Lecture notes on

- Numerical Analysis

Hui Wei

Only for foreign graduate students of AUST

Contents

1 Computational Error **1**

1.1 Error definition 1
1.2 Types of error 2
1.3 Significant Digit 3
1.4 Rules for Rounding off numbers 3
1.5 Exercises for Sect.1-4 6
1.6 Two important theorems 7
1.7 Miscellaneous examples. 10
1.8 Exercises 13

2 Solutions of System of Linear Equations **15**

2.1 Direct Methods 16
2.1.1 Gauss elimination method 16
2.1.2 Gauss Elimination in matrix notation 20
2.1.3 Simple Partial Pivoting 23
2.1.4 Exercises 25
2.1.5 Crout's method 26
2.2 Iterative Methods 32
2.2.1 Jacobi's Method 32
2.2.2 Gauss Seidel Method 36
2.2.3 Exercises 39
2.3 What is a norm? 41

3 Eigenvalues and Eigenvectors **47**

3.1 Power method 48
3.2 Exercises 51
3.3 Inverse power method 52
3.4 Exercises 54

4 Roots of Non-linear Equations **55**
4.1 Introduction 55
4.2 Bisection method 56
4.3 Fixed Point Iteration 59
4.4 Newton-Raphson Method 63

5 Interpolation **69**
5.1 Lagrangian Polynomial 70
5.1.1 Lagrange Interpolating Polynomials 70
5.2 Finite Differences 74
5.2.1 Introduction 74
5.2.2 Operators 74
5.2.3 Exercises 80
5.2.4 Divided Differences 81
5.3 Interpolation in Newton's Polynomial 83

6 References **87**

Chapter 1

Computational Error

Lecture 1

1.1 Error definition

Numerical methods are to provide practical procedures for obtaining the numerical solutions of problems to a specified degree of accuracy. In numerical analysis, besides the study of the methods, one studies the errors involving in the methods and in the final results.

There are many kinds of error, which we shall discuss in detail later on. At present the meaning of error can be taken as follows; let an approximate value of a number be x^*, whose actual value is x. The value difference $|x^* - x|$ is called the absolute error in x^*, denoted by

$$e_A(x^*) = |x^* - x|.$$

$e_r(x^*) = \frac{|x^*-x|}{x}$ is called the relative error in x^*. In general, the real value x is unknown, thus the relative error is replace by

$$e_r(x^*) = \frac{|x^*-x|}{x^*}$$

Then, the percentage error in x^* is 100 times its relative error. That is, percentage error in $e_p = e_A \times 100\%$.

1.2 Types of error

In general, the errors in a practical problem may get introduced into four forms as follows:

I. Initial error/Error of the problem:

These are involved in the statement of problem itself. In fact, the statement of a problem generally gives an idealized model and not the exact picture of the actual phenomena. For example, in the calculation of the value of earths gravitational force g by simple pendulum, the experiment is based upon certain axioms: such as

(i) bob is weight less;

(ii) the motion of the bob is linear, that is, in a straight line; which are not true, in fact. So the value of the parameter (s) involved can only be determined approximately.

II. Residual error or truncation error:

This error occurs when mathematical functions like

$$sinx = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ... \text{ and } e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

whose infinite series expansion exist, are used in the calculations. Because, incalculating the value of such function for an assigned value of x, only a finite 3number of terms can be taken, an error get introduced for not considering the remaining terms.

III. Rounding error/Round-off error:

When the rational numbers like $1/3; 22/7; 5/9; 8/9$ etc, whose decimal representation involve infinite number of digits, are involved in our calculations, we are forced to take only a few number of digits from their decimal expression and thus an error named round-off error gets involved. There are universal rules for rounding a number as rounding rules.

IV. Inherent error /error of the operation

When performing computations with algebraic operations among approximate numbers, we naturally carry to some extent the errors of the original data into final result. Such errors are called inherent error/error of the operation. For example, let $x = 0.3333$ and $y = 3.1416$ be two approximate numbers for the exact number $1/3$ and π. Obviously, if we perform an algebraic operation between these two approximate numbers, the error will introduce in the final

result accordingly.

1.3 Significant Digit

In the decimal representation of a number, a digit is said to be significant if it is either a non-zero digit or any zero (s) lying between two non-zero digits are used as a placeholder, to indicate a retrained place. All other zeros used to fix-up the position of the decimal point are not to be counted as significant digit.

The number of significant digits in a number will be counted from the left-most non-zero digit towards right. Thus, the numbers 0.7452 and 0.007452 both have four significant digits. Similarly,

- 0.00400300 has 6 significant digits
- 0.4003000 has 7 significant digits
- 0.30040000 has 8 significant digits

1.4 Rules for Rounding off numbers

While performing any algebraic operation between two or more numbers written in the decimal number system, it is often required to round-off these numbers, that is, replace each of them having a smaller of significant digits. The rule for doing this as follows:

To round off a number to n significant figures, discard all digits to the right of the n^{th} place; if the discarded number is less than half a unit in the $(n+1)^{th}$ place, leave the n^{th} digit unchanged; if the discarded number is greater than half a unit in the $(n+1)^{th}$ place, add 1 to the n^{th} digit; if the discarded number is exactly half a unit in the $(n+1)^{th}$ place, leave the n^{th} digit unaltered if it is an even number, but increase it by 1 if it is an odd number.

Correct digit: In the decimal representation of an approximate number, the n^{th} digit after decimal is said to be correct if the absolute error does not exceed one half unit in the n^{th} place. For example,

let an exact number be $A = 35.974$ and its approximate value $a = 36.0$,
then the absolute error $|A - a| = 0.026$

Since $0.026 < 0.05 = \frac{1}{2} \times 10^{-1} \leq 0.5 \times 10^{-n}$ for $n = 1$, the approximation 36.0 for 35.974 is correct to one decimal place.

Example 1.1 *Round off 37.897456 correct to 5 significant figures.*

Solution: *Discard all digits to the right of the* 5^{th} *place, which is in this case 456. For convenience and better understanding, assume that the discarded number is* 0.456 *(whatever digits are being discarded, put a decimal before that). Now this number* 0.456 *is less than half a unit, which is* 0.5*;*

That is $0.456 < 0.5$ *and hence leaving the* 5^{th} *place digit unchanged 37.897456 become* 37.897 *(correct to 5 significant figures).*

Example 1.2 *Round off 28.244 795 correct to 5 significant figures.*

Solution: *Discard all the digits to the right of the 5th place, which is in this case 795. For convenience and better understanding, assume that the discarded number is* 0.795 *(whatever digits are being discarded, put a decimal before that). Now this number* 0.795 *is greater than half a unit, which is* 0.5*;*

That is, $0.795 > 0.5$ *and hence we add* 1 *to the* 5^{th} *place digit.* 28.244795 *become* 28.245 *(correct to 5 significant figures).*

Example 1.3 *Round off* 6.000559 *correct to* 4 *significant figures.*

Solution: *Discard all the digits to the right of the* 4^{th} *place, which is in this case* 559*. For convenience and better understanding, assume that the discarded number is* 0.559 *(whatever digits are being discarded, put a decimal before that). Now this number 0.559 is greater than half a unit, which is* 0.5*;*

That is, $0.559 > 0.5$ *and hence we add* 1 *to the* 4^{th} *place digit.* 6.000559 *become* 6.001 *(correct to 4 significant figures).*

Example 1.4 *Round off* 6.002500 *correct to* 4 *significant figures.*

Solution: *Discard all the digits to the right of the* 4^{th} *place, which is in this case* 500*. For convenience and better understanding, assume that the discarded number is* 0.500 *(whatever digits are being discarded, put a decimal before that). Now this number* 0.500 *is exactly equal to half a unit, which is* 0.5*;*

That is $0.500 = 0.5$ *and since the* 4^{th} *place digit is* 2 *(even), we keep the* 4^{th} *significant digit unaltered. That is* 6.002500 *become* 6.002 *(correct to 4 significant figures).*

Example 1.5 *Round off 5.001500 correct to 4 significant figures.*

Solution: *Discard all the digits to the right of the* 4^{th} *place, which is in this*

case 500. Assume that the discarded number is 0.500 (whatever digits are being discarded, put a decimal before that). Now this number 0.500 is exactly equal to half a unit, which is 0.5;

That is, 0.500=0.5 and since the 4th place digit is 1(odd), we add 1 to the 4^{th} *place digit. That is, 5.001500 become 5.002 (correct to 4 significant figures).*

1.5 Exercises for Sect.1-4

1. Round off 0.070000123456 correct to 4 significant figures.

2. Round-off the following numbers correct to 4- significant figures:
   (i) 4.79132, (ii) 23.2975, (iii) 0.000956754, (iv) 0.0082665, (v) 4378.562, (vi) 2.000469, (vii) 3.35008, (viii) 87.5555, (ix) 4.000559, (x) 0.0000300085

3. Round-off the following numbers correct to 4- significant figures:
   (i) 3.96312, (ii) 49.00088, (iii) 0.0000656. (iv) 76.000654, (v) 28.555555, (vi) 5.00109600, (vii) 538.29059, (viii) 0.0129456, (ix) 1.45008, (x) 8.999999

4. If $f(x) = 5tan(x) - 9x$, find the percentage error in $f(x)$ for $x = \frac{\pi}{4}$, if the error in x is 0.003.

5. Find the sum of the following approximate numbers, if the numbers are correct to the last digit:
   a) 8.37642, 6.876, 2.896555, 0.22359;
   b) 0.00024356, 19.42, 8.987, 0.35793, 173682, 3.73928.

6. a) Find the absolute error (E_A), relative error (E_R) and percentage error (E_P) of the numbers whose true value (V_T) and approximate value (V_A) are given:
   i) $(V_T) = 5.86593$, (V_A)=5.866;
   ii) $(V_T) = 2.55555$, $(V_A) = 2.556$;
   iii) (V_T)= 9.75600, (V_A)=9.757.

   b) Given $E_A = 0.510^{-3}$, $(E_R) = 0.3710^{-5}$, find (V_T).
   c) If $V_A = 0.468, EP = 6\%$, find (V_T) .

Lecture 2
(a) Important theorems on absolute error and relative error with proof
(b) General expression for error.

1.6 Two important theorems

Theorem 1.1 *The maximum absolute error of an algebraic sum or difference of several approximate numbers does not exceed to the sum of absolute error of the numbers, that is, if* $x_i(i = 1, 2, \cdots, n)$ *be the* n *approximate numbers and* u *is their algebraic sum or difference, then*

$$\triangle u \leq \triangle x_1 + \triangle x_2 + ... + \triangle x_n$$

We prove the results for two numbers:

a)For sum:
Let n_1, n_2 be approximate numbers to N_1, N_2 with errors E_1, E_2, so that
$N_1 = n_1 + E_2, N_2 = n_2 + E_2$, then
$N_1 + N_2 = (n_1 + E_1) + (n_1 + E_2) = (n_1 + n_2) + (E_1 + E_2)$
Let $N_1 + N_2 = N, n_1 + n_2 = n$ and $N = n + E$, then
$N = n + (E_1 + E_2)$
$N - n = E_1 + E_2$
$E = E_1 + E_2$
or, $|E| = |E_1 + E_2|$
or, $|E| \leq |E_1| + |E_2| \cdots\cdots\cdots (A)$

b)For difference:
$N_1 - N_2 = (n_1 - n_2) + (E_1 - E_2)$
Let $N_1 - N_2 = N, n_1 - n_2 = n$ and $N = n + E$,
then $N = n + (E_1 - E_2)$
$N - n = E_1 - E_2$
$E = E_1 + (-E_2)$
$|E| \leq |E_1| + |-E_2|$
or $|E| \leq |E_1| + |E_2| \cdots\cdots\cdots (B)$

(A) and (B) show that the absolute errors in the sum or difference of two numbers does not exceed the sum of their absolute errors. Similarly the result can be extended for three and more approximate numbers.

Theorem 1.2 *The maximum relative error for both multiplication and division does not exceed to the algebraic sum of their relative errors.*

Proof:

(a) For multiplication:

Let n_1, n_2 be approximate numbers toN_1, N_2 with errors E_1, E_2, so that $N_1N_2 = N, n_1n_2 = n, N = n + E$. Then,

$N_1N_2 = (n_1 + E_1)(n_2 + E_2) = n_1n_2 + E_1n_2 + n_2E_2 + E_1E_2,$

$N_1N_2 - n_1n_2 \equiv E = E_1n_2 + E_2n_2 + E_1E_2.$

Neglecting the second order term E_1E_2 and dividing both sides by n_1n_2 i.e. n, one gets

$$\frac{E}{n} = \frac{E_1}{n_1} + \frac{E_2}{n_2} \Rightarrow |\frac{E}{n}| \leq |\frac{E_1}{n_1}| + |\frac{E_2}{n_2}| \tag{1.1}$$

(b) For division:

let $\frac{N_1}{N_2} = N$, $\frac{n_1}{n_2} = n$ and $N - n = E$, then

$$\frac{N_1}{N_2} = \frac{n_1 + E_1}{n_2 + E2} = \frac{n_1}{n_2}(1 + \frac{E_1}{n_1})(1 + \frac{E_2}{n_2})^{-1}$$

$$N \cong n(1 + \frac{E_1}{n_1})(1 - \frac{E_2}{n_2}) = n(1 + \frac{E_1}{n_1} - \frac{E_2}{n_2})$$

where we have used the binomial theorem and neglected the second and higher order terms 9assuming they are small). Then

$$\frac{N - n}{n} \equiv \frac{E}{n} = \frac{E_1}{n_1} - \frac{E_2}{n_2}$$

$$\Rightarrow |\frac{E}{n}| = |\frac{E_1}{n_1} - \frac{E_2}{n_2}| \leq |\frac{E_1}{n_1}| + |\frac{E_2}{n_2}| \tag{1.2}$$

Inequalities (1.1) and (1.2) shows that relative error in the product of two approximate numbers is always less than the algebraic sum of their relative errors.

Thus we can evaluate the errors in an arithmetic operation between two or more numbers/quantities. But the method becomes difficult when the number of 10quantities are large and many algebraic operations, together with mathematical functions, are involved in the calculation. **For such cases, we proceed as follows:**

Let $u = f(x_1, x_2, ..., x_n)$ be a function of n variables $x_1, x_2, ..., x_n$. Let ε_i be the error in the variable x_i for $i = 1, 2, ..., n$ and be the change, ie, error in u,

when x_i becomes $x_i + \varepsilon_i$ for $i = 1, 2, ..., n$. Then,

$$\begin{aligned} u + \varepsilon &= f(x_1 + \varepsilon_1, x_2\varepsilon_2, ..., x_n\varepsilon_n) \\ &= f(x_1, x_2, ..., x_n) + (\varepsilon_1 \frac{\partial}{\partial x_1} + \varepsilon_2 \frac{\partial}{\partial x_2} + ... + \varepsilon_n \frac{\partial}{\partial x_n}) f(x_1, x_2, ..., x_n) \\ &+ \frac{1}{2!}(\varepsilon_1 \frac{\partial}{\partial x_1} + \varepsilon_2 \frac{\partial}{\partial x_2} + ... + \varepsilon_n \frac{\partial}{\partial x_n})^2 f(x_1, x_2, ..., x_n) + ... \end{aligned}$$

(Taylor's series expansion of n- variables)

As ε_i $(i = 1, 2, ..., n)$ is small, neglecting their second and higher order terms, one gets

$$\varepsilon \approx \varepsilon_1 \frac{\partial f}{\partial x_1} + \varepsilon_2 \frac{\partial f}{\partial x_2} + ... + \varepsilon_n \frac{\partial f}{\partial x_n}$$

$$\Rightarrow |\varepsilon| \leq |\varepsilon_1 \frac{\partial f}{\partial x_1}| + |\varepsilon_2 \frac{\partial f}{\partial x_2}| + ... + |\varepsilon_n \frac{\partial f}{\partial x_n}|$$

Now, the maximum absolute error in u is given by

$$max|\varepsilon| \leq |\varepsilon_1||\frac{\partial f}{\partial x_1}| + |\varepsilon_2||\frac{\partial f}{\partial x_2}| + ... + |\varepsilon_n||\frac{\partial f}{\partial x_n}|$$

that is,

$$max|\varepsilon| \leq \sum_{i=1}^{n} |\varepsilon_i||\frac{\partial f}{\partial x_i}|$$

The relative error in u will be between $\frac{|\varepsilon|}{u \pm \varepsilon|}$ and maximum relative error will be between $\frac{max|\varepsilon|}{u \pm max\varepsilon|}$.

1.7 Miscellaneous examples.

Lecture 3

Example 1.6 *Round off 7.001500002 correct to 4 significant figures*

Solution: Discard all the digits to the right of the nth place, which is in this case 500002. For convenience and better understanding, assume that the discarded number is 0.500002 (whatever digits are being discarded, put a decimal before that). Now this number 0.500002 is greater than half a unit, which is 0.5;

That is $0.500002 > 0.5$ and hence we add 1 to the nth place digit. 7.001500002 become 7.002 (correct to 4 significant figures).

Example 1.7 *Round off 0.000000123456 correct to 4 significant figures.*

Solution: First of all in this example the number starts with zeros and hence they are not significant. Discard all the digits to the right of the nth place, which is in this case 56 (the starting zeros are not considered). For convenience and better understanding, assume that the discarded number is 0.56 (whatever digits are being discarded, put a decimal before that). Now this number 0.56 is greater than half a unit, which is 0.5;

That is $0.56 > 0.5$ and hence we add 1 to the nth place digit. 0.000000123456 become 0.0000001235 (correct to 4 significant figures).

Example 1.8 *Let $u = \frac{x^2y^3}{z}$ given that $x = 2.0, y = 3.0, z = 4.0$ and errors are $e_x = 0.02, e_y = 0.03, e_z = 0.04$. Find the absolute and maximum absolute error in u. Also find the range of relative and maximum relative error in u.*

Solution: We have $|\varepsilon| = |e_x \frac{\partial f}{\partial x} + e_y \frac{\partial f}{\partial y} + e_z \frac{\partial f}{\partial z}|$, where
$\frac{\partial f}{\partial x} = \frac{2xy^3}{z} = 4 \times 27 = 108$, $\frac{\partial f}{\partial y} = \frac{3x^2y^2}{z} = 108$, $\frac{\partial f}{\partial z} = -\frac{x^2y^3}{z^2} = -108$,
$|\varepsilon| = |(0.02 - 0.03 - 0.04) \times 108| = |-0.5 \times 108| = |-5.40| = 5.4$
$max|\varepsilon| \leq |0.9 \times 108| = 9.72$.
For given x, y and z,the exact value$u = 108$,
Relative error in u lies between $\frac{5.4}{108 \pm 5.4}$, that is, between 0.048 and 0.053; max., relative error lies between $\frac{9.72}{108 \pm 9.72}$, that is, between 0.0826 and 0.0990.

