Lecture Notes for Computational Methods in Chemical Engineering (CL 701)

Sachin C. Patwardhan,

Department of Chemical Engineering,

Indian Institute of Technology, Bombay,

Mumbai 400076, India.

Contents

CHAPTER 1

Mathematical Modeling and Simulation

1. Introduction

A modern chemical plant consists of interconnected units such as heat exchangers, reactors, distillation columns, mixers etc. with high degree of integration to achieve energy efficiency. Design and operation of such complex plants is a challenging problem. Mathematical modeling and simulation is a cost effective method of designing or understanding behavior of these chemical plants when compared to study through experiments. Mathematical modeling cannot substitute experimentation, however, it can be effectively used to plan the experiments or creating scenarios under different operating conditions. Thus, best approach to solving most chemical engineering problems involves judicious combination of mathematical modeling and carefully planned experiments.

To begin with, let us look at types of problems that can arise in context of modeling and simulation. Consider a typical small chemical plant consisting of a reactor and a distillation column, which is used to separate the product as overhead (see Figure 1). The reactants, which are separated as bottom product of the distillation column, are recycled to the reactor. We can identify following problems

- **Process Design problem**

Given: Desired product composition, raw material composition and availability.

- *To Find:* Raw material flow rates, reactor volume and operating conditions (temperature, pressure etc.), distillation column configuration (feed locations and product draws), reboiler,condenser sizes and operating conditions (recycle and reflux flows, steam flow rate, operating temperatures and pressure etc.)
- **Process Retrofitting:** Improvements in the existing set-up or operating conditions

  Plant may have been designed for certain production capacity and assuming certain raw material quality. We are often required to assess whether

  - Is it possible to operate the plant at a different production rate?
  - What is the effect of changes in raw material quality?
  - Is it possible to make alternate arrangement of flows to reduce energy consumption?
- **Dynamic behavior and operability analysis:** Any plant is designed by assuming certain ideal composition of raw material quality, temperature and operating temperatures and pressures of utilities. In practice, however, it is impossible to maintain all the operating conditions exactly at the nominal design conditions. Changes in atmospheric conditions of fluctuations in steam header pressure, cooling water temperature, feed quality fluctuations, fouling of catalysts, scaling of heat transfer surfaces etc. keep perturbing the plant from the ideal operating condition. Thus, it becomes necessary to understand transient behavior of the system in order to
  - reject of effects of disturbances on the key operating variables such as product quality
  - achieve transition from one operating point to an economically profitable operating point.
  - carry out safety and hazard analysis

In order to solve process design or retrofitting problems, mathematical models are developed for each unit operation starting from first principles. Such mechanistic (or first principles) models in Chemical Engineering are combination of mass, energy and momentum balances together with associated rate equations, equilibrium relation and equations of state.

- **Mass balances**: overall, component.
- **Rate equations**: mass, heat and momentum transfer rates (constitutive equations.), rate of chemical reactions
- **Equilibrium principles** : physical(between phases) and chemical (reaction rate equilibrium).
- **Equations of state**: primarily for problems involving gases.

From mathematical viewpoint, these models can be classified into two broad classes

- *Distributed parameter model:* These models capture the relationship between the variables involved as functions of time and space.
- *Lumped parameter models*: These models lump all spatial variation and all the variables involved are treated as functions time alone.

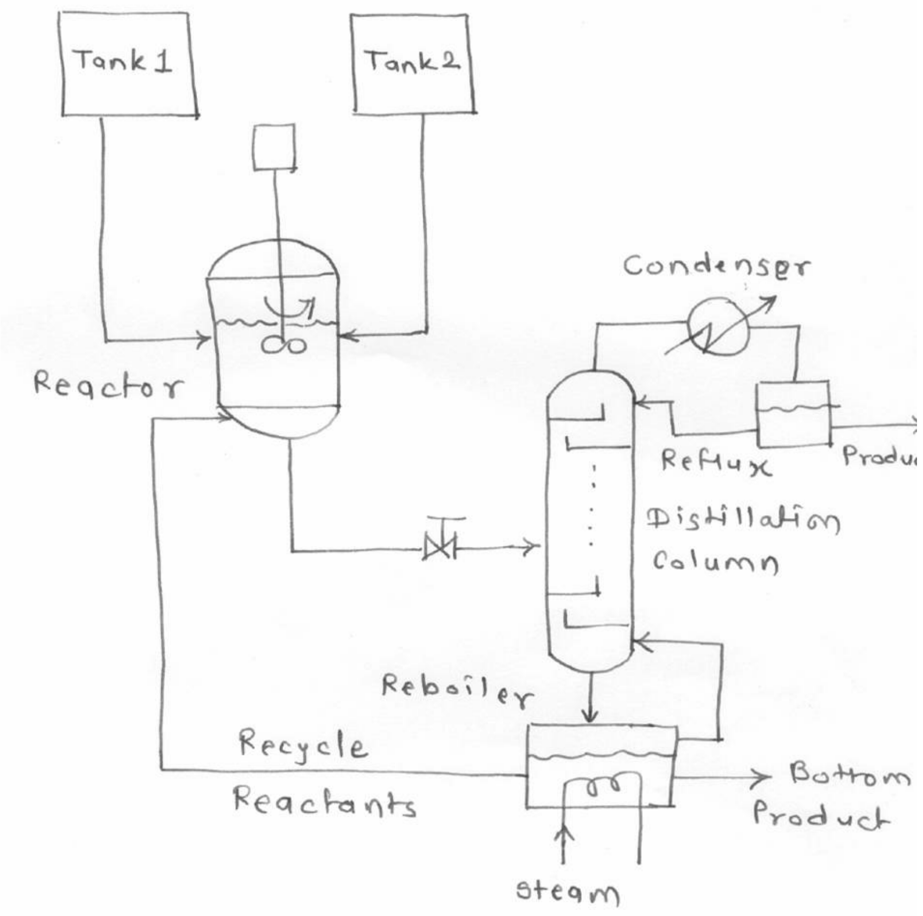


FIGURE 1. Typical processing plant: Schematic daigram

The above two classes of models together with the various scenarios under consideration give rise to different types of equation forms such as linear / non-linear algebraic equations, ordinary differential equations or partial differential equations. In order to provide motivation for studying these different kinds of equation forms, we present examples of different models in chemical engineering and derive abstract equation forms in the following section.

2. Mechanistic Models and Abstract Equation Forms

2.1. Linear Algebraic Equations. Plant wide or section wide mass balances are carried out at design stage or later during operation for keeping material audit. These models are typical examples of systems of simultaneous linear algebraic equations..

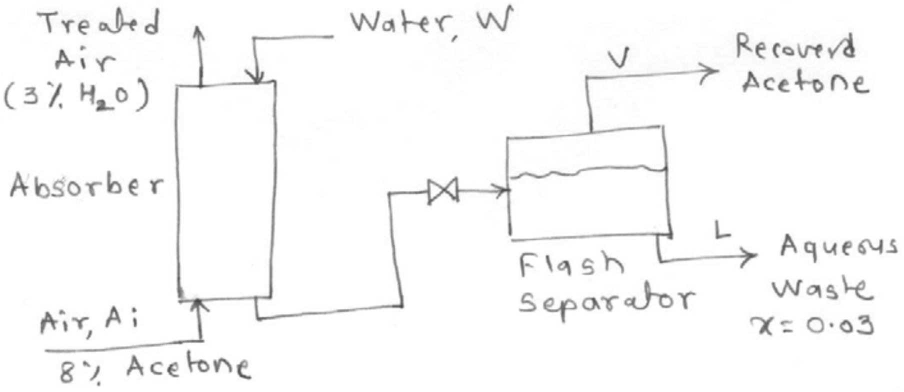


FIGURE 2

EXAMPLE 1. *Recovery of acetone from air -acetone mixture is achieved using an absorber and a flash separator (Figure 2). A model for this system is developed under following conditions*

- *All acetone is absorbed in water*
- *Air entering the absorber contains no water vapor*
- *Air leaving the absorber contains 3 mass % water vapor*

The flash separator acts as a single equilibrium stage such that acetone mass fraction in vapor and liquid leaving the flash separator is related by relation

$$y = 20.5x \tag{2.1}$$

where y mass fraction of the acetone in the vapor stream and x mass fraction of the acetone in the liquid stream. Operating conditions of the process are as follows

- *Air in flow: 600 lb /hr with 8 mass % acetone*
- *Water flow rate: 500 lb/hr*

It is required that the waste water should have acetone content of 3 mass % and we are required to determine concentration of the acetone in the vapor stream and flow rates of the product streams.

Mass Balance:

$$\textit{Air} \ : \ 0.92Ai = 0.97Ao \tag{2.2}$$

$$\textit{Acetone} \ : \ 0.08\,Ai = 0.03\,L + y\,V \tag{2.3}$$

$$Water \ : \ W = 0.03\,Ao + (1-y)V + 0.97L \tag{2.4}$$

$$\textit{Design requirement} \ : \ x = 0.03 \tag{2.5}$$

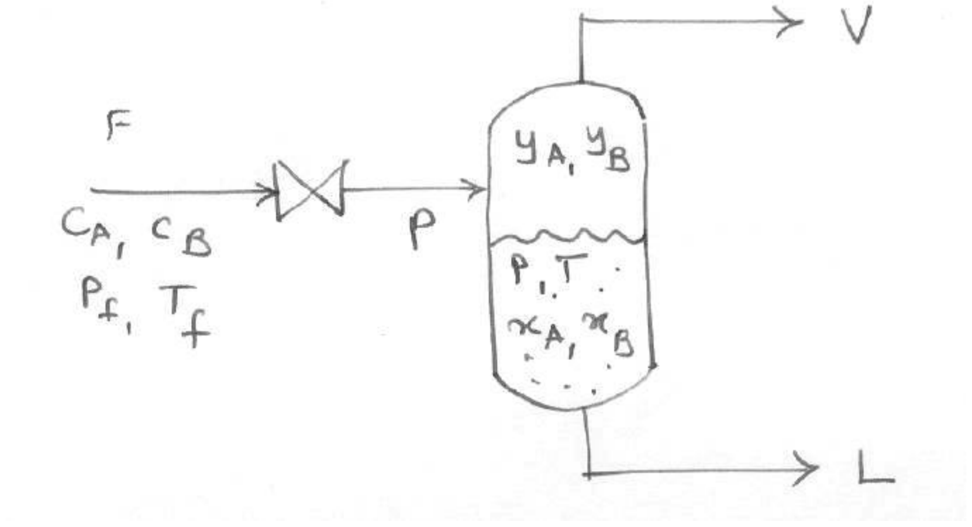


Figure 3. Flash Drum: Schematic Diagram

Equilibrium Relation:

$$y = 20.5\,x \tag{2.6}$$

$$\Rightarrow \quad y = 20.5 \times 0.03 = 0.615 \tag{2.7}$$

Substituting for all the known values and rearranging, we have

$$\begin{bmatrix} 0.97 & 0 & 0 \\ 0 & 0.03 & 0.615 \\ 0.03 & 0.385 & 0.97 \end{bmatrix} \begin{bmatrix} Ao \\ L \\ V \end{bmatrix} = \begin{bmatrix} 0.92 \times 600 \\ 0.08 \times 600 \\ 500 \end{bmatrix} \tag{2.8}$$

The above model is a typical example of system of linear algebraic equations, which have to be solved simultaneously. The above equation can be represented in abstract form set of linear algebraic equations

$$A\mathbf{x} = \mathbf{b} \tag{2.9}$$

where $\mathbf{x}$ and $\mathbf{b}$ are a $(n \times 1)$ vectors (i.e. $\mathbf{x}, \mathbf{b} \in R^n$) and A is a $(n \times n)$ matrix.

2.2. Nonlinear Algebraic Equations. Consider a stream of two components A and B at a high pressure P_f and temperature T_f as shown in Figure 3. If the P_f is greater than the bubble point pressure at T_f, no vapor will be present. The liquid stream passes through a restriction (valve) and is flashed in the drum, i.e. pressure is reduced from P_f to P. This abrupt expansion takes place under constant enthalpy. If the pressure P in the flash drum is less than the bubble point pressure of the liquid feed at T_f, the liquid will partially

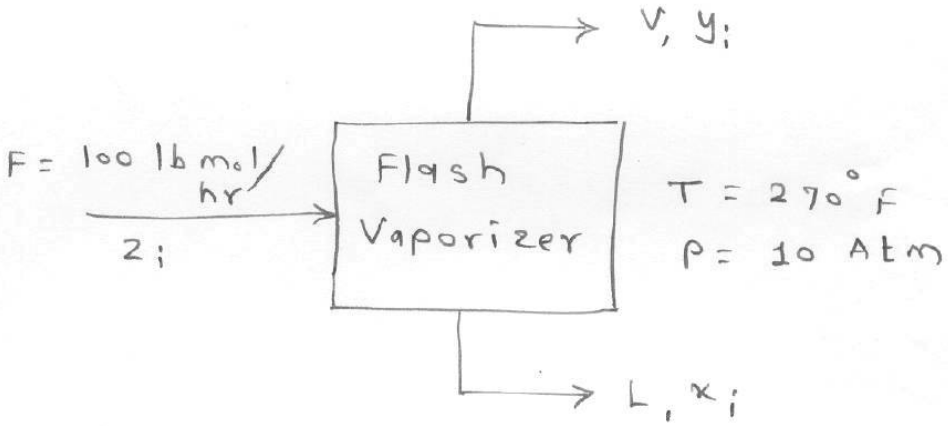


FIGURE 4

vaporize and two phases at the equilibrium with each other will be present in the flash drum. The equilibrium relationships are

- Temperature of the liquid phase = temperature of the vapor phase.
- Pressure of the liquid phase = pressure of the vapor phase.
- Chemical potential of the $i'th$ component in the liquid phase = Chemical potential of the $i'th$ component in the vapor phase

EXAMPLE 2. *Consider flash vaporization unit shown in Figure 4. A hydrocarbon mixture containing 25 mole % of n butane, 45 mole %of n -hexane is to be separated in a simple flash vaporization process operated at 10 atm. and 270^0F. The equilibrium k- values at this composition are*

Component	$\mathbf{z}_i$	$\mathbf{k}_i$
n-butane	*0.25*	*2.13*
n-pentane	*0.45*	*1.10*
n-hexane	*0.30*	*0.59*

Let x_i represent mole fraction of the component i in liquid phase and y_i represent mole fraction of the component i in vapor phase. Model equations for the flash vaporizer are

- *Equilibrium relationships*

$$k_i = y_i/x_i \qquad (i = 1, 2, 3) \tag{2.10}$$

- *Overall mass balance*

$$F = L + V \tag{2.11}$$

- *Component balance*

$$z_i * F = x_i * L + y_i * V \qquad (i = 1, 2, 3) \tag{2.12}$$

$$= x_i * L + k_i * x_i * V \tag{2.13}$$

$$\sum x_i = 1 \tag{2.14}$$

Note that this results in a set of simultaneous 5 nonlinear algebraic equations in 5 unknowns Equations (2.11-2.14) can be written in abstract form as follows

$$f_1(x_1, x_2, x_3, L, V) = 0 \tag{2.15}$$

$$f_2(x_1, x_2, x_3, L, V) = 0 \tag{2.16}$$

$$\text{............................} = 0$$

$$f_5(x_1, x_2, x_3, L, V) = 0 \tag{2.17}$$

which represent coupled nonlinear algebraic equations. These equations have to be solved simultaneously to find solution vector

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & L & V \end{bmatrix}^T \tag{2.18}$$

The above 5 equations can also be further simplified as follows

$$x_i = z_i / \left[1 + \left(\frac{V}{F} \right) (k_i - 1) \right]$$

Using $\sum x_i = 1$,*we have*

$$f\left(V/F\right) = \sum \frac{zi}{1 + \left(V/F\right)\left(ki - 1\right)} - 1 = 0 \tag{2.19}$$

In general, we encounter n nonlinear algebraic equations in n variables, which have to be solved simultaneously. These can be expressed in the following abstract form

$$f_1(x_1, x_2, x_{3,}......x_n) = 0 \tag{2.20}$$

$$f_2(x_1, x_2, x_{3,}......x_n) = 0 \tag{2.21}$$

$$\text{............................} = 0$$

$$f_n(x_1, x_2, x_{3,}......x_n) = 0 \tag{2.22}$$

Using vector notation, we can write

$$F(\mathbf{x}) = \overline{0} \qquad ; \qquad \mathbf{x} \in R^n \tag{2.23}$$

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}^T$$

where $\overline{0}$ represents $n \times 1$ zero vector. Here $F(\mathbf{x}) \in R^n$ represents n dimensional function vector defined as

$$F(\mathbf{x}) = \left[\begin{array}{cccc} f_1(\mathbf{x}) & f_2(\mathbf{x}) & ... & f_n(\mathbf{x}) \end{array} \right]^T \tag{2.24}$$

2.3. Optimization Based Formulations. Variety of modeling and design problems in chemical engineering are formulated as optimization problems.

Example 3. *Consider a simple reaction*

$$A \rightarrow B$$

modelled using the following reaction rate equation

$$-r_a = -dC_a/dt = k_o(C_a)^n \exp(\frac{-E}{RT}) \tag{2.25}$$

carried out in a batch reactor. It is desired to find the kinetic parameters k_o, E and n from the experimental data. The following data is collected from batch experiments in a reactor at different temperatures

Reaction Rate	Concentration	Temperature
$-r_{a1}$	C_{a1}	T_1
$-r_{a2}$	C_{a2}	T_2
....	....	....
$-r_{aN}$	C_{aN}	T_N

Substituting these values in the rate equation will give rise to N equations in three unknowns, which is an overdetermined set equations. Due to experimental errors in the measurements of temperature and reaction rate, it may not be possible to find a set of values of $\{k_o, E, n\}$such that the reaction rate equation is satisfied at all the data points. However one can decide to select $\{Vo, E, n\}$such that the quantity

$$\Phi = \sum_{i=1}^{N} \left[-r_{ai} \; - k_o(C_{ai})^n \exp(\frac{-E}{RT_i}) \right]^2 \tag{2.26}$$

Suppose we use $-r_{aie}$ to denote the estimated reaction rate

$$-r_{aie} = k_o \; C_{ai}^n \quad \exp(\frac{-E}{R * T_i}) \tag{2.27}$$

then, the problem is to choose parameters $\{k_o, E, n\}$ such that the sum of the square of errors between the measured and estimated rates is minimum, i.e.

$$\underset{k_o, E, n}{Min} \quad \Phi(k_o, E, n) = \sum_{i=1}^{N} \left[-r_{ai} \; - (-r_{aie})\right]^2 \tag{2.28}$$

Example 4. *Cooling water is to be allocated to three distillation columns. Up to 8 million liters per day are available, and any amount up to this limit may be use. The costs of supplying water to each equipment are*

Equip. 1: $f_1 = |1 - D_1| - 1 \quad for \quad 0 \leq D_1 \leq 2$

$\qquad = 0 \quad (otherwise)$

Equip. 2: $f_2 = -\exp(\frac{-1}{2}(D_2 - 5)^2) \quad for \quad 0 \leq D_2 \leq \infty$

Equip. 2: $f_2 = D_3^2 - 6D_3 + 8 \quad for \quad 0 \leq D_3 \leq 4$

Minimize $\Phi = \sum f_i$ *to find* $D_1, D_2,$ *and* D_3

Note that this is an example of a typical multi-dimensional optimization problem, which can be expressed in abstract form

$$\underset{\mathbf{x}}{Min} \quad \Phi(\mathbf{x}) \tag{2.29}$$

where $\mathbf{x} \in R^n$ and $f(\mathbf{x}) : R^n \rightarrow R$ is a scalar objective function. A general problem of this type may include constraints on $\mathbf{x}$ or functions of $\mathbf{x}$.

2.4. ODE - Initial Value Problem (ODE-IVP).

For most of the processing systems of interest to the chemical engineer, there are three fundamental quantities :mass, energy and momentum. These quantities are can be characterized by variables such as density, concentration, temperature, pressure and flow rate. These characterizing variables are called as state of the processing system. The equations that relate the state variables (dependent variables) to the independent variables are derived from application of conservation principle on the fundamental quantities and are called the state equations.

Let quantity S denote any one of the fundamental quantities

- Total mass
- Mass of the individual components
- Total energy.
- Momentum

Then, the principles of the conservation of the quantity S states that:

$$\frac{[\text{Accumulation of S within a system}]}{\text{time period}} = \frac{[\text{Flow of S in the system}]}{\text{time period}}$$

$$- \frac{[\text{Flow of S out of the system}]}{\text{time period}}$$

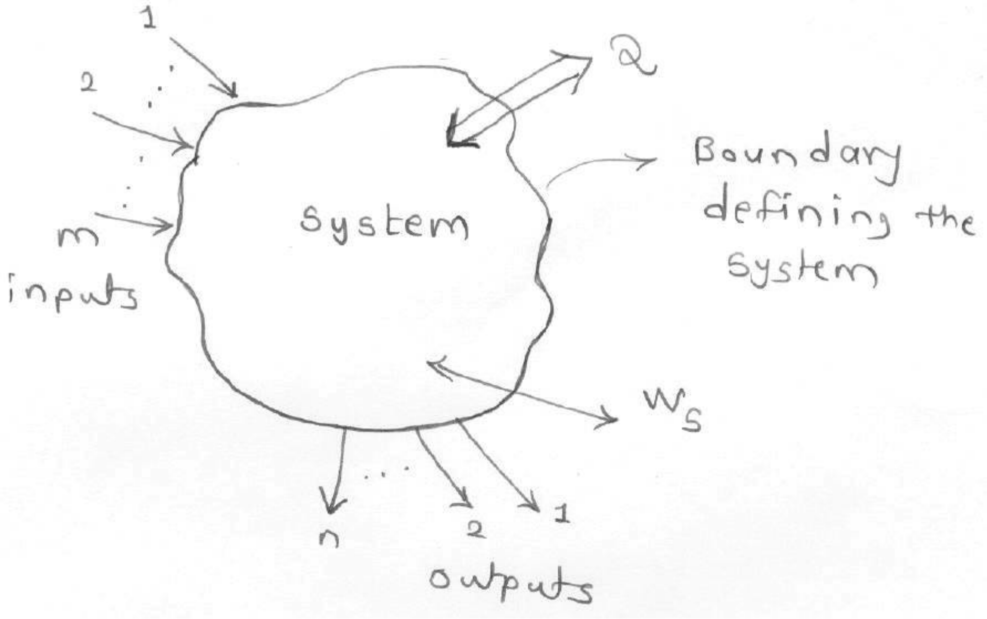


FIGURE 5. General lumped parameter system

$$+\ \frac{\text{[Amount of S generated within the system]}}{\text{time period}}$$

$$-\frac{\text{[Amount of S consumed by the system]}}{\text{time period}}$$

Figure 5 shows schematic diagram of a general system and its interaction with external world. Typical dynamic model equations are as follows

Total Mass Balance

$$d(\rho V) = \sum_{i:inlet} \rho_i F_i - \sum_{j:outlet} \rho_j F_j$$

Mass Balance of the component A

$$\frac{dn_a}{dt} = \frac{d(C_a V)}{dt} = \sum C_{ai} F_i - \sum C_{aj} F_i \ + \ or - rV \tag{2.30}$$

Total energy Balance

$$\frac{dE}{dt} = \frac{d(U+K+P)}{dt} = \sum \rho_i F_i h_i - \sum \rho_j F_j h_j \pm Q \pm W_S \simeq \frac{dH}{dt} \tag{2.31}$$

$$i \ : \ inlet \qquad j : outlet$$

where

ρ : density of the material in the system

ρ_i : density of the material in the i'th inlet stream

ρ_j : density of the material in the j'th outlet stream

V : Total volume of the system

$F_{i:}$ Volumetric flow rate of the i'th inlet stream

$F_{j:}$ Volumetric flow rate of the j'th outlet stream

n_a : number of moles of the component A in the system

C_A : Molal concentration (moles /volume)of A in the system

C_{Ai} : Molal concentration (moles /volume)of A in the i'th inlet stream

C_{Aj} : Molal concentration (moles /volume)of A in the j'th outlet stream

r : reaction rate per unit volume of the component A in the system.

$h_{i:}$ specific enthalpy of the material in the i'th inlet stream

$h_{i:}$ specific enthalpy of the material in the j'th outlet stream

U, K, P : internal, kinetic and potential energies of the system, respectively.

Q : Amount of the heat exchanged between the system and the surroundings per unit time

W_S : Shaft work exchanged between the system and its surroundings.

By convention, a quantity is considered positive if it flows in and negative if it flows out. The state equations with the associated variables constitute the 'lumped parameter mathematical model' of a process, which yields the dynamic or static behavior of the process. The application of the conservation principle stated above will yield a set of differential equations with the fundamental quantities as the dependent variables and time as independent variable. The solution of the differential equations will determine how the state variables change with time i.e., it will determine the dynamic behavior of the process. The process is said to be at the steady state if the state variables do not change with time. In this case, the rate of accumulation of the fundamental quantity S is zero and the resulting balance yields a set of algebraic equations

Example 5. ***Stirred Tank Heater (STH) System (Figure 6):*** *Total momentum of the system remains constant and will not be considered. Total mass balance: Total mass in the tank at any time $t = \rho V = \rho A h$ where A represents cross sectional area.*

$$\frac{d(\rho A h)}{dt} = \rho F_i - \rho F \tag{2.32}$$

Assuming that the density is independent of the temperature,

$$A\frac{dh}{dt} = F_i - F \tag{2.33}$$

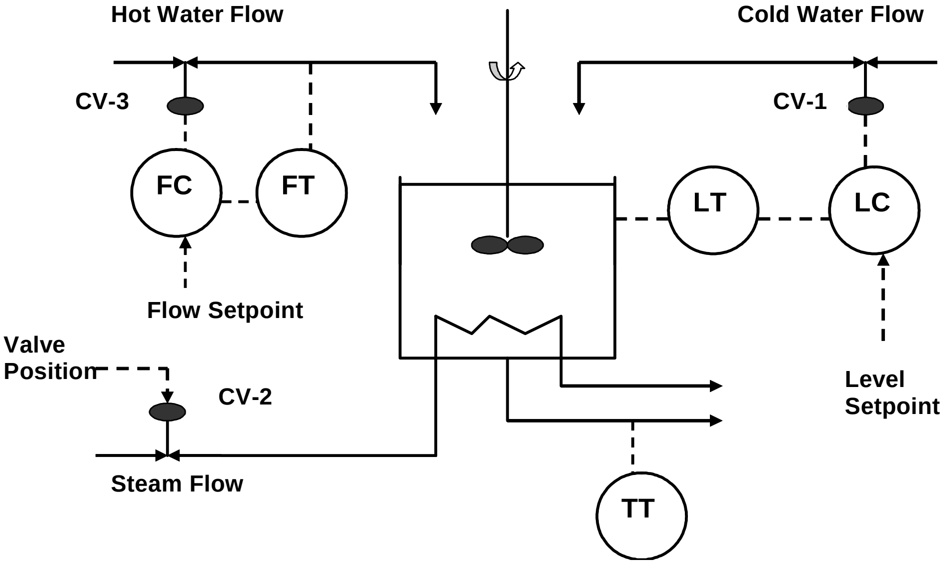


FIGURE 6. Stitted Tank Heater (STH) System

Now, flow out due to the gravity is also a function of height

$$F = k\sqrt{h}$$

Thus,

$$A\frac{dh}{dt} + k\sqrt{h} = F_i \tag{2.34}$$

Total energy of liquid in the tank is given by

$$E = U + k + P$$

However, since tank does not move

$$\frac{dk}{dt} = \frac{dP}{dt} = 0\ ; \quad \frac{dE}{dt} = \frac{dU}{dt}$$

For liquid systems

$$\frac{dU}{dt} \approx \frac{dH}{dt} \tag{2.35}$$

where H is total enthalpy of the liquid in the tank.

$$H = \rho V C_p(T - T_{ref}) = \rho A h C_p(T - T_{ref}) \tag{2.36}$$

T_{ref} *represents reference temperature where the specific enthalpy of the liquid is assumed to be zero. Now, using the energy conservation principle*

$$\frac{d\left(\rho A h C_p(T - T_{ref})\right)}{dt} = \rho F_i C_p(T_i - T_{ref}) - \rho F C_p(T - T_{ref}) + Q \tag{2.37}$$

where Q is the amount of heat supplied by the steam per unit time. Assuming $T_{ref} = 0$, we have

$$A\frac{d(hT)}{dt} = F_iT_i - FT + \frac{Q}{\rho C_p} \tag{2.38}$$

$$\begin{aligned} A\frac{d(hT)}{dt} &= Ah\frac{dT}{dt} + AT\frac{dh}{dt} \\ &= Ah\frac{dT}{dt} + T(F_i - F) \\ &= F_iT_i - FT + \frac{Q}{\rho C_p} \end{aligned}$$

Or

$$Ah\frac{dT}{dt} = F_i(T_i - T) + \frac{Q}{\rho C_p}$$

Summarizing modelling steps

$$\frac{dh}{dt} = \frac{1}{A}(F_i - F\,) = \frac{1}{A}(F_i - k\sqrt{h}) \tag{2.39}$$

$$\frac{dT}{dt} = \frac{F_i}{Ah}(T_i - T) + \frac{Q}{Ah\rho C_p} \tag{2.40}$$

The associated variables can be classified as

- ***state(****or dependent) variables : h, T*
- ***Input (****or independent) variables :T_i, F_i, Q*
- ***Parameters****: A, ρ, C_p*

Steady state behavior can be computed by solving following two equations

$$\frac{dh}{dt} = F_i - k\sqrt{h} = 0 \tag{2.41}$$

$$\frac{dT}{dt} = \frac{F_i}{Ah}(T_i - T) + \frac{Q}{\rho C_p} = 0 \tag{2.42}$$

Once we choose independent variables $F_i = \overline{F}_i, T_i = \overline{T}_i$ and $Q = \overline{Q}$, the steady state $h = \overline{h}$ and $T = \overline{T}$ can be computed by simultaneously solving nonlinear algebraic equations (2.41-2.42).

