

# TE2401 Linear Algebra & Numerical Methods

Ben M. Chen

Associate Professor

National University of Singapore

Office: E4-6-7

Phone: 874-2289

Email: [bmchen@nus.edu.sg](mailto:bmchen@nus.edu.sg)

<http://vlab.ee.nus.edu.sg/~bmchen>

# Course Outlines

## Part I: Linear Algebra

**Introduction to Matrices:** Definitions of matrices, sub-matrices, square matrices, lower and upper triangular matrices, diagonal matrices, identity matrices, symmetric matrices, skew-symmetric matrices.

**Matrix Operations:** Matrix transpose, addition, multiplication.

**Linear Systems:** Definition, homogeneous systems, elementary row operations, echelon form of a matrix, row echelon form, reduced row echelon form.

**Vector Algebra:** Linear combination, Linear independence, vector space, vector subspaces, dimension and basis of a vector space, null space, determinant, inverse and rank of matrices.

## Course Outlines (cont.)

**Eigenanalysis of Matrices:** Eigenvalues and eigenvectors, characteristic equation, matrix diagonalization, quadratic forms.

**Introduction to a Software Package - MATLAB**

### **Part II: Numerical Methods**

**Introduction to Numerical Methods:** Numerical errors, absolute and relative errors, stability and convergence of numerical algorithms.

**Computation of Zeros:** Bisection method, false position method, Newton method, secant method and fixed-point method.

**Interpolation:** Lagrangian polynomials, Newton's divided difference method, least square approximation.

## Course Outlines (cont.)

**Numerical Integration:** Newton-Cotes Method, trapezoidal rule, Simpson's 1/3 rule and Simpson's 3/8 rule.

**Numerical Solutions to Ordinary Differential Equations:** Taylor series method, Euler method, Runge-Kutta method.

**Numerical Solutions to Partial Differential Equations:** Classification of 2nd order quasilinear PDE, numerical solutions.

## Lab and Final Examination

There will be a lab session for every student. It is to learn how to use MATLAB, which is capable of realizing all computations and algorithms covered in this course. The lab sessions will be held in CAD/CAM center. Students are required to submit their lab reports right after the lab session.

There will be a final examination at the end of the semester.

Your final grade of this course will be awarded as follows:

**Final Grade = Lab Report Marks (max.=10) + 90% of Exam Marks.**

## Lectures

Lectures will follow closely (but not 100%) the materials in the lecture notes.

However, certain parts of the lecture notes will not be covered and examined and this will be made known during the classes.

Attendance is essential.

ASK any question at any time during the lecture.

## Tutorials

The tutorials will start on Week 4 of the semester.

Although you should make an effort to attempt each question before the tutorial, it is NOT necessary to finish all the questions.

Some of the questions are straightforward, but quite a few are difficult and meant to serve as a platform for the introduction of new concepts.

**ASK your tutor any question related to the tutorials and the course.**

## Reference Textbooks

- E. Kreyszig, *Advanced Engineering Mathematics*, Any Ed., Wiley.
- B. A. Ayyub and R. H. McCuen, *Numerical Methods for Engineers*, Prentice Hall, 1996.



## Linear Equations and Matrices

Linear equations arise frequently in the analysis, design, and synthesis of engineering systems. It is these equations that form the focus of our study in the first half of this course. The objective is two folds. The students are exposed to **systematic methods** and the associated algorithms for some of the most widely computational tasks in linear algebra. Also, the occurrence of these tasks in engineering systems is observed via **simple examples**.

## Definitions:

The simplest type of a linear equation is given by

$$a x = b$$

where  $a$  and  $b$  are given and known, and  $x$  is unknown variable to be determined. This equation is linear as it contains only  $x$  and nothing else.

It is simple to see that the equation has:

- (a) a unique solution if  $a \neq 0$
- (b) no solution if  $a = 0$  and  $b \neq 0$
- (c) multiple solutions if  $a = 0$  and  $b = 0$

## Example:

The relationship between the voltage and current of a resistor,  $IR = V$ .

A simple generalization of the one-equation-one-unknown linear system is the two-equations-two-unknowns linear system:

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

where  $a_{11}$ ,  $a_{12}$ ,  $a_{21}$ ,  $a_{22}$ ,  $b_1$  and  $b_2$  are known constants and  $x_1$  and  $x_2$  are unknown variables to be solved. In general, a linear system with  $m$  equations and  $n$  unknowns can be written as the following form:

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

$$\vdots$$

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i$$

$$\vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

We can re-write this set of linear equations in a compact form, i.e., a matrix form:

$$\begin{array}{c}
 \begin{matrix} \text{coefficient} \\ \text{matrix} \end{matrix} \left[ \begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array} \right] \begin{matrix} \left( \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) \\ \mathbf{x} \end{matrix} = \begin{matrix} \left( \begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_m \end{array} \right) \\ \mathbf{b} \end{matrix} \\
 \begin{matrix} \text{data} \\ \text{vector} \end{matrix} \\
 \mathbf{A} \mathbf{x} = \mathbf{b} \\
 \begin{matrix} \text{vector of unknowns} \end{matrix}
 \end{array}$$

Matrix  $\mathbf{A}$  has  $m$  rows and  $n$  columns. Such a matrix is called an  $m \times n$  matrix.

Each of the numbers that constitute the matrix is called an element of  $\mathbf{A}$ .

The element sitting on the  $i$ -th row and  $j$ -th column, or simply the  $(i, j)$ -th element, is denoted by  $a_{ij}$ .

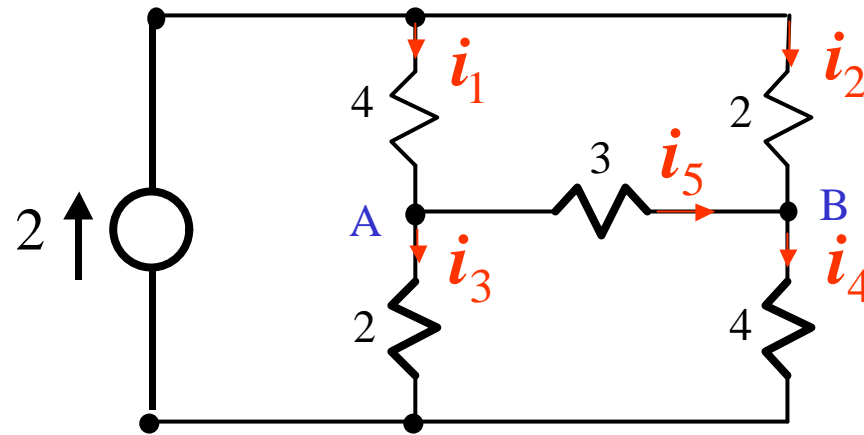
A matrix with one column or one row is also called a vector. For example,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \text{ is a column vector of length } n$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} \text{ is a column vector of length } m$$

$$\mathbf{c} = [1 \quad 3 \quad 2 \quad 4] \text{ is a row vector of length } 4$$

Example:



KCL at Node A:  $i_1 - i_3 - i_5 = 0$

KCL at Node B:  $i_2 - i_4 + i_5 = 0$

KVL to left loop:  $4i_1 + 2i_3 = 2$

KVL to right upper loop:

$$4i_1 - 2i_2 + 3i_5 = 0$$

KVL to right lower loop:

$$2i_3 - 4i_4 - 3i_5 = 0$$

In a matrix form:

$$\begin{bmatrix} 1 & 0 & -1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 \\ 4 & 0 & 2 & 0 & 0 \\ 4 & -2 & 0 & 0 & 3 \\ 0 & 0 & 2 & -4 & -3 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

## More Definitions

**Sub-matrix:** A sub-matrix of  $\mathbf{A}$  is a matrix obtained from  $\mathbf{A}$  by deleting some selected rows and columns.

**Example:** Given a matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix} \text{ are submatrices of } \mathbf{A}$$

**Square matrix:** An  $m \times n$   $\mathbf{A}$  is called a square matrix if  $m = n$ .

**Example:** The above matrix  $\mathbf{A}$  and  $\mathbf{A}_1$  are a square one while  $\mathbf{A}_2$  is not.

**Diagonal Elements:** All those elements of a matrix  $\mathbf{A}$ ,  $a_{ij}$  with  $i = j$  are called main diagonal elements.

**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 2 \\ 7 & 8 & 9 & 3 \end{bmatrix}$$

Diagonal elements of  $\mathbf{A}$  are 1, 5 and 9.

**Lower and Upper Triangular Matrices:** A matrix is called a **lower triangular** matrix if all the elements above the main diagonal are zero. Similarly, a matrix is called an **upper triangular** matrix if all the elements below the main diagonal are zero.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 4 & 5 & 0 & 0 \\ 7 & 8 & 9 & 0 \end{bmatrix}$$

is a lower triangular matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 & 8 \\ 0 & 5 & 9 & 7 \\ 0 & 0 & 9 & 3 \end{bmatrix}$$

is an upper triangular matrix

16



**Diagonal Matrix:** A matrix that is both a lower and an upper triangular is called a **diagonal matrix**.

**Examples:**

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 9 & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$$

**Identity Matrix:** An identity matrix that is a square diagonal matrix with all its diagonal elements equal to 1. It is denoted by **I**.

**Examples:**

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

**Transpose of a Matrix:** Given an  $m \times n$  matrix  $\mathbf{A}$ , the  $n \times m$  matrix obtained by listing the  $i$ -th row of  $\mathbf{A}$  as its  $i$ -th column for all  $i = 1, 2, \dots, m$ , is called the transpose of  $\mathbf{A}$  and is denoted by  $\mathbf{A}^T$

Examples:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1n} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{m1} & \mathbf{a}_{m2} & \cdots & \mathbf{a}_{mn} \end{bmatrix} \quad \mathbf{A}^T = \begin{bmatrix} \mathbf{a}_{11} & \mathbf{a}_{21} & \cdots & \mathbf{a}_{m1} \\ \mathbf{a}_{12} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{1n} & \mathbf{a}_{2n} & \cdots & \mathbf{a}_{nm} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 2 \\ 7 & 8 & 9 & 3 \end{bmatrix} \quad \mathbf{A}^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

**Symmetric Matrix:** A matrix  $\mathbf{A}$  is said to be symmetric if  $\mathbf{A} = \mathbf{A}^T$ . Note that a symmetric matrix must be square.

Examples:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 3 & 6 & 9 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

**Skew-symmetric Matrix:** A matrix  $\mathbf{A}$  is said to be skew-symmetric if  $\mathbf{A} = -\mathbf{A}^T$ . Note that a skew-symmetric matrix must be square as well, and all its **diagonal elements** must be equal to **zero**.

Examples:

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 6 \\ -3 & -6 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 & -c \\ c & 0 \end{bmatrix}$$

## Matrix Operations:

**Equality of Matrices:** Two matrices **A** and **B** are equal if they have the same size and are equal element by element, i.e., they are identical.

**Addition of Matrices:** Two matrices can be added only if they have the same size. The sum of two matrices is performed by taking the sum of the corresponding elements.

### Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 2 \\ 7 & 8 & 9 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 & 2 & 7 & 7 \\ 1 & 1 & 1 & 5 \\ 7 & 3 & 5 & 2 \end{bmatrix} \quad \mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 4 & 10 & 8 \\ 5 & 6 & 7 & 7 \\ 14 & 11 & 14 & 5 \end{bmatrix}$$

**Scalar Multiplication:** The product of a matrix  $\mathbf{A}$  with a scalar  $c$  is a matrix whose  $(i,j)$ -th element is  $c a_{ij}$ .

**Examples:**

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 7 & 7 \\ 1 & 1 & 1 & 5 \\ 7 & 3 & 5 & 2 \end{bmatrix} \quad \Rightarrow \quad 5\mathbf{A} = \begin{bmatrix} 0 & 10 & 35 & 35 \\ 5 & 5 & 5 & 25 \\ 35 & 15 & 25 & 10 \end{bmatrix}$$

$$\mathbf{A} + (-1)\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$\mathbf{A} + (-1)\mathbf{A} = \mathbf{0}$ , a zero matrix, is true in general for any matrix  $\mathbf{A}$ .

## Some Properties of Matrix Additions and Scalar Multiplications:

1. Commutative law of addition:  $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
2. Associative law of addition:  $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
3.  $\mathbf{A} + \mathbf{0} = \mathbf{A}$ , where  $\mathbf{0}$  is a zero matrix with a same size as  $\mathbf{A}$  has.
4.  $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$
5.  $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$ , where  $c$  is a scalar
6.  $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$ , where  $c$  and  $d$  are scalars
7.  $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
8.  $(c\mathbf{A})^T = c\mathbf{A}^T$ , where  $c$  is a scalar

Inner Product of Two Vectors: Given two vectors  $\mathbf{a}$  and  $\mathbf{b}$  with same length of  $n$ , the inner product is defined as

$$\mathbf{a} \bullet \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Multiplication of Matrices: Given an  $m \times n$  matrix  $\mathbf{A}$  and an  $n \times p$  matrix  $\mathbf{B}$ ,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{bmatrix}, \quad \text{where}$$

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj} = a_{i1} b_{1j} + \cdots + a_{in} b_{nj},$$

$$i = 1, \cdots, m; \quad j = 1, \cdots, p$$

Example:

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 7 & 7 \\ 1 & 1 & 1 & 5 \\ 7 & 3 & 5 & 2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix}$$

$$\begin{aligned} \mathbf{C} = \mathbf{AB} &= \begin{bmatrix} 0 \cdot 1 + 2 \cdot 3 + 7 \cdot 5 + 7 \cdot 7 & 0 \cdot 2 + 2 \cdot 4 + 7 \cdot 6 + 7 \cdot 8 \\ 1 \cdot 1 + 1 \cdot 3 + 1 \cdot 5 + 5 \cdot 7 & 1 \cdot 2 + 1 \cdot 4 + 1 \cdot 6 + 5 \cdot 8 \\ 7 \cdot 1 + 3 \cdot 3 + 5 \cdot 5 + 2 \cdot 7 & 7 \cdot 2 + 3 \cdot 4 + 5 \cdot 6 + 2 \cdot 8 \end{bmatrix} \\ &= \begin{bmatrix} 90 & 106 \\ 44 & 52 \\ 55 & 72 \end{bmatrix} \end{aligned}$$

Note that  $\mathbf{BA}$  is not defined for the above matrices. Thus, in general,

$$\mathbf{AB} \neq \mathbf{BA}$$



Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 5 & 6 \\ 7 & 10 \end{bmatrix} \neq \mathbf{BA} = \begin{bmatrix} 7 & 6 \\ 8 & 8 \end{bmatrix}$$

Example:

$$\mathbf{A} = \begin{bmatrix} 0 & 2 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \neq \mathbf{BA} = \begin{bmatrix} 0 & 6 \\ 0 & 0 \end{bmatrix}$$

Note that the product of two matrices can be equal to a zero matrix even though none of them is a zero matrix.

## Properties of Matrix Multiplication

1.  $(c A) B = A (c B)$

2.  $A (B C) = (A B) C$

3.  $(A + B) C = A C + B C$

4.  $C (A + B) = C A + C B$

5.  $(A B)^T = B^T A^T$

**Solutions to Linear Systems:** Let us recall the following linear system,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$\mathbf{A} \mathbf{x} = \mathbf{b}$$

If data vector  $\mathbf{b} = \mathbf{0}$ , then the above linear system, i.e.,  $\mathbf{A}\mathbf{x} = \mathbf{0}$ , is called a **homogeneous system**. It is easy to see that  $\mathbf{x} = \mathbf{0}$  is a solution to the homogeneous system. Such a solution in fact is called a **trivial solution**. Any non-zero solution of the homogeneous system is called a **non-trivial solution**.

## Augmented Matrix of Linear System:

$$\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix}$$

The above augmented matrix contains all information about the given system.

**Example:** Given a linear system

$$\begin{bmatrix} 0 & 2 & 7 & 7 \\ 1 & 1 & 1 & 5 \\ 7 & 3 & 5 & 2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \Rightarrow \quad \tilde{\mathbf{A}} = \begin{bmatrix} 0 & 2 & 7 & 7 & 1 \\ 1 & 1 & 1 & 5 & 2 \\ 7 & 3 & 5 & 2 & 3 \end{bmatrix}$$

## A Basic Approach for Solving a Linear System:

Step 1. Use the first equation to eliminate  $x_1$  from all the other equations, i.e., make the coefficients of  $x_1$  equal to 0 in all equations except the first one.

Step 2. In the resulting system, use the second equation to eliminate  $x_2$  from all equations that follow it, i.e., make the coefficients of  $x_2$  equal to 0 in the 3rd, 4th, ... equations.

Similarly, carry out steps 3, 4, ..., and so on, each time eliminating one variable from all the equations below the equation being used. At the end of these steps, we will have: the 1st equation contains all unknowns; 2nd equation contains  $x_2, x_3, \dots, x_n$ ; and finally the last equation contains only  $x_n$ .

Then, one can solve for  $x_n$ , and then  $x_{n-1}$ , and so on.

## Elementary Operations for Equations

There are three elementary operations:

1. Interchange of two equations
2. Multiplication of an equation by a non-zero constant
3. Addition of a multiple of one equation to another equation.

If any of these three operations is performed on a linear system, we obtain a new system (and hence a new augmented matrix), but the overall solution  $\mathbf{x}$  does not change. If  $\mathbf{x}$  is a solution to the linear system, then  $\mathbf{x}$  is also a solution to the linear system after any of these three operations is performed. Thus, we may perform any of these operations in order to compute the solution.

## Elementary Row Operations (ERO) for Augmented Matrix

Since each equation in linear system is represented by a row in the augmented matrix, corresponding to the elementary operations for equations, we have the following three elementary row operations for the augmented matrix:

ERO 1. Interchange of two rows

ERO 2. Multiplication of a row by a non-zero constant

ERO 3. Addition of a multiple of one row to another row.

Two augmented matrices are said to be **equivalent** if one can be obtained from another by a series of elementary row operations. It is clear that if two augmented matrices are equivalent, then their corresponding linear systems have the same solution. Thus, the basic idea to use these EROs to simplify the augmented matrix of a linear system in order to obtain its solution.

Example: Let the linear system be

$$\begin{aligned} -x_1 + x_2 + 2x_3 &= 2 \\ 3x_1 - x_2 + x_3 &= 6 \\ -x_1 + 3x_2 + 4x_3 &= 4 \end{aligned} \quad \Rightarrow \quad \tilde{\mathbf{A}} = \begin{bmatrix} -1 & 1 & 2 & 2 \\ 3 & -1 & 1 & 6 \\ -1 & 3 & 4 & 4 \end{bmatrix}$$

If the elementary row operation is to replace the 2nd row by **2nd row + 3 × 1st row**, then the augmented matrix is simplified to

$$\begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ -1 & 3 & 4 & 4 \end{bmatrix} \begin{array}{l} \text{3rd row} - \text{1st row} \\ \\ \end{array} \Rightarrow \begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ 0 & 2 & 2 & 2 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

$-x_1 + (-1) + 2 \cdot 2 = 2 \Rightarrow x_1 = 1$   
 $2x_2 + 7 \cdot 2 = 12 \Rightarrow x_2 = -1$   
 $x_3 = 2$



## Leading Entry in a Vector:

The first non-zero element in a vector is called its **leading entry**. A vector with all its elements equal to zero is said to have no leading entry.

**Example:** In the following matrix

$$\tilde{\mathbf{A}} = \begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The first row has a leading entry equal to  $-1$ . The 2nd row has a leading entry equal to  $2$  and the last one has no leading entry.

## Echelon Form of a Matrix:

A matrix is said to be in an echelon form if it satisfies the following:

- If there is a row containing all zeros, then it is put below the non-zero rows.
- The leading entry of a row is to the right of the leading entry of any other row above it.

**Example:** Consider the following matrices

$$\begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

in an echelon form

$$\begin{bmatrix} -1 & 1 & 2 & 2 \\ 1 & 0 & 7 & 12 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

not in an echelon form

$$\begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 7 & 12 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

The variables corresponding to the leading entries are called **leading variables**. All other variables are called **free variables**. Given the augmented matrix of a linear system, we will use elementary row operations to reduce it to its echelon form and then compute its solution.

**Example:** Consider the echelon form

$$\begin{bmatrix} -1 & 1 & 2 & 2 \\ 0 & 2 & 7 & 12 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

All the three variables are leading variables and the solution is  $x_3 = 2$ ,  $x_2 = -1$ , and  $x_1 = 1$ . Note the solution is unique (since all the variables are leading variables and there are no free variables).

**Example:** Consider the echelon form

$$\begin{bmatrix} 1 & 1 & 7 & 12 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In this case,  $x_1$  and  $x_3$  are leading variables and  $x_2$  is a free variable. The solution is written as  $x_3 = 1$  and  $x_1 = -x_2 + 5$ .  $x_2$  being a free variable can take any value and all such values are admissible as solutions. Thus, in this case we have an infinite number of solutions.

**Example:** Consider the echelon form

$$\begin{bmatrix} 1 & 1 & 7 & 12 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 9 \end{bmatrix}$$

In this case,  $x_1$  and  $x_2$  are leading variables and  $x_3$  is a free variable. There is no solution in this case as the last equation implies  $0x_3 = 9$ .

## Row Echelon Form of a Matrix

A matrix is said to be in a **row echelon form** if

- it is in an echelon form, and
- the leading entries are all equal to 1

Example: Consider the echelon form

$$\begin{bmatrix} 1 & 0 & 7 & 12 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & -5 & -10 \end{bmatrix}$$

Dividing the last row by  $-5$ , we get the row echelon form as

$$\begin{bmatrix} 1 & 0 & 7 & 12 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

## Reduced Row Echelon Form of a Matrix

A matrix is said to be in a **reduced row echelon form** if

- it is in a row echelon form, and
- each column that contains a leading entry 1 has zeros everywhere else except those coefficients of free variables.

Examples: Consider the row echelon forms

$$\begin{bmatrix} 1 & 0 & 7 & 12 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 7 & 12 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix}$$

$$\begin{bmatrix} 1 & 1 & 7 & 12 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & 0 & -2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -2 - a \\ a \\ 2 \end{pmatrix}$$

# Vector Algebra

## Linear Combination of a Set of Vectors

Given a set of  $n$  vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , which have a same length of  $m$ , a linear combination of these vectors is defined as the vector

$$\mathbf{V} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_n \mathbf{v}_n$$

Here  $c_1, c_2, \dots, c_n$  are scalars.

**Example:** Consider a set of vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}$$

Can we express the following vectors

$$\mathbf{u} = \begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 4 \\ -1 \\ 8 \end{pmatrix}$$

as a linear combinations of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ ?

For the first vector  $\mathbf{u}$  to be a linear combination of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ , we must have

$$\begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + c_2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix} \Rightarrow \begin{bmatrix} 1 & 6 \\ 2 & 4 \\ -1 & 2 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix}$$

which is a linear system. The augmented matrix is

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1 & 6 & 9 \\ 2 & 4 & 2 \\ -1 & 2 & 7 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 9 \\ 0 & -8 & -16 \\ 0 & 8 & 16 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 9 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus,

$$\begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + c_2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix} = -3 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + 2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}$$



For the second vector  $\mathbf{w}$  to be a linear combination of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ , we must have

$$\begin{pmatrix} 4 \\ -1 \\ 8 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + c_2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix} \Rightarrow \begin{bmatrix} 1 & 6 \\ 2 & 4 \\ -1 & 2 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ -1 \\ 8 \end{pmatrix}$$

which is a linear system. The augmented matrix is

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1 & 6 & 4 \\ 2 & 4 & -1 \\ -1 & 2 & 8 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 4 \\ 0 & -8 & -9 \\ 0 & 8 & 12 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 4 \\ 0 & -8 & -9 \\ 0 & 0 & 3 \end{bmatrix}$$

Thus,  $\mathbf{w}$  **cannot** be expressed as a linear combination of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ .

## Linear Dependence

Consider a set of  $n$  vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , which have a same length of  $m$ . They are said to be **linearly dependent** if there exist  $n$  scalars,  $c_1, c_2, \dots, c_n$ , not all zeros such that

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_n \mathbf{v}_n = \mathbf{0}$$

It is clear that  $c_1 = c_2 = \dots = c_n = \mathbf{0}$  satisfy the above equation trivially. Thus, for linear dependence, we require at least one of them to be non-zero. If say  $c_1 \neq \mathbf{0}$ , then we can write  $\mathbf{v}_1 = \left( -\frac{c_2}{c_1} \right) \mathbf{v}_2 + \dots + \left( -\frac{c_n}{c_1} \right) \mathbf{v}_n$

## Linear Independence

Consider a set of  $n$  vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , which have a same length of  $m$ . They are said to be **linearly independent** if

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_n \mathbf{v}_n = \mathbf{0}$$

has only a trivial solution  $c_1 = c_2 = \dots = c_n = \mathbf{0}$ .

## How to Check for Linear Dependence/Independence?

Let

$$\mathbf{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{m2} \end{pmatrix}, \quad \dots, \quad \mathbf{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{mn} \end{pmatrix}$$

Then

$$c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix} c_1 + \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{m2} \end{pmatrix} c_2 + \dots + \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{mn} \end{pmatrix} c_n = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1} & v_{m2} & \dots & v_{mn} \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \mathbf{0}$$

which is an  $m \times n$  **homogeneous system** and it has a trivial solution. If this **trivial solution** is the only solution, then the given vectors are **linearly independent**. If there are **non-trivial solutions**, then the vectors are **linearly dependent**.

**Example:** Determine whether the vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 5 \\ 6 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

are linearly independent or not.

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} c_1 + \begin{pmatrix} 5 \\ 6 \\ -1 \end{pmatrix} c_2 + \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} c_3 = \begin{bmatrix} 1 & 5 & 3 \\ -2 & 6 & 2 \\ 3 & -1 & 1 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The augmented matrix is

$$\begin{aligned} \tilde{\mathbf{A}} &= \begin{bmatrix} 1 & 5 & 3 & 0 \\ -2 & 6 & 2 & 0 \\ 3 & -1 & 1 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 5 & 3 & 0 \\ 0 & 16 & 8 & 0 \\ 0 & -16 & -8 & 0 \end{bmatrix} \\ &\Rightarrow \begin{bmatrix} 1 & 5 & 3 & 0 \\ 0 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0.5 & 0 \\ 0 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

$c_3$  is a free variable

$$c_1 = c_2 = -0.5c_3$$

There are non-trivial solutions. Hence, they are linearly dependent.

## Vector Space

Given a set of  $n$  vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , which have a same length of  $m$ , the set  $V$  containing  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  and their linear combinations is said to form a **vector space**.

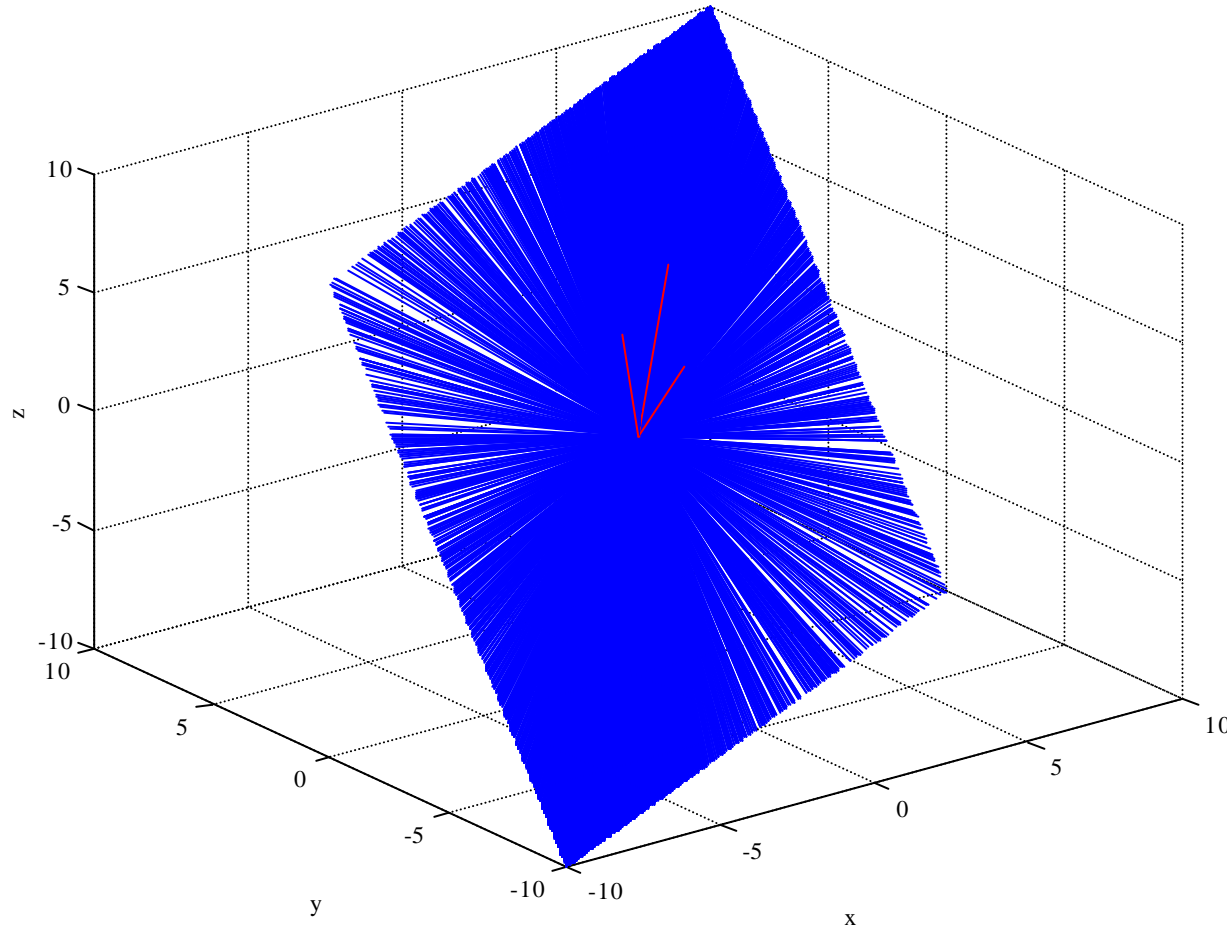
The vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are said to span the vector space  $V$ .

In essence a vector space is a collection of vectors that satisfy the following **two properties**:

1. **Addition:** If  $\mathbf{u}$  and  $\mathbf{w}$  belong to  $V$  then so does  $\mathbf{u} + \mathbf{w}$ .
2. **Scalar Multiplication:** If  $\mathbf{u}$  belongs to  $V$ , then so does  $k\mathbf{u}$  for all arbitrary scalars  $k$ .

It is clear that  $\mathbf{0}$  is a vector in all  $V$  ( $k = 0$ ).

**Example:** Picture a vector space spanned by  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix} = \mathbf{v}_1 + \mathbf{v}_2$



**DEMO**

## Vector Subspace

A vector space  $W$  is said to be a **vector subspace** of a vector space  $V$  if all the vectors in  $W$  are also contained in  $V$ .

**Example:** Consider all possible solutions to the homogeneous system  $A\mathbf{x} = \mathbf{0}$ . If  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are two solutions then so is  $\mathbf{x}_1 + \mathbf{x}_2$ , as

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2 = \mathbf{0} + \mathbf{0} = \mathbf{0}$$

Similarly, if  $\mathbf{x}$  is a solution, then so is  $k\mathbf{x}$ . As a result, all the possible solutions to  $A\mathbf{x} = \mathbf{0}$  constitute a vector space called the **solution space**.

It is standard notation to write  $\hat{A}^n$  to denote the real  $n$ -dimensional vector space, that is, all the vectors of length  $n$  having real numbers as their components. It is clear  $\hat{A}$  or  $\hat{A}^1$  denotes the real numbers,  $\hat{A}^2$  the 2-D plane and  $\hat{A}^3$  the 3-D space and in general  $\hat{A}^n$ , the  $n$ -dimensional space.

Example: Let

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$$

Determine whether these vectors span  $\hat{\mathbf{A}}^3$ . In other words, **determine if every vector in  $\hat{\mathbf{A}}^3$ , say  $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$  can be expressed as a linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  or not.** Let

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} c_1 + \begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix} c_2 + \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} c_3 = \begin{bmatrix} 1 & -2 & 1 \\ -1 & 3 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

The augmented matrix

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1 & -2 & 1 & x_1 \\ -1 & 3 & 2 & x_2 \\ 0 & 1 & 4 & x_3 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -2 & 1 & x_1 \\ 0 & 1 & 3 & x_1 + x_2 \\ 0 & 0 & 1 & x_3 - x_1 - x_2 \end{bmatrix}$$

A solution exists

and hence the

vectors span  $\hat{\mathbf{A}}^3$ .



## Dimension and Basis of a Vector Space

We have described a vector space as the set of vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  called the spanning vectors and all their linear combinations. Therefore, any arbitrary vector, say  $\mathbf{v}$ , in  $V$  can be expressed as

$$\mathbf{v} = c_1 \mathbf{v}_1 + \dots + c_{n-1} \mathbf{v}_{n-1} + c_n \mathbf{v}_n$$

In addition, if  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are linearly dependent, we can simplify the above equation as follows. If  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are linearly dependent, then we have

$$\mathbf{a}_1 \mathbf{v}_1 + \dots + \mathbf{a}_{n-1} \mathbf{v}_{n-1} + \mathbf{a}_n \mathbf{v}_n = \mathbf{0}$$

with at least one of the coefficients is non-zero. For simplicity, we let  $\mathbf{a}_n \neq 0$ .

Then

$$\mathbf{v}_n = \left( -\frac{\mathbf{a}_1}{\mathbf{a}_n} \right) \mathbf{v}_1 + \dots + \left( -\frac{\mathbf{a}_{n-1}}{\mathbf{a}_n} \right) \mathbf{v}_{n-1}$$

and  $\mathbf{v} = \left( c_1 - \frac{c_n \mathbf{a}_1}{\mathbf{a}_n} \right) \mathbf{v}_1 + \dots + \left( c_{n-1} - \frac{c_n \mathbf{a}_{n-1}}{\mathbf{a}_n} \right) \mathbf{v}_{n-1} = \mathbf{b}_1 \mathbf{v}_1 + \dots + \mathbf{b}_{n-1} \mathbf{v}_{n-1}$  49

It is clear that the same vector space  $V$  can now be expressed as vectors  $\mathbf{v}_1, \dots, \mathbf{v}_{n-1}$  and all their possible linear combinations.

Now, suppose  $\mathbf{v}_1, \dots, \mathbf{v}_{n-1}$  are again linearly dependent. Then we have

$$a_1 \mathbf{v}_1 + \dots + a_{n-1} \mathbf{v}_{n-1} = \mathbf{0}$$

with at least one of the coefficients is non-zero. For simplicity, we let  $a_{n-1} \neq 0$ .

Then

$$\mathbf{v}_{n-1} = \left( -\frac{a_1}{a_{n-1}} \right) \mathbf{v}_1 + \dots + \left( -\frac{a_{n-2}}{a_{n-1}} \right) \mathbf{v}_{n-2}$$



$$\mathbf{v} = \left( \mathbf{b}_1 - \frac{\mathbf{b}_{n-1} a_1}{a_{n-1}} \right) \mathbf{v}_1 + \dots + \left( \mathbf{b}_{n-2} - \frac{\mathbf{b}_{n-1} a_{n-2}}{a_{n-1}} \right) \mathbf{v}_{n-2} = \mathbf{g}_1 \mathbf{v}_1 + \dots + \mathbf{g}_{n-2} \mathbf{v}_{n-2}$$



Using the same approach, we can get rid of all those vectors in the description of  $V$  that are linearly **dependent**. In the end, we are left only linearly independent vectors, say,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ , and now  $V$  can be expressed as  $\mathbf{v}_1, \dots, \mathbf{v}_d$  and all their possible linear combinations. We cannot reduce this set any more as it is impossible to find a non-zero scalar in the following equation:

$$c_1 \mathbf{v}_1 + \dots + c_d \mathbf{v}_d = \mathbf{0}$$

Thus, we have that

1.  $\mathbf{v}_1, \dots, \mathbf{v}_d$  are linearly independent, and
2.  $\mathbf{v}_1, \dots, \mathbf{v}_d$  span  $V$ .

Such vectors are said to constitute a basis for the vector space  $V$ . The number  $d$ , the largest number of linearly independent vectors in  $V$ , is called the dimension of  $V$ . Note that the basis is non-unique but  $d$  is fixed for a given  $V$ .

**Example:** Determine a basis and the dimension of the solution space of the following homogeneous system,

$$2x_1 + 2x_2 - 3x_3 + 0x_4 + x_5 = 0$$

$$-x_1 - x_2 + 2x_3 - 3x_4 + x_5 = 0$$

$$x_1 + x_2 - 2x_3 + 0x_4 - x_5 = 0$$

$$0x_1 + 0x_2 + x_3 + x_4 + x_5 = 0$$

The augmented matrix is

$$\tilde{\mathbf{A}} = \begin{bmatrix} 2 & 2 & -3 & 0 & 1 & 0 \\ -1 & -1 & 2 & -3 & 1 & 0 \\ 1 & 1 & -2 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 2 & -3 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & -3 & 1.5 & 0 \\ 0 & 0 & 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 \end{bmatrix}$$

Thus,  $x_5$ ,  $x_4$ ,  $x_3$  and  $x_1$  are leading variables and  $x_5 = x_4 = x_3 = 0$ .  $x_2$  is a free variable. Let  $x_2 = a$ . Then we have  $x_1 = -x_2 = -a$  all possible vectors look like

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -a \\ a \\ 0 \\ 0 \\ 0 \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ is a basis vector and } d = 1.$$

## Row Space of a Matrix

Given an  $m \times n$  matrix  $\mathbf{A}$ , the vector space spanned by the rows of  $\mathbf{A}$  is called the row space of  $\mathbf{A}$ .

Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 3 \\ 2 & 6 & 2 \end{bmatrix} \Rightarrow \text{Row Space of } \mathbf{A} = \{c_1[1 \ 5 \ 3] + c_2[2 \ 6 \ 2]\} \\ = \{[c_1 + 2c_2 \quad 5c_1 + 6c_2 \quad 3c_1 + 2c_2]\}^{53}$$

## Column Space of a Matrix

Given an  $m \times n$  matrix  $\mathbf{A}$ , the vector space spanned by the columns of  $\mathbf{A}$  is called the column space of  $\mathbf{A}$ .

Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 3 \\ 2 & 6 & 2 \end{bmatrix} \Rightarrow \text{Column Space of } \mathbf{A} = \left\{ c_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} 5 \\ 6 \end{bmatrix} + c_3 \begin{bmatrix} 3 \\ 2 \end{bmatrix} \right\}$$
$$= \left\{ \begin{bmatrix} c_1 + 5c_2 + 3c_3 \\ 2c_1 + 6c_2 + 2c_3 \end{bmatrix} \right\}$$

**Remarks:** Row space is a set of row vectors and column space is a set of column vectors. Thus, in general the row space and column space of a given matrix is different.

## Null Space of a Matrix

Given an  $m \times n$  matrix  $\mathbf{A}$ , the vector space consists of all possible solutions of  $\mathbf{Ax} = \mathbf{0}$  is called the null space of  $\mathbf{A}$ .

**Example:** Consider

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 1 & 5 & 3 \\ 2 & 6 & 2 \end{bmatrix} \Rightarrow \tilde{\mathbf{A}} = \begin{bmatrix} 1 & 5 & 3 & 0 \\ 2 & 6 & 2 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 5 & 3 & 0 \\ 0 & -4 & -4 & 0 \end{bmatrix} \\ \Rightarrow \begin{bmatrix} 1 & 5 & 3 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} &\Rightarrow \begin{bmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \Rightarrow x_2 = -x_3 \ \& \ x_1 = 2x_3 \end{aligned}$$

$$\Rightarrow \mathbf{x} = \begin{pmatrix} 2x_3 \\ -x_3 \\ x_3 \end{pmatrix} = x_3 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \Rightarrow \text{Null Space of } \mathbf{A} = \left\{ x_3 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, x_3 \text{ is free} \right\}$$

Theorem:

The row elementary operations (EROs) **do not** change the row space and null space of a given matrix. It **might** change the column space of a matrix.

### Rank and Nullity of a Matrix

The maximum number of linearly independent vectors of a matrix  $\mathbf{A}$ , which is also the dimension of the row space of  $\mathbf{A}$ , is called the row rank or simply the rank of  $\mathbf{A}$ .

Theorem:

The dimension of the row space of  $\mathbf{A}$  = rank of  $\mathbf{A}$  = the number of non-zero rows in the echelon form of  $\mathbf{A}$ . In fact, the non-zero rows of the echelon form of  $\mathbf{A}$  form a basis of the row space of  $\mathbf{A}$ .



**Example:** Determine the rank and a basis of the row space of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 6 & 9 \\ 2 & 4 & 2 \\ -1 & 2 & 7 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 9 \\ 0 & -8 & -16 \\ 0 & 8 & 16 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 9 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

Then rank of  $\mathbf{A}$  equal to 2. The row space of  $\mathbf{A}$  are spanned by

$$\mathbf{v}_1 = [1 \ 6 \ 9], \quad \mathbf{v}_2 = [0 \ 1 \ 2]$$

### Procedure for Determining Basis and Dimension of a Vector Space $V$

Given a vector space spanned by a set of column vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , which are not necessarily linearly independent, the following procedure can be used to obtain a basis for it as well as its dimension:

Arrange

$$\mathbf{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{m2} \end{pmatrix}, \quad \dots, \quad \mathbf{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{mn} \end{pmatrix}$$

to form a matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{m1} \\ v_{12} & v_{22} & \cdots & v_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \cdots & v_{mn} \end{bmatrix}$$

Then, Dimension of  $\mathbf{V}$  = dimension of row space of  $\mathbf{A}$  = rank of  $\mathbf{A}$ .

Basis vectors of  $\mathbf{V}$  = transpose of non-zero rows in the echelon form of  $\mathbf{A}$ .

If dimension of  $\mathbf{V} = n$ , then  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are linearly independent.

**Example:** Find the dimension and a set of basis vectors for the vector space

$V$  spanned by

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$$

For a matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ -2 & 3 & 1 \\ -1 & 2 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

The dimension of  $V$  is equal to 2 (the rank of  $\mathbf{A}$  is equal 2) and it has a basis

$$\mathbf{a}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

Suppose now a given vector space  $V$  is spanned by a set of row vectors

$$\begin{aligned}\mathbf{a}_1 &= (a_{11} \quad a_{12} \quad \cdots \quad a_{1m}) \\ \mathbf{a}_2 &= (a_{21} \quad a_{22} \quad \cdots \quad a_{2m}) \\ &\vdots \\ \mathbf{a}_n &= (a_{n1} \quad a_{n2} \quad \cdots \quad a_{nm})\end{aligned}$$

Form a matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

Then, the vector space  $V$  = the row space of  $\mathbf{A}$ .

Dimension of  $V$  = dimension of row space of  $\mathbf{A}$  = rank of  $\mathbf{A}$ .

Basis vectors of  $V$  = non-zero rows in the echelon form of  $\mathbf{A}$ .

**Example:** Find the dimension and a set of basis vectors for the vector space

$V$  spanned by

$$\mathbf{a}_1 = (1 \ 2 \ 0 \ -1)$$

$$\mathbf{a}_2 = (2 \ 6 \ -3 \ -3)$$

$$\mathbf{a}_3 = (3 \ 10 \ -6 \ -6)$$

For a matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 6 & -3 & -3 \\ 3 & 10 & -6 & -6 \end{bmatrix} \xrightarrow{\text{ERO}} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & -3 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The dimension of  $V$  is equal to the rank of  $\mathbf{A}$ , which is equal 2 and  $V$  has a basis

$$\mathbf{u}_1 = (1 \ 2 \ 0 \ -1)$$

$$\mathbf{u}_2 = (0 \ 2 \ -3 \ -1)$$

## Nullity of a Matrix

Nullity of a matrix  $\mathbf{A}$  is defined as the dimension of the null space of  $\mathbf{A}$ .

Example: Find the nullity of

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 6 & -3 & -3 \\ 3 & 10 & -6 & -5 \end{bmatrix} \xrightarrow{\text{ERO}} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & -3 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Clearly,  $x_3$  and  $x_4$  are free variables and

$$2x_2 = 3x_3 + x_4 \Rightarrow x_2 = 1.5x_3 + 0.5x_4, \quad x_1 = -2x_2 + x_4 = -3x_3 - x_4 + x_4 = -3x_3$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -3x_3 \\ 1.5x_3 + 0.5x_4 \\ x_3 \\ x_4 \end{pmatrix} = x_3 \begin{pmatrix} -3 \\ 1.5 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 0.5 \\ 0 \\ 1 \end{pmatrix}$$

basis vectors  
for null space

Nullity = 2

Theorem:

Numbers of linearly independent rows of  $\mathbf{A}$  = numbers of linearly independent columns of  $\mathbf{A}$ . In other words, row rank of  $\mathbf{A}$  = column rank of  $\mathbf{A}$  = rank of  $\mathbf{A}$ .

Theorem:

For an  $m \times n$  matrix  $\mathbf{A}$ ,

$$\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = n \quad \text{or} \quad \text{nullity}(\mathbf{A}) = n - \text{rank}(\mathbf{A})$$

Example: Consider the previous example,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & 6 & -3 & -3 \\ 3 & 10 & -6 & -5 \end{bmatrix} \xrightarrow{\text{ERO}} \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & -3 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$n = 4$ , rank  $(\mathbf{A}) = 2$  and nullity  $(\mathbf{A}) = 2$ . It is verified.

## Existence and Uniqueness of Solutions of Linear Systems

Consider a linear

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

The above linear system has a solution if and only if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \text{ and } \tilde{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix}$$

has the same rank. Thus, if the above two matrices has different ranks, no solution exists for the given system.



If the rank of  $\mathbf{A}$  and the rank of the augmented matrix have the same rank, say  $k$ , and if  $k = n$ , the number of unknowns, then the linear system has a **unique solution** (exactly one solution).

If  $k < n$ , the given system has infinitely many solutions. In this case,  $k$  unknowns (leading variables) can be determined in terms of  $n - k$  unknowns (free variables) that can be assigned arbitrary values.

## Homogeneous System

The homogeneous system is a special class of systems with  $\mathbf{b} = \mathbf{0}$ , or  $\mathbf{A} \mathbf{x} = \mathbf{0}$ . Let  $k = \text{rank}(\mathbf{A})$ . Then the homogeneous system has a unique solution if and only if  $k = n$ , the number of unknowns. The solution is given by  $\mathbf{x} = \mathbf{0}$ . If  $k < n$ , the homogeneous system always has infinitely many solutions.

Note that for homogeneous system,  $n \geq k$  (always true).