Example 1.9 *Find the sum of the following approximate numbers, if the numbers are correct to the last digit:*
(i). 4.23721, 3.96, 0.00123.0.12359,
(ii). 432.9618, 29.42, 321.0, 68.243, 17.482

Solution:
(i) Since the second number is known only to the decimal places it is unwise to retain more than four decimal in the other numbers (two decimal places more relative to 0.396). We add the numbers by rounding-off to two decimal and then rounding off the sum to one decimal place.
$4.2372 + 3.96 + 0.0012 + 0.1236 = 8.3220 = 8.32$

(ii) Since the third number is known only to the first decimal place it is unwise to retain more than three decimal in the other number (two decimal places more relative to 321.0). We add the numbers by rounding-off to two decimals and then rounding-off the sum to one decimal place.
$423.962 + 29.42 + 321.0 + 68.243 + 17.482 = 869.107 = 869.1$

Example 1.10 *Find the absolute error* (E_A)*, relative error* (E_R) *and percentage error* (E_P) *of the number whose true value* (V_T) *and approximate value* (V_A) *are given:*
(i) $V_T = 8.42759, V_A = 8.428$*,*
(ii) $V_T = 5.81728, V_A = 5.817$*.*

Solution:
(i) E_A = Absolute error =|True Value −approximate value|,
so that, $E_A = |V_T - V_A| = |8.42759 - 8.428| = |-0.00041| = 0.00041$.
E_R = relative error=$\frac{Absolute\ error}{True\ Value} = |\frac{V_T - V_A}{V_T}| = \frac{0.00041}{8.42759} = 0.00004865$,
E_p=Percentage error = Relative error$\times 100 = E_R \times 100 = 0.004865$.

(ii)E_A = Absolute error =$|V_T - V_A| = |5.81728 - 5.817| = |0.00028| = 0.00028$,
E_R = relative error=$|\frac{V_T - V_A}{V_T}| = \frac{0.00028}{5.81728} = 0.00004813$,
E_p=Percentage error = Relative error$\times 100 = E_R \times 100 = 0.004813$.

Example 1.11 *If* $V_T = 2.73456, E_A = 0.210^{-2}$*, find* E_R, E_P*.*

Solution:
$E_R = |\frac{V_T - V_A}{V_T}| = \frac{0.210^{-2}}{2.73456} = 0.000731$,
$E_p = E_R \times 100 = 0.0731$.

Example 1.12 *Find the relative error in computation of x+y for x = 20.01 and y=1.89 having absolute errors x = 0.005 and y = 0.002 respectively.*

Solution:
Here $x + y = 20.01 + 1.89 = 22.90$, which has an absolute error

$$\triangle x + \triangle y = 0.005 + 0.002 = 0.007,$$

Therefore, the relative error in $x+y$: $\varepsilon_R(x+y) = \frac{\triangle x+\triangle y}{x+y} = \frac{0.007}{22.90} = 0.000305676 \approx 0.0003057$.

Example 1.13 *Find the relative error in computation $x - y$ for x=10.91 and y= 4.37 having absolute error $\triangle x = 0.008$ and $\triangle y = 0.003$ respectively.*

Solution:
Here the absolute error in x: $\triangle x = 0.008$,
Therefore, the relative error in x: $\varepsilon_x = \frac{\triangle x}{x} = \frac{0.008}{10.91} = 0.0007333$,
Similarly, the relative error in y: $\varepsilon_y = \frac{\triangle y}{y} = \frac{0.003}{4.37} = 0.0006865$
Therefore, the relative error in $x - y$: $\varepsilon = 0.0007333 - 0.0006865 = 0.0000468$.

Example 1.14 *If $y = 7x^7 - 3x^3$, find the percentage error in y at x=1, if the error in x is $\epsilon_x = 0.05$.*

Solution:
$\triangle y = \epsilon_x \times (49x^6 - 9x^2) = 0.05 \times (49 - 9) = 2$,
The percentage error in y: $E_p = \frac{\triangle y}{y} \times 100 = \frac{2}{4} \times 100 = 50\%$.

Example 1.15 *Evaluate $f(x) = \frac{1}{\sqrt{1+x^2}-\sqrt{1-x^2}}$ for small $x = 0.02$ using 3 digit arithmetic with rounding.*

Solution:
Here, $x^2 = 0.0004, 1 + x_2 = 1.0004 = 1.000$, taking 3 significant digits.
Similarly,
$1 - x^2 = 1 - 0.0004 = 0.9996 = 1.000$, by 3 digits with rounding, then
$\sqrt{1+x^2} - \sqrt{1-x^2} = 1 - 1 = 0$.
Therefore, $f(x) = \frac{1}{\sqrt{1+x^2}-\sqrt{1-x^2}}$ cannot be computed in this present form because denominator is zero, which is due to subtraction of nearly equal numbers.

To remove this, we rewrite the function in equivalent mathematical form, which does not involve subtraction of nearly equal numbers. Thus $f(x)$ can be written as
$f(x) = \frac{\sqrt{1+x^2}+\sqrt{1-x^2}}{(1+x^2)-(1-x^2)} = \frac{\sqrt{1+x^2}+\sqrt{1-x^2}}{2x^2} = \frac{1+1}{2\times 0.0004} = \frac{1}{0.0004} = 2500$

1.8 Exercises

1. Find the relative error in computation of $x+y$ for $x = 13.64$ and $y = 5.28$ having absolute errors $x = 0.005$ and $y = 0.002$ respectively
2. Find the relative error in computation of $x-y$ for $x = 20.91$ and $y = 9.27$ having absolute errors $x = 0.003$ and $y = 0.001$ respectively.
3. Find the value of $\sqrt(10) - \pi$ correct to five significant figures.
4. Solve the equation $x^2 + 9.9x - 1 = 0$ by using two digit arithmetic with rounding.

Chapter 2

Solutions of System of Linear Equations

Lecture 1

In this section, we shall discuss about the numerical computation of the solution of a system of n linear equations of the form

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

where a_{ij}'s $(i, j = 1, 2, ..., n)$ are the coefficients of the unknowns $x_1, x_2, ..., x_n$ and b_i's are constants.

In matrix notation, this can be written in the form

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}$ is a $n \times n$ matrix, $\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}$ is a $n \times 1$ matrix of unknowns and $\mathbf{b}=\begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{pmatrix}$ is a $n \times 1$ matrix of prescribed constants.

We assume that $det(\mathbf{A}) \neq 0$, so that the system of n linear equations in n-unknowns has a unique solution. Our aim is to compute n-unknown components $x_1, x_2, ..., x_n$ up to desired degree of accuracy.

The methods for solving the system of linear equation can be categorized into two groups:

1. **Direct Method**(or exact method): where we obtain the solution through a finite number of arithmetic operations, for example, Gauss Elimination method, Crout's method.

2. **Iterative Method**: where a sequence of successive approximations, obtained, which converges to the required solution, up to some desired degree of accuracy, for example, Jacobi's method, Gauss Seidel method.

2.1 Direct Methods

2.1.1 Gauss elimination method

It is a direct method for finding the solution or the values of unknown of a system of linear equations and is based on the principle of elimination of unknown in successive steps. We first discuss the method considering n-equations and then we shall consider in particular, 3-equations with 3-unknowns.

We consider a system of n-linear equations with n-unknowns as:

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 ... + a_{1,n-1}^{(1)}x_{n-1} + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 + ... + a_{2,n-1}^{(1)}x_{n-1} + a_{2n}^{(1)}x_n &= b_2^{(1)} \\
a_{31}^{(1)}x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 + ... + a_{3,n-1}^{(1)}x_{n-1} + a_{3n}^{(1)}x_n &= b_3^{(1)} \\
&..... \\
a_{n-1,1}^{(1)}x_1 + a_{n-1,2}^{(1)}x_2 + a_{n-1,3}^{(1)}x_3 + a_{n-1,4}^{(1)}x_4 + ... + a_{n-1,n-1}^{(1)}x_{n-1} + a_{n-1,n}^{(1)}x_n &= b_{n-1}^{(1)} \\
a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + a_{n4}^{(1)}x_4 + ... + a_{n,n-1}^{(1)}x_{n-1} + a_{nn}^{(1)}x_n &= b_n^{(1)}
\end{aligned} \tag{2.1}$$

where $a_{ij}^{(1)}$'s $(i, j = 1, 2, ..., n)$ are the coefficients of the unknowns $x_1, x_2, ..., x_n$ and $b_i^{(1)}$'s $(i = 1, 2, ..., n)$ are prescribed constants.

Let $a_{11}^{(1)} \neq 0$. Now multiplying the first equation successively by $-\frac{a_{21}^{(1)}}{a_{11}^{(1)}} = m_{21}, -\frac{a_{31}^{(1)}}{a_{11}^{(1)}} = m_{31}, -\frac{a_{41}^{(1)}}{a_{11}^{(1)}} = m_{41}, ..., -\frac{a_{n-1,1}^{(1)}}{a_{11}^{(1)}} = m_{n-1,1}, -\frac{a_{n1}^{(1)}}{a_{11}^{(1)}} = m_{n1}$ and adding respectively to $2^{nd}, 3^{rd}, 4^{th}, ..., (n-1)^{th}$ and n^{th} equations of the

system (2.1), we get

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 ... + a_{1,n-1}^{(1)}x_{n-1} + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + a_{24}^{(1)}x_4 + ... + a_{2,n-1}^{(2)}x_{n-1} + a_{2n}^{(2)}x_n &= b_2^{(2)} \\
a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 + ... + a_{3,n-1}^{(2)}x_{n-1} + a_{3n}^{(2)}x_n &= b_3^{(2)} \\
&..... \\
a_{n-1,2}^{(2)}x_2 + a_{n-1,3}^{(2)}x_3 + a_{n-1,4}^{(2)}x_4 + ... + a_{n-1,n-1}^{(2)}x_{n-1} + a_{n-1,n}^{(2)}x_n &= b_{n-1}^{(2)} \\
a_{n2}^{(2)}x_2 + a_{n3}^{(2)}x_3 + a_{n4}^{(2)}x_4 + ... + a_{n,n-1}^{(2)}x_{n-1} + a_{nn}^{(2)}x_n &= b_n^{(2)}
\end{aligned} \tag{2.2}$$

where

$a_{22}^{(2)} = a_{22}^{(1)} - \frac{a_{12}^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}} = a_{22}^{(1)} + m_{21}a_{12}^{(1)}$, $a_{23}^{(2)} = a_{23}^{(1)} - \frac{a_{13}^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}} = a_{22}^{(1)} + m_{21}a_{13}^{(1)}, ...$

$a_{32}^{(2)} = a_{32}^{(1)} - \frac{a_{12}^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}} = a_{32}^{(1)} + m_{31}a_{12}^{(1)}$, $a_{33}^{(2)} = a_{33}^{(1)} - \frac{a_{13}^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}} = a_{33}^{(1)} + m_{31}a_{13}^{(1)}, ...$

.....

$a_{n2}^{(2)} = a_{n2}^{(1)} - \frac{a_{12}^{(1)} \cdot a_{n1}^{(1)}}{a_{11}^{(1)}} = a_{n2}^{(1)} + m_{n1}a_{12}^{(1)}$, $a_{n3}^{(2)} = a_{n3}^{(1)} - \frac{a_{13}^{(1)} \cdot a_{n1}^{(1)}}{a_{11}^{(1)}} = a_{n3}^{(1)} + m_{n1}a_{13}^{(1)}, ...$

$b_2^{(2)} = b_2^{(1)} - \frac{b_1^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}} = b_2^{(1)} + m_{21}b_1^{(1)}$, $b_3^{(2)} = b_3^{(1)} - \frac{b_1^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}} = b_3^{(1)} + m_{31}b_1^{(1)}, ...$

From the system (2.2), it is clear that, except the first equation, the rest $(n-1)$ equations are free from the unknown x_1.

Again assuming $a_{22}^{(1)} \neq 0$, multiplying second equation of the system (2.2) successively by

$-\frac{a_{32}^{(2)}}{a_{22}^{(2)}} = m_{32}, -\frac{a_{42}^{(2)}}{a_{22}^{(2)}} = m_{42}, ..., -\frac{a_{n-1,1}^{(2)}}{a_{22}^{(2)}} = m_{n-1,2}, -\frac{a_{n1}^{(2)}}{a_{22}^{(2)}} = m_{n2}$

and adding respectively to $2^{nd}, 3^{rd}, 4^{th}, ..., (n-1)^{th}$ and n^{th} equations of the system (2.2), we get

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 ... + a_{1,n-1}^{(1)}x_{n-1} + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + a_{24}^{(1)}x_4 + ... + a_{2,n-1}^{(2)}x_{n-1} + a_{2n}^{(2)}x_n &= b_2^{(2)} \\
a_{33}^{(3)}x_3 + a_{34}^{(3)}x_4 + ... + a_{3,n-1}^{(3)}x_{n-1} + a_{3n}^{(3)}x_n &= b_3^{(3)} \\
&..... \\
a_{n-1,3}^{(3)}x_3 + a_{n-1,4}^{(3)}x_4 + ... + a_{n-1,n-1}^{(3)}x_{n-1} + a_{n-1,n}^{(3)}x_n &= b_{n-1}^{(3)} \\
a_{n3}^{(3)}x_3 + a_{n4}^{(3)}x_4 + ... + a_{n,n-1}^{(3)}x_{n-1} + a_{nn}^{(3)}x_n &= b_n^{(3)}
\end{aligned} \tag{2.3}$$

Here also, we observe that $3^{rd}, 4^{th}$ up to n^{th} equations of the system (2.3) are free from the unknowns x_1, x_2.

Repeating the same procedure of elimination of the unknowns, lastly we get a system of equation which is equivalent to the system (2.1) as:

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 ... + a_{1,n-1}^{(1)}x_{n-1} + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + a_{24}^{(1)}x_4 + ... + a_{2,n-1}^{(2)}x_{n-1} + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 + a_{34}^{(3)}x_4 + ... + a_{3,n-1}^{(3)}x_{n-1} + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ &..... \\ a_{n-1,n-1}^{(n-1)}x_{n-1} + a_{n-1,n}^{(n-1)}x_n &= b_{n-1}^{(n-1)} \\ a_{nn}^{(n)}x_n &= b_n^{(n)} \end{aligned} \tag{2.4}$$

The non-zero (by assumption) coefficients $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, ..., a_{nn}^{(n)}$ of the system of equations are known as ***pivots*** and the corresponding equations are known as ***pivotal equations***.

Please note if any of the coefficients $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, ..., a_{nn}^{(n)}$ are zeros, then the system has to be reshuffled so that they are non-zeros.

Now we can get easily calculate the solution of the system of equations (2.4) as follows: First we find x_n from n^{th} equation, then x_{n-1} from $(n-1)^{th}$ equation after substituting x_n and then successively we shall get all the unknowns $x_1, x_2, ..., x_n$ (by the method of back substitution).

Gauss Elimination Method (Particular Case)
In this article we now consider a system of 3-equations with 3-unknowns for better illustration to the readers.

A system of 3-equations with 3-unknowns is given by

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{31}^{(1)}x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 &= b_3^{(1)} \end{aligned} \tag{2.5}$$

where $a_{ij}^{(1)}$'s $(i, j = 1, 2, 3)$ and $b_i^{(1)}$'s $(i = 1, 2, 3)$ are known constants.

Let $a_{11}^{(1)} \neq 0$. Multiplying the first equation of (2.5) successively by $-\frac{a_{21}^{(1)}}{a_{11}^{(1)}}$ and $-\frac{a_{31}^{(1)}}{a_{11}^{(1)}}$, adding respectively with 2^{nd} and 3^{rd} equation we get, the system

as:

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 &= b_3^{(2)} \end{aligned} \tag{2.6}$$

where

$a_{22}^{(2)} = a_{22}^{(1)} - \frac{a_{12}^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}}$, $a_{23}^{(2)} = a_{23}^{(1)} - \frac{a_{13}^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}}$, $b_2^{(2)} = b_2^{(1)} - \frac{b_1^{(1)} \cdot a_{21}^{(1)}}{a_{11}^{(1)}}$,

$a_{32}^{(2)} = a_{32}^{(1)} - \frac{a_{12}^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}}$, $a_{33}^{(2)} = a_{33}^{(1)} - \frac{a_{13}^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}}$,$b_3^{(2)} = b_3^{(1)} - \frac{b_1^{(1)} \cdot a_{31}^{(1)}}{a_{11}^{(1)}}$

Let $a_{22}^{(2)} \neq 0$. Multiplying the second equation by $-\frac{a_{32}^{(2)}}{a_{22}^{(2)}}$ and adding with the 3^{rd} equation of the system (2.6) we get,

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ a_{33}^{(3)}x_3 &= b_3^{(3)} \end{aligned} \tag{2.7}$$

where $a_{33}^{(3)} = a_{33}^{(2)} - \frac{a_{23}^{(2)} \cdot a_{32}^{(2)}}{a_{22}^{(2)}}$.

The non-zero constants (by assumption) $a_{11}^{(1)}, a_{22}^{(2)}$ and $a_{33}^{(3)}$ are called *pivots* and the corresponding equations are called the pivotal equations.

Now the value of the unknown x_3 can be obtain easily from the third equation, which can be substituted in the second equation to obtain x_2 . Substituting x_3, x_2 in the first equation, x_1 also be determined. Thus, all the unknown are completely known by the method of back substitution.

Lecture 2

2.1.2 Gauss Elimination in matrix notation

In matrix-form, the system of linear equations can be written in the form

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A}=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & ... & a_{2n}^{(1)} \\ ... & ... & ... & ... \\ a_{n1}^{(1)} & a_{n2}^{(1)} & ... & a_{nn}^{(1)} \end{pmatrix}$ is a $n \times n$ matrix, $\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}$ is a $n \times 1$ matrix of unknowns and $\mathbf{b}=\begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ ... \\ b_n^{(1)} \end{pmatrix}$ is a $n \times 1$ matrix of prescribed constants.