The system will be disturbed from the steady state if the input variables suddenly change value at $t = 0$. Consider following two situations in which we need to investigate transient behavior of the above process

- *T_i decreases by 10% from its steady state value $\overline{T}_i$ at $t = 0$. Liquid level remains at the same steady state value as T_i does not influence the total mass in tank. The temperature T in the tank will start decreasing with time (see Figure 7). How $T(t)$ changes with time is determined by the*

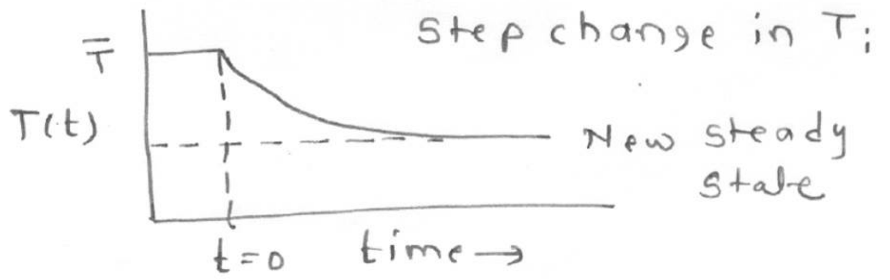


FIGURE 7. System response to step change in T_i

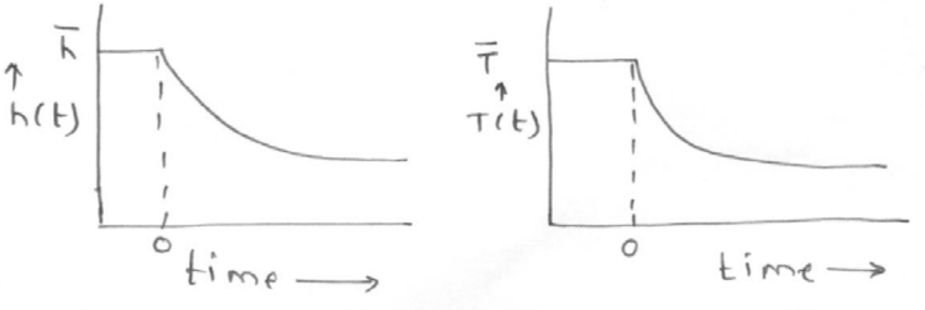


FIGURE 8. System reponse for step change in F_i

solution of the equation (2.39) using the initial as condition $T(0) = \overline{T}$*, the steady state value of* T*.*

- F_i is decreased by 10% from its steady state value $\overline{F}_i$: Since F_i appears in both the dynamic equations, the temperature and the liquid level will start changing simultaneously and the dynamics will be governed by simultaneous solution of coupled nonlinear differential equations (2.39-2.40) starting with initial conditions $T(0) = \overline{T}, h(0) = \overline{h}$.

Figures 8 show schematic diagrams of the process responses for step change in F_i.

It is also possible to investigate response of the system for more complex inputs, such as

$$T_i(t) = \overline{T}_i + \Delta T_i \sin(\omega t)$$

where above function captures daily variation of cooling water inlet temperature. In each case, the transient behavior $T(t)$ and $h(t)$ is computed by solving the system of ODEs subject to given initial conditions and time variation of independent inputs (i.e. forcing functions).

The model we considered above did not contain variation of the variables with respect to space. Such models are called as 'Lumped parameter models' and are described by ordinary differential equations of the form

$$\frac{d\mathbf{x}_1}{dt} = f_1\left[\mathbf{x}_1(t),\, \mathbf{x}_2(t), ..., \mathbf{x}_n(t),\, \mathbf{u}_1(t),\, .., \mathbf{u}_m(t)\right] \tag{2.43}$$

$$..$$

$$\frac{d\mathbf{x}_n}{dt} = f_n\left[\mathbf{x}_1(t),\, \mathbf{x}_2(t), ..., \mathbf{x}_n(t),\, \mathbf{u}_1(t), .., \mathbf{u}_m(t)\right] \tag{2.44}$$

$$\mathbf{x}_1(0) = \overline{\mathbf{x}}_1, ...\mathbf{x}_n(0) = \overline{\mathbf{x}}_n \quad \text{(Initial conditions)}$$

where $\{\mathbf{x}_i(t)\}$ denote the state (or dependent) variables and $\{\mathbf{u}_i(t)\}$ denote independent inputs (or forcing functions) specified for $t \geq 0$. Using vector notation, we can write the above set of ODEs in more compact form

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}, \mathbf{u}) \tag{2.45}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{2.46}$$

where

$$\mathbf{x}(t) = \left[\mathbf{x}_1(t)......\mathbf{x}_n(t)\right]^T \in R^n \tag{2.47}$$

$$\mathbf{u}(t) = \left[\mathbf{u}_1(t)......\mathbf{u}_n(t)\right]^T \in R^m \tag{2.48}$$

$$F(\mathbf{x}, \mathbf{u}) = \left[f_1(\mathbf{x}, \mathbf{u})........f_n(\mathbf{x}, \mathbf{u})\right]^T \in R^n \tag{2.49}$$

and $\mathbf{u}(t)$ is a forcing function vector defined over $t \geq 0$.

- *Steady State Simulation Problem:* If we fix independent inputs to some constant value, say $\mathbf{u}(t) = \overline{\mathbf{u}}$ for $t \geq= 0$,then we can find a steady state solution $\mathbf{x} = \overline{\mathbf{x}}$ corresponding to these constant inputs by simultaneously solving n nonlinear algebraic equations

$$F(\mathbf{x}, \overline{\mathbf{u}}) = \overline{0} \tag{2.50}$$

  obtained by setting $d\mathbf{x}/dt = \overline{0}$ where $\overline{0}$ represents $n \times 1$ zero vector.

- *Dynamic Simulation Problem:* Given input trajectories

$$\mathbf{u}(t) = \left[\mathbf{u}_1(t) \quad \mathbf{u}_2(t)......\mathbf{u}_m(t)\right]^T \tag{2.51}$$

  as a function of time for $t \geq= 0$ and with the initial state $\mathbf{x}(0)$, integrate

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}, \mathbf{u}(t)) \tag{2.52}$$

  over interval $0 \leq t \leq t_f$ to determine state trajectories

$$\mathbf{x}(t) = \left[x_1(t) \quad x_2(t)..........x_n(t)\right]^T \tag{2.53}$$

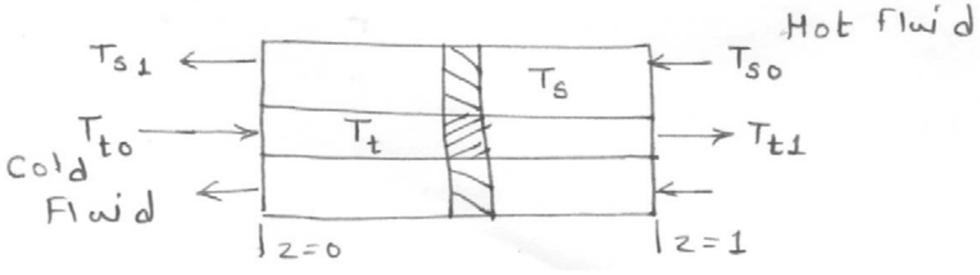


FIGURE 9. Shell and tube heat exchanger

Since $\mathbf{u}(t)$ is a known function of time, we re-state the above problem as

$$\frac{d\mathbf{x}}{dt} = F_{\mathbf{u}}(\mathbf{x}, t) \quad ; \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.54}$$

$F_{\mathbf{u}}(\mathbf{x}, t)\, (= F(\mathbf{x}, \mathbf{u}(t)))$ denotes $F()$ with the given $\mathbf{u}(t)$.

2.5. PDEs and ODE-Boundary value Problems. Most of the systems encountered in chemical engineering are distributed parameter systems. Even though behavior of some of these systems can be adequately represented by lumped parameter models, such simplifying assumptions may fail to provide accurate picture of system behavior in many situations and variations of variables along time and space have to be considered while modeling. This typically result in a set of partial differential equations.

EXAMPLE 6. *Consider the double pipe heat exchanger in which a liquid flowing in the inner tube is heated by steam flowing countercurrently around the tube (Figure 10). The temperature in the pipe changes not only with time but also along the axial direction z. While developing the model, it is assumed that the temperature does not change along the radius of the pipe. Consequently , we have only two independent variables, i.e. z and t. To perform the energy balance,we consider an element of length Δz as shown in the figure. For this element, over a period of time Δt*

$$\begin{aligned} \rho C_p A\Delta z[(T\ )_{t+\Delta t} - (T\ )_t] &= \rho C_p VA(T)_z\Delta t - \rho C_p VA(T)_{z+\Delta z}\Delta t \qquad (2.55) \\ &\quad +Q\Delta t(\pi D\Delta z) \qquad (2.56) \end{aligned}$$

This equation can be explained as

[accumulation of the enthalpy during the time period Δt]
= [flow in of the enthalpy during Δt] - [flow out of the enthalpy during Δt]
[enthalpy transferred from steam to the liquid through wall during Δt]

where

Q : *amount of heat transferred from the steam to the liquid per unit time and per unit heat transfer area.*

A : *cross section area of the inner tube.*

V : *average velocity of the liquid(assumed constant).*

D : *external diameter of the inner tube.*

Dividing both the sides by $(\Delta z \Delta t)$ *and taking limit as* $\Delta t \to 0$ *and* $\Delta z \to 0$*, we have*

$$\rho C_p A \frac{\partial T(z,t)}{\partial t} = -\rho C_p V A \frac{\partial T(z,t)}{\partial z} + \pi D Q \tag{2.57}$$

$$Q = U[T_{st} - T] \tag{2.58}$$

Boundary conditions:

$$T(t, z = 0) = T_1 for t \geq 0$$

Initial condition

$$T(t = 0, z) = T_0\,(0, z) \tag{2.59}$$

Steady State Simulation: *Find* $T(z)$ *given* $T(z = 0) = T_1$ *when* $\partial T/\partial t = 0$*, i.e. solve for*

$$\rho C_p V A \frac{\partial T}{\partial z} = \pi D Q = \pi D Q U(T_{st} - T) \tag{2.60}$$

$$T(0) = T_1 \tag{2.61}$$

This results in a ODE-IVP, which can be solved to obtain steady state profiles $T(z)$ *for specified heat load and liquid velocity.*

Dynamic Simulation

$$\rho C_p A \frac{\partial T}{\partial t} = -\rho C_p V A \frac{\partial T}{\partial z} + \pi D Q \tag{2.62}$$

with

$$T(t, 0) = T_1 at\ z = 0\ \textit{and}\ t \succeq 0 \ : \textit{Boundary condition} \tag{2.63}$$

$$T(0, z) = T_0(z) \quad : \textit{Initial temperature profile} \tag{2.64}$$

This results in a Partial Differential Equation (PDE) model for the distributed parameter system.

Example 7. *Now, let us consider the situation where the some hot liquid is used on the shell side to heat the tube side fluid (see Figure 10). The model*

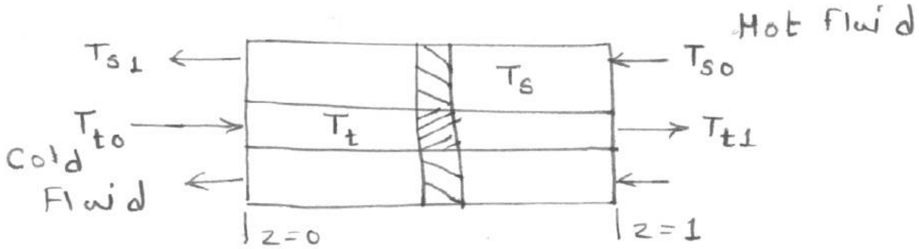


FIGURE 10. Double Pipe Heat Exchanger

equations for this case can be stated as

$$\rho_t C_{pt} A_t \frac{\partial T_t(z,t)}{\partial t} = -\rho_t C_{pt} V_t A_t \frac{\partial T_t(z,t)}{\partial z} + \pi D Q(z,t) \tag{2.65}$$

$$\rho_s C_{pt} A_s \frac{\partial T_s(z,t)}{\partial t} = \rho_s C_{ps} V_s A_s \frac{\partial T_s(z,t)}{\partial z} - \pi D Q(z,t) \tag{2.66}$$

$$Q(z,t) = U[T_s(z,t) - T_t(z,t)] \tag{2.67}$$

where subscript t denotes tube side and subscript s denotes shell side. The initial and boundary conditions become

$$T_t(t,0) = T_{t0} \; at \; z=0 \; and \; t \succeq 0 \quad : \; Boundary \; condition \tag{2.68}$$

$$T(0,z) = T_{t0}(z) \quad : \; Initial \; temperature \; profile \tag{2.69}$$

$$T_s(t,1) = T_{s1} at \; z=1 \; and \; t \succeq 0 \quad : \; Boundary \; condition \tag{2.70}$$

$$T(0,z) = T_{s0}(z) \quad : \; Initial \; temperature \; profile \tag{2.71}$$

These are coupled PDEs and have to be solved simultaneously to understand the transient behavior. The steady state problem can be stated as

$$\rho_t C_{pt} V_t A_t \frac{dT_t(z,t)}{dz} = \pi D U[T_s(z) - T_t(z)] \tag{2.72}$$

$$\rho_s C_{ps} V_s A_s \frac{dT_s(z,t)}{dz} = \pi D U[T_s(z) - T_t(z)] \tag{2.73}$$

$$T_t(0) = T_{t0} \quad at \; z=0 \tag{2.74}$$

$$T_s(1) = T_{s1} \quad at \; z=1 \tag{2.75}$$

Equations (2.72-2.73) represent coupled ordinary differential equations. The need to compute steady state profiles for the counter-current double pipe heat exchanger results in a boundary value problem (ODE-BVP) as one variable is specified at $z=0$ while the other is specified at at $z=1$.

Typical partial differential equations we come across in engineering applications are of the form

$$\nabla^2 u = a\frac{\partial u}{\partial t} + b\frac{\partial^2 u}{\partial t^2} + cu + f(x_1, x_2, x_3, t) \tag{2.76}$$

subject to appropriate boundary conditions and initial conditions. This PDE is solved in a three dimensional region V, which can be bounded or unbounded. The boundary of V is denoted by S. On the spatial surface S, we have boundary conditions of the form

$$\left(\alpha(s,t)\,\widehat{n}\right).\nabla u + \beta(s,t)u = h(s,t) \tag{2.77}$$

where $\widehat{n}$ is the outward normal direction to S and s represents spatial coordinate along S. We can classify the PDEs as follows

- Elliptic: $a = b = 0$
- Parabolic: $a \neq 0, b = 0$
- Hyperbolic: $b > 0$

3. Summary

These lecture notes introduces various basic forms of equations that appear in steady state and dynamic models of simple unit operations. Following generic forms or problem formulations have been identified

- Linear algebraic equations
- Nonlinear algebraic equations
- Unconstrained optimization
- Ordinary Differential Equations : Initial Value Problem (ODE-IVP)
- Ordinary Differential Equations : Boundary Value Problem (ODE-BVP)
- Partial Differential Equations (PDEs)

Methods for dealing with numerical solutions of these generic forms / formulations will be discussed in the later parts.

4. Appendix: Basic Concepts of Partial Differential Equations

Definition 1. ***Order of PDE:*** *Order of a PDE is highest order of derivative occurring in PDE.*

Definition 2. ***Degree of PDE:*** *Power to which highest order derivative is raised.*

Example 8. *Consider PDE*

$$\partial u/\partial t+(d^2u/dz^2)^n=u^3 \tag{4.1}$$

Here the $Oredr = 2$ and $Degree = n.$

Solutions of PDEs are sought such that it is satisfied in the domain and on the boundary conditions are satisfied. A problem is said to be well posed when the solution uniquely determined and it is sufficient smooth and differentiable function of the independent variables. The boundary conditions have to be consistent with one another in order for a problem to be well posed. This implies that at the points common to boundaries, the conditions should not violet each other.

A linear PDE can be classified as:

- Homogeneous equations:Differential equation that does not contain any terms other than dependent variables and their derivatives.

$$\partial u/\partial t=\partial u^2/\partial x^2 \tag{4.2}$$

$$\partial^2u/\partial x^2+\partial^2u/\partial y^2=0$$

- Non homogeneous equations: Contain terms other than dependent variables

$$\partial u/\partial t=\partial^2u/\partial x^2+\sin x \tag{4.3}$$

$$\partial^2u/\partial x^2+\partial^2u/\partial y^2=\sin x\ \ \sin y \tag{4.4}$$

The boundary conditions can be similarly homogeneous or non homogeneous depending on whether they contain terms independent of dependent variables.

The PDEs typically encountered in engineering applications are 2^{nd}order PDEs (reaction-diffusion systems, heat transfer, fluid-flow etc.)

Classification of 2^{nd} order PDEs:

Consider a 2^{nd}order PDE in n independent variables $(x_1,x_2,x_3,x_4)=(x,y,z,t)$. This can be written as

$$\sum_{i=1}^{4}\sum_{j=1}^{4}a_{ij}\frac{\partial^2u}{\partial x_i\partial x_j}=f\left[\partial u/\partial x_1,......\partial u/\partial x_4,,u,x_1,.........,x_4\right] \tag{4.5}$$

a_{ij} are assumed to be independent of $'u'$and its derivative. They can be functions of (x_i). a_{ij} can always be written as $a_{ij} = a_{ji}$ for i $\neq$ j as

$$\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_j \partial x_i} \tag{4.6}$$

Thus, $a_{ij}$are elements of a real symmetric matrix A. Obviously A has real eigen values. The PDE is called

- **Elliptic:** if all eigenvalues are +ve or-ve.
- **Hyperbolic:** if some eigenvalues are +ve and rest are -ve.
- **Parabolic:** if at-least one eigen value is zero.

The classification is global if a_{ij} are independent of x_i, else it is local. Elliptic Problems typically arise while studying steady-state behavior of diffusive systems. Parabolic or hyperbolic problems typically arise when studying transient behavior of diffusive systems.

CHAPTER 2

Fundamentals of Functional Analysis

1. Introduction

When we begin to use concept of vectors in formulating mathematical models for engineering systems, we tend to associate the concept of a vector space with the three dimensional coordinate space. The three dimensional space we are familiar with can be looked upon as a set of objects called vectors, which satisfy certain generic properties. While working with mathematical modeling, however, we need to deal with variety of such sets containing such objects. It is possible to 'distill' essential properties satisfied by vectors in the three dimensional vector space and develop a more general concept of a vector space, which is collection of objects satisfying these properties. Such a generalization can provide a unified view of problem formulations and the solution techniques.

Generalization of the concept of vector and vector space to any general set other than collection of vectors in three dimensions is not sufficient. In order to work with these sets of generalized vectors, we need various algebraic and geometric structures on these sets, such as norm of a vector, angle between two vectors or convergence of a sequence of vectors. To understand why these structures are necessary, consider the fundamental equation that arises in numerical analysis

$$F(\mathbf{x}) = \overline{0} \tag{1.1}$$

where $\mathbf{x}$ is a vector and $F(.)$ represents some linear or nonlinear operator, which when operates on $\mathbf{x}$ yields the zero vector $\overline{0}$. In order to generate a numerical approximation to the solution of equation (6.2), this equation is further transformed to formulate an iteration sequence as follows

$$\mathbf{x}^{(k+1)} = G\left[\mathbf{x}^{(k)}\right] \quad ; \quad k = 0, 1, 2, ... \tag{1.2}$$

where $\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$ is sequence of vectors in vector space under consideration. The iteration equation is formulated in such a way that the solution $\mathbf{x}^{*}$ of equation **(**1.2**),** i.e.

$$\mathbf{x}^{*} = G\left[\mathbf{x}^{*}\right]$$

is also a solution of equation (6.2). Here we consider two well known examples of such formulation.

EXAMPLE 9. ***Iterative schemes for solving single variable nonlinear algebraic equation:*** *Consider one variable nonlinear equation*

$$f(x) = x^3 + 7x^2 + x\sin(x) - e^x = 0$$

This equation can be rearranged to formulate an iteration sequence as

$$x^{(k+1)} = \frac{\exp(x^{(k)}) - \left[x^{(k)}\right]^3 - 7\left[x^{(k)}\right]^2}{\sin(x^{(k)})} \tag{1.3}$$

Alternatively, using Newton-Raphson method for single variable nonlinear equations, iteration sequence can be formulated as

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \frac{f(x^{(k)})}{\left[df(x^{(k)})/dx\right]} \\ &= x^{(k)} - \frac{\left[x^{(k)}\right]^3 + 7\left[x^{(k)}\right]^2 + x^{(k)}\sin(x^{(k)}) - \exp(x^{(k)})}{3\left[x^{(k)}\right]^2 + 14x^{(k)} + \sin(x^{(k)}) + x^{(k)}\cos(x^{(k)}) - \exp(x^{(k)})} \qquad (1.4) \end{aligned}$$

Both the equations (1.3) and (1.4) are of the form given by equation (1.2).

EXAMPLE 10. ***Solving coupled nonlinear algebraic equations by successive substitution***

Consider three coupled nonlinear algebraic equations

(1.5)

$$F(x,y,z,w) = \begin{bmatrix} f_1(x,y,z,w) \\ f_2(x,y,z,w) \\ f_3(x,y,z,w) \\ f_4(x,y,z,w) \end{bmatrix} = \begin{bmatrix} xy - xz - 2w \\ y^2x + zwy - 1 \\ z\sin(y) - z\sin(x) - ywx \\ wyz - \cos(x) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

which have to be solved simultaneously.. One possible way to transform the above set of nonlinear algebraic equations to the form given by equation (1.2) is as follows as

(1.6)

$$\mathbf{x}^{(k+1)} \equiv \begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \\ z^{(k+1)} \\ w^{(k+1)} \end{bmatrix} = \begin{bmatrix} 2w^{(k)}/(y^{(k)} - z^{(k)}) \\ 1/\left(y^{(k)}x^{(k)} + z^{(k)}w^{(k)}\right) \\ y^{(k)}x^{(k)}w^{(k)}/\left(\sin(y^{(k)}) - \sin(x^{(k)})\right) \\ \cos(x^{(k)})/\left(y^{(k)}z^{(k)}\right) \end{bmatrix} \equiv G\left[\mathbf{x}^{(k)}\right]$$

EXAMPLE 11. ***Picard's Iteration:*** *Consider single variable ODE-IVP*

$$F\left[x(z)\right] = \frac{dx}{dz} - f(x,z) = 0 \;\; ; \;\; x(0) = x_0 \;\; ; \;\; 0 \leq z \leq 1 \tag{1.7}$$

which can be solved by formulating the Picard's iteration scheme as follows

$$x^{(k+1)}(z) = x_0 + \int_0^z f\left[x^{(k)}(q), q\right] dq \tag{1.8}$$

Note that this procedure generates a sequence of functions $x^{(k+1)}(z)$ over the interval $0 \leq z \leq 1$ starting from an initial guess solution $x^{(0)}(z)$. If we can treat a function $x^{(k)}(z)$ over the interval $0 \leq z \leq 1$ as a ***vector****, then it is easy to see that the equation (1.8) is of the form given by equation (1.2).*

Fundamental concern in such formulations is, whether the sequence of **vectors**

$$\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$$

converges to the solution $\mathbf{x}^*$. Another important question that naturally arises while numerically computing the sequences of vectors in all the above examples is that when do we terminate these iteration sequences. In order to answer such questions while working with general vector spaces, we have to define concepts such as norm of a vector in such spaces and use it to generalize the notion of convergence of a sequence.

A branch of mathematics called *functional analysis* deals with generalization of geometric concepts, such as length of a vector, convergence, orthogonality etc. used in the three dimensional vector spaces to more general finite and infinite dimensional vector spaces. Understanding fundamentals of functional analysis helps understand basis of various seemingly different numerical methods. In this part of the lecture notes we introduce some concepts from functional analysis and linear algebra, which help develop a unified view of all the numerical methods. In the next section, we review fundamentals of functional analysis, which will be necessary for developing a good understanding of numerical analysis. Detailed treatment of these topics can be found in Luenberger [**11**] and Kreyzig [**7**].

A word of advice before we begin to study these grand generalizations. While dealing with the generic notion of vector spaces, it is difficult to visualize shapes as we can do in the three dimensional vector space. However, if you know the concepts from three dimensional space well, then you can develop an understanding of the corresponding concept in any general space. It is enough to know the Euclidean geometry well in the three dimensions.

2. Vector Spaces

Associated with every vector space is a set of scalars F (also called as *scalar field* or *coefficient field*) used to define scalar multiplication on the space. In functional analysis, the scalars will be always taken to be the set of real numbers (R) or complex numbers (C).

