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y. \tag{1.8}$$

The error in arithmetic operations can be explained with the help of formula (1.8) considering δx and δy as errors in x and y respectively:

(*i*) & (*ii*) Addition/Subtraction

$$z = f(x,\ y) = x \pm y;\ \ \frac{\partial f}{\partial x} = 1,\ \frac{\partial f}{\partial y} = \pm 1.$$

a.e. $\quad = |\delta z| \leq |\delta x| + |\delta y|$

(*iii*) Multiplication

$$z = f(x,\ y) = xy;\ \ \frac{\partial f}{\partial x} = y,\ \frac{\partial f}{\partial y} = x.$$

a.e. $\quad = |\delta z| \leq |y\delta x| + |x\delta y|$ and r.e. $= \left|\frac{\delta z}{z}\right| \leq \left|\frac{\delta x}{x}\right| + \left|\frac{\delta y}{y}\right|.$

(*iv*) Division

$$z = f(x,\ y) = \frac{x}{y};\ \ \frac{\partial f}{\partial z} = \frac{1}{y},\ \frac{\partial f}{\partial y} = -\frac{x}{y^2}.$$

$$\delta z = \frac{\delta x}{y} - \frac{x\delta y}{y^2} \text{ or } \frac{\delta z}{z} = \frac{\delta x}{x} - \frac{\delta y}{y}$$

$$\text{r.e.} = \left|\frac{\delta z}{z}\right| \leq \left|\frac{\delta x}{x}\right| + \left|\frac{\delta y}{y}\right|.$$

Note: The analysis can be extended for n variables $x_1,\ x_2,\ \ldots x_n$.

If $\ z = f(x_1,\ x_2,\ \ldots x_n)$ then

$$\delta z\ = \frac{\partial f}{\partial x_1}\cdot\delta x_1 + \frac{\partial f}{\partial x_2}\cdot\delta x_2 + \ldots + \frac{\partial f}{\partial x_n}\cdot\delta x_n.$$

1.10 Errors in Evaluation of Some Standard Functions

Let y be a function of x defined by $y = f(x)$. Then for a small change δx in x, the corresponding change δy in y is given by,

$$\delta y = \frac{df}{dx} \cdot \delta x. \tag{1.9}$$

Let us apply formula (1.9) to compute the errors in some functions of one variable.

(*i*) Power Function $f(x) = x^a$

Differentiating $y = x^a$, we get using (1.9)

$$\text{a.e.} = |\delta y| = |ax^{a-1} \cdot \delta x|$$

$$\text{r.e.} = \left|\frac{\delta y}{y}\right| = \left|a \cdot \frac{\delta x}{x}\right| \leq |a| \left|\frac{\delta x}{x}\right|$$

$$\leq |a|. \text{ relative error in } x. \tag{1.10}$$

It means that the relative error in x^2 will be twice and in $\sqrt{x}$ it will be half that of the relative error in x. In x^{-1}, the r.e. will be same as in x and twice in x^{-2}.

(*ii*) Exponential Function $f(x) = a^x$

$$y = a^x; \ \frac{df}{dx} = a^x \cdot \ln a.$$

$$\text{a.e.} \quad = |a^x \ln a.\ \delta x| \text{ and r.e.} \quad = |\ln a.\delta x| = |\ln a||\delta x|. \tag{1.11}$$

(*iii*) Logarithmic Function $f(x) = \ln x$

$$y = \ln x; \ \frac{df}{dx} = \frac{1}{x}$$

$$\text{a.e.} \quad = |\delta y| = \left|\frac{\delta x}{x}\right|. \tag{1.12}$$

It shows that the absolute error in $\ln x$ will be same as the relative error in x.

(*iv*) Trigonometric Function $f(x) = \sin x (\text{or } \cos x)$

$$y = \sin x; \ \frac{df}{dx} = \cos x$$

$$\text{a.e.} \quad = |\delta y| = |\cos x \cdot \delta x| \leq |\cos x| \cdot |\delta x| \leq |\delta x|. \tag{1.13}$$

Since the value of $\cos x$ (or $\sin x$) does not exceed 1, the absolute error in $\sin x$(or $\cos x$) does not exceed the absolute error in x.

Example 1.1

Find the sum of the following 10 approximate numbers: 0.248, 1.1524, 31.3, 9.75, 74.2, 8.14, 0.0767, 1.00621, 1.000245, 14.8 and round the sum to one place of decimal. Also add up the rounding errors in all these numbers up to 4 decimal places only.

Solution These numbers have different rounding errors. The max. error is in numbers rounded to one decimal place which is in $\frac{1}{2} \times 10^{-1} = 0.05$ in each of the numbers 31.3, 74.2 and 14.8. The min. rounded error is $\frac{1}{2} \times 10^{-6} = 0.0000005$ in 1.000245. The final sum should be computed up to one decimal only. For this it may be sufficient to retain 2 decimal digits (or at the most 3 in a number since the numbers are only 10.

Number rounded to 3 decimals	*Rounding error up to 4 decimal*
0.248	0.0005
1.152	0.0005
31.3- -	0.05- -
9.75-	0.005
74.2- -	0.05- -
8.14-	0.005-
0.077	0.0005
1.006	0.0005
1.000	0.0000
14.8- -	0.05- -
141.673	0.1620

$$\text{sum} = 141.673 \simeq 141.7$$

Max. Rounding error in the sum $= 0.162 \simeq 0.16$ or $\simeq 0.2$.
Actual error in the sum due to rounding is

$$141.7 - 141.673 = 0.027 \text{ which is much less than } 0.2 \text{ (or } 0.16)$$

$$\text{r.e.} = \frac{0.027}{141.67} \simeq 0.0002$$

Show that it will be less than the max. r.e. of any of the given number.

Example 1.2

Two numbers x_1 and x_2 are given which are correct up to their last digit: $x_1 = 12.47$, $x_2 = 10.3$. Estimate the max. absolute error, relative error and percentage error in computing $x_1 - x_2$.

Solution Max rounding error in 12.47 is $\pm$ 0.005
Max rounding error in 10.3 is $\pm$ .05

$$x_1 - x_2 = 12.47 - 10.3 = 2.17$$

$$\text{a.e.} = 0.005 + 0.05 = 0.055$$

$$\text{r.e.} = \frac{0.055}{2.17} \simeq 0.022$$

$$\text{p.e.} = 0.022 \times 100 = 2.2\%.$$

Example 1.3

Multiply two numbers $x_1 = 2.47$ and $x_2 = 1.6$ and estimate the relative error in the product. The numbers are correct to their last digit. Also compute absolute error and percentage error.

Solution Max. rounding error in $x_1 = 2.47$ is $\pm$ 0.005
Max. rounding error in $x_2 = 1.6$ is $\pm$ 0.05

$$x_1 x_2 = 2.47 \times 1.6 = 3.952 \simeq 3.95$$

$$\text{r.e.} = \frac{0.005}{2.47} + \frac{0.05}{1.6}$$

$$\simeq 0.002 + 0.031$$

$$\simeq 0.033$$

$$\text{a.e.} = (\text{r.e.}) \times x_1 x_2$$

$$= 0.033 \times 3.95 = 0.130$$

$$\text{p.e.} = (\text{r.e.}) \times 100$$

$$= 0.033 \times 100 = 3.3\%$$

Example 1.4

Two approximate numbers x_1 and x_2 are given correct to their last digit, $x_1 = 5.16$ and $x_2 = 1.2$. Find the r.e., a.e. and p.e. in computing x_1/x_2.

Solution Max. error in 5.16 is $\pm$ 0.005
Max. error in 1.2 is $\pm$ 0.05

$$\frac{x_1}{x_2} = \frac{5.16}{12} = 4.3$$

$$\text{r.e.} = \frac{0.005}{5.16} + \frac{0.05}{1.2}$$

$$= 0.00097 + 0.0417$$

$$= 0.0427$$

$$\text{p.e.} = 0.0427 \times 100 = 4.27\%$$

$$\text{a.e.} = 0.0427 \times 4.3$$

$$\simeq 0.184.$$

Example 1.5

To what accuracy can we expect the number x to be correct if its logarithm $(\ln x)$ is read from a four-place log table for $x < 100$.

Solution When $y = \ln x$

$$\delta y = \frac{\delta x}{x} \text{ or } \delta x = x \cdot \delta y.$$

If $\ln x$ is correct up to 4 decimals the error in it will be 0.00005 $(= \delta y)$.

$$\delta x_{\max} = \delta y \cdot x = 0.00005 \times 100 = 0.005$$

For the error to be less than 0.005 the value of x may be expected to be correct up to 2 decimals.

Example 1.6

What will be the percentage error in the area of a rectangle if there is an error of 1% in the measurement of its sides?

Solution Let the accurate sides of the rectangle be x and y, then its area $A = xy$. If there is error δx in x and δy in y, the corresponding error in A in given by,

$$\delta A = x\delta y + y\delta x$$

or

$$\frac{\delta A}{A} = \frac{\delta x}{x} + \frac{\delta y}{y}$$

$$\text{Percentage error in } A = \frac{\delta \mathrm{A}}{\mathrm{A}} \times 100$$

Thus we get the percentage error in A as

$$\frac{\delta \mathrm{A}}{\mathrm{A}} \times 100 = \frac{\delta x}{x} \times 100 + \frac{\delta y}{y} \times 100$$

$$= 1 + 1 = 2\%.$$

Example 1.7

What will be the percentage error in the time period T of a pendulum where $\mathrm{T} = 2\pi\sqrt{\frac{l}{g}}$ if there is an error of 1% in l and 2% in g.

Solution $$\mathrm{T} = 2\pi\sqrt{\frac{l}{g}}$$

Taking log on both sides we get

$$\ln \mathrm{T} = \ln 2\pi + \frac{1}{2}[\ln l - \ln g]$$

Its differential will give,

$$\frac{1}{\mathrm{T}}\delta \mathrm{T} = 0 + \frac{1}{2}\left[\frac{\delta l}{l} - \frac{\delta g}{g}\right]$$

or $$\frac{\delta \mathrm{T}}{\mathrm{T}} \times 100 = \frac{1}{2}\left[\frac{\delta l}{l} \times 100 - \frac{\delta g}{g} \times 100\right]$$

Max. Percentage error in T is given by

$$\frac{\delta \mathrm{T}}{\mathrm{T}} \times 100 = \frac{1}{2}[(\pm 1) - (\pm 2)]$$

$$= \frac{1}{2}[3] = 1.5\%.$$

1.11 Truncation Error and Taylor's Theorem

Quite often a function or a formula is expressed in the form of a series which may contain finite or infinite number of terms. Depending on the accuracy warranted by the problem or due to practical considerations, this series is truncated consisting of a first few terms of

the original series, neglecting all the remaining terms. The difference between the original series and the truncated series is known as 'truncation error' or 'remainder'. Obviously, if the series is infinite, it should be convergent, otherwise truncation error will be infinite.

Let us consider the exponential function e^x which is expressed by an infinite series as,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \tag{1.14}$$

Suppose we want to evaluate e^x at $x = -0.1$ from (1.14), then

$$e^{-0.1} = 1 - 0.1 + 0.005 - 0.000167 + 0.0000042 \ldots . \tag{1.15}$$

Truncated up to four terms it gives,

$$e^{-0.1} = 1 - 0.1 + 0.005 - 0.000167 = 0.904843. \tag{1.16}$$

The infinite series (1.15) is a convergent series with alternating signs. Hence, if it is truncated up to four terms (1.16), the next neglected term will give the maximum error (truncation) in the truncated series. That means, the maximum error in the computation, $e^{-0.1} = 0.904843$ can be at the most 0.0000042, implying that our result is correct up to five places of decimal.

The infinite series (1.14) truncated to four terms may be written as,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}. \tag{1.17}$$

Again, let us evaluate e^x at $x = 0.1$ from the truncated series (1.17).

$$e^{0.1} = 1 + 0.1 + 0.005 + 0.000167 = 1.105167. \tag{1.18}$$

Although series (1.14) converges for all finite values of x, we can not ascertain the degree of accuracy of the result of (1.18), i.e., the magnitude of truncation error in the computation of $e^{0.1}$ from a truncated series (1.17). Secondly we may be interested to know the range of x, for which the truncated series (1.17) will provide the value of e^x which is correct up to a certain number of decimals, say for example, up to 4 places of decimal which means the truncation error is not greater than $\frac{1}{2} \times 10^{-4}$. The answer to these questions may be found in the Taylor's Theorem (formula) stated below:

Let $f(x)$ be a function of x, possessing derivatives of all orders up to $(n+1)$ in an interval I. If x_0 is a point in I, then for each x in I,

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^n(x_0) + R_{n+1}(x) \quad (1.19)$$

$$\text{where, } R_{n+1}(x) = \frac{(x-x_0)^{n+1}}{(n+1)!}f^{n+1}(\xi),\ \xi \text{ lies between } x_0 \text{ and } x. \quad (1.20)$$

The representation of the function $f(x)$ as power series in powers of $(x-x_0)$ as shown by (1.19) is known is Taylor's Theorem (formula) with the associated remainder term of order $(n+1)$, or error term, $R_{n+1}(x)$ as given by (1.20).