The augmented matrix is given by

$$(\mathbf{A},\mathbf{b})=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & ... & a_{2n}^{(1)} & b_2^{(1)} \\ ... & ... & ... & ... & ... \\ a_{n1}^{(1)} & a_{n2}^{(1)} & ... & a_{nn}^{(1)} & b_n^{(1)} \end{pmatrix}$$

If $a_{11}^{(1)} \neq 0$ then multiplying the first row by $m_{i1} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ $(i = 2, 3, ..., n)$, which is called the **multiplier**, we add them to the corresponding elements of the i^{th} row $(i = 2, 3, ..., n)$, which means that the new values are gives as follows:

$$new\ a_{ij} = old\ a_{ij} + m_{i1}a_{1j}$$
$$new\ b_i = old\ b_i + m_{i1}b_1$$
$$new\ a_{i1} = 0, i, j = 2, 3, ..., n$$

Thus, the new augmented matrix is given by

$$(\mathbf{A},\mathbf{b}) \rightarrow \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & ... & a_{2n}^{(2)} & b_2^{(2)} \\ ... & ... & ... & ... & ... \\ 0 & a_{n2}^{(2)} & ... & a_{nn}^{(2)} & b_n^{(2)} \end{pmatrix}$$

If $a_{22}^{(2)} \neq 0$, then we multiply the second row by the multiplier $m_{i2} = -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}} (i = 3, 4, ..., n)$ and add them to the corresponding element of the i^{th} row $(i =$

$3, 4, ..., n)$ respectively and like before,

$$new\ a_{ij} = old\ a_{ij} + m_{i2}a_{2j}$$
$$new\ b_i = old\ b_i + m_{i2}b_2$$
$$new\ a_{i2} = 0, i, j = 3, 4, ..., n$$

The process is repeated and after $n-1$ steps we get the augmented matrix as

$$(\mathbf{A,b}) \rightarrow \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & ... & a_{2n}^{(2)} & b_2^{(2)} \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & a_{nn}^{(n)} & b_n^{(n)} \end{pmatrix}$$

and we can get the value of the unknowns $x_1, x_2, ..., x_n$ by the method of back substitution.

Example 2.1 *Solve the system of linear equations using Gauss Elimination method:*

$$\begin{aligned} 0.4x + 5y + 2z &= 7.4 \\ 6x - 0.3y + 4z &= 9.7 \\ -5x + 2y + 0.8z &= -2.2 \end{aligned}$$

Solution: In matrix form, the equations can be written as **Ax=b**, where

$\mathbf{A}= \begin{pmatrix} 0.4 & 5 & 2 \\ 6 & -0.3 & 4 \\ -5 & 2 & 0.8 \end{pmatrix}$,$\mathbf{x}= \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $\mathbf{b}= \begin{pmatrix} 7.4 \\ 9.7 \\ -2.2 \end{pmatrix}$

The augmented matrix is given by

$(\mathbf{A,b}) = \begin{pmatrix} 0.4 & 5 & 2 & 7.4 \\ 6 & -0.3 & 4 & 9.7 \\ -5 & 2 & 0.8 & -2.2 \end{pmatrix}$, $(a_{11}^{(1)} = 0.4 \neq 0, we\ divide\ the\ first\ row\ by\ 0.4)$

$\rightarrow \begin{pmatrix} 1 & 12.5 & 5 & 18.5 \\ 6 & -0.3 & 4 & 9.7 \\ -5 & 2 & 0.8 & -2.2 \end{pmatrix}$ (Notation: $R_1^{'} = \frac{1}{0.4}R_1$)

$\rightarrow \begin{pmatrix} 1 & 12.5 & 5 & 18.5 \\ 0 & -75.3 & -26 & -101.3 \\ 0 & 64.5 & 25.8 & 90.3 \end{pmatrix}$ $(R_2^{'} = R_2 - 6R_1, R_3^{'} = R_3 + 5R_1)$

$\rightarrow \begin{pmatrix} 1 & 12.5 & 5 & 18.5 \\ 0 & 1 & 0.3453 & 1.3453 \\ 0 & 64.5 & 25.8 & 90.3 \end{pmatrix}$ $(R_2^{'} = \frac{-1}{75.3}R_2)$

$$\rightarrow \begin{pmatrix} 1 & 12.5 & 5 & 18.5 \\ 0 & 1 & 0.3453 & 1.3453 \\ 0 & 0 & 3.52815 & 3.52815 \end{pmatrix} (R_3^{'} = R_3 - 64.5R_2)$$

The pivotal equations are

$$\begin{aligned} x + 12.5y + 5z &= 18.5 \\ y + 0.3453z &= 1.3453 \\ 3.52815z &= 3.52815 \end{aligned}$$

By back substitution, we get, $z = 1, y = 1, x = 1$.

Example 2.2 *Solve the system of linear equations by Gauss Elimination method:*

$$\begin{aligned} x + 2y + 3z &= 10 \\ x + 3y - 2z &= 7 \\ 2x - y + z &= 5 \end{aligned}$$

Solution: In matrix form, the equations can be written as **Ax=b**, where

$\mathbf{A}= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & -2 \\ 2 & -1 & 1 \end{pmatrix}$,$\mathbf{x}= \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $\mathbf{b}= \begin{pmatrix} 10 \\ 7 \\ 5 \end{pmatrix}$

The augmented matrix is given by

$$(\mathbf{A},\mathbf{b}) = \begin{pmatrix} 1 & 2 & 3 & 10 \\ 1 & 3 & -2 & 7 \\ 2 & -1 & 1 & 5 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 3 & 10 \\ 1 & 3 & -2 & 7 \\ 2 & -1 & 1 & 5 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 3 & 10 \\ 0 & 1 & -5 & -3 \\ 0 & -5 & -5 & -15 \end{pmatrix} (R_2^{'} = R_2 - R_1, R_3^{'} = R_3 - 2R_1)$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 3 & 10 \\ 0 & 1 & -5 & -3 \\ 0 & 0 & -30 & -30 \end{pmatrix} (R_3^{'} = R_3 + 5R_2))$$

The pivotal equations are

$$\begin{aligned} x + 2y + 3z &= 10 \\ y - 5z &= -3 \\ -30z &= -30 \end{aligned}$$

By back substitution, we get, $z = 1, y = 2, x = 3$.

2.1.3 Simple Partial Pivoting

Suppose $a_{kk}^{(k)} = 0$. In that case, $a_{kk}^{(k)}$ cannot be used as a pivotal element, that is, it is not possible for this number to eliminate other elements of the first column. Again, if at any step, the pivotal element is very small in magnitude, then the corresponding multiplier is very large numerically. Multiplying the pivoted equation by such large value increase sound off errors and other computational errors in the coefficients and the constants.

Simple partial pivoting takes care of all these problems. In this process, from all the equations, we choose the one with numerically largest coefficient of x_1 and name it as the first pivotal equation, with non-zero pivot $a_{11}^{(1)} \neq 0$. The order of other equations are kept arbitrary. We then eliminate x_1 from the last $(n-1)$ linear equations using the pivotal equation and obtain a system of $(n-1)$ linear equations in $(n-1)$ unknowns, namely, $x_1, x_2, ..., x_n$. We then look for the numerically largest coefficient of x_2 (non-zero) in the reduced system of $(n-1)$ linear equations and name it as the second pivotal equation. The same process is repeated at every subsequent step.

Example 2.3 *Solve the system of linear equations by Gauss Elimination method(Using 4 digits for algebraic operations):*

$$\begin{aligned} 10^{-5}x + 2y &= 1 \\ 2x + 3y &= 2 \end{aligned}$$

Solution: In matrix form, the equations can be written as **Ax=b**, where

$\mathbf{A}= \begin{pmatrix} 10^{-5} & 2 \\ 2 & 3 \end{pmatrix}$,$\mathbf{x}=\begin{pmatrix} x \\ y \end{pmatrix}$, $\mathbf{b}=\begin{pmatrix} 1 \\ 2 \end{pmatrix}$

The augmented matrix is given by

$(\mathbf{A},\mathbf{b}) =\begin{pmatrix} 10^{-5} & 2 & 1 \\ 2 & 3 & 2 \end{pmatrix}$

$\rightarrow \begin{pmatrix} 10^{-5} & 2 & 1 \\ 0 & -4 \times 10^5 & -2 \times 10^5 \end{pmatrix}$ $(R_2^{'} = R_2 - 2 \times 10^5 R_1)$

The pivotal equations are

$$\begin{aligned} 10^{-5}x + 2y &= 1 \\ -4 \times 10^5 y &= -2 \times 10^5 \end{aligned}$$

By back substitution, we get, $y = 0.5, x = 0$. Whereas the true solution is $x = 0.2500, y = 0.5000$. This has happened because the pivotal element 10^{-5} is too small relative to the other coefficients.

We now use (simple) partial pivoting to obtain the solution, that is, we interchange first and second equation's, avoiding 10^{-5} for pivotal element. The second equation is chosen because the coefficient of x is 2, greater than 10^{-5} . Thus, the new augmented matrix is

$$(\mathbf{A},\mathbf{b}) = \begin{pmatrix} 2 & 3 & 2 \\ 10^{-5} & 2 & 1 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 2 & 3 & 2 \\ 0 & 2 & 1 \end{pmatrix} \; (R_2' = R_2 - 0.5 \times 10^{-5} R_1)$$

The pivotal equations are

$$\begin{aligned} 2x + 3y &= 2 \\ 2y &= 1 \end{aligned}$$

By back substitution, we get $y = 0.5, x = 0.25$ (true solution), which shows the usefulness of partial pivoting.

2.1.4 Exercises

1. Solve the linear system by Gauss Elimination method

$$\begin{aligned} x + 3y + 2z &= 5 \\ x - y + z &= -1 \\ x + 2y + 3z &= 2 \end{aligned}$$

$(Ans : x = 1, y = 2, z = -1)$

2. Solve the following system of equations by Gauss-Elimination method, correct to three places of decimals:
   (a)

$$\begin{aligned} 5.091x + 3.455y + 1.091z &= 1.276 \\ 2.818x + 6.455y - 4.273z &= 4.654 \\ 1.273x - 3.091y + 7.545z &= 2.187 \end{aligned}$$

$(Ans : x = -1.992, y = 2.751, z = 1.753)$

   (b)

$$\begin{aligned} 1.660x + 0.684y + 0.820z + 0.380w &= -4.925 \\ 0.784x + 1.690y + 1.396z + 0.492w &= 6.105 \\ 0.754x + 1.602y + 1.608z + 0.456w &= 7.325 \\ 0.442x + 0.570y + 0.338z + 1.398w &= -4.175 \end{aligned}$$

$(Ans : x = -6.069, y = 2.929, z = 5.502, w = -3.592)$

3. Solve , by Gauss-Elimination method, the system (using three digits)

$$\begin{aligned} 0.003x + 4.00y + 5.00z &= 9.003 \\ -3.00x + 3.85y - 6.75z &= -5.900 \\ 4.00x - 5.25y + 3.50z &= -4.750 \end{aligned}$$

   Explain why the solution deviates from true solution $(1, 1, 1)^T$. Use simple partial pivoting and solve the system again. Did you see any difference in the solutions?

Lecture 3

2.1.5 Crout's method

It is a distinct method of solving a system of linear equations of the form **Ax=b**, where the matrix **A** is decomposed into a product of a lower triangular matrix **L** and an upper triangular matrix **U**, that is **A=LU**, thus Crout's method also is called LU decomposition method.

Explicitly, we can write it as

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ ... & ... & ... & ... & ... \\ a_{n1} & a_{n2} & a_{n3} & ... & a_{nn} \end{pmatrix}=\begin{pmatrix} l_{11} & 0 & 0 & ... & 0 \\ l_{21} & l_{22} & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ l_{n1} & l_{n2} & l_{n3} & ... & l_{nn} \end{pmatrix}\begin{pmatrix} 1 & u_{12} & u_{13} & ... & u_{1n} \\ 0 & 1 & u_{23} & ... & u_{2n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{pmatrix}$$

Therefore, by LU-decomposition, the system of linear equations **Ax=b** can be solved in three steps:

1. Construct the lower triangular matrix **L** and upper triangular matrix **U**, then **Ax=b** can be rewritten as **LUx=b**, set **Ux=y**, then **Ly=b**;

2. Using forward substitution, solve **Ly=b** to obtain the solution vector **y**;

3. Solve **Ux=y** by backward substitution method, then the solutions of the original linear can be obtained.

We further elaborate the process by considering a 3×3 matrix **A**.We consider solving the system of equation of the form **Ax=b** , where,

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \mathbf{b}=\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

The matrix **A** is factorized as a product of two matrices **L** (lower triangular matrix) and **U** (upper triangular matrix) as follows:

$$\begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix}\begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

$$\Rightarrow \begin{pmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12}+l_{22} & l_{21}u_{13}+l_{22}u_{23} \\ l_{31} & l_{31}u_{12}+l_{32} & l_{31}u_{13}+l_{32}u_{23}+l_{33} \end{pmatrix}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

This implies

$$
\begin{aligned}
&l_{11} = a_{11}, l_{21} = a_{21}, l_{31} = a_{31};\\
&l_{11}u_{12} = a_{12} \Rightarrow u_{12} = \frac{a_{12}}{l_{11}} = \frac{a_{12}}{a_{11}};\\
&l_{11}u_{13} = a_{13} \Rightarrow u_{13} = \frac{a_{13}}{l_{11}} = \frac{a_{13}}{a_{11}};\\
&l_{21}u_{12} + l_{22} = a_{22} \Rightarrow l_{22} = a_{22} - l_{21}u_{12};\\
&l_{21}u_{13} + l_{22}u_{23} = a_{23} \Rightarrow u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}};\\
&l_{31}u_{12} + l_{32} = a_{32} \Rightarrow l_{32} = a_{32} - l_{31}u_{12};\\
&l_{31}u_{13} + l_{32}u_{23} + l_{33} = a_{33} \Rightarrow l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}.
\end{aligned}
$$

Once all the value of l_{ij}'s and u_{ij}'s are obtained, we can write

$$\mathbf{Ax} = \mathbf{b}\ as\ \mathbf{LUx} = \mathbf{b},$$

Let **Ux=y**, then **Ly=b**:

$$
\begin{aligned}
&\Rightarrow \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}\\
&\Rightarrow \begin{pmatrix} l_{11}y_1 \\ l_{21}y_1 + l_{22}y_2 \\ l_{31}y_1 + l_{32}y_2 + l_{33}y_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}\\
&\Rightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{b_1}{l_{11}} \\ \frac{1}{l_{22}}(b_2 - l_{21}y_1) \\ \frac{1}{l_{33}}(b_3 - l_{31}y_1 - l_{32}y_2) \end{pmatrix}
\end{aligned}
$$

For **Ux=y**, by forward substitution we obtain

$$
\begin{aligned}
&\Rightarrow \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}\\
&\Rightarrow \begin{pmatrix} x_1 + u_{12}x_2 + u_{13}x_3 \\ x_2 + u_{23}x_3 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}
\end{aligned}
$$

By back substitution we get,

$$
\begin{aligned}
&x_3 = y_3\\
&x_2 + u_{23}x_3 = y_2 \Rightarrow x_2 = y_2 - u_{23}x_3\\
&x_1 + u_{12}x_2 + u_{13}x_3 = y_1 \Rightarrow x_1 = y_1 - u_{12}x_2 - u_{13}x_3
\end{aligned}
$$

Example 2.4 *Solve the following system of linear equations, by Crout's method:*

$$10x_1 + 3x_2 + 4x_3 = 15$$
$$2x_1 - 10x_2 + 3x_3 = 37$$
$$3x_1 + 2x_2 - 10x_3 = -10$$

Solution: In matrix form, the given system of equation can be written as

$$\begin{pmatrix} 10 & 3 & 4 \\ 2 & -10 & 3 \\ 3 & 2 & -10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 15 \\ 37 \\ -10 \end{pmatrix}$$

Let **A**=**LU**,which implies

$$\begin{pmatrix} 10 & 3 & 4 \\ 2 & -10 & 3 \\ 3 & 2 & -10 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{pmatrix}$$

$l_{11} = 10, l_{21} = 2, l_{31} = 3;$

$l_{11}u_{12} = 3 \Rightarrow u_{12} = \dfrac{3}{10}$

$l_{11}u_{13} = 4 \Rightarrow u_{13} = \dfrac{4}{10}$

$l_{21}u_{12} + l_{22} = -10 \Rightarrow l_{22} = -10 - 2 \times \dfrac{3}{10} = -\dfrac{106}{10};$

$l_{21}u_{13} + l_{22}u_{23} = 3 \Rightarrow u_{23} = \dfrac{3 - 2 \times \frac{4}{10}}{-\frac{106}{10}} = -\dfrac{11}{53};$

$l_{31}u_{12} + l_{32} = 2 \Rightarrow l_{32} = 2 - 3 \times \dfrac{3}{10} = \dfrac{11}{10};$

$l_{31}u_{13} + l_{32}u_{23} + l_{33} = -10 \Rightarrow l_{33} = -10 - 3 \times \dfrac{4}{10} - \dfrac{11}{10} \times (-\dfrac{11}{53}) = -\dfrac{1163}{106}.$

Therefore, we get

$$\mathbf{L} = \begin{pmatrix} 10 & 0 & 0 \\ 2 & -\frac{106}{10} & 0 \\ 3 & \frac{11}{10} & -\frac{1163}{106} \end{pmatrix}, \ and \ \mathbf{U} = \begin{pmatrix} 1 & \frac{3}{10} & \frac{4}{10} \\ 0 & 1 & -\frac{11}{53} \\ 0 & 0 & 1 \end{pmatrix}$$

Now let **Ux=y**, then **Ly=b**:

$$\begin{pmatrix} 10 & 0 & 0 \\ 2 & -\frac{106}{10} & 0 \\ 3 & \frac{11}{10} & -\frac{1163}{106} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 15 \\ 37 \\ -10 \end{pmatrix}$$

This implies

$$10y_1 = 15 \Rightarrow y_1 = \frac{15}{10} = \frac{3}{2}$$
$$2y_1 - \frac{106}{10}y_2 = 37 \Rightarrow y_2 = (37 - 2 \times \frac{3}{2}) \times (-\frac{10}{106}) = -\frac{170}{53}$$
$$3y_1 + \frac{11}{10}y_2 - \frac{1163}{106}y_3 = -10 \Rightarrow y_3 = (-10 - 2 \times \frac{3}{2} - \frac{11}{10} \times (-\frac{170}{53})) \times (-\frac{106}{1163}) = 1$$

Thus

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ -\frac{170}{53} \\ 1 \end{pmatrix}$$

Then **Ux=y** is given as

$$\begin{pmatrix} 1 & \frac{3}{10} & \frac{4}{10} \\ 0 & 1 & -\frac{11}{53} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ -\frac{170}{53} \\ 1 \end{pmatrix}$$

which implies

$$x_1 + \frac{3}{10}x_2 + \frac{4}{10}x_3 = \frac{3}{2}$$
$$x_2 - \frac{11}{53}x_3 = -\frac{170}{53}$$
$$x_3 = 1$$

By back substitution, we get

$$x_3 = 1$$
$$x_2 = -\frac{170}{53} + \frac{11}{53} \times 1 = -3$$
$$x_1 = \frac{3}{2} - \frac{3}{10} \times (-3) - \frac{4}{10} \times 1 = 2$$

Therefore, the required solution by Crout's method (LU decomposition method) is $x_1 = 2, x_2 = -3, x_3 = 1$.