DEFINITION 3. ***(Vector Space):*** *A vector space X is a set of elements called vectors and scalar field F together with two operations . The first operation is called addition which associates with any two vectors $\mathbf{x}, \mathbf{y} \in X$ a vector $\mathbf{x}+\mathbf{y} \in X$, the sum of $\mathbf{x}$ and $\mathbf{y}$. The second operation is called scalar multiplication, which associates with any vector $\mathbf{x} \in X$ and any scalar α a vector $\alpha\mathbf{x}$ (a scalar multiple of $\mathbf{x}$ by α). The set X and the operations of addition and scalar multiplication are assumed to satisfy the fallowing axioms for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$ ·*

(1) $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$ (commutative law)
(2) $\mathbf{x}+(\mathbf{y}+\mathbf{z})=(\mathbf{x}+\mathbf{y})+\mathbf{z}$ (associative law)
(3) There exists a null vector $\overline{0}$ such that $\mathbf{x}+\overline{0}=\mathbf{x}$ for all $\mathbf{x} \in X$
(4) $\alpha(\mathbf{x}+\mathbf{y})=\alpha\mathbf{x}+\alpha\mathbf{y}$
(5) $(\alpha+\beta)\,\mathbf{x}=\alpha\mathbf{x}+\beta\mathbf{x}$ (4,5 are distributive laws)
(6) $\alpha\beta\,(\mathbf{x})=\alpha\,(\beta\mathbf{x})$
(7) $\alpha x=\overline{0}$ when $\alpha=0$.
$\alpha x=\mathbf{x}$ when $\alpha=1$.
(8) For convenience $-1\mathbf{x}$ is defined as $-\mathbf{x}$ and called as negative of a vector we have

$\mathbf{x}+(-\mathbf{x})=\overline{0}$

In short, given any vectors $\mathbf{x}, \mathbf{y} \in X$ and any scalars $\alpha, \beta \in R$,we can write that $\alpha\mathbf{x}+\beta\mathbf{y} \in X$ when X is a linear vector space.

EXAMPLE 12. ($X \equiv R^n, F \equiv R$) : *$n-$ dimensional real coordinate space. A typical element $\mathbf{x} \in X$ can be expressed as*

$$\mathbf{x}=\left[\begin{array}{cccc} x_1 & x_2 & ..... & x_n \end{array}\right]^T \tag{2.1}$$

where x_i denotes the i'th element of the vector.

EXAMPLE 13. *($X \equiv C^n, F \equiv C$) : $n-$ dimensional complex coordinate space.*

EXAMPLE 14. *($X \equiv R^n, F \equiv C$) : It may be noted that this combination of set X and scalar field F does not form a vector space. For any $\mathbf{x} \in X$ and any $\alpha \in C$ the vector $\alpha\mathbf{x} \notin X$.*

Example 15. $(X \equiv l_\infty, F \equiv R)$ *: Set of all infinite sequence of real numbers. A typical vector* $\mathbf{x}$ *of this space has form* $\mathbf{x} = (\zeta_1, \zeta_2, ..., \zeta_k, ...)$.

Example 16. $(X \equiv C[a,b],\ F \equiv R)$ *: Set of all continuous functions over an interval* $[a,b]$ *forms a vector space. We write* $\mathbf{x} = \mathbf{y}$ *if* $\mathbf{x}(t) = \mathbf{y}(t)$ *for all* $t \in [a,b]$ *The null vector* $\overline{0}$ *in this space is a function which is zero every where on* $[a,b]$ *,i.e*

$$f(t) = 0 \text{ for all } t \in [a,b]$$

If $\mathbf{x}$ *and* $\mathbf{y}$ *are vectors from this space and* α *is real scalar then* $(\mathbf{x}+\mathbf{y})(t) = \mathbf{x}(t) + \mathbf{y}(t)$ *and* $(\alpha\mathbf{x})(t) = \alpha\mathbf{x}(t)$ *are also elements of* $C[a,b]$.

Example 17. $\left(X \equiv C^{(n)}[a,b],\ F \equiv R\right)$ *: Set of all continuous and n times differentiable functions over an interval* $[a,b]$ *forms a vector space.*

Example 18. $X \equiv$ *set of all real valued polynomial functions defined on interval* $[a,b]$ *together with* $F \equiv R$ *forms a vector space.*

Example 19. *The set of all functions* $f(t)$ *for which*

$$\int_a^b |f(t)|^p \, dt < \infty$$

is a linear space L_p.

Example 20. $(X \equiv R^m \times R^n,\ F \equiv R)$ *: Set of all* $m \times n$ *matrices with real elements. Note that a* vector *in this space is a* $m \times n$ *matrix and the null vector corresponds to* $m \times n$ *null matrix. It is easy to see that, if* $A, B \in X$, *then* $\alpha A + \beta B \in X$ *and* X *is a linear vector space.*

Definition 4. ***(Subspace):*** *A non-empty subset* M *of a vector space* X *is called subspace of* X *if every vector* $\alpha\mathbf{x} + \beta\mathbf{y}$ *is in* M *wherever* $\mathbf{x}$ *and* $\mathbf{y}$ *are both in* M. *Every subspace always contains the null vector, I.e. the origin of the space* $\mathbf{x}$

Example 21. ***Subspaces***

1. *Two dimensional plane passing through origin of three dimensional coordinate space. (Note that a plane which does not pass through the origin is not a sub-space.)*
2. *A line passing through origin of* R^n
3. *The set of all real valued n'th order polynomial functions defined on interval* $[a,b]$ *is a subspace of* $C[a,b]$.

Thus, the fundamental property of objects (elements) in a vector space is that they can be constructed by simply adding other elements in the space. This property is formally defined as follows.

Definition 5. ***(Linear Combination):*** *A linear combination of vectors* $\mathbf{x}^{(1)},\mathbf{x}^{(2)}_{,},.......\mathbf{x}^{(m)}$ *in a vector space is of the form* $\alpha_1\mathbf{x}^{(1)}+\alpha_2\mathbf{x}^{(2)}+.............+\alpha_m\mathbf{x}^{(m)}$ *where* $(\alpha_1,...\alpha_m)$ *are scalars.*

Note that we are dealing with set of vectors

$$\left\{\mathbf{x}^{(k)}:k=1,2,......m.\right\} \tag{2.2}$$

. Thus, if $X=R^n$ and $\mathbf{x}^{(k)}\in R^n$ represents k'th vector in the set, then it is a vector with n components i.e.

$$\mathbf{x}^{(k)}=\left[\begin{array}{cccc} x_1^{(k)} & x_2^{(k)} & .... & x_n^{(k)}\end{array}\right]^T \tag{2.3}$$

Similarly, if $X=l_\infty$ and $\mathbf{x}^{(k)}\in l_\infty$ represents k'th vector in the set, then $\mathbf{x}^{(k)}$ represents a sequence of infinite components

$$\mathbf{x}^{(k)}=\left[\begin{array}{cccc} x_1^{(k)} & .... & x_i^{(k)} & ......\end{array}\right]^T \tag{2.4}$$

Definition 6. ***(Span of Set of Vectors):*** *Let S be a subset of vector space X. The set generated by all possible linear combinations of elements of S is called as span of S and denoted as $[S]$. Span of S is a subspace of X.*

Definition 7. ***(Linear Dependence):*** *A vector* $\mathbf{x}$ *is said to be linearly dependent up on a set S of vectors if* $\mathbf{x}$ *can be expressed as a linear combination of vectors from S. Alternatively,* $\mathbf{x}$ *is linearly dependent upon S if* $\mathbf{x}$ *belongs to span of S, i.e.* $\mathbf{x}\in[S]$. *A vector is said to be linearly independent of set S, if it not linearly dependent on S . A necessary and sufficient condition for the set of vectors* $\mathbf{x}^{(1)},\mathbf{x}^{(2)},.....\mathbf{x}^{(m)}$ *to be linearly independent is that expression*

$$\sum_{i=1}^{m}\alpha_i\mathbf{x}^{(i)}=\overline{0} \tag{2.5}$$

implies that $\alpha_i=0$ for all $i=1,2......m$.

Definition 8. ***(Basis):*** *A finite set S of linearly independent vectors is said to be basis for space X if S generates X i.e. $X=[S]$*

A vector space having finite basis (spanned by set of vectors with finite number of elements) is said to be finite dimensional. All other vector spaces are said to be infinite dimensional. We characterize a finite dimensional space by number of elements in a basis. Any two basis for a finite dimensional vector space contain the same number of elements.

Example 22. ***Basis of a vector space***

(1) *Let $S = \{\mathbf{v}\}$ where $\mathbf{v} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}^T$ and let us define span of S as $[S] = \alpha\mathbf{v}$ where $\alpha \in R$ represents a scalar. Here, $[S]$ is one dimensional vector space and subspace of R^5*

(2) *Let $S = \left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right\}$ where*

$$\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \quad ; \ \mathbf{v}^{(2)} = \begin{bmatrix} 5 \\ 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} \tag{2.6}$$

Here span of S (i.e. $[S]$) is two dimensional subspace of R^5.

(3) *Consider set of n^{th} order polynomials on interval $[0, 1]$. A possible basis for this space is*

$$p^{(1)}(z) = 1; \ \ p^{(2)}(z) = z; \ \ p^{(3)}(z) = z^2, ..., p^{(n+1)}(z) = z^n \tag{2.7}$$

Any vector $p(t)$ from this pace can be expressed as

$$\begin{aligned} p(z) &= \alpha_0 p^{(1)}(z) + \alpha_1 p^{(2)}(z) + ... + \alpha_n p^{(n+1)}(z) \\ &= \alpha_0 + \alpha_1 z + ... + \alpha_n z^n \end{aligned} \tag{2.8}$$

Note that $[S]$ in this case is $(n + 1)$ dimensional subspace of $C[a, b]$.

(4) *Consider set of continuous functions over interval, i.e. $C[-\pi, \pi]$. A well known basis for this space is*

$$x^{(0)}(z) = 1; \ x^{(1)}(z) = \cos(z); \ \ x^{(2)}(z) = \sin(z), \tag{2.9}$$

$$x^{(3)}(z) = \cos(2z), \ \ x^{(4)}(z) = \sin(2z), ... \tag{2.10}$$

It can be shown that $C[-\pi, \pi]$ is an infinite dimensional *vector space.*

3. Normed Linear Spaces and Banach Spaces

In three dimensional space, we use lengths to compare any two vectors. Generalization of the concept of length of a vector in three dimensional vector space to an arbitrary vector space is achieved by defining a scalar valued function called *norm* of a vector.

Definition 9. ***(Normed Linear Vector Space):*** *A normed linear vector space is a vector space X on which there is defined a real valued function which maps each element $\mathbf{x} \in X$ into a real number $\|\mathbf{x}\|$ called norm of $\mathbf{x}$. The norm satisfies the fallowing axioms.*

(1) $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in X$; $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \overline{0}$ (*zero vector*)
(2) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$for each $\mathbf{x}, \mathbf{y} \in X$. (*triangle inequality*).
(3) $\|\alpha\mathbf{x}\| = |\alpha| \,.\, \|\mathbf{x}\|$ for all scalars α and each $\mathbf{x} \in X$

The above definition of norm is an abstraction of usual concept of length of a vector in three dimensional vector space.

Example 23. ***Vector norms:***

(1) $(R^n, \|.\|_1)$:*Euclidean space* R^n *with 1-norm:* $\|\mathbf{x}\|_1 = \sum\limits_{i=1}^{N} |x_i|$
(2) $(R^n, \|.\|_2)$:*Euclidean space* R^n *with 2-norm:*

$$\|\mathbf{x}\|_2 = \left[\sum_{i=1}^{N} (x_i)^2\right]^{\frac{1}{2}}$$

(3) $(R^n, \|.\|_\infty)$:*Euclidean space* R^n *with* $\infty-$*norm:* $\|\mathbf{x}\|_\infty = \max |x_i|$
(4) $\left(R^n, \|.\|_p\right)$:*Euclidean space* R^n *with p-norm,*

$$\|\mathbf{x}\|_p = \left[\sum_{i=1}^{N} |x_i|^p\right]^{\frac{1}{p}} \tag{3.1}$$

, where p is a positive integer

(5) *n-dimensional complex space* (C^n) *with p-norm,*

$$\|\mathbf{x}\|_p = \left[\sum_{i=1}^{N} |x_i|^p\right]^{\frac{1}{p}} \tag{3.2}$$

, where p is a positive integer

(6) *Space of infinite sequences* (l_∞)*with p-norm: An element in this space, say* $\mathbf{x} \in l_\infty$*, is an infinite sequence of numbers*

$$\mathbf{x} = \{x_{1,}x_2, ..., x_n, ...\} \tag{3.3}$$

such that p-norm is bounded

$$\|\mathbf{x}\|_p = \left[\sum_{i=1}^{\infty} |x_i|^p\right]^{\frac{1}{p}} < \infty \tag{3.4}$$

for every $\mathbf{x} \in l_\infty$*, where p is an integer.*

Example 24. $(C[a,b], \|\mathbf{x}(t)\|_\infty)$: *The normed linear space* $C[a,b]$ *together with infinite norm*

$$\|\mathbf{x}(t)\|_\infty = \begin{matrix} \max \\ a \leq t \leq b \end{matrix} \; |\mathbf{x}(t)| \tag{3.5}$$

It is easy to see that $\|\mathbf{x}(t)\|_{\infty}$ *defined above qualifies to be a norm*

$$max\,|\mathbf{x}(t)+\mathbf{y}(t)| \leq \max[|\mathbf{x}(t)|+|\mathbf{y}(t)|] \quad \leq \max|\mathbf{x}(t)|+\max|\mathbf{y}(t)| \tag{3.6}$$

$$max\,|\alpha\mathbf{x}(t)| = \max|\alpha|\,|\mathbf{x}(t)| = |\alpha|\max|\mathbf{x}(t)| \tag{3.7}$$

Other types of norms, which can be defined on the set of continuous functions over $[a,b]$ *are as follows*

$$\|\mathbf{x}(t)\|_1 = \int_a^b |\mathbf{x}(t)|\,dt \tag{3.8}$$

$$\|\mathbf{x}(t)\|_2 = \left[\int_a^b |\mathbf{x}(t)|^2\,dt\right]^{\frac{1}{2}} \tag{3.9}$$

Once we have defined norm of a vector in a vector space, we can proceed to generalize the concept of convergence of sequence of vectors. Concept of convergence is central to all iterative numerical methods.

Definition 10. ***(Convergence):*** *In a normed linear space an infinite sequence of vectors* $\left\{\mathbf{x}^{(k)} : k = 1,2,.......\right\}$ *is said to converge to a vector* $\mathbf{x}^*$ *if the sequence* $\left\{\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|,\ k = 1,2,...\right\}$ *of real numbers converges to zero. In this case we write* $\mathbf{x}^{(k)} \rightarrow \mathbf{x}^*$.

In particular, a sequence $\left\{\mathbf{x}^{(k)}\right\}$ in R^nconverges if and only if each component of the vector sequence converges. If a sequence converges, then its limit is unique.

Definition 11. ***(Cauchy sequence):*** *A sequence* $\left\{\mathbf{x}^{(k)}\right\}$ *in normed linear space is said to be a Cauchy sequence if* $\left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\| \rightarrow 0$ *as* $n,m \rightarrow \infty.i.e.$ *given an* $\varepsilon > 0$ *there exists an integer* N *such that* $\left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\| < \varepsilon$ *for all* $n,m \geq N$

Example 25. ***Convergent sequences:*** *Consider the sequence of vectors represented as*

$$\mathbf{x}^{(k)} = \begin{bmatrix} 1+(0.2)^k \\ -1+(0.9)^k \\ 3/\left(1+(-0.5)^k\right) \\ (0.8)^k \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ -1 \\ 3 \\ 0 \end{bmatrix} \tag{3.10}$$

with respect to any p-norm defined on R^4. *It can be shown that it is a Cauchy sequence. Note that each element of the vector converges to a limit in this case.*

When we are working in R^n or C^n,all convergent sequences are Cauchy sequences and vice versa. However, all Cauchy sequences in a general vector space need not be convergent. Cauchy sequences in some vector spaces exhibit such strange behavior and this motivates the concept of completeness of a vector space.

Definition 12. ***(Banach Space):*** *A normed linear space X is said to be complete if every Cauchy sequence has a limit in X. A complete normed linear space is called Banach space.*

Examples of Banach spaces are

$$(\mathbf{R}^n, \|.\|_1), \ (\mathbf{R}^n, \|.\|_2), (\mathbf{R}^n, \|.\|_\infty)$$
$$(\mathbf{C}^n, \|.\|_1), (\mathbf{C}^n, \|.\|_2), (l_\infty, \|.\|_1), (l_\infty, \|.\|_2)$$

etc. Concept of Banach spaces can be better understood if we consider an example of a vector space where a Cauchy sequence is not convergent, i.e. the space under consideration is an incomplete normed linear space. Note that, even if we find one Cauchy sequence in this space which does not converge, it is sufficient to prove that the space is not complete.

Example 26. *Let $X = (Q, \|.\|_1)$ i.e. set of rational numbers (Q) with scalar field also as the set of rational numbers (Q) and norm defined as*

$$\|x\|_1 = |x| \tag{3.11}$$

A vector in this space is a rational number. In this space, we can construct Cauchy sequences which do not converge to a rational numbers (or rather they converge to irrational numbers). For example, the well known Cauchy sequence

$$\begin{aligned} x^{(1)} &= 1/1 \\ x^{(2)} &= 1/1 + 1/(2!) \\ &\ldots\ldots\ldots \\ x^{(n)} &= 1/1 + 1/(2!) + \ldots.. + 1/(n!) \end{aligned}$$

converges to e, which is an irrational number. Similarly, consider sequence

$$x^{(n+1)} = 4 - (1/x^{(n)})$$

Starting from initial point $x^{(0)} = 1$, we can generate the sequence of rational numbers

$$3/1, 11/3, 41/11, ...$$

which converges to $2 + \sqrt{3}$ as $n \to \infty$.Thus, limits of the above sequences is outside the space X and the space is incomplete.

Example 27. *Consider sequence of functions in the space of twice differentiable continuous functions* $C^{(2)}(-\infty,\infty)$

$$f^{(k)}(t)=\frac{1}{2}+\frac{1}{\pi}\tan^{-1}(kt)$$

defined in interval $-\infty<t<\infty$*, for all integers* k*. The range of the function is (0,1). As* $k\rightarrow\infty$*, the sequence of continuous function converges to a discontinuous function*

$$\begin{aligned} u^{(*)}(t) &= 0 \qquad -\infty<t<0 \\ &= 1 \qquad\quad 0<t<\infty \end{aligned}$$

Example 28. *Let* $X=(C[0,1],\|.\|_1)$ *i.e. space of continuous function on* $[0,1]$ *with one norm defined on it i.e.*

$$\|\mathbf{x}(t)\|_1=\int\limits_0^1|\mathbf{x}(t)|\,dt \tag{3.12}$$

and let us define a sequence [**11**]

$$\mathbf{x}^{(n)}(t)=\left\{\begin{array}{ll} 0 & (0\leq t\leq(\frac{1}{2}-\frac{1}{n}) \\ n(t-\frac{1}{2})+1 & (\frac{1}{2}-\frac{1}{n})\leq t\leq\frac{1}{2}) \\ 1 & (t\geq\frac{1}{2}) \end{array}\right\} \tag{3.13}$$

Each member is a continuous function and the sequence is Cauchy as

$$\left\|\mathbf{x}^{(n)}-\mathbf{x}^{(m)}\right\|=\frac{1}{2}\left|\frac{1}{n}-\frac{1}{m}\right|\rightarrow 0 \tag{3.14}$$

However, as can be observed from Figure1, the sequence does not converge to a continuous function.

The concepts of convergence, Cauchy sequences and completeness of space assume importance in the analysis of iterative numerical techniques. Any iterative numerical method generates a sequence of vectors and we have to assess whether the sequence is Cauchy to terminate the iterations.

4. Inner Product Spaces and Hilbert Spaces

Concept of norm explained in the last section generalizes notion of length of a vector in three dimensional Euclidean space. Another important concept in three dimensional space is angle between any two vectors. Given any two unit

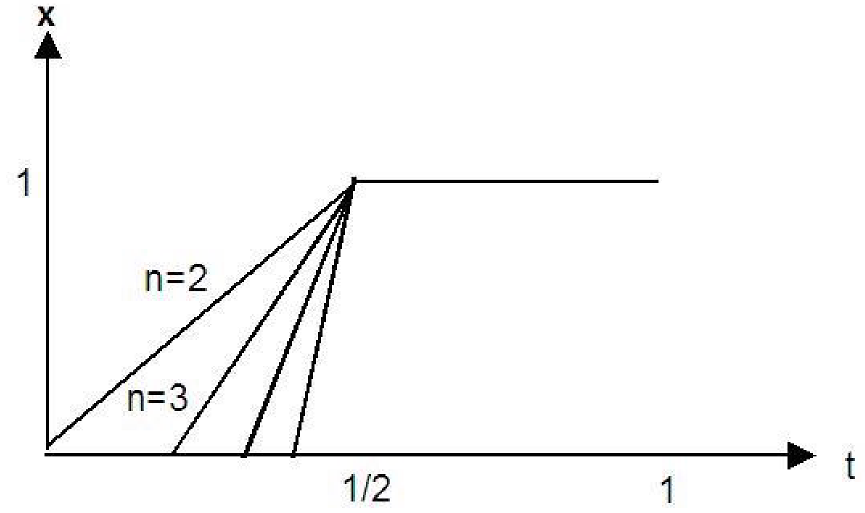


FIGURE 1. Sequence of continuous functions

vectors in R^3, say $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$,the angle between these two vectors is defined using inner (or dot) product of two vectors as

$$\cos(\theta) = (\widehat{\mathbf{x}})^T \widehat{\mathbf{y}} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right)^T \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \tag{4.1}$$

$$= \widehat{x}_1\widehat{y}_1 + \widehat{x}_2\widehat{y}_2 + \widehat{x}_3\widehat{y}_3 \tag{4.2}$$

The fact that cosine of angle between any two unit vectors is always less than one can be stated as

$$|\cos(\theta)| = |\langle\widehat{\mathbf{x}}, \widehat{\mathbf{y}}\rangle| \leq 1 \tag{4.3}$$

Moreover, vectors $\mathbf{x}$ and $\mathbf{y}$ are called orthogonal if $(\mathbf{x})^T\ \mathbf{y} = 0$. Orthogonality is probably the most useful concept while working in three dimensional Euclidean space. Inner product spaces and Hilbert spaces generalize these simple geometrical concepts in three dimensional Euclidean space to higher or infinite dimensional spaces.

DEFINITION 13. ***(Inner Product Space):*** *An inner product space is a linear vector space X together with an inner product defined on $X \times X$. Corresponding to each pair of vectors $\mathbf{x}, \mathbf{y} \in X$ the inner product $\langle\mathbf{x}, \mathbf{y}\rangle$ of $\mathbf{x}$ and $\mathbf{y}$ is a scalar. The inner product satisfies following axioms.*

(1) $\langle\mathbf{x}, \mathbf{y}\rangle = \overline{\langle\mathbf{y}, \mathbf{x}\rangle}$ (complex conjugate)
(2) $\langle\mathbf{x} + \mathbf{y}, \mathbf{z}\rangle = \langle\mathbf{x}, \mathbf{z}\rangle + \langle\mathbf{y}, \mathbf{z}\rangle$

(3) $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \overline{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle$
$\langle \mathbf{x}, \lambda \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$
(4) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \overline{0}$

Axioms 2 and 3 imply that the inner product is linear in first entry. The quantity $\langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}$ is a candidate function for defining norm on the inner product space. Axioms 1 and 3 imply that $\|\alpha \mathbf{x}\| = |\alpha| \ \|\mathbf{x}\|$ and axiom 4 implies that $\|\mathbf{x}\| > 0$ for $\mathbf{x} \neq \overline{0}$. If we show that $\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$satisfies triangle inequality, then $\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ defines a norm on space X . We first prove Cauchy-Schwarz inequality, which is generalization of equation (4.3), and proceed to show that $\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ defines the well known 2-norm on X, i.e. $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Lemma 1. ***(Cauchey- Schwarz Inequality):*** *Let X denote an inner product space. For all $\mathbf{x}, \mathbf{y} \in X$,the following inequality holds*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq [\langle \mathbf{x}, \mathbf{x} \rangle]^{1/2} [\langle \mathbf{y}, \mathbf{y} \rangle]^{1/2} \tag{4.4}$$

The equality holds if and only if $\mathbf{x} = \lambda \mathbf{y}$ or $\mathbf{y} = \overline{0}$

Proof: If $\mathbf{y} = \overline{0}$, the equality holds trivially so we assume $\mathbf{y} \neq \overline{0}$. Then, for all scalars λ,we have

$$0 \leq \langle \mathbf{x} - \lambda \mathbf{y}, \mathbf{x} - \lambda \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle - \lambda \langle \mathbf{x}, \mathbf{y} \rangle - \overline{\lambda} \langle \mathbf{y}, \mathbf{x} \rangle + |\lambda|^2 \langle \mathbf{y}, \mathbf{y} \rangle \tag{4.5}$$

In particular, if we choose $\lambda = \dfrac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle}$, then, using axiom 1 in the definition of inner product, we have

$$\overline{\lambda} = \frac{\overline{\langle \mathbf{y}, \mathbf{x} \rangle}}{\langle \mathbf{y}, \mathbf{y} \rangle} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{4.6}$$

$$\Rightarrow \quad -\lambda \langle \mathbf{x}, \mathbf{y} \rangle - \overline{\lambda} \langle \mathbf{y}, \mathbf{x} \rangle = -\frac{2 \langle \mathbf{x}, \mathbf{y} \rangle \ \langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{4.7}$$

$$= \quad -\frac{2 \langle \mathbf{x}, \mathbf{y} \rangle \overline{\langle \mathbf{x}, \mathbf{y} \rangle}}{\langle \mathbf{y}, \mathbf{y} \rangle} = -\frac{2 |\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{4.8}$$

$$\Rightarrow 0 \leq \langle \mathbf{x}, \mathbf{x} \rangle - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{4.9}$$

$$\text{or} \ \ | \ \langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle}$$

The triangle inequality can be can be established easily using the Cauchy-Schwarz inequality as follows

$$\langle \mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{y}\rangle + \langle \mathbf{y},\mathbf{x}\rangle + \langle \mathbf{y},\mathbf{y}\rangle . \tag{4.10}$$

$$\leq \langle \mathbf{x},\mathbf{x}\rangle + 2\left|\langle \mathbf{x},\mathbf{y}\rangle\right| + \langle \mathbf{y},\mathbf{y}\rangle \tag{4.11}$$

$$\leq \langle \mathbf{x},\mathbf{x}\rangle + 2\sqrt{\langle \mathbf{x},\mathbf{x}\rangle\langle \mathbf{y},\mathbf{y}\rangle} + \langle \mathbf{y},\mathbf{y}\rangle \tag{4.12}$$

$$\sqrt{\langle \mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle} \leq \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} + \sqrt{\langle \mathbf{y},\mathbf{y}\rangle} \tag{4.13}$$

Thus, the candidate function $\sqrt{\langle \mathbf{x},\mathbf{x}\rangle}$ satisfies all the properties necessary to define a norm, i.e.

$$\sqrt{\langle \mathbf{x},\mathbf{x}\rangle} \geq 0 \; \forall \; \mathbf{x} \in X \text{ and } \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} = 0 \;\; iff \;\; \mathbf{x} = \overline{0} \tag{4.14}$$

$$\sqrt{\langle \alpha\mathbf{x},\alpha\mathbf{x}\rangle} = |\alpha| \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} \tag{4.15}$$

$$\sqrt{\langle \mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle} \leq \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} + \sqrt{\langle \mathbf{y},\mathbf{y}\rangle} \qquad \text{(Triangle inequality)} \tag{4.16}$$

Thus, the function $\sqrt{\langle \mathbf{x},\mathbf{x}\rangle}$ indeed defines a norm on the inner product space X. In fact the inner product defines the well known 2-norm on X, i.e.