If a function is approximated by the first $(n+1)$ terms of (1.19), i.e., a polynomial of degree n, then the error in the polynomial approximation would be given by (1.20) which will be of order $(n+1)$ while the formula will be of order n.

If $R_{n+1}(x)$ tends to zero as $n \to \infty$ for all x in the interval I, then formula (1.19) may be represented by an infinite series which converges to $f(x)$, i.e.,

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \frac{(x-x_0)^3}{3!}f'''(x_0) + \cdots \quad (1.21)$$

The series given by (1.21) is known as Taylor's series which is also called Taylor's expansion of the function $f(x)$ about the point x_0.

Alternatively, if we put $x = x_0 + h,\ h > 0$, the Taylor's formula can be written as,

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f'''(x_0) + \cdots \frac{h^n}{n!}f^n(x_0) + R_{n+1} \quad (1.22)$$

where the remainder term,

$$R_{n+1}(h) = \frac{h^{n+1}}{(n+1)!}f^{n+1}(\xi), \qquad x_0 \le \xi \le x_0 + h. \quad (1.23)$$

Similarly, expansion for $f(x_0-h)$ can be written as,

$$f(x_0-h) = f(x_0) - hf'(x_0) + \frac{h^2}{2!}f''(x_0) - \frac{h^3}{3!}f'''(x_0) + \cdots (-1)^n\frac{h^n}{n!}f^n(x_0) + R_{n+1}, \quad (1.24)$$

with associated remainder term,

$$R_{n+1}(h) = (-1)^{n+1}\frac{h^{n+1}}{(n+1)!}f^{n+1}(\xi), \quad x_0 - h \le \xi \le x_0. \quad (1.25)$$

If point x_0 is taken at the origin in formula (1.19), it is called Maclaurin's formula which can be written as,

$$f(x) = f_0 + xf_0' + \frac{x^2}{2!}f_0'' + \frac{x^3}{3!}f_0''' + \cdots + \frac{x^n}{n!}f_0^n + R_{n+1}(x), \tag{1.26}$$

where $$R_{n+1}(x) = \frac{x^{n+1}}{(n+1)!}f^{n+1}(\xi), \qquad 0 \le \xi \le x, \tag{1.27}$$

and $f_0^k,\ k = 1(1)n$, denotes the kth derivative of $f(x)$ at $x = 0$.

When a function $f(x)$ is approximated by a polynomial $P_n(x)$ of degree n by Taylor's formula, the truncation error is given by the remainder term $R_{n+1}(x)$, i.e.,

$$f(x) = P_n(x) + R_{n+1}(x)$$

or $$R_{n+1}(x) = f(x) - P_n(x). \tag{1.28}$$

But we can not compute $R_{n+1}(x)$ as its value is dependent on ξ which is a point we have no knowledge of. Therefore, we compute the upper bound of $R_{n+1}(x)$ to give the magnitude of truncation error, denoted by $R(x)$, i.e.,

$$R(x) = \max |R_{n+1}(x)|, \text{ over domain of } x. \tag{1.29}$$

Now, we revert back to our problem of computing the truncation error in the computation of e^x, at $x = 0.1$, by the truncated series (1.17).

We see that the expansion (1.16) of e^x is a Taylor's (Maclaurin's) series for $f(x) = e^x$. The truncation error in (1.17) will be given by,

$$R(x) = \left|\frac{x^4}{4!}f^{iv}(\xi)\right|,\ 0 \le \xi \le 0.1.$$

$f^{\text{iv}}(x) = e^x$, has maximum value at $x = 0.1$, for $0 \le \xi \le 0.1$ giving $f^{\text{iv}}(0.1) = e^{0.1} = 1.1052$. Thus the truncation error is given by

$$R = \frac{(0.1)^4}{24} \times 1.1052 = 0.000004605 = 0.46 \times 10^{-5}.$$

As the truncation error is less than $\frac{1}{2} \times 10^{-5}$, the value of $e^{0.1}$ computed from (1.17) is correct up to five decimal places, i.e., $e^{0.1} = 1.10517$.

Suppose we want range of x for which the truncated formula provides values of e^x correct up to 4 places of decimal, then we solve,

$$|R(x)| \le \frac{1}{2} \times 10^{-4} \text{ or } \left|\frac{x^4}{24} \times f^{iv}(x)\right| \le \frac{1}{2} \times 10^{-4} \text{ giving}$$

$x^4 \leq 0.0012 \times e^{-x}$ or $x \leq 0.1861 \times e^{-x/4}$. $\quad \because e^{\xi}$ will be maximum at x.

To get an approximate value of x we can solve,

$$x = 0.1861\left(1 - \frac{x}{4}\right) \text{ giving } x \simeq 0.1778.$$

Thus for values approximately $0 \leq x \leq 0.18$, the truncation error will be less than 0.00005.

Example 1.8

Compute $(1.1)^{-1}$ from Taylor's expansion of the function $f(x) = \frac{1}{x}$ about $x_0 = 1$, truncated up to four terms. Also compute the truncation error and compare your result with the exact value.

Solution

$$f(x) = \frac{1}{x};\ f'(x) = -\frac{1}{x^2},\ f''(x) = +\frac{2}{x^3},\ f'''(x) = -\frac{6}{x^4},\ f^{iv}(x) = \frac{24}{x^5}.$$

$$f(1) = 1;\ f'(x) = -1,\ f''(1) = 2,\ f'''(1) = -6.$$

The truncated series about $x_0 = 1$, up to four terms is,

$$f(x) = f(1) + (x-1)f'(1) + \frac{(x-1)^2}{2}f''(1) + \frac{(x-1)^3}{6}f'''(1)$$

$$f(1.1) = 1 + 0.1 \times (-1) + \frac{0.01}{2} \times 2 + \frac{0.001}{6} \times (-6) = 0.909$$

The truncation error is given by,

$$R = \max\left|\frac{(x-1)^4}{24} \times f^{iv}(\xi)\right|,\ 1 \leq \xi \leq 1.1$$

$$= \frac{(1.1-1)^4}{24} \times 24 = 0.0001, \qquad \because f^{\text{iv}}(x) \text{ is maximum at } x = 1.$$

Exact value $= 1/1.1 = 0.909090$ (90 recurring)
Computed value $= 0.909$
Actual error $= 0.909090 - 0.909 = 0.00009$.
The actual error is less than the truncation error 0.0001.

Example 1.9

Using Taylor's expansion for $f(x) = \frac{1}{x}$ about $x_0 = 1$, truncated up to four terms, compute inverse of 2. Discuss the result.

Solution Taylor's series up to four terms is,

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \frac{(x-x_0)^3}{3!}f'''(x_0)$$

$$x_0 = 1;\ f(x_0) = 1,\ f'(x_0) = -1,\ f''(x_0) = 2,\ f'''(x_0) - 6.$$

$$f(2) = 1 + 1(-1) + \frac{1}{2} \times 2 - \frac{1}{6} \times 6$$

$$= 0$$

Truncation error is given by,

$$R(x) = \left| \frac{(x-1)^4}{4!} \times 24 \right|$$

For $x = 2, |R| = 1.$

Exact value $= \frac{1}{2} = 0.5$

Computed value $= 0$

Truncation error $= 1$

Thus we see that the truncation error is very large; the difference between the exact value and computed value could be as large as 1.0. Further, even if we take infinite number of terms, the sum will oscillate between 0 and 1 since $f(2) = 1 - 1 + 1 - 1 + \ldots$. The point x should be close to x_0, i.e. $(x - x_0)$ should not be too large.

Exercise 1

1.1 Find the range of number x, if it has been (*i*) rounded off to 3.14 (*ii*) chopped off to 3.14.

1.2 Express the number 0.007856 in floating point form; round the number to two significant digits and find the absolute error.

1.3 Let $x = 9.5$ be an approximate number which has an error of at most 5%. Find the range of the exact number.

1.4 Find the maximum value of the expression given below when all the numbers have been rounded,

$$x = \frac{1.25(4.0 - 2.25)}{10}$$

(Hint: For computing maximum value of x, take largest numerator and smallest denominator)

1.5 The diameter and height of a right circular cylinder are measured as 4 cm and 10 cm. respectively. If the possible error in each measurement is 0.1 cm, find the maximum absolute error in its volume. $(\pi = 3.14)$.

1.6 Obtain a quadratic approximation for e^x near $x = 0$ by truncating the Taylor's series. Use the approximation to find the range of x so that the error does not exceed 0.005 (or approximation computes values correct up to 2 decimal places).
Compute your answer correct up to one place of decimal only.

1.7 If $y = x^{1/3}$, show that the relative error in y will be $\frac{1}{3}$rd of the relative error in x. Hence compute $(1003)^{1/3}$.
[Hint: Take $x = 1000,\ \delta x = 3$].

References and Some Useful Related Books/Papers

1. Hartree, D.R., *Numerical Analysis*, Oxford University Press.
2. Hildebrand, F.B., *Introduction to Numerical Analysis*, Tata McGraw-Hill.
3. Scarborough, J.B., *Numerical Mathematical Analysis*, Oxford Book Company.

2

Linear Equations and Eigenvalue Problem

2.1 Introduction

Let us consider the following system of equations,

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots. + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots. + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots. + a_{3n}x_n &= b_3 \\ \vdots \qquad \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots. + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{2.1}$$

The system of equations given by (2.1) is a set of n algebraic equations which are linear in $x_1,\ x_2 \ldots.\ x_n$; while the values of a_{ij} and $b_i,\ i = 1(1)n,\ j = 1(1)n$ are prescribed, the values of the unknowns $x_1,\ x_2,\ \ldots.\ x_n$ are to be determined such that all the n equations are satisfied simultaneously.

The system of equations (2.1) can be expressed in matrix form as,

$$\mathrm{A}\boldsymbol{x} = \boldsymbol{b}, \tag{2.2}$$

where $$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \tag{2.3}$$

$$\boldsymbol{x}^{\mathrm{T}} = (x_1 \;\; x_2 \;\; x_3 \;\; .... \;\; x_n), \tag{2.3a}$$

$$\boldsymbol{b}^{\mathrm{T}} = (b_1 \;\; b_2 \;\; b_3 \;\; .... \;\; b_n), \tag{2.3b}$$

the superscript T denotes transpose of a matrix. We assume that not all the elements of b are zero; that is, $\boldsymbol{b}$ is not a 'null' vector ($\boldsymbol{b} \neq 0$).

The matrix A is called coefficient matrix; the column vector (matrix) $\boldsymbol{x}$ which is to be determined, is the solution vector and $\boldsymbol{b}$ is the given right side.

2.2 Ill-conditioned Equations

When a minor change in the given value (s) results in a drastic change in the true solution of a system, it is called 'ill-conditioned'. Such a set of simultaneous equations is very sensitive to small errors.

2.3 Inconsistency of Equations

When a system of equations provide contradicting solutions or the equations themselves are self-contradictory we say that the equations are inconsistent.

The problem of inconsistency may arise when the system is over-determined, i.e., when there are more equations than the number of unknowns. For example let us consider following three equations in two unknowns;

$$2x_1 + x_2 = 4,\; 3x_1 - x_2 = 1,\; x_1 + x_2 = 7.$$

From the first two equations we get $x_1 = 1,\; x_2 = 2$ but they do not satisfy the third. Similarly, the second and third equations give $x_1 = 2,\; x_2 = 5$ which do not satisfy the first. Thus, there is no solution which can satisfy all the three equations simultaneously; hence the equations are 'inconsistent'. If however, there exists a solution which satisfies all the given equations, then the system would be called 'consistent'. The two equations $x_1 + x_2 = 1$ and $x_1 + x_2 = 2$ are inconsistent.

2.4 Linear Dependence

If some of the equations in a system are linearly related, the equations of such a system are said to be 'linearly dependent'; otherwise linearly independent. That means, in a linearly dependent system, at least one equation can be expressed as linear sum of some other equations. For example, consider the following equations,

$$2x_1 + 3x_2 + x_3 = 4,\; 5x_1 + 2x_2 + 2x_3 = 5,\; 3x_1 + 10x_2 + 2x_3 = 11.$$

We see that the third equation can be expressed as, four times the first minus the second. Or, first equation can be obtained by adding second and third equations and then dividing by four. As there exists a linear relation between the equations, they are not linearly independent. Effectively, there are only two linearly independent equations as third can be expressed in terms of the other two. In such cases an arbitrary value can be assigned to one of the variables, say, $x_3 = k$; then values of x_1 and x_2 can be computed in terms of k. Thus, the system will have infinite number of solutions depending on k.