Example 2.5 *Solve the following system of linear equations by Crout's Method (LU factorization or decomposition method):*

$$9x_1 + 3x_2 + 3x_3 + 3x_4 = 24$$
$$3x_1 + 10x_2 - 2x_3 - 2x_4 = 17$$
$$3x_1 - 2x_2 + 18x_3 + 10x_4 = 45$$
$$3x_1 - 2x_2 + 10x_3 + 10x_4 = 29$$

Solution: The given system of equation can be written in matrix form as

$$\begin{pmatrix} 9 & 3 & 3 & 3 \\ 3 & 10 & -2 & -2 \\ 3 & -2 & 18 & 10 \\ 3 & -2 & 10 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 24 \\ 17 \\ 45 \\ 29 \end{pmatrix}$$

Let

$$\begin{pmatrix} 9 & 3 & 3 & 3 \\ 3 & 10 & -2 & -2 \\ 3 & -2 & 18 & 10 \\ 3 & -2 & 10 & 10 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} & l_{11}u_{14} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} & l_{21}u_{14} + l_{22}u_{24} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} & l_{31}u_{14} + l_{32}u_{24} + l_{33}u_{34} \\ l_{41} & l_{41}u_{12} + l_{42} & l_{41}u_{13} + l_{42}u_{23} + l_{43} & l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + l_{44} \end{pmatrix}$$

Comparing, we get

$l_{11} = 9, l_{21} = 3, l_{31} = 3, l_{41} = 3;$

$l_{11}u_{12} = 3 \Rightarrow u_{12} = \dfrac{1}{3},\ similarly, u_{13} = u_{14} = \dfrac{1}{3}$

$l_{21}u_{12} + l_{22} = 10 \Rightarrow l_{22} = 10 - 3 \times \dfrac{1}{3} = 9;$

$l_{31}u_{12} + l_{32} = -2 \Rightarrow l_{32} = -2 - 3 \times \dfrac{1}{3} = -3;$

$l_{41}u_{12} + l_{42} = -2 \Rightarrow l_{42} = -2 - 3 \times \dfrac{1}{3} = -3;$

$l_{21}u_{13} + l_{22}u_{23} = -2 \Rightarrow u_{23} = \dfrac{-2 - 3 \times \frac{1}{3}}{9} = -\dfrac{1}{3};$

$l_{31}u_{13} + l_{32}u_{23} + l_{33} = 18 \Rightarrow l_{33} = 18 - 3 \times \dfrac{1}{3} - (-3) \times (-\dfrac{1}{3}) = 16;$

$l_{41}u_{13} + l_{42}u_{23} + l_{43} = 10 \Rightarrow l_{43} = 10 - 3 \times \dfrac{1}{3} - (-3) \times (-\dfrac{1}{3}) = 8.$

$l_{21}u_{14} + l_{22}u_{24} = -2 \Rightarrow u_{24} = \dfrac{-2 - 3 \times \frac{1}{3}}{9} = -\dfrac{1}{3};$

$l_{31}u_{14} + l_{32}u_{24} + l_{33}u_{34} = 10 \Rightarrow u_{34} = \dfrac{10 - 3 \times \frac{1}{3} - (-3) \times (-\frac{1}{3})}{16} = \dfrac{1}{2};$

$l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + l_{44} = 10 \Rightarrow l_{44} = 10 - 3 \times \dfrac{1}{3} - (-3) \times (-\dfrac{1}{3}) - 8 \times \dfrac{1}{2} = 4;$

Therefore,we get

$$\mathbf{L} = \begin{pmatrix} 9 & 0 & 0 & 0 \\ 3 & 9 & 0 & 0 \\ 3 & -3 & 16 & 0 \\ 3 & -3 & 8 & 4 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Forward substitution gives (**Ly=b**)

$$\begin{aligned} &9y_1 = 24 \Rightarrow y_1 = \frac{8}{3}, \\ &3y_1 + 9y_2 = 17 \Rightarrow y_2 = \frac{17 - 3 \times y_1}{9} = 1 \\ &3y_1 - 3y_2 + 16y_3 = 45 \Rightarrow y_3 = \frac{5}{2}, \\ &3y_1 - 3y_2 + 8y_3 + 4y_4 = 29 \Rightarrow y_4 = 1 \end{aligned}$$

Thus **Ux=y** gives

$$\begin{pmatrix} 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \frac{8}{3} \\ 1 \\ \frac{5}{2} \\ 1 \end{pmatrix}$$

By back substitution we get,

$$\begin{aligned} &x_4 = 1 \\ &x_3 + \frac{1}{2}x_4 = \frac{5}{2} \Rightarrow x_3 = 2 \\ &x_2 - \frac{1}{3}x_3 - \frac{1}{3}x_4 = 1 \Rightarrow x_2 = 2, \\ &x_1 + \frac{1}{3}x_2 + \frac{1}{3}x_3 + \frac{1}{3}x_4 = \frac{3}{3} \Rightarrow x_1 = 1 \end{aligned}$$

Therefore, the required solution by Crout's method is

$$x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 1.$$

Lecture 4

2.2 Iterative Methods

(a)Jacobi's method
(b)Gauss-Seidel method

2.2.1 Jacobi's Method

Carl Gustav Jacob Jacobi (1804-1851,German) gave an indirect method for finding the solution of a system of linear equations, which is based on the successive better approximations of the values of the unknowns, using an iterative procedure.

The sufficient condition for the convergence of Jacobi method to solve **Ax=b** is that the coefficient matrix **A** is strictly diagonally row dominant, i.e.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix},$$

if $|a_{ii}| > \sum_{j=1(j\neq i)}^{n} |a_{ij}|, \ for\ i = 1, 2, ..., n$ holds, we call the coefficient matrix **A** is strictly diagonally row dominant.

It should be noted that this method makes **two assumptions**:

First, the system of linear equations to be solved, must have a unique solution;

Second, there should not be any zeros on the main diagonal of the coefficient matrix A. In case, there exist zeros on its main diagonal, then rows must be interchanged to obtain a coefficient matrix that does not have zero entries on the main diagonal.

Consider a system of n linear equations in n unknowns, which are strictly

diagonally row dominant, as follows:

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

Since the system is strictly diagonally row dominant,$a_{ii} \neq 0$.

Therefore, the system of equations is rewritten as

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1 - a_{12}x_2 - ... - a_{1n}x_n) \\ x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - 0 \cdot x_2 - ... - a_{2n}x_n) \\ &..... \\ x_n &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1 - a_{n2}x_2 - ... - 0 \cdot x_n) \end{aligned} \tag{2.8}$$

We then consider an arbitrary initial guess of the solution as

$$\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})^T$$

which are row substituted to the right hand side of the rewritten equations to obtain the first approximation as

$$\begin{aligned} x_1^{(1)} &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1^{(0)} - a_{12}x_2^{(0)} - ... - a_{1,n-1}x_{n-1}^{(0)} - a_{1n}x_n^{(0)}) \\ x_2^{(1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(0)} - 0 \cdot x_2^{(0)} - ... - a_{2,n-1}x_{n-1}^{(0)} - a_{2n}x_n^{(0)}) \\ &..... \\ x_n^{(1)} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{(0)} - a_{n2}x_2^{(0)} - ... - a_{n,n-1}x_{n-1}^{(0)} - 0 \cdot x_n^{(0)}) \end{aligned}$$

This process is repeated by substituting the first approximate solution

$$\mathbf{x}^{(1)} = (x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)})^T$$

to the rewritten equations (2.8), we get

$$\begin{aligned} x_1^{(2)} &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1^{(1)} - a_{12}x_2^{(1)} - ... - a_{1,n-1}x_{n-1}^{(1)} - a_{1n}x_n^{(1)}) \\ x_2^{(2)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(1)} - 0 \cdot x_2^{(1)} - ... - a_{2,n-1}x_{n-1}^{(1)} - a_{2n}x_n^{(1)}) \\ &..... \\ x_n^{(2)} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{(1)} - a_{n2}x_2^{(1)} - ... - a_{n,n-1}x_{n-1}^{(1)} - 0 \cdot x_n^{(1)}) \end{aligned}$$

By repeated iteration, we get the required solution up to the desired level of the accuracy ε. Stop the iterative process by $||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| < \varepsilon$, where $||\cdot||$ denotes a norm of a vector, which should be introduced latter.

Example 2.6 *Solve the following system of linear equations by Jacobi's Method:*

$$\begin{aligned} x_1 + x_2 + 4x_3 &= 9 \\ 8x_1 - 3x_2 + 2x_3 &= 20 \\ 4x_1 + 11x_2 - x_3 &= 33 \end{aligned}$$

Solution:

The given system of equations is not diagonally row dominant as $|a_{11}| < |a_{12}| + |a_{13}|$. Therefore, we re-arrange the system as

$$\begin{aligned} 8x_1 - 3x_2 + 2x_3 &= 20 \\ 4x_1 + 11x_2 - x_3 &= 33 \\ x_1 + x_2 + 4x_3 &= 9 \end{aligned}$$

Here,

$$\begin{aligned} &|8| > |-3| + |2|, \\ &|11| > |4| + |-1|, \\ &|4| > |1| + |1|. \end{aligned}$$

Thus, the system is diagonally row dominant. We now re-write the system as

$$\begin{aligned} x_1 &= \frac{1}{8}(20 + 3x_2 - 2x_3) \\ x_2 &= \frac{1}{11}(33 - 4x_1 + x_3) \\ x_3 &= \frac{1}{4}(9 - x_1 - x_2) \end{aligned}$$

Then the iterative formula in Jacobi's sense is

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{8}(20 + 3x_2^{(k)} - 2x_3^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{11}(33 - 4x_1^{(k)} + x_3^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{4}(9 - x_1^{(k)} - x_2^{(k)}) \end{aligned}$$

Let the initial guess be

$$\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^T = (1, 1, 0)^T$$

Then, the first approximation to the solution is given by

$$x_1^{(1)} = \frac{1}{8}(20 + 3 \times 1 - 2 \times 0) = 2.875$$
$$x_2^{(1)} = \frac{1}{11}(33 - 4 \times 1 + 0) = 2.636$$
$$x_3^{(1)} = \frac{1}{4}(9 - 1 - 1) = 1.75$$

Second approximation

$$x_1^{(2)} = \frac{1}{8}(20 + 3 \times 2.636 - 2 \times 1.75) = 3.051$$
$$x_2^{(2)} = \frac{1}{11}(33 - 4 \times 2.875 + 1.75) = 2.114$$
$$x_3^{(2)} = \frac{1}{4}(9 - 2.875 - 2.636) = 0.872$$

Third approximation

$$x_1^{(3)} = \frac{1}{8}(20 + 3 \times 2.114 - 2 \times 0.872) = 3.075$$
$$x_2^{(3)} = \frac{1}{11}(33 - 4 \times 3.051 + 0.872) = 1.969$$
$$x_3^{(3)} = \frac{1}{4}(9 - 3.051 - 2.114) = 0.959$$

Fourth approximation

$$x_1^{(4)} = \frac{1}{8}(20 + 3 \times 1.969 - 2 \times 0.959) = 2.999$$
$$x_2^{(4)} = \frac{1}{11}(33 - 4 \times 3.075 + 0.959) = 1.969$$
$$x_3^{(4)} = \frac{1}{4}(9 - 3.075 - 1.969) = 0.989$$

Fifth approximation

$$x_1^{(5)} = 2.991$$
$$x_2^{(5)} = 1.999$$
$$x_3^{(5)} = 1.008$$

Sixth approximation

$$x_1^{(6)} = 2.997$$
$$x_2^{(6)} = 2.004$$
$$x_3^{(6)} = 1.002$$

Therefore, correct to two significant figures, the solution can be $x_1 = 3.0, x_2 = 2.0, x_3 = 1.0$.

2.2.2 Gauss Seidel Method

Gauss Seidel iteration method for solving a system of n-linear equations in n unknowns is a modified Jacobis method. Therefore, all the conditions that is true for Jacobis method, also holds for Gauss Seidel method. As before, the system of linear equations are rewritten as

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1 - a_{12}x_2 - ... - a_{1n}x_n) \\ x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - 0 \cdot x_2 - ... - a_{2n}x_n) \\ &..... \\ x_n &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1 - a_{n2}x_2 - ... - 0 \cdot x_n) \end{aligned} \tag{2.9}$$

If $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})^T$ be the initial guess of the solution, which is arbitrary, then the first approximation to the solution is obtained as

$$\begin{aligned} x_1^{(1)} &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1^{(0)} - a_{12}x_2^{(0)} - a_{13}x_3^{(0)} - ... - a_{1,n-1}x_{n-1}^{(0)} - a_{1n}x_n^{(0)}) \\ x_2^{(1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(1)} - 0 \cdot x_2^{(0)} - a_{23}x_3^{(0)} - ... - a_{2,n-1}x_{n-1}^{(0)} - a_{2n}x_n^{(0)}) \\ x_3^{(1)} &= \frac{1}{a_{33}}(b_2 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)} - 0 \cdot x_3^{(0)} - ... - a_{3,n-1}x_{n-1}^{(0)} - a_{2n}x_n^{(0)}) \\ &..... \\ x_n^{(1)} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{(1)} - a_{n2}x_2^{(1)} - a_{n3}x_3^{(1)} - ... - a_{n,n-1}x_{n-1}^{(1)} - 0 \cdot x_n^{(0)}) \end{aligned}$$

Please note, while calculating $x_2^{(1)}$, the value of x_1 is replaced by $x_1^{(1)}$, not by $x_1^{(0)}$. This is the basic difference of Gauss Seidel with Jocobi's method.

The successive iterations are generated by the scheme called iteration formulae of Gauss-Seidel method, which is as follows:

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}(b_1 - 0 \cdot x_1^{(k)} - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - ... - a_{1,n-1}x_{n-1}^{(k)} - a_{1n}x_n^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - 0 \cdot x_2^{(k)} - a_{23}x_3^{(k)} - ... - a_{2,n-1}x_{n-1}^{(k)} - a_{2n}x_n^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{a_{33}}(b_2 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)} - 0 \cdot x_3^{(k)} - ... - a_{3,n-1}x_{n-1}^{(k)} - a_{2n}x_n^{(k)}) \\ &..... \\ x_n^{(k+1)} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - a_{n3}x_3^{(k+1)} - ... - a_{n,n-1}x_{n-1}^{(k+1)} - 0 \cdot x_n^{(k)}) \end{aligned}$$

The number of iterations (k) required depends upon the desired degree of accuracy.

Example 2.7 *Solve the system of linear equations by Gauss Seidel method:*

$$\begin{aligned} x_1 + x_2 + 4x_3 &= 9 \\ 8x_1 - 3x_2 + 2x_3 &= 20 \\ 4x_1 + 11x_2 - x_3 &= 33 \end{aligned}$$

Solution:

The given system of equations is not diagonally row dominant as $|a_{11}| < |a_{12}| + |a_{13}|$. Therefore, we re-arrange the system as

$$\begin{aligned} 8x_1 - 3x_2 + 2x_3 &= 20 \\ 4x_1 + 11x_2 - x_3 &= 33 \\ x_1 + x_2 + 4x_3 &= 9 \end{aligned}$$

Here,

$$\begin{aligned} &|8| > |-3| + |2|, \\ &|11| > |4| + |-1|, \\ &|4| > |1| + |1|. \end{aligned}$$

Thus, the system is diagonally row dominant. We now re-write the system as

$$\begin{aligned} x_1 &= \frac{1}{8}(20 + 3x_2 - 2x_3) \\ x_2 &= \frac{1}{11}(33 - 4x_1 + x_3) \\ x_3 &= \frac{1}{4}(9 - x_1 - x_2) \end{aligned}$$

Thus, the Gauss-Seidel iterative formula can be written as

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{8}(20 + 3x_2^{(k)} - 2x_3^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{11}(33 - 4x_1^{(k+1)} + x_3^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{4}(9 - x_1^{(k+1)} - x_2^{(k+1)}) \end{aligned}$$

Let the initial guess be

$$\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^T = (1, 1, 0)^T$$

Then, the first approximation to the solution is given by

$$x_1^{(1)} = \frac{1}{8}(20 + 3 \times 1 - 2 \times 0) = 2.875$$
$$x_2^{(1)} = \frac{1}{11}(33 - 4 \times 2.875 + 0) = 1.955$$
$$x_3^{(1)} = \frac{1}{4}(9 - 2.875 - 1.955) = 1.043$$

2nd approximation

$$x_1^{(2)} = \frac{1}{8}(20 + 3 \times 1.955 - 2 \times 1.043) = 2.972$$
$$x_2^{(2)} = \frac{1}{11}(33 - 4 \times 2.972 + 1.043) = 2.014$$
$$x_3^{(2)} = \frac{1}{4}(9 - 2.972 - 2.014) = 1.004$$

3rd approximation

$$x_1^{(3)} = \frac{1}{8}(20 + 3 \times 2.014 - 2 \times 1.004) = 3.004$$
$$x_2^{(3)} = \frac{1}{11}(33 - 4 \times 3.004 + 1.043) = 1.999$$
$$x_3^{(3)} = \frac{1}{4}(9 - 3.004 - 1.999) = 0.999$$

4th approximation

$$x_1^{(4)} = 3.00$$
$$x_2^{(4)} = 2.00$$
$$x_3^{(4)} = 1.00$$

Therefore, correct to two significant figures, the solution can be $x_1 = 3.0, x_2 = 2.0, x_3 = 1.0$.

2.2.3 Exercises

1. Use Jacobis method to solve the following system of equations, with $x^{(0)} = (1, 1, 1)^T$ as initial approximation, correct to 2 significant figures.

$$x - 10y + 3z = 39$$
$$10x - 2y - 5z = 26$$
$$4x - 5y + 10z = 47$$

   What is the minimum number of iterations required to get 5 significant digit accuracy, if 5 arithmetic digits is used.

   (Ans: True solution $(3, -3, 2)^T$; number of iteration required=36)

2. Do three iterations of Jacobis method to solve

$$-2x + 3y + 10z = 22$$
$$10x + 2y + z = 29$$
$$x + 10y - z = -22$$

   with $x^{(0)} = (1, -1, 1)^T$ T as starting vector. What is the minimum number of iterations required, so that the solution is correct to 4 decimal places.

   (Ans: True solution $(1, -2, 3)^T$; number of iteration required =17)

3. Solve, by Gauss-Seidal iteration method, the system of linear equations

$$3x + 9y - 2z = 11$$
$$4x + 2y + 13z = 24$$
$$4x - 2y + z = -8$$

   correct up to four significant figures.

   (Ans: x=1.423,y=2.131, z=1.956)

4. Compute the solution of the system of linear equations by Gauss-Seidal iteration method

$$6.7x + 1.1y + 2.2z = 20.5$$
$$3.1x + 9.4y - 1.5z = 22.9$$
$$2.1x - 1.5y + 8.4z = 28.8$$

   correct up to 3-significant figures.

(Ans: x=1.50, y=2.50,z= 3.50)

5. Do five iterations of each Jacobi's and Gauss Seidel method to solve

$$2x + 3y + 7z = 16$$
$$3x + y + z = 6$$
$$x + 5y + 3z = 10$$

with starting initial guess as $(x, y, z) = (1, 1, 1)$. What is the minimum number of iterations required, so that the solutions correct to 8 significant figures?