**Example:** Consider the following linear system

$$2x_1 + 2x_2 - 3x_3 = 1$$

$$-x_1 - x_2 + 2x_3 = 2$$

$$x_1 + x_2 - x_3 = 3$$

The coefficient and the augmented matrices are given by

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & -3 \\ -1 & -1 & 2 \\ 1 & 1 & -1 \end{bmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 2 & 2 & -3 & 1 \\ -1 & -1 & 2 & 2 \\ 1 & 1 & -1 & 3 \end{bmatrix}$$

Using the EROs, the echelon forms are obtained as

$$\begin{bmatrix} 2 & 2 & -3 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 2 & -3 & 1 \\ 0 & 0 & 0.5 & 2.5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

It is clear that above two matrices have the same rank of 2, which is less than the number of unknowns. Hence, the system has **infinitely many solutions**, all which can be obtained by choosing arbitrary values for free variable.

**Example:** Consider the following linear system

$$2x_1 + 2x_2 - 3x_3 = 1$$

$$-x_1 - x_2 + 2x_3 = 2$$

$$x_1 + x_2 - x_3 = 0$$

The coefficient and the augmented matrices are given by

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & -3 \\ -1 & -1 & 2 \\ 1 & 1 & -1 \end{bmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 2 & 2 & -3 & 1 \\ -1 & -1 & 2 & 2 \\ 1 & 1 & -1 & 0 \end{bmatrix}$$

Using the EROs, the echelon forms are obtained as

$$\begin{bmatrix} 2 & 2 & -3 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 2 & -3 & 1 \\ 0 & 0 & 0.5 & 2.5 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

It is clear that above two matrices have ranks of 2 and 3, respectively. Hence, the system has **no solution** at all.

**Example:** Consider the following linear system

$$2x_1 + 2x_2 - 3x_3 = 1$$

$$-x_1 + x_2 + 2x_3 = 2$$

$$x_1 + x_2 + x_3 = 0$$

The coefficient and the augmented matrices are given by

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & -3 \\ -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 2 & 2 & -3 & 1 \\ -1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Using the EROs, the echelon forms are obtained as

$$\begin{bmatrix} 2 & 2 & -3 \\ 0 & 2 & 0.5 \\ 0 & 0 & 2.5 \end{bmatrix}, \quad \begin{bmatrix} 2 & 2 & -3 & 1 \\ 0 & 2 & 0.5 & 2.5 \\ 0 & 0 & 2.5 & -0.5 \end{bmatrix}$$

It is clear that above two matrices have the same rank = 3 = the number of unknowns. Hence, the system has **a unique solution**, and the solution is given by  $x_1 = -1.1, x_2 = 1.3, x_3 = -0.2$ .

## Determinant of a Matrix

Given a square  $n \times n$  matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \begin{bmatrix} a_{22} & \cdots & a_{2n} \end{bmatrix} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \begin{bmatrix} a_{n2} & \cdots & a_{nn} \end{bmatrix} \end{bmatrix}$$

$M_{11}$  is the det of this submatrix.

define  $M_{ij}$  as the determinant of an order  $(n-1) \times (n-1)$  matrix obtained from  $\mathbf{A}$  by deleting its  $i$ -th row and  $j$ -th column. The determinant of  $\mathbf{A}$  is given by

$$\det(\mathbf{A}) = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in}$$

where  $C_{ij} = (-1)^{i+j} M_{ij}$  is called the **co-factor** of  $a_{ij}$  and  $M_{ij}$  is called the **minor** of  $a_{ij}$ .

## Determinant of a 2 x 2 Matrix

Given a 2 x 2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow \det(\mathbf{A}) = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

## Determinant of a 3 x 3 Matrix

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{23}a_{12}$$

$$-a_{13}a_{22}a_{31} - a_{21}a_{12}a_{33} - a_{11}a_{23}a_{32}$$

The diagram shows a 3x3 matrix with elements  $a_{11}, a_{12}, a_{13}$  in the first row,  $a_{21}, a_{22}, a_{23}$  in the second row, and  $a_{31}, a_{32}, a_{33}$  in the third row. Red solid lines connect  $(a_{11}, a_{22}, a_{33})$ ,  $(a_{21}, a_{32}, a_{13})$ , and  $(a_{31}, a_{23}, a_{12})$ , representing the positive terms in the expansion. Blue dashed lines connect  $(a_{13}, a_{22}, a_{31})$ ,  $(a_{21}, a_{12}, a_{33})$ , and  $(a_{11}, a_{23}, a_{32})$ , representing the negative terms.

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

All products of red-lined (solid) carry a positive sign and those of blue-lined (dashed) carry a negative one.

Example: Compute the determinant of

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ -2 & 3 & 1 \\ -1 & 2 & 1 \end{bmatrix}$$

Solution:

$$\begin{aligned} \det(\mathbf{A}) &= \det \begin{bmatrix} 1 & -1 & 1 \\ -2 & 3 & 1 \\ -1 & 2 & 1 \end{bmatrix} \\ &= 1 \cdot 3 \cdot 1 + (-2) \cdot 2 \cdot 1 + (-1) \cdot 1 \cdot (-1) - (-1) \cdot 3 \cdot 1 - 2 \cdot 1 \cdot 1 - 1 \cdot (-1) \cdot (-2) \\ &= -1 \end{aligned}$$

or

$$\begin{aligned} \det(\mathbf{A}) &= 1 \cdot (-1)^{1+1} \det \begin{bmatrix} 3 & 1 \\ 2 & 1 \end{bmatrix} - 2 \cdot (-1)^{2+1} \det \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} - 1 \cdot (-1)^{3+1} \det \begin{bmatrix} -1 & 1 \\ 3 & 1 \end{bmatrix} \\ &= (3 - 2) + 2(-1 - 2) - (-1 - 3) = -1 \end{aligned}$$

## Properties of Determinant

1.  $\det(\mathbf{A}) = \det(\mathbf{A}^T)$
2. If all the elements of a row (or a column) of  $\mathbf{A}$  are multiplied by a scalar  $k$ , then  $\det.$  of the resulting matrix  $= k \det(\mathbf{A})$ .
3. If two rows of  $\mathbf{A}$  are interchanged, then the determinant of the resulting matrix is  $-\det(\mathbf{A})$ .
4. If two rows (or two columns) of  $\mathbf{A}$  are equal, then  $\det(\mathbf{A}) = 0$ .
5. For any  $n \times n$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$ .
6. For an upper or a lower triangular matrix or a diagonal matrix  
**determinant = product of the diagonal elements.**



**Example:** Verify that  $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$  with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

**Solution:**

$$\det(\mathbf{A}) = \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 4 - 6 = -2, \quad \det(\mathbf{B}) = \det \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = 40 - 42 = -2$$

$$\Rightarrow \det(\mathbf{A}) \det(\mathbf{B}) = 4$$

$$\det(\mathbf{AB}) = \det \left( \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \right) = \det \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} = 19 \times 50 - 22 \times 43 = 4$$

**Example:**

$$\det \begin{bmatrix} 1 & 10000 & 10000000 \\ 0 & 2 & 9999 \\ 0 & 0 & 3 \end{bmatrix} = \det \begin{bmatrix} 1 & 0 & 0 \\ 12345 & 2 & 0 \\ 6789 & 987654321 & 3 \end{bmatrix} = 1 \times 2 \times 3 = 6$$

## Inverse of a Matrix

For a  $n \times n$  matrix  $\mathbf{A}$ , if there exists a square matrix  $\mathbf{B}$  such that

$$\mathbf{AB} = \mathbf{I}$$

then  $\mathbf{B}$  is called the inverse of  $\mathbf{A}$  and we write  $\mathbf{B} = \mathbf{A}^{-1}$ . If we let

$$\mathbf{B} = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_n]$$

then we have

$$\mathbf{Ab}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{Ab}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots \quad \mathbf{Ab}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Note that we need to solve  $n$  linear systems. In fact, we can solve them all together by defining a combined augmented matrix:

$$\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Reduce the above augmented matrix to a reduced row echelon form,

$$\Rightarrow \begin{bmatrix} 1 & 0 & \cdots & 0 & b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & 1 & \cdots & 0 & b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} = [\mathbf{I} \quad \mathbf{B}]$$

Then, the inverse of the given matrix  $\mathbf{A}$  is given by matrix  $\mathbf{B}$ .

Note that in order for the inverse of  $\mathbf{A}$  to be existent,

$\mathbf{A}$  must has a rank of  $n$  or determinant of  $\mathbf{A}$  must be non-zero.

Example: Find the inverse of

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ -2 & 3 & 1 \\ -1 & 2 & 1 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} -1 & -3 & 4 \\ -1 & -2 & 3 \\ 1 & 1 & -1 \end{bmatrix}$$

Form an augmented matrix

$$\begin{aligned} \tilde{\mathbf{A}} &= \begin{bmatrix} 1 & -1 & 1 & 1 & 0 & 0 \\ -2 & 3 & 1 & 0 & 1 & 0 \\ -1 & 2 & 1 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 & 0 & 1 \end{bmatrix} \\ &\Rightarrow \begin{bmatrix} 1 & -1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 0 & -1 & -1 & -1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & -1 \end{bmatrix} \\ &\Rightarrow \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & -1 & -2 & 3 \\ 0 & 0 & 1 & 1 & 1 & -1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & -1 & -3 & 4 \\ 0 & 1 & 0 & -1 & -2 & 3 \\ 0 & 0 & 1 & 1 & 1 & -1 \end{bmatrix} \end{aligned}$$

76

## Eigenvalues and Eigenvectors of a Matrix

Given an **square**  $n \times n$  matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

consider the vector equation

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \quad \Rightarrow \quad (\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}, \quad \mathbf{x} \neq \mathbf{0}$$

where  $\lambda$  is a scalar quantity. In other words, we seek all those values of  $\lambda$  and non-zero vectors of  $\mathbf{x}$  such that the above equation is satisfied.

All those values of  $\lambda$  that satisfy the above equation are called the **eigenvalues** of  $\mathbf{A}$  and their corresponding vectors  $\mathbf{x}$  are called **eigenvectors**.

77

## Remarks:

1. Eigenvalues and eigenvectors are only defined for square matrices.
2.  $\mathbf{x} = \mathbf{0}$  is a trivial solution for the eigenvector. We will discard such a solution. In other words, we seek only those  $\mathbf{x}$  that are non-zero.
3. If  $\mathbf{x}$  is eigenvector, then so is  $\mathbf{y} = k \mathbf{x}$  as

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{y} = k(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

Hence, eigenvectors are non-unique.

4. Eigenvalues can be complex valued even for real valued matrices.

## Procedure for Determining Eigenvalues and Eigenvectors

Recall that the eigenvalues and eigenvectors are defined as the solutions to

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

This is a homogeneous system with a coefficient matrix  $\mathbf{A} - \lambda \mathbf{I}$ . Recall that a homogeneous system has a non-zero solution if and only if

$$\text{rank of the coefficient matrix} < \text{number of unknowns} = n$$

Hence, non-zero solution will exist for all those values of  $\lambda$  for which

$$\text{rank}[\mathbf{A} - \lambda \mathbf{I}] < n$$

This is equivalent to saying that non-zero solutions will exist for all those values of  $\lambda$  for which

$$c(\lambda) = \det[\mathbf{A} - \lambda \mathbf{I}] = 0$$

## Characteristic Polynomial

$$c(\lambda) = \det[\mathbf{A} - \lambda \mathbf{I}]$$

which is a  $n$ -th degree polynomial of  $\lambda$ , is called the characteristic polynomial of the matrix  $\mathbf{A}$ . And then the eigenvalues of  $\mathbf{A}$  are given by the roots of this characteristic polynomial. Thus, we can compute the eigenvalues and eigenvectors of a given matrix  $\mathbf{A}$  as follows:

**Step 1:** Compute the characteristic polynomial of  $\mathbf{A}$ .

**Step 2:** Find all  $n$  roots for the polynomial obtained in **Step 1** and label them as  $\lambda_1, \lambda_2, \dots, \lambda_n$ .

**Step 3:** Find corresponding eigenvectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  by solving the homogeneous system

$$(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_i = \mathbf{0}, \quad i = 1, 2, \dots, n$$



Example: Consider a 3 x 3 matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \\ 0 & 6 & 5 \end{bmatrix}$$

Step 1: Find the characteristic polynomial

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) &= \det \begin{bmatrix} 1-\lambda & 2 & 3 \\ 0 & 5-\lambda & 6 \\ 0 & 6 & 5-\lambda \end{bmatrix} = (1-\lambda)[(5-\lambda)^2 - 36] \\ &= (1-\lambda)(-1-\lambda)(11-\lambda) \end{aligned}$$

Step 2: The eigenvalues are  $\lambda_1 = 1$ ,  $\lambda_2 = -1$ , and  $\lambda_3 = 11$ .

Step 3: Compute the corresponding eigenvectors

For  $\lambda_1 = 1$ , we have

$$(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x}_1 = \begin{bmatrix} 1-1 & 2 & 3 \\ 0 & 5-1 & 6 \\ 0 & 6 & 5-1 \end{bmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix} = \begin{bmatrix} 0 & 2 & 3 \\ 0 & 4 & 6 \\ 0 & 6 & 4 \end{bmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix} = \mathbf{0}$$
$$\Rightarrow \mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \text{ as } x_1 \text{ is a free variable.}$$

For  $\lambda_2 = -1$ , we have

$$(\mathbf{A} - \lambda_2 \mathbf{I})\mathbf{x}_2 = \begin{bmatrix} 2 & 2 & 3 \\ 0 & 6 & 6 \\ 0 & 6 & 6 \end{bmatrix} \begin{pmatrix} x_{21} \\ x_{22} \\ x_{23} \end{pmatrix} = \mathbf{0} \Rightarrow \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$$

For  $\lambda_3 = 11$ , we have

$$(\mathbf{A} - \lambda_3 \mathbf{I})\mathbf{x}_3 = \begin{bmatrix} -10 & 2 & 3 \\ 0 & -6 & 6 \\ 0 & 6 & -6 \end{bmatrix} \begin{pmatrix} x_{31} \\ x_{32} \\ x_{33} \end{pmatrix} = \mathbf{0} \Rightarrow \mathbf{x}_3 = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

## Further Properties of Eigenvalues and Eigenvectors

1. Eigenvalues of  $\mathbf{A}$  and  $\mathbf{A}^T$  are the same.

Proof.  $\det(\mathbf{A} - \lambda \mathbf{I}) = \det(\mathbf{A} - \lambda \mathbf{I})^T = \det(\mathbf{A}^T - \lambda \mathbf{I}^T) = \det(\mathbf{A}^T - \lambda \mathbf{I})$

2. If a matrix is real valued, then either all its eigenvalues are real or they occur in complex conjugate pairs.

3.  $c(\lambda) = \det[\mathbf{A} - \lambda \mathbf{I}] = \det(\mathbf{A})$

Also, let  $c(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)$ , we have

$$c(0) = (\lambda_1 - 0)(\lambda_2 - 0) \cdots (\lambda_n - 0) = \lambda_1 \lambda_2 \cdots \lambda_n = \det(\mathbf{A})$$

4. Thus, the inverse of  $\mathbf{A}$  exists if and only if  $\mathbf{A}$  has no zero eigenvalues.

5. Eigenvalues of an upper or lower diagonal matrix, or a diagonal matrix are equal to the diagonal elements of the given matrix.

## Summary of Some Useful Properties Square Matrices:

For an  $n \times n$  square matrix  $A$ , the following statements are equivalent:

$A$  is non-singular.

The inverse of  $A$  exists.

$\det(A)$  is nonzero.

$\text{rank}(A) = n$

$A$  has no eigenvalues at 0.

The following statement are also equivalent:

$A$  is singular.

The inverse of  $A$  does not exist.

$\det(A)$  is zero.

$\text{rank}(A) < n$

$A$  has at least one eigenvalue at 0.

## Eigenvalues and Eigenvectors of an Orthogonal Matrix

An orthogonal matrix is defined as a matrix, say  $\mathbf{A}$ , which satisfies the property

$$\mathbf{A}^{-1} = \mathbf{A}^T \quad \Rightarrow \quad \mathbf{A}\mathbf{A}^T = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Thus, for an orthogonal matrix

$$\det(\mathbf{A}\mathbf{A}^T) = \det(\mathbf{A})\det(\mathbf{A}^T) = \det(\mathbf{A})^2 = \det(\mathbf{I}) = 1 \quad \Rightarrow \quad \det(\mathbf{A}) = \pm 1$$

**Theorem:** The eigenvalues of an orthogonal matrix have a absolute value equal to 1.

**Example:** Verify the following matrix is orthogonal

$$\begin{aligned} \mathbf{A} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} &\Rightarrow \det(\mathbf{A} - \lambda \mathbf{I}) = \det \begin{bmatrix} -\lambda & 0 & 1 \\ 0 & 1-\lambda & 0 \\ -1 & 0 & -\lambda \end{bmatrix} \\ &= 1 - \lambda + \lambda^2 - \lambda^3 \quad \Rightarrow \quad \lambda_1 = 1, \lambda_2 = i, \lambda_3 = -i \end{aligned}$$

## Symmetric Matrices

If  $\mathbf{A}$  is a symmetric matrix, i.e,  $\mathbf{A} = \mathbf{A}^T$ , then

1. All its eigenvalues are real, and
2. Its eigenvectors are real valued and are orthogonal to each other, i.e.,

$$x_i^T x_j = 0, \quad i \neq j$$

**Proof.** Let  $\lambda$  and  $\mathbf{x}$  be the eigenvalue and eigenvector of  $\mathbf{A}$ , i.e.,

$$\mathbf{Ax} = l\mathbf{x} \Rightarrow \mathbf{x}^T \mathbf{A}^T = (\mathbf{Ax})^T = (l\mathbf{x})^T = l^* \mathbf{x}^T \Rightarrow \mathbf{x}^T \mathbf{A}^T \mathbf{x} = \mathbf{x}^T \mathbf{Ax} = l^* \mathbf{x}^T \mathbf{x}$$

$$\Rightarrow \mathbf{Ax} = l\mathbf{x} \Rightarrow \mathbf{x}^T \mathbf{Ax} = l \mathbf{x}^T \mathbf{x} \Rightarrow l \mathbf{x}^T \mathbf{x} = l^* \mathbf{x}^T \mathbf{x}$$

$$\Rightarrow l = l^* \Rightarrow \text{Hence } l \text{ is real}$$

To show that the eigenvectors are orthogonal, we let

$$\mathbf{A}\mathbf{x}_i = l_i\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j = l_j\mathbf{x}_j \quad l_i \neq l_j$$

Thus,

$$\begin{array}{c} \downarrow \\ \mathbf{x}_i^T \mathbf{A} = l_i \mathbf{x}_i^T \end{array}$$

and

$$\begin{aligned} \mathbf{x}_i^T \mathbf{A} \mathbf{x}_j &= l_i \mathbf{x}_i^T \mathbf{x}_j \quad \Rightarrow \quad \mathbf{x}_i^T l_j \mathbf{x}_j = l_i \mathbf{x}_i^T \mathbf{x}_j \\ &\Rightarrow \quad (l_i - l_j) \mathbf{x}_i^T \mathbf{x}_j = 0 \end{aligned}$$



$$\mathbf{x}_i^T \mathbf{x}_j = 0$$

Hence, they are orthogonal.

## Norm of a Vector

The norm of a vector is defined as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} = \left( \begin{bmatrix} x_1^* & x_2^* & \cdots & x_n^* \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right)^{\frac{1}{2}} = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$$

It is clear that norm is non-zero if the vector is non-zero.

## Normalization of Eigenvectors

We normalize the eigenvectors by dividing their respective norms. Suppose  $\mathbf{x}_1, \mathbf{x}_2, \dots$  and  $\mathbf{x}_n$  are the eigenvectors of a symmetric matrix, the normalized eigenvectors are given by

$$\mathbf{e}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}, \quad \mathbf{e}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}, \quad \dots, \quad \mathbf{e}_n = \frac{\mathbf{x}_n}{\|\mathbf{x}_n\|} \quad \Rightarrow \quad \|\mathbf{e}_i\| = 1$$



Let us define an eigenvector matrix for a symmetric matrix

$$\mathbf{P} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n]$$

It is simple to see that

$$\begin{aligned} \mathbf{P}^T \mathbf{P} &= [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n]^T [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n] = \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n] \\ &= \begin{bmatrix} \mathbf{e}_1^T \mathbf{e}_1 & \mathbf{e}_1^T \mathbf{e}_2 & \cdots & \mathbf{e}_1^T \mathbf{e}_n \\ \mathbf{e}_2^T \mathbf{e}_1 & \mathbf{e}_2^T \mathbf{e}_2 & \cdots & \mathbf{e}_2^T \mathbf{e}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_n^T \mathbf{e}_1 & \mathbf{e}_n^T \mathbf{e}_2 & \cdots & \mathbf{e}_n^T \mathbf{e}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I} \end{aligned}$$

Hence  $\mathbf{P}^{-1} = \mathbf{P}^T$   $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} \Rightarrow \sqrt{\mathbf{e}_1^T \cdot \mathbf{e}_1} = \|\mathbf{e}_1\| \Rightarrow \mathbf{e}_1^T \cdot \mathbf{e}_1 = \|\mathbf{e}_1\|^2 = 1$

**Example:** Consider the following symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 3 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

Its characteristic polynomial is

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \det \begin{bmatrix} 4 - \lambda & 0 & 2 \\ 0 & 3 - \lambda & 0 \\ 2 & 0 & 4 - \lambda \end{bmatrix} = 36 - 36\lambda + 11\lambda^2 - \lambda^3$$

The roots of this characteristic polynomial are

$$\lambda_1 = 3, \quad \lambda_2 = 6, \quad \lambda_3 = 2$$

and their corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

The norms of these eigenvectors are

$$\|x_1\| = 1, \quad \|x_2\| = \sqrt{2}, \quad \|x_3\| = \sqrt{2}$$

Thus, the normalized eigenvectors are

$$\mathbf{e}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix}$$

The eigenvector matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

and it is simple to verify that

$$\mathbf{P}\mathbf{P}^T = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

## Skew-Symmetric Matrix

The eigenvalues of a skew-symmetric matrix  $\mathbf{A}$ , i.e.,  $\mathbf{A} = -\mathbf{A}^T$ , is purely imaginary or zero, i.e., they are sitting on the imaginary axis.

The proof of the above result is similar to that for symmetric matrices.

**Example:** Consider the following skew-symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 18 & -24 \\ -18 & 0 & 40 \\ 24 & -40 & 0 \end{bmatrix}$$

The roots of the characteristic polynomial are

$$I_1 = 0, \quad I_2 = 50i, \quad I_3 = -50i$$

They are either 0 or purely imaginary, as expected.

## Similarity of Matrices

Two matrices **A** and **B** are said to be similar if there exists a matrix **S** such that  $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ . Of course,  $\mathbf{S}^{-1}$  must exist for the quantities to be defined.

Theorem:

**A** and **B** have the same eigenvalues.

Proof. Given  $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ , we define a new vector  $\mathbf{y} = \mathbf{S}^{-1}\mathbf{x}$   $\mathbf{P}\ \mathbf{x} = \mathbf{S}\mathbf{y}$  to get

$$\mathbf{A}\mathbf{S}\mathbf{y} = \lambda\mathbf{S}\mathbf{y} \Rightarrow (\mathbf{S}^{-1}\mathbf{A}\mathbf{S})\mathbf{y} = \mathbf{B}\mathbf{y} = \lambda\mathbf{y}$$

Hence  $\lambda$  is also an eigenvalue of **B**.

Note that this result can be used to find a matrix **S** for a given matrix **A** such that its similarity or the transformed matrix **B** is a diagonal matrix.

Such a technique is very useful in solving some complicated problems.