It is easy to visualise that since the rows of the coefficient matrix of linearly dependent system are linearly related, its determinant will vanish. In order to know as how many of rows are linearly independent, we are lead to the notion of 'rank' of a matrix. A square matrix whose determinant vanishes is called 'singular'; otherwise 'non-singular' or 'regular'.

2.5 Rank of a Matrix

Let A be a $m \times n$ matrix. From matrix A we can form square submatrices by removing some of its rows and/or some columns including the matrix A itself (when $m = n$). The matrix A is said to have rank k if it has at least one submatrix of order k which is non-singular while all submatrices of order greater than k are singular. It is denoted as,

rank (A) or $r(\mathrm{A}) = \mathrm{k}$.

It may be noted that the number linearly independent equations in a system of equations is equal to the rank of its coefficient matrix.

2.6 Augmented Matrix

Referring (2.3) and (2.3b), when the matrix A is augmented by adjoining vector b as $(n+1)^{\text{th}}$ column, we call matrix A as augmented matrix and is denoted as,

$$\text{aug } (\mathrm{A}/\boldsymbol{b}) = \left[\begin{array}{ccccc:c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{3n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right]. \tag{2.4}$$

As regards existence/uniqueness of the solution of the system of equations (2.2), following three cases arise; for brevity/clarity we have used aug A for aug $(\mathrm{A}/\boldsymbol{b})$:

(*i*) $r(\mathrm{A}) = r\ (\text{aug A}) = \text{number of unknowns}$
The solution is unique. The number of linearly independent equations is same as the number of unknowns.

(*ii*) $r(\mathrm{A}) = r\ (\text{aug A}) < \text{number of unknowns}$
There are infinite number of solutions. The number of linearly independent equations is less than the number of unknowns.

(*iii*) $r(\mathrm{A}) < r\ (\text{aug A})$
There exists no solution. There are more linearly independent equations than there are unknowns.

Note: In the ensuing discussions we will assume that a unique solution exists of the system (2.2) which means A is regular and its inverse exists.

2.7 Methodology for Computing A^{-1} by Solving $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$

Let A be a regular (non-singular) matrix of order n and its inverse denoted as

$$\mathrm{A}^{-1} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \cdots & \alpha_{2n} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \cdots & \alpha_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \alpha_{n3} & \cdots & \alpha_{nn} \end{bmatrix}.$$

Then from $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$, i.e., from (2.2) we can write,

$$\boldsymbol{x} = \mathrm{A}^{-1}\boldsymbol{b} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \cdots & \alpha_{2n} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \cdots & \alpha_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \alpha_{n3} & \cdots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

We observe that if $\boldsymbol{b}$ is chosen to be such that $b_1 = 1$ and $b_2 = b_3 = \ldots. = b_n = 0$, then $\boldsymbol{x}$ is simply the first column of A^{-1}. Thus if we solve (2.2) by taking $\boldsymbol{b}$ as the first column of a unit/Identity matrix, the solution vector will give the first column of A^{-1}. Similarly, if we solve $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$ by choosing vector $\boldsymbol{b}$ as , i.e., $b_2 = 1$ and $b_1 = b_3 = b_4 = \ldots = b_n = 0$, then,

the solution vector $\boldsymbol{x}$ will provide the second column of A^{-1}; and so on. Thus solving the system of equations $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$, n times, by taking $\boldsymbol{b}$ to be the various columns of a unit matrix will provide respective columns of A^{-1}.

2.8 Cramer's Rule

Denoting the determinant of a square matrix A by $|\mathrm{A}|$ or $\det \mathrm{A}$, the solution to the system of equations (2.2) by Cramer's rule, is given as,

$$x_j = \frac{|\mathrm{A}_j|}{|\mathrm{A}|},\ j = 1(1)n \tag{2.5}$$

where $|\mathrm{A}_j|$ denotes the determinant of the matrix A with its jth column replaced by the right side $\boldsymbol{b}$. Since (2.5) involves evaluation of determinants, the method is not suitable for larger systems.

2.9 Inverse of Matrix by Cofactors

The adjoint of a matrix A is defined as,

$$\text{adj A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{21} & \mathrm{A}_{31} & \cdots & \mathrm{A}_{n1} \\ \mathrm{A}_{12} & \mathrm{A}_{22} & \mathrm{A}_{32} & \cdots & \mathrm{A}_{n2} \\ \mathrm{A}_{13} & \mathrm{A}_{23} & \mathrm{A}_{33} & \cdots & \mathrm{A}_{n3} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathrm{A}_{1n} & \mathrm{A}_{2n} & \mathrm{A}_{3n} & \cdots & \mathrm{A}_{nn} \end{bmatrix} \tag{2.6}$$

where A_{ij} is the cofactor of a_{ij}. We know that if M_{ij} is the minor of a_{ij}, then $\mathrm{A}_{ij} = (-1)^{i+j}\mathrm{M}_{ij}$. The inverse of matrix A is given by,

$$\mathrm{A}^{-1} = \frac{1}{|\mathrm{A}|}.\ \text{adj A}. \tag{2.7}$$

After computing A^{-1} the solution of $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$ may be obtained from $\boldsymbol{x} = \mathrm{A}^{-1}\boldsymbol{b}$. But again, the evaluation minors is a very time-consuming process. Therefore this method is also of limited use and is not suitable for larger systems.

The method described in Sec 2.7 will be taken up in detail in Sec 2.17 along with other computer-oriented methods. Now, let us present a brief review of some special square matrices and properties associated with them.

2.10 Definitions of Some Matrices

(*i*) Symmetric Matrix
A matrix A is called 'symmetric' if

$$\mathrm{A} = \mathrm{A}^{\mathrm{T}} \text{ or } [a_{ij}] = [a_{ji}]$$

Also if $\mathrm{A} = -\mathrm{A}^{\mathrm{T}}$, then it is called skew-symmetric.

(*ii*) Diagonal Matrix
A matrix A is called 'diagonal' if all its off-diagonal elements are zero, i.e.,

$$a_{ij} = 0,\ i \neq j.$$

When all the diagonal elements are unity, it is called unit or Identity matrix and when the diagonal elements are also zero it is known as Null matrix. The diagonal matrix is generally denoted by symbol D, a unit/Identity matrix by I while a Null matrix by letter O and a null vector by $\mathbf{0}$.

(*iii*) Lower Triangular/Upper Triangular Matrices
When all the elements in a matrix above its main diagonal are zero, it is called Lower Triangular; on the other hand if all the elements below the main diagonal are zero, the matrix is called Upper Triangular. The forms of 4×4 Lower Triangular matrix (L) and Upper Triangular matrix (U) are shown below:

$$\mathrm{L} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix},\ \mathrm{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

If the diagonal elements are 1's, the matrices are called unit lower triangular and unit upper triangular. When diagonal elements are zero, they are called 'strictly' lower and upper triangular.

(*iv*) Tri-diagonal Matrix/Band matrices
A matrix a is called tri-diagonal if, for any i,

$$a_{ij} = 0,\ j < i-1 \text{ and } j > i+1.$$

Thus, there are at the most three non-zero terms in each row$-$diagonal term, one before the diagonal term and one after. The first and last rows have two terms only. A (5×5) tri-diagonal matrix will look like the following

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 \\ 0 & 0 & a_{43} & a_{44} & a_{45} \\ 0 & 0 & 0 & a_{54} & a_{55} \end{bmatrix}.$$

Symbolically, it is also represented as,

$$A = \begin{bmatrix} & & \circ \\ & & \\ & & \\ & & \\ \circ & & \end{bmatrix}$$

The big zeros show the zero elements and the straight lines along the main diagonal and paralled to it (sub-diagonal and super-diagonal) show the elements which may or may not be zero. We may also have matrices when there are at the most five non-zero terms in any row in the above fashion. Such matrices are called 'Penta-diagonal'. In general, when there are non-zero terms along the main diagonal, sub-diagonal and super-diagonal while all the remaining terms are zero in a matrix it is called 'Band-matrix'. These type of matrices arise in solving ordinary and partial differential equations.

(*v*) Sparse Matrix

A matrix is called 'sparse' if most of its elements are zero.

(*vi*) Orthogonal Matrix

Matrix A is said to be orthogonal, if

$$A^T = A^{-1}, \text{ implying } AA^T = A^TA = I.$$

For example, following matrix is orthogonal

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

(*vii*) Hermitian Matrix

If elements of a matrix A are complex numbers and $\overline{A}$ represents complex conjugate of A, i.e., when all the elements of A are replaced by their complex conjugates, then

A is Hermitian matrix, if

$(\overline{\mathrm{A}})^{\mathrm{T}} = \mathrm{A}^* = \mathrm{A}$, where $*$ denotes conjugation and transposition (tranjugation) of matrix A.

It is equivalent to symmetric matrix if A is real.

(*viii*) Unitary Matrix
Matrix A is called Unitary, if

$$\mathrm{A}^* = \mathrm{A}^{-1}$$

It is equivalent to orthogonal if A is real.

(*ix*) Involutary Matrix

If $\mathrm{A} = \mathrm{A}^{-1}$, implying $\mathrm{A}^2 = \mathrm{I}$.

(*x*) Positive Definite Matrix
Matrix A is said to be positive definite if for any non-zero column vector $\boldsymbol{x}$,

$$\boldsymbol{x}^{\mathrm{T}}\mathrm{A}\boldsymbol{x} > 0.$$

Sometimes we use condition $\boldsymbol{x}^{\mathrm{T}}\mathrm{A}\boldsymbol{x} \geq 0$ for positive definite and $\boldsymbol{x}^{\mathrm{T}}\mathrm{A}\boldsymbol{x} > 0$ for strictly positive definite matrix. If all the leading principal minors of a matrix are positive (+ve), the matrix will be positive definite.

2.11 Properties of Matrices

We discuss below some of the properties of $(n \times n)$ square matrices. Let it be reminded that the determinant of a matrix, say, A is denoted by $|\mathrm{A}|$.

(*i*) $|\mathrm{A}| = |\mathrm{A}^{\mathrm{T}}|$

(*ii*) $|\mathrm{A} \cdot \mathrm{B}| = |\mathrm{A}| \cdot |\mathrm{B}| = |\mathrm{B} \cdot \mathrm{A}|$.
This can be generalised to any number of matrices.

(*iii*) The determinant of a Lower Triangular/Upper Triangular matrix is equal to product of its diagonal elements. That is, if L is a Lower Triangular matrix with its diagonal elements as $l_{11},\ l_{22} \ldots l_{nn}$, then

$$|\mathrm{L}| = l_{11} l_{22} \ldots l_{nn}.$$

Similarly if U is an Upper Triangular matrix with its diagonal elements as, $u_{11},\ u_{22},\ \ldots u_{nn}$, then

$$|\mathrm{U}| = u_{11} u_{22} \ldots u_{nn}.$$

(*iv*) Transpose of product of matrices is equal to the product of their transposes taken in reverse order, i.e.,

$$(A \cdot B)^T = B^T \cdot A^T$$

$$(A \cdot B \cdot C)^T = C^T B^T A^T \text{ etc.}$$

(*v*) Inverse of product of matrices is equal to the product of their inverses taken in reverse order, i.e.,

$$(A \cdot B)^{-1} = B^{-1} A^{-1}$$

$$(A \cdot B \cdot C)^{-1} = C^{-1} B^{-1} A^{-1} \text{ etc.}$$

(*vi*) Inverse of a transpose of a matrix is the same as transpose of its inverse, i.e.,

$$(A^T)^{-1} = (A^{-1})^T.$$

This can be verified by multiplying both sides by A^T and using property (*iv*) which gives $I = I$.

(*vii*) If A is symmetric then A^{-1} is also symmetric, for if $A = A^T$, then

$$A^{-1} = (A^T)^{-1} = (A^{-1})^T \text{ from property (vi).}$$

(*viii*) If A is an orthogonal matrix, then

$$(A^T)^{-1} = A, \text{ since } A^T = A^{-1} \text{ and } (A^{-1})^{-1} = A.$$

(*ix*) Product of Lower Triangular matrices is also a Lower Triangular matrix. Similarly product of upper triangular matrices is an Upper Triangular matrix.

(*x*) Inverse of a lower triangular matrix is also a lower triangular and similarly the inverse of an upper triangular matrix is also an upper triangular matrix.

2.12 Elementary Transformations

The elementary operations on a matrix like, interchanging its rows (or columns) or adding a multiple of a row (or column) to another are known as elementary transformation on the matrix. The desired elementary row transformations can be performed on a matrix by premultiplying it by a unit matrix which has undergone the same transformation. Similarly elementary column operations can be performed by postmultiplying the matrix by a unit matrix with same operations.