(Ans: True solution: x =1.2, y=0.8, z=1.6)

Lecture 5

2.3 What is a norm?

Mathematically a norm is a total size or length of all vectors in a vector space or matrices. For simplicity, we can say that the higher the norm is, the bigger the (value in) matrix or vector is. Norm may come in many forms and many names, including these popular name: Euclidean distance, Mean-squared Error,etc.

Most of the time you will see the norm appears in a equation like this: $||\mathbf{x}||$, where $\mathbf{x}$ can be a vector or a matrix.

For example, a Euclidean norm of a vector $\mathbf{x} = \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$ is

$$||\mathbf{x}||_2 = \sqrt{1^2 + 3^2 + (-2)^2} = \sqrt{14}.$$

which is the size of vector $\mathbf{x}$.

The above example shows how to compute a Euclidean norm, or formally called an l_2- norm. There are many other types of norm that beyond our explanation here, actually for every single *real number*, there is a norm correspond to it (Notice the emphasised word *real number*, that means it not limited to only integer.)

Formally the l_p- norm of $\mathbf{x}$ is defined as:

$$||\mathbf{x}||_p = \sqrt[p]{\sum_i |x_i|^p}, where\ p \in \Re$$

That's it! A p-th-root of a summation of all elements to the p-th power is what we call a norm.

The interesting point is even though every l_p-norm is all look very similar to each other, their mathematical properties are very different and thus their application are dramatically different too. Hereby we are going to look into some of these norms in details.

Norm of a vector

l_1-norm

Following the definition of norm, l_1-norm of $\mathbf{x}$ is defined as

$$||\mathbf{x}||_1 = \sum_i |x_i|$$

This norm is quite common among the norm family. It has many names and many forms among various fields, namely **Manhattan norm** is its nickname. If the l_1-norm is computed for a difference between two vectors or matrices, that is

$$||\mathbf{x}_1 - \mathbf{x}_2||_1 = \sum_i |x_{1_i} - x_{2_i}|$$

it is called Sum of Absolute Difference (SAD) among computer vision scientists.

l_2-norm

The most popular of all norm is the l_2-norm. It is used in almost every field of engineering and science as a whole. Following the basic definition, l_2-norm is defined as

$$||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$$

l_2-norm is well known as a Euclidean norm, which is used as a standard quantity for measuring a vector difference. As in l_2-norm, if the Euclidean norm is computed for a vector difference, it is known as a Euclidean distance:

$$||\mathbf{x}_1 - \mathbf{x}_2||_2 == \sqrt{\sum_i (x_{1_i} - x_{2_i})^2}$$

Its most well known application in the signal processing field is the Mean-Squared Error (MSE) measurement, which is used to compute a similarity, a quality, or a correlation between two signals.

l-infinity norm

As always, the definition for l_∞-norm is

$$||\mathbf{x}||_\infty = \max_i |x_i|$$

Example: Let $\mathbf{x} = \begin{pmatrix} -1 \\ -6 \\ 0 \\ 2 \end{pmatrix}$, find $||\mathbf{x}||_1, ||\mathbf{x}||_2, ||\mathbf{x}||_\infty$.

Solution:

$$||\mathbf{x}||_1 = \sum_{i=1}^{4} |x_i| = |-1| + |-6| + |0| + |2| = 9$$

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{4} x_i^2} = \sqrt{(-1)^2 + (-6)^2 + 0^2 + 2^2} = \sqrt{41}$$

$$||\mathbf{x}||_\infty = \max_{1 \le i \le 4} |x_i| = \max\{|-1|, |-6|, 0, 2\} = 6$$

Infinity Norm of Matrix

The infinity norm of $(n \times n)$ matrix is given by

$$||\mathbf{A}||_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$$

This norm is also called maximum row sum norm, since, by definition, it is the maximum the sum of the absolute values of elements in each row. For example, consider the matrix

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 5 \\ 3 & 0 & 9 \\ -8 & 4 & 1 \end{pmatrix}$$

then,

$$\sum_{j=1}^{n} |a_{1j}| = |2| + |-1| + |5| = 8,$$

$$\sum_{j=1}^{n} |a_{2j}| = |3| + |0| + |9| = 12,$$

$$\sum_{j=1}^{n} |a_{3j}| = |-8| + |4| + |1| = 13.$$

Thus

$$||\mathbf{A}||_\infty = \max\{8, 12, 13\} = 13.$$

Therefore, the infinity norm of the given matrix $\mathbf{A}$ is 13.

Conditional Number

The conditional number of an invertible square matrix $\mathbf{A}$ is given by

$$cond(\mathbf{A}) = ||\mathbf{A}||_\infty ||\mathbf{A}^{-1}||_\infty$$

However, if $\mathbf{A}$ is singular, then $cond(\mathbf{A}) = +\infty$. So, what this conditional number do? It actually gives a measure of stability or sensitivity of a matrix to

numerical operations. It the conditional number of a matrix is near 1, it is said to be well conditioned matrix and it is called ill-conditioned if its conditional number is much higher than 1.

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 3 & 2.01 \end{pmatrix}$$

then

$$\mathbf{A}^{-1} = \begin{pmatrix} 67 & -66.67 \\ -100 & 100 \end{pmatrix}$$

Therefore, $cond(\mathbf{A}) = ||\mathbf{A}||_\infty ||\mathbf{A}^{-1}||_\infty = 5.01 \times 200 = 1002$, which is much greater than 1. Thus, the matrix A is called ill-conditioned.

ILL Conditioned Equations

A system of equation for which small changes in the coefficients and/or constants produce substantial changes in the solution is called ill-conditioned. Consider a system of two linear equations in two unknowns:

$$\begin{pmatrix} 500 & -201 \\ -1000 & 401 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 300 \\ -300 \end{pmatrix}$$

The solution of this system is $x = 120, y = 300$. We now make a small change in the coefficient of a_{12} from 201 to 202. The solution of the system changes to $x = 39.8, y = 100$, which shows that the solution is very sensitive to the value of the coefficient matrix $\mathbf{A}$. Thus, a small change in one of the element of the coefficient $\mathbf{A}$ produce large change in the solution of the system of linear equations, which are called ill-conditioned.

For this system, the coefficient matrix $\mathbf{A} = \begin{pmatrix} 500 & -201 \\ -1000 & 401 \end{pmatrix}$, and the inversion matrix is $\mathbf{A}^{-1} = -\frac{1}{500} \begin{pmatrix} 401 & 201 \\ 1000 & 500 \end{pmatrix}$,thus the conditional number of $\mathbf{A}$ is

$$cond(\mathbf{A}) = ||\mathbf{A}||_\infty ||\mathbf{A}^{-1}||_\infty = 1401 \times 3 = 4203,$$

which is much greater than 1. Therefore, we can call this linear system as an ill-conditioned system, which is very sensitive to the small changing value of the coefficient matrix $\mathbf{A}$.

Reference and further reading:

Mathematical Norm - wikipedia
Mathematical Norm -MathWorld
Michael Elad-Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.
Linear Programming - MathWorld
Compressive Sensing - Rice University

Chapter 3

Eigenvalues and Eigenvectors

Lecture 1

1. Introduction

Let $\mathbf{A}$ be a $n \times n$ matrix. A non-zero vector $\mathbf{x}$ is an eigenvector of A if there exists a scalar λ such that

$$\mathbf{Ax} = \lambda\mathbf{x}$$

The scalar λ is called the eigenvalue of the matrix $\mathbf{A}$, corresponding to the eigenvector $\mathbf{x}$.

Let $\lambda_1, \lambda_2, ..., \lambda_n$ be the eigenvalues of a matrix $\mathbf{A}$. If $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$, the λ_1 is called the dominant eigenvalue of $\mathbf{A}$ and the eigenvectors corresponding to λ_1, are called dominant eigenvectors.

Let us consider a system

$$\mathbf{Ax} = \lambda\mathbf{x},$$

for which we want to find the eigenvalues and eigenvectors. The standard method for that is to solve for the roots of λ of the characteristic equation

$$|\mathbf{A} - \lambda I| = 0.$$

When $\mathbf{A}$ is large, this method is totally impractical. Evaluating the determinant of a $n \times n$ matrix is a huge task, when n is large and solving the resulting n-th degree polynomial equation for λ is another additional task on top of that.

The power method is a simple iteration method that can be used to find the dominant eigenvalue λ_1 and dominant eigenvector v_1 for a given matrix $\mathbf{A}$, where λ_1 is the largest eigenvalue and v_1 is the corresponding eigenvector.

Similarly, the inverse power method is used to find the smallest eigenvalue and its corresponding vector, which is very similar to power method.

3.1 Power method

We first assume that the matrix $\mathbf{A}$ has a dominant eigenvalue with the corresponding dominant eigenvectors. As stated before, the power method for approximating eigenvalues is iterative. Hence, we start with an initial approximation x_0 of the dominant eigenvector of $\mathbf{A}$, which must be non-zero. Thus, we obtain a sequence of eigenvectors given by

$$\begin{aligned}
&\mathbf{x}_1 = \mathbf{A}\tilde{\mathbf{x}}_0 \\
&\mathbf{x}_2 = \mathbf{A}\tilde{\mathbf{x}}_1 = \mathbf{A}(\mathbf{A}\tilde{\mathbf{x}}_0) = \mathbf{A}^2\tilde{\mathbf{x}}_0 \\
&\mathbf{x}_3 = \mathbf{A}\tilde{\mathbf{x}}_2 = \mathbf{A}(\mathbf{A}^2\tilde{\mathbf{x}}_0) = \mathbf{A}^3\tilde{\mathbf{x}}_0 \\
&\ldots\ldots \\
&\mathbf{x}_k = \mathbf{A}\tilde{\mathbf{x}}_{k-1} = \mathbf{A}(\mathbf{A}^{k-1}\tilde{\mathbf{x}}_0) = \mathbf{A}^k\tilde{\mathbf{x}}_0
\end{aligned}$$

Where $\tilde{\mathbf{x}}_k = \frac{1}{\mu}\mathbf{x}_k$, $\mu = sgn(x_{k_i})(\max\limits_{1\le i\le n}\{|x_{k_i}|\}$, x_{k_i} is the i-th component of the n-dimensional vector $\mathbf{x}_k$.

When k is large, we can obtain a good approximation of the dominant eigenvector of $\mathbf{A}$ by properly scaling the sequence.

Example 3.1 *Use power method to approximate a dominant eigenvalue and the corresponding eigenvector of* $\boldsymbol{A} = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}$ *correct to 3-significant figures, after 10 iterations.*

Solution:

We begin with an initial non-zero approximation of dominant eigenvector $\mathbf{x}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and obtain the following approximations as

$$\mathbf{x}_1 = \mathbf{A}\tilde{\mathbf{x}}_0 = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 4.00\begin{pmatrix} 1.00 \\ 0.500 \end{pmatrix}$$

From the above step, it is clear that after performing $\mathbf{A}$, the dominant element of the matrix $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ is taken out (which is 4.00 in this case, correct to 3- significant figures). The corresponding vector $\begin{pmatrix} 1.00 \\ 0.500 \end{pmatrix}$ will be the new initial vector for the next approximation.

We now obtain the series of approximations as follows:

$$\mathbf{x}_2 = \mathbf{A}\tilde{\mathbf{x}}_1 = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} 1.00 \\ 0.500 \end{pmatrix} = \begin{pmatrix} 1.50 \\ 0.500 \end{pmatrix} = 1.50 \begin{pmatrix} 1.00 \\ 0.333 \end{pmatrix}$$

$$\mathbf{x}_3 = \mathbf{A}\tilde{\mathbf{x}}_2 = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} 1.00 \\ 0.333 \end{pmatrix} = \begin{pmatrix} 2.335 \\ 1.001 \end{pmatrix} = 2.335 \begin{pmatrix} 1.00 \\ 0.4286 \end{pmatrix}$$

$$\mathbf{x}_4 = \mathbf{A}\tilde{\mathbf{x}}_3 = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} 1.00 \\ 0.4286 \end{pmatrix} = \begin{pmatrix} 1.8757 \\ 0.7142 \end{pmatrix} = 1.857 \begin{pmatrix} 1.00 \\ 0.3846 \end{pmatrix}$$

$$\mathbf{x}_5 = \mathbf{A}\tilde{\mathbf{x}}_4 = \begin{pmatrix} 2.077 \\ 0.8462 \end{pmatrix} = 2.077 \begin{pmatrix} 1.0000 \\ 0.4074 \end{pmatrix}$$

$$\mathbf{x}_6 = 1.963 \begin{pmatrix} 1.0000 \\ 0.3962 \end{pmatrix}, \mathbf{x}_7 = 2.019 \begin{pmatrix} 1.0000 \\ 0.4019 \end{pmatrix}, \mathbf{x}_8 = 1.991 \begin{pmatrix} 1.0000 \\ 0.3991 \end{pmatrix}$$

$$\mathbf{x}_9 = 2.005 \begin{pmatrix} 1.0000 \\ 0.4005 \end{pmatrix}, \mathbf{x}_{10} = 1.998 \begin{pmatrix} 1.0000 \\ 0.3998 \end{pmatrix}$$

Therefore, the dominant eigenvalue, correct to 3-significant figure is 2.00 and the corresponding eigenvector is $\begin{pmatrix} 1.00 \\ 0.400 \end{pmatrix}$.

Example 3.2 *Use power method to approximate a dominant eigenvalue and the corresponding eigenvector of* $\boldsymbol{A} = \begin{pmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{pmatrix}$ *correct to 3-significant figures, after 9 iterations.*

Solution:

We begin with an initial non-zero approximation of dominant eigenvector $\mathbf{x}_0 =$

$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and obtain the following approximations as

$$\mathbf{x}_1 = \mathbf{A}\tilde{\mathbf{x}}_0 = \begin{pmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \\ 12 \end{pmatrix} = 12.00 \begin{pmatrix} 0.5000 \\ 0.6667 \\ 1.000 \end{pmatrix}$$

As explained in **Example 3.2**, we take out the largest element of the resultant matrix(which is 12 in this case) and $\begin{pmatrix} 0.5000 \\ 0.6667 \\ 1.000 \end{pmatrix}$ will be our new initial vector.
Proceeding in this manner we obtain a series of approximations as follows:

$$\mathbf{x}_2 = \mathbf{A}\tilde{\mathbf{x}}_1 = \begin{pmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{pmatrix} \begin{pmatrix} 0.5000 \\ 0.6667 \\ 1.000 \end{pmatrix} = \begin{pmatrix} 2.3333 \\ 3.3333 \\ 5.3333 \end{pmatrix} = 5.3333 \begin{pmatrix} 0.4375 \\ 0.6250 \\ 1.000 \end{pmatrix}$$

$$\mathbf{x}_3 = \begin{pmatrix} 1.875 \\ 2.750 \\ 4.500 \end{pmatrix} = 4.500 \begin{pmatrix} 0.4167 \\ 0.6111 \\ 1.000 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 1.722 \\ 2.555 \\ 4.222 \end{pmatrix} = 4.222 \begin{pmatrix} 0.4079 \\ 0.6053 \\ 1.000 \end{pmatrix}$$

$$\mathbf{x}_4 = 4.105 \begin{pmatrix} 0.4038 \\ 0.6026 \\ 1.000 \end{pmatrix}, \mathbf{x}_5 = 4.051 \begin{pmatrix} 0.4019 \\ 0.6013 \\ 1.000 \end{pmatrix}, \mathbf{x}_6 = 4.025 \begin{pmatrix} 0.4009 \\ 0.6006 \\ 1.000 \end{pmatrix}$$

$$\mathbf{x}_7 = 4.013 \begin{pmatrix} 0.4005 \\ 0.6003 \\ 1.000 \end{pmatrix}, \mathbf{x}_8 = 4.006 \begin{pmatrix} 0.4002 \\ 0.6001 \\ 1.000 \end{pmatrix}, \mathbf{x}_9 = 4.003 \begin{pmatrix} 0.4001 \\ 0.6001 \\ 1.000 \end{pmatrix}$$

Therefore, the dominant eigenvalue is 4.00 and the corresponding eigenvector is $\begin{pmatrix} 0.400 \\ 0.600 \\ 1.00 \end{pmatrix}$, correct to 3-significant figure.

3.2 Exercises

Using Power Method, find the dominant eigenvalue and the corresponding eigenvector of the following matrices:

1. $\begin{pmatrix} 1 & -5 \\ -3 & -1 \end{pmatrix}$
2. $\begin{pmatrix} -4 & 10 \\ 7 & 5 \end{pmatrix}$
3. $\begin{pmatrix} 1 & 2 & -2 \\ -2 & 5 & -2 \\ -6 & 6 & -3 \end{pmatrix}$
4. $\begin{pmatrix} 3 & 2 & -3 \\ -3 & -4 & 9 \\ -1 & -2 & 5 \end{pmatrix}$

Lecture 2

3.3 Inverse power method

Inverse power method, which is similar to power method, is used to calculate the smallest eigenvalue and its corresponding eigenvector. Here, we simply use the property that if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \textit{then } \mathbf{A}^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}$$

Thus, in inverse power method, we first take the inverse of the given matrix $\mathbf{A}$ (provided inverse exists) and then apply the standard power method to $\mathbf{A}^{-1}$, which will return the largest eigenvalue of $\mathbf{A}^{-1}$(say, $\hat{\lambda}$) and the corresponding eigenvector (say, $\hat{\mathbf{v}}$).

The inverse of this eigenvalue, that is, $1/\hat{\lambda}$ will give the smallest eigenvalue of the matrix $\mathbf{A}$. Please note that the eigenvector corresponding to $\hat{\lambda}$ in $\mathbf{A}^{-1}$, that is, $\hat{\mathbf{v}}$ remains same as the eigenvector for $1/\hat{\lambda}$ in $\mathbf{A}$.

Example 3.3 *Use inverse power method to approximate the least eigenvalue and the corresponding eigenvector of* $\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 5 & 6 & 0 \end{pmatrix}$*, correct to 3-significant figures, after 6 iterations.*

Solution: We first obtain the inverse of the given matrix A, which can be easily calculated as

$$\mathbf{A}^{-1} = \begin{pmatrix} -24 & 18 & 5 \\ 20 & -15 & -4 \\ -5 & 4 & 1 \end{pmatrix} = \mathbf{B}$$

We now calculate the dominant eigenvalue and its corresponding eigenvector of matrix $\mathbf{B}$ by power method (as described in earlier examples) by taking an initial non-zero approximation $\mathbf{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$. The series of approximations are

obtained as follows:

$$\mathbf{x}_1 = \mathbf{B}\tilde{\mathbf{x}}_0 = \begin{pmatrix} -24 & 18 & 5 \\ 20 & -15 & -4 \\ -5 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1.000 \\ 1.000 \\ -0.000 \end{pmatrix} = 1.000 \begin{pmatrix} -1.000 \\ 1.000 \\ -0.000 \end{pmatrix}$$

$$\mathbf{x}_2 = \mathbf{B}\tilde{\mathbf{x}}_1 = 42.00 \begin{pmatrix} 1.000 \\ -0.8333 \\ 0.2143 \end{pmatrix}, \mathbf{x}_3 = \mathbf{B}\tilde{\mathbf{x}}_2 = -37.93 \begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2141 \end{pmatrix},$$

$$\mathbf{x}_4 = -37.95 \begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2143 \end{pmatrix}, \mathbf{x}_5 = -37.95 \begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2141 \end{pmatrix}, \mathbf{x}_6 = -37.95 \begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2141 \end{pmatrix}$$

Thus, the dominant eigenvalue of $\mathbf{B} = \mathbf{A}^{-1}$ is -37.95 and the corresponding eigenvector is $\begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2141 \end{pmatrix}$. Therefore, by inverse power method, the least eigenvalue of $\mathbf{A}$ is $\frac{1}{-37.95} = -0.02635$ and its corresponding eigenvector is $\begin{pmatrix} 1.000 \\ -0.8343 \\ 0.2141 \end{pmatrix}$.