## Diagonalization of a Matrix

### Theorem:

Consider a  $n \times n$  square matrix  $\mathbf{A}$  has distinct eigenvalues. Let  $\mathbf{P}$  be its eigenvector matrix. Then we have  $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$  is a diagonal matrix.

Proof. Let  $\lambda_1, \lambda_2, \dots, \lambda_n$  be the eigenvalues of  $\mathbf{A}$ . Thus

$$\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1, \quad \Rightarrow \quad \mathbf{A}\mathbf{x}_2 = \lambda_2\mathbf{x}_2, \quad \dots, \quad \mathbf{A}\mathbf{x}_n = \lambda_n\mathbf{x}_n$$

or in a matrix form

$$\begin{aligned} \mathbf{A}[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_n] &= [\lambda_1\mathbf{x}_1 \quad \lambda_2\mathbf{x}_2 \quad \dots \quad \lambda_n\mathbf{x}_n] \\ &= \underbrace{[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_n]}_{\mathbf{P}} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}}_{\mathbf{D}} \end{aligned} \quad \begin{aligned} &\Rightarrow \mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D} \\ &\Rightarrow \mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \end{aligned}$$

**Example:** Diagonalize the following symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 3 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

We have computed its normalized eigenvector matrix in the previous example

$$\mathbf{P} = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

It is simple to verify that

$$\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 4 & 0 & 2 \\ 0 & 3 & 0 \\ 2 & 0 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

$= \begin{bmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ , which is a diagonal matrix with its diagonal elements being the eigenvalues of  $\mathbf{A}$ .

**Example:** Diagonalize the following non-symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & -5 \\ 1 & 0 & 1 \\ 1 & 0 & 6 \end{bmatrix}$$

It has three eigenvalues at  $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 5$  and corresponding eigenvectors

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} -5 \\ 4 \\ -1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 0 \\ -3 \end{pmatrix} \Rightarrow \mathbf{P} = \begin{bmatrix} 0 & -5 & 3 \\ 1 & 4 & 0 \\ 0 & -1 & -3 \end{bmatrix}$$

It is tedious to compute that

$$\mathbf{P}^{-1} = \frac{1}{18} \begin{bmatrix} 12 & 18 & 12 \\ -3 & 0 & -3 \\ 1 & 0 & -5 \end{bmatrix} \Rightarrow \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_1 & 0 & 0 \\ 0 & \mathbf{I}_2 & 0 \\ 0 & 0 & \mathbf{I}_3 \end{bmatrix} = \mathbf{D}$$



## Quadratic Forms

**Quadratic forms** arise in many system analysis techniques. We study in the next the behavior of quadratic forms in term of eigen-distribution of an associated matrix.

A quadratic form of two variables  $x_1$  and  $x_2$  is

$$Q(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2$$

A quadratic form of three variables  $x_1$ ,  $x_2$  and  $x_3$  is

$$Q(x_1, x_2, x_3) = ax_1^2 + bx_2^2 + cx_3^2 + dx_1x_2 + ex_1x_3 + fx_2x_3$$

A quadratic form of  $n$  variables  $x_1, x_2, \dots, x_n$  is

$$Q(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$$

A quadratic form can be expressed in terms of a matrix form

$$\begin{aligned} Q(x_1, x_2, \dots, x_n) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \mathbf{x}^T \mathbf{A} \mathbf{x} \end{aligned}$$

In general, one can choose  $\mathbf{A}$  such that it is a symmetric matrix. For example, for  $n = 2$ ,

$$Q(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

For  $n = 3$ ,

$$\begin{aligned} Q(x_1, x_2, x_3) &= ax_1^2 + bx_2^2 + cx_3^2 + dx_1x_2 + ex_1x_3 + fx_2x_3 \\ &= (x_1 \quad x_2 \quad x_3) \begin{bmatrix} a & d/2 & e/2 \\ d/2 & b & f/2 \\ e/2 & f/2 & c \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \end{aligned}$$

Example: Given

$$Q(x_1, x_2, x_3) = x_1^2 + x_2^2 + 3x_3^2 + 6x_1x_2 + 4x_1x_3 - 10x_2x_3$$

express it as a  $\mathbf{x}^T \mathbf{A} \mathbf{x}$  with  $\mathbf{A}$  being symmetric.

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 1 & -5 \\ 2 & -5 & 3 \end{bmatrix}$$

**Definitions:** A quadratic form  $Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$  is said to be

1. **positive definite** if  $Q(\mathbf{x}) > 0$  for all values of  $\mathbf{x}$  except for  $\mathbf{x} = \mathbf{0}$ .

The corresponding  $\mathbf{A}$  is also said to be **positive definite**.

2. **positive semi-definite** if  $Q(\mathbf{x}) \geq 0$  for all values of  $\mathbf{x}$ .

The corresponding  $\mathbf{A}$  is also said to be **positive semi-definite**.

3. **negative semi-definite** if  $Q(\mathbf{x}) \leq 0$  for all values of  $\mathbf{x}$ .

The corresponding  $\mathbf{A}$  is also said to be **negative semi-definite**.

4. **negative definite** if  $Q(\mathbf{x}) < 0$  for all values of  $\mathbf{x}$  except for  $\mathbf{x} = \mathbf{0}$ .

The corresponding  $\mathbf{A}$  is also said to be **negative definite**.

5. **indefinite** if  $Q(\mathbf{x}) > 0$  for some  $\mathbf{x}$  and  $Q(\mathbf{x}) < 0$  for some other  $\mathbf{x}$ .

The corresponding  $\mathbf{A}$  is also said to be **indefinite**.

## Diagonal Quadratic Forms

Quadratic form  $Q(\mathbf{x}) = \mathbf{x}^T \mathbf{D} \mathbf{x}$  is said to be a diagonal quadratic form if  $\mathbf{D}$  is a diagonal matrix.

$$\begin{aligned} Q(x_1, x_2, \dots, x_n) &= \mathbf{x}^T \mathbf{D} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= d_{11}x_1^2 + d_{22}x_2^2 + \cdots + d_{nn}x_n^2 \end{aligned}$$

It is clear that the eigenvalues of  $\mathbf{D}$  are  $d_{11}, d_{22}, \dots$ , and  $d_{nn}$ .

## How to Determine the Definiteness of a Diagonal Quadratic Form?

A quadratic form  $Q(\mathbf{x}) = \mathbf{x}^T \mathbf{D} \mathbf{x}$  with  $\mathbf{D}$  being a diagonal matrix is

1. **positive definite** if and only if all the eigenvalues of  $\mathbf{D}$  are positive.
2. **positive semi-definite** if and only if all the eigenvalues of  $\mathbf{D}$  are non-negative.
3. **negative semi-definite** if and only if all the eigenvalues of  $\mathbf{D}$  are non-positive.
4. **negative definite** if and only if all the eigenvalues of  $\mathbf{D}$  are negative.
5. **indefinite** if and only if some eigenvalues of  $\mathbf{D}$  are positive and some are negative.

## Diagonalization of a General Quadratic Form

Given a quadratic form  $Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$  with  $\mathbf{A}$  being a symmetric matrix, we have shown that  $\mathbf{A}$  has all real eigenvalues and eigenvectors. For an normalized eigenvector matrix  $\mathbf{P}$ , then  $\mathbf{P}^{-1} = \mathbf{P}^T$  and  $\mathbf{D} = \mathbf{P}^T \mathbf{A} \mathbf{P}$  is a diagonal matrix.

We define a new variable  $\mathbf{y} = \mathbf{P}^T \mathbf{x}$  or equivalently  $\mathbf{x} = \mathbf{P} \mathbf{y}$

$$Q(x_1, x_2, \dots, x_n) = \mathbf{x}^T \mathbf{A} \mathbf{x} = (\mathbf{y}^T \mathbf{P}^T) \mathbf{A} (\mathbf{P} \mathbf{y}) = \mathbf{y}^T (\mathbf{P}^T \mathbf{A} \mathbf{P}) \mathbf{y} = \mathbf{y}^T \mathbf{D} \mathbf{y} = Q(\mathbf{y})$$

$$\Rightarrow Q(y_1, y_2, \dots, y_n) = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= d_{11} y_1^2 + d_{22} y_2^2 + \cdots + d_{nn} y_n^2$$

## How to Determine the Definiteness of a General Quadratic Form?

A quadratic form  $Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$  with  $\mathbf{A}$  being a symmetric matrix is

1. **positive definite** if and only if all the eigenvalues of  $\mathbf{A}$  are positive.
2. **positive semi-definite** if and only if all the eigenvalues of  $\mathbf{A}$  are non-negative.
3. **negative semi-definite** if and only if all the eigenvalues of  $\mathbf{A}$  are non-positive.
4. **negative definite** if and only if all the eigenvalues of  $\mathbf{A}$  are negative.
5. **indefinite** if and only if some eigenvalues of  $\mathbf{A}$  are positive and some are negative.



**Example:** Show whether the following quadratic form is positive semi-definite

$$Q(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2^2$$

We rewrite

$$Q(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 \quad x_2) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

and obtain the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \Rightarrow \det(\mathbf{A} - I\mathbf{I}) = \det \begin{bmatrix} 1-I & 1 \\ 1 & 1-I \end{bmatrix} = I(2-I)$$

$$\Rightarrow I_1 = 0, \quad I_2 = 2, \quad \mathbf{e}_1 = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$$

$$\Rightarrow \mathbf{P} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

Let  $\mathbf{y} = \mathbf{P}^T \mathbf{x}$ , we have  $Q(y_1, y_2) = (y_1 \quad y_2) \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 2y_2^2 \geq 0$

**Example:** Show that the following quadratic is indefinite

$$Q(\mathbf{x}) = x_1^2 - x_3^2 - 4x_1x_2 + 4x_2x_3$$

It can be written as

$$Q(\mathbf{x}) = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{bmatrix} 1 & -2 & 0 \\ -2 & 0 & 2 \\ 0 & 2 & -1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

The eigenvalues and eigenvectors of matrix  $\mathbf{A}$  are given by

$$\lambda_1 = 0, \quad \mathbf{e}_1 = \begin{pmatrix} 2/3 \\ 1/3 \\ 2/3 \end{pmatrix}, \quad \lambda_2 = -3, \quad \mathbf{e}_2 = \begin{pmatrix} -1/3 \\ -2/3 \\ 2/3 \end{pmatrix}, \quad \lambda_3 = 3, \quad \mathbf{e}_3 = \begin{pmatrix} -2/3 \\ 2/3 \\ 1/3 \end{pmatrix}$$

Let  $\mathbf{y} = \mathbf{P}^T \mathbf{x}$ , we have

$$Q(\mathbf{y}) = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = -3y_2^2 + 3y_3^2, \text{ indefinite !}$$

# Part 2: Numerical Methods

## ISSUES IN NUMERICAL ANALYSIS

- **WHAT IS NUMERICAL ANALYSIS?**

- It is a way to do highly complicated mathematics problems on a computer.
- It is also known as a technique widely used by scientists and engineers to solve their problems.

- **TWO ISSUES OF NUMERICAL ANALYSIS:**

- How to compute? This corresponds to algorithmic aspects;
- How accurate is it? That corresponds to error analysis aspects.

- **ADVANTAGES OF NUMERICAL ANALYSIS:**

- It can obtain numerical answers of the problems that have no “analytic” solution.
- It does NOT need special substitutions and integrations by parts. It needs only the basic mathematical operations: addition, subtraction, multiplication and division, plus making some comparisons.

- **IMPORTANT NOTES:**

- Numerical analysis solution is always numerical.
- Results from numerical analysis is an approximation.

- **NUMERICAL ERRORS**

When we get into the **real world** from an **ideal world** and **finite** to **infinite**, errors arise.

- **SOURCES OF ERRORS:**

- Mathematical problems involving quantities of infinite precision.
- Numerical methods bridge the precision gap by putting errors under firm control.
- Computer can only handle quantities of finite precision.

## – TYPES OF ERRORS:

- Truncation error (finite speed and time) - An example:

$$\begin{aligned} e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!} = \left( 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} \right) + \sum_{n=4}^{\infty} \frac{x^n}{n!} \\ &= p_3(x) + \sum_{n=4}^{\infty} \frac{x^n}{n!} \end{aligned}$$

- Round-off error (finite word length): All computing devices represent numbers with some imprecision, except for integers.
- Human errors: (a) Mathematical equation/model. (b) Computing tools/machines. (c) Error in original data. (d) Propagated error.

– MEASURE OF ERRORS:

Let  $a$  be a scalar to be computed and let  $\bar{a}$  be its approximation.

Then, we define

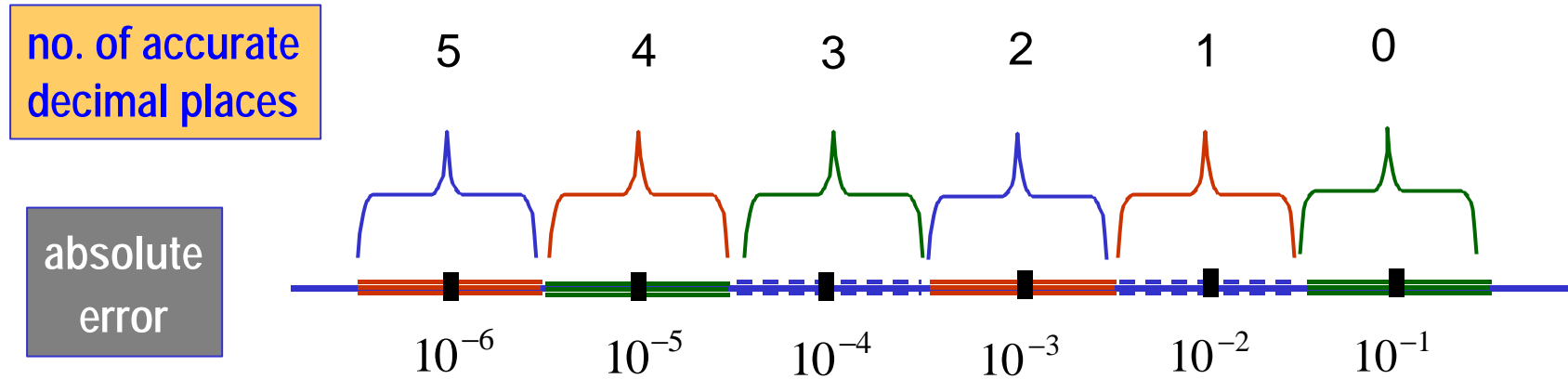
- Absolute error = | true value – approximated value |.

$$e = | a - \bar{a} |$$

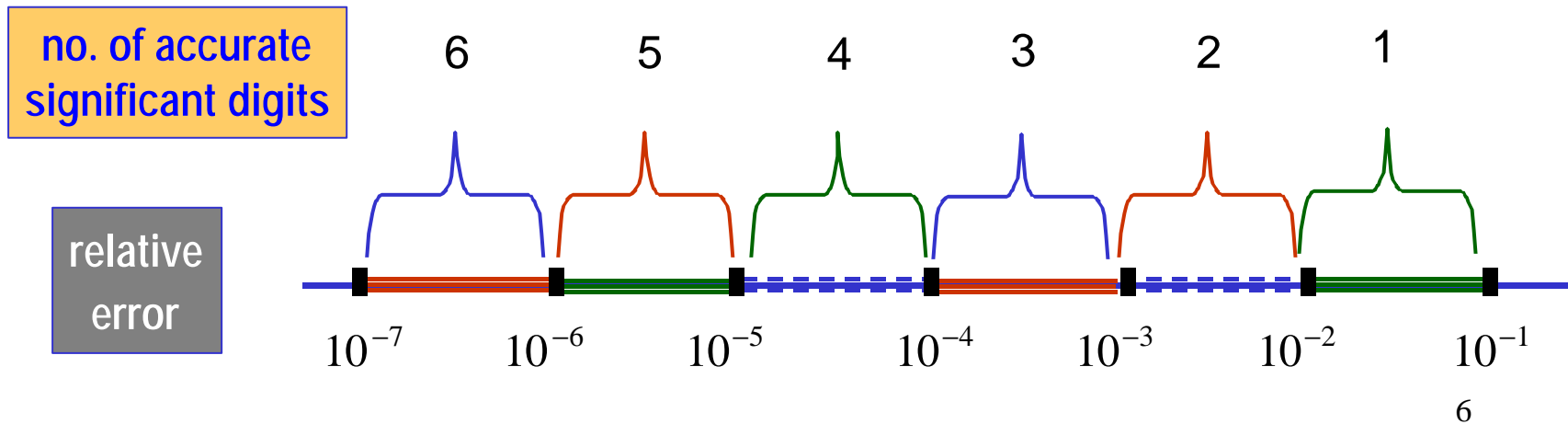
- Relative error =  $\left| \frac{\text{true value} - \text{approximated value}}{\text{true value}} \right|$

$$e_r = \left| \frac{a - \bar{a}}{a} \right|$$

## Absolute Error and Accuracy in Decimal Places



## Relative Error and Accuracy in Significant Digits





**Example:** Let the true value of  $\pi$  be 3.1415926535898 and its approximation be 3.14 as usual. Compute the absolute error and relative error of such an approximation.

The absolute error:

$$e = |p - \bar{p}| = |3.1415926535898 - 3.14| = 0.0015926535898$$

which implies that the approximation is accurate up to 2 decimal places.

The relative error:

$$e_r = \left| \frac{p - \bar{p}}{p} \right| = \frac{0.0015926535898}{3.1415926535898} = 0.000506957382897$$

which implies that the approximation has a accuracy of 3 significant figures.

- **STABILITY AND CONVERGENCE**

- **STABILITY** in numerical analysis refers to the trend of error change iterative scheme. It is related to the concept of convergence.

It is stable if initial errors or small errors at any time remain small when iteration progresses. It is unstable if initial errors or small errors at any time get larger and larger, or eventually get unbounded.

- **CONVERGENCE**: There are two different meanings of convergence in numerical analysis:

a. If the discretized interval is getting finer and finer after discretizing the continuous problems, the solution is convergent to the true solution.

b. For an iterative scheme, convergence means the iteration will get closer to the true solution when it progresses.

## Solutions to Nonlinear Equations (Computing Zeros)

- **Problem:** Given a function  $f(x)$ , which normally is nonlinear, the problem of “computing zeros” means to find all possible points, say

$$\tilde{x}_0, \tilde{x}_1, \dots, \tilde{x}_n$$

such that

$$f(\tilde{x}_0) = f(\tilde{x}_1) = \dots = f(\tilde{x}_n) = 0$$

However, it is often that we are required to find a single point  $\tilde{x}_0$  in certain interval, say  $[a,b]$  such that

$$f(\tilde{x}_0) = 0$$

General strategy is to design an iterative process of the form

$$x_{n+1} = g(x_n)$$

with some starting point  $x_0$ . So that the numerical solution as

$$x_n \rightarrow \tilde{x}_0, \quad \text{as } n \rightarrow \infty$$

Thus, instead of finding the exact solution, we find an approximation.

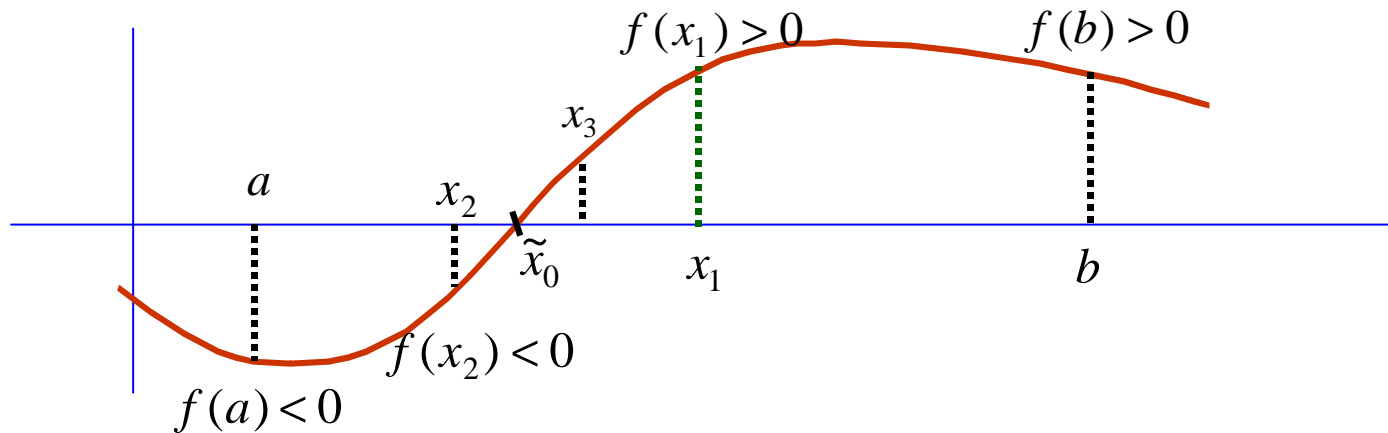
We focus on the following methods for this subject:

**Bisection Method** + **False Position Method** + **Newton Method** +

**Secant Method** + **Fixed Point Method** + Your Own Method

## BISECTION METHOD

Given a function  $f(x)$  in  $[a, b]$  satisfying  $f(a)f(b) < 0$ , find a zero of  $f(x)$  in  $[a, b]$ .



Step 0: Let  $x_a := a$ ;  $x_b := b$ ;  $n = 1$ .

Step 1: Cut the interval in the middle, i.e., find  $x_n := \frac{x_a + x_b}{2}$

Step 2: Define 
$$\begin{cases} x_a := x_a; & x_b := x_n; & \text{if } f(x_a)f(x_n) < 0 \\ x_a := x_n; & x_b := x_b; & \text{if } f(x_b)f(x_n) < 0 \end{cases}$$

Step 3: If  $x_n$  is close enough to  $\tilde{x}_0$ , stop. Otherwise,  $n := n + 1$  & go to Step 1.

## Advantages:

1. It is guaranteed to work if  $f(x)$  is continuous in  $[a, b]$  and a zero actually exists.
2. A specific accuracy of iterations is known in advance. Few other root-finding methods share this advantage.

## Disadvantages:

- a. It requires the values of  $a$  and  $b$ .
- b. The convergence of interval halving is very slow.
- c. Multiple zeros between  $a$  and  $b$  can cause problem.