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} \tag{4.17}$$

and the triangle inequality can be stated as

$$\|\mathbf{x}+\mathbf{y}\|_2^2 \leq \|\mathbf{x}\|_2^2 + 2\,\|\mathbf{x}\|_2 \,.\, \|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 \,. = \left[\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2\right]^2 \tag{4.18}$$

$$\text{or } \|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2 \tag{4.19}$$

Definition 14. *(**Angle**) The angle θ between any two vectors in an inner product space is defined by*

$$\theta = \cos^{-1}\left[\frac{\langle \mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2}\right] \tag{4.20}$$

Example 29. ***Inner Product Spaces***

(1) *Space $X \equiv R^n$ with $\mathbf{x} = (\xi_{1,}\xi_{2,}\xi_{3,}.......\xi_{n,})$ and $\mathbf{y} = (\eta_1, \eta_2........\eta_n)$*

$$\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} \xi_i\eta_i \tag{4.21}$$

$$\langle \mathbf{x},\mathbf{x}\rangle = \sum_{i=1}^{n} (\xi_i)^2 = \|\mathbf{x}\|_2^2 \tag{4.22}$$

(2) *Space $X \equiv R^n$ with $\mathbf{x} = (\xi_{1,}\xi_{2,}\xi_{3,}.......\xi_{n,})$ and $\mathbf{y} = (\eta_1, \eta_2........\eta_n)$*

$$\langle \mathbf{x},\mathbf{y}\rangle_W = \mathbf{x}^T W \mathbf{y} \tag{4.23}$$

where W is a positive definite matrix. The corresponding 2-norm is defined as $\|\mathbf{x}\|_{W,2} = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle_W} = \sqrt{\mathbf{x}^T W \mathbf{x}}$

(3) *Space $X \equiv C^n$ with $\mathbf{x} = (\xi_{1,}\xi_{2,}\xi_{3,}.......\xi_{n,})$ and $\mathbf{y} = (\eta_1, \eta_2........\eta_n)$*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} \overline{\xi}_i \eta_i \tag{4.24}$$

$$\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{n} \overline{\xi}_i \xi_i = \sum_{i=1}^{n} |\xi_i|^2 = \|\mathbf{x}\|_2^2 \tag{4.25}$$

(4) *Space $X \equiv L_2[a, b]$ of real valued square integrable functions on $[a, b]$ with inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_a^b \mathbf{x}(t)\mathbf{y}(t)dt \tag{4.26}$$

is an inner product space and denoted as $L_2[a, b]$. Well known examples of spaces of this type are the set of continuous functions on $[-\pi, \pi]$ or $[0, 2\pi]$,which we consider while developing Fourier series expansions of continuous functions on these intervals using $sin(n\pi)$ and $cos(n\pi)$ as basis functions.

(5) *Space of polynomial functions on $[a, b]$with inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_a^b \mathbf{x}(t)\mathbf{y}(t)dt \tag{4.27}$$

is a inner product space. This is a subspace of $L_2[a, b]$.

(6) *Space of complex valued square integrable functions on $[a, b]$ with inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_a^b \overline{\mathbf{x}}(t)\mathbf{y}(t)dt \tag{4.28}$$

is an inner product space.

Definition 15. ***(Hilbert Space):*** *A complete inner product space is called as an Hilbert space.*

Definition 16. ***(Orthogonal Vectors):*** *In a inner product space X two vector $\mathbf{x}, \mathbf{y} \in X$ are said to be orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0.$We symbolize this by $\mathbf{x} \perp \mathbf{y}$. A vector $\mathbf{x}$ is said to be orthogonal to a set S (written as $\mathbf{x} \perp s$) if $\mathbf{x} \perp \mathbf{z}$ for each $\mathbf{z} \in S$.*

Just as orthogonality has many consequences in plane geometry, it has many implications in any inner-product space [**11**]. The Pythagoras theorem, which is probably the most important result the plane geometry, is true in any inner product space.

LEMMA 2. *If* $\mathbf{x}\perp\mathbf{y}$ *in an inner product space then* $\|\mathbf{x}+\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2$
.

Proof: $\|\mathbf{x}+\mathbf{y}\|_2^2 = \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{x}\rangle$.

DEFINITION 17. ***(Orthogonal Set):*** *A set of vectors* S *in an inner product space* X *is said to be an orthogonal set if* $\mathbf{x}\perp\mathbf{y}$ *for each* $\mathbf{x}, \mathbf{y} \in S$ *and* $\mathbf{x} \neq \mathbf{y}$*. The set is said to be orthonormal if, in addition each vector in the set has norm equal to unity.*

Note that an orthogonal set of nonzero vectors is linearly independent set. We often prefer to work with an orthonormal basis as any vector can be uniquely represented in terms of components along the orthonormal directions. Common examples of such orthonormal basis are (a) unit vectors along coordinate directions in R^n (b) $sin(nt)$ and $cos(nt)$ functions in $C[0, 2\pi]$.

4.1. Gram-Schmidt procedure. Given any linearly independent set in an inner product space, it is possible to construct an orthonormal set. This procedure is called Gram-Schmidt procedure. Consider a linearly independent set of vectors $\left\{\mathbf{x}^{(i)}; i = 1, 2, 3.....n\right\}$ in a inner product space we define $\mathbf{e}^{(1)}$ as

$$\mathbf{e}^{(1)} = \frac{\mathbf{x}^{(1)}}{\|\mathbf{x}^{(1)}\|_2} \tag{4.29}$$

We form unit vector $\mathbf{e}^{(2)}$ in two steps.

$$\mathbf{z}^{(2)} = \mathbf{x}^{(2)} - \left\langle \mathbf{x}^{(2)}, \mathbf{e}^{(1)} \right\rangle \mathbf{e}^{(1)} \tag{4.30}$$

where$\left\langle \mathbf{x}^{(2)}, \mathbf{e}^{(1)} \right\rangle$ is component of $\mathbf{x}^{(2)}$ along $\mathbf{e}^{(1)}$.

$$\mathbf{e}^{(2)} = \frac{\mathbf{z}^{(2)}}{\|\mathbf{z}^{(2)}\|_2} \tag{4.31}$$

.By direct calculation it can be verified that $\mathbf{e}^{(1)}\perp\mathbf{e}^{(2)}$. The remaining orthonormal vectors $\mathbf{e}^{(i)}$ are defined by induction. The vector $\mathbf{z}^{(k)}$ is formed according to the equation

$$\mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \sum_{i=1}^{k-1} \left\langle \mathbf{x}^{(k)}, \mathbf{e}^{(i)} \right\rangle .\mathbf{e}^{(i)} \tag{4.32}$$

and

$$\mathbf{e}^{(k)} = \frac{\mathbf{z}^{(k)}}{\left\|\mathbf{z}^{(k)}\right\|_2} \quad ; \qquad k = 1, 2, ...n \tag{4.33}$$

Again it can be verified by direct computation that $\mathbf{z}^{(k)} \perp \mathbf{e}^{(i)}$ for all $i < k$.

EXAMPLE 30. ***Gram-Schmidt Procedure in*** $\boldsymbol{R}^3$: *Consider* $X = R^3$ *with* $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$. *Given a set of three linearly independent vectors in* R^3

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} ; \ \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} ; \ \mathbf{x}^{(3)} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \tag{4.34}$$

we want to construct and orthonormal set. Applying Gram Schmidt procedure,

$$\mathbf{e}^{(1)} = \frac{\mathbf{x}^{(1)}}{\left\|\mathbf{x}^{(1)}\right\|_2}. = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \tag{4.35}$$

$$\begin{aligned} \mathbf{z}^{(2)} &= \mathbf{x}^{(2)} - \left\langle \mathbf{x}^{(2)}, \mathbf{e}^{(1)} \right\rangle .\mathbf{e}^{(1)} \\ &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ -\frac{1}{2} \end{bmatrix} \end{aligned} \tag{4.36}$$

$$\mathbf{e}^{(2)} = \frac{\mathbf{z}^{(2)}}{\left\|\mathbf{z}^{(2)}\right\|_2}. = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{4.37}$$

$$\begin{aligned} \mathbf{z}^{(3)} &= \mathbf{x}^{(3)} - \left\langle \mathbf{x}^{(3)}, \mathbf{e}^{(1)} \right\rangle .\mathbf{e}^{(1)} - \left\langle \mathbf{x}^{(3)}, \mathbf{e}^{(2)} \right\rangle .\mathbf{e}^{(2)} \\ &= \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} - \sqrt{2} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} - \sqrt{2} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \end{aligned} \tag{4.38}$$

$$\mathbf{e}^{(3)} = \frac{\mathbf{z}^{(3)}}{\left\|\mathbf{z}^{(3)}\right\|_2}. = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$$

Note that the vectors in the orthonormal set will depend on the definition of inner product. Suppose we define the inner product as follows

$$\langle \mathbf{x}, \mathbf{y} \rangle_W = \mathbf{x}^T W \mathbf{y} \tag{4.39}$$

$$W = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}$$

where W is a positive definite matrix. Then, length of $\left\|\mathbf{x}^{(1)}\right\|_{W,2} = \sqrt{6}$ and the unit vector $\widehat{\mathbf{e}}^{(1)}$ becomes

$$\widehat{\mathbf{e}}^{(1)} = \frac{\mathbf{x}^{(1)}}{\|\mathbf{x}^{(1)}\|_{W,2}}. = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ 0 \\ \frac{1}{\sqrt{6}} \end{bmatrix} \tag{4.40}$$

The remaining two orthonormal vectors have to be computed using the inner product defined by equation 4.39.

Example 31. ***Gram-Schmidt Procedure in C[a,b]***: *Let X represent set of continuous functions on interval $-1 \leq t \leq 1$ with inner product defined as*

$$\langle \mathbf{x}(t), \mathbf{y}(t)\rangle = \int_{-1}^{1} \mathbf{x}(t)\mathbf{y}(t)dt \tag{4.41}$$

Given a set of four linearly independent vectors

$$\mathbf{x}^{(1)}(t) = 1; \quad \mathbf{x}^{(2)}(t) = t; \quad \mathbf{x}^{(3)}(t) = t^2; \quad \mathbf{x}^{(4)}(t) = t^3 \tag{4.42}$$

we intend to generate an orthonormal set. Applying Gram-Schmidt procedure

$$\mathbf{e}^{(1)}(t) = \frac{\mathbf{x}^{(1)}(t)}{\|\mathbf{x}^{(1)}(t)\|} = \frac{1}{\sqrt{2}} \tag{4.43}$$

$$\left\langle \mathbf{e}^{(1)}(t), \mathbf{x}^{(2)}(t)\right\rangle = \int_{-1}^{1} \frac{t}{2} dt = 0 \tag{4.44}$$

$$\mathbf{z}^{(2)}(t) = t - \left\langle \mathbf{x}^{(2)}, \mathbf{e}^{(1)}\right\rangle .\mathbf{e}^{(1)} = t = \mathbf{x}^{(2)}(t) \tag{4.45}$$

$$\mathbf{e}^{(2)} = \frac{\mathbf{z}^{(2)}}{\|\mathbf{z}^{(2)}\|} \tag{4.46}$$

$$\left\|\mathbf{z}^{(2)}(t)\right\|^2 = \int_{-1}^{1} t^2 dt = \left[\frac{t^3}{3}\right]_{-1}^{1} = \frac{2}{3} \tag{4.47}$$

$$\left\|\mathbf{z}^{(2)}(t)\right\| = \sqrt{\frac{2}{3}} \tag{4.48}$$

$$\mathbf{e}^{(2)}(t) = \sqrt{\frac{3}{2}}.t \tag{4.49}$$

$$\begin{aligned}
\mathbf{z}^{(3)}(t) &= \mathbf{x}^{(3)}(t) - \left\langle \mathbf{x}^{(3)}(t), \mathbf{e}^{(1)}(t) \right\rangle .\mathbf{e}^{(1)}(t) - \left\langle \mathbf{x}^{(3)}(t), \mathbf{e}^{(2)}(t) \right\rangle .\mathbf{e}^{(2)}(t) \\
&= t^2 - \frac{1}{2}\left(\int_{-1}^{1} t^2 dt \right) \mathbf{e}^{(1)}(t) - \left(\sqrt{\frac{3}{2}} \int_{-1}^{1} t^3 dt \right) \mathbf{e}^{(2)}(t) \\
&= t^2 - \frac{1}{3} - 0 = t^2 - \frac{1}{3} \qquad (4.50)
\end{aligned}$$

$$\mathbf{e}^{(3)}(t) = \frac{\mathbf{z}^{(3)}(t)}{\left\| \mathbf{z}^{(3)}(t) \right\|} \qquad (4.51)$$

$$\begin{aligned}
\textit{where } \left\| \mathbf{z}^{(3)}(t) \right\|^2 &= \left\langle \mathbf{z}^{(3)}(t), \mathbf{z}^{(3)}(t) \right\rangle = \int_{-1}^{1} \left(t^2 - \frac{1}{3} \right)^2 dt \qquad (4.52) \\
&= \int_{-1}^{1} \left(t^4 - \frac{2}{3}t^2 + \frac{1}{9} \right) dt = \left[\frac{t^5}{5} - \frac{2t^3}{9} + \frac{t}{9} \right]_{-1}^{1} \\
&= \frac{2}{3} - \frac{4}{9} + \frac{2}{9} = \frac{18 - 10}{45} = \frac{8}{45}
\end{aligned}$$

$$\left\| \mathbf{z}^{(3)}(t) \right\| = \sqrt{\frac{8}{45}} = \frac{2}{3}\sqrt{\frac{2}{5}} \qquad (4.53)$$

The orthonormal polynomials constructed above are well known Legandre polynomials. It turns out that

$$\mathbf{e}_n(t) = \sqrt{\frac{2n+1}{2}} p_n(t) \quad ; \quad (n = 0, 1, 2.......) \qquad (4.54)$$

where

$$P_n(t) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dt^n} \left\{ \left(1 - t^2 \right)^n \right\} \qquad (4.55)$$

are Legandre polynomials. It can be shown that this set of polynomials forms a orthonormal basis for the set of continuous functions on [-1,1].

Example 32. ***Gram-Schmidt Procedure in other Spaces***

(1) ***Shifted Legandre polynomials:*** *$X = C[0, 1]$ and inner product defined as*

$$\langle \mathbf{x}(t), \mathbf{y}(t) \rangle = \int_{0}^{1} \mathbf{x}(t)\mathbf{y}(t) dt \qquad (4.56)$$

(2) ***Hermite Polynomials:*** *$X \equiv L^2(-\infty,\infty)$, i.e. space of continuous functions over $(-\infty,\infty)$ with 2 norm defined on it and*

$$\langle \mathbf{x}(t), \mathbf{y}(t)\rangle = \int_{-\infty}^{\infty} \mathbf{x}(t)\mathbf{y}(t)dt \tag{4.57}$$

Apply Gram-Schmidt to the following set of vectors in $L^2(-\infty,\infty)$

(3) ***Laguerre Polynomials:*** *$X \equiv L^2(0,\infty)$, i.e. space of continuous functions over $(0,\infty)$ with 2 norm defined on it and*

$$\langle \mathbf{x}(t), \mathbf{y}(t)\rangle = \int_{0}^{\infty} \mathbf{x}(t)\mathbf{y}(t)dt \tag{4.58}$$

Apply Gram-Schmidt to the following set of vectors in $L^2(0,\infty)$

$$\mathbf{x}^{(1)}(t) = \exp(-\frac{t}{2}) \;;\; \mathbf{x}^{(2)}(t) = tx^{(1)}(t)\,; \tag{4.59}$$

$$\mathbf{x}^{(3)}(t) = t^2\mathbf{x}^{(1)}(t)\,; ......\mathbf{x}^{(k)}(t) = t^{k-1}\mathbf{x}^{(1)}(t)\,; .... \tag{4.60}$$

The first few Laguerre polynomials are as follows

$$L_0(t) = 1 \;\;;\;\; L_1(t) = 1 - t \;\;;\;\; L_2(t) = 1 - 2t + (1/2)t^2....$$

5. Problem Formulation, Transformation and Convergence

5.1. One Fundamental Problem with Multiple Forms. Using the generalized concepts of vectors and vector spaces discussed above, we can look at mathematical models in engineering as transformations, which map a subset of vectors from one vector space to a subset in another space.

Definition 18. ***(Transformation):****Let X and Y be linear spaces and let M be subset of X. A rule which associates with every element $\mathbf{x} \in M$ to an element $\mathbf{y} \in Y$ is said to be transformation from X to Y with domain M. If $\mathbf{y}$ corresponds to $\mathbf{x}$ under the transformation we write $\mathbf{y} = \mathcal{F}(\mathbf{x})$ where $\mathcal{F}(.)$ is called an operator.*

The set of all elements for which and operator $\mathcal{F}$ is defined is called as *domain* of $\mathcal{F}$ and set of all elements generated by transforming elements in domain by $\mathcal{F}$ are called as range of $\mathcal{F}$. If for every $\mathbf{y} \in Y$, there is utmost one $\mathbf{x} \in M$ for which $\mathcal{F}(\mathbf{x}) = \mathbf{y}$, then $\mathcal{F}(.)$ is said to be one to one. If for every $\mathbf{y} \in Y$ there is at least one $\mathbf{x} \in M$, then $\mathcal{F}$ is said to map M onto Y.

Definition 19. ***(Linear Transformations):*** *A transformation $\mathcal{F}$ mapping a vector space X into a vector space$Y$ is said to be linear if* ***for every*** *$\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in X$ and all scalars α, β we have*

$$\mathcal{F}(\alpha x^{(1)} + \beta x^{(2)}) = \alpha\mathcal{F}(\mathbf{x}^{(1)}) + \beta\mathcal{F}(\mathbf{x}^{(2)}). \tag{5.1}$$

Note that any transformation that does not satisfy above definition is not a linear transformation.

Definition 20. ***(Continuous Transformation):*** *A transformation $F : X \rightarrow Y$ is continuous at print $\mathbf{x}^{(0)} \in X$ if and only if $\left\{\mathbf{x}^{(n)}\right\} \rightarrow \mathbf{x}^{(0)}$ implies $F(\mathbf{x}^{(n)}) \rightarrow F\left(\mathbf{x}^{(0)}\right)$. If $F(.)$ is continuous at each $\mathbf{x}^{(0)} \in X$, then we say that the function is a continuous function.*

Example 33. ***Operators***

(1) *Consider transformation*

$$\mathbf{y} = A\mathbf{x} \tag{5.2}$$

where $\mathbf{y} \in \mathbf{R}^m$ and $\mathbf{x} \in \mathbf{R}^n$ and $A \in \mathbf{R}^m \times \mathbf{R}^n$. Whether this mapping in onto R^n depends on the rank of the matrix. It is easy to check that A is a linear operator. However, contrary to the common perception, the transformation

$$\mathbf{y} = A\mathbf{x} + \mathbf{b} \tag{5.3}$$

does not satisfy equation (2.1) and does not qualify as a linear transformation.

(2) *$d/dt(.)$ is an operator from the space of continuously differentiable functions to the space of continuous function.*

(3) *The operator $\int_0^1 [.]dt$ maps space of integrable functions into R.*

A large number of problems arising in applied mathematics can be stated as follows [**9**]: Solve equation

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) \tag{5.4}$$

where $\mathbf{x} \in X, \mathbf{y} \in Y$ are linear vector spaces and operator $\mathcal{F} : X \rightarrow Y.$In engineering parlance, $\mathbf{x}, \mathbf{y}$ and $\mathcal{F}$ represent input, output and model, respectively. Linz [**9**] proposes following broad classification of problems encountered in computational mathematics

- **Direct Problems:** Given operator $\mathcal{F}$ and $\mathbf{x}$, find $\mathbf{y}$.In this case, we are trying to compute output of a given system given input. The computation of definite integrals is an example of this type.

- **Inverse Problems:**Given operator $\mathcal{F}$ and $\mathbf{y}$, find $\mathbf{x}$. In this case we are looking for input which generates observed output. Solving system of simultaneous (linear / nonlinear) algebraic equations, ordinary and partial differential equations and integral equations are examples of this category
- **Identification problems:**Given operator $\mathbf{x}$ and $\mathbf{y}$, find $\mathcal{F}$. In this case, we try to find the laws governing systems from the knowledge of relation between in the inputs and outputs.

The direct problems can be treated relatively easily. Inverse problems and identification problems are relatively difficult to solve and form the central theme of the numerical analysis.

Once we understand the generalized concepts of vector spaces, norms, operators etc., we can see that the various inverse problems under consideration are not fundamentally different. For example,

- Linear algebraic equations : $X \equiv R^n$

$$A\mathbf{x} = \mathbf{b} \tag{5.5}$$

can be rearranged as

$$F(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = \overline{0} \tag{5.6}$$

- Nonlinear algebraic equations: $X \equiv R^n$

$$\begin{aligned} F(\mathbf{x}) &= \overline{0} \\ F(\mathbf{x}) &= \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & ... & f_n(\mathbf{x}) \end{bmatrix}_{n\times 1}^T \end{aligned} \tag{5.7}$$

- ODE–IVP: $X \equiv C^n[0, \infty)$

$$\begin{bmatrix} \dfrac{d\mathbf{x}(t)}{dt} - F(\mathbf{x}(t), t) \\ \mathbf{x}(0) - \mathbf{x}_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{5.8}$$

can be expressed in an abstract form as

$$\mathcal{F}[\mathbf{x}(t)] = \overline{0} \quad ; \quad \mathbf{x}(t) \in C^n[0, \infty) \tag{5.9}$$

- ODE-BVP: $X \equiv C^{(2)}[0, 1]$

$$\Psi[d^2u/dz^2, du/dz, u, z] = 0 \;\; ; \qquad (0 < z < 1) \tag{5.10}$$

Boundary Conditions

$$f_1[du/dz, u, z] = 0 \; at \; z = 0 \tag{5.11}$$

$$f_2[du/dz, u, z] = 0 \;\; at\; z = 1 \tag{5.12}$$

which can be written in abstract form

$$\mathcal{F}\left[u(t)\right] = \;\; \overline{0} \quad ; \quad u(t) \in C^{\,(2)}[0,1]$$

Here, the operator $\mathcal{F}\left[u(t)\right]$ consists of the differential operator $\Psi(.)$ together with the boundary conditions and $C^{\,(2)}[0,1]$ represents set of twice differentiable continuous functions.

As evident from the above abstract representations, all the problems can be reduced to one fundamental equation of the form

$$\mathcal{F}(\mathbf{x}) = \overline{0} \tag{5.13}$$

where $\mathbf{x}$ represents a vector in the space under consideration. It is one fundamental problem, which assumes different forms in different context and different vector spaces. Viewing these problems in a unified framework facilitates better understanding of the problem formulations and the solution techniques.

5.2. Numerical Solution. Given any general problem of the form (5.13), the approach to compute a solution to this problem typically consists of two steps

- **Problem Transformation:** In this step, the original problem is transformed into a one of the known standard forms, for which numerical tools are available.
- **Computation of Numerical Solution:** Use a standard tool or a combination of standard tools to construct a numerical solution to the transformed problem. The three most commonly used tools are (a) linear algebraic equation solvers (b) ODE-IVPs solvers (c) numerical optimization. A schematic diagram of the generic solution procedure is presented in Figure (2). In the sub-section that follows, we will describe two most commonly used tools for problem transformation. The tools used for constructing solutions and their applications in different context will form the theme for the rest of the book.

5.3. Tools for Problem Transformation. As stated earlier, given a system of equations (5.13), the main concern in numerical analysis is to come up with an iteration scheme of the form

$$\mathbf{x}^{(k+1)} = G\left[\mathbf{x}^{(k)}\right] \quad ; \quad k = 0, 1, 2, ... \tag{5.14}$$

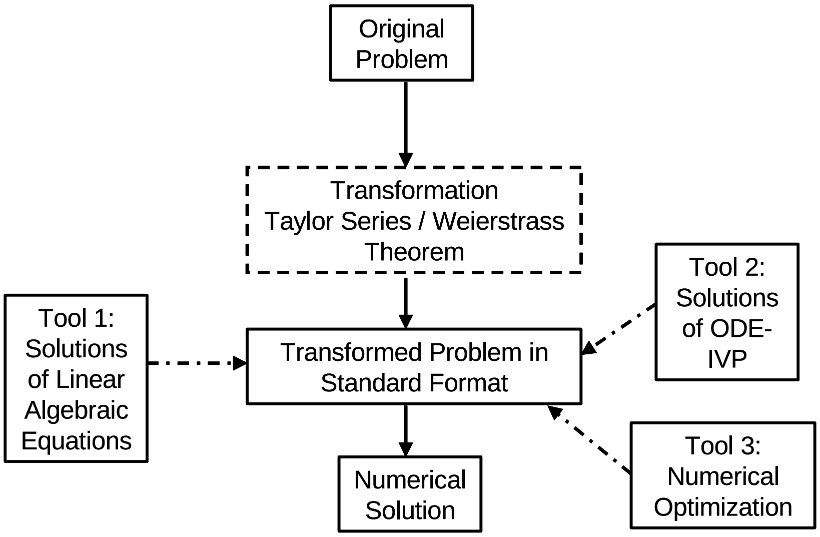

Original Problem
Transformation Taylor Series / Weierstrass Theorem
Tool 1: Solutions of Linear Algebraic Equations
Transformed Problem in Standard Format
Tool 2: Solutions of ODE-IVP
Tool 3: Numerical Optimization
Numerical Solution

FIGURE 2

where in such a way that the solution $\mathbf{x}^*$ of equation (6.5), i.e.

$$\mathbf{x}^* = G\left[\mathbf{x}^*\right] \tag{5.15}$$

satisfies

$$\mathcal{F}(\mathbf{x}^*) = \overline{0} \tag{5.16}$$

Key to such transformations is approximation of continuous functions using polynomials. The following two important results from functional analysis play pivotal role in the development of many iterative numerical schemes. The Taylor series expansion helps us develop local linear or quadratic approximations of a function vector in the neighborhood of a point in a vector space. The Weierstrass approximation theorem provides basis for approximating any arbitrary continuous function over a finite interval by polynomial approximation. These two results are stated here without giving proofs.

5.3.1. *Local approximation by Taylor series expansion.* Any scalar function $f(\mathbf{x}) : R \rightarrow R$, which is continuously differentiable n times at $\mathbf{x} = \overline{\mathbf{x}}$, the Taylor series expansion of this function in the neighborhood the point $\mathbf{x} = \overline{\mathbf{x}}$ can be expressed as

$$f(\mathbf{x}) = f(\overline{\mathbf{x}}) + \left[\frac{\partial f(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right]\delta\mathbf{x} + \frac{1}{2!}\left[\frac{\partial^2 f(\overline{\mathbf{x}})}{\partial \mathbf{x}^2}\right](\delta\mathbf{x})^2 + ... \tag{5.17}$$

$$.... + \frac{1}{n!}\left[\frac{\partial^n f(\overline{\mathbf{x}})}{\partial \mathbf{x}^n}\right].(\delta\mathbf{x})^n + \mathbf{R}_n(\overline{\mathbf{x}},\delta\mathbf{x}) \tag{5.18}$$

$$R_n(\overline{\mathbf{x}},\delta\mathbf{x}) = \frac{1}{(n+1)!}\frac{\partial^{n+1} f(\overline{\mathbf{x}}+\lambda\delta\mathbf{x})}{\partial \mathbf{x}^{n+1}}(\delta\mathbf{x})^{n+1} \quad ; \quad (0<\lambda<1)$$

While developing numerical methods, we require more general Taylor series expansions for multi-dimensional cases. We consider following two multidimensional cases

- **Case A: Scalar Function** $f(\mathbf{x}) : R^n \rightarrow R$

$$
\begin{aligned}
&(5.19) & f(\mathbf{x}) &= f(\overline{\mathbf{x}}) + \left[\nabla f(\overline{\mathbf{x}})\right]^T \delta\mathbf{x} \\
&(5.20) & & +\frac{1}{2!}\delta\mathbf{x}^T \left[\nabla^2 f(\overline{\mathbf{x}})\right] \delta\mathbf{x} + R_3(\overline{\mathbf{x}}, \delta\mathbf{x}) \\
& & \nabla f(\overline{\mathbf{x}}) &= \left[\frac{\partial f(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right] \\
&(5.21) & &= \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} & \ldots\ldots & \dfrac{\partial f}{\partial x_n} \end{bmatrix}^T_{\mathbf{x}=\overline{\mathbf{x}}} \\
& & \nabla^2 f(\overline{\mathbf{x}}) &= \left[\frac{\partial^2 f(\overline{\mathbf{x}})}{\partial \mathbf{x}^2}\right] \\
&(5.22) & &= \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \ldots\ldots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \ldots\ldots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \ldots\ldots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}_{\mathbf{x}=\overline{\mathbf{x}}} \\
& & R_3(\overline{\mathbf{x}}, \delta\mathbf{x}) &= \frac{1}{3!}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^3 f(\overline{\mathbf{x}} + \lambda\delta\mathbf{x})}{\partial x_i \partial x_j \partial x_k}\delta x_i \delta x_j \delta x_k \qquad ; \qquad (0 < \lambda < 1)
\end{aligned}
$$

Note that the gradient $\nabla f(\overline{\mathbf{x}})$ is an $n \times 1$ vector and the Hessian $\nabla^2 f(\overline{\mathbf{x}})$ is an $n \times n$ matrix..