Suppose we want to interchange the first row of a 4×4 matrix A by its third row, we can choose a 4×4 unit matrix with its first and third rows interchanged, say, I_{13}, where

$$I_{13} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The matrix obtained after pre-multiplication of A by I_{13}, i.e., $I_{13}A$ will be matrix A with its first and third rows interchanged. Similarly the matrix obtained after postmultiplication of A by I_{13}, i.e., AI_{13} will be matrix A with its first and third columns interchanged. The matrix which is used for interchanging rows (or columns) is known as permutation matrix.

Now suppose we want to add p times of the second row of A to its third row and q times of the second row to its fourth row. This can be achieved by elementary transformation by choosing a matrix I_{2R} (say) obtained by doing the same operations on a unit matrix, i.e.,

$$I_{2R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & p & 1 & 0 \\ 0 & q & 0 & 1 \end{bmatrix}.$$

Then we will have,

$$I_{2R}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} + pa_{21} & a_{32} + pa_{22} & a_{33} + pa_{23} & a_{34} + pa_{24} \\ a_{41} + qa_{21} & a_{42} + qa_{22} & a_{43} + qa_{23} & a_{44} + qa_{24} \end{bmatrix}.$$

Similarly if we want to add p times of the second column of A to its third column and q times of its second column to the fourth column, then the desired transformation matrix will be,

$$I_{2C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & p & q \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The matrix $\mathrm{AI}_{2\mathrm{C}}$ will be the matrix with desired result, i.e.,

$$\mathrm{AI}_{2\mathrm{C}} = \begin{bmatrix} a_{11} & a_{12} & a_{13}+pa_{12} & a_{14}+qa_{12} \\ a_{21} & a_{22} & a_{23}+pa_{22} & a_{24}+qa_{22} \\ a_{31} & a_{32} & a_{33}+pa_{32} & a_{34}+qa_{32} \\ a_{41} & a_{42} & a_{43}+pa_{42} & a_{44}+qa_{42} \end{bmatrix}.$$

2.13 Methods for Solving Equations (Direct Methods)

There are two approaches for solving the system of equations $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$, known as (1) Direct Methods and (2) Iterative Methods. In the 'Direct methods' the solution is obtained in some definite number of steps while in the 'iterative methods' the process is started from an initial guess (usually taking $\boldsymbol{x} = 0$) which is improved in an interative manner until the solution agrees in two successive iterations within the desired accuracy. Thus the number of steps (iterations) in an iterative method can not be predicted beforehand.

Let us first discuss the Direct Methods.

2.13.1 Gaussian elimination method (Basic)

We describe the Gaussian Elimination method, in its basic form by taking a 4×4 system of equations, i.e.,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}. \tag{2.8}$$

Stage 1. It consists of three steps:

(i) Take a multiplier $m_{21} = -\dfrac{a_{21}}{a_{11}}$.

Multiply 1^{st} row of (2.8) by m_{21} and add to the 2^{nd}.

i.e., $\mathrm{R}_2 \leftarrow \mathrm{R}_2 + m_{21}\mathrm{R}_1$.

(ii) Take a multiplier $m_{31} = -\dfrac{a_{31}}{a_{11}}$.

Multiply 1^{st} row by m_{31} and add to 3^{rd}.

i.e., $\mathrm{R}_3 \leftarrow \mathrm{R}_3 + m_{31} \cdot \mathrm{R}_1$.

(*iii*) Take a multiplier $m_{41} = -\dfrac{a_{41}}{a_{11}}$.

Multiply 1^{st} row by m_{41} and add to 4^{th}

i.e., $\mathrm{R}_4 \leftarrow \mathrm{R}_4 + m_{41} \cdot \mathrm{R}_1$.

It may be noted that each time, first row is being multiplied by appropriate multiplier and added to different rows. The row which is being multiplied (first in this case) is known as 'pivotal' row and the element a_{11}, the divisor, is called the 'pivotal element' or simply 'pivot'. Next, we need to do operations of multiplications and additions from second column onwards since elements in the first column are made zero.

After execution of the first stage, the system will have the following form,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} \\ 0 & a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(1)} \\ b_4^{(1)} \end{bmatrix}. \tag{2.9}$$

The elements with superscript 1 indicate that they have changed after stage 1.

Stage 2. We omit the first row and first column of the coefficient matrix in (2.9), thereby dealing with a 3×3 system of equations. Thus stage 2 has two steps.

(*i*) Take multiplier $m_{32} = -\dfrac{a_{32}^{(1)}}{a_{22}^{(1)}}$.

Multiply 2^{nd} row of (2.9) by m_{32} and add to the 3^{rd}.

i.e., $\mathrm{R}_3 \leftarrow \mathrm{R}_3 + m_{32} \cdot \mathrm{R}_2$.

(*ii*) Take multiplier $m_{42} = -\dfrac{a_{42}^{(1)}}{a_{22}^{(1)}}$.

Multiply 2^{nd} row by m_{42} and add to 4^{th}.

i.e., $\mathrm{R}_4 \leftarrow \mathrm{R}_4 + m_{42} \cdot \mathrm{R}_2$

Here second row is the pivotal row and $a_{22}^{(1)}$, the pivot.

Note that we have to do the operations of multiplications and additions from the third column onwards, since the elements in the first columns are zero which are not affected by these operations and elements of second column are reduced to zero due to choice of multipliers.

After Stage 2, the system (2.9) will assume the following form,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(2)} \\ b_4^{(2)} \end{bmatrix}. \tag{2.10}$$

Here, the elements which have changed during Stage 2 are shown with superscript 2.

Stage 3. We ignore first two rows and first two columns in (2.10) and deal with a 2×2 system having superscript 2. This stage has only one step.

Take the multiplier $m_{43} = -\dfrac{a_{43}^{(2)}}{a_{33}^{(2)}}$.

Multiply the 3rd row of (2.10) by m_{43} and add to 4th

i.e., $\quad R_4 \leftarrow R_4 + m_{43} \cdot R_3$.

This operation has to be made only on the fourth column and on the right side, of course. Third row is pivotal row and $a_{33}^{(2)}$ is the pivot. After stage 3, the system (2.10) finally reduces to the following form,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(2)} \\ b_4^{(3)} \end{bmatrix} \tag{2.11}$$

We see that the coefficient matrix in (2.11) has been reduced to an Upper Triangular matrix. This process of reduction of the coefficient matrix to an Upper Triangular form is known as 'Gaussian Elimination' or 'pivotal condensation'.

Having got the final reduced form (2.11), the solution is obtained by the process of 'back-substitution'. That is, we compute the value of x_4 from the last equation as,

$$x_4 = b_4^{(3)} / a_{44}^{(3)}.$$

Substituting the value of x_4 in the third equation, gives the value of x_3. After putting the values of x_3 and x_4 in the second equation we get x_2. And finally the value of x_1 is obtained from first equation, using the values of x_2, x_3 and x_4.

Caution: Some caution has to be exercised while using the basic Gaussian Elimination method in that the pivotal element should not be zero (or very small) since it will result in infinitely large multipliers. In order to avoid it, the pivotal row should be exchanged with the next row such that the pivotal element does not remain too small.

Note for Computer Algorithm: We see that the zeros once produced do not play any role in the future computations. Therefore, in order to save computer space, the multipliers may be stored in their place. For example, the multipliers m_{21}, m_{31} and m_{41} may be stored in the space previously occupied by a_{21}, a_{31} and a_{41} respectively. Similarly, m_{32} and m_{42} may be stored in the space previously occupied by $a_{32}^{(1)}$ and $a_{42}^{(1)}$ respectively. Finally, m_{43} may be stored in the space earlier occupied by $a_{43}^{(2)}$.

Thus in a compact form, we can store the coefficient matrix and the multipliers inside the computer in a 4×4 array in the following manner,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ m_{21} & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ m_{31} & m_{32} & a_{33}^{(2)} & a_{34}^{(2)} \\ m_{41} & m_{42} & m_{43} & a_{44}^{(3)} \end{bmatrix}. \tag{2.12}$$

The method described above for (4×4) system can be easily generalised for $(n \times n)$ system. There will be $(n-1)$ stages to reduce the original coefficient matrix to upper triangular form. At the kth stage, $k = 1(1)n-1$, we compute in the following manner:

For the rows, $i = k+1(1)n$,

$$\begin{aligned} m_{ik} &= -{a_{ik}}^{(k-1)} / {a_{kk}}^{(k-1)} \\ {a_{ij}}^{(k)} &= {a_{ij}}^{(k-1)} + m_{ik} {a_{kj}}^{(k-1)}, \ j = k+1(1)n, \\ {b_i}^{(k)} &= {b_i}^{(k-1)} + m_{ik} {b_k}^{(k-1)}, \\ {a_{ik}}^{(k)} &= m_{ik} (\text{optional}). \end{aligned}$$

The superscript zero corresponds to original values of A and $\boldsymbol{b}$. The basic concept in the method is that the elements once made zero, remain zero throughout the subsequent operations.

2.13.2 Gaussian elimination (with row interchanges)

Gaussian Elimination with row interchanges, called 'partial pivoting' ensures that all the multipliers are less than 1 in absolute value. This is achieved as described below:

At the kth stage, all the elements in the k^{th} column, i.e., $a_{ik},\ i = k(1)n$ are scanned and the numerically largest element is selected; suppose it is a_{pk}, i.e., element in the p^{th} row. Then kth row is interchanged with the p^{th} row so that the pivotal element is numerically largest, rendering the multipliers less than 1 in absolute value. The objective of affecting the interchanges is to reduce the rounding errors in the subsequent computations. When a number is rounded it has certain rounding error and when this number is multiplied by a multiplier greater than 1, the rounding error also increases but if the number is multiplied by a multiplier smaller than 1, the rounding error does not increase. Moreover, the partial pivoting automatically excludes the possibility of the pivotal element being zero.

A further modification to the method is 'complete pivoting' in that all the elements $a_{i,j}^{(k-1)},\ i = k(1)n, j = k(1)n$ are scanned and numerically largest element $a_{pq}^{(k-1)}$, is selected; then kth row is interchanged with the p^{th} row and k^{th} column with the q^{th} column. In this manner element $a_{pq}^{(k-1)}$ is brought at the $(k,\ k)$ position. Obviously, it requires recording of interchanges of rows and particularly of columns at each stage; however, it may not be worth except in some rare cases.

Example 2.1

Solve the following system of simultaneous equations by Gaussian elimination method,

$$3x_1 + 2x_2 + x_3 - 4x_4 = 5$$

$$x_1 - 5x_2 + 2x_3 + x_4 = 18$$

$$5x_1 + x_2 - 3x_3 + 2x_4 = -4$$

$$2x_1 + 3x_2 + x_3 + 5x_4 = 11$$

Compute up to 3 places of decimal without using fractions. Round the final answer to two decimals.

Solution Since vector x plays no part in computations we may work out with numbers only.

$$\begin{array}{cccc|c} x_1 & x_2 & x_3 & x_4 & \boldsymbol{b} \\ \hline 3 & 2 & 1 & -4 & 5 \\ 1 & -5 & 2 & 1 & 18 \\ 5 & 1 & -3 & 2 & -4 \\ 2 & 3 & 1 & 5 & 11 \end{array}$$

Forward Elimination:

$$m_{21} = -\frac{1}{3} = -0.333; \qquad \mathrm{R}_2 \leftarrow \mathrm{R}_2 - 0.333\mathrm{R}_1$$

$$m_{31} = -\frac{5}{3} = -1.667; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 1.667\mathrm{R}_1$$

$$m_{41} = -\frac{2}{3} = -0.667; \qquad \mathrm{R}_4 \leftarrow \mathrm{R}_4 - 0.667\mathrm{R}_1$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc:c} 3 & 2 & 1 & -4 & 5 \\ 0 & -5.666 & 1.667 & 2.332 & 16.335 \\ 0 & -2.334 & -4.667 & 8.668 & -12.335 \\ 0 & 1.666 & 0.333 & 7.668 & 7.665 \end{array}\right] \end{array}$$

$$m_{32} = -\frac{2.334}{5.666} = -0.412; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 0.412\mathrm{R}_2$$

$$m_{42} = \frac{1.666}{5.666} = 0.294; \qquad \mathrm{R}_4 \leftarrow \mathrm{R}_4 + 0.294\mathrm{R}_2$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc:c} 3 & 2 & 1 & -4 & 5 \\ 0 & -5.666 & 1.667 & 2.332 & 16.335 \\ 0 & 0 & -5.354 & 7.707 & -19.065 \\ 0 & 0 & 0.823 & 8.354 & 12.467 \end{array}\right] \end{array}$$

$$m_{43} = \frac{0.823}{5.354} = 0.154; \qquad \mathrm{R}_4 \leftarrow \mathrm{R}_4 + 0.154\mathrm{R}_3$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc:c} 3 & 2 & 1 & -4 & 5 \\ 0 & -5.666 & 1.667 & 2.335 & 16.335 \\ 0 & 0 & -5.354 & 7.707 & -19.065 \\ 0 & 0 & 0 & 9.541 & 9.531 \end{array}\right] \end{array}$$

Back Substitution:

$$x_4 = \frac{9.531}{9.541} = 0.999$$

$$x_3 = (-19.065 - 7.707 \times 0.999)/(-5.354) = 4.999$$

$$x_2 = (16.335 - 2.332 \times 0.999 - 1.667 \times 4.999)/(-5.666) = -1.001$$

$$x_1 = (5 + 4 \times 0.999 - 1 \times 4.999 - 2 \times (-1.001)/3 = 2.000$$

Final answer rounded to two places of decimal is:

$$x_1 = 2.00,\ x_2 = -1.00,\ x_3 = 5.00 \ \text{ and } \ x_4 = 1.00.$$

(exact answer is: $x_1 = 2,\ x_2 = -1,\ x_3 = 5,\ x_4 = 1$)

Note: If we compute $a_{21} \leftarrow a_{21} + m_{21} \times a_{11}$, we get $a_{21} = 1 - 0.333 \times 3 = 0.0001$, which is not zero although we have assumed it to be so. Similar argument holds for the other zeros produced. Thus the equations are satisfied approximately. If higher accuracy is required, we may work with more decimal places and in double precision on a computer, if needed.