**Example:** Let  $f(x) = x^2 - 1$ . Find its zero in  $[0, 1.5]$

Of course, we know  $f(x)$  has a root at  $\tilde{x}_0 = 1$ . Let us find it using the Bisection Method:

Step 0:  $x_a = 0, \quad x_b = 1.5, \quad n = 1$

Step 1:  $x_1 = \frac{0+1.5}{2} = 0.75$

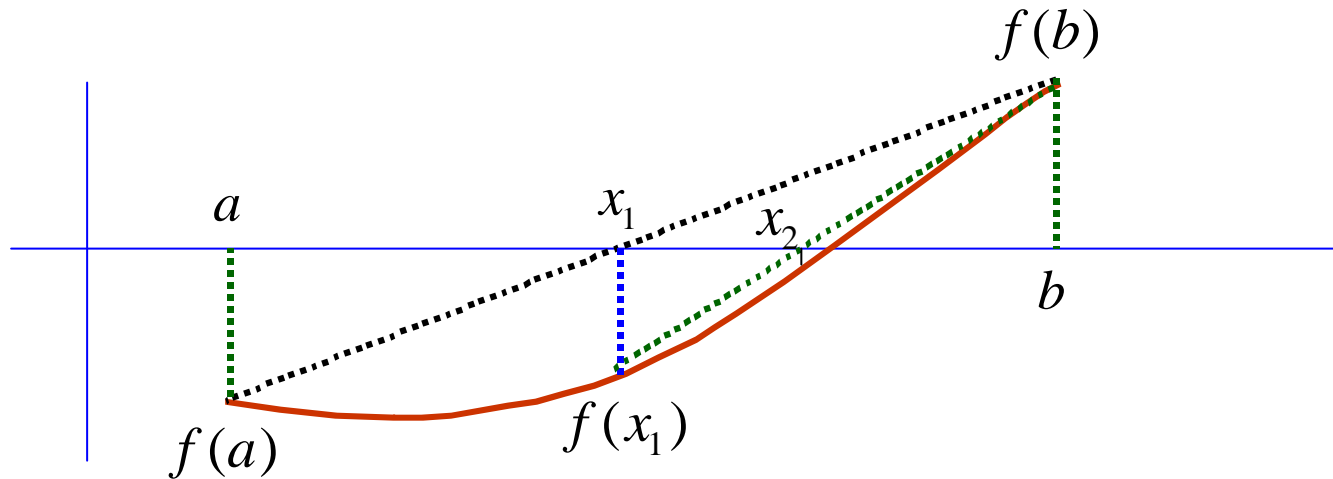
Step 2:  $f(x_b)f(x_1) = (1.5^2 - 1)(0.75^2 - 1) = -0.546875 < 0$

$$\Rightarrow x_2 = \frac{0.75+1.5}{2} = 1.125$$

$$\Rightarrow x_3 = \frac{0.75+1.125}{2} = 0.9375 \quad \Rightarrow \quad x_4 = 1.03125 \quad \dots$$

## FALSE POSITION METHOD

The graph used in this method is shown in the following figure.



The key idea is to approximate the curve by a straight line within the interval and identify a “false” position  $x_1$ , which of course may not be a true solution. We can keep repeating this procedure to get approximations of the solution,  $x_2, x_3, \dots$ . Mathematically,

$$x_{n+1} = x_n - \frac{b - x_n}{f(b) - f(x_n)} f(x_n), \quad x_0 = a$$



## Advantage:

Convergence is faster than bisection method.

## Disadvantages:

1. It requires  $a$  and  $b$ .
2. The convergence is generally slow.
3. It is only applicable to  $f(x)$  of certain fixed curvature in  $[a, b]$ .
4. It cannot handle multiple zeros.

**Example:**  $f(x) = x^2 - 1$ . Find its root in  $[0, 1.5]$

$$\begin{aligned}x_1 &= x_0 - \frac{b - a}{f(b) - f(a)} f(a) \\ &= 0 - \frac{1.5 - 0}{1.25 - (-1)} (-1) = 0.6667\end{aligned}$$

$$\begin{aligned}x_2 &= x_1 - \frac{b - x_1}{f(b) - f(x_1)} f(x_1) \\ &= 0.6667 - \frac{1.5 - 0.6667}{1.25 - (-0.5556)} (-0.5556) \\ &= 0.9231\end{aligned}$$

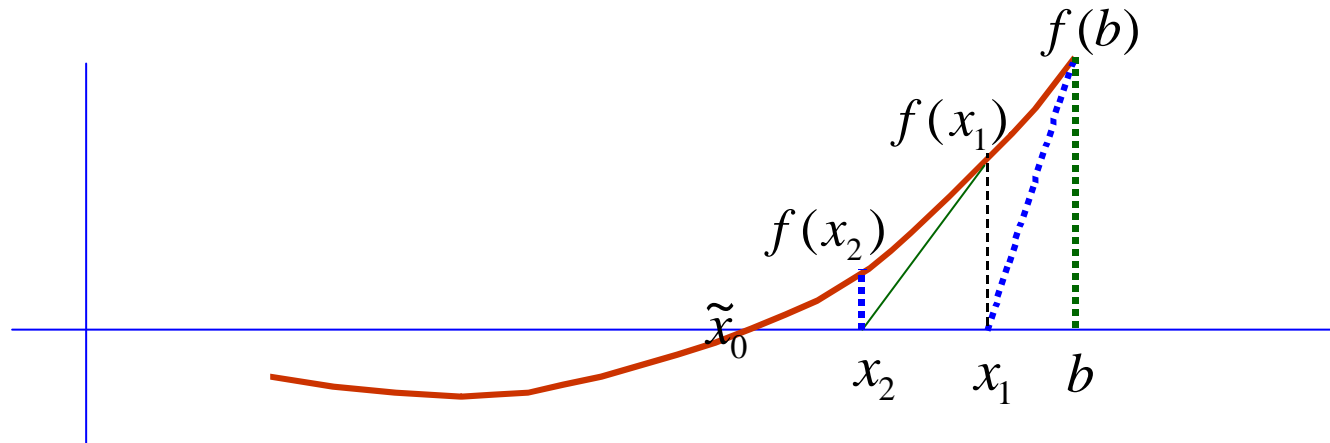
$$\begin{aligned}x_3 &= x_2 - \frac{b - x_2}{f(b) - f(x_2)} f(x_2) \\ &= 0.9231 - \frac{1.5 - 0.9231}{1.25 - (-0.1479)} (-0.1479) \\ &= 0.9841\end{aligned}$$

...

$$x_n \rightarrow 1 \quad \text{as} \quad n \rightarrow \infty$$

## Newton Method

Assume that  $f'(x)$  exists and nonzero at  $x_n$  for all  $n$ .



**Zero-finding:** The linear approximation based on one point  $(x_1, y_1)$  only is given by

$$y = y_1 + (x - x_1)f'(x)$$

We look for a point  $x$  for which  $y = 0$ . As such we have the following iteration:

$$y_{n+1} = y_n + (x_{n+1} - x_n)f'(x_n) = 0 \quad \Rightarrow \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

### Advantages:

1. Starting point  $x_1$  can be arbitrary.
2. The convergence is faster than the previous two methods.

### Disadvantages:

1. It needs  $f'(x)$ .
2. The divergence may occur.

**Example:** Find zero of  $f(x) = x^2 - 1$  in  $[0, 1.5]$  using Newton's Method

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

Starting with any initial

point, say  $x_0 = 0.1$ , we

have

$$x_1 = \frac{1}{2 \times 0.1}(0.1^2 + 1) = 5.05$$

$$x_2 = \frac{1}{2 \times 5.05}(5.05^2 + 1) = 2.624$$

$$x_3 = \frac{1}{2 \times 2.624}(2.624^2 + 1) = 1.5026$$

$$x_4 = \frac{1}{2 \times 1.5026}(1.5026^2 + 1) = 1.084$$

$$x_5 = \frac{1}{2 \times 1.084}(1.084^2 + 1) = 1.003$$

...

Again,  $x_n \rightarrow 1$  as  $n \rightarrow \infty$

## Secant Method

Secant Method is a modified version of Newton's method in which  $f'(x_n)$  is approximated by

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

Substituting this into the iteration scheme of Newton's method, we obtain

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

- Advantage: 1) Convergence is fast and 2) it does not need derivative.
- Disadvantage: The method may fail.

## Fixed Point Method

Start from  $f(x) = 0$  and derive a relation

$$x = g(x)$$

Example: Compute zero for  $f(x)$  with  $f(x) = e^x - 4 - 2x$  or find  $x$  such that

$$e^x - 4 - 2x = 0$$

$$\Rightarrow x = \frac{1}{2}(e^x - 4) = g(x) \quad (1)$$

$$\Rightarrow e^x = 4 + 2x \Rightarrow x = \ln(4 + 2x) = g(x) \quad (2)$$

The fixed-point method is simply given by

$$x_{n+1} = g(x_n)$$

Q: Does it work? Does it converge? A: Maybe yes and maybe not.

Q: When does it converge?

A: Convergence Theorem

Consider a function  $f(x)$  and suppose it has a zero on the interval  $[a, b]$ .

Also, consider the iteration scheme

$$x_{n+1} = g(x_n)$$

derived using fixed-point method. Then this scheme converges, i.e.,

$$x_n \rightarrow \tilde{x}_0 \quad \text{as} \quad n \rightarrow \infty$$

if the following conditions are satisfied:

(1)  $|g'(x)| < 1$  for all  $x \in [a, b]$ .  $g(x)$  is also said to be contraction in  $[a, b]$ .

(2) Start any initial point  $x_0 \in [a, b]$ .

**Remark:** If the above conditions are not satisfied, the iteration scheme might still converge as the above theorem only gives sufficient conditions.



**Example:** Compute zeros of  $f(x) = e^x - 4 - 2x$

**Scheme 1:**

$$x = g(x) = \frac{1}{2}(e^x - 4)$$

$$x_{n+1} = \frac{1}{2}(e^{x_n} - 4)$$

$$g'(x) = \frac{1}{2}e^x$$

with  $|g'(x)| < 1$  for  $-\infty < x < 0.693$

Let us choose  $x_0 = -2$ ,

$$\Rightarrow x_1 = \frac{1}{2}(e^{-2} - 4) = -1.9323$$

$$x_2 = \frac{1}{2}(e^{-1.9323} - 4) = -1.9276$$

$$x_3 = -1.9273, x_4 = -1.9272,$$

$$x_5 = -1.9272$$

**Scheme 2:**

$$x = g(x) = \ln(4 + 2x)$$

$$x_{n+1} = \ln(4 + 2x_n), \quad x > -2$$

$$g'(x) = \frac{2}{4 + 2x} \Rightarrow$$

$$|g'(x)| < 1, x \in (-\infty, -3)$$

$$|g'(x)| < 1, x \in (-1, \infty)$$

Let  $x_0 = 0$ ,  $\Rightarrow$

$$x_1 = 1.3863, x_2 = 1.9129,$$

$$x_3 = 2.0574, x_4 = 2.0937,$$

$$x_5 = 2.1026, x_6 = 2.1048,$$

$$x_7 = 2.1053, x_8 = 2.1054,$$

$$x_9 = 2.1054$$

Applications: Compute  $\sqrt{3}$ . (Actual value = 1.73205)

Solution: Let  $x = \sqrt{3} \Rightarrow x^2 = 3 \Rightarrow f(x) = x^2 - 3 = 0$

The problem is transformed to a problem of finding zero (or root) for  $f(x)$  in  $[0, \infty)$ .

We use Newton's Method with  $x_0=1$ .

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

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

$$= \frac{1}{2x_n}(x_n^2 + 3)$$

$$x_1 = \frac{1}{2x_0}(x_0^2 + 3) = 2$$

$$x_2 = \frac{1}{2x_1}(x_1^2 + 3) = 1.75$$

$$x_3 = \frac{1}{2x_2}(x_2^2 + 3) = 1.73214$$

$$x_4 = \frac{1}{2x_3}(x_3^2 + 3) = 1.73205$$

(good enough)

## SUMMARY

- Bisection Method

**Condition:** Continuous function  $f(x)$  in  $[a, b]$  satisfying  $f(a)f(b) < 0$ .

**Convergence:** Slow but sure. Linear.

- False position method

**Condition:** Continuous function  $f(x)$  in  $[a, b]$  satisfying  $f(a)f(b) < 0$ .

**Convergence:** Slow (linear).

- Newton Method

**Condition:** Existence of nonzero  $f'(x)$

**Convergence:** Fast (quadratic).

- Secant Method

**Condition:** Existence of nonzero  $f(x_{n+1}) - f(x_n)$

**Convergence:** Fast (quadratic).

- Fixed-point Method

**Condition:** Contraction of  $g(x)$ .

**Convergence:** Varying with the nature of  $g(x)$ .

## Interpolation

**Problem:** Given a set of measured data, say  $n + 1$  pairs,

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

the problem of interpolation is to find a function  $f(x)$  such that

$$f(x_i) = y_i, \quad i = 0, 1, \dots, n$$

- $x_i$  is called nodes;
- $f(x)$  is said to interpolate the data and is called interpolation function.
- $f(x)$  is said to approximate  $g(x)$  if the data are from a function  $g(x)$ .
- It is called interpolate (or extrapolate) if  $f(x)$  gives values within (or outside)  $[x_0, x_n]$ .

A simple choice for  $f(x)$  is a polynomial of degree  $n$ :

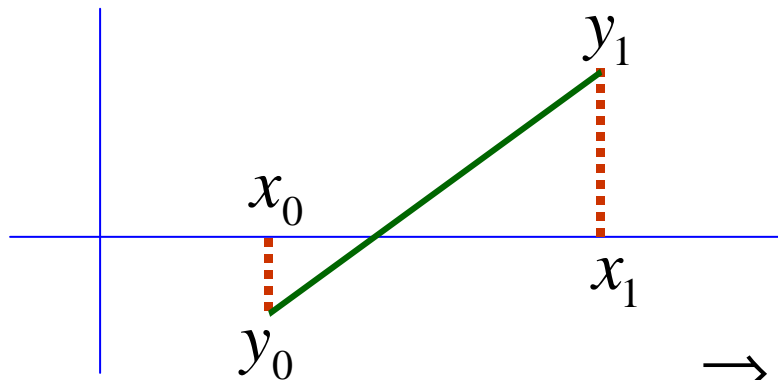
$$f(x) = p_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

The existence and uniqueness have been verified. There is only one polynomial existing for the interpolation.

## LAGRANGIAN POLYNOMIALS

### (a) Fitting Two Points

Fit the linear polynomial for two given points  $(x_0, y_0)$  and  $(x_1, y_1)$ .

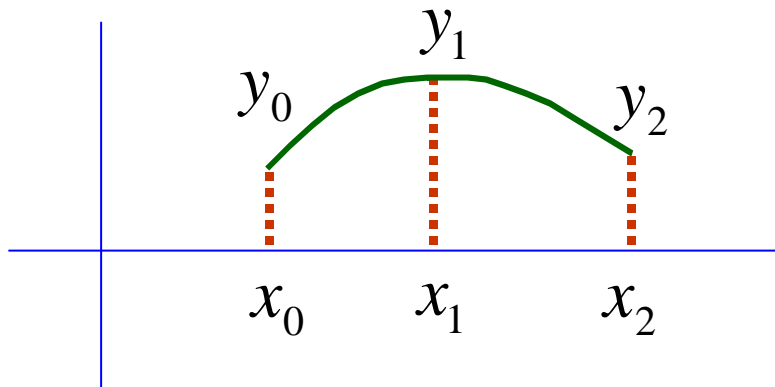


$$\begin{aligned} p_1(x) &= L_0(x)y_0 + L_1(x)y_1 \\ &= \left( \frac{x - x_1}{x_0 - x_1} \right) y_0 + \left( \frac{x - x_0}{x_1 - x_0} \right) y_1 \end{aligned}$$

$$\Rightarrow p_1(x_0) = y_0, \quad p_1(x_1) = y_1 \quad 27$$

## (b) Fitting Three Points

Fit the quadratic polynomial for three given points.



$$\begin{aligned} \Rightarrow p_2(x_0) &= y_0 \\ p_2(x_1) &= y_1 \\ p_2(x_2) &= y_2 \end{aligned}$$

$$\begin{aligned} p_2(x) &= \left( \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \right) y_0 + \left( \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \right) y_1 + \left( \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \right) y_2 \\ &= L_0(x)y_0 + L_1(x)y_1 + L_2(x)y_2 \end{aligned}$$

$$L_0(x) = \begin{cases} 1, & x = x_0 \\ 0, & x = x_1 \\ 0, & x = x_2 \end{cases} \quad L_1(x) = \begin{cases} 0, & x = x_0 \\ 1, & x = x_1 \\ 0, & x = x_2 \end{cases} \quad L_2(x) = \begin{cases} 0, & x = x_0 \\ 0, & x = x_1 \\ 1, & x = x_2 \end{cases}$$

### (c) Fitting $n+1$ Points

Lagrangian polynomial for fitting  $n + 1$  given points

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n).$$

is given by

$$p_n(x) = \sum_{k=0}^n L_k(x) y_k = \sum_{k=0}^n \frac{l_k(x)}{l_k(x_k)} y_k$$

where

$$l_k(x) = \frac{1}{x - x_k} (x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n);$$

$$L_k(x) = \begin{cases} 1, & x = x_k \\ 0, & x = x_j, \quad j \neq k \end{cases}$$

## NEWTON'S DIVIDED DIFFERENCE METHOD

The following two disadvantages of Lagrangian polynomial method lead us to develop a new method for the interpolation. They are:

- (1) it involves more arithmetic operations; and
- (2) we essentially need to start over the computation if we desire to add or subtract a point from the set of data.

### The Basic Idea of Divided Difference:

Consider the  $n$ -th-degree polynomial written in a special way:

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

The key idea is to find  $a_0, \dots, a_n$  so that  $P_n$  interpolates the given data:

$$(x_0, f_0), (x_1, f_1), \dots, (x_n, f_n).$$



Define the first order divided difference between two notes  $x_i$  and  $x_{i+1}$  as

$$f[x_i, x_{i+1}] = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = f_i^{[1]} = f[x_{i+1}, x_i]$$

the second order divided difference as

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i} = f_i^{[2]}$$

and the higher order divided differences as

$$f[x_i, x_{i+1}, \dots, x_{i+m}] = \frac{f[x_{i+1}, \dots, x_{i+m}] - f[x_i, \dots, x_{i+m-1}]}{x_{i+m} - x_i} = f_i^{[m]}$$

as well as zero-th order divided difference:

$$f[x_i] = f_i = f_i^{[0]}$$

### Divided Difference Table:

$x_i$	$f_i$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	
$x_0$	$f_0$	$f_0^{[1]}$	$f_0^{[2]}$		
$x_1$	$f_1$	$f_1^{[1]}$	$f_1^{[2]}$	$f_0^{[3]}$	
$x_2$	$f_2$	$f_2^{[1]}$	$f_2^{[2]}$	$f_1^{[3]}$	$f_0^{[4]}$
$x_3$	$f_3$	$f_3^{[1]}$			
$x_4$	$f_4$				

### Example:

$x_i$	$f_i$	1st Order	2nd Order	3rd Order	4th Order
3.2	22.0	8.400			
2.7	17.8	2.118	2.856	-0.5280	
1.0	14.2	6.342	2.012	0.0865	0.2560
4.8	38.3	16.750	2.263		
5.6	51.7				

**Idea:** If

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

is an interpolation of the given data:

$$(x_0, f_0), (x_1, f_1), \dots, (x_n, f_n),$$

then we have

$$P_n(x_i) = f_i, \quad i = 0, 1, \dots, n$$

Thus

$$P_n(x_0) = a_0 + (x_0 - x_0)a_1 + \dots + (x_0 - x_0)\dots(x - x_{n-1})a_n = a_0 = f_0 = f_0^{[0]}$$

$$P_n(x_1) = a_0 + (x_1 - x_0)a_1 = f_0 + (x_1 - x_0)a_1 = f_1 \quad \Rightarrow \quad a_1 = \frac{f_1 - f_0}{x_1 - x_0} = f_0^{[1]}$$

In general, we can show that

$$a_k = f_0^{[k]}, \quad k = 1, 2, \dots, n$$

Thus, given a set of data:

$$(x_0, f_0), (x_1, f_1), \dots, (x_n, f_n),$$

their  $n$ -th degree polynomial interpolation is given by

$$P_n(x) = f_0^{[0]} + (x - x_0)f_0^{[1]} + (x - x_0)(x - x_1)f_0^{[2]} + \dots \\ + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f_0^{[n]}.$$

The advantage of the above method is that there is no need to start all over again if their additional pairs of data are added. We simply need to compute additional divided differences.

Since  $n$ -th order polynomial interpolation of a given  $(n + 1)$  pairs of data is unique, thus the above polynomial and Lagrangian polynomial are exactly the same.

**Example:** Interpolate the following set of data

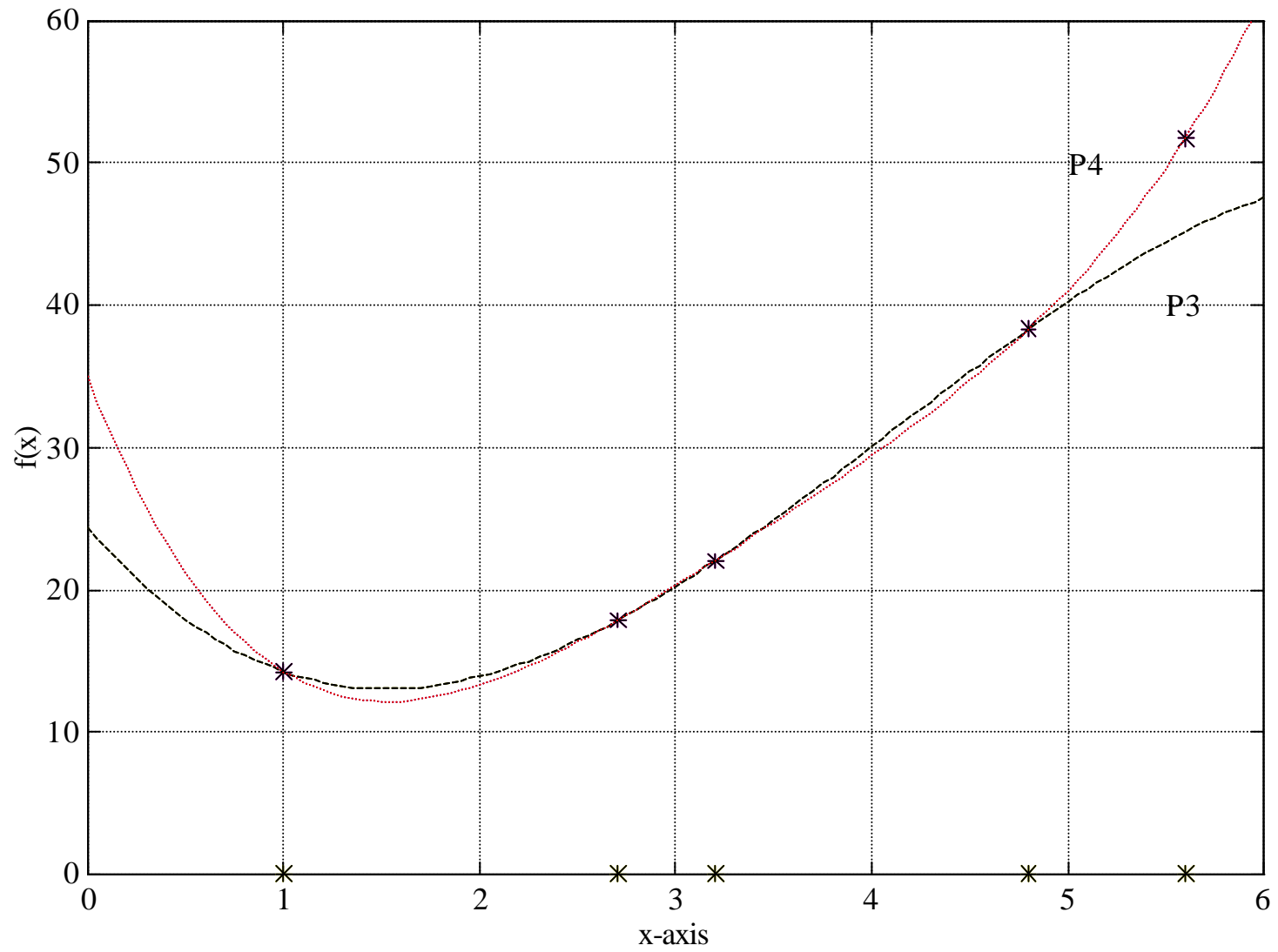
$x_i$	$f_i$	1st Order	2nd Order	3rd Order	4th Order
3.2	22.0	8.400	2.856		
2.7	17.8	2.118		-0.5280	
1.0	14.2	6.342	2.012		0.2560
4.8	38.3	16.750	2.263	0.0865	
5.6	51.7				

Interpolate from  $x_0$  to  $x_3$ :

$$P_3(x) = 22.0 + 8.400(x - 3.2) + 2.856(x - 3.2)(x - 2.7) - 0.528(x - 3.2)(x - 2.7)(x - 1.0)$$

Interpolate from  $x_0$  to  $x_4$ :

$$P_4(x) = P_3(x) + 0.256(x - 3.2)(x - 2.7)(x - 1.0)(x - 4.8)$$



## Evenly Spaced Data

The problem of interpolation from tabulated data is considerably simplified if the values of the function are given at evenly spaced intervals of the independent variable.