Example 34. *Consider the function vector* $f(\mathbf{x}) : R^2 \rightarrow R$

$$f(\mathbf{x}) = x_1^2 + x_2^2 + e^{(x_1+x_2)}$$

which can be approximated in the neighborhood of $\overline{\mathbf{x}} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ *using the Taylor series expansion as*

$$\begin{aligned} f(\mathbf{x}) &= f(\overline{\mathbf{x}}) + \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \end{bmatrix}_{\mathbf{x}=\overline{\mathbf{x}}} \delta\mathbf{x} \\ & + [\delta\mathbf{x}]^T \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} \end{bmatrix}_{\mathbf{x}=\overline{\mathbf{x}}} \delta\mathbf{x} + R_3(\overline{\mathbf{x}}, \delta\mathbf{x}) \end{aligned} \tag{5.23}$$

$$\begin{aligned} &= 2(1+e^2) + \begin{bmatrix} (2+e^2) & (2+e^2) \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \\ & + \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix}^T \begin{bmatrix} (2+e^2) & e^2 \\ e^2 & (2+e^2) \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + R_3(\overline{\mathbf{x}}, \delta\mathbf{x}) \end{aligned} \tag{5.24}$$

- ***Case B: Function vector*** $F(\mathbf{x}) : R^n \to R^n$

$$F(\mathbf{x}) = F(\overline{\mathbf{x}}) + \left[\frac{\partial F(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right] \delta\mathbf{x} + R_2(\overline{\mathbf{x}}, \delta\mathbf{x}) \tag{5.25}$$

$$\left[\frac{\partial F(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right] = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \ldots\ldots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \ldots\ldots & \dfrac{\partial f_2}{\partial x_n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \ldots\ldots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}_{\mathbf{x}=\overline{\mathbf{x}}}$$

Here matrix $\left[\frac{\partial F(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right]$ *is called as Jackobian.*

Example 35. *Consider the function vector* $F(\mathbf{x}) \in R^2$

$$F(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 + 2x_1 x_2 \\ x_1 x_2 e^{(x_1+x_2)} \end{bmatrix}$$

which can be approximated in the neighborhood of $\overline{\mathbf{x}} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ *using the Taylor series expansion as*

$$\begin{aligned} F(\mathbf{x}) &= \begin{bmatrix} f_1(\overline{\mathbf{x}}) \\ f_2(\overline{\mathbf{x}}) \end{bmatrix} + \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix}_{\mathbf{x}=\overline{\mathbf{x}}} \delta\mathbf{x} + R_2(\overline{\mathbf{x}}, \delta\mathbf{x}) \\ &= \begin{bmatrix} 4 \\ \mathbf{e}^2 \end{bmatrix} + \begin{bmatrix} 4 & 4 \\ 2\mathbf{e}^2 & 2\mathbf{e}^2 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + R_2(\overline{\mathbf{x}}, \delta\mathbf{x}) \end{aligned}$$

A classic application of this theorem in the numerical analysis is Newton-Raphson method for solving set of simultaneous nonlinear algebraic equations. Consider set of n nonlinear equations

$$f_i(\mathbf{x}) = 0\,; \quad i = 1, ..., n \tag{5.26}$$

$$\text{or } F(\mathbf{x}) = \overline{0} \tag{5.27}$$

which have to be solved simultaneously. Suppose $\mathbf{x}^*$ is a solution such that $F(\mathbf{x}^*) = \overline{0}$. If each function $f_i(\mathbf{x})$ is continuously differentiable, then, in the neighborhood of $\mathbf{x}^*$ we can approximate its behavior by Taylor series, as

$$\mathbf{F}(\mathbf{x}^*) = \mathbf{F}\left[\mathbf{x}^{(k)} + \left(\mathbf{x}^* - \mathbf{x}^{(k)}\right)\right] \cong F(\mathbf{x}^{(k)}) + \left[\frac{\partial F}{\partial x}\right]_{\mathbf{x}=\mathbf{x}^{(k)}} \left[\Delta \mathbf{x}^{(k)}\right] = \overline{0} \tag{5.28}$$

Solving above linear equation yields the iteration sequence

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[\left[\frac{\partial F}{\partial x}\right]_{\mathbf{x}=\mathbf{x}^{(k)}}\right]^{-1} F\left[\mathbf{x}^{(k)}\right] \tag{5.29}$$

We will discuss this approach in greater detail in the next chapter.

5.3.2. *Polynomial approximation over an interval.* Given an arbitrary continuous function over an interval, can we approximate it with another "**simple**" function with arbitrary degree of accuracy? This question assumes significant importance while developing many numerical methods. In fact, this question can be posed in any general vector space. We often use such **simple** approximations while performing computations. The classic examples of such approximations are use of a rational number to approximate an irrational number (e.g. 22/7 is used in place of π or series expansion of number e) or polynomial approximation of a continuous function. This subsections discusses rationale behind such approximations.

DEFINITION 21. ***(Dense Set)*** *A set D is said to be dense in a normed space X, if for each element $\mathbf{x} \in X$ and every $\varepsilon > 0$, there exists an element $\mathbf{d} \in D$ such that $\|\mathbf{x} - \mathbf{d}\| < \varepsilon$.*

Thus, if set D is dense in X, then there are points of D arbitrary close to any element of X. Given any $\mathbf{x} \in X$, a sequence can be constructed in D which converges to $\mathbf{x}$. Classic example of such a dense set is the set of rational numbers in the real line. Another dense set, which is widely used for approximations, is the set of polynomials. This set is dense in $C[a,b]$ and any continuous function $f(t) \in C[a,b]$ can be approximated by a polynomial function $p(t)$ with an arbitrary degree of accuracy

Theorem 1. *(**Weierstrass Approximation Theorem**): Consider space $C[a,b]$, the set of all continuous functions over interval $[a,b]$, together with $\infty-$norm defined on it as*

$$\|f(t)\| = \begin{matrix} \max \\ t \in [a,b] \end{matrix} \ |f(t)| \tag{5.30}$$

Given any $\varepsilon > 0,$ for every $f(t) \in C[a,b]$ there exists a polynomial $p(t)$ such that $|f(t) - p(t)| < \varepsilon$ for all $t \in [a,b]$.

This fundamental result forms the basis of many numerical techniques.

Example 36. *We routinely approximate unknown continuous functions using polynomial approximation. For example, specific heat of a pure substance at constant pressure (C_p) is approximated as a polynomial function of temperature*

$$C_p \simeq a + bT + cT^2 \tag{5.31}$$

$$or \quad C_p \simeq a + bT + cT^2 + dT^3 \tag{5.32}$$

While developing such empirical correlations over some temperature interval $[T_1, T_2]$, it is sufficient to know that (C_p) is some continuous function of temperature over this interval. Even though we do not know exact form of this functional relationship, we can invoke Weierstrass theorem and approximate it as a polynomial function.

Example 37. *Consider a first order ODE-IVP*

$$\frac{dy}{dz} = -5y + 2y^3 \,; \;\; y(0) = y_0 \tag{5.33}$$

and we want to generate profile $y^*(z)$ over interval $[0,1]$. It is clear that the true solution to the problem $y^*(z) \in C^{(1)}[0,1]$, i.e. the set of once differentiable continuous functions over interval $[0,1]$. Suppose the true $y^*(z)$ is difficult to generate analytically and we want to compute an approximate solution, say $y(z)$, for the ODE-IVP under consideration. Applying Weierstrass theorem, we can choose a polynomial approximation to $y^*(z)$ as

$$y(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 \tag{5.34}$$

Substituting (5.34) in ODE, we get

$$\begin{aligned} a_1 + 2a_2 z + 3a_3 z^2 &= -5(a_0 + a_1 z + a_2 z^2 + a_3 z^3) \\ &\quad +2(a_0 + a_1 z + a_2 z^2 + a_3 z^3)^3 \end{aligned} \tag{5.35}$$

Now, using the initial condition, we have

$$y(0) = y_0 \Rightarrow a_0 = y_0$$

In order to estimate the remaining three coefficients, we force residual $R(z)$ obtained by rearranging equation (5.35)

$$\begin{aligned} R(z) &= 5a_0 + (2a_2 + 5a_1)z + (3a_3 + 5a_2)z^2 \\ &\quad +5a_3 z^3 - 2(a_0 + a_1 z + a_2 z^2 + a_3 z^3)^3 \end{aligned} \tag{5.36}$$

zero at two intermediate points, say $z = z_1$ and $z = z_2$, and at $z = 1$, i.e.

$$R(z_1) = 0\,; \quad R(z_2) = 0\,; \, R(1) = 0\,;$$

This gives nonlinear equations in three unknowns, which can be solved simultaneously to compute $a_{1,}a_2$ and a_3. In effect, the Weierstrass theorem has been used to convert an ODE-IVP to a set of nonlinear algebraic equations.

6. Summary

In this chapter, we review important concepts from functional analysis and linear algebra, which form the basis of synthesis and analysis the numerical methods. We begin with the concept of a general vector space and define various algebraic and geometric structures like norm and inner product. We also interpret the notion of orthogonality in a general inner product space and develop Gram-Schmidt process, which can generate an orthonormal set from a linearly independent set. We later introduce two important results from analysis, namely the Taylor's theorem and Weierstrass approximation theorems, which play pivotal role in formulation of iteration schemes . We then proceed to develop theory for analyzing convergence of linear and nonlinear iterative schemes using eigen value analysis and contraction mapping principle, respectively. In the end, we establish necessary and sufficient conditions for optimality of a scalar valued function, which form basis of the optimization based numerical approaches.

7. Exercise

(1) While solving problems using a digital computer, arithmetic operations can be performed only with a limited precision due to finite word length. Consider the vector space $X \equiv R$ and discuss which of the laws of algebra (associative, distributive, commutative) are not satisfied for the floating point arithmetic in a digital computer.

(2) Show that the solution of the differential equation

$$\frac{d^2x}{dt^2} + x = 0$$

is a linear space. What is the dimension of this space?

(3) Show that functions 1, exp(t), exp(2t), exp(3t) are linearly independent over any interval [a,b].

(4) Does the set of functions of the form

$$f(t) = 1/(a + bt)$$

constitute a linear vector space?

(5) Give an example of a function which is in $\mathbf{L}_1[0,1]$ but not in $\mathbf{L}_2[0,1]$.

(6) Decide linear dependence or independence of
  (a) (1,1,2), (1,2,1), (3,1,1)
  (b) $\left(\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\right), \left(\mathbf{x}^{(2)} - \mathbf{x}^{(3)}\right), \left(\mathbf{x}^{(3)} - \mathbf{x}^{(4)}\right), \left(\mathbf{x}^{(4)} - \mathbf{x}^{(1)}\right)$ for any $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}$
  (c) (1,1,0), (1,0,0), (0,1,1), (x,y,z) for any scalars x,y,z

(7) Describe geometrically the subspaces of R^3 **spanned** by following sets
  (a) (0,0,0), (0,1,0), (0,2,0)
  (b) (0,0,1), (0,1,1), (0,2,0)
  (c) all six of these vectors
  (d) set of all vectors with positive components

(8) Consider the space X of all $n \times n$ matrices. Find a basis for this vector space and show that set of all lower triangular $n \times n$ matrices forms a subspace of X.

(9) Determine which of the following definitions are valid as definitions for norms in $\mathbf{C}^{(2)}[a,b]$
  (a) $max\,|\mathbf{x}(t)| + \max|\mathbf{x}'(t)|$
  (b) $\max|\mathbf{x}'(t)|$
  (c) $|\mathbf{x}(a)| + \max|\mathbf{x}'(t)|$
  (d) $|\mathbf{x}(a)| \max|\mathbf{x}(t)|$

(10) In a normed linear space X the set of all vectors $\mathbf{x} \in X$ such that $\|\mathbf{x} - \overline{\mathbf{x}}\| \leq 1$ is called unit ball centered at $\overline{\mathbf{x}}$.
  (a) Sketch unit balls in R^2 when 1, 2 and ∞ norms are used.
  (b) Sketch unit ball in C[0,1] when maximum norm is used.
  (c) Can you draw picture of unit ball in $\mathrm{L}_2[0,1]$?

(11) Two norms $\|.\|_a$ and $\|.\|_b$ are said to be equivalent if there exists two positive constants c_1 and c_2,independent of $\mathbf{x}$, such that

$$c_1 \|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq c_2 \|\mathbf{x}\|_a$$

Show that in R^n the 2 norm (Euclidean norm) and $\infty-$norm (maximum norm) are equivalent.

(12) How that

$$|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\|$$

(13) A norm $\|.\|_a$is said to be stronger than another norm $\|.\|_b$if

$$\begin{array}{c}\lim\\ k\rightarrow\infty\end{array}\left\|\mathbf{x}^{(k)}\right\|_a=0\Rightarrow\begin{array}{c}\lim\\ k\rightarrow\infty\end{array}\left\|\mathbf{x}^{(k)}\right\|_b=0$$

but not vice versa. For C[0,1], show that the maximum norm is stronger than 2 norm.

(14) Show that function $\|\mathbf{x}\|_{2,W}:R^n\rightarrow R$ defined as

$$\|\mathbf{x}\|_{2,W}=\sqrt{\mathbf{x}^TW\mathbf{x}}$$

defines a norm on when W is a positive definite matrix.

(15) Show that function $\langle\mathbf{x},\mathbf{y}\rangle:R^n\times R^n\rightarrow R$ defined as

$$\langle\mathbf{x},\mathbf{y}\rangle=\mathbf{x}^TW\mathbf{y}$$

defines an inner product on when W is a positive definite matrix. The corresponding 2-norm is defined as $\|\mathbf{x}\|_{2,W}=\sqrt{\mathbf{x}^TW\mathbf{x}}$.

(16) Consider $X=R^3$ with $\langle\mathbf{x},\mathbf{y}\rangle=\mathbf{x}^TW\mathbf{y}$. Given a set of three linearly independent vectors in R^3

$$\mathbf{x}^{(1)}=\begin{bmatrix}1\\2\\1\end{bmatrix};\ \mathbf{x}^{(2)}=\begin{bmatrix}3\\2\\1\end{bmatrix};\ \mathbf{x}^{(3)}=\begin{bmatrix}1\\2\\3\end{bmatrix}$$

we want to construct and orthonormal set. Applying Gram Schmidt procedure,

$$\langle\mathbf{x},\mathbf{y}\rangle_W=\mathbf{x}^TW\mathbf{y}$$

$$W=\begin{bmatrix}2&-1&1\\-1&2&-1\\1&-1&2\end{bmatrix}$$

(17) **Gram-Schmidt Procedure in C[a,b]**: Let X represent set of continuous functions on interval $0\leq t\leq 1$ with inner product defined as

$$\langle\mathbf{x}(t),\mathbf{y}(t)\rangle=\int_0^1 w(t)\mathbf{x}(t)\mathbf{y}(t)dt$$

Given a set of four linearly independent vectors

$$\mathbf{x}^{(1)}(t)=1;\quad\mathbf{x}^{(2)}(t)=t;\quad\mathbf{x}^{(3)}(t)=t^2;$$

find orthonormal set of vectors if (a) $w(t)=1$ (Shifted Legandre Polynomials) (b) $w(t)=t(1-t)$ (Jacobi polynomials).

(18) Show that in C[a,b] with maximum norm, we cannot define an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ such that $\langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = \|\mathbf{x}\|_{\infty}$. In other words, show that in $C[a, b]$ the following function

$$\langle f(t),g(t) \rangle = \begin{array}{c} \max \\ t \end{array} |x(t)y(t)|$$

cannot define an inner product.

(19) In $C^{(1)}[a, b]$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_a^b \mathbf{x}'(t)\mathbf{y}'(t)dt + \mathbf{x}(a)\mathbf{y}(a)$$

an inner product?

(20) Show that in $C^{(1)}[a, b]$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_a^b w(t)\mathbf{x}(t)\mathbf{y}(t)dt$$

with $w(t) > 0$ defines an inner product.

(21) Show that parallelogram law holds in any inner product space.

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$$

Does it hold in C[a,b] with maximum norm?

(22) The triangle inequality asserts that, for any two vectors $\mathbf{x}$ and $\mathbf{y}$ belonging to an inner product space

$$\|\mathbf{x} + \mathbf{y}\|_2 \leq ||\mathbf{y}||_2 + ||\mathbf{x}||_2$$

After squaring both the sides and expanding,reduce this to Schwarz inequality. Under what condition Schwarz inequality becomes an equality?

(23) Show that operator $d/dx(.) : C^{(1)}[a, b] \rightarrow C[a, b]$ is onto $C[a, b]$ but not one to one.

(24) If L is a linear operator, show that $L(\overline{0}) = \overline{0}$.

(25) If L is a linear operator, show that range of L is a linear vector space. Show by example that this is not necessarily true for nonlinear operator.

CHAPTER 3

Linear Algebraic Equations and Related Numerical Schemes

1. Solution of Ax = b and Fundamental Spaces of A

The central problem of linear algebra is solution of linear equations of type

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \tag{1.1}$$

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \tag{1.2}$$

$$.. = ... \tag{1.3}$$

$$a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \tag{1.4}$$

which can be expressed in vector notation

$$A\mathbf{x} = \mathbf{b} \tag{1.5}$$

$$A = \begin{bmatrix} a_{11} & . & . & a_{1n} \\ . & . & . & . \\ . & . & . & . \\ a_{m1} & . & . & a_{mn} \end{bmatrix} \tag{1.6}$$

$\mathbf{x} \in R^n; \mathbf{b} \in R^m$ and $A \in R^m \times R^n$. Here m represents number of equations while n represents number of unknowns. Three possible situations arise while developing mathematical models

- **Case** $(m = n)$: system may have a unique solution / no solution / multiple solutions depending on rank of matrix A and vector $\mathbf{b}$
- **Case** $(m > n)$: system may have no solution or may have multiple solutions
- **Case** $(m < n)$: multiple solution

In these lecture notes, we are interested in the first case , i.e. $m = n$, particularly when the number of equations are large.

Before we discuss the numerical methods to solve equation (1.5), let us briefly review some geometric concepts associated with this equation. Consider

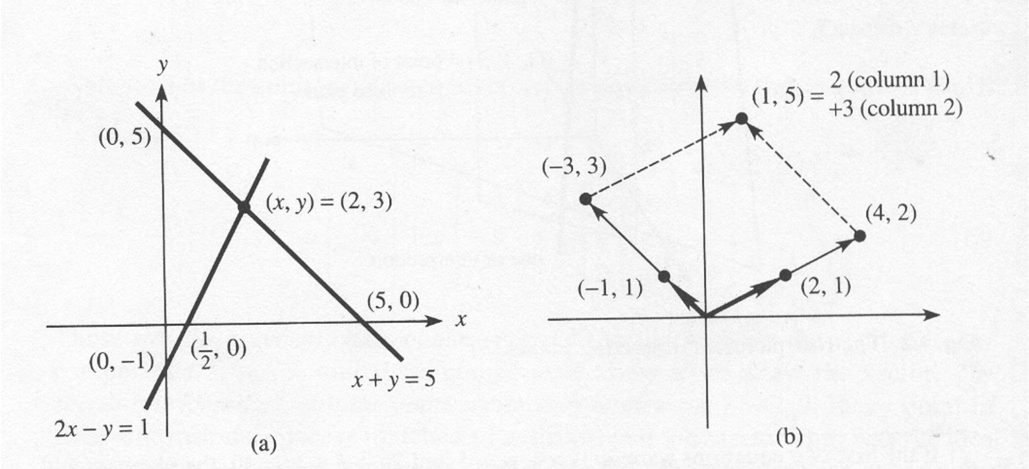


FIGURE 1

the following system of equations

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \tag{1.7}$$

There are two ways of interpreting the above matrix vector equation geometrically.

- **Row picture :** If we consider two equations separately as

$$2x - y = \begin{bmatrix} 2 \\ -1 \end{bmatrix}^T \begin{bmatrix} x \\ y \end{bmatrix} = 1 \tag{1.8}$$

$$x + y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} x \\ y \end{bmatrix} = 5 \tag{1.9}$$

then, each one is a line in x-y plane and solving this set of equations simultaneously can be interpreted as finding the point of their intersection (see Figure 1 (a)).

- **Column picture** : We can interpret the equation as linear combination of column vectors, i.e. as vector addition

$$x_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \tag{1.10}$$

Thus, the system of simultaneous equations can be looked upon as one vector equation i.e. addition or linear combination of two vectors (see Figure 1 (b)).

Now consider the following set of equations

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \tag{1.11}$$

In row picture, this is clearly an inconsistent case $(0 = 1)$ and has no solution as the row vectors are linearly dependent. In column picture, no scalar multiple of

$$\mathbf{v} = [1\ 2]^T$$

can found such that $\alpha\mathbf{v} = [2\ 5]^T$. Thus, in a singular case

$$\text{Row picture fails} \quad \Longleftrightarrow \quad \text{Column picture fails}$$

i.e. if two lines fail to meet in the row picture then vector $\mathbf{b}$ cannot be expressed as linear combination of column vectors in the column picture.

Now, consider a general system of linear equations $A\mathbf{x} = \mathbf{b}$ where A is an $n \times n$ matrix.

Row picture : Let A be represented as

$$A = \begin{bmatrix} \left(\mathbf{r}^{(1)}\right)^T \\ \left(\mathbf{r}^{(2)}\right)^T \\ .... \\ \left(\mathbf{r}^{(n)}\right)^T \end{bmatrix} \tag{1.12}$$

where $\left(\mathbf{r}^{(i)}\right)^T$ represents i'th row of matrix A. Then $A\mathbf{x} = \mathbf{b}$ can be written as n equations

$$\begin{bmatrix} \left(\mathbf{r}^{(1)}\right)^T \mathbf{x} \\ \left(\mathbf{r}^{(2)}\right)^T \mathbf{x} \\ .... \\ \left(\mathbf{r}^{(n)}\right)^T \mathbf{x} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ .... \\ b_n \end{bmatrix} \tag{1.13}$$

Each of these equations $\left(\mathbf{r}^{(i)}\right)^T \mathbf{x} = b_i$ represents a hyperplane in R^n (i.e. line in R^2 ,plane in R^3 and so on). Solution of $A\mathbf{x} = \mathbf{b}$ is the point $\mathbf{x}$ at which all these hyperplanes intersect (if at all they intersect in one point).

Column picture : Let A be represented as $A = [\ \mathbf{c}^{(1)}\ \mathbf{c}^{(2)}.........\mathbf{c}^{(n)}\ ]$ where $\mathbf{c}^{(i)}$ represents i^{th}column of A. Then we can look at $A\mathbf{x} = \mathbf{b}$ as one vector equation

$$x_1\mathbf{c}^{(1)} + x_2\mathbf{c}^{(2)} + ... + x_n\mathbf{c}^{(n)} = \mathbf{b} \tag{1.14}$$

Components of solution vector $\mathbf{x}$ tells us how to combine column vectors to obtain vector $\mathbf{b}$.In singular case, the n hyperplanes have no point in common or equivalently the n column vectors are not linearly independent. Thus, both these geometric interpretations are consistent with each other.

Now, to understand behavior of solutions of type (1.5), we can define four fundamental spaces associated with matrix A

Definition 22. *(**Column Space**): The space spanned by column vectors of matrix A is defined as column space of the matrix and denoted as $R(A)$.*

Definition 23. *(**Row Space**): The space spanned by row vectors of matrix A is called as row space of matrix$A$ and denoted as $R(A^T)$.*

Definition 24. *(**Null space**): The set of all vectors $\mathbf{x}$ such that $A\mathbf{x} = \bar{0}$ is called as null space of matrix A and denoted as $N(A)$.*

A non-zero null space is obtained only when columns of A are linearly dependent. If columns of A are linearly independent, then $N(A) \equiv \{\bar{0}\}$.

Definition 25. *(**Left Null Space**) :The set of all vectors $\mathbf{y}$ such that $A^T\mathbf{y} = \bar{0}$ is called as null space of matrix A and denoted as $N(A^T)$.*

A non-zero left null space is obtained only when rows of A are linearly dependent. If rows of A are linearly independent, then $N(A^T) \equiv \{\bar{0}\}$.

The following fundamental result, which relates dimensions of row and column spaces with the rank of a matrix, holds true for any $m \times n$ matrix A.

Theorem 2. *(**Fundamental Theorem of Linear Algebra**): Given a $m \times n$ matrix A*

$$\begin{aligned} \dim[R(A)\,] &= \textit{Number of linearly independent columns of } A = rank(A) \\ \dim[N(A)] &= n - rank(A) \end{aligned}$$

$$\begin{aligned} \dim[R(A^T)\,] &= \textit{Number of linearly independent rows of } A = rank(A) \\ \dim[N(A^T)\,] &= m - rank(A) \end{aligned}$$

In other words, number of linearly independent columns of A equals number of linearly independent rows of A.

Note that

- When matrix A operates on vector $\mathbf{x} \in R(A^T)$ (i.e. a vector belonging to row space of A) it produces a vector $A\mathbf{x} \in R(A)$ (i.e. a vector in column space of A)

- The system of equations $A\mathbf{x} = \mathbf{b}$ can be solved if and only if $\mathbf{b}$ belongs to the column space of A. i.e., $\mathbf{b} \in R(A)$. the solution is unique only if $N(A) \equiv \{\bar{0}\}$. If $N(A) \neq \{\bar{0}\}$ i.e. if columns of A are linearly dependent and $\mathbf{b} \in R(A)$, then we can find infinite solutions to $A\mathbf{x} = \mathbf{b}$.

Example 38. *Consider the following set of equations*

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \tag{1.15}$$

It is easy to see that $\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ *is a solution*

$$(1) \begin{bmatrix} 1 \\ 2 \end{bmatrix} + (1) \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \tag{1.16}$$

as the vector on R.H.S. belongs to $R(A)$*. But, this is not the only solution. We can write*

$$3 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \tag{1.17}$$

This implies that $\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 3 & -1 \end{bmatrix}^T$ *is also a solution to the above problem. Why does this happen and how can we characterize all possible solutions to this problem? To answer this question, let us find null space of matrix A. In this particular case, by simple visual inspection, we can find that* $\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$ *is a vector belonging to the null space of A.*

$$(1) \begin{bmatrix} 1 \\ 2 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.18}$$

In fact, null space of A can be written as $N(A) = \alpha \begin{bmatrix} 1 & -1 \end{bmatrix}^T$ *for any real scalar* α*. Thus,*

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.19}$$

This implies that, if $\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ *is a solution to (1.15), then any vector*

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{1.20}$$

is also a solution to (1.15).