Example 2.2

Solve the following system of simultaneous equations by Gaussian elimination method,

$$4.3x_1 - 3.5x_1 - 1.2x_3 = 10.90$$

$$18.4x_1 + 2.1x_2 - x_3 = 7.80$$

$$7.2x_1 + 1.8x_2 + 3.4x_3 = 23.22$$

Perform computations up to 2 decimal places. Round the final answer to one decimal.

Solution

$$\begin{array}{cccc} x_1 & x_2 & x_3 & \boldsymbol{b} \\ \left[\begin{array}{ccc:c} 4.3 & -3.5 & -1.2 & 10.9 \\ 18.4 & 2.1 & -1.0 & 7.8 \\ 7.2 & 1.8 & 3.4 & 23.22 \end{array}\right] \end{array}$$

$$m_{21} = -\frac{18.4}{4.3} = -4.28; \qquad \mathrm{R}_2 \leftarrow \mathrm{R}_2 - 4.28\mathrm{R}_1.$$

$$m_{31} = -\frac{7.2}{4.3} = -1.67; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 1.67\mathrm{R}_1.$$

$$\begin{array}{cccc} x_1 & x_2 & x_3 & \boldsymbol{b} \\ \left[\begin{array}{ccc:c} 4.3 & -3.5 & -1.2 & 10.90 \\ 0 & 17.08 & 4.14 & -38.85 \\ 0 & 7.64 & 5.40 & 5.02 \end{array}\right] \end{array}$$

$$m_{32} = -\frac{7.64}{17.08} = -0.45; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 0.45\mathrm{R}_2.$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{ccc:c} 4.3 & -3.5 & -1.2 & 10.90 \\ 0 & 17.08 & 4.14 & -38.85 \\ 0 & 0 & 3.54 & 22.50 \end{array}\right] \end{array}$$

$x_3 = 6.36;\ x_2 = -3.82;\ x_1 = 1.20.$

Rounding to one decimal,

$$x_1 = 1.2,\ x_2 = -3.8,\ x_3 = 6.4$$

(exact answer is $x_1 = 1.2,\ x_2 = -3.8,\ x_3 = 6.3$)

Example 2.3

Solve the following system of simultaneous equations by Gaussian elimination method:

$$0.12x_2 + 0.15x_3 = 0.33$$
$$0.56x_1 + 0.40x_2 - 0.18x_3 = 2.34$$
$$0.20x_1 + 0.71x_2 + x_3 = 2.04$$

Compute up to 3 places of decimal and round the final answer to two decimal places.

Solution

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{ccc:c} 0 & 0.12 & 0.15 & 0.33 \\ 0.56 & 0.40 & -0.18 & 2.34 \\ 0.20 & 0.71 & 1.0 & 2.04 \end{array}\right] \end{array}$$

Here first equation can not be treated as pivotal equation as $m_{21} \to \infty$. Therefore, we interchange it with the next row which has non-zero as its first element. Thus we write,

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{ccc:c} 0.56 & 0.40 & -0.18 & 2.34 \\ 0 & 0.12 & -0.15 & 0.33 \\ 0.20 & 0.71 & 1.0 & 2.04 \end{array}\right] \end{array}$$

$m_{21} = 0, \qquad \mathrm{R}_2 \leftarrow \mathrm{R}_2$ (remains unchanged)

$$m_{31} = -\frac{0.20}{0.56} = -0.357; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 0.357\mathrm{R}_1$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.56 & 0.40 & -0.18 & 2.34 \\ 0 & 0.12 & 0.15 & 0.33 \\ 0 & 0.567 & 1.064 & 1.205 \end{array}\right] \end{array}$$

$$m_{32} = -\frac{0.567}{0.12} = -4.725; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 4.725\mathrm{R}_2$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.56 & 0.40 & -0.18 & 2.34 \\ 0 & 0.12 & -0.15 & 0.33 \\ 0 & 0 & 0.355 & -0.354 \end{array}\right] \end{array}$$

$x_3 = -0.997;\ x_2 = 3.996;\ x_1 = 1.004$

After rounding to two decimals

$$x_1 = 1.00,\ x_2 = 4.00,\ x_3 = -1.00$$

(exact values are $x_1 = 1,\ x_2 = 4,\ x_3 = -1$)

Example 2.4

Solve the following system of simultaneous equations using Gaussian elimination with row interchanges:

$$0.3x_1 + 2.6x_2 + 1.3x_3 = 7.65$$

$$8.3x_1 + 8.2x_2 + 5.6x_3 = 43.17$$

$$12.7x_2 + 3.5x_2 + 7.4x_3 = 49.68$$

Compute up to two significant figures after decimal. Also solve the system without interchanging the rows.

Solution (by row interchanges)

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.3 & 2.6 & 1.3 & 7.65 \\ 8.3 & 8.2 & 5.6 & 43.17 \\ 12.7 & 3.5 & 7.4 & 49.68 \end{array}\right] \end{array}$$

Since $|12.7|$ is largest in absolute value in the first column, we interchange first row with third.

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 12.7 & 3.5 & 7.4 & 49.68 \\ 8.3 & 8.2 & 5.6 & 43.17 \\ 0.3 & 2.6 & 1.3 & 7.65 \end{array}\right] \end{array}$$

$m_{21} = -\dfrac{8.3}{12.7} = -0.65, \qquad \mathrm{R}_2 \leftarrow \mathrm{R}_2 - 0.65\mathrm{R}_1$

$m_{31} = -\dfrac{0.3}{12.7} = -0.024$, (two significant figures after decimal)

$\mathrm{R}_3 \leftarrow \mathrm{R}_3 - 0.024\mathrm{R}_1$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 12.7 & 3.5 & 7.4 & 49.68 \\ 0 & 5.92 & 0.79 & 10.88 \\ 0 & 2.52 & 1.12 & 6.46 \end{array}\right] \end{array}$$

Since $|5.92|$ is larger than $|2.52|$, there is no need to interchange rows.

$m_{32} = -\dfrac{2.52}{5.92} = -0.42; \qquad \mathrm{R}_3 \leftarrow \mathrm{R}_3 - 0.42\mathrm{R}_2$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 12.7 & 3.5 & 7.4 & 49.68 \\ 0 & 5.92 & 0.79 & 10.88 \\ 0 & 0 & 0.79 & 1.89 \end{array}\right] \end{array}$$

By back substitution,

$$x_3 = 2.39,\ x_2 = 1.52,\ x_1 = 2.10$$

(exact solution is $x_1 = 2.1,\ x_2 = 1.5,\ x_3 = 2.4$)

Solution (without row interchanges)

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.3 & 2.6 & 1.3 & 7.65 \\ 8.3 & 8.2 & 5.6 & 43.17 \\ 12.7 & 3.5 & 7.4 & 49.68 \end{array}\right] \end{array}$$

$m_{21} = -\dfrac{8.3}{0.3} = -27.67; \qquad \mathrm{R}_2 \leftarrow \mathrm{R}_2 - 27.67\mathrm{R}_1$

$$m_{31} = -\frac{12.7}{0.3} = -42.33; \qquad R_3 \leftarrow R_3 - 42.33R_1$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.3 & 2.6 & 1.3 & 7.65 \\ 0 & -63.74 & -30.37 & -168.50 \\ 0 & -106.56 & -47.63 & -274.14 \end{array}\right] \end{array}$$

$$m_{32} = -\frac{106.56}{63.74} = -1.67; \qquad R_3 \leftarrow R_3 - 1.67R_2$$

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & \boldsymbol{b} \end{matrix} \\ \left[\begin{array}{cccc} 0.3 & 2.6 & 1.3 & 7.65 \\ 0 & -63.74 & -30.37 & -168.50 \\ 0 & 0 & 3.09 & 7.26 \end{array}\right] \end{array}$$

$$x_3 = 2.35,\ x_2 = 1.52,\ x_1 = 2.14$$

(exact solution is, $x_1 = 2.1,\ x_2 = 1.5,\ x_3 = 2.4$)

2.14 LU Decomposition/Factorisation

A square matrix A can be decomposed/factorised (with conditions) into a product of two matrices L and U, i.e., $A = LU$, where L and U are lower triangular and upper triangular matrices respectively. We will discuss three methods in this regard (*i*) By Gaussian Elimination method (*ii*) Crout's method and (*iii*) Cholesky's method.

2.14.1 By Gaussian elimination method

We assume that no interchange of rows has taken place at any stage. Let us recall that in Gaussian Elimination method, matrix A is reduced to an upper triangular matrix U by a series of elementary row operations/transformations. For a 4×4 matrix, these transformations may be expressed in following way,

$$L_3L_2L_1A = U \tag{2.13}$$

where L_1, L_2 and L_3 are lower triangular matrices as,

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix}, \; L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & 0 & 1 \end{bmatrix}, \; L_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & m_{43} & 1 \end{bmatrix}.$$

The matrix A is original matrix and U the final upper triangular, i.e.,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}, U = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} \end{bmatrix}.$$

From (2.13) we have,

$$A = (L_3 L_2 L_1)^{-1} U = L_1^{-1} L_2^{-1} L_3^{-1} \; U.$$

It is easy to see that the inverses of L_1, L_2 and L_3 can be obtained simply by changing the signs of multipliers. It can also be shown by the logic of elementary transformations that,

$$L = L_1^{-1} L_2^{-1} L_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -m_{43} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & -m_{32} & 1 & 0 \\ -m_{41} & -m_{42} & -m_{43} & 1 \end{bmatrix}. \tag{2.14}$$

Thus we have the desired decomposition A = LU where L is given by (2.14).

The above may be generalised for $n \times n$ matrix straight away.

After reducing the matrix to LU form we can solve the system of equations $A\boldsymbol{x} = \boldsymbol{b}$ in the following manner:

$$A\boldsymbol{x} = LU\boldsymbol{x} = \boldsymbol{b} \tag{2.15}$$

Put $\quad \mathrm{U}\boldsymbol{x} = \boldsymbol{y}$, in (2.15); (2.16)

then, $\mathrm{L}\boldsymbol{y} = \boldsymbol{b}$, from (2.15). (2.17)

Solve (2.17) for $\boldsymbol{y}$, then solve (2.16) to obtain the required solution $\boldsymbol{x}$. The system (2.17) will be solved in a top-down manner while (2.16) in a bottom-up manner. Evidently the pivotal element should not be zero or very small since it will give rise to infinitely large multiplier, an element of L.

2.14.2 Crout's method

Let us consider each of L, U and A as 4×4 matrices, i.e.,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{2.18}$$

On the L.H.S. there are 16 unknowns, 6 elements of L and 10 of U. In order to evaluate them, we perform the product LU and equate, term by term, with the 16 elements of A. This is done in a systematic manner given as under:

Compute row-wise in the following order by equating the corresponding terms of LU and A:

(*i*) $\quad l_{11} = 1;\ u_{11},\ u_{12},\ u_{13},\ u_{14}$

(*ii*) $\quad l_{21},\ l_{22} = 1;\ u_{22},\ u_{23},\ u_{24}$

(*iii*) $\quad l_{31},\ l_{32},\ l_{33} = 1,\ u_{33},\ u_{34}$

(*iv*) $\quad l_{41},\ l_{42},\ l_{43},\ l_{44} = 1,\ u_{44}.$

Note: In practical computations we put $l_{11} = l_{22} = l_{33} = l_{44} = 1$ to start with. Thus we get,

(*i*) $\quad u_{11} = a_{11},\ u_{12} = a_{12},\ u_{13} = a_{13},\ u_{14} = a_{14}$

(ii) $\quad l_{21} \cdot u_{11} = a_{21}$ giving $l_{21} = a_{21}/u_{11}$; $u_{11}(= a_{11})$ should not be zero.