### Difference Table:

Assume that the given set of data is evenly spaced, i.e.,

$$x_{i+1} - x_i = h$$

(a) The first order differences of the functions are defined as:

$$\Delta f_0 = f_1 - f_0 \quad \text{at } x_0$$

$$\Delta f_1 = f_2 - f_1 \quad \text{at } x_1$$

$$\vdots$$

$$\Delta f_i = f_{i+1} - f_i \quad \text{at } x_i$$

(b) The second order differences of the functions are given by:

$$\Delta^2 f_0 = \Delta(\Delta f_0) = \Delta f_1 - \Delta f_0 \quad \text{at } x_0$$

$$\Delta^2 f_1 = \Delta(\Delta f_1) = \Delta f_2 - \Delta f_1 \quad \text{at } x_1$$

$\vdots$

$$\Delta^2 f_i = \Delta(\Delta f_i) = \Delta f_{i+1} - \Delta f_i \quad \text{at } x_i$$

(c) The n-th order differences of the functions are given by:

$$\Delta^n f_i = \Delta^{n-1} f_{i+1} - \Delta^{n-1} f_i$$



## Newton Forward Method for Evenly Spaced Data:

Given a set of measured data

$$(x_0, f_0), (x_1, f_1), \dots, (x_n, f_n),$$

in which  $x_{i+1} - x_i = h$ , then the Newton forward interpolation polynomial is

given by

$$P_n(x) = f_0 + \binom{s}{1} \Delta f_0 + \binom{s}{2} \Delta^2 f_0 + \dots + \binom{s}{n} \Delta^n f_0$$

where

$$s = \frac{x - x_0}{h} \quad \text{and} \quad \binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!}, \quad k = 1, 2, \dots, n$$

and  $\Delta^k f_0$  is the  $k$ -th order difference of the given data.

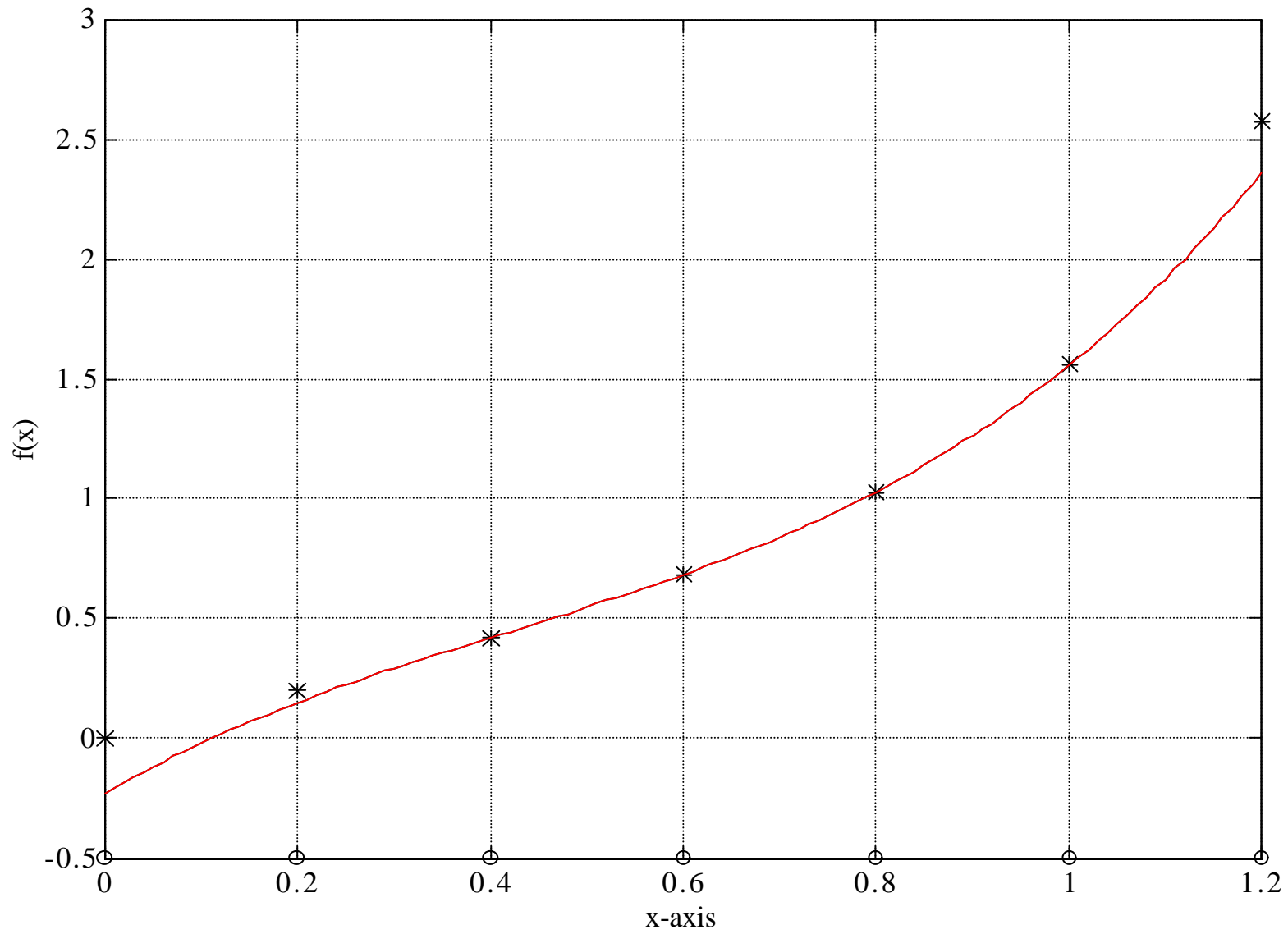
**Example:** Interpolate the following set of data using Newton Forward Method

$x_i$	$f_i$	1st Order	2nd Order	3rd Order	4th Order
0.0	0.000	0.203	0.017		
0.2	0.203	0.220	0.041	0.024	
0.4	0.423	0.261	0.085	0.044	0.020
0.6	0.684	0.346	0.181	0.096	0.052
0.8	1.030	0.527	0.488	0.307	0.211
1.0	1.557	1.015			
1.2	2.572				

Assume that we only want to interpolate from 0.4 ( $x_0$ ) to 1.0 ( $x_3$ ):

$$P_3(x) = 0.423 + 0.261 \binom{s}{1} + 0.085 \binom{s}{2} + 0.096 \binom{s}{3}$$

$$s = \frac{x - 0.4}{0.2} = 5x - 2_{40}$$



## Least Squares Approximation

**Least Squares Linear Fitting:** If we are given a set of data points,

$$(x_i, y_i) \quad i = 1, 2, \dots, n$$

can we use a line to fit these data points? The answer is positive.

If the line is expressed as

$$y = a_0 + a_1 x$$

where  $a_0$  and  $a_1$  are the two best values to be determined. Obviously, the

error  $e_i$  of each point  $(x_i, y_i)$  with respect to  $y = a_0 + a_1 x$  will be

$$e_i = y_i - y|_{x=x_i} = y_i - (a_0 + a_1 x_i)$$

The least squares criterion requires that

$$S = e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_1 x_i - a_0)^2$$

be a minimum.

At a minimum for  $S$ , the two derivatives  $\partial S / \partial a_0$  and  $\partial S / \partial a_1$  will both be zero:

$$\left. \begin{aligned} \frac{\partial S}{\partial a_1} &= \sum_{i=1}^n 2(y_i - a_1 x_i - a_0)(-x_i) = 0, \\ \frac{\partial S}{\partial a_0} &= \sum_{i=1}^n 2(y_i - a_1 x_i - a_0)(-1) = 0, \end{aligned} \right\} \begin{aligned} a_1 \sum_{i=1}^n x_i^2 + a_0 \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i, \\ a_1 \sum_{i=1}^n x_i + a_0 n &= \sum_{i=1}^n y_i \end{aligned}$$

Thus,  $a_0$  and  $a_1$  can be obtained so that the data points are linearly fitted.

In fact, we can write the above equations into a linear system:

$$\begin{bmatrix} \sum_{i=1}^n (x_i)^0 & \sum_{i=1}^n (x_i)^1 \\ \sum_{i=1}^n (x_i)^1 & \sum_{i=1}^n (x_i)^2 \end{bmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

for solve it for  $a_0$  and  $a_1$ .

## Least Squares Polynomials:

Instead of matching the data in every node, the least square method is trying to fit  $n$  pairs of data by a polynomial of a pre-determined degree, say  $m$ ,

$$y = a_0 + a_1x + a_2x^2 + \dots + a_mx = \sum_{j=0}^m a_j x^j$$

We define the fitting errors

$$e_i = y_i - \sum_{j=0}^m a_j x_i^j \quad S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( y_i - \sum_{j=0}^m a_j x_i^j \right)^2$$

In order to achieve minimal error  $S$  (least square error), all the partial derivatives

$$\frac{\partial S}{\partial a_0}, \frac{\partial S}{\partial a_1}, \dots, \frac{\partial S}{\partial a_m}$$

must equal to 0. Writing the equations for these given  $m + 1$  equations:

$$\left. \begin{aligned}
 \frac{\partial S}{\partial a_0} &= \sum_{i=1}^n 2 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) (-1) = 0 \\
 \frac{\partial S}{\partial a_1} &= \sum_{i=1}^n 2 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) (-x_i) = 0 \\
 &\vdots \\
 \frac{\partial S}{\partial a_m} &= \sum_{i=1}^n 2 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) (-x_i^m) = 0
 \end{aligned} \right\}
 \begin{aligned}
 a_0 n + a_1 \sum x_i + \cdots + a_n \sum x_i^m &= \sum y_i \\
 a_0 \sum x_i + a_1 \sum x_i^2 + \cdots + a_n \sum x_i^{m+1} &= \sum x_i y_i \\
 a_0 \sum x_i^2 + a_1 \sum x_i^3 + \cdots + a_n \sum x_i^{m+2} &= \sum x_i^2 y_i \\
 &\vdots \\
 a_0 \sum x_i^n + a_1 \sum x_i^{n+1} + \cdots + a_n \sum x_i^{2m} &= \sum x_i^m y_i
 \end{aligned}$$

Or solving the following system

$$\begin{pmatrix}
 n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^m \\
 \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{m+1} \\
 \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{m+2} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 \sum x_i^m & \sum x_i^{m+1} & \sum x_i^{m+2} & \cdots & \sum x_i^{2m}
 \end{pmatrix}
 \begin{pmatrix}
 a_0 \\
 a_1 \\
 a_2 \\
 \vdots \\
 a_m
 \end{pmatrix}
 =
 \begin{bmatrix}
 \sum y_i \\
 \sum x_i y_i \\
 \sum x_i^2 y_i \\
 \vdots \\
 \sum x_i^m y_i
 \end{bmatrix}$$

for  $a_0, a_1, \dots, a_m$ .

### Example:

To demonstrate how the method is used, we would fit a quadratic to the following data:

$$\begin{array}{l} x_i: 0.05 \quad 0.11 \quad 0.15 \quad 0.31 \quad 0.46 \quad 0.52 \quad 0.70 \quad 0.74 \quad 0.82 \quad 0.98 \quad 1.17 \\ y_i: 0.956 \quad 0.890 \quad 0.832 \quad 0.717 \quad 0.571 \quad 0.539 \quad 0.378 \quad 0.370 \quad 0.306 \quad 0.242 \quad 0.104 \end{array}$$

These data are actually a perturbation of the relation

$$y = 1 - x + 0.2x^2$$

Obviously we have

$$\begin{array}{l} \Sigma x_i = 6.01 \quad \Sigma x_i^2 = 4.6545 \quad \Sigma x_i^3 = 4.1150 \quad \Sigma x_i^4 = 3.9161 \\ n = 11 \quad \Sigma y_i = 5.9050 \quad \Sigma x_i y_i = 2.1839 \quad \Sigma x_i^2 y_i = 1.3357 \end{array}$$

Thus the equation system to be solved is:



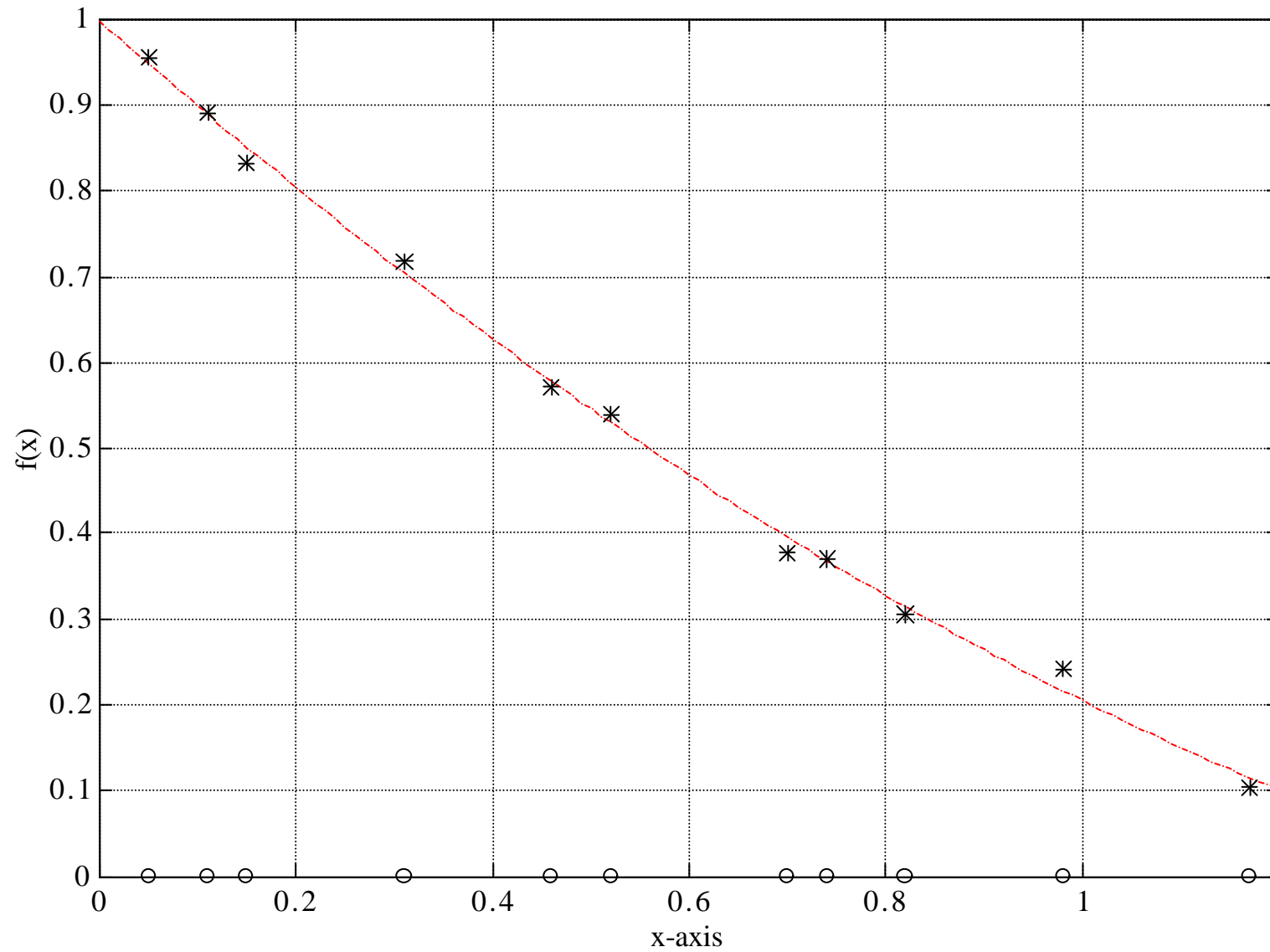
$$\begin{array}{r}
 11.0000a_0 + 6.0100a_1 + 4.6545a_2 = 5.9050 \\
 6.0100a_0 + 4.6545a_1 + 4.1150a_2 = 2.1839 \\
 4.6545a_0 + 4.1150a_1 + 3.9161a_2 = 1.3357
 \end{array}
 \left. \vphantom{\begin{array}{r} 11.0000a_0 + 6.0100a_1 + 4.6545a_2 = 5.9050 \\ 6.0100a_0 + 4.6545a_1 + 4.1150a_2 = 2.1839 \\ 4.6545a_0 + 4.1150a_1 + 3.9161a_2 = 1.3357 \end{array}} \right\}
 \begin{array}{l}
 a_0 = 0.998 \\
 a_1 = -1.018 \\
 a_2 = 0.225
 \end{array}$$

The above linear system can be solved using methods given in Part I of this course or using MATLAB software package.

Thus, the least square quadratic fit is given by

$$y = 0.998 - 1.018x + 0.225x^2$$

Compare this to  $y = 1 - x + 0.2x^2$ . We do not expect to reproduce the coefficients exactly because of the error in the data. Figure of next page shows a plot of the data and its fitting-curve.



## Numerical Integration

Given a function  $f(x)$  and an interval, say  $[a, b]$ , we want to find an algorithm to approximate

$$\int_a^b f(x) dx \cong ?$$

### Newton-Cotes Integration Method

Using Newton-Gregory Forward Polynomial

$$P_n(x) = f_0 + s\Delta f_0 + \frac{s(s-1)}{2!} \Delta^2 f_0 + \dots$$

to interpolate (approximate)  $f(x)$  in  $[a, b]$ , i.e.,

$$\int_a^b f(x) dx \cong \int_a^b P_n(x) dx$$

## NEWTON-COTES INTEGRATION

Although the analytical procedure could be used to find out the expression of integrals, a large number of integrals do not have solutions in closed form. Numerical integration applies regardless of the complexity of the integrand or the existence of a closed form for the integral.

### General Consideration:

A very simple method used in the numerical integration is the Newton-Cotes forward polynomial, particularly the polynomial of degrees 1, 2 and 3;

i.e;

$$\int_a^b f(x)dx \cong \int_a^b P_n(x_s)dx$$

- Let us now develop our three important Newton-Cotes formulas. During the integration, we will need to change the variable of integration from  $x$  to  $s$ , since our polynomials are expressed in terms of  $s$ . Observe that

$$s = \frac{x - x_0}{h} \iff dx = h \cdot ds \implies$$

For  $n = 1$ , we have

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &\cong \int_{x_0}^{x_1} (f_0 + s\Delta f_0) dx = h \int_0^1 (f_0 + s\Delta f_0) ds \\ &= \left[ hf_0 s + h\Delta f_0 \frac{s^2}{2} \right]_0^1 = h \left( f_0 + \frac{1}{2} \Delta f_0 \right) \\ &= \frac{h}{2} [2f_0 + (f_1 - f_0)] = \frac{h}{2} (f_0 + f_1) \end{aligned}$$

$$\begin{aligned} x &= x_0 \\ &\iff \\ s &= \frac{x - x_0}{h} = 0 \end{aligned}$$

$$\begin{aligned} x &= x_1 = x_0 + h \\ &\iff \\ s &= \frac{x_0 + h - x_0}{h} = 1 \end{aligned}$$

For  $n = 2$ , we have

$$\begin{aligned}
 x &= x_2 = x_0 + 2h \\
 &\Downarrow \\
 s &= \frac{x_0 + 2h - x_0}{h} = 2
 \end{aligned}$$

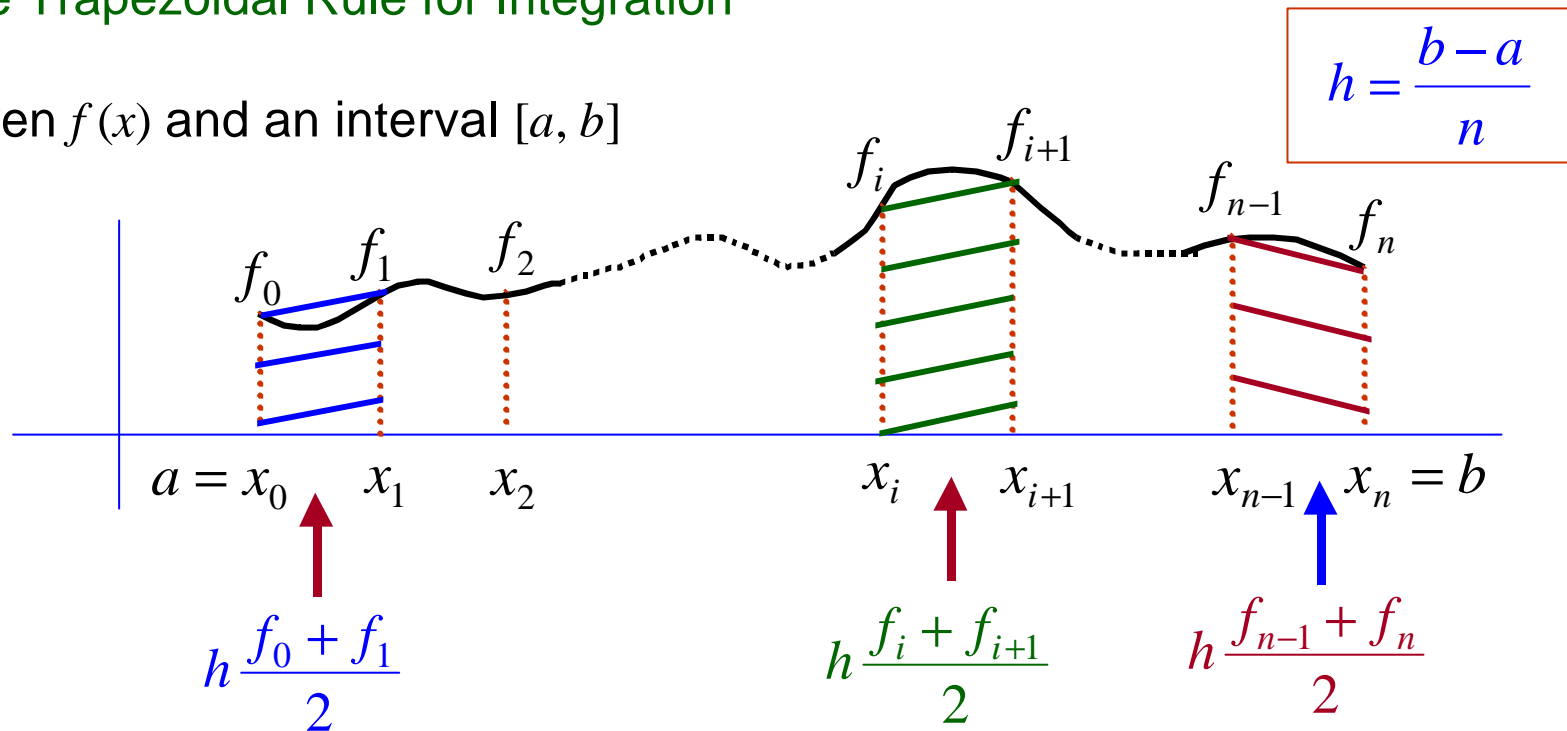
$$\begin{aligned}
 \int_{x_0}^{x_2} f(x) dx &\cong \int_{x_0}^{x_2} \left( f_0 + s\Delta f_0 + \frac{s(s-1)}{2} \Delta^2 f_0 \right) dx \\
 &= h \int_0^2 \left( f_0 + s\Delta f_0 + \frac{s(s-1)}{2} \Delta^2 f_0 \right) ds \\
 &= \left[ hf_0 s + h\Delta f_0 \frac{s^2}{2} + h\Delta^2 f_0 \left( \frac{s^3}{6} - \frac{s^2}{4} \right) \right]_0^2 \\
 &= h \left( 2f_0 + 2\Delta f_0 + \frac{1}{3} \Delta^2 f_0 \right) \\
 &= \frac{h}{3} (f_0 + 4f_1 + f_2)
 \end{aligned}$$