3.4 Exercises

Using Inverse Power Method, find the least eigenvalue and the corresponding eigenvector of the following matrices:

1. $\begin{pmatrix} 4 & 5 \\ 6 & 5 \end{pmatrix}$
2. $\begin{pmatrix} 2 & -12 \\ 1 & -5 \end{pmatrix}$
3. $\begin{pmatrix} -1 & -6 & 0 \\ 2 & 7 & 0 \\ 1 & 2 & -1 \end{pmatrix}$
4. $\begin{pmatrix} 1 & 2 & -2 \\ -2 & 5 & -2 \\ -6 & 6 & -3 \end{pmatrix}$

Chapter 4

Roots of Non-linear Equations

Lecture 1

4.1 Introduction

In this chapter, we solve the equation of the form $f(x) = 0$. The equation $f(x) = 0$, will be called algebraic or transcendental according as $f(x)$ is purely a polynomial in x or contains some other functions, such as trigonometric, logarithmic or exponential functions etc. For example,

$$x^3 - x + 2 = 0, \ (\textit{purely a polynomial in } x)$$

is an algebraic equation. And a transcendental equation, for example,

$$\sin x + x^2 - \log x = 0,$$

contains trigonometric, logarithmic and exponential function. Normally, it is assumed that $f(x)$ is a continuous function in our discussion.

By finding the solution of the equation $f(x) = 0$, we mean to find those values of x for which $f(x) = 0$. Such values of x are called the roots or zeros of the equation $f(x) = 0$. In this part, we will only concentrate on finding the real roots of the equation $f(x) = 0$. Geometrically, by solving $f(x) = 0$, we wish to find those points where the graph of $f(x)$ crosses the x-axis. In other words, we

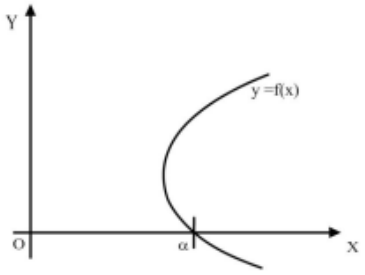

Figure 4.1: the point of intersection of the graph of $f(x)$ with the x-axis

are finding the points of intersection of the graph of with the x-axis (see Figure 4.1 below).

Location of Roots
There are various methods of solving the equation of the form $f(x) = 0$, whether algebraic or transcendental. With the help of these methods, we first find an approximate value of the root of the equation and then successively improve it. Here, we discuss only one method, namely the method of tabulation, to obtain the location of a root.

Method of Tabulation: The method of tabulation is the application of the **Bolzano's theorem** on continuity which states that:

if the function $f(x)$ is continuous in the closed interval $[a, b]$ and if $f(a)$ and $f(b)$ are of opposite signs ($f(a)f(b) < 0$), then there exists at least one real root of $f(x) = 0$ between a and b, i.e. $\exists \xi \in (a, b), s.t\ f(\xi) = 0$.

Geometrical Interpretation of the Method of Tabulation:

If the graph of a continuous function lies above the x-axis at one end of an interval $[a, b]$ and below the x-axis at another end, then it must cross the x-axis somewhere in between at least once.

4.2 Bisection method

This method is also based on Bolzano's theorem on continuity. Assume $f(x) = 0$ hasa root in $[a, b]$, and the function $f(x)$ is continuous in $[a, b]$. And also, $f(a)$

and $f(b)$ are of opposite signs, i.e., $f(a)f(b) < 0$.

Let $x_1 = \frac{a+b}{2}$ be the middle point of the interval $[a, b]$. If $f(x_1) = 0$, then x_1 is the root of $f(x) = 0$. Otherwise,
either $f(a)f(x_1) < 0$, implying that the root lies in the interval $[a, x_1]$;
or $f(x_1)f(b) < 0$, implying that the root lies in the interval $[x_1, b]$.
Thus, Thus, the interval is reduced from $[a, b]$ to either $[a, x_1]$ or$[x_1, b]$. We rename it $[a_1, b_1]$.

Let $x_2 = \frac{a_1+b_1}{2}$ be the middle point of the interval $[a_1, b_1]$. If $f(x_2) = 0$, then x_2 is the root of $f(x) = 0$. Otherwise,
either $f(a)f(x_2) < 0$, implying that the root lies in the interval $[a, x_2]$;
or $f(x_2)f(b) < 0$, implying that the root lies in the interval $[x_2, b]$.
Thus, Thus, the interval is reduced from $[a, b]$ to either $[a_1, x_2]$ or$[x_2, b_1]$. We rename it $[a_2, b_2]$.

We continue in this manner and the process is repeated until the root is obtained to the desired accuracy.

Example 4.1 *Find a real positive root of* $x^3 - 2x + 1 = 0$ *by bisection method.*

Solution: We first get an approximate location by the method of tabulation. Here

$$f(x) = x^3 - 5x + 1, f(0) = 2, f(1) = -3, \Rightarrow f(0)f(1) < 0,$$

so that there exists a root α lying in $(0, 1)$.

Computation of $\alpha(0 < \alpha < 1)$

n	$a_n(+)$	$b_n(-)$	x_{n+1}	$f(x_{n+1)}$
0	0	1	0.5	-1.3750
1	0	0.5	0.25	-0.2344
2	0	0.25	0.125	0.3770
3	0.125	0.25	0.1875	0.0691
4	0.1875	0.25	0.2188	-0.0835
5	0.1875	0.2188	0.2032	-0.0076
6	0.1875	0.2032	0.1954	0.0305
7	0.1954	0.2032	0.1993	0.0114
8	0.1993	0.2032	0.2012	0.0021
9	0.2012	0.2032	0.2022	-0.0027
10	0.2012	0.2022	0.2017	2.9426e-04

In the 10th step, f(0.2017)=2.9426e-04 is very near zero. Thus, the approximate value of the root α is 0.2017, namely $\alpha \approx 0.202$, correct to 3-significant figures.

Example 4.2 *Find a real positive root of* $\sin x + x - 1 = 0$ *by bisection method.*

Solution: We first get an approximate location by the method of tabulation. Here

$$f(x) = \sin x + x - 1, f(0) = -1, f(1) = 0.8415, \Rightarrow f(0)f(1) < 0,$$

so that there exists a root α lying in $(0, 1)$.

Computation of $\alpha(0 < \alpha < 1)$

n	$a_n(-)$	$b_n(+)$	x_{n+1}	$f(x_{n+1})$
0	0	1	0.5	-0.0206
1	0.5	1	0.75	0.4316
2	0.5	0.75	0.6250	0.2101
3	0.5	0.6250	0.5625	0.0958
4	0.5	0.5625	0.5313	0.0379
5	0.5	0.5313	0.5156	0.0088
6	0.5	0.5156	0.5078	-0.0059
7	0.5078	0.5156	0.5117	0.0014
8	0.5078	0.5117	0.5098	-0.0023
9	0.5098	0.5117	0.5108	-4.1833e-04

In the 9th step, f(0.5108)=-4.1833e-04 is very near zero. Thus, the approximate value of the root α is 0.5108, namely $\alpha \approx 0.511$, correct to 3-significant figures.

Exercises

1. Find a real root of $x^3 + x^2 + x + 7 = 0$ by the method of bisection, correct to three significant figures. (Ans:-2.11)
2. Find a real root of $\sin x = 10(x - 1)$ by the method of bisection, correct to three significant figures.(Ans: 1.09)

Lecture 2

4.3 Fixed Point Iteration

Fixed point iteration method is one of iteration methods which is based on the principle of finding a sequence $\{x_n\}$, each element of which successively approximates a real root α of equation $f(x) = 0$, for $x \in [a, b]$. We re-write $f(x) = 0$ as $x = \varphi(x)$. For example,

$$f(x) = x^3 - x + 2 = 0, \Rightarrow x = x^3 + 2, \ i.e. \ \varphi(x) = x^3 + 2.$$

Thus, a root α of the given equation $f(x) = 0$ satisfies $\alpha = \varphi(\alpha)$. In another word, the point α remains fixed under the mapping φ and so a root of the equation $f(x) = 0$ is a fixed point of the function $\varphi(x)$.

Here, $\varphi(x)$ is called the **iteration function**, assume it is continuously differentiable in $[a, b]$.

Using graphical or tabulation method, we first find a location or crude approximation of a real root α of $f(x) = 0$ in an interval saying $[a_0, b_0]$. Let $x_0 \in [a_0, b_0]$ be the initial approximation of the real root α. Whereas, α satisfies the equation

$$x = \varphi(x)$$

Putting the initial approximation x_0 in $x = \varphi(x)$, we get first approximation x_1 of α as

$$x_1 = \varphi(x_0)$$

And then the successive approximations are calculated as

$$\begin{aligned} x_2 &= \varphi(x_1) \\ x_3 &= \varphi(x_2) \\ x_4 &= \varphi(x_3) \\ &..... \\ x_n &= \varphi(x_{n-1}) \\ x_{n+1} &= \varphi(x_n) \end{aligned}$$

The above iteration is generated by the formula $x_{n+1} = \varphi(x_n)$, which is so-called the **iteration formula**, where x_n means the n^{th} approximation of the root α of the equation $f(x) = 0$. The sequence $\{x_n\}$ of iterations or the successive better approximations may or may not converge to a limit. If $\{x_n\}$ converges, then

it converges to the root α and also the number of iterations required depends upon the desired degree of accuracy of the root α.

Convergence of Method of Iteration

The presentation of $f(x) = 0$ as is not unique, for example,

$$f(x) = x^3 - x + 2 = 0, \Rightarrow x = x^3 + 2, \ i.e. \ \varphi(x) = x^3 + 2,$$
$$\Rightarrow x = \sqrt[3]{x-2}, \ i.e. \ \varphi(x) = \sqrt[3]{x-2}.$$

Therefore, the convergence of $\{x_n\}$ depends upon the nature of $\varphi(x)$. Now, we investigate about the nature of $\varphi(x)$ which yields a convergent sequence $\{x_n\}$.

Berfore discussing the convergence of the fixed point iteration method, we firstly review the **Mean Value Theorem** introduced by Lagrange.

Theorem 4.1 *Assume $f(x)$ is a continuous function in $[a, b]$, then a number ξ exists between a and b, such that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi)$$

holds, namely,

$$f(b) - f(a) = f'(\xi)(b - a).$$

Now by Lagrange's mean value theorem, we have

$$|\alpha - x_1| = |\varphi(\alpha) - \varphi(x_0)| = |\alpha - x_0||\varphi'(\varepsilon_0)|, \varepsilon_0 \in (x_0, \alpha) \ or \ (\alpha, x_0)$$
$$|\alpha - x_2| = |\varphi(\alpha) - \varphi(x_1)| = |\alpha - x_1||\varphi'(\varepsilon_1)|, \varepsilon_1 \in (x_1, \alpha) \ or \ (\alpha, x_1)$$
$$|\alpha - x_3| = |\varphi(\alpha) - \varphi(x_2)| = |\alpha - x_2||\varphi'(\varepsilon_2)|, \varepsilon_2 \in (x_2, \alpha) \ or \ (\alpha, x_2)$$
$$.....$$
$$|\alpha - x_{n+1}| = |\varphi(\alpha) - \varphi(x_n)| = |\alpha - x_n||\varphi'(\varepsilon_n)|, \varepsilon_n \in (x_n, \alpha) \ or \ (\alpha, x_n)$$

Hence,

$$\begin{aligned} |\alpha - x_{n+1}| &= |\alpha - x_n||\varphi'(\varepsilon_n)| \\ &= |\alpha - x_{n-1}||\varphi'(\varepsilon_{n-1})||\varphi'(\varepsilon_n)| \\ &= |\alpha - x_{n-2}||\varphi'(\varepsilon_{n-2})||\varphi'(\varepsilon_{n-1})||\varphi'(\varepsilon_n)| \\ &= ..... \\ &= |\alpha - x_0||\varphi'(\varepsilon_0)||\varphi'(\varepsilon_1)||\varphi'(\varepsilon_2)|...|\varphi'(\varepsilon_n)|. \end{aligned}$$

Assuming $|\varphi'(x)| \leq \rho$ $(a_0 \leq x \leq b_0)$, then $|\varphi'(\varepsilon_i)| \leq \rho$ $(i = 0, 1, 2, ..., n)$, we have

$$|\alpha - x_{n+1}| \leq |\alpha - x_0|\rho^{n+1}.$$

Thus,

$$\lim_{n\to\infty} |\alpha - x_{n+1}| \leq |\alpha - x_0| \lim_{n\to\infty} \rho^{n+1},$$

for $\lim_{n\to\infty} \rho^{n+1}$,

$$\lim_{n\to\infty} \rho^{n+1} = \{ \begin{array}{l} 0,\ if\ \rho < 1,\ i.e.,\ |\varphi'(x)| < 1 \\ \infty,\ if\ \rho > 1,\ i.e.,\ |\varphi'(x)| > 1 \end{array} .$$

Therefore,

$$\lim_{n\to\infty} x_{n+1} = \alpha,\ if\ and\ only\ if\ |\varphi'(x)| \leq \rho < 1 for\ x \in [a_0, b_0]$$

Examples

Example 4.3 *Find the root of $x^2 + \ln x - 2 = 0$, which lies between 1 and 2, by fixed-point iteration method, correct to four decimal places.*

Solution:

Let $f(x) = x^2 - \ln x - 2 = 0$. Now, $f(1) = -1 < 0$, $f(2) = 2.69 > 0$, therefore, there exists a root α in the interval $[1, 2]$.

We can re-write the equation as:

$$x = \sqrt{2 - \ln x} = \varphi(x),\ \varphi'(x) = -\frac{1}{2x\sqrt{2 - \ln x}}$$

so, $\varphi''(x) > 0 (x \in [1, 2])$, which is a monotonic increasing function in the interval $[1, 2]$. Then $|\varphi'(x)| < 1$, because

$$\varphi'(x) \in [\varphi'(1), \varphi'(2)] = [-0.3536, -0.2187],\ \max|\varphi'(x)| = |\varphi'(2)| = 0.3536 < 1.$$

Therefore, $\varphi(x) = \sqrt{2 - \ln x}$ gives us a convergent sequence of iteration, namely, the iteration formula $x_{n+1} = \varphi(x_n)$ is convergent. We take the initial guess value as $x_0 = 1$.

Computation of α $(1 < \alpha < 2)$

n	x_n	$\varphi(x_n) = \sqrt{2 - \ln x_n}$
0	1	1.4142
1	1.4142	1.2859
2	1.2859	1.3223
3	1.3223	1.3117
4	1.3117	1.3148
5	1.3148	1.3139
6	1.3139	1.3142
7	1.3142	1.3141
8	1.3141	1.3141

Thus, $\alpha \approx$ **1.3141** is the root of the equation between 1 and 2, correct up to four decimal places.

Example 4.4 *Find the root of the equation* $\sin x + 5x - 1 = 0$ *by the iteration method, correct to four significant figures.*

Solution:

We first get an approximate location by the method of tabulation. Here

$$f(x) = \sin x + 5x - 1, f(0) = -1, f(1) = 5.8415, \Rightarrow f(0)f(1) < 0,$$

so that there exists a root α lying in $(0, 1)$.

We rewrite the $f(x) = \sin x + 5x - 1 = 0$ as

$$x = \varphi(x) = \frac{1 - \sin x}{5},$$

then, $\varphi'(x) = -\frac{1}{5}\cos x$. Therefore, $|\varphi'(x)| = \frac{1}{5}|cosx| < 1 \; for \; x \in (0, 1)$ and the convergence criteria is satisfied.

The successive approximations of the root are computed in tabular form as follows (the initial guess values are different in two folowing cases):

Computation of α $(0 < \alpha < 1)$, $x_0 = 0.5$.

n	x_n	$\varphi(x_n) = 1 - \sin x_n$
0	0.5	0.1041
1	0.1041	0.1792
2	0.1792	0.1644
3	0.1644	0.1673
4	0.1673	0.1667
5	0.1667	0.1668
6	0.1668	0.1668

Computation of α $(0 < \alpha < 1)$, $x_0 = 0.5$.

n	x_n	$\varphi(x_n) = 1 - \sin x_n$
0	0	0.2000
1	0.2000	0.1603
2	0.1603	0.1681
3	0.1681	0.1665
4	0.1665	0.1669
5	0.1669	0.1668
6	0.1668	0.1668

Thus, $\alpha \approx 0.1668$ is a root of the equation with different initial guess values, correct up to four significant figures.

Exercises

1. Find a root of the equation $\tan x + x = 0$ by fixed point iteration method, correct to three significant figures. (Ans: 2.03)

2. Find a root of the equation $10^x + x - 4 = 0$ by fixed point iteration method, correct to four significant figures. (Ans: 0.5392)

4.4 Newton-Raphson Method

Newtons (or the Newton-Raphson) method is one of the most powerful and well-known numerical methods for solving a root-finding problem. There are many ways of introducing Newtons method. If we only want an algorithm, we can consider the technique graphically, as is often done in calculus. Another way of introducing Newtons method, which is discussed next, is based on Taylor polynomials. We will see there that this particular derivation produces not only the method, but also a bound for the error of the approximation.

Suppose that $f \in C^2[a, b]$. Let x_0 be an approximation of the root of $f(x) = 0$, whose real root is α. Thus, $\alpha = x_0 + h$, where h is the correction (small) to be applied to x_0 to give the exact value of the root. Therefore,

$$f(\alpha) = f(x_0 + h) = 0$$

Consider the first Taylor polynomial for $f(x)$ expanded about x_0 and evaluated at $x = \alpha$,

$$f(\alpha) = f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(\xi_\alpha)$$

where ξ_α lies between x_0 and α. Since $f(\alpha) = 0$, this equation gives

$$f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(\xi_\alpha) = 0$$

Since h is small, the term involving h^2 is much smaller, so

$$f(x_0) + hf'(x_0) \approx 0$$

Solving for h gives

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

Substituting this value of h into $\alpha = x_0 + h$ we get a better approximation to the root α of $f(x) = 0$ as

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

This sets the stage for Newtons method, which starts with an initial approximation x_0 and generates the sequence $\{x_n\}_{n=0}^{\infty}$, by

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})},\ for\ n \geq 1$$

This formula is known as the iteration formula for Newton-Raphson Method.