$l_{21}a_{12} + u_{22} = a_{22}$, giving $u_{22} = a_{22} - l_{21}a_{12}$; u_{22} should not be zero.

$l_{21}a_{13} + u_{23} = a_{23}$, giving $u_{23} = a_{23} - l_{21} \cdot a_{13}$

$l_{21}a_{14} + u_{24} = a_{24}$, giving $u_{24} = a_{24} - l_{21}a_{14}$

(*iii*) $$l_{31}u_{11} = a_{31}, \text{ giving } l_{31} = a_{31}/u_{11}$$

$$l_{31}u_{12} + l_{32}u_{22} = a_{32}, \text{ giving } l_{32} = (a_{32} - l_{31}u_{12})/u_{22};\ u_{22} \neq 0.$$

$$l_{31}u_{13} + l_{32}u_{23} + u_{33} = a_{33}, \text{ giving } u_{33} = a_{33} - (l_{31}u_{13} + l_{32}u_{23});\ u_{33} \text{ should not be zero.}$$

$$l_{31}u_{14} + l_{32}u_{24} + u_{34} = a_{34}, \text{ giving } u_{34} = a_{34} - (l_{31}u_{14} + l_{32}u_{24})$$

(*iv*) $$l_{41}u_{11} = a_{41}, \text{ giving } l_{41} = a_{41}/u_{11}$$

$$l_{41}u_{12} + l_{42}u_{22} = a_{42}, \text{ giving } l_{42} = (a_{42} - l_{41}u_{12})/u_{22}$$

$$l_{41}u_{13} + l_{42}u_{23} + l_{43}u_{33} = a_{43}, \text{ giving } l_{43} = (a_{43} - l_{41}u_{13} - l_{42}u_{23})/u_{33};\ u_{33} \neq 0.$$

$$l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + u_{44} = a_{44}, \text{ giving } u_{44} = a_{44} - (l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34}).$$

It may be noted that each element is computed before it is used.

The above can be generalised when A is an $(n \times n)$ matrix:

$$l_{ii} = 1, \qquad i = 1(1)n.$$

$$u_{1j} = a_{1j}, \quad j = 1(1)n.$$

For the i^{th} row, $i = 2(1)n$,

$$l_{ij} = [a_{ij} - \{l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{i,i-1}u_{i-1,j}\}]/u_{jj}$$

$$= \left[a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\right] \Big/ u_{jj}, \quad j = 1(1)i-1. \tag{2.19}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \quad j = i(1)n. \tag{2.20}$$

Note: (*i*) If it is required that A = LU where L is a Lower triangular and U a Unit upper triangular matrix, then we can proceed as follows:

Put $\quad u_{ii} = 1,\ i = 1(1)n.$

Compute the elements column-wise, i.e., elements of k^{th} column of U and elements of k^{th} column of L in that order for $k = 1(1)n$.

(*ii*) Matrix A can also be reduced to the form

$$\text{A} = \text{LDU}$$

where L and U are unit lower and unit upper triangular matrices and D, a diagonal matrix. Let A be a matrix of order n. It is easy to see that there will be $\frac{(n-1)n}{2}$ elements in the matrices L and U each and n elements in D. Thus there are total n^2 unknowns in the product LDU which can be matched with the n^2 elements of A. In practice however, we can reduce the matrix A to LU form where L is a unit lower triangular matrix and U, an upper triangular. We choose the elements of the diagonal matrix D to be the diagonal elements of U, i.e., $d_{ii} = u_{ii}$ and divide each element of U in the i^{th} row by u_{ii}. It may again be emphasised that value of none of the dividing element u_{ii} should be zero; otherwise the process will break down.

This method is also attributed to Doolittle.

2.14.3 Cholesky's method

The Cholesky's method deals with a special case when the given matrix is symmetric and positive definite. If A is a symmetric matrix, $\mathrm{A} = \mathrm{A}^{\mathrm{T}}$ then it can be expressed as product of two matrices L and L^{T} where L is a lower triangular matrix, i.e.,

$$\mathrm{A} = \mathrm{LL}^{\mathrm{T}} \text{ or } (\mathrm{U}^{\mathrm{T}}\mathrm{U}). \tag{2.21}$$

For example, for a 4×4 matrix,

$$\underset{\text{L or } \mathrm{U}^{\mathrm{T}}}{\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}} \underset{\mathrm{L}^{\mathrm{T}} \text{ or U}}{\begin{bmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ 0 & l_{22} & l_{32} & l_{42} \\ 0 & 0 & l_{33} & l_{43} \\ 0 & 0 & 0 & l_{44} \end{bmatrix}} = \underset{\text{A}}{\begin{bmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}} \tag{2.22}$$

In (2.22), there are 10 elements in matrix L and 16 in A. But due to symmetry six elements above the diagonal are same as below the diagonal, i.e., $a_{ij} = a_{ji},\ i \neq j$. Thus equating the corresponding terms of LL^{T} and A gives the elements of L. When A is an $(n \times n)$ matrix, the number of elements to be evaluated are $n(n+1)/2$.

We evaluate the elements of L by equating the corresponding elements column-wise (or row-wise since LL^{T} is a symmetric matrix). Let us say, we are equating column-wise.

For each value of $j = 1(1)n$, compute for $i = j(1)n$, the values of l_{ij} by the following formulae:

$$\text{for } i = j, \quad l_{jj} = \left[a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right]^{1/2}, \tag{2.23}$$

$$\text{for } i > j, \quad l_{ij} = \left[a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk} \right] \Big/ l_{jj} \tag{2.24}$$

In case of 4×4 matrix, i.e., (2.22) we compute the elements of L using formulae (2.23) and (2.24) in the following manner:

For $j = 1$:

$$i = j = 1 \Rightarrow l_{11} = \sqrt{a_{11}},$$

$$i = 2(j = 1) \Rightarrow l_{21} = a_{21}/l_{11},$$

$$i = 3(j = 1) \Rightarrow l_{31} = a_{31}/l_{11},$$

$$i = 4(j = 1) \Rightarrow l_{41} = a_{41}/l_{11}.$$

For $j = 2$:

$$i = j = 2 \Rightarrow l_{22} = [(a_{22} - l_{21})^2]^{1/2},$$

$$i = 3(j = 2) \Rightarrow l_{32} = (a_{23} - l_{31} \cdot l_{21})/l_{22},$$

$$i = 4(j = 2) \Rightarrow l_{42} = (a_{24} - l_{41} \cdot l_{21})/l_{22}.$$

For $j = 3$:

$$i = j = 3 \Rightarrow l_{33} = [a_{33} - (l_{31}^2 + l_{32}^2)]^{1/2},$$

$$i = 4(j = 3) \Rightarrow l_{43} = [a_{43} - (l_{41} l_{31} + l_{42} \cdot l_{32})]/l_{33}.$$

For $j = 4$:

$$i = j = 4 \Rightarrow l_{44} = [a_{44} - (l_{41}^2 + l_{42}^2 + l_{43}^2)]^{1/2}.$$

Example 2.5

Reduce the following matrix A to LU form by Gaussian elimination where L is a Unit Lower Triangular matrix and U is an upper triangular; also solve the system $A\boldsymbol{x} = \boldsymbol{b}$ where,

$$\mathrm{A} = \begin{bmatrix} 3 & 2 & 1 & -4 \\ 1 & -5 & 2 & 1 \\ 5 & 1 & -3 & 2 \\ 2 & 3 & 1 & 5 \end{bmatrix},$$

$\boldsymbol{x}^{\mathrm{T}} = (x_1 \;\; x_2 \;\; x_3 \;\; x_4)$ and $\boldsymbol{b}^{\mathrm{T}} = (5 \;\; 18 \;\; -4 \;\; 11)$.

Solution It is same coefficient matrix as in Example 2.1.

From (2.14) and (2.13),

$$\mathrm{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.333 & 1 & 0 & 0 \\ 1.667 & 0.412 & 1 & 0 \\ 0.667 & -0.294 & -0.0154 & 1 \end{bmatrix} \text{ and } \mathrm{U} = \begin{bmatrix} 3 & 2 & 1 & -4 \\ 0 & -5.666 & 1.667 & 2.332 \\ 0 & 0 & -5.354 & 7.707 \\ 0 & 0 & 0 & 9.541 \end{bmatrix}$$

In $\mathrm{LU}\boldsymbol{x} = \boldsymbol{b}$, putting $\mathrm{U}\boldsymbol{x} = \boldsymbol{y}$, the system becomes $\mathrm{L}\boldsymbol{y} = \boldsymbol{b}$, where $\boldsymbol{y}^{\mathrm{T}} = (y_1 \;\; y_2 \;\; y_3 \;\; y_4)$. Solving $\mathrm{L}\boldsymbol{y} = \boldsymbol{b}$ by forward substitution gives,

$$y_1 = 5,\ y_2 = 16.335,\ y_3 = -19.065,\ y_4 = 9.531$$

Finally on solving $\mathrm{U}\boldsymbol{x} = \boldsymbol{y}$, by backward substitution, we get,

$$x_4 = 0.999,\ x_3 = 4.999,\ x_2 = -1.001,\ x_1 = 2.000$$

Example 2.6

Decompose the following matrix A to LU form by Crout's method where L is a unit lower triangular and U an upper triangular matrix,

$$\mathrm{A} = \begin{bmatrix} 3 & 2 & 1 & -4 \\ 1 & -5 & 2 & 1 \\ 5 & 1 & -3 & 2 \\ 2 & 3 & 1 & 5 \end{bmatrix}.$$

Solution Let us assume,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 & -4 \\ 1 & -5 & 2 & 1 \\ 5 & 1 & -3 & 2 \\ 2 & 3 & 1 & 5 \end{bmatrix}.$$

1st row:

$$u_{11} = 3,\ u_{12} = 2,\ u_{13} = 1,\ u_{14} = -4$$

2nd row:

$$l_{21} = a_{21}/u_{11} = 1/3 = 0.333$$

$$u_{22} = a_{22} - l_{21} \cdot u_{12} = -5 - 0.333 \times 2 = -5.666$$

$$u_{23} = a_{23} - l_{21} \cdot u_{13} = 2 - 0.333 \times 1 = 1.667$$

$$u_{24} = a_{24} - l_{21} \cdot u_{14} = 1 - 0.333 \times (-4) = 2.332$$

3rd row:

$$l_{31} = a_{31}/u_{11} = 5/3 = 1.667$$

$$l_{32} = (a_{32} - l_{31} \cdot u_{12})/u_{22} = (1 - 1.667 \times 2)/(-5.667) = 0.412$$

$$u_{33} = a_{33} - (l_{31}u_{13} + l_{32}u_{23}) = -3 - (1.667 \times 1 + 0.412 \times 1.667) = -5.354$$

$$u_{34} = a_{34} - (l_{31}u_{14} + l_{32}u_{24}) = 2 - (1.667 \times (-4) + 0.412 \times 2.332) = 7.707$$

4th row:

$$l_{41} = a_{41}/u_{11} = 2/3 = 0.667$$

$$l_{42} = (a_{42} - l_{41} \cdot u_{12})/u_{22} = (3 - 0.667 \times 2)/(-5.667) = -0.294$$

$$l_{43} = [a_{43} - (l_{41} \cdot u_{13} + l_{42} \cdot u_{23})]/u_{33}$$

$$= [1 - (0.667 \times 1 - 0.294 \times 1.667)]/(-5.354) = -0.154$$

$$u_{44} = a_{44} - (l_{41} \cdot u_{14} + l_{42} \cdot u_{24} + l_{43} \cdot u_{34})$$

$$= 5 - (-0.667 \times 4 - 0.294 \times 2.332 - 0.154 \times 7.707)$$

$$= 9.540$$

$$\mathrm{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.333 & 1 & 0 & 0 \\ 1.667 & 0.412 & 1 & 0 \\ 0.667 & -0.294 & -0.154 & 1 \end{bmatrix}; \mathrm{U} = \begin{bmatrix} 3 & 2 & 1 & -4 \\ 0 & -5.666 & 1.667 & 2.332 \\ 0 & 0 & -5.354 & 7.707 \\ 0 & 0 & 0 & 9.540 \end{bmatrix}$$

Note: This example is same as 2.5. There is a difference of .001 in u_{44} due to rounding, which may be expected.