For  $n = 3$ , we have

$$\int_{x_0}^{x_3} f(x) dx \cong \frac{3h}{8} (f_0 + 3f_1 + 3f_2 + f_3)$$

## The Trapezoidal Rule for Integration

Given  $f(x)$  and an interval  $[a, b]$



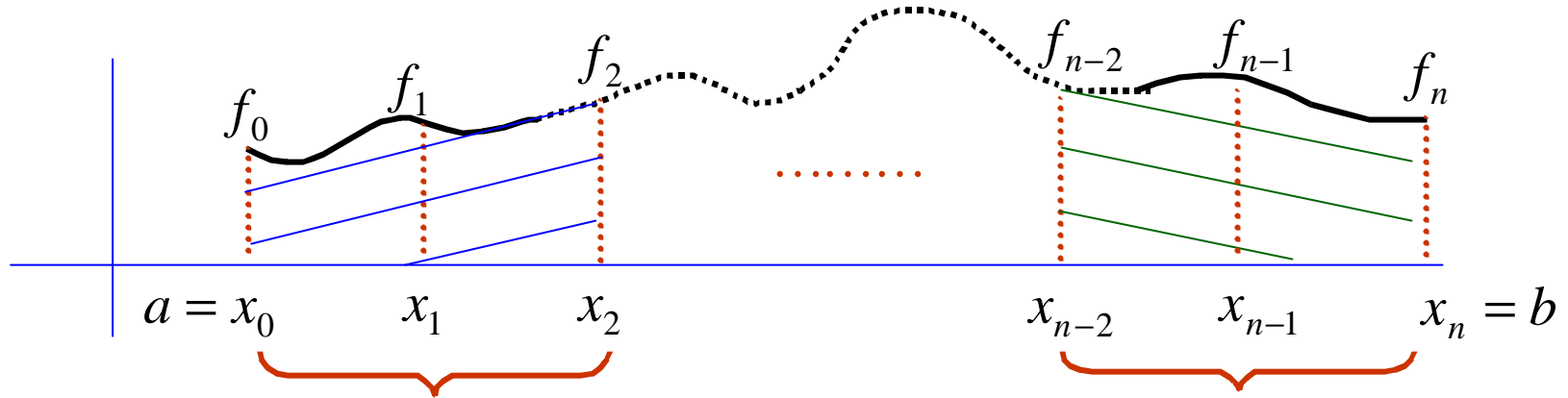
Thus,

$$\begin{aligned} \int_a^b f(x) dx &\cong h \frac{f_0 + f_1}{2} + h \frac{f_1 + f_2}{2} + \dots + h \frac{f_i + f_{i+1}}{2} + \dots + h \frac{f_{n-1} + f_n}{2} \\ &= \frac{h}{2} (f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n) \end{aligned}$$

# The Simpson's $\frac{1}{3}$ Rule for Integration

$$h = \frac{b-a}{n}, \quad n \text{ is even}$$

Given  $f(x)$  and an interval  $[a, b]$



Newton-Cotes:  $\frac{h}{3}(f_0 + 4f_1 + f_2)$       .....       $\frac{h}{3}(f_{n-2} + 4f_{n-1} + f_n)$

Thus,

$$\int_a^b f(x) dx \cong \frac{h}{3}(f_0 + 4f_1 + f_2) + \dots + \frac{h}{3}(f_{n-2} + 4f_{n-1} + f_n)$$

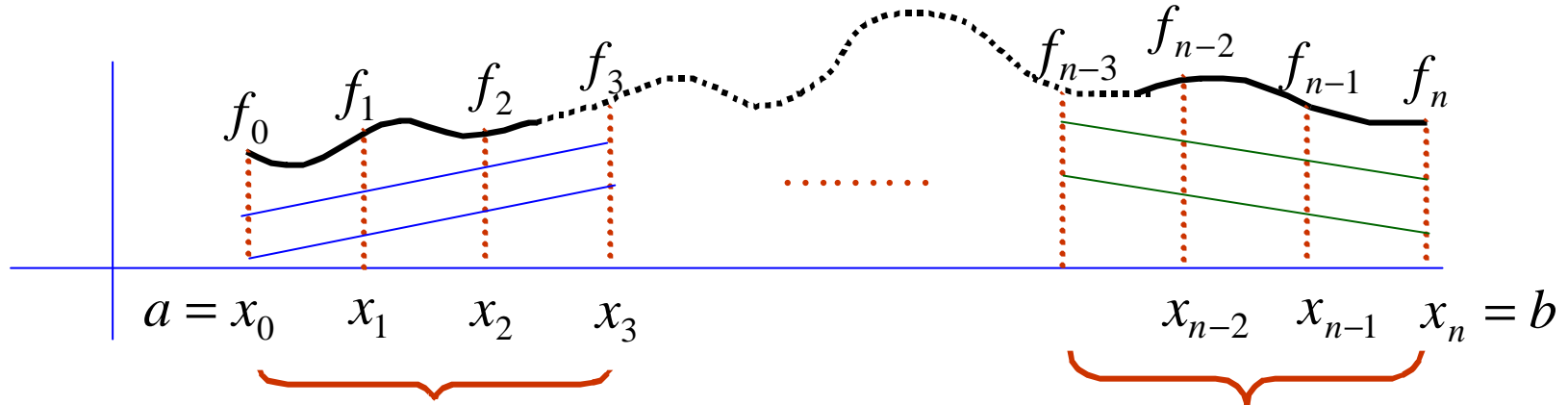
$$= \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \dots + 2f_{n-2} + 4f_{n-1} + f_n)$$



# The Simpson's $\frac{3}{8}$ Rule for Integration

$$h = \frac{b-a}{n}, \quad n = 3m$$

Given  $f(x)$  and an interval  $[a, b]$



Newton-Cotes:  $\frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) \quad \dots \quad \frac{3h}{8}(f_{n-3} + 3f_{n-2} + 3f_{n-1} + f_n)$

Thus,

$$\begin{aligned} \int_a^b f(x) dx &\cong \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) + \dots + \frac{3h}{8}(f_{n-3} + 3f_{n-2} + 3f_{n-1} + f_n) \\ &= \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + 2f_3 + 3f_4 + \dots + 2f_{n-3} + 3f_{n-2} + 3f_{n-1} + f_n) \end{aligned}$$

Example: Evaluate

$$\int_1^2 x^2 \cos x dx, \quad f(x) = x^2 \cos x \quad \text{with} \quad h = \frac{2-1}{6} = \frac{1}{6}. \quad [a,b] = [1, 2]$$

True value:

$$\begin{aligned} \int_1^2 x^2 \cos x dx &= \int_1^2 x^2 d \sin x = (x^2 \sin x) \Big|_1^2 - \int_1^2 \sin x \times 2x dx = 2.7957 + 2 \int_1^2 x d \cos x \\ &= 2.7957 + 2x \cos x \Big|_1^2 - 2 \int_1^2 \cos x dx = 0.0505 - 2 \sin x \Big|_1^2 = -0.0851 \end{aligned}$$

Trapezoidal Rule:

$$\begin{aligned} \int_1^2 x^2 \cos x dx &= \frac{h}{2} [0.5403 + 2 \times 0.5352 + 2 \times 0.4182 + 2 \times 0.1592 - 2 \times 0.2659 - 2 \times 0.8723 - 1.6646] \\ &= -0.09796 \end{aligned}$$

Simpson  $\frac{1}{3}$  Rule:

$$\begin{aligned} \int_a^b f(x) dx &= \frac{h}{3} [f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + 4f_5 + f_6] \\ &= \frac{1}{18} [0.5403 + 4 \times 0.5352 + 2 \times 0.4182 + 4 \times 0.1592 - 2 \times 0.2659 - 4 \times 0.8723 - 1.6646] \\ &= -0.08507 \end{aligned}$$

Simpson's  $\frac{3}{8}$  rule:

$$\begin{aligned}\int_a^b f(x)dx &= \frac{3h}{8} [f_0 + 3f_1 + 3f_2 + 2f_3 + 3f_4 + 3f_5 + f_6] \\ &= \frac{3}{48} [0.5403 + 3 \times 0.5352 + 3 \times 0.4182 + 2 \times 0.1592 \\ &\quad - 3 \times 0.2659 - 3 \times 0.8723 - 1.6646] \\ &= -0.08502\end{aligned}$$

Simpson's  $\frac{1}{3}$  rule gives the best result!

In general, Simpson's  $\frac{3}{8}$  rule would give the best results.

## NUMERICAL SOLUTIONS TO ORDINARY DIFFERENTIAL EQUATIONS

If the equation contains derivatives of an  $n$ -th order, it is said to be an  $n$ -th order differential equation. For example, a second-order equation describing the oscillation of a weight acted upon by a spring, with resistance motion proportional to the square of the velocity, might be

$$\frac{d^2 x}{dt^2} + 4\left(\frac{dx}{dt}\right)^2 + 0.6x = 0$$

where  $x$  is the displacement and  $t$  is time.

The solution to a differential equation is the function that satisfies the differential equation and that also satisfies certain initial conditions on the function. The analytical methods are limited to a certain special forms of the equations. Elementary courses normally treat only linear equations with constant coefficients.

Numerical methods have no such limitations to only standard forms. We obtain the solution as a tabulation of the values of the function at various values of the independent variable, however, and not as a functional relationship.

Our procedure will be to explore several methods of solving first-order equations, and then to show how these same methods can be applied to systems of simultaneous first-order equations and to higher-order differential equations. We will use the following form

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

for our typical first-order equation.

## THE TAYLOR-SERIES METHOD

The Taylor-series method serves as an introduction to the other techniques we will study although it is not strictly a numerical method. Consider the example problem

$$\frac{dy}{dx} = -2x - y, \quad y(0) = -1, x_0 = 0$$

(This particularly simple example is chosen to illustrate the method so that you can really check the computational work. The analytical solution,

$$y(x) = -3e^{-x} - 2x + 2$$

is obtained immediately by application of standard methods and will be compared with our numerical results to show the error at any step.)

## Taylor Series Expansion:

We develop the relation between  $y$  and  $x$  by finding the coefficients of the Taylor series expanded at  $x_0$

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

If we let  $x - x_0 = h$ , we can write the series as

$$y(x) = y(x_0) + y'(x_0)h + \frac{y''(x_0)}{2!}h^2 + \frac{y'''(x_0)}{3!}h^3 + \dots$$

## Iterative Procedure:

Since  $y(x_0)$  is our initial condition, the first term is known from the initial condition  $y(x_0) = -1$ . We get the coefficient of the second term by substituting  $x = 0$ ,  $y = -1$  in the equation for the first derivative

$$y'(x_0) = \left. \frac{dy}{dx} \right|_{x=x_0} = (-2x - y) \Big|_{x=x_0} = -2x_0 - y(x_0) = -2(0) - (-1) = 1$$

Similarly, we have

$$y''(x) = \frac{d}{dx} \left( \frac{dy}{dx} \right) = \frac{d}{dx} (-2x - y) = -2 - y'(x) \Rightarrow y''(x_0) = -3$$

$$\Rightarrow y'''(x_0) = 3 \qquad \Rightarrow y^{(4)}(x_0) = -3$$

We then write our series solution for  $y$ , letting  $x = h$  be the value at which we wish to determine  $y$ :

$$y(h) = -1 + 1.0h - 1.5h^2 + 0.5h^3 - 0.125h^4 + \text{error term}$$

Here shown is a case whose function is so simple that the derivatives of different orders can be obtained easily. However, the differentiation of  $f(x,y)$  could be very messy, say, those of  $x / (y - x^2)$ .



## EULER METHOD

As shown previously, the Taylor-series method may be awkward to apply if the derivatives becomes complicated and in this case the error is difficult to determine. In fact, we may only need a few terms of the Taylor series expansion for good accuracy if we make  $h$  small enough. The Euler method follows this idea to the extreme for first-order differential equations: it uses only the first two terms of the Taylor series!

### Iterative Procedure:

Suppose that we have chosen  $h$  small enough that we may truncate after the first-derivative term. Then

$$y(x_0 + h) = y(x_0) + y'(x_0)h + \frac{y''(\mathbf{z})}{2}h^2$$


where we have written the usual form of the error term for the truncated Taylor-series.

The Euler Method Iterative Scheme is given by

$$\begin{cases} y'_n = f(x_n, y_n), & y_0 = y(x_0) \\ y_{n+1} = y_n + h \cdot y'_n \end{cases}$$

Example: Using Euler Method with  $h = 0.1$ , find solution to the following o.d.e.

$$\frac{dy}{dx} = f(x, y) = -2x - y, \quad y(0) = -1, x_0 = 0$$

  $\begin{cases} y'_n = -2x_n - y_n, & y_0 = -1, & x_0 = 0 \\ y_{n+1} = y_n + 0.1 \cdot (-2x_n - y_n) = 0.9y_n - 0.2x_n \end{cases}$

$$y_1 = 0.9y_0 - 0.2x_0 = -0.9 \quad (-0.9145)$$

$$y_2 = 0.9y_1 - 0.2 \times 0.1 = -0.83 \quad (-0.8562)$$

$$y_3 = 0.9y_2 - 0.2 \times 0.2 = -0.787 \quad (-0.8225)$$

$$y_4 = 0.9y_3 - 0.2 \times 0.3 = -0.7683 \quad (-0.8110)$$

} (•) are true values

Example (cont.): Let us choose  $h = 0.001$

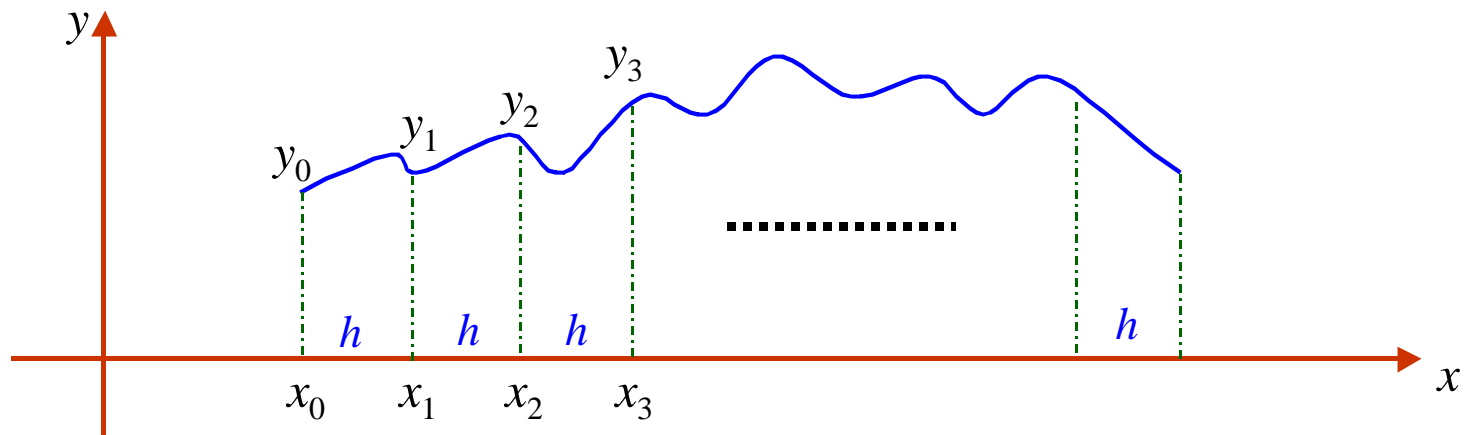


$$\begin{cases} y'_n = -2x_n - y_n, & y_0 = -1, & x_0 = 0 \\ y_{n+1} = y_n + 0.001 \cdot (-2x_n - y_n) = 0.999y_n - 0.002x_n \end{cases}$$

$$y_1 = 0.999(-1) - 0 = -0.999 \quad (-0.999002)$$

$$y_2 = 0.999(-0.999) - 0.002 \times 0.001 = -0.998003 \quad (-0.998006)$$

Quite accurate, right? What is the price we pay for accuracy? Consider  $y(10)$ , for  $h = 0.1$ , we need to compute it in 100 steps. For  $h = 0.001$ , we will have to calculate it in 10000 steps. No free lunch as usual.



## THE MODIFIED EULER METHOD

In the Euler method, we use the slope at the beginning of the interval,  $y'_n$  to determine the increment to the function. This technique would be correct only if the function were linear. What we need instead is the correct average slope within the interval. This can be approximated by the mean of the slopes at both ends of the interval.

### Modified Euler Iteration:

Given an o.d.e.

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

The modified Euler iteration is:

$$\left\{ \begin{array}{l} y'_n = f(x_n, y_n) \quad y_0 = y(x_0) \\ z_{n+1} = y_n + hy'_n \\ z'_{n+1} = f(x_{n+1}, z_{n+1}) \\ y_{n+1} = y_n + \frac{y'_n + z'_{n+1}}{2} h \end{array} \right.$$

The key idea is to fine-tune  $y'_n$  by using  $\frac{y'_n + z'_{n+1}}{2}$

**Example: Solve o.d.e.**  $\frac{dy}{dx} = -2x - y$ ,  $y(0) = -1$ ,  $x_0 = 0$  **with**  $h = 0.1$

$$\text{Step 1: } \begin{cases} y'_0 = f(x_0, y_0) = -2x_0 - y_0 = -2 \times 0 - (-1) = 1 \\ z_1 = y_0 + hy'_0 = (-1) + 0.1 \times 1 = -0.9 \\ z'_1 = f(x_1, z_1) = -2 \times 0.1 - (-0.9) = 0.7 \\ y_1 = y_0 + \frac{y'_0 + z'_1}{2} h = -1 + \frac{1 + 0.7}{2} \times 0.1 = -0.915 \quad (-0.9145) \end{cases}$$

$$\text{Step 2: } \begin{cases} y'_1 = f(x_1, y_1) = -2x_1 - y_1 = -2 \times 0.1 + 0.915 = 0.715 \\ z_2 = y_1 + hy'_1 = -0.915 + 0.1 \times 0.715 = -0.8435 \\ z'_2 = f(x_2, z_2) = -2x_2 - z_2 = -0.4 + 0.8435 = 0.4435 \\ y_2 = y_1 + \frac{h}{2}(y'_1 + z'_2) = -0.915 + 0.05(0.715 + 0.4435) = -0.8571 \quad (-0.8562) \end{cases}$$

$$\text{Step 3: } \begin{cases} y'_2 = f(x_2, y_2) = -2x_2 - y_2 = -2 \times 0.2 + 0.8571 = 0.4571 \\ z_3 = y_2 + hy'_2 = -0.8571 + 0.1 \times 0.4571 = -0.8114 \\ z'_3 = f(x_3, z_3) = -2x_3 - z_3 = -0.6 + 0.8114 = 0.2114 \\ y_3 = y_2 + \frac{h}{2}(y'_2 + z'_3) = -0.8571 + 0.05(0.4571 + 0.2114) = -0.8237 \quad (-0.8225) \end{cases}$$

## THE RUNGE-KUTTA METHODS

The two numerical methods of the last two sections, though not very impressive, serve as a good approximation to our next procedures. While we can improve the accuracy of those two methods by taking smaller step sizes, much greater accuracy can be obtained more efficiently by a group of methods named after two German mathematicians, Runge and Kutta. They developed algorithms that solve a differential equation efficiently and yet are the equivalent of approximating the exact solution by matching the first  $n$  terms of the Taylor-series expansion. We will consider only the fourth- and fifth-order Runge-Kutta methods, even though there are higher-order methods. Actually, the modified Euler method of the last section is a second-order Runge-Kutta method.

## Fourth-Order Runge-Kutta Method:

**Problem:** To solve differential equation,

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

**Algorithm:**

$$\left\{ \begin{array}{l} y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad y_0 = y(x_0) \quad \text{with} \\ k_1 = hf(x_n, y_n) \\ k_2 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right) \\ k_3 = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2\right) \\ k_4 = hf(x_n + h, y_n + k_3) \end{array} \right.$$

**Proof:** 1) Read textbook, or 2) forget about it.



**Example:** Solve the following o.d.e. using Fourth-Order Runge-Kutta Method

$$\frac{dy}{dx} = -2x - y, \quad y(0) = -1, \quad x_0 = 0, \quad h = 0.1$$

**Step 1:**

$$k_1 = hf(x_0, y_0) = 0.1(-2 \times 0 + 1) = 0.1$$

$$k_2 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right) = 0.1\left(-2 \times \frac{1}{2} \times 0.1 - (-1) - \frac{1}{2} \times 0.1\right) = 0.85 \times 0.1 = 0.085$$

$$k_3 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2\right) = 0.1 \times \left(-2 \times \frac{1}{2} \times 0.1 + 1 - \frac{1}{2} \times 0.085\right) = 0.08575$$

$$k_4 = hf(x_0 + h, y_0 + k_3) = 0.1 \times (-2 \times 0.1 + 1 - 0.08575) = 0.071425$$

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = -0.9145125 \quad \boxed{(-0.9145123)}$$

true values

**Step 2:**

$$k_1 = 0.0715 \quad k_2 = 0.0579 \quad k_3 = 0.0586 \quad k_4 = 0.0456 \quad y_2 = -0.8562 \quad \boxed{(-0.85619)}$$

# Numerical Solutions to Partial Differential Equations

## Introduction

General partial differential equations (PDE) is hard to solve! We shall only treat some special types of PDE's that are useful and easier to be solved.

## Classification of 2nd order quasi-linear PDE's

General form

$$a(x, y) \frac{\partial^2 u}{\partial x^2} + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y} + c(x, y) \frac{\partial^2 u}{\partial y^2} = F \left( x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right)$$

**quasi-linear** — linear in highest order derivatives

$u = u(x, y)$  — unknown functions to be solved.

$x, y$  — independent variable  $x$  and  $y$ .

## Some standard notations

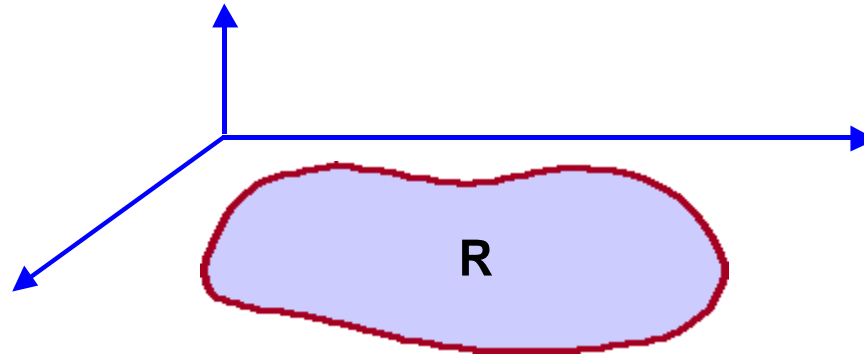
$$u_x := \frac{\partial u}{\partial x}, \quad u_y = \frac{\partial u}{\partial y}, \quad u_{xx} := \frac{\partial^2 u}{\partial x^2}, \quad u_{yy} := \frac{\partial^2 u}{\partial y^2}, \quad u_{xy} := \frac{\partial^2 u}{\partial x \partial y}$$

## Types of equations

Type	Condition	Example
elliptic	$b^2 - ac < 0$	Laplace equation: $u_{xx} + u_{yy} = 0$ $\{a = 1, b = 0, c = 1\}$
parabolic	$b^2 - ac = 0$	Heat equation: $k^2 u_{xx} = u_t$ $\{a = k^2, b = 0, c = 0\}$
hyperbolic	$b^2 - ac > 0$	Wave equation: $A^2 u_{xx} = u_{tt}$ $\{a = A^2, b = 0, c = -1\}$

Methods of solutions depended on the type of equations.