Thus, if we add any vector from the null space of **A** to the solution of (1.5), then we get another solution to equation (1.5). If $N(A) \equiv \{\bar{0}\}$ and a solution exists, i.e. $\mathbf{b} \in R(A)$, then the solution is unique. If $N(A) \neq \{\bar{0}\}$ and $\mathbf{b} \in R(A)$, then there are infinite solutions to equation (1.5).

Methods for solving linear algebraic equations can be categorized as (a) direct or Gaussian elimination based schemes and (b) iterative schemes. In the sections that follow, we discuss these techniques in detail.

2. Direct Solution Techniques

There are several methods which directly solve equation (1.5). Prominent among these are such as Cramer's rule, Gaussian elimination, Gauss-Jordan method and LU decomposition. We assume that you have some exposure to these method in earlier courses on engineering mathematics. Let φ denote the number of divisions and multiplications required for generating solution by a particular method. We first compare various methods on the basic of φ.[**6**]

- **Cramers Rule:**

$$\varphi(estimated) = (n-1)(n+1)(n!) + n \cong n^2 * n! \tag{2.1}$$

For a problem of size $n = 100$ we have $\varphi \cong 10^{162}$ and the time estimate for solving this problem on DEC1090 is approximately 10^{149}years.

- **Gaussian Elimination and Backward Sweep:** By maximal pivoting and row operations, we first reduce the system (1.5) to

$$U\mathbf{x} = \widehat{\mathbf{b}} \tag{2.2}$$

where U is a upper triangular matrix and then use backward sweep to solve (2.2) for **x**.For this scheme, we have

$$\varphi = \frac{n^3 + 3n^2 - n}{3} \cong \frac{n^3}{3} \tag{2.3}$$

For $n = 100$ we have $\varphi \cong 3.3 * 10^5$.

- **LU-Decomposition:** LU decomposition is used when equation (1.5) is to be solved for several different values of vector **b**, i.e.,

$$A\mathbf{x}^{(k)} = \mathbf{b}^{(k)} \quad ; \quad k = 1, 2, ...N \tag{2.4}$$

The sequence of computations is as follows

$$A = LU \quad \text{(Solved only once)} \tag{2.5}$$

$$L\mathbf{y}^{(k)} = \mathbf{b}^{(k)} \; ; \; k = 1, 2, ...N \tag{2.6}$$

$$U\mathbf{x}^{(k)} = \mathbf{y}^{(k)} \; ; \; k = 1, 2, ...N \tag{2.7}$$

For N different $\mathbf{b}$ vectors,

$$\boldsymbol{\varphi} = \frac{n^3 - n}{3} + Nn^2 \tag{2.8}$$

- **Gauss_Jordon Elimination:** In this case, we start with $[A : I : b]$ and by sequence row operations we reduce it to $[I : A^{-1} : x]$ i.e.

$$\left[A : I : \mathbf{b}^{(k)}\right] \quad \underset{\text{Row Operations}}{\overset{\text{Sequence of}}{\rightarrow}} \quad \left[I : A^{-1} : \mathbf{x}^{(k)}\right] \tag{2.9}$$

For this scheme, we have

$$\boldsymbol{\varphi} = \frac{n^3 + (N-1)n^2}{2} \tag{2.10}$$

Thus, Cramer's rule is certainly not suitable for numerical computations. The later three methods require significantly smaller number of multiplication and division operations when compared to Cramer's rule and are best suited for computing numerical solutions moderately large ($n \approx 1000$) systems. When number of equations is significantly large ($n \approx 10000$), even the Gaussian elimination and related methods can turn out to be computationally expensive and we have to look for alternative schemes that can solve (1.5) in smaller number of steps. When matrices have some **simple structure** (few non-zero elements and large number of zero elements), direct methods tailored to exploit sparsity and perform efficient numerical computations. Also, iterative methods give some hope as an approximate solution $\mathbf{x}$ can be calculated quickly using these techniques. In these lecture notes we describe direct methods for sparse linear systems and iterative computing schemes, which are useful when system dimension is large.

3. Solutions of Sparse Linear Systems

A system of Linear equations given by (1.5) is called Sparse if only a relatively small number of its matrix elements (a_{ij}) are nonzero. The sparse patterns that frequently occur are

- Tridiagonal
- Bond Diagonal with bond width M
- Block Diagonal

It is wasteful to apply general linear algebra methods on these problems. Special methods are evolved for solving such sparse systems that achieve considerable reduction in computation time and memory space requirements. In this section, we first provide motivation for looking at sparse matrix forms. Later, some of the sparse matrix algorithms are discussed in detail. This is meant to be a brief introduction to sparse matrix computations and the treatment of the topic is, by no means, exhaustive.

3.1. Origin of Sparse Linear Systems.

3.1.1. *Solutions ODE-BVP using finite difference method.* Consider the following general form of 2^{nd}order ODE-BVP problem frequently encountered in engineering problems

ODE:

$$\Psi[d^2y/dz^2, dy/dz, y, z] = 0 \;\; ; \qquad (0 < z < 1) \tag{3.1}$$

Boundary Conditions

$$f_1[dy/dz, y, z] = 0 \; at \; z = 0 \tag{3.2}$$

$$f_2[dy/dz, y, z] = 0 \;\; at \; z = 1 \tag{3.3}$$

Let $y^*(z) \in C^{(2)}[0,1]$ denote true solution to the above ODE-BVP. Depending on the nature of operator Ψ,it may or may not be possible to find the true solution to the problem. In the present case, however, we are interested in finding an approximate numerical solution, say $y(z)$, to the above ODE-BVP.

The basic idea in finite difference approach is to convert the ODE-BVP to a set of linear or nonlinear algebraic equations using Taylor series expansion as basis. In order to achieve this, the domain $0 \leq z \leq 1$ is divided into $(n+1)$ equidistant grid points $z_0, z_1......, z_n$ located such that

$$z_i = i(\Delta z) = i/(n) \text{ for } \; i = 0, 1, ...n$$

Let the value of y at location z_i be defined as $y_i = y(z_i)$. Using the Taylor Series expansion $y_{i+1} = y(z_{i+1}) = y(z_i + \Delta z)$ can be written as

(3.4)

$$y_{i+1} = y_i + (dy/dz)_i \;\; (\Delta z) \; + (1/2!) y_i^{(2)} (\Delta z)^2 \;\; + (1/3!) \; y_i^{(3)} (\Delta z)^3 + ...$$

Where

$$y_i^{(k)} = (d^k y/dz^k)_{z=zi} \tag{3.5}$$

So similarly we can write $y_{i-1} = y(z_{i-1}) = y(z_i - \Delta z)$ as

(3.6)

$$y_{i-1} = y_i - (dy/dz)_i \, (\Delta z_i) + (1/2!) \; y_i^{(2)} (\Delta z)^2 - (1/3!) \; y_i^{(3)} (\Delta z)^3 + ...$$

From equations (3.4) and (3.6) we can arrive at several expressions for y'_{i-1}. From equation (3.4) we get [**6**]

$$(dy/dz)_i = \frac{(y_{i+1} - y_i)}{\Delta z} - \left[y_i^{(2)} (\Delta\ z/2) + ... \right] \tag{3.7}$$

From equation (3.6) we get

$$(dy/dz)_i = \frac{(y_i - y_{i-1})}{\Delta z} + \left[y_i^{(2)} (\Delta\ z/2) - .... \right] \tag{3.8}$$

Combining equations (3.4) and (3.6) we get

$$(dy/dz)_i = \frac{(y_{i+1} - y_{i-1})}{2(\Delta z)} - \left[y_i^{(3)} (\Delta\ z^2/3!) + ... \right] \tag{3.9}$$

The first two formulae are accurate to $O(\Delta z)$ while the last one is accurate to $O[(\Delta z)^2]$ and so is more commonly used. Equation (3.7) and (3.8) can be combined to give an equation for second derivatives $y_i^{(2)}$ at location i:

$$y_i^{(2)} = \frac{(y_{i+1} - 2y_i + y_{i-1})}{(\Delta z)^2} - \left[2y_i^{(4)} (\Delta\ z^2/4!) + ... \right] \tag{3.10}$$

Note that approximations (3.9) and (3.10) both are of order $O[\Delta z)^2]$. One usually avoids using formulae having different accuracies for the same independent variable [**6**]. These equations can be used to reduce the ODE-BVP to set of algebraic equations as follows:

- **Step 1 :** Force residual R_i at each internal grid point to zero,i.e.,

$$R_i = \Psi \left[\frac{(y_{i+1} - 2y_i + y_{i-1})}{(\Delta z)^2}, \frac{(y_{i+1} - y_{i-1})}{2(\Delta z)},\ y_i,\ z_i \right] = 0 \tag{3.11}$$

$$i = 1, 2, 3 ....n - 1. \tag{3.12}$$

  This gives $(n-1)$ equations in $(n+1)$ unknowns. ODE-BVP is satisfied exactly at internal grid points.

- **Step 2:** Use B.C. to generate remaining equations
  - **Approach 1:** Use one-sided derivatives only at the boundary points, i.e.,

$$f_1[\frac{(y_1 - y_0)}{(\Delta z)}, y_0, 0] = 0 \tag{3.13}$$

$$f_2[\frac{(y_n - y_{n-1})}{(\Delta z)}, y_0, 0] = 0 \tag{3.14}$$

    This gives remaining two equations.

– **Approach 2:**

$$f_1[\frac{(y_1 - y_{-1})}{(2\Delta z)}, y_0, 0] = 0 \tag{3.15}$$

$$f_2[\frac{y_{n+1} - y_{n-1}}{(\Delta z)}, y_0, 0] = 0 \tag{3.16}$$

This approach introduces two more variables y_{-1} and y_{n+1} at hypothetical grid points. Thus we have $n+3$ variables and $n+1$ equations, two more algebraic equations can be generated by setting residual at zero at boundary points,i.e., at z_0 and z_n,i.e.,

$$R_0 = 0 \; and \; R_n = 0$$

This results in $(n+3)$ equations in $(n+3)$ unknowns.

Remark 1. *Forcing residual $R_i = 0$ at internal grid points implies choosing the values of y_i such that the ODE-BVP is satisfied exactly at the internal grid points. Obviously larger the value of N, the closer is the numerical solution expected to be near exact solution. Note that $R_i = 0$ at internal grid points does not imply y_i equals y_i^* (exact solution) at that point, i.e. $y(z_i) \neq y^*(z_i)$.*

Example 39. *Consider steady state heat transfer/conduction in a slab of thickness L, in which energy is generated at a constant rate of q W/m^3. The boundary at $z = 0$ is maintained at a constant temperature T_0,while the boundary at $z = L$ dissipates heat by convection with a heat transfer coefficient h into the ambient temperature at T_∞. The mathematical formulation of the conduction problem is given as*

$$kd^2T/dz^2 + q = 0 \tag{3.17}$$

$$0 < z < L$$

$$B.C. \quad : \quad T(0) = T_0 \; at \; z = 0 \tag{3.18}$$

$$k\left[\frac{dT}{dz}\right]_{z=L} = h\left[T_\infty - T(L)\right] \tag{3.19}$$

(Note that this problem can be solved analytically.) Dividing the region $0 \leq z \leq L$ into n equal subregions and setting residuals zero at the internal grid points, we have

$$(T_{i-1} - 2T_i + 2T_{i+1})/(\Delta z)^2 + q/k = 0 \tag{3.20}$$

$$i = 1, 2, ...(.n-1).$$

Or

$$(T_{i-1} - 2T_i + T_{i+1}) = -(\Delta z)^2 q/k \tag{3.21}$$

Note that at $i = 1$ *we have*

$$(T_0 - 2T_1 + T_2) = -(\Delta z)^2 q/k \tag{3.22}$$

Using B.C. (3.18) we have

$$-2T_1 + T_2 = -(\Delta z)^2 q/k - T_0 \tag{3.23}$$

Using one sided derivative at $z = 1$*and using B.C. (3.19)*

$$k(T_n - T_{n-1})/(\Delta z) = h((T_\infty - T_n) \tag{3.24}$$

or

$$T_n - T_{n-1} = h\Delta z(T_\infty - T_n)/k \tag{3.25}$$

$$T_n(1 + h\Delta z/k) - T_{n-1} = h\Delta z T_\infty/k \tag{3.26}$$

Rearranging the equations in matrix form

$$\begin{bmatrix} -2 & 1 & 0 & 0....................... \\ 1 & -2 & 1 & 0....................... \\ 0 & 1 & -2 & 1...................... \\ . & . & . & .... \\ . & . & . & .... -2.......1..... \\ 0 & 0 & 0 & ....-1.. \quad (1 + h\Delta z/k) \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ . \\ . \\ T_n \end{bmatrix} = \begin{bmatrix} -(\Delta z)^2 q/k - T_0 \\ -(\Delta z)^2 q/k \\ . \\ . \\ . \\ -h(\Delta z)T_\infty/k \end{bmatrix} \tag{3.27}$$

we get a sparse tridiagonal matrix.

3.1.2. *Solution of PDE using Finite Difference Method* [**6**]. Consider elliptic PDEs described by

$$\begin{aligned} \nabla^2 u &= cu + f(x, y, z) \\ \text{or } \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} &= cu + f(x, y, z) \end{aligned}$$

which is solved on a 3 dimensional bounded region V with boundary S. The boundary conditions on spatial surface S are given as

$$\langle \alpha(s)\mathbf{n}, \nabla u \rangle + \beta(s)u = h(s)$$

These equations can be transformed into a set of linear (or nonlinear) algebraic equations by using Taylor's theorem and approximating

$$\begin{aligned}
\left(\frac{\partial^2 u}{\partial x^2}\right)_{ijk} &= \frac{(u_{i+1,j,k} - 2u_{i,j,k} + u_{i-1,j,k})}{(\Delta x)^2} \\
\left(\frac{\partial^2 u}{\partial y^2}\right)_{ijk} &= \frac{(u_{i,j+1,k} - 2u_{i,j,k} + u_{i,j-1,k})}{(\Delta y)^2} \\
\left(\frac{\partial^2 u}{\partial z^2}\right)_{ijk} &= \frac{(u_{i,j,k+1} - 2u_{i,j,k} + u_{i,j,k-1})}{(\Delta z)^2}
\end{aligned}$$

and so on.

Example 40. *Laplace equation represents a prototype for steady state diffusion processes. For example 2-dimensional Laplace equation*

$$\partial^2 T/\partial x^2 + \partial^2 T/\partial y^2 = 0 \tag{3.28}$$

represents a description of 2-dimensional steady state heat conduction in a solid , where T is temperature and x, y *are space coordinates. Equations similar to this arise in many problems of fluid mechanics, heat transfer and mass transfer. In the present case , we consider conduction in a rectangular plate of dimension* $L_x \times L_{y.}$*The boundary conditions are as follows:*

$$x = 0 : T = T_1; \;\; x = L_x : T = T_3 \tag{3.29}$$

$$y = 0 : T = T_2; \;\; y = L_y : T = T_4 \tag{3.30}$$

Construct the 2 -dimensional grid with $(n_x + 1)$ *equispaced grid lines parallel to y axis and* $(n_y + 1)$ *equispaced grid lines parallel to x axis. The temperature T at* (i, j) th *grid point is denoted as* $T_{ij} = T(x_{i,}y$*We force the residual to be zero at each internal grid point to obtain the following set of equations:*

$$(T_{i+1,j} - 2T_{i,j} + T_{i-1,j})/(\Delta x)^2 + (T_{i,j+1} - 2T_{i,j} + T_{i,j-1})/(\Delta y)^2 = 0 \tag{3.31}$$

for $(i = 1, 2, ...n_x - 1)$ *and* $(\ j = 1, 2, ...n_y - 1)$. *As the boundary temperatures are known, number of variables exactly equals no of interior nodes. Note that regardless of the size of the system, each equation contains not more than 5 unknowns, resulting in a sparse linear algebraic system. Consider the special case when*

$$\Delta x = \Delta y$$

For this case the above equations can be written as

$$T_{i-1,j} + T_{i,j-1} - 4T_{i,j} + T_{i,j+1} + T_{i+1,j} = 0 \tag{3.32}$$

$$for\ (i = 1, 2, ...n_x - 1)\ and\ \ (j = 1, 2, ...n_y - 1)$$

Using boundary conditions we have

$$\begin{aligned} T_{i,0} &= T_2\ \ ;\ \ T_{i,n_y} = T_4 \qquad (3.33)\\ i &= 0, 1, ...n_x \\ T_{0,j} &= T_1\ \ ;\ \ T_{n_x,j} = T_3 \qquad (3.34)\\ j &= 0, 1, ...n_y \end{aligned}$$

Now we define

$$\mathbf{x} = [T_{11}\ T_{12}............T_{1,n_y-1},.........., T_{n_x-1,1}.............T_{n_x-1,n_y-1}]^T \qquad (3.35)$$

And rearrange the above set of equations in form of $A\mathbf{x} = \mathbf{b}$*, then* A *turns out to be a large sparse matrix. Even for 10 internal grid lines in each direction we would get a* 100×100 *sparse matrix associated with* 100 *variables.*

3.1.3. *Cubic Spline Interpolation* [**6**]. Suppose we are given $n+1$ values $\{y_0, y_1, ...y_n\}$ some dependent variable $y(z)$ corresponding to values $\{z_0, z_1, ...z_n\}$ of independent variable z over some finite interval. We have to find a continuous function $f(z)$ that passes through each of these points. Invoking Weierstarss theorem, we propose a $n'th$ degree polynomial function

$$y = f(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + ... + \alpha_n z^n$$

Note that values of y and x are exactly known and this is a problem of finding an n'th degree interpolation polynomial that passes through all the points. The coefficients of this polynomial can be easily found by solving equation

$$\begin{bmatrix} 1 & z_0 & ... & (z_0)^n \\ 1 & z_1 & ... & (z_1)^2 \\ ... & ... & ... & ..... \\ 1 & z_n & ... & (z_n)^2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ ..... \\ \alpha_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ .... \\ y_n \end{bmatrix} \qquad (3.36)$$

$$\text{or}\quad A\boldsymbol{\alpha} = \mathbf{y} \qquad (3.37)$$

If matrix A is nonsingular, coefficient vector can be simply computed as $\boldsymbol{\alpha} = A^{-1}\mathbf{y}$.

The matrix A appearing in the above equation is known as *Vandermonde matrix*. Larger matrices of this type tend to become numerically ill-conditioned. As a consequence, if the number of data points is large, fitting a large order polynomial can result in a polynomial which exhibits unexpected oscillatory behavior. In order to avoid such oscillations and difficulties due to matrix ill conditioning, the data is divided into sub-intervals and a lower order (say cubic) spline approximation is developed on each interval. For example, n cubic splines fitting $n+1$ data points can be expressed as

$$\mathbf{p}_0(z) = \alpha_{0,0} + \alpha_{1,0}(z - z_0) + \alpha_{2,0}(z - z_0)^2 + \alpha_{3,0}(z - z_0)^3 \tag{3.38}$$

$$(z_0 \leq z \leq z_1) \tag{3.39}$$

$$\mathbf{p}_1(z) = \alpha_{0,1} + \alpha_{1,1}(z - z_1) + \alpha_{2,1}(z - z_1)^2 + \alpha_{3,1}(z - z_1)^3 \tag{3.40}$$

$$(z_1 \leq z \leq z_2) \tag{3.41}$$

$$....... = ...$$

$$\mathbf{p}_{n-1}(z) = \alpha_{0,n-1} + \alpha_{1,n-1}(z - z_{n-1})$$

$$+\alpha_{2,n-1}(z - z_{n-1})^2 + \alpha_{3,n-1}(z - z_{n-1})^3 \tag{3.42}$$

$$(z_{n-1} \leq z \leq z_n) \tag{3.43}$$

There are total $4n$ unknown coefficients $\{\alpha_{0,0}, \alpha_{1,0}.......\alpha_{3,n-1}\}$ to be determined. In order to ensure continuity and smoothness of the approximation, the following conditions are imposed

$$\mathbf{p}_i(z_i) = y_i \quad ; \quad i = 0, 1, 2, ..., n-1 \tag{3.44}$$

$$\mathbf{p}_{n-1}(z_n) = y_n \tag{3.45}$$

$$\mathbf{p}_i(z_{i+1}) = \mathbf{p}_{i+1}(z_{i+1}) \quad ; \quad i = 0, 1, ...n-2 \tag{3.46}$$

$$\frac{d\mathbf{p}_i(z_{i+1})}{dz} = \frac{d\mathbf{p}_{i+1}(z_{i+1})}{dz} \quad ; \quad i = 0, 1, ...n-2 \tag{3.47}$$

$$\frac{d^2\mathbf{p}_i(z_{i+1})}{dz^2} = \frac{d^2\mathbf{p}_{i+1}(z_{i+1})}{dz^2} \quad ; \quad i = 0, 1, ...n-2 \tag{3.48}$$

which result in $4n-2$ conditions. Two additional conditions are imposed at the boundary points

$$\frac{d^2\mathbf{p}_0(z_0)}{dz^2} = \frac{d^2\mathbf{p}_{n-1}(z_n)}{dz^2} = 0 \tag{3.49}$$

which are referred to as free boundary conditions. If the first derivatives at the boundary points are known,

$$\frac{d\mathbf{p}_0(z_0)}{dz} = d_0 \ ; \ \frac{d^2\mathbf{p}_{n-1}(z_n)}{dz^2} = d_n \tag{3.50}$$

then we get the clamped boundary conditions.

Using constraints (3.44-3.48), together with free boundary conditions, we get

$$\alpha_{0,i} = y_i \quad ; \quad (i = 0, 1, 2, ..., n-1) \tag{3.51}$$

$$\alpha_{0,n-1} + \alpha_{1,n-1}\left(\Delta z_{n-1}\right) + \alpha_{2,n-1}\left(\Delta z_{n-1}\right)^2 + \alpha_{3,n-1}\left(\Delta z_{n-1}\right)^3 = y_n \tag{3.52}$$

$$\alpha_{0,i}+\alpha_{1,i}\left(\Delta z_i\right)+\alpha_{2,i}\left(\Delta z_i\right)^2+\alpha_{3,i}\left(\Delta z_i\right)^3 = \alpha_{0,i+1} \tag{3.53}$$

$$\alpha_{1,i}+2\alpha_{2,i}\left(\Delta z_i\right)+3\alpha_{3,i}\left(\Delta z_i\right)^2 = \alpha_{1,i+1} \tag{3.54}$$

$$\alpha_{2,i}+3\alpha_{3,i}\left(\Delta z_i\right) = \alpha_{2,i+1}$$

$$for(\quad i=0,1,2,...,n-2\ )$$

$$\alpha_{2,0} = 0 \tag{3.55}$$

$$\alpha_{2,n-1}+3\alpha_{3,n-1}\left(\Delta z_{n-1}\right) = 0 \tag{3.56}$$

Eliminating $\alpha_{3,i}$ using

$$\alpha_{3,i} = \frac{\alpha_{2,i+1}-\alpha_{2,i}}{3\left(\Delta z_i\right)} \quad \text{for} \ (\quad i=0,1,2,...,n-2\ ) \tag{3.57}$$

$$\alpha_{3,n-1} = \frac{-\alpha_{2,n-1}}{3\left(\Delta z_i\right)} \tag{3.58}$$

and eliminating $\alpha_{1,n-1}$using equation 3.52

$$\alpha_{1,i} = \frac{1}{\Delta z_i}(\alpha_{0,i+1}-\alpha_{0,i})-\frac{\Delta z_i}{3}(2\alpha_{2,i}+\alpha_{2,i+1}) \tag{3.59}$$

$$\text{for} \ (\quad i = 0,1,2,...,n-2\ )$$

$$\alpha_{1,n-1} = \frac{y_n-\alpha_{0,n-1}}{\Delta z_{n-1}}-\left(\Delta z_{n-1}\right)\alpha_{2,n-1}-\alpha_{3,n-1}\left(\Delta z_{n-1}\right)^2 \tag{3.60}$$

we get only $\{\alpha_{1,i}:i=0,1,...n-1\}$ as unknowns

$$\alpha_{2,0} = 0 \tag{3.61}$$

$$\left(\Delta z_{i-1}\right)\alpha_{2,i-1}+2(\Delta z_i+\Delta z_{i-1})\alpha_{2,i}+\left(\Delta z_{i-1}\right)\alpha_{2,i+1} = b_i \tag{3.62}$$

$$\text{for} \ (\quad i=1,2,...,n-2\ )$$

where

$$b_i=\frac{3(\alpha_{0,i+1}-a_{0,1})}{\Delta z_i}-\frac{3(\alpha_{0,i}-a_{0,i-1})}{\Delta z_{i-1}}=\frac{3(y_{i+1}-y_1)}{\Delta z_i}-\frac{3(y_i-y_{i-1})}{\Delta z_{i-1}}$$

$$\text{for} \ (\quad i=1,2,...,n-2\ )$$

$$\frac{1}{3}\left(\Delta z_{n-2}\right)\alpha_{2,n-2}+\frac{2}{3}(\Delta z_{n-2}+\Delta z_{n-1})\alpha_{2,n-1}=b_n \tag{3.63}$$

$$b_n=\frac{y_n}{\Delta z_{n-1}}-\left(\frac{1}{\Delta z_{n-1}}+\frac{1}{\Delta z_{n-2}}\right)y_{n-1}+\frac{y_{n-2}}{\Delta z_{n-2}} \tag{3.64}$$

Defining vector $\boldsymbol{\alpha}_2$ as

$$\boldsymbol{\alpha}_2=\left[\begin{array}{cccc}\alpha_{2,0} & \alpha_{2,1} & .... & \alpha_{2,n}\end{array}\right]^T$$

the above set of n equations can be rearranged as

$$A\boldsymbol{\alpha}_2 = \mathbf{b} \tag{3.65}$$

where A is a $(n+1)\times(n+1)$ and Y is $(n+1)$ vector. Elements of A and $\mathbf{b}$ can be obtained from equations (3.61-3.63). Note that matrix A will be a near tridiagonal matrix.

3.2. Algorithms for Solving Sparse Linear Systems [5].

3.2.1. *Thomas Algorithm for Tridiagonal and Block Tridiagonal Matrices.* Consider system of equation given by following equation

$$\begin{bmatrix} b_1 & c_1 & 0 & ... & ... & ... & ... & 0 \\ a_2 & b_2 & c_2 & 0 & ... & ... & ... & 0 \\ 0 & a_3 & b_3 & c_3 & ... & ... & ... & 0 \\ 0 & 0 & a_4 & b_4 & c_4. & ... & ... & ... \\ ... & ... & .. & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & c_{n-2} & 0 \\ ... & ... & ... & ... & ... & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & ... & .0 & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ .... \\ .... \\ .... \\ .... \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ .... \\ .... \\ .... \\ .... \\ .... \\ d_n \end{bmatrix} \tag{3.66}$$

where matrix A is a tridiagonal matrix.

Step 1:Triangularization: Forward sweep with normalization

$$\gamma_1 = c_1/b_1 \tag{3.67}$$

$$\gamma_k = \frac{c_k}{b_k - a_k\gamma_{k-1}} \quad ; \quad k = 2,3,....(n-1) \tag{3.68}$$

$$\beta_1 = d_1/b_1 \tag{3.69}$$

$$\beta_k = \frac{(d_k - a_k\beta_{k-1})}{(b_k - a_k\gamma_{k-1})} \quad ; \quad k = 2,3,....n \tag{3.70}$$

This sequence of operations finally results in the following system of equations

$$\begin{bmatrix} 1 & \gamma_1 & 0 & .... & 0 \\ 0 & 1 & \gamma_2 & .... & 0 \\ ... & 0 & 1 & .... & . \\ .... & .... & .... & .... & \gamma_{n-1} \\ 0 & 0 & ... & .... & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ .... \\ .... \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_n \end{bmatrix}$$

Step 2: Backward sweep leads to solution vector

$$x_n = \beta_n$$

$$\begin{aligned} x_k &= \beta_k - \gamma_k x_{k+1} && (3.71) \\ k &= (n-1), .(n-2), ..., 1 && (3.72) \end{aligned}$$

Total no of multiplications and divisions

$$\boldsymbol{\varphi} = 5n - 8$$

Which is far smaller than the $n^3/3$ operations(approximately) for Gaussian elimination and backward sweep required for dense matrices.