Example 2.7

Using the lower and upper triangular matrices obtained in Example 2.6, reduce the matrix A to LDU form where L and U are unit lower and unit upper triangular matrices respectively.

Further, using LDU, express A = LU where L is a lower triangular and U a unit upper triangular matrix.

Solution In order to express A = LDU, we divide each row of U by its diagonal element, i.e., we divide the i^{th} row by u_{ii}. The elements of D are given by $d_{ii} = u_{ii}$. Thus we have,

$$\text{LDU} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.333 & 1 & 0 & 0 \\ 1.667 & 0.412 & 1 & 0 \\ 0.667 & -0.294 & -0.154 & 1 \end{bmatrix}}_{\text{L}} \underbrace{\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & -5.666 & 0 & 0 \\ 0 & 0 & -5.354 & 0 \\ 0 & 0 & 0 & 9.540 \end{bmatrix}}_{\text{D}}$$

$$\underbrace{\begin{bmatrix} 1 & 0.667 & 0.333 & -1.333 \\ 0 & 1 & -0.294 & -0.412 \\ 0 & 0 & 1 & -1.439 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{U}}$$

To further reduce it to A = LU form, we have to multiply the elements of the j^{th} column of unit lower triangular matrix by $d_{jj},\ j = 1(1)4$. Thus we get

$$\text{LU} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0.999 & -5.666 & 0 & 0 \\ 5.001 & -2.334 & -5.354 & 0 \\ 2.001 & 1.666 & 0.824 & 9.540 \end{bmatrix} \begin{bmatrix} 1 & 0.667 & 0.333 & -1.333 \\ 0 & 1 & -0.294 & -0.412 \\ 0 & 0 & 1 & -1.439 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

2.14.4 Reduction to PA = LU

In Sec. 2.14.1, we had reduced matrix A to LU form by Gaussian Elimination assuming there were no interchanges of rows at any stage. But if interchanges are employed, then the matrix A can be reduced to the form PA = LU where P is a permutation matrix, i.e., a matrix obtained from a unit matrix with its rows interchanged (permuted). We illustrate the reduction by considering a 4×4 matrix A.

Let I_{pq} denote a unit matrix with its p^{th} and q^{th} rows interchanged and that matrix A has been reduced to an upper triangular matrix U under following transformations,

$$\text{L}_3\text{I}_{34}\text{L}_2\text{I}_{24}\text{L}_1\text{I}_{13}\text{A} = \text{U}, \tag{2.25}$$

where L_1, L_2 and L_3 are unit lower triangular matrices as given in (2.14).
First we should note that $\text{I}_{pq}^{-1} = \text{I}_{pq}$, i.e., I_{pq} is an involutary matrix.
From (2.25) we can write,

$$I_{34}I_{24}I_{13}A = \text{I}_{34} \cdot \text{I}_{24}\text{L}_1^{-1}\text{I}_{24}^{-1}\text{L}_2^{-1}\text{I}_{34}^{-1}\text{L}_3^{-1}\text{U}$$

$$= I_{34}I_{24}L_1^{-1}I_{24}L_2^{-1}I_{34}L_3^{-1}U$$

$$= [I_{34}\{(I_{24}L_1^{-1}I_{24})L_2^{-1}\}I_{34}]L_3^{-1}U. \tag{2.26}$$

We see that,

$$I_{24}L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 0 & 0 & 1 \\ -m_{31} & 0 & 1 & 0 \\ -m_{21} & 1 & 0 & 0 \end{bmatrix}, \; I_{24}L_1^{-1}I_{24} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{21} & 0 & 0 & 1 \end{bmatrix}$$

Now, the bracketed term,

$$(I_{24}L_1^{-1}I_{24})_{L_2}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{21} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{31} & -m_{32} & 1 & 0 \\ -m_{21} & -m_{42} & 0 & 1 \end{bmatrix}.$$

The term within square bracket in (2.26) becomes,

$$I_{34}\{(I_{24}L_1^{-1}I_{24})L_2{}^{-1}\}I_{34} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{21} & -m_{42} & 1 & 0 \\ -m_{31} & -m_{32} & 0 & 1 \end{bmatrix}$$

Finally the right side of (2.26) will be,

$$\left[I_{34}\left\{\left(I_{24}L_1^{-1}I_{24}\right)L_2{}^{-1}\right\}I_{34}\right]L_3^{-1}U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{21} & -m_{42} & 1 & 0 \\ -m_{31} & -m_{32} & -m_{43} & 1 \end{bmatrix} U.$$

The left side of (2.26) will be,

$$I_{34}I_{24}I_{13}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} A$$

$$= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} A$$

Thus we have the reduction,

$$PA = LU \tag{2.27}$$

where P is the permutation matrix and L is a unit lower triangular matrix given as

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \; L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{41} & 1 & 0 & 0 \\ -m_{21} & -m_{42} & 1 & 0 \\ -m_{31} & -m_{32} & -m_{43} & 1 \end{bmatrix}. \tag{2.28}$$

The reduction $PA = LU$ can also be expressed as $A = P^{-1}LU$ where $P^{-1} = I_{13}I_{24}I_{34}$.

In practice, to compute the permutation matrix P, we maintain the record of the interchanges of rows of the unit matrix by a one-dimensional array (vector) storing the row number in it. In the above case of 4×4 matrix we may have an array row number $(1 : 4)$ having values 1, 2, 3, 4 initially. We change these numbers according to change of rows. In the present case I_{13} will change the order to 3, 2, 1, 4. Then I_{24} will change it to 3, 4, 1, 2 and finally I_{34} will change it to 3, 4, 2, 1. Then the four rows of matrix P can be written as row number 3, 4, 2 and 1 of the unit matrix. For complete details of computational procedure see Example 2.8.

Example 2.8

Reduce the following matrix A in the form PA = LU, by Gaussian Elimination method with partial pivoting, where L is a unit lower triangular and U an upper triangular matrix; P is a permutation matrix.

$$\mathrm{A} = \begin{bmatrix} 2 & 5 & 1 & 8 \\ 1 & 6 & 3 & 5 \\ 7 & 2 & 6 & 3 \\ 4 & 8 & 1 & 2 \end{bmatrix}$$

Also solve the system of equations $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{b}^{\mathrm{T}} = (5\ 11\ 14\ 19)$. Compute up to two decimals only.

Solution

Row number of unit matrix	Cols of multipliers	Transformed matrix	$\boldsymbol{b}$
$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$	—	$\begin{bmatrix} 2 & 5 & 1 & 8 \\ 1 & 6 & 3 & 5 \\ 7 & 2 & 6 & 3 \\ 4 & 8 & 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 5 \\ 11 \\ 14 \\ 19 \end{bmatrix}$

Interchanging 1st row by 3rd,

$\begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix}$	—	$\begin{bmatrix} 7 & 2 & 6 & 3 \\ 1 & 6 & 3 & 5 \\ 2 & 5 & 1 & 8 \\ 4 & 8 & 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 14 \\ 11 \\ 5 \\ 19 \end{bmatrix}$

Elimination of coeff. of x_1;

$$\begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \quad \begin{matrix} 1 \\ -0.14 \\ -0.28 \\ -0.57 \end{matrix} \quad \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 5.72 & 2.16 & 4.58 \\ 0 & 4.44 & -0.68 & 7.16 \\ 0 & 6.86 & -2.42 & 0.29 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 9.04 \\ 1.08 \\ 11.02 \end{bmatrix}$$

Interchanging 2nd by 4th row,

$$\begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix} \quad \begin{matrix} 1 \\ -0.57 \\ -0.14 \\ -0.14 \end{matrix} \quad \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 6.86 & -2.42 & 0.29 \\ 0 & 4.44 & -0.68 & 7.16 \\ 0 & 5.72 & 2.16 & 4.58 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 11.02 \\ 1.08 \\ 9.04 \end{bmatrix}$$

Elimination of coeff. of x_2;

$$\begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix} \quad \begin{matrix} 1 & 0 \\ -0.57 & 1 \\ -0.28 & -0.65 \\ -0.14 & -0.83 \end{matrix} \quad \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 6.86 & -2.42 & 0.29 \\ 0 & 0 & 0.89 & 6.97 \\ 0 & 0 & 4.17 & 4.34 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 11.02 \\ -6.08 \\ -0.11 \end{bmatrix}$$

Interchanging 3rd by 4th row,

$$\begin{bmatrix} 3 \\ 4 \\ 2 \\ 1 \end{bmatrix} \quad \begin{matrix} 1 & 0 \\ -0.57 & 1 \\ -0.14 & -0.83 \\ -0.28 & -0.65 \end{matrix} \quad \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 6.86 & -2.42 & 0.29 \\ 0 & 0 & 4.17 & 4.34 \\ 0 & 0 & 0.89 & 6.97 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 11.02 \\ -0.11 \\ -6.08 \end{bmatrix}$$

Eliminating coeff. of x_3,

$$\begin{bmatrix} 3 \\ 4 \\ 2 \\ 1 \end{bmatrix} \quad \begin{matrix} 1 & 0 & 0 \\ -0.57 & 1 & 0 \\ -0.14 & -0.83 & 1 \\ -0.28 & -0.65 & -0.21 \end{matrix} \quad \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 6.86 & -2.42 & 0.29 \\ 0 & 0 & 4.17 & 4.34 \\ 0 & 0 & 0 & 6.06 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 11.02 \\ 0.11 \\ -6.06 \end{bmatrix}$$

Solving by back-substitution gives

$x_4 = -1.00,\ x_3 = 1.01,\ x_2 = 2.00,\ x_1 = 0.99$

(exact solution is $x_1 = 1,\ x_2 = 2,\ x_3 = 2,\ x_3 = 1,\ x_4 = -1$)

$$\mathrm{P} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \leftarrow 3^{\text{rd}} \text{ row of I} \\ \leftarrow 4^{\text{th}} \text{ row of I} \\ \leftarrow 2^{\text{nd}} \text{ row of I} \\ \leftarrow 1^{\text{st}} \text{ row of I} \end{matrix} ;\ \mathrm{PA} = \begin{bmatrix} 7 & 2 & 6 & 3 \\ 4 & 8 & 1 & 2 \\ 1 & 6 & 3 & 5 \\ 2 & 5 & 1 & 8 \end{bmatrix};$$

$$\mathrm{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.57 & 1 & 0 & 0 \\ 0.14 & 0.83 & 1 & 0 \\ 0.28 & 0.65 & 0.21 & 1 \end{bmatrix}; \mathrm{U} = \begin{bmatrix} 7 & 2 & 6 & 3 \\ 0 & 6.86 & -2.42 & 0.29 \\ 0 & 0 & 4.17 & 4.34 \\ 0 & 0 & 0 & 6.06 \end{bmatrix}.$$

Note: Multipliers can be stored in space occupied by zeros in the appropriate column.

2.15 Gauss–Jordan (or Jordan's) Method

Let us again consider solving the system of equations (2.2), i.e., $A\boldsymbol{x} = \boldsymbol{b}$. In the Gaussian elimination method we reduce the system to an upper triangular form by making elementary operations on the rows below the pivotal row and then obtain the solution by the process of back-substitution. In the Jordan's method operations are made on all the rows below the pivotal row as well as on the rows above it such that the system reduces to an Identity matrix form, $I\boldsymbol{x} = \boldsymbol{b}'$ and the solution is straight away given by the transformed right side $\boldsymbol{b}'$. In order to achieve this, the pivotal row is divided by the pivotal element throughout, thus making the pivot as unity. Then operations are made, choosing suitable multipliers so that all the elements below the pivot as well as above it are reduced to zero. Of course same operations are to be made simultaneously on the right side. For solving an $n \times n$ system, there will be $(n-1)$ stages and in each stage there will be $(n-1)$ steps to be performed to make $(n-1)$ elements zero. The computations are made on the augmented matrix aug $(A/\boldsymbol{b})$ as shown in the following example.