Figure 4.2 illustrates how the approximations are obtained using successive tangents. The figure represents a magnified view of the graph where it crosses the x-axis at $x = \alpha$. Starting with the initial approximation x_0, the approximation x_1 is the x-intercept of the tangent line to the graph of $f(x)$ at $(x_0, f(x_0))$. The approximation x_2 is the x-intercept of the tangent line to the graph of f at $(x_1, f(x_1))$ and so on.

Notes:

1. The method fails if $f'(x) = 0$ or very small in the neighbourhood of the root.

2. The sufficient condition for convergence of Newton-Raphson method is $|f(x)f''(x)| < [f'(x)]^2$

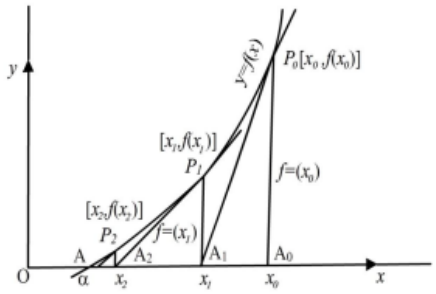

Figure 4.2: Newton Raphson method: Geometrical significance

3. The Newton Raphson method is said to have a quadratic rate of convergence.

Example 4.5 *By using Newton-Raphson Method, find the real root of the following equation*

$$\sin x + 5x - 1 = 0.$$

Solution: Let $f(x) = \sin x + 5x - 1$. Since $f(0.1) = -0.40, f(0.5) = 1.98$, then there exists a real root of $f(x) = 0$, which lies in $[0.1, 0.5]$.

Now, $f'(x) = \cos x + 5$, and $f'(0.2) = 5.98$.

Taking $x_0 = 0.2$, the successive approximations of the root are computed in the following table:

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n = -\frac{f(x_n)}{f'(x_n)}$	$x_{n+1} = x_n + h_n$
0	0.2	0.1987	5.9801	-0.0332	0.1668
1	0.1668	2.7617e-05	5.9861	-4.6135e-06	0.1668

Therefore, 0.167 is the root of $f(x) = 0$, correct up to three decimal places.

Example 4.6 *Find a real root of* $\ln x + x - 4 = 0$*, by Newton-Raphson method, correct to three decimal places.*

Solution:

Let $f(x) = \ln x + x - 4$. Since $f(1) = -3, f(2) = -1.306, f(3) = 0.0986$, then there exists a real root of $f(x) = 0$, which lies in $[2, 3]$.

Now, $f'(x) = \frac{1}{x} + 1$, and $f'(2) = 1.5$.

Taking $x_0 = 2$, the successive approximations of the root are computed in the following table:

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n = -\frac{f(x_n)}{f'(x_n)}$	$x_{n+1} = x_n + h_n$
0	2	-1.3069	1.5000	0.8712	2.8712
1	2.8712	-0.0741	1.3483	0.0549	2.9261
2	2.9261	-2.2952e-04	1.3418	1.7106e-04	2.9263
3	2.9263	3.8826e-05	1.3417	-2.8938e-05	2.9263

Therefore, 2.926 is the root of $f(x) = 0$, correct up to three decimal places.

Example 4.7 *Find the cube of 10, that is, $\sqrt[3]{10}$, correct to 4-significant figures.*

Solution:

Let $x = \sqrt[3]{10}$, then $f(x) = x^3 - 10 = 0$, $f'(x) = 3x^2$, the Newton-Raphson's iterative formula gives

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^3 - 10}{3x_n^2} = \frac{1}{3}(2x_n + \frac{10}{x_n^2})$$

Since $f(2) = -2, f(3) = 17$, then there exists a real root of $f(x) = 0$ lying in $[2, 3]$.

$f'(2) = 12$, taking $x_0 = 2$, the successive approximations of the root are computed as the following:

$$x_1 = \frac{1}{3}(2x_0 + \frac{10}{x_0^2}) = \frac{1}{3}(2 \times 2 + \frac{10}{2^2}) = 2.1667$$
$$x_2 = \frac{1}{3}(2x_1 + \frac{10}{x_1^2}) = \frac{1}{3}(2 \times 2.1667 + \frac{10}{2.1667^2}) = 2.1545$$
$$x_3 = \frac{1}{3}(2 \times 2.1545 + \frac{10}{2.1545^2}) = 2.1544$$
$$x_4 = \frac{1}{3}(2 \times 2.1544 + \frac{10}{2.1544^2}) = 2.1544$$

Therefore, $\sqrt[3]{10} \approx 2.154$, correct to 4-significant figures.

Exercises

1. Find a real root of $x^5 + 2x^4 - 12x + 1 = 0$ by Newton-Raphson method, correct to three significant figures.

2. Find a real root of $3\cos x - x + 1 = 0$ by Newton-Raphson method, correct to three significant figures

3. Find the square of 7 by Newton-Raphson method, that is, $\sqrt{7}$, correct to 4-significant figures.

Chapter 5

Interpolation

Lecture 1

A census of the population of the United States is taken every 10 years. The following table lists the population, in thousands of people, from 1950 to 2000, and the data are also represented in the figure.

Year	1950	1960	1970	1980	1990	2000
Population (in thousands)	151,326	179,323	203,302	226,542	249,633	281,422

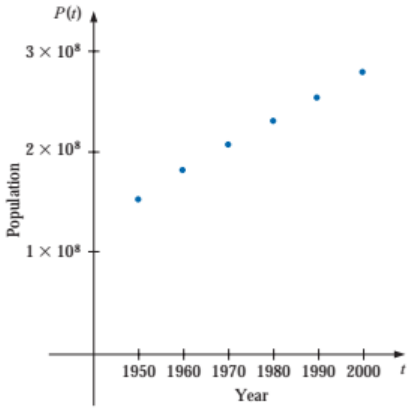

Figure 5.1: Population of the United States from 1950 to 2000

In reviewing these data, we might ask whether they could be used to provide a reasonable estimate of the population, say, in 1975 or even in the year 2020. Predictions of this type can be obtained by using a function that fits the given data. This process is called interpolation and is the subject of this chapter.

5.1 Lagrangian Polynomial

One of the most useful and well-known classes of functions mapping the set of real numbers into itself is the algebraic polynomials, the set of functions of the form

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$$

where n is a nonnegative integer and $a_0, a_1, ..., a_n$ are real constants. One reason for their importance is that they uniformly approximate continuous functions. By this we mean that given any function, defined and continuous on a closed and bounded interval, there exists a polynomial that is as close to the given function as desired.

Another important reason for considering the class of polynomials in the approximation of functions is that the derivative and indefinite integral of a polynomial are easy to determine and are also polynomials. For these reasons, polynomials are often used for approximating continuous functions.

5.1.1 Lagrange Interpolating Polynomials

The problem of determining a polynomial of degree one that passes through the distinct points (x_0, y_0) and (x_1, y_1) is the same as approximating a function f for which $f(x_0) = y_0$ and $f(x_1) = y_1$ by means of a first-degree polynomial **interpolating**, or agreeing with, the values of f at the given points. Using this polynomial for approximation within the interval given by the endpoints is called **polynomial interpolation**.

Define the functions

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}, l_1(x) = \frac{x - x_0}{x_1 - x_0}$$

The linear Lagrange interpolating polynomial through (x_0, y_0) and (x_1, y_1) is

$$P_1(x) = l_0(x)f(x_0) + l_1(x)f(x_1) = \frac{x - x_1}{x_0 - x_1}f(x_0) + \frac{x - x_0}{x_1 - x_0}f(x_1).$$

Note that

$$l_0(x_0) = 1, l_0(x_1) = 0; \text{ and } l_1(x_0) = 0, l_1(x_1) = 1,$$

which implies that

$$P_1(x_0) = 1 \times f(x_0) + 0 \times f(x_1) = f(x_0) = y_0$$

and

$$P_1(x_1) = 0 \times f(x_0) + 1 \times f(x_1) = f(x_1) = y_1$$

So $P_1(x)$ is the unique polynomial of degree at most one that passes through (x_0, y_0) and (x_1, y_1).

Example 5.1 *Determine the linear Lagrange interpolating polynomial that passes through the points (2, 4) and (5, 1).*

Solution: In this case, $x_0 = 2, x_1 = 5, y_0 = 4, y_1 = 1$, then we have

$$l_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x - 5}{2 - 5} = -\frac{1}{3}(x - 5),$$
$$l_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - 2}{5 - 2} = \frac{1}{3}(x - 2)$$

so

$$P_1(x) = l_0(x)f(x_0) + l_1(x)f(x_1) = -\frac{1}{3}(x - 5) \times 4 + \frac{1}{3}(x - 1) \times 1 = -x + 6.$$

The graph of y = P(x) is shown in Figure 5.2.

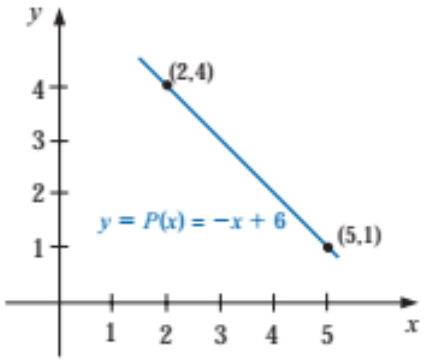

Figure 5.2: The graph of $y = -x + 6$

To generalize the concept of linear interpolation, consider the construction of a polynomial of degree at most n that passes through the $n + 1$ points, see Figure 5.3,

$$(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n)).$$

In this case we first construct, for each $k = 0, 1, 2, .., n$, a function $l_k(x)$ with the property that $l_k(x_i) = 0$ when $i \neq k$ and $l_k(x_k) = 1$. To satisfy $l_k(x_i) = 0$

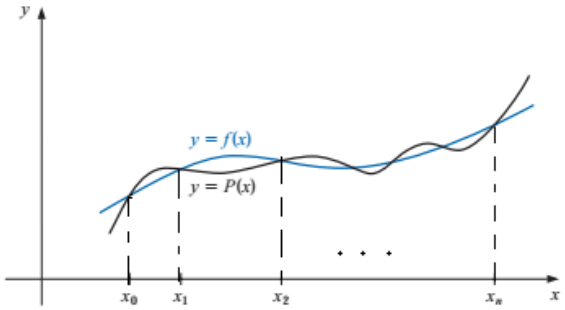

Figure 5.3: $P_n(x)$ passes through the $n+1$ points

for each $i \neq k$ requires that the numerator of $l_k(x)$ contain the term

$$(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n).$$

To satisfy $l_k(x_k) = 1$, the denominator of $l_k(x)$ must be this same term but evaluated at $x = x_k$. Thus

$$l_k(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}$$

The interpolating polynomial is easily described once the form of $l_k(x)$ is known. This polynomial, called the ***nth Lagrange interpolating polynomial***, is defined in the following theorem.

Theorem 5.1 *If $x_0, x_1, \ldots, x_n$ are $n+1$ distinct numbers and f is a function whose values are given at these numbers, then a unique polynomial $P_n(x)$ of degree at most n exists with $f(x_k) = P_n(x_k)$, for each $k = 0, 1, ..., n.$, this polynomial is given by*

$$P_n(x) = l_0(x)f(x_0) + l_1(x)f(x_1) + \cdots + l_n(x)f(x_n) = \sum_{k=0}^{n} l_k(x)f(x_k) \equiv L_n(x),$$

where
$l_k(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}$, *for* $k = 0, 1, 2, .., n$

The interpolation formula named for Joseph Louis Lagrange (1736-1813) was likely known by Isaac Newton around 1675, but it appears to first have been published in 1779 by Edward Waring (1736-1798). Lagrange wrote extensively on the subject of interpolation and his work had significant influence on later mathematicians. He published this result in 1795.

Example 5.2 *(a) Use the numbers (called nodes)* $x_0 = 2, x_1 = 2.75$*, and* $x_2 = 4$ *to find the second Lagrange interpolating polynomial for* $f(x) = 1/x$*.*
(b) Use this polynomial to approximate $f(3) = 1/3$*.*

Solution: (a) We first determine the coefficient polynomials $l_0(x), l_1(x)$, and $l_2(x)$. In nested form they are

$$l_0(x) = \frac{(x-2.75)(x-4)}{(2-2.75)(2-4)} = \frac{2}{3}(x-2.75)(x-4),$$
$$l_1(x) = \frac{(x-2)(x-4)}{(2.75-2)(2.75-4)} = -\frac{16}{15}(x-2)(x-4),$$
$$l_2(x) = \frac{(x-2)(x-2.75)}{(4-2)(4-2.75)} = \frac{2}{5}(x-2)(x-2.75).$$

Also, $f(x_0) = f(2) = 1/2, f(x_1) = f(2.75) = 4/11$, and $f(x_2) = f(4) = 1/4$, so

$$\begin{aligned} L_2(x) &= l_0(x)f(x_0) + l_1(x)f(x_1) + l_2(x)f(x_2) \\ &= \frac{2}{3}(x-2.75)(x-4) \times \frac{1}{2} - \frac{16}{15}(x-2)(x-4) \times \frac{4}{11} + \frac{2}{5}(x-2)(x-2.75) \times \frac{1}{4} \\ &= \frac{1}{22}x^2 - \frac{35}{88}x + \frac{49}{44} \end{aligned}$$

(b) An approximation to $f(3) = 1/3$ is

$$f(3) \approx P_2(3) = \frac{1}{22} \times 3^2 - \frac{35}{88} \times 3 + \frac{49}{44} = \frac{29}{88} \approx 0.32955.$$

Exercises

1. For the given functions $f(x)$, let $x_0 = 0, x_1 = 0.6$, and $x_2 = 0.9$. Construct interpolation polynomials of degree at most one and at most two to approximate $f(0.45)$, and find the absolute error.

   (a) $f(x) = cosx$;

   (b) $f(x) = \ln(x+1)$;

   (c) $f(x) = \sqrt{x+1}$.

5.2 Finite Differences

5.2.1 Introduction

When a function $f(x) = 0$ is known explicitly, it is easy to calculate the value (or values) of $f(x)$, corresponding to a fixed given value x. However, when the explicit form of the function $f(x) = 0$ is not known, it is possible to obtain an approximate value of the function up to a desired level of accuracy with the help of finite differences. A function $f(x) = 0$, x being an independent variable and y, a dependent variable, is considered. Let x takes equidistant values $a, a+h, a+2h, a+3h, ...$ (which are finite in numbers); h is the equal spacing, then $f(a), f(a+h), f(a+2h), f(a+3h), ...$ are the corresponding values of y. The values of the independent variable x are termed as **arguments** and the corresponding values of the dependent variable y are called **entries**.

5.2.2 Operators

1. **Forward Difference** (Δ)
The forward difference, denoted by Δ, is defined as

$$\Delta y = \Delta f(x) = f(x+h) - f(x)$$

h is called the interval of differencing; $\Delta f(x)$ is the first order differences. We get the second order differences (denoted by Δ^2) when Δ is operated twice on $f(x)$, Thus

$$\begin{aligned}\Delta^2 y = \Delta^2 f(x) = \Delta[\Delta f(x)] &= \Delta[f(x+h) - f(x)] \\ &= \Delta f(x+h) - \Delta f(x) \\ &= [f(x+2h) - f(x+h)] - [f(x+h) - f(x)] \\ &= f(x+2h) - 2f(x+h) + f(x)]\end{aligned}$$

Similarly, Δ^3, Δ^4 may be calculated.

Forward Difference Table with 4 arguments

x	y	1st Differences Δy	Second Differences $\Delta^2 y$	Third Differences $\Delta^3 y$
$x_0 = a$	$y_0 = f(a)$			
$x_1 = a+h$	$y_1 = f(a+h)$	$\Delta y_0 = y_1 - y_0$		
$x_2 = a+2h$	$y_2 = f(a+2h)$	$\Delta y_1 = y_2 - y_1$	$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$	
$x_3 = a+3h$	$y_3 = f(a+3h)$	$\Delta y_2 = y_3 - y_2$	$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$	$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0$

Note: If $f(x)$ is a polynomial of degree n in x, then $f(x)$ is a constant and $\Delta^{n+1} f(x)$ is zero. Conversely, if the $(n+1)$th difference is zero, then the polynomial is less or equal to degree n.

Example 5.3 *Let* $f(x) = x^2 + 8x - 5$*, then*

$$\begin{aligned}\Delta f(x) &= f(x+h) - f(x)\\ &= [(x+h)^2 + 8(x+h) - 5] - [x^2 + 8x - 5]\\ &= 2xh + h^2 + 8h\end{aligned}$$

and

$$\begin{aligned}\Delta^2 f(x) &= \Delta f(x+h) - \Delta f(x)\\ &= [2(x+h)h + h^2 + 8h] - [2xh + h^2 + 8h]\\ &= 2h^2, (which is a constant)\end{aligned}$$

Hence

$$\Delta^3 f(x) = \Delta^2 f(x+h) - \Delta^2 f(x) = 2h^2 - 2h^2 = 0.$$

2. **Backward Difference** (∇)

We now define a difference operator, known as backward difference operator, given by

$$\nabla f(x+h) = f(x+h) - f(x)$$

Please note that the backward difference of $f(x+h)$ is same as the forward difference of $f(x)$, that is,

$$\nabla f(x+h) = f(x+h) - f(x) = \Delta f(x)$$

Backward Difference Table with 4 arguments

x	y	1st Differences ∇y	Second Differences $\nabla 2y$	Third Differences $\nabla^3 y$
$x_0 = a$	$y_0 = f(a)$			
$x_1 = a + h$	$y_1 = f(a+h)$	$\nabla y_1 = y_1 - y_0$		
$x_2 = a + 2h$	$y_2 = f(a+2h)$	$\nabla y_2 = y_2 - y_1$	$\nabla^2 y_2 = \nabla y_2 - \nabla y_1$	
$x_3 = a + 3h$	$y_3 = f(a+3h)$	$\nabla y_3 = y_3 - y_2$	$\nabla^2 y_3 = \nabla y_3 - \nabla y_2$	$\nabla^3 y_3 = \nabla^2 y_3 - \nabla^2 y_2$

3. **E-Operator** (E)

We now define the operator as

$$Ef(x) = f(x+h),$$

that is, the operator gives an increment of h to the argument.
Note: $E^{-1} f(x) = f(x-h)$.

4. **Relation between E and Δ**

Let $y = f(x)$ be a function of an independent variable x and the dependent variable y. We have

$$\Delta f(x) = f(x+h) - f(x) = Ef(x) - f(x) = (E-1)f(x) \Rightarrow \Delta \equiv E - 1 \Rightarrow E = 1 + \Delta$$

where h is the interval of differencing.