Block Thomas Algorithm: Consider block triangular system of the form

$$\begin{bmatrix} B_1 & C_1 & [0]. & ..... & .... & [0]. \\ A_2 & B_2 & C_2 & .... & ..... & ..... \\ .... & ..... & ..... & ..... & ..... & [0]. \\ ..... & ..... & ..... & ..... & .B_{n-1} & C_{n-1} \\ [0] & ..... & ..... & [0] & A_n & B_n \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ . \\ . \\ \mathbf{x}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{d}^{(1)} \\ \mathbf{d}^{(2)} \\ . \\ . \\ \mathbf{d}^{(n)} \end{bmatrix} \qquad (3.73)$$

where A_i, B_i and C_i are matrices and $\mathbf{x}_i$ and $\mathbf{d}_i$ represent vectors of appropriate dimensions. Thomas algorithm can be developed for such systems in a analogous manner.

- **Step 1:Block Triangularization**

$$\Gamma_1 = [B_1]^{-1} C_1$$

$$\Gamma_k = [B_k - A_k\Gamma_{k-1}]^{-1} C_k \quad ; \quad k = 2, 3, ...(n-1) \qquad (3.74)$$

$$\boldsymbol{\beta}^{(1)} = [B_1]^{-1} \mathbf{d}^{(1)} \qquad (3.75)$$

$$\boldsymbol{\beta}^{(k)} = [B_k - A_k\Gamma_{k-1}]^{-1} (\mathbf{d}^{(k)} - A_k\boldsymbol{\beta}^{(k-1)}) \quad ; \quad k = 2, 3, ...n$$

- **Step 2: Backward sweep**

$$\mathbf{x}^{(n)} = \boldsymbol{\beta}^{(n)} \qquad (3.76)$$

$$\begin{aligned} \mathbf{x}^{(k)} &= \boldsymbol{\beta}^{(k)} - \Gamma_k \mathbf{x}^{(k+1)} && (3.77) \\ k &= (n-1), .(n-2), ..., 1 && (3.78) \end{aligned}$$

3.2.2. *Triangular and Block Triangular Matrices.* A triangular matrix is a sparse matrix with zero-valued elements above the diagonal,i.e.,

$$L = \begin{bmatrix} l_{11} & 0 & . & . & 0 \\ l_{12} & l_{22} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ l_{\ n1} & . & . & . & l_{\ nn} \end{bmatrix}$$

To solve a system $Lx = b$, the following algorithm is used

$$x_1 = b_1/l_{11} \tag{3.79}$$

$$x_i = \frac{[b_i - \sum\limits_{j=1}^{i-1} l_{ij} x_j]}{l_{ii}} \quad ; \quad i = 2, 3, ...n \tag{3.80}$$

The operational count $\boldsymbol{\varphi}$ i.e., the number of multiplications and divisions, for this elimination process is

$$\boldsymbol{\varphi} = n(n+1)/2 \tag{3.81}$$

which is considerably smaller than the Gaussian elimination..

In some applications we encounter equations with a block triangular matrices. For example,

$$\begin{bmatrix} A_{11} & [0] & .... & [0] \\ A_{12} & A_{22} & .... & [0] \\ ..... & . & .... & ..... \\ A_{n1} & A_{n2} & .... & A_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}^{(1)} \\ \boldsymbol{\eta}^{(2)} \\ .... \\ \boldsymbol{\eta}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ .... \\ \mathbf{b}^{(n)} \end{bmatrix}$$

Where A_{ij} are $m \times m$ sub-matrices while $\boldsymbol{\eta}^{(i)} \in R^m$ and $\mathbf{b}^{(i)} \in R^m$ are subvectors for $i = 1, 2, ..n$. The solution of this type of systems is completely analogous to that of lower triangular systems,except that sub-matrices and subvectors are used in place of scalers. The block counterpart for lower triangular system is

$$\boldsymbol{\eta}^{(i)} = (A_{11})^{-1}\mathbf{b}^{(1)} \tag{3.82}$$

$$\boldsymbol{\eta}^{(i)} = (A_{ii})^{-1}[\mathbf{b}^{(i)} - \sum_{j=1}^{i-1}(A_{ij}\boldsymbol{\eta}^{(i)})] \quad ; \quad i = 2, 3, ...n \tag{3.83}$$

The above form does not imply that the inverse $(A_{ii})^{-1}$should be compared explicitly. For example we can find $\boldsymbol{\eta}^{(1)}$ by Gaussian elimination to solve the system $A_{11}\boldsymbol{\eta}^{(1)} = \mathbf{b}^{(1)}$

3.2.3. *Solution of a Large System By Partitioning.* If matrix A is equation (1.5) is very large, then we can partition matrix A and vector $\mathbf{b}$ as

$$A\mathbf{x} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \end{bmatrix}$$

where A_{11} is a $(m \times m)$ square matrix. this results in two equations

$$A_{11}\mathbf{x}^{(1)} + A_{12}\mathbf{x}^{(2)} = \mathbf{b}^{(1)} \tag{3.84}$$

$$A_{21}\mathbf{x}^{(1)} + A_{22}\,\mathbf{x}^{(2)} = \mathbf{b}^{(2)} \tag{3.85}$$

which can be solved sequentially as follows

$$\mathbf{x}^{(1)} = [A_{11}]^{-1}\,[\mathbf{b}^{(1)} - A_{12}\mathbf{x}^{(2)}] \tag{3.86}$$

$$A_{21}\,[A_{11}]^{-1}\,[\mathbf{b}^{(1)} - A_{12}\mathbf{x}^{(2)}] + A_{22}\mathbf{x}^{(2)} = \mathbf{b}^{(2)} \tag{3.87}$$

$$\left[A_{22} - A_{21}\,[A_{11}]^{-1}\,A_{12}\right]\mathbf{x}^{(2)} = \mathbf{b}^{(2)} - (A_{21}A_{11}^{-1})\mathbf{b}^{(1)} \tag{3.88}$$

or

$$\mathbf{x}^{(2)} = \left[A_{22} - A_{21}\,[A_{11}]^{-1}\,A_{12}\right]^{-1}\left[\mathbf{b}^{(2)} - (A_{21}A_{11}^{-1})\mathbf{b}^{(1)}\right]$$

It is also possible to work with higher number of partitions equal to say 9, 16 .... and solve the given system.

4. Iterative Solution Techniques for solving $\mathbf{Ax} = \mathbf{b}$

By this approach, we starts from any initial guess, say $\mathbf{x}^{(0)}$,and generate an improved estimate $\mathbf{x}^{(k+1)}$ from previous approximation $\mathbf{x}^{(k)}$. This sequence is terminated when some norm of the residue $\left\|\mathbf{r}^{(k)}\right\| = \left\|A\mathbf{x}^{(k)} - \mathbf{b}\right\|$ becomes sufficiently small. Such a method can be developed by splitting matrix A. If A is expressed as

$$A = S - T \tag{4.1}$$

then, equation (1.5) can be written as

$$S\mathbf{x} = T\mathbf{x} + \mathbf{b}$$

Thus, starting from a guess solution

$$\mathbf{x}^{(0)} = [x_1^{(0)}........x_n^{(0)}] \tag{4.2}$$

we can generate a sequence of approximate vectors as follows

$$\mathbf{x}^{(k+1)} = S^{-1}[T\mathbf{x}^{(k)} + \mathbf{b}] \quad ; \quad (k = 0, 1, 2, ...) \tag{4.3}$$

Requirements on S and T matrices are as follows [**15**] : matrix A should be decomposed into $A = S - T$ such that

- Matrix S should be easily invertible
- Sequence $\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$ should converge to $\mathbf{x}^*$ where $\mathbf{x}^*$ is the solution of $A\mathbf{x} = \mathbf{b}$.

At the k'th iteration step, we can write

$$\mathbf{x}^{(k)} = \left(S^{-1}T\right)^k \mathbf{x}^{(0)} + \left[\left(S^{-1}T\right)^{k-1} + \left(S^{-1}T\right)^{k-2} + ... + S^{-1}T + I\right] S^{-1}\mathbf{b} \tag{4.4}$$

If we select $(S^{-1}T)$ such that

$$\begin{matrix}\lim \\ k \to \infty\end{matrix} \left(S^{-1}T\right)^k = [\mathbf{0}] \tag{4.5}$$

where $[\mathbf{0}]$ represents null matrix, then, using identity

$$\left[I - \left(S^{-1}T\right)\right]^{-1} = I + \left(S^{-1}T\right) + ... + \left(S^{-1}T\right)^{k-1} + \left(S^{-1}T\right)^k + ...$$

we can write

$$\mathbf{x}^{(k)} \to \left[I - \left(S^{-1}T\right)\right]^{-1} S^{-1}\mathbf{b} = [S - T]^{-1}\,\mathbf{b} = A^{-1}\mathbf{b}$$

for large k. The above expression clearly explains how the iteration sequence generates a numerical approximation to $A^{-1}\mathbf{b}$, provided condition (4.5) is satisfied.

Let D, L and U be diagonal, strictly lower triangular and strictly upper triangular parts of A, i.e.,

$$A = L + D + U \tag{4.6}$$

There are three popular iterative formulations [**15**]

- **Jacobi Method:**

$$S = D \tag{4.7}$$

$$T = -(L + U) \tag{4.8}$$

- **Gauss-Seidel Method**

$$S = L + D \tag{4.9}$$

$$T = -U \tag{4.10}$$

- **Relaxation Method:**

$$S = L + \left(\frac{1}{\omega}\right) D \tag{4.11}$$

$$T = \left(\frac{1 - \omega}{\omega}\right) D - U \tag{4.12}$$

where $0 < \omega < 2$

The above formulations in vector-matrix notation, though helps in analyzing these algorithms, is not suitable for developing computer programs. We will derive computationally efficient algorithms and convergence criteria for each of these iteration schemes in the following subsections.

4.1. Jacobi-Method. In order to understand the rationale behind formulation of Jacobi iterations, consider the first equation in the set of equations $A\mathbf{x} = \mathbf{b}$, i.e.,

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \tag{4.13}$$

Rearranging this equation, we can arrive at a iterative formula for computing $x_1^{(k+1)}, as$

$$x_1^{(k+1)} = [b_1 - a_{12}x_2^{(k)}....... - a_{1n}x_n^{(k)}]/a_{11} \tag{4.14}$$

Similarly, using second equation from $A\mathbf{x} = \mathbf{b}$, we can derive

$$x_2^{(k+1)} = [b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}........ - a_{2n}x_n^{(k)}]/a_{22} \tag{4.15}$$

and, in general, using ithrow

$$x_i^{(k+1)} = [b_2 - a_{i1}x_1^{(k)}...... - a_{i,i-1}x_{i-1}^{(k)} - a_{i,i+1}x_{i+1}^{(k)}..... - a_{i,n}x_n^{(k)}]/a_{ii} \tag{4.16}$$

Arranging the above equations in the matrix form yields

$$\mathbf{x}^{(k+1)} = -D^{-1}(L+U)\mathbf{x}^{(k)} + D^{-1}\mathbf{b} \tag{4.17}$$

$$D\mathbf{x}^{(k+1)} = -(L+U)\mathbf{x}^{(k)} + \mathbf{b} \tag{4.18}$$

In the above derivation, it is implicitly assumed that $a_{ii} \neq 0$. If this is not the case, simple row exchange is often sufficient to satisfy this condition. Suppose we denote residue vector $\mathbf{r}$ as

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} \tag{4.19}$$

$$\text{i.e. } r_i^{(k)} = b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)} \tag{4.20}$$

then, the standard termination criterion is

$$\frac{\left\|\mathbf{r}^{(k)}\right\|}{\|\mathbf{b}\|} < \varepsilon$$

where ε is an arbitrarily small number (such as 10^{-6} or 10^{-8}). Another standard termination criterion can be

$$\frac{\left\|x^{(k)} - x^{(k+1)}\right\|}{\left\|x^{(k+1)}\right\|} < \varepsilon$$

This condition is *practically* equivalent to the previous condition. Equation (4.16) is more suitable from the view point of programming than equation (4.3) and the algorithm can be stated as follows:

Jacobi Algorithm

```
INITIALIZE :$\mathbf{b}, A, \mathbf{x}^{(0)}, k_{\max}, \varepsilon$
            $k = 0$
            $\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon)\ \ AND\ \ (k < k_{\max})]$
    FOR $i = 1 : n$
        $r_i = b_i - \sum_{j=1}^{n} a_{ij} x_j^{(0)}$
        $x_i^{(1)} = x_i^{(0)} + (r_i/a_{ii})$
    END FOR
    $\delta = \|\mathbf{r}\| \,/\, \|\mathbf{b}\|$
    $k = k + 1$
    $\mathbf{x}^{(0)} = \mathbf{x}^{(1)}$
END WHILE
```

4.2. Gauss-Seidel Method. When matrix A is large, there is a practical difficulty with the Jacobi method. It required to store all components of $\mathbf{x}^{(k)}$ in the computer memory (as a separate variables) until calculations of $\mathbf{x}^{(k+1)}$ is over. The Gauss-Seidel method overcomes this difficulty by using $x_i^{(k+1)}$ immediately in the next equation while computing $x_{i+1}^{(k+1)}$.This modification leads to the following set of equations

$$x_1^{(k+1)} = [b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}......a_{1n}x_n^{(k)}]/a_{11} \tag{4.21}$$

$$x_2^{(k+1)} = [b_2 - \left\{a_{21}x_1^{(k+1)}\right\} - \left\{a_{23}x_3^{(k)} + ... + a_{2n}x_n^{(k)}\right\}]/a_{22} \tag{4.22}$$

$$x_3^{(k+1)} = [b_3 - \left\{a_{31}x_1^{(k+1)} + a_{32}x_2^{(k+1)}\right\} - \left\{a_{34}x_4^{(k)} + ... + a_{3n}x_n^{(k)}\right\}]/a_{33} \tag{4.23}$$

$$\begin{aligned} ...... &= .... \\ x_n^{(k+1)} &= [b_n - a_{n1}x_1^{(k+1)}......a_{n,n-1}x_{n-1}^{(k+1)}]/a_{nn} \end{aligned} \tag{4.24}$$

Again it is implicitly assumed that pivots a_{ii} are non zero. Now in the above set of equations, if we move all terms involving $x_i^{(k+1)}$ from R.H.S. to L.H.S. , we get

$$\begin{bmatrix} a_{11} & 0 & 0 & . \\ a_{21} & a_{22} & . & . \\ . & . & . & . \\ a_{n1} & . & . & a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ . \\ . \\ x_n^{(k+1)} \end{bmatrix} \tag{4.25}$$

$$= \begin{bmatrix} 0 & -a_{12} & . & . & -a_{1n} \\ . & . & . & . & . \\ . & . & . & .. & -a_{n-1,n} \\ 0 & . & . & . & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ . \\ . \\ x_n^{(k)} \end{bmatrix} + \begin{bmatrix} b_1 \\ . \\ . \\ b_n \end{bmatrix} \tag{4.26}$$

$$\text{or} \quad (L + D)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b} \tag{4.27}$$

Thus, we have

$$\begin{aligned} S &= L + D \\ T &= -U \end{aligned}$$

and, using vector-matrix notation, the Gauss-Seidel method can be stated as

$$\text{or} \quad \mathbf{x}^{(k+1)} = (L + D)^{-1} \left[-U\mathbf{x}^{(k)} + \mathbf{b}\right] \tag{4.28}$$

Gauss-Seidel Algorithm

```
INITIALIZE :$\mathbf{b}, A, \mathbf{x}, k_{\max}, \varepsilon$
            $k = 0$
            $\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon)\ \ AND\ \ (k < k_{\max})]$
    FOR $i = 1 : n$
        $r_i = b_i - \sum_{j=1}^{n} a_{ij} x_j$
        $x_i = x_i + (r_i / a_{ii})$
    END FOR
    $T_c = \|\mathbf{r}\| \,/\, \|\mathbf{b}\|$
    $k = k + 1$
END WHILE
```

4.3. Relaxation Method. Suppose we have a starting value say y, of a quantity and we wish to approach a target value y^* by some method. Let application of the method change the value from y to $\widehat{y}$. If $\widehat{y}$ is between y and $\widetilde{y}$, which is even closer to y^*, then we can approach $\widetilde{y}$ faster by magnifying the change $(\widehat{y} - y)$ [**15**]. In order to achieve this, we need to apply a magnifying factor $\omega > 1$ and get

$$\widetilde{y} - y = \omega\,(\widehat{y} - y) \tag{4.29}$$

$$\text{or} \quad \widetilde{y} = \omega\,\widehat{y} + (1-\omega)\,y \tag{4.30}$$

This amplification process is an extrapolation and is an example of **over-relaxation**. If the intermediate value $\widehat{y}$ tends to overshoot target y^*, then we may have to use $\omega < 1$; this is called **under-relaxation**.

Application of **over-relaxation** to Gauss-Seidel method leads to the following set of equations

$$\begin{aligned} \tilde{x}_i^{(k+1)} &= x_i^{(k)} + \omega[x_i^{(k+1)} - x_i^{(k)}] \\ i &= 1, 2, ...n \end{aligned} \tag{4.31}$$

where $x_i^{(k+1)}$ are generated by Gauss-Seidel method, i.e.,

$$\begin{aligned} x_i^{(k+1)} &= \left(\frac{1}{a_{ii}}\right)\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right] \\ i &= 1, 2, ...n \end{aligned} \tag{4.32}$$

With some algebraic manipulations, the above equations can be rearranged in vector matrix form as follows

$$(D + \omega L)\mathbf{x}^{(k+1)} = [(1-\omega)D - \omega U]\mathbf{x}^{(k)} + \omega\,\mathbf{b} \tag{4.33}$$

```
Relaxation Algorithm
INITIALIZE :$\mathbf{b}, A, \mathbf{x}, k_{\max}, \varepsilon, \omega$
            $k = 0$
            $\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon)\ AND\ (k < k_{\max})]$
    FOR $i = 1 : n$
        $q_i = b_i - \sum_{j=1}^{n} a_{ij}x_j$
        $z_i = x_i + (q_i/a_{ii})$
        $x_i = \omega z_i + (1-\omega)x_i$
    END FOR
    $\mathbf{r} = \mathbf{b} - A\mathbf{x}$
    $T_c = \|\mathbf{r}\| / \|\mathbf{b}\|$
    $k = k + 1$
END WHILE
```

4.4. Convergence of Iterative Methods [15, 5]. In order to solve equation (1.5), we have formulated an iterative scheme

$$\mathbf{x}^{(k+1)} = \left(S^{-1}T\right)\mathbf{x}^{(k)} + S^{-1}\mathbf{b} \tag{4.34}$$

Let the true solution equation (1.5) be

$$\mathbf{x}^{*} = \left(S^{-1}T\right)\mathbf{x}^{*} + S^{-1}\mathbf{b} \tag{4.35}$$

Defining error vector

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^{*} \tag{4.36}$$

and subtracting equation (4.35) from equation (4.34), we get

$$\mathbf{e}^{(k+1)} = \left(S^{-1}T\right)\mathbf{e}^{(k)} \tag{4.37}$$

Thus, if we start with some $\mathbf{e}^{(0)}$, then after k iterations we have

$$\begin{aligned}
\mathbf{e}^{(1)} &= \left(S^{-1}T\right)\mathbf{e}^{(0)} && (4.38)\\
\mathbf{e}^{(2)} &= \left(S^{-2}T^{2}\right)\mathbf{e}^{(1)} = [S^{-1}T]^{2}\mathbf{e}^{(0)} && (4.39)\\
\ldots\ldots\ldots &= \ldots\ldots\ldots && (4.40)\\
\mathbf{e}^{(k)} &= [S^{-1}T]^{k}\mathbf{e}^{(0)} && (4.41)
\end{aligned}$$

The convergence of the iterative scheme is assured if

$$\begin{aligned}
\begin{matrix} lim \\ k \rightarrow \infty \end{matrix}\ \mathbf{e}^{(k)} &= \overline{0} && (4.42)\\
\text{i.e.}\quad \begin{matrix} lim \\ k \rightarrow \infty \end{matrix}\ [S^{-1}T]^{k}\mathbf{e}^{(0)} &= \overline{0} && (4.43)
\end{aligned}$$

for **any** initial guess vector $\mathbf{e}^{(0)}$. It mat be noted that equation (4.37) is a linear difference equation of form

$$\mathbf{z}^{(k+1)} = \mathbf{B}\mathbf{z}^{(k)} \tag{4.44}$$

subject to initial condition $\mathbf{z}^{(0)}$.Here $\mathbf{z} \in \mathbf{R}^{n}$ and $\mathbf{B}$ is a $n \times n$ matrix. In the next sub-section, we analyze behavior of the solutions of linear difference equations of type (4.44). We then proceed with applying these general results to the specific problem at hand. i.e. convergence of iteration schemes for solving linear algebraic equations.

4.4.1. *Eigenvalue Analysis.* To begin with, let us consider scalar linear iteration scheme

$$\mathbf{z}^{(k+1)} = b\mathbf{z}^{(k)} \tag{4.45}$$

where $\mathbf{z}^{(k)} \in R$ and b is a real scalar. It can be seen that

$$\mathbf{z}^{(k)} = (b)^k \mathbf{z}^{(0)} \rightarrow 0 \text{ as } k \rightarrow \infty \tag{4.46}$$

if and only if $|b| < 1$.To generalize this notation to a multidimensional case, consider equation of type (4.44) where $\mathbf{z}^{(k)} \in R^n$. Taking motivation from the scalar case, we propose a solution to equation (4.44) of type

$$\mathbf{z}^{(k)} = \lambda^k \mathbf{v} \tag{4.47}$$

where λ is a scalar and $\mathbf{v} \in R^n$ is a vector. Substituting equation (4.47) in equation (6.54), we get

$$\begin{aligned} \lambda^{k+1}\mathbf{v} &= \mathbf{B}(\lambda^k \mathbf{v}) && (4.48) \\ \text{or } \lambda^k \left(\lambda I - \mathbf{B}\right) \mathbf{v} &= \overline{0} && (4.49) \end{aligned}$$

Since we are interested in a non-trivial solution, the above equation can be reduced to

$$\left(\lambda I - \mathbf{B}\right) \mathbf{v} = \overline{0} \tag{4.50}$$

where $\mathbf{v} \neq \overline{0}$. Note that the above set of equations has n equations in $(n+1)$ unknowns (λ and n elements of vector $\mathbf{v}$). Moreover, these equations are nonlinear. Thus, we need to generate an additional equation to be able to solve the above set exactly. Now, the above equation can hold only when the columns of matrix $(\lambda I - \mathbf{B})$ are linearly dependent and $\mathbf{v}$ belongs to null space of $(\lambda I - \mathbf{B})$. If columns of matrix $(\lambda I - \mathbf{B})$ are linearly dependent, matrix $(\lambda I - \mathbf{B})$ is singular and we have

$$\det\left(\lambda I - \mathbf{B}\right) = 0 \tag{4.51}$$

Note that equation (4.51) is nothing but the characteristic polynomial of matrix A and its roots are called eigenvalues of matrix A. For each eigenvalue λ_i we can find the corresponding eigen vector $\mathbf{v}^{(i)}$ such that

$$\mathbf{B}\mathbf{v}^{(i)} = \lambda_i \mathbf{v}^{(i)} \tag{4.52}$$

Thus, we get n fundamental solutions of the form $(\lambda_i)^k \mathbf{v}^{(i)}$ to equation (6.54) and a general solution to equation (6.54) can be expressed as linear combination of these fundamental solutions

$$\mathbf{z}^{(k)} = \alpha_1 \left(\lambda_1\right)^k \mathbf{v}^{(1)} + \alpha_2 \left(\lambda_2\right)^k \mathbf{v}^{(2)} + ... + \alpha_n \left(\lambda_n\right)^k \mathbf{v}^{(n)} \tag{4.53}$$

Now, at $k = 0$ this solution must satisfy the condition

$$\begin{aligned}
\mathbf{z}^{(0)} &= \alpha_1 \mathbf{v}^{(1)} + \left(\alpha_2\right)^k \mathbf{v}^{(2)} + ... + \alpha_n \mathbf{v}^{(n)} && (4.54)\\
&= \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & .... & \mathbf{v}^{(n)} \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 & .... & \alpha_n \end{bmatrix}^T && (4.55)\\
&= \Psi \boldsymbol{\alpha} && (4.56)
\end{aligned}$$

where Ψ is a $n \times n$ matrix with eigenvectors as columns and $\boldsymbol{\alpha}$ is a $n \times 1$ vector of n coefficients. Let us consider the special case when the eigenvectors are linearly independent. Then, we can express $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha} = \Psi^{-1} \mathbf{z}^{(0)} \tag{4.57}$$

Behavior of equation (4.53) can be analyzed as $k \rightarrow \infty$.Contribution due to the i'th fundamental solution $\left(\lambda_i\right)^k \mathbf{v}^{(i)} \rightarrow \overline{0}$ if and only if $|\lambda_i| < 1$.Thus, $\mathbf{z}^{(k)} \rightarrow \overline{0}$ as $k \rightarrow \infty$ if and only if

$$|\lambda_i| < 1 \text{ for } i = 1, 2, ...n \tag{4.58}$$

If we define **spectral radius** of matrix A as

$$\rho(\mathbf{B}) = \begin{array}{c} \max \\ i \end{array} |\lambda_i| \tag{4.59}$$

then, the condition for convergence of iteration equation (6.54) can be stated as

$$\rho(\mathbf{B}) < 1 \tag{4.60}$$

Equation (4.53) can be further simplified as

$$\mathbf{z}^{(k)} = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & .... & \mathbf{v}^{(n)} \end{bmatrix} \begin{bmatrix} \left(\lambda_1\right)^k & 0 & ..... & 0 \\ 0 & \left(\lambda_2\right)^k & 0 & ... \\ .... & .... & ..... & .... \\ 0 & .... & 0 & \left(\lambda_n\right)^k \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_n \end{bmatrix} \tag{4.61}$$

$$= \Psi \begin{bmatrix} \left(\lambda_1\right)^k & 0 & ..... & 0 \\ 0 & \left(\lambda_2\right)^k & 0 & ... \\ .... & .... & ..... & .... \\ 0 & .... & 0 & \left(\lambda_n\right)^k \end{bmatrix} \Psi^{-1} \mathbf{z}^{(0)} = \Psi \left(\Lambda\right)^k \Psi^{-1} \mathbf{z}^{(0)} \tag{4.62}$$

where Λ is the diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & ..... & 0 \\ 0 & \lambda_2 & 0 & ... \\ .... & .... & ..... & .... \\ 0 & .... & 0 & \lambda_n \end{bmatrix} \tag{4.63}$$

Now, consider set of n equations

$$\mathbf{B}\mathbf{v}^{(i)} = \lambda_i \mathbf{v}^{(i)} \quad \text{for} \quad (i = 1, 2, ...n) \tag{4.64}$$

which can be rearranged as

$$\Psi = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & .... & \mathbf{v}^{(n)} \end{bmatrix}$$

$$\mathbf{B}\Psi = \begin{bmatrix} \lambda_1 & 0 & ..... & 0 \\ 0 & \lambda_2 & 0 & ... \\ .... & .... & ..... & .... \\ 0 & .... & 0 & \lambda_n \end{bmatrix} \Psi \tag{4.65}$$

$$\text{or } \mathbf{B} = \Psi\Lambda\Psi^{-1} \tag{4.66}$$

Using above identity, it can be shown that

$$\mathbf{B}^k = \left(\Psi\Lambda\Psi^{-1}\right)^k = \Psi\left(\Lambda\right)^k \Psi^{-1} \tag{4.67}$$

and the solution of equation (6.54) reduces to

$$\mathbf{z}^{(k)} = \mathbf{B}^k \mathbf{z}^{(0)} \tag{4.68}$$

and $\mathbf{z}^{(k)} \rightarrow \overline{0}$ as $k \rightarrow \infty$ if and only if $\rho(\mathbf{B}) < 1$. The largest magnitude eigen value, i.e., $\rho(\mathbf{B})$ will eventually dominate and determine the rate at which $\mathbf{z}^{(k)} \rightarrow \bar{0}$. The result proved in this section can be summarized as follows:

Theorem 3. *A sequence of vectors* $\left\{\mathbf{z}^{(k)} : k = 0, 1, , 2, ...\right\}$ *generated by the iteration scheme*

$$\mathbf{z}^{(k+1)} = \mathbf{B}\mathbf{z}^{(k)}$$

where $\mathbf{z} \in R^n$ *and* $\mathbf{B} \in R^n \times R^n$, *starting from any arbitrary initial condition* $\mathbf{z}^{(0)}$ *will converge to limit* $\mathbf{z}^* = \bar{0}$ *if and only if*

$$\rho(\mathbf{B}) < 1$$

Note that computation of eigenvalues is a computationally intensive task. The following theorem helps derive a sufficient condition for convergence of linear iterative equations.