Example 2.9

Solve by Jordan's method $A\boldsymbol{x} = \boldsymbol{b}$ where

$$A = \begin{bmatrix} 2 & 5 & 1 & 8 \\ 1 & 6 & 3 & 5 \\ 7 & 2 & 6 & 3 \\ 4 & 8 & 1 & 2 \end{bmatrix}; \boldsymbol{b}^{T} = (5 \;\; 11 \;\; 14 \;\; 19)$$

Solution

$$\left[\begin{array}{cccc:c} 2 & 5 & 1 & 8 & 5 \\ 1 & 6 & 3 & 5 & 11 \\ 7 & 2 & 6 & 3 & 14 \\ 4 & 8 & 1 & 2 & 19 \end{array}\right] \xrightarrow{R_1 \leftarrow R_1/2} \left[\begin{array}{cccc:c} 1 & 2.5 & 0.5 & 4 & 2.5 \\ 1 & 6 & 3 & 5 & 11 \\ 7 & 2 & 6 & 3 & 14 \\ 4 & 8 & 1 & 2 & 19 \end{array}\right]$$

$R_2 \leftarrow R_2 - R_1;\ R_3 \leftarrow R_3 - 7R_1;\ R_4 \leftarrow R_4 - 4R_1$

$$\left[\begin{array}{cccc:c} 1 & 2.5 & 0.5 & 4 & 2.5 \\ 0 & 3.5 & 2.5 & 1 & 8.5 \\ 0 & -15.5 & 2.5 & -25 & -3.5 \\ 0 & -2 & -1 & -14 & 9 \end{array}\right] \xrightarrow{R_2 \leftarrow R_2/3.5} \left[\begin{array}{cccc:c} 1 & 2.5 & 0.5 & 4 & 2.5 \\ 0 & 1 & 0.71 & 0.28 & 2.43 \\ 0 & -15.5 & 2.5 & -25 & -3.5 \\ 0 & -2 & -1 & -14 & 9 \end{array}\right]$$

$R_1 \leftarrow R_1 - 2.5R_2;\ R_3 \leftarrow R_3 + 15.5R_2;\ R_4 \leftarrow R_4 + 2R_2$

$$\left[\begin{array}{cccc|c} 1 & 0 & -1.28 & 3.3 & -3.58 \\ 0 & 1 & 0.71 & 0.28 & 2.43 \\ 0 & 0 & 13.50 & -20.66 & 34.16 \\ 0 & 0 & 0.42 & -13.44 & 13.86 \end{array}\right] \xrightarrow{R_3 \leftarrow R_3/13.50} \left[\begin{array}{cccc|c} 1 & 0 & -1.28 & 3.3 & -3.58 \\ 0 & 1 & 0.71 & 0.28 & 2.43 \\ 0 & 0 & 1 & -1.53 & 2.53 \\ 0 & 0 & 0.42 & -13.44 & 13.86 \end{array}\right]$$

$R_1 \leftarrow R_1 + 1.28R_1;\ R_2 \leftarrow R_2 - 0.71R_3;\ R_4 \leftarrow R_4 - 0.42R_3$

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 1.34 & -0.34 \\ 0 & 1 & 0 & 1.37 & 0.63 \\ 0 & 0 & 1 & -1.53 & 12.53 \\ 0 & 0 & 0 & -12.80 & 12.80 \end{array}\right] \xrightarrow{R_4 \leftarrow R_4/(-12.8)} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1.34 & -0.34 \\ 0 & 1 & 0 & 1.37 & 0.63 \\ 0 & 0 & 1 & -1.53 & 2.53 \\ 0 & 0 & 0 & 1 & -1.00 \end{array}\right]$$

$R_1 \leftarrow R_1 - 1.34R_4;\ R_2 \leftarrow R_2 - 1.37R_4;\ R_3 \leftarrow R_3 - 1.53R_4$

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 & 1.00 \\ 0 & 1 & 0 & 0 & 2.00 \\ 0 & 0 & 1 & 0 & 1.00 \\ 0 & 0 & 0 & 1 & -1.00 \end{array}\right].$$

$x_1 = 1,\ x_2 = 2,\ x_3 = 1,\ x_4 = -1$

(correct answer is $x_1 = 1,\ x_2 = 2,\ x_3 = 1,\ x_4 = -1$)

Note: Change of rows may be performed if required.

2.16 Tridiagonal System

The solution of second order boundary value problems by numerical method reduces to solving a system of linear equations which is tridiagonal in nature. A (4×4) tri-diagonal system of equations may be written as,

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}. \tag{2.29}$$

From Gaussian elimination or Crout's method, it is easy to see that the coefficient matrix in (2.29) can be reduced to LU where L and U have following forms,

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ 0 & l_{32} & 1 & 0 \\ 0 & 0 & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & 0 & 0 \\ 0 & u_{22} & u_{23} & 0 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

We can easily find the elements of L and U as given below:

$$u_{11} = a_{11},\ u_{12} = a_{12};\ a_{11} \neq 0.$$

$$l_{21} = a_{21}/a_{11};\ u_{22} = a_{22} - \frac{a_{12} \cdot a_{21}}{u_{11}};\ u_{23} = a_{23}.$$

$$l_{32} = \frac{a_{32}}{a_{22}};\ u_{33} = a_{33} - \frac{a_{23} \cdot a_{32}}{u_{22}};\ u_{34} = a_{34}.$$

$$l_{43} = \frac{a_{43}}{a_{33}};\ u_{44} = a_{44} - \frac{a_{34} \cdot a_{43}}{u_{33}}.$$

We solve A$\boldsymbol{x} = \boldsymbol{b}$ or LU$\boldsymbol{x} = \boldsymbol{b}$ by putting U$\boldsymbol{x} = \boldsymbol{y}$ so that L$\boldsymbol{y} = \boldsymbol{b}$ has to be solved first for $\boldsymbol{y}$ giving,

$$y_1 = b_1,\ y_2 = b_2 - \frac{a_{21}}{u_{11}} y_1;\ y_3 = b_3 - \frac{a_{32}}{u_{22}} y_2,\ y_4 = b_4 - \frac{a_{43}}{u_{33}} y_3.$$

Further, on solving U$\boldsymbol{x} = \boldsymbol{y}$, we get,

$$x_4 = y_4/u_{44};\ x_3 = (y_3 - a_{34}x_4)/u_{33}$$

$$x_2 = (y_2 - a_{23} \cdot x_3)/u_{22};\ x_1 = (y_1 - a_{12} \cdot x_2)/u_{11}.$$

Thus we are required to compute the following in that order:

(1) u_{11} u_{22} u_{33} u_{44}

(2) y_1 y_2 y_3 y_4

(3) x_4 x_3 x_2 x_1

However in order to save memory space in the computer, we may store the elements of A by three one-dimensional arrays for storing the diagonal, subdiagonal and super diagonal elements. Let us now consider an $n \times n$ tridiagonal system which is stored as follows:

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & \cdots & 0 \\ 0 & a_3 & b_3 & c_3 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & 0 & & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_n \end{bmatrix} \tag{2.30}$$

Compute the following in that order:

$$(1) \quad u_1 = b_1;\ u_i = b_i - \frac{a_i \cdot c_{i-1}}{u_{i-1}},\ i = 2(1)n. \tag{2.30a}$$

$$(2) \quad y_1 = r_1;\ y_i = r_i - \frac{a_i y_{i-1}}{u_{i-1}},\ i = 2(1)n. \tag{2.30b}$$

$$(3) \quad x_n = y_n/u_n;\ x_i = (y_i - c_i x_{i+1})/u_i,\ i = n-1(-1)1. \tag{2.30c}$$

It may be noted that instead of $n \times n$, only $3n-2$ computer locations are required for storing the elements of the coefficient matrix.

Example 2.10

Solve the following tridiagonal system of equations,

$$\begin{aligned} 1.98x_1 - 1.01x_2 &= 0.985 \\ 0.98x_1 - 1.98x_2 + 1.02x_3 &= 0.01 \\ 0.97x_2 - 1.98x_3 + 1.03x_4 &= 0.015 \\ 0.96x_3 - 1.98x_4 &= -1.540 \end{aligned}$$

Compute up to four places of decimal.

Solution

$$u_1 = 1.98$$

$$u_2 = -1.98 - \frac{0.98 \times (-1.01)}{1.98} = -1.4801$$

$$u_3 = -1.98 - \frac{0.97 \times (1.02)}{-1.4801} = -1.3115$$

$$u_4 = -1.98 - \frac{0.96 \times (1.03)}{-1.3115} = -1.2260$$

$$y_1 = 0.985$$

$$y_2 = 0.01 - \frac{0.98}{1.98} \times 0.985 = -0.4775$$

$$y_3 = 0.015 - \frac{0.97 \times (-0.4775)}{-1.4801} = -0.2979$$

$$y_4 = -1.540 - \frac{0.96 \times (-0.2979)}{-1.3115} = -1.7580$$

$$x_4 = \frac{-1.7580}{-1.2260} = 1.4339$$

$$x_3 = (-0.2979 - 1.03 \times 1.4339)/(-1.3115) = 1.3533$$

$$x_2 = (-0.4775 - 1.02 \times 1.3533)/(-1.4801) = 1.2552$$

$$x_1 = \{0.985 - (-1.01) \times 1.2552\}/1.98 = 1.1378$$

Note: The above equations have arrived at in solving a differential equation described in Chapter 7.

2.17 Inversion of Matrix

We can find the solution of $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$ using A^{-1} as $\boldsymbol{x} = \mathrm{A}^{-1}\boldsymbol{b}$, although it will not be wise to first compute A^{-1} and then form $\mathrm{A}^{-1}\boldsymbol{b}$. However, if we have got a software/subroutine/procedure for A^{-1}, we can use it easily on any number of different $\boldsymbol{b}$ vectors. We had also said in Sec 2.9 that to compute A^{-1} by adj A/|A| is most uneconomical computation-wise. We had also mentioned in Sec 2.7 that A^{-1} can be found by solving $\mathrm{A}\boldsymbol{x} = \boldsymbol{b}$, taking $\boldsymbol{b}$ as different columns of a unit matrix. That is, solution $\boldsymbol{x}$ of $\mathrm{A}\boldsymbol{x} = \mathrm{I}_k$, where I_k is the k^{th} column of the unit/identity matrix I, renders the k^{th} column of A^{-1}. Thus we have to work with the augmented matrix aug (A/I). When matrix A is of order 4, we have,

$$\text{aug } (\mathrm{A/I}) = \left[\begin{array}{cccc:cccc} x_1 & x_2 & x_3 & x_4 & \mathrm{I}_1 & \mathrm{I}_2 & \mathrm{I}_3 & \mathrm{I}_4 \\ a_{11} & a_{12} & a_{13} & a_{14} & 1 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & 0 & 1 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} & 0 & 0 & 1 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} & 0 & 0 & 0 & 1 \end{array}\right].$$

All the methods discussed before can be employed to solve the 4×4 system with four right hand sides, namely (*i*) Gauss Elimination (*ii*) LU decomposition and (*iii*) Jordan's. See Examples 2.11 and 2.12.

Example 2.11

Find the inverse of the following matrix A by Gaussian elimination method, where

$$A = \begin{bmatrix} 4.3 & -3.5 & -1.2 \\ 18.4 & 2.1 & -1.0 \\ 7.2 & 1.8 & 3.4 \end{bmatrix}.$$

Using A^{-1}, compute $\boldsymbol{x}$ form $A\boldsymbol{x} = \boldsymbol{b}$, when $\boldsymbol{b}^T = (10.90 \quad 7.80 \quad 23.22)$. Compute up to 2 decimals only.

Solution We apply Gaussian elimination method taking $\boldsymbol{b}$ as columns of unit matrix I, i.e., considering augmented matrix:

$$\left[\begin{array}{ccc|ccc} 4.3 & -3.5 & -1.2 & 1 & 0 & 0 \\ 18.4 & 2.1 & -1.0 & 0 & 1 & 0 \\ 7.2 & 1.8 & 3.4 & 0 & 0 & 1 \end{array}\right]$$

$R_2 \leftarrow R_2 - 4.28R_1$; $R_3 \leftarrow R_3 - 1.67R_1$

$$\left[\begin{array}{ccc|ccc} 4.3 & -3.5 & -1.2 & 1 & 0 & 0 \\ 0 & 17.08 & 4.14 & -4.28 & 1 & 0 \\ 0 & 7.64 & 5.40 & -1.67 & 0 & 1 \end{array}\right]$$

$R_3 \leftarrow R_3 - 0.45R_2$

$$\left[\begin{array}{ccc|ccc} 4.3 & -3.5 & -1.2 & 1 & 0 & 0 \\ 0 & 17.08 & 4.14 & -4.28 & 1 & 0 \\ 0 & 0 & 3.54 & 0.26 & -0.45 & 1 \end{array}\right]$$

Solving by back-substitution, for 3 right sides, we get the respective 3 columns of A^{-1}, i.e.,

$$A^{-1} = \begin{bmatrix} 0.03 & 0.04 & 0.02 \\ -0.27 & 0.09 & -0.07 \\ 0.07 & -0.13 & 0.28 \end{bmatrix}.$$

For given $\boldsymbol{b}^T$,

$$\boldsymbol{x} = \begin{bmatrix} 0.03 & 0.04 & 0.02 \\ -0.27 & 0.09 & -0.07 \\ 0.07 & -0.13 & 0.28 \end{bmatrix} \begin{bmatrix} 10.90 \\ 7.80 \\ 23.22 \end{bmatrix} = \begin{bmatrix} 1.10 \\ -3.87 \\ 6.25 \end{bmatrix}.$$