Geometrically



Type may not be constant over  $\mathbf{R}$  because  $a, b, c$  can vary over  $\mathbf{R}$ , e.g, elliptic in one part of  $\mathbf{R}$  and parabolic in the other part of  $\mathbf{R}$ .

Example:

$$\underbrace{(\sin^2 y)}_a u_{xx} + u_{yy} = -[u + (\sin y)u_x]$$

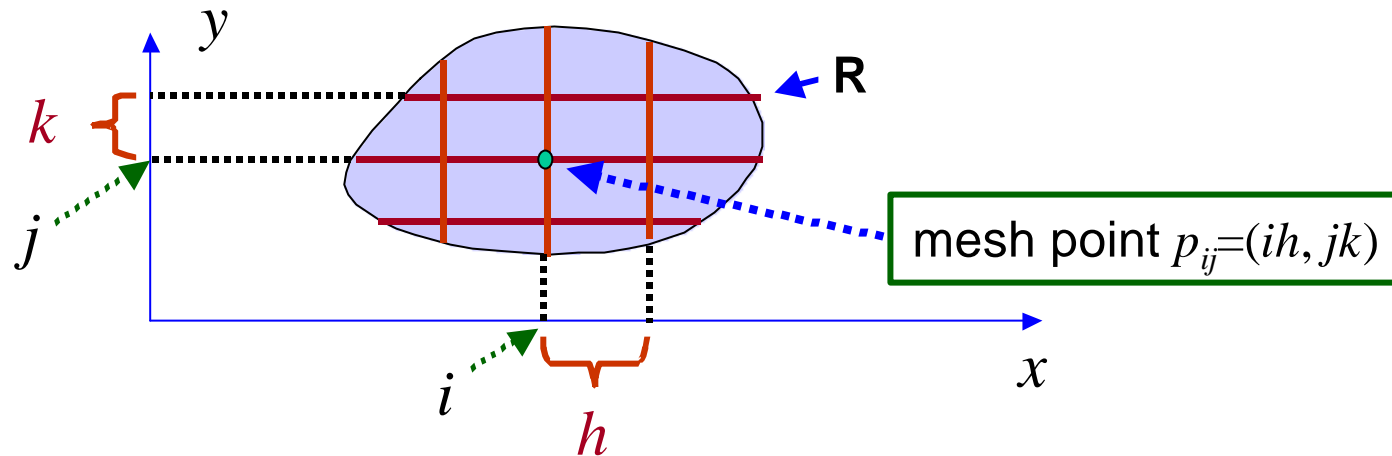
$$\mathbf{R} : -3 \leq x \leq 3, \quad -3 \leq y \leq 3$$

$$\Rightarrow a = \sin^2 y, \quad b = 0, \quad c = 1 \Rightarrow b^2 - ac = -\sin^2 y \leq 0 \quad \begin{cases} = 0 & \text{parabolic} \\ < 0 & \text{elliptic} \end{cases}$$

74

## General Approach to the Solutions of PDEs

Step 1: Define a grid on  $\mathbf{R}$  with “mesh points”



Step 2: Approximate derivatives at mesh points by central difference quotients

$$u_x(ih, jk) = \frac{u_{i+1,j} - u_{i-1,j}}{2h}, \quad u_y(ih, jk) = \frac{u_{i,j+1} - u_{i,j-1}}{2k},$$
$$u_{xx}(ih, jk) = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad u_{yy}(ih, jk) = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2}$$

These will bring a PDE to a difference equation relating  $u_{ij}$  to its neighbouring points in the grid.

For example,

$$u_{xx} + u_{yy} = 0 \Rightarrow \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = 0$$

$$k^2 u_{i+1,j} + k^2 u_{i-1,j} + h^2 u_{i,j+1} + h^2 u_{i,j-1} - 2(k^2 + h^2) u_{i,j} = 0$$

Step 3: Arrange the resulting difference equation into a system of linear equations

$$\begin{bmatrix} * & * & \dots & * \\ * & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & * \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ * \end{bmatrix} = \begin{bmatrix} * \\ * \\ \vdots \\ * \end{bmatrix}$$

Taking into consideration of boundary conditions and solve it for  $u_{11}, u_{12}, \dots$

Step 4: change grid size for a more accurate approximation.

$$h \rightarrow \frac{h}{2} \rightarrow \frac{h}{4} \dots \quad k \rightarrow \frac{k}{2} \rightarrow \frac{k}{4} \dots$$

## Solution to Elliptic Type's PDE

The general approach will be followed to solve these types of problems by taking into account various kinds of boundary conditions in form of the system of linear equations. We will illustrate this using the following PDE:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = u_{xx} + u_{yy} = f(x, y) \equiv 0$$

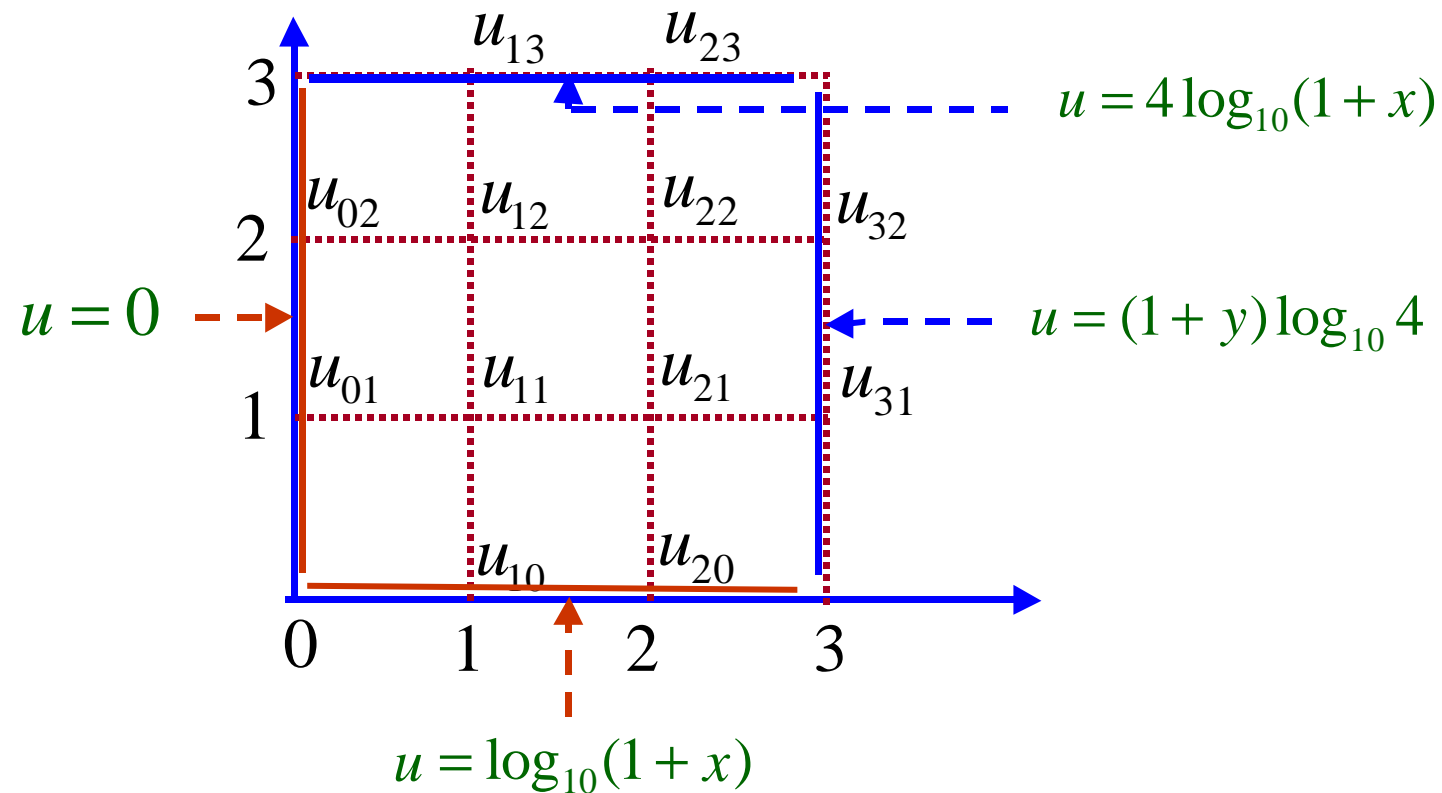
$$\mathbf{R} = \{ (x, y), 0 \leq x \leq 3, 0 \leq y \leq 3 \}$$

**Boundary condition**  $u(x, y) = (1 + y) \log_{10}(1 + x)$

We follow the step-by-step procedure given in the previous section.

Step 1: Define a grid along with an order of mesh-points inside  $\mathbf{R}$ . (We have to be clear about  $\mathbf{R}$  and  $h, k$ )

First, let us start with a crude grid  $h = k = 3/N$ , for  $N=3 \rightarrow h = k = 1$



knowns:  $u_{13}, u_{23}, u_{32}, u_{31}, u_{02}, u_{01}, u_{10}, u_{20}$

unknowns:  $u_{11}, u_{12}, u_{21}, u_{22}$  78



Step 2: Approximate derivatives at mesh-points

$$u_{xx} + u_{yy} = f(x, y) = 0 \Rightarrow$$

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = f_{ij} \equiv 0 \quad i=1,2,3 \quad j=1,2,3$$

$$\Rightarrow u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0$$

At mesh-point  $(i, j)$  where  $u_{i,j}$  is unknown:

@ (1,1):  $0 = u_{01} - 2u_{11} + u_{21} + u_{10} - 2u_{11} + u_{12} = -4u_{11} + u_{01} + u_{21} + u_{10} + u_{12}$

@ (2,1):  $0 = u_{11} + u_{31} + u_{20} + u_{22} - 4u_{21}$

@ (1,2):  $0 = u_{11} + u_{13} + u_{02} + u_{22} - 4u_{12}$

@ (2,2):  $0 = u_{21} + u_{23} + u_{12} + u_{32} - 4u_{22}$

Boundary values  
are known

Step 3: Arrange the equation into matrix form

$$\begin{pmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 1 & 0 & -4 & 1 \\ 0 & 1 & 1 & -4 \end{pmatrix} \begin{pmatrix} u_{12} \\ u_{22} \\ u_{11} \\ u_{21} \end{pmatrix} = \begin{pmatrix} -1.204 \\ -3.714 \\ -0.301 \\ -1.681 \end{pmatrix}$$

Solve the equations for

$$\begin{pmatrix} u_{12} \\ u_{22} \\ u_{11} \\ u_{21} \end{pmatrix} = \begin{pmatrix} 0.756 \\ 1.336 \\ 0.483 \\ 0.875 \end{pmatrix}$$

Step 4: Refine the step-size by choosing smaller  $h, k$ .

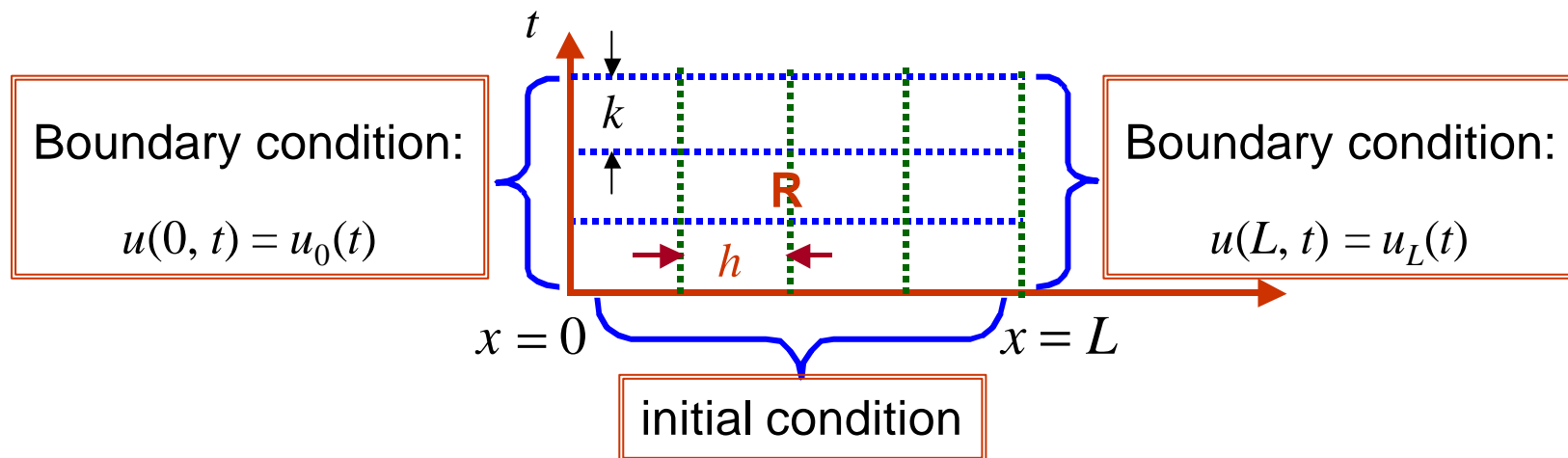
## Parabolic and Hyperbolic Types

**Parabolic:** Example — heat equation

$$Du_{xx} = u_t, \quad \text{where } D \text{ is heat diffusion coefficient}$$

**Hyperbolic:** Example — wave equation

$$C^2 u_{xx} = u_{tt} \quad \text{where } C^2 \text{ is wave propagation velocity}$$



We will use parabolic type  $u_{xx} = u_t$  to illustrate the solution method, which carries over to the hyperbolic type as well!

Notations:

$$x_i = i \cdot h, \quad i = 0, 1, \dots, N, \quad L = Nh \quad \Rightarrow \quad h = \frac{L}{N}$$

$$t_j = j \cdot k, \quad j = 0, 1, \dots$$

$$u_{i,j} = u(x_i, t_j)$$

$$u_t(x_i, t_j) = \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} = \frac{u_{i,j+1} - u_{i,j}}{k}$$

$$u_{xx}(x_i, t_j) = \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j))}{h^2} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

Then

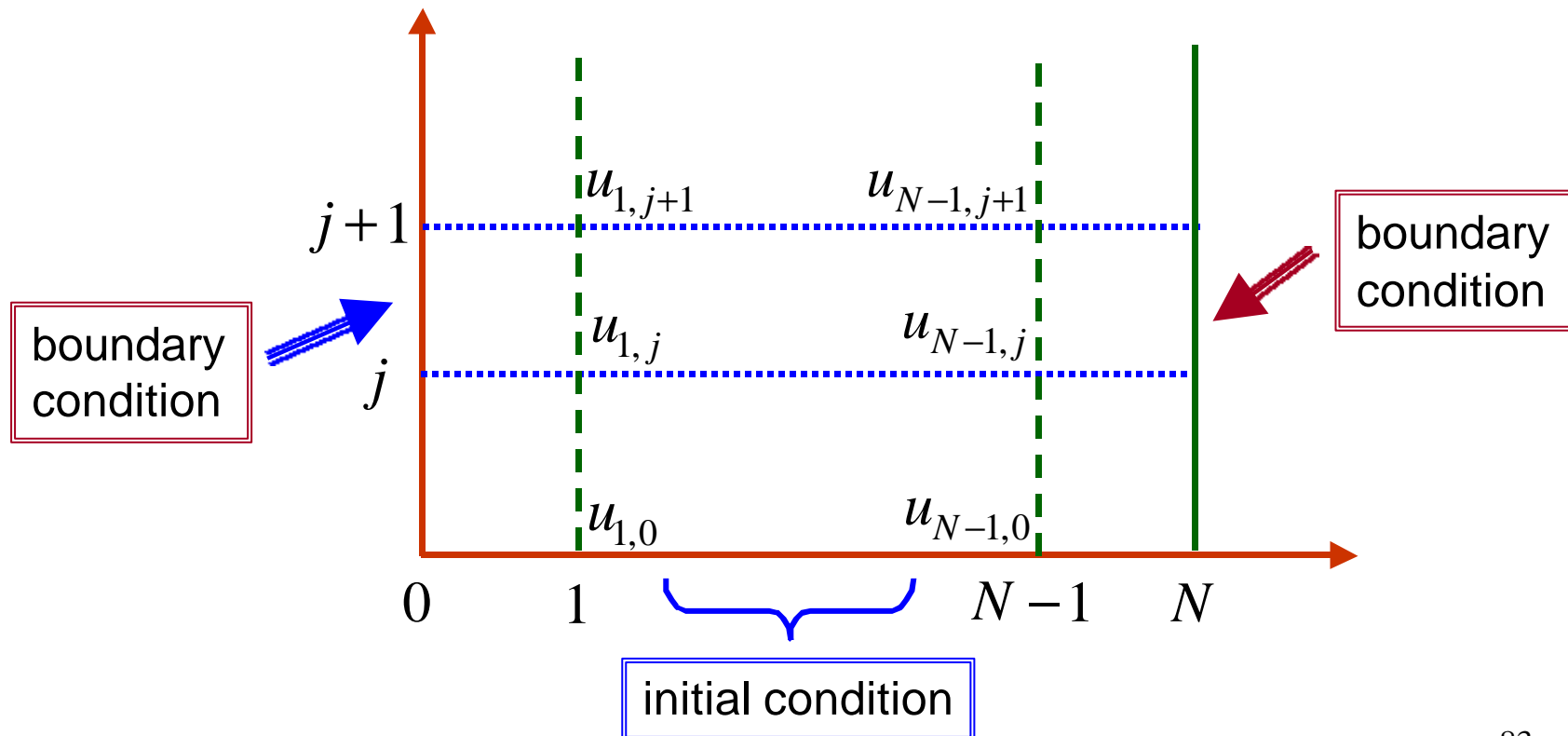
$$u_t = D u_{xx} \quad \Rightarrow \quad u_{i,j+1} - u_{i,j} = \frac{kD}{h^2} (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \quad (\clubsuit)$$

To solve the equation, we start with  $j = 0$ , then  $u_{i,0}$ 's are given as initial conditions and can be used to solve for  $u_{i,1}$ ,  $i = 1, \dots, N - 1$

Rewrite equation ( $\clubsuit$ ) as

$$u_{i,j+1} = g \cdot u_{i+1,j} + (1 - 2g)u_{i,j} + g \cdot u_{i-1,j} \quad \text{with } g = \frac{kD}{h^2}$$

In general, we can solve for  $u_{i,j+1}$ ,  $i = 1, \dots, N$ , if we know the  $j$ -th row.



Example: Solve the following boundary value problem,

$$u_t = u_{xx}, \quad 0 \leq x \leq 1, \quad D = 1, \quad L = 1$$

initial condition:  $u(x,0) = \sin \frac{\pi x}{2}$       boundary condition:  $u(0,t) = 0, \quad u(1,t) = 1$

We choose  $N = 3$  and hence  $h = 1/3$  and choose two different  $k$ :

$k = 0.05 \mapsto g = 0.45$					$k = 0.1 \mapsto g = 0.9$		
$jk \backslash i$	$u_{0,j}$	$u_{1,j}$	$u_{2,j}$	$u_{3,j}$	$jk \backslash i$	$u_{1,j}$	$u_{2,j}$
0.00	0	0.500	0.866	1	0.0	0.500	0.866
0.05	0	0.434	0.762	1	0.1	0.379	0.657
0.10	0	0.387	0.724	1	0.2	0.288	0.716
0.15	0	0.364	0.696	1	0.3	0.414	0.587
0.20	0	0.350	0.684	1	0.4	0.197	0.803
0.25	0	0.343	0.676	1	0.5	0.505	0.435
0.30	0	0.338	0.672	1	0.6	-0.061	1.061
0.35	0	0.336	0.667	1	0.7	1.003	-0.003
0.40	0	0.335	0.668	1	0.8	-0.804	1.805
0.45	0	0.334	0.668	1	0.9	2.269	-1.269
0.50	0	0.333	0.667	1	1.0	-2.957	3.457

unstable  
case

## A short discussion about hyperbolic type PDE:

$$\text{PDE:} \quad u_{tt} = C^2(x,t)u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0$$

$$\text{Initial conditions:} \quad u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x)$$

$$\text{Boundary conditions:} \quad u(0,t) = g_0(t), \quad u(1,t) = g_1(t)$$

Following the usual procedure, we obtain an approximation:

$$u_{i,j+1} = (2 - 2g^*)u_{i,j} + g^* \cdot u_{i-1,j} + g^* \cdot u_{i+1,j} - u_{i,j-1} \quad \text{with } g^* = \frac{C^2 k^2}{h^2}$$

Note that at  $j=0$ , we have to deal with  $u_{i,-1}$ , which are not readily available.

Thus, we will have to compute these terms first.

$$u_t(x,0) = f_2(x) \quad \Rightarrow \quad u_{i,1} - u_{i,-1} = 2kf_2(x_i) \quad \Rightarrow \quad u_{i,-1} = u_{i,1} - 2kf_2(x_i)$$

The difference equation can then be solved by using the direct method, e.g,

$$\begin{aligned}u_{i,1} &= (2 - 2\mathbf{g}^*)u_{i,0} + \mathbf{g}^* u_{i-1,0} + \mathbf{g}^* u_{i+1,0} - u_{i,-1} \\ &= (2 - 2\mathbf{g}^*)u_{i,0} + \mathbf{g}^* u_{i-1,0} + \mathbf{g}^* u_{i+1,0} - u_{i,1} + 2kf_2(x_i)\end{aligned}$$



$$u_{i,1} = (1 - \mathbf{g}^*)f_1(x_i) + \frac{1}{2}\mathbf{g}^* f_1(x_{i-1}) + \frac{1}{2}\mathbf{g}^* f_1(x_{i+1}) + kf_2(x_i), \quad i = 1, 2, \dots, N-1$$

For  $j > 1$ , we still use

$$u_{i,j+1} = (2 - 2\mathbf{g}^*)u_{i,j} + \mathbf{g}^* \cdot u_{i-1,j} + \mathbf{g}^* \cdot u_{i+1,j} - u_{i,j-1}$$

The rest of computational procedure is exactly the same as that in the parabolic case.



Example: Solve

$$PDE : \quad u_{tt} = u_{xx} \quad 0 \leq x \leq 1, \quad t \geq 0$$

$$Initial\ conditions : \quad u(x,0) = f_1(x) = 0$$

$$u_t(x,0) = f_2(x) = x + \sin(\mathbf{p}x)$$

$$Boundary\ conditions : \quad u(0,t) = g_0(t) = 0$$

$$u(1,t) = g_1(t) = \frac{1}{\mathbf{p}} \sin(\mathbf{p}t)$$

Let us choose  $h = k = 0.25$  so that  $\gamma^* = 1$

Determine  $u_{i,-1}$  to start the solution or use formula on the previous page

to compute  $u_{i,1}$ ,  $i = 1, 2, 3$ , first, i.e.,

$$u_{i,1} = 0 + k \cdot f_2(x_i) = 0.25 [x_i + \sin(\delta x_i)] \Rightarrow \begin{cases} u_{1,1} = 0.239 \\ u_{2,1} = 0.375 \\ u_{3,1} = 0.364 \end{cases}$$

D.I.Y. to complete the solutions up to  $t = 2$ .