THEOREM 4. *For a $n \times n$ matrix $\mathbf{B}$, the following inequality holds for any induced matrix norm*

$$\rho(\mathbf{B}) \leq \|\mathbf{B}\| \tag{4.69}$$

Proof: *Let λ be eigen value of $\mathbf{B}$ and $\mathbf{v}$ be the corresponding eigenvector. Then, we can write*

$$\|\mathbf{Bv}\| = \|\lambda\mathbf{v}\| = |\lambda| \, \|\mathbf{v}\| \tag{4.70}$$

By definition of induced matrix norm

$$\|\mathbf{Bz}\| \leq \|\mathbf{B}\| \, \|\mathbf{z}\| \tag{4.71}$$

Thus, $|\lambda| \leq \|\mathbf{B}\|$ for any $\mathbf{z} \in R^n$. This inequality is true for all λ and this implies

$$\rho(\mathbf{B}) \leq \|\mathbf{B}\| \tag{4.72}$$

Using above theorem, a sufficient condition for convergence of iterative scheme can be derived as

$$\|\mathbf{B}\| < 1 \tag{4.73}$$

as $\rho(\mathbf{B}) \leq \|\mathbf{B}\| < 1 \Rightarrow \rho(\mathbf{B}) < 1$.

The above sufficient condition is more useful from the viewpoint of computations as $\|\mathbf{B}\|_1$ and $\|\mathbf{B}\|_\infty$ can be computed quite easily. On the other hand, the spectral radius of a large matrix can be comparatively difficult to compute.

4.4.2. *Convergence Criteria for Iteration Schemes.* The criterion for convergence of iteration equation (4.37) can be derived using results derived above. The necessary and sufficient condition for convergence of (4.37) can be stated as

$$\rho(S^{-1}T) < 1$$

i.e. the spectral radius of matrix $S^{-1}T$ should be less than one.

The necessary and sufficient condition for convergence stated above requires computation of eigen values of $S^{-1}T$, which is a computationally demanding task when the matrix dimension is large. If for a large dimensional matrix, we could check this condition before starting iterations, then we might as well solve the problem by a direct method rather than using iterative approach to save computations. Thus, there is a need to derive some alternate criteria for convergence, which can be checked easily before starting iterations. For example, using Theorem 3.3, we can obtain sufficient conditions for convergence

$$\left\|S^{-1}T\right\|_1 < 1 \quad \text{OR} \quad \left\|S^{-1}T\right\|_\infty < 1$$

which are significantly easy to evaluate than the spectral radius. Also, if the matrix A has some special properties, such as diagonal dominance or symmetry and positive definiteness, then we can derive easily computable criteria by exploiting these properties.

DEFINITION 26. *A matrix A is called diagonally dominant if*

$$\sum_{j=1(j\neq i)}^{n}|a_{ij}|<|a_{ii}| \quad \textit{for}\ \ i=1,2.,...n \tag{4.74}$$

THEOREM 5. [**5**] *A sufficient condition for the convergence of Jacobi and Gauss-Seidel methods is that the matrix A of linear system $Ax=b$ is diagonally dominant.*

Proof: *See Appendix.*

THEOREM 6. [**4**] *The Gauss-Seidel iterations converge if matrix A an symmetric and positive definite.*

Proof: *See Appendix.*

THEOREM 7. [**15**] *For an arbitrary matrix A, the necessary condition for the convergence of relaxation method is $0<\omega<2$.*

Proof: *The relaxation iteration equation can be given as*

$$\mathbf{x}^{(k+1)}=(D+\omega L)^{-1}\left[[(1-\omega)D-\omega U]\mathbf{x}^{(k)}+\omega\,\mathbf{b}\right] \tag{4.75}$$

Defining

$$B_{\omega} = (D+\omega L)^{-1}\left[(1-\omega)D-\omega U\right] \tag{4.76}$$

$$\det(B_{\omega}) = \det\left[(D+\omega L)^{-1}\right]\det\left[(1-\omega)D-\omega U\right] \tag{4.77}$$

Now, using the fact that the determinant of a triangular matrix is equal to multiplication of its diagonal elements, we have

$$\det(B_{\omega})=\det\left[D^{-1}\right]\det\left[(1-\omega)D\right]=(1-\omega)^{n} \tag{4.78}$$

Using the result that product of eigenvalues of B_{ω} is equal to determinant of B_{ω},we have

$$\lambda_1\lambda_2...\lambda_n=(1-\omega)^{n} \tag{4.79}$$

where $\lambda_i\,(i=1,2...n)$ denote eigenvalues of B_{ω}.

$$|\lambda_1\lambda_2...\lambda_n|=|\lambda_1|\,|\lambda_2|\,....\,|\lambda_n|=|(1-\omega)^{n}| \tag{4.80}$$

It is assumed that iterations converge. Now, convergence criterion requires

$$\lambda_i(B_{\omega})<1 \ \ \textit{for}\ i=1,2,...n \tag{4.81}$$

$$\Rightarrow \quad |\lambda_1| \, |\lambda_2| \, ... \, |\lambda_n| < 1 \tag{4.82}$$

$$\Rightarrow \quad |(1-\omega)^n| < 1 \tag{4.83}$$

This is possible only if

$$0 < \omega < 2 \tag{4.84}$$

The optimal or the best choice of the ω is the one that makes spectral radius $\rho(B_\omega)$smallest possible and gives fastest rate of convergence.

Theorem 8. [**5**] *A sufficient condition for the convergence of relaxation methods when matrix A of linear system $Ax = b$ is strictly diagonally dominant is that* $0 < \omega \leq 1$.

Proof: *Left to reader as an exercise.*

Theorem 9. [**5**]*For an symmetric and positive definite matrix A, the relaxation method converges if and only if* $0 < \omega < 2$.

The theorems 3 and 6 guarantees convergence of Gauss-Seidel method or relaxation method when matrix A is symmetric and positive definite. Now, what do we do if matrix A in $A\mathbf{x} = \mathbf{b}$ is not symmetric and positive definite? Can we transform the problem to another equivalent problem such that conditions for above theorem are satisfied? Well, if matrix A is non-singular, we can multiply both the sides of the equation by A^T and transform the original problem as

$$\left(A^T A\right) \mathbf{x} = \left(A^T \mathbf{b}\right) \tag{4.85}$$

The resulting matrix $\left(A^T A\right)$ is always symmetric and positive definite as

$$\mathbf{x}^T \left(A^T A\right) \mathbf{x} = (A\mathbf{x})^T \left(A\mathbf{x}\right) > 0 \text{ for any } \mathbf{x} \neq \overline{0} \tag{4.86}$$

Now, for the transformed problem, we are guaranteed convergence if we use Gauss-Seidel method. Thus, instead of working with the original system of equations $A\mathbf{x} = \mathbf{b}$, it is always better to formulate the iteration scheme for the transformed problem (4.85).

Example 41. *Consider system $A\mathbf{x} = \mathbf{b}$ where*

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \tag{4.87}$$

For Jacobi method

$$S^{-1}T \quad = \quad \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \tag{4.88}$$

$$\rho(S^{-1}T) \quad = \quad 1/2 \tag{4.89}$$

Thus, the error norm at each iteration is reduced by factor of 0.5
For Gauss-Seidel method

$$S^{-1}T = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix} \tag{4.90}$$

$$\rho(S^{-1}T) = 1/4 \tag{4.91}$$

Thus, the error norm at each iteration is reduced by factor of 1/4. This implies that, for the example under consideration

$$\textit{1 Gauss Seidel iteration} \equiv \textit{2 Jacobi iterations} \tag{4.92}$$

For relaxation method,

$$S^{-1}T = \begin{bmatrix} 2 & 0 \\ -\omega & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2(1-\omega) & \omega \\ 0 & 2(1-\omega) \end{bmatrix} \tag{4.93}$$

$$= \begin{bmatrix} (1-\omega) & (\omega/2) \\ (\omega/2)(1-\omega) & (1-\omega+\dfrac{\omega^2}{4}) \end{bmatrix} \tag{4.94}$$

$$\lambda_1\lambda_2 = \det(S^{-1}T) = (1-\omega)^2 \tag{4.95}$$

$$\lambda_1+\lambda_2 = trace(S^{-1}T) \tag{4.96}$$

$$= 2-2\omega+\frac{\omega^2}{4} \tag{4.97}$$

Now, if we plot $\rho(S^{-1}T)$ v/s ω, then it is observed that $\lambda_1 = \lambda_2$ at $\omega = \omega_{opt}$.From equation (4.95), it follows that

$$\lambda_1 = \lambda_2 = \omega_{opt} - 1 \tag{4.98}$$

at optimum ω.Now,

$$\lambda_1+\lambda_2 = 2(\omega_{opt}-1) \tag{4.99}$$

$$= 2-2\omega_{opt}+\frac{\omega_{opt}^2}{4} \tag{4.100}$$

$$\Rightarrow \omega_{opt} = 4(2-\sqrt{3}) \cong 1.07 \tag{4.101}$$

$$\Rightarrow \rho(S^{-1}T) = \lambda_1 = \lambda_2 \cong 0.07 \tag{4.102}$$

This is a major reduction in spectral radius when compared to Gauss-Seidel method. Thus, the error norm at each iteration is reduced by factor of 1/16 ($\cong 0.07$) if we choose $\omega = \omega_{opt}$.

Example 42. *Consider system $A\mathbf{x} = \mathbf{b}$ where*

$$A = \begin{bmatrix} 4 & 5 & 9 \\ 7 & 1 & 6 \\ 5 & 2 & 9 \end{bmatrix} \quad ; \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{4.103}$$

If we use Gauss-Seidel method to solve for $\mathbf{x}$, *the iterations do not converge as*

$$S^{-1}T = \begin{bmatrix} 4 & 0 & 0 \\ 7 & 1 & 0 \\ 5 & 2 & 9 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -5 & -9 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.104}$$

$$\rho(S^{-1}T) = 5.9 > 1 \tag{4.105}$$

Now, let us modify the problem by pre-multiplying $A\mathbf{x} = \mathbf{b}$ by A^T on both the sides, i.e. the modified problem is $\left(A^TA\right)\mathbf{x} = \left(A^T\mathbf{b}\right)$. The modified problem becomes

$$A^TA = \begin{bmatrix} 90 & 37 & 123 \\ 37 & 36 & 69 \\ 123 & 69 & 198 \end{bmatrix} \quad ; \quad A^T\mathbf{b} = \begin{bmatrix} 16 \\ 8 \\ 24 \end{bmatrix} \tag{4.106}$$

The matrixA^TA is symmetric and positive definite and, according to Theorem 3, the iterations should converge if Gauss-Seidel method is used. For the transformed problem, we have

$$S^{-1}T = \begin{bmatrix} 90 & 0 & 0 \\ 37 & 36 & 0 \\ 123 & 69 & 198 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -37 & -123 \\ 0 & 0 & -69 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.107}$$

$$\rho(S^{-1}T) = 0.96 < 1 \tag{4.108}$$

and within 220 iterations (termination criterion 1×10^{-5}), we get following solution

$$\mathbf{x} = \begin{bmatrix} 0.0937 \\ 0.0312 \\ 0.0521 \end{bmatrix} \tag{4.109}$$

which is close to the solution

$$\mathbf{x}^* = \begin{bmatrix} 0.0937 \\ 0.0313 \\ 0.0521 \end{bmatrix} \tag{4.110}$$

computed as $\mathbf{x}^* = A^{-1}\mathbf{b}$.

From example 3, we can clearly see that the rate of convergence depends on $\rho(S^{-1}T)$. From analysis of some simple problems, we can generate the following table [**5**]

Method	Convergence Rate	No. of iterations
Jacobi	$O(1/2n^2)$	$O(2n^2)$
Gauss_Seidel	$O(1/n^2)$	$O(n^2)$
Relaxation with optimal ω	$O(2/n)$	$O(n/2)$

5. Well Conditioned and Ill-Conditioned Problems

One of the important issue in computing solutions of large dimensional linear system of equations is the round-off errors caused by the computer. Some matrices are *well conditioned* and the computations proceed smoothly while some are inherently *ill conditioned,* which imposes limitations on how accurately the system of equations can be solved using any computer or solution technique. Before we discuss solution techniques for (1.5), we introduce measures for assessing whether a given system of linear algebraic equations is inherently *ill conditioned* or *well conditioned*.

Normally any computer keeps a fixed number of significant digits. For example, consider a computer that keeps only first three significant digits. Then, adding

$$0.234 + 0.00231 \rightarrow 0.236$$

results in loss of smaller digits in the smaller number. When a computer can commits millions of such errors in a complex computation, the question is, how do these individual errors contribute to the final error in computing the solution? As mentioned earlier, the set of all $m \times n$ matrices together with real scalars defines a linear vector space. Suppose we solve for $A\mathbf{x} = \mathbf{b}$ using LU decomposition, the elimination algorithm actually produce approximate factors L' and U' . Thus, we end up solving the problem with a wrong matrix, i.e.

$$A + \delta A = L'U' \tag{5.1}$$

instead of right matrix $A = LU$. In fact, due to round off errors inherent in any computation using computer, we actually end up solving the equation

$$(A + \delta A)(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b}+\delta\mathbf{b} \tag{5.2}$$

The question is, how serious are the errors $\delta\mathbf{x}$ in solution $\mathbf{x}$, due to round off errors in matrix A and vector $\mathbf{b}$? Can these errors be avoided by rearranging computations or are the computations inherent ill-conditioned? In order to answer these questions, we need to develop some quantitative measure for **matrix conditioning**.

The following section provides motivation for developing a quantitative measure for matrix conditioning. In order to develop such a index, we need to define the concept of norm of a $m \times n$ matrix. The formal definition of matrix norm and method of computing 2-norm of a matrix is explained in the later sub-sections. Concept of condition number of a matrix is introduced next.

5.1. Motivation for looking at Matrix Conditioning. In many situations, if the system of equations under consideration is **numerically well conditioned,** then it is possible to deal with the menace of round off errors by re-arranging the computations. If the system of equations is inherently an **ill conditioned** system, then the rearrangement trick does not help. Let us try and understand this by considering to simple examples and a computer that keeps only three significant digits.

Consider the system (**System-1**)

$$\begin{bmatrix} 0.0001 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{5.3}$$

If we proceed with Guassian elimination without maximal pivoting , then the first elimination step yields

$$\begin{bmatrix} 0.0001 & 1 \\ 0 & -9999 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -9998 \end{bmatrix} \tag{5.4}$$

and with back substitution this results in

$$x_2 = 0.999899 \tag{5.5}$$

which will be rounded off to

$$x_2 = 1 \tag{5.6}$$

in our computer which keeps only three significant digits. The solution then becomes

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 0.0 & 1 \end{bmatrix} \tag{5.7}$$

However, using maximal pivoting strategy the equations can be rearranged as

$$\begin{bmatrix} 1 & 1 \\ 0.0001 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \tag{5.8}$$

and the Guassian elimination yields

$$\begin{bmatrix} 1 & 1 \\ 0 & 0.9999 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0.9998 \end{bmatrix} \tag{5.9}$$

and again due to three digit round off in our computer, the solution becomes

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Thus, A is a well conditioned numerically only thing that can cause calculation blunders is wrong pivoting strategy. If maximum pivoting is used then natural resistance to round off the errors is no longer compromised.

Now, in order to understand difficulties with **ill conditioned** systems, consider another system (**System-2**)

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \tag{5.10}$$

By Guassian elimination

$$\begin{bmatrix} 1 & 1 \\ 0 & 0.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \Longrightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \tag{5.11}$$

If we change R.H.S. of the system 2 by a small amount

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix} \tag{5.12}$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0.0001 \end{bmatrix} \Longrightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{5.13}$$

Note that change in the fifth digit of second element of vector **b** was amplified to change in the first digit of the solution.

Thus, matrix A in the system 2 is **ill conditioned** as it is near singular. Hence, no numerical method can avoid sensitivity of the system 2 to small permutations. The ill conditioning can be shifted from one plane to another but it cannot be eliminated.

5.2. Induced Matrix Norms. We have already mentioned that set of all $m \times n$ matrices with real entries (or complex entries) can be viewed a linear vector space. In this section, we introduce the concept of *induced norm* of a matrix, which plays a vital role in the numerical analysis. A norm of a matrix can be interpreted as *amplification power* of the matrix. To develop a numerical measure for ill conditioning of a matrix, we first have to quantify this *amplification power* of the matrix.

Definition 27. ***(Induced Matrix Norm):*** *The induced norm of a $m \times n$ matrix A is defined as mapping from $R^m \times R^n \rightarrow R^+$ such that*

$$\|A\| = \underset{\mathbf{x} \neq 0}{Max} \ \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \tag{5.14}$$

In other words, $\|A\|$ bounds the amplification power of the matrix i.e.

$$\|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\| \ \text{ for all } \ \mathbf{x} \in R^n \tag{5.15}$$

The equality holds for at least one non zero vector $\mathbf{x} \in R^n$. An alternate way of defining matrix norm is as follows

$$\|A\| = \underset{\mathbf{x} \neq \overline{0}, \|\mathbf{x}\| \leq 1}{Max} \ \|A\mathbf{x}\| \tag{5.16}$$

The following conditions are satisfied for any matrices $A, B \in R^m \times R^n$

1. $\|A\| > 0$ if $A \neq [0]$ and $\|[0]\| = 0$
2. $\|\alpha A\| = |\alpha|.\|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|AB\| \leq \|A\| \, \|B\|$

The induced norms, i.e. norm of matrix induced by vector norms on R^m and R^n, can be interpreted as maximum **gain** or **amplification factor** of the matrix.

Computation of Matrix Norms

Consider **2-norm** of a matrix, which can be defined as

$$||A||_2 = \max_{\mathbf{x} \neq \, 0} \frac{||A\mathbf{x}||_2}{||\mathbf{x}||_2} \tag{5.17}$$

Squaring both sides

$$\begin{aligned} ||A||_2^2 &= \max_{\mathbf{x} \neq \, 0} \frac{||A\mathbf{x}||_2^2}{||\mathbf{x}||_2^2} \\ &= \max_{\mathbf{x} \neq \, 0} \frac{(A\mathbf{x})^T (A\mathbf{x})}{(\mathbf{x}^T \mathbf{x})} \\ &= \max_{\mathbf{x} \neq \, 0} \frac{\mathbf{x}^T B \mathbf{x}}{(\mathbf{x}^T \mathbf{x})} \end{aligned} \tag{5.18}$$

Where $B = A^T A$ is a symmetric and positive definite matrix. Positive definiteness of matrix B implies

$$\mathbf{x}^T B\mathbf{x} > 0 \; if \; \mathbf{x} \neq \; \overline{0} \quad \text{and} \;\; \mathbf{x}^T B\mathbf{x} = \overset{-}{0} \text{ if and only if } \; \mathbf{x} = \overline{0} \tag{5.19}$$

Obviously, if A is nonsingular

$$\mathbf{x}^T B\mathbf{x} = \mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T (A\mathbf{x}) > 0 \; if \; \mathbf{x} \neq \overline{0} \tag{5.20}$$

Now, a positive definite symmetric matrix can be diagonalized as

$$B = \Psi\Lambda\Psi^T \tag{5.21}$$

Where Ψ is matrix with eigen vectors as columns and Λ is the diagonal matrix with eigenvalues of $B\,(= A^TA)$ on the diagonal. Note that in this case Ψ is unitary matrix ,i.e.,

$$\Psi\Psi^T = I \ or \ \Psi^T = \Psi^{-1} \tag{5.22}$$

and eigenvectors are orthogonal. Using the fact that Ψ is unitary, we can write

$$\mathbf{x}^T\mathbf{x} = \mathbf{x}^T\Psi\Psi^T\mathbf{x} = \mathbf{y}^T\mathbf{y} \tag{5.23}$$

$$or \quad \frac{\mathbf{x}^TB\mathbf{x}}{(\mathbf{x}^T\mathbf{x})} = \frac{\mathbf{y}^T\Lambda\mathbf{y}}{(\mathbf{y}^T\mathbf{y})} \text{ where } \mathbf{y} = \Psi^T\mathbf{x} \tag{5.24}$$

Suppose eigenvalues $\lambda_i$of A^TA are numbered such that

$$0 \leq \lambda_1 \leq \lambda_2 \leq .... \leq \lambda_n \tag{5.25}$$

Then

$$\frac{\mathbf{y}^T\Lambda\mathbf{y}}{(\mathbf{y}^T\mathbf{y})} = \frac{(\lambda_1\mathbf{y}_1^2 + ... + \lambda_n\mathbf{y}_n^2)}{(\mathbf{y}_1^2 + ... + \mathbf{y}_n^2)} \leq \lambda_n \tag{5.26}$$

This implies

$$\frac{\mathbf{y}^T\Lambda\mathbf{y}}{(\mathbf{y}^T\mathbf{y})} = \frac{\mathbf{x}^TB\mathbf{x}}{(\mathbf{x}^T\mathbf{x})} = \frac{\mathbf{x}^T(A^TA)\mathbf{x}}{(\mathbf{x}^T\mathbf{x})} \leq \lambda_n \tag{5.27}$$

The equality holds only at the corresponding eigenvector of A^TA, i.e.,

$$\frac{\left[\mathbf{v}^{(n)}\right]^T(A^TA)\mathbf{v}^{(n)}}{\left[\mathbf{v}^{(n)}\right]^T\mathbf{v}^{(n)}} = \frac{\left[\mathbf{v}^{(n)}\right]^T\lambda_n\mathbf{v}^{(n)}}{\left[\mathbf{v}^{(n)}\right]^T\mathbf{v}^{(n)}} = \lambda_n \tag{5.28}$$

Thus,

$$||A||_2^2 = \max_{\mathbf{x}\neq\, 0} ||Ax||^2/||\mathbf{x}||^2 = \lambda_{\max}(A^TA) \tag{5.29}$$

or

$$||A||_2 = [\lambda_{max}(A^TA)]^{1/2} \tag{5.30}$$

where $\lambda_{max}(A^TA)$ denotes maximum magnitude eigenvalue or spectral radius of A^TA. Other commonly used matrix norms are

- **1-norm**: Maximum over column sums

$$||A||_1 = \begin{matrix} \max \\ 1 \leq j \leq n \end{matrix} \left[\sum_{i=1}^{n}|a_{ij}|\right] \tag{5.31}$$

- $\infty-$**norm:** Maximum over row sums

$$||A||_\infty = \begin{array}{c} \max \\ 1 \leq i \leq n \end{array} \left[\sum_{j=1}^{n} |a_{ij}| \right] \tag{5.32}$$

Remark 2. *There are other matrix norms, such as Frobenious norm, which are not induced matrix norms. Frobenious norm is defined as*

$$||A||_F = \left[\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right]^{1/2}$$

5.3. Condition Number: A Measure to Quantify Matrix Ill-conditioning. Consider system of equations given as $A\mathbf{x} = \mathbf{b}$. We examine two situations: (a) errors in representation of vector $\mathbf{b}$ and (b) errors in representation of matrix A.

5.3.1. *Case: Perturbation in vector b.* Consider the case when there is a change in $\mathbf{b}$ to $\mathbf{b} + \delta\mathbf{b}$. Such an error might come from experimental data or from round off error. Such a perturbation causes a change in solution from $\mathbf{x}$ to $\mathbf{x} + \delta\mathbf{x}$.

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \tag{5.33}$$

By subtracting $Ax = \mathbf{b}$ from the above equation we have :

$$A\delta\mathbf{x} = \delta\mathbf{b} \tag{5.34}$$

To develop a measure for conditioning of a matrix we compare relative change/error in solution,i.e. $||\delta\mathbf{x}|| \,/\, ||\mathbf{x}||$ to relative change in $\mathbf{b}$,i.e. $||\delta\mathbf{b}|| \,/\, ||\mathbf{b}||$. Now

$$\delta\mathbf{x} = A^{-1}\delta\mathbf{b} \Rightarrow ||\delta\mathbf{x}|| \leq ||A^{-1}||\ ||\delta\mathbf{b}|| \tag{5.35}$$

Also,

$$A\mathbf{x} = \mathbf{b} \Rightarrow ||\mathbf{b}|| = ||A\mathbf{x}|| \leq ||A||\ ||\mathbf{x}|| \tag{5.36}$$

Combining the above two inequalities, we can write

$$||\delta\mathbf{x}||\ ||\mathbf{b}|| \leq ||A^{-1}||\ ||A||\ ||\mathbf{x}||\ ||\delta\mathbf{b}|| \tag{5.37}$$

$$\Rightarrow \frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \leq (||A^{-1}||\ ||A||)\ \frac{||\delta\mathbf{b}||}{||\mathbf{b}||} \tag{5.38}$$

The above inequality holds for every $\mathbf{b}$ and $\delta\mathbf{b}$ vector. The number

$$C(A) = ||A^{-1}||\ ||A|| \tag{5.39}$$

is called as **condition number** of matrix A. Thus the condition number

$$\frac{||\delta\mathbf{x}||/||\mathbf{x}||}{||\delta\mathbf{b}||/||\mathbf{b}||} \leq C(A) = ||A^{-1}||\ ||A|| \tag{5.40}$$

gives an upper bound on the possible amplification of errors in $\mathbf{b}$ while computing the solution.

5.3.2. *Case: Perturbation in matrix A.* Suppose ,instead of solving for $A\mathbf{x} = \mathbf{b}$ due to truncation errors we end up solving

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} \tag{5.41}$$

Then by subtracting $A\mathbf{x} = \mathbf{b}$ from the above equation we obtain

$$A\delta\mathbf{x} + \delta A(\mathbf{x} + \delta\mathbf{x}) = \overline{0} \tag{5.42}$$

or

$$\delta\mathbf{x} = -A^{-1}\delta A(\mathbf{x} + \delta\mathbf{x}) \tag{5.43}$$

Taking norm on both the sides, we have

$$\begin{aligned}
&& ||\delta\mathbf{x}|| &= ||A^{-1}\delta A(\mathbf{x} + \delta\mathbf{x})|| && (5.44)\\
\text{or} && ||\delta\mathbf{x}|| &\leq ||A^{-1}||\ ||\delta A||\ ||\mathbf{x} + \delta\mathbf{