Relation between operator of finite differences and differential operator of differential calculus

We know

$$\frac{df(x)}{dx} = f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = Df(x+h)$$

where $D \equiv \frac{d}{dx}$.

Now, by Taylor series expansion, we have

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + ... \\ &= f(x) + hDf(x) + \frac{h^2}{2!}D^2 f(x) + ... \\ &= [1 + hD + \frac{h^2}{2!}D^2 + ...]f(x) \end{aligned}$$

or

$$\begin{aligned} E[f(x)] &= e^{hD} f(x) \\ &\Rightarrow E \equiv e^{hD} \equiv 1 + \Delta, (E = 1 + \Delta) \end{aligned}$$

or

$$hD \equiv log(1+\Delta) \equiv \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - ...$$

or

$$D \equiv \frac{1}{h}(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - ...)$$

Example 5.4 *Evaluate the following:*

(1) $\Delta \log[f(x)]$

Solution:

$$\begin{aligned} \Delta \log[f(x)] &= \log[f(x+h)] - \log[f(x)] \\ &= \log[\frac{f(x+h)}{f(x)}] = \log[\frac{Ef(x)}{f(x)}] \\ &= \log[\frac{(1+\Delta)f(x)}{f(x)}] = \log[\frac{f(x) + \Delta f(x)}{f(x)}] \\ &= \log[1 + \frac{\Delta f(x)}{f(x)}] \end{aligned}$$

(2) $\Delta^2(ab^{cx})$

Solution:

$$\begin{aligned}\Delta^2(ab^{cx}) = a\Delta^2(b^{cx}) &= a\Delta(b^{c(x+h)} - b^{cx}) \\ &= a\Delta b^{c(x+h)} - a\Delta b^{cx} \\ &= a[b^{c(x+2h)} - b^{c(x+h)}] - a[b^{c(x+h)} - b^{cx}] \\ &= a[b^{c(x+2h)} - 2b^{c(x+h)} - b^{cx}] \\ &= ab^{cx}[b^{2hc} - 2b^{hc} + 1] \\ &= ab^{cx}[b^{hc} - 1]^2\end{aligned}$$

The interval of differencing being h.

(3) $\Delta^2(2^x)$

Solution:

$$\begin{aligned}\Delta^2(2^x) &= \Delta[\Delta(2^x)] \\ &= \Delta[2^{x+h} - 2^x] = \Delta 2^{x+h} - \Delta 2^x \\ &= [2^{x+2h} - 2^{x+h}] - [2^{x+h} - 2^x] \\ &= 2^{x+2h} - 2 \cdot 2^{x+2h} + 2^x\end{aligned}$$

(4) $\nabla^2(2^x)$

Solution:

$$\begin{aligned}\nabla^2(2^x) &= \nabla[\nabla(2^x)] \\ &= \nabla[2^x - 2^{x-h}] = \nabla 2^x - \nabla 2^{x-h} \\ &= [2^x - 2^{x-h}] - [2^{x-h} - 2^{x-2h}] \\ &= 2^x - 2 \cdot 2^{x-h} + 2^{x-2h}\end{aligned}$$

(5) $E^2(e^x)$

Solution:

$$E^2(e^x) = E[E(e^x)] = E[(e^{x+h})] = e^{x+2h}$$

(6) $(\frac{\Delta^2}{E})f(x)$, where $f(x) = x^3$

Solution:

$$\begin{aligned}(\frac{\Delta^2}{E})f(x) &= (\frac{\Delta^2}{E})x^3 \\ &= \Delta^2 E^{-1} x^3 \\ &= \Delta^2 (x-h)^3 \\ &= \Delta[x^3 - (x-h)^3] = \Delta x^3 - \Delta(x-h)^3 \\ &= [(x+h)^3 - x^3] - [x^3 - (x-h)^3] \\ &= 6xh^2\end{aligned}$$

(7) $\frac{\Delta^2 f(x)}{Ef(x)}$, where $f(x) = x^3$

Solution:

$$\frac{\Delta^2 f(x)}{Ef(x)} = \frac{\Delta^2 x^3}{Ex^3} = \frac{\Delta[(x+h)^3 - x^3]}{(x+h)^3}$$
$$= \frac{[(x+2h)^3 - (x+h)^3] - [(x+h)^3 - x^3]}{(x+h)^3}$$
$$= \frac{6h^3 + 6xh^2}{(x+h)^3}$$

Example 5.5 *Prove the following relations:*

(1) $E\nabla \equiv \nabla E \equiv \Delta$,
(2) $(1+\Delta)(1-\nabla) \equiv 1$.

Solution:

(1) Let $y = f(x)$ be a function of an independent variable x and the dependent variable y. Now, $E\nabla f(x) = E[f(x) - f(x-h)]$, h being the interval of differencing

$$E\nabla f(x) = E[f(x) - f(x-h)]$$
$$= Ef(x) - Ef(x-h)$$
$$= f(x+h) - f(x) = \Delta f(x) \Rightarrow E\nabla = \Delta$$

Again

$$\nabla E f(x) = \nabla f(x+h)$$
$$= f(x+h) - f(x) = \Delta f(x) \Rightarrow \nabla E = \Delta$$

Hence

$$E\nabla \equiv \nabla E \equiv \Delta$$

(2) Let $y = f(x)$ be a function of an independent variable x and the dependent variable y. Then,

$$(1+\Delta)(1-\nabla)f(x) = (1+\Delta)[f(x) - \nabla f(x)]$$
$$= (1+\Delta)f(x) - [f(x) - f(x-h)]$$
$$= (1+\Delta)f(x-h)$$
$$= f(x-h) + \Delta f(x-h)$$
$$= f(x-h) + [f(x) - f(x-h)] = f(x)$$
$$\Rightarrow (1+\Delta)(1-\nabla) \equiv 1$$

Example 5.6 *$f(x)$ is polynomial in x with the following functional values: $f(2) = f(3) = 27, f(4) = 78, f(5) = 169$. Find the function $f(x)$.*

Solution:

Since four entries (i.e., four functional values) are given, $f(x)$ can be represented by a polynomial of degree 3. Let $f(x) = a + bx + cx^2 + dx^3$, where a, b, c, d are constants to be determined. Now,

$$f(2) = a + 2b + 4c + 8d = 27$$
$$f(3) = a + 3b + 9c + 27d = 27$$
$$f(4) = a + 4b + 16c + 64d = 78$$
$$f(5) = a + 5b + 25c + 125d = 169$$

Solving these equations, we get

$$a = 224, b = -\frac{1051}{6}, c = 42, d = -\frac{11}{6},$$

Therefore, the required function is $f(x) = 224 - \frac{1051}{6}x + 42x^2 - \frac{11}{6}x^3$.

Example 5.7 *Compute the missing terms in the following table:*

x	*2*	*3*	*4*	*5*	*6*	*7*	*8*
$f(x)$	*0.135*	-	*0.111*	*0.100*	-	*0.082*	*0.074*

Solution: Since five entries are given, f(x) can be represented by a polynomial of degree four. Hence $\Delta^4 f(x) = constant$ and $\Delta^5 f(x) = 0$, i.e.

$$\Rightarrow (E-1)^5 f(x) = 0$$
$$\Rightarrow (E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)f(x) = 0$$
$$\Rightarrow E^5 f(x) - 5E^4 f(x) + 10E^3 f(x) - 10E^2 f(x) + 5Ef(x) - f(x) = 0$$
$$\Rightarrow f(x+5) - 5f(x+4) + 10f(x+3) - 10f(x+2) + 5f(x+1) - f(x) = 0$$

where the interval of differencing is $h = 1$. Now, putting $x = 2$ and $x = 3$, we get

$$f(7) - 5f(6) + 10f(5) - 10f(4) + 5f(3) - f(2) = 0$$
$$f(8) - 5f(7) + 10f(6) - 10f(5) + 5f(4) - f(3) = 0$$

Substituting the values of $f(8), f(7), f(5), f(4)$ and $f(2)$, we get,

$$f(3) - f(6) = 0.0326$$
$$10f(6) - f(3) = 0.781$$

Solving we get $f(3) = 0.123, f(6) = 0.0904$.

5.2.3 Exercises

1. Evaluate the following:
(1) $(\frac{\Delta^2}{E})e^x$; (2) $\frac{Ee^x}{\Delta^2 e^x}$; (3) $\frac{\Delta}{E}\sin 2x$.
2. Find $f(1.1)$ from the following table:

x	1	2	3	4	5
$f(x)$	7	12	29	64	123

3. Prove that $e^{-hD} \equiv 1 - \nabla$.
(Hint: Already proved $\nabla E = \Delta$, therefore, $E = 1 + \Delta = 1 + \nabla E$)

5.2.4 Divided Differences

When the arguments are not equi-spaced, we use divided differences. Let $y = f(x)$ be a function whose functional form is not known but its values at (n+1) points, namely, $x_0, x_1, ..., x_n$ are known.

The divided difference of first order for the points x_0, x_1 is defined as

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

For the points x_0, x_1, x_2, the divided difference of second order is defined as

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_0, x_1] - f[x_1, x_2]}{x_0 - x_2} \\ &= \frac{1}{x_0 - x_2}[\frac{f(x_1) - f(x_0)}{x_1 - x_0} - \frac{f(x_2) - f(x_1)}{x_2 - x_1}] \\ &= \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)} + \frac{f(x_1)}{(x_0 - x_1)(x_2 - x_1)} + \frac{f(x_2)}{(x_0 - x_2)(x_1 - x_2)} \end{aligned}$$

In the similar manner, the divided difference of n^{th} order for the points $x_0, x_1, ..., x_n$ is defined as

$$\begin{aligned} f[x_0, x_1, x_2, ..., x_n] &= \frac{f[x_0, x_1, ..., x_{n-1}] - f[x_1, x_2, ..., x_n]}{x_0 - x_n} \\ &= \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)...(x_n - x_0)} + \frac{f(x_1)}{(x_0 - x_1)(x_2 - x_1)...(x_n - x_1)} \\ &+ ... + \frac{f(x_n)}{(x_0 - x_n)(x_1 - x_n)...(x_{n-1} - x_n)} \end{aligned}$$

Property 1: $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = f[x_1, x_0]$.
This implies divided difference is symmetric.

Property 2: If the arguments are equi-spaced, then

$$f[x_1, x] = \frac{f(x) - f(x_1)}{x - x_1} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{f(x + \varepsilon) - f(x)}{\varepsilon}$$

where we set $x_1 = x + \varepsilon$. Since, the arguments are equi-spaced, $\varepsilon \to 0$ and we get

$$\lim_{\varepsilon \to 0} f[x + \varepsilon, x] = \lim_{\varepsilon \to 0} \frac{f[x + \varepsilon) - f(x)]}{\varepsilon} \Rightarrow f[x, x] = f'(x)$$

provided $f(x)$ is differentiable.

The divided difference tables with 5 arguments are as follows:

x	y	1st	2nd	3rd	4th
x_0	y_0				
x_1	y_1	$f[x_0, x_1]$			
x_2	y_2	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
x_3	y_3	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$	
x_4	y_4	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$

Example 5.8 *Obtain the divided difference table for the function* $y = f(x)$ *given by*

x	-1	1	4	6
y	1	-3	21	127

Solution: The divided difference tables with 4 arguments are as follows:

x	y	1st	2nd	3rd
-1	1			
1	-3	$\frac{-3-1}{1-(-1)} = -2$		
4	21	$\frac{21-(-3)}{4-1} = 8$	$\frac{8-(-2)}{4-(-1)} = 2$	
6	127	$\frac{127-21}{6-4} = 53$	$\frac{53-8}{6-1} = 9$	$\frac{9-2}{6-(-1)} = 1$

Exercise

1. Obtain the divided difference table for the function $y = f(x)$ given by

x	-1	-2	2	4
y	-1	-9	11	69

2. Obtain the divided difference table for the function $y = f(x)$ given by

x	0	1	3	4
y	0	2	8	9

5.3 Interpolation in Newton's Polynomial

The problem of determining a polynomial of degree one that passes through the distinct points (x_0, y_0) and (x_1, y_1) is the same as approximating a function f for which $f(x_0) = y_0$ and $f(x_1) = y_1$ by means of a first-degree polynomial **interpolating**, or agreeing with, the values of f at the given points. From Section 5.1, we have the lagrange's polynomial as follow:

$$P_1(x) = l_0(x)y_0 + l_1(x)y_1 = \frac{x - x_1}{x_0 - x_1}y_0 + \frac{x - x_0}{x_1 - x_0}y_1 = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

We rewrite in the following form

$$P_1(x) = f(x_0) + f[x_0, x_1](x - x_0)$$

For three distinct points x_0, x_1, x_2, the interpolating polynomial of degree 2:

$$P_2(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1)$$

with the conditions

$$P_2(x_0) = y_0; P_2(x_1) = y_1; P_2(x_2) = y_2.$$

Then, we get

$$a_0 = f(x_0); a_1 = f[x_0, x_1]; a_2 = f[x_0, x_1, x_2].$$

Therefore, the Newton's polynomial can be obtained

$$P_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \equiv N_2(x).$$

For $(n + 1)$ points: $(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))$, the interpolating polynomial of degree n:

$$N_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + ... + a_n(x - x_0)(x - x_1)...(x - x_{n-1})$$

with the conditions

$$N_n(x_0) = y_0; N_n(x_1) = y_1; N_n(x_2) = y_2; ...; N_n(x_n) = y_n.$$

Hence

$$N_n(x_0) = a_0 = y_0$$
$$N_n(x_1) = a_0 + a_1(x_1 - x_0) = y_1$$
$$...$$
$$N_n(x_n) = a_0 + a_1(x_n - x_0) + ... + a_n(x_n - x_0)(x_n - x_1)...(x_n - x_{n-1}) = y_n$$

Solving the above equations, we have

$$a_0 = f(x_0); a_1 = f[x_0, x_1]; a_2 = f[x_0, x_1, x_2], ..., a_n = f[x_0, x_1, ..., x_n].$$

Therefore, the Newton's polynomial (degree n) can be obtained

$$N_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + ... + f[x_0, x_1, ..., x_n](x - x_0)(x - x_1)...(x - x_{n-1})$$

Example 5.9 *suppose we are given the following data:*

x	0	5	10	15	20
y	10	13	17	20	25

find $f(8)$ and $f(17)$ with the help of given data by using Newton's polynomial.
Solution: The difference table of the given data is as follows:

x	y	1st	2nd	3rd	4th
0	10				
5	13	$\frac{13-10}{5-0} = 0.6$			
10	17	$\frac{17-13}{10-5} = 0.8$	$\frac{0.8-0.6}{10-0} = 0.02$		
15	20	$\frac{20-17}{15-10} = 0.6$	$\frac{0.6-0.8}{15-5} = -0.02$	$\frac{-0.02-0.02}{15-0} = -0.0027$	
20	25	$\frac{25-20}{20-15} = 1$	$\frac{1-0.6}{20-10} = 0.04$	$\frac{0.04-(-0.02)}{20-5} = 0.004$	$\frac{0.004+0.0027}{20-0} = 0.00034$

Thus, the Newton's polynomial (degree 4) can be

$$\begin{aligned} N_4(x) =& f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &+ f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\ &+ f[x_0, x_1, x_2, x_3, x_4](x - x_0)(x - x_1)(x - x_2)(x - x_3) \\ =& 10 + 0.6x + 0.02x(x - 5) - 0.0027x(x - 5)(x - 10) + 0.0034x(x - 5)(x - 10)(x - 15). \end{aligned}$$

Hence

$$\begin{aligned} f(8) \approx N_4(8) =& 10 + 0.6 * 8 + 0.02 * 8 * (8 - 5) - 0.0027 * 8 * (8 - 5)(8 - 10) \\ &+ 0.0034 * 8 * (8 - 5)(8 - 10)(8 - 15) = 16.5520 \end{aligned}$$

$$\begin{aligned} f(17) \approx N_4(17) =& 10 + 0.6 * 17 + 0.02 * 17 * (17 - 5) - 0.0027 * 17 * (17 - 5)(17 - 10) \\ &+ 0.0034 * 17 * (17 - 5)(17 - 10)(17 - 15) = 30.1348 \end{aligned}$$

Example 5.10 *Given*

x	3	4	5	6	7	8
y	27	64	125	216	343	512

Estimate $f(7.5)$.

Solution: The divided difference table of the given data is as follows :

x	y	1st	2nd	3rd	4th	5th
3	27					
4	64	$\frac{64-27}{4-3} = 37$				
5	125	$\frac{125-64}{5-4} = 61$	$\frac{61-37}{5-3} = 12$			
6	216	$\frac{216-125}{6-5} = 91$	$\frac{91-61}{6-4} = 15$	$\frac{15-12}{6-3} = 1$		
7	343	$\frac{343-216}{7-6} = 127$	$\frac{127-91}{7-5} = 18$	$\frac{18-15}{7-4} = 1$	0	
8	512	$\frac{512-343}{8-7} = 169$	$\frac{169-127}{8-6} = 21$	$\frac{21-18}{8-5} = 1$	0	0

$$N_5(x) = 27 + 37(x-3) + 12(x-3)(x-4) + (x-3)(x-4)(x-5) + 0,$$

$$f(7.5) \approx N_5(7.5) = 27+37(7.5-3)+12(7.5-3)(7.5-4)+(7.5-3)(7.5-4)(7.5-5) = 421.8750$$

Exercises

1. From the following data, estimate the number of students who obtained marks between 40 and 45:

Marks less than x	40	50	60	70	80
Number of students y	31	73	124	159	190

2. The area A of a circle of diameter d is given for the following values:

d	80	85	90	95	100
A	5026	5674	8362	7088	7854

Although the road is endless and faraway, I still want to pursue the truth in the world.

Chapter 6

References

1. Richard L. Burden, J. Douglas Faires. Numerical Analysis,ninth edition. 2011

2. Allgower, E. and K. Georg, Numerical continuation methods: an introduction, Springer-Verlag, New York, 1990, 388 pp. QA377.A56 668, 669

3. DeFranza, J. and D. Gagliardi, Introduction to linear algebra, McGraw-Hill, New York, 2009, 488 pp. QA184.2.D44

4. Neville, E. H. Iterative Interpolation, J. Indian Math Soc. 20: 87-120 (1934)

5. Saad, Y., Numerical methods for large eigenvalue problems, Halsted Press, New York, 1992, 346 pp. QA188.S18

6. Saad, Y., Iterative methods for sparse linear systems, (Second Edition), SIAM, Philadelphia, PA 2003, 528

pp. QA188.S17

7. Wilkinson, J. H., The algebraic eigenvalue problem, Clarendon Press, Oxford, 1965, 662 pp. QA218.W5 476, 580, 586, 593, 604, 611, 627

8. Traub, J. F., Iterative methods for the solution of equations, Prentice-Hall, Englewood Cliffs, NJ, 1964, 310 pp. QA297.T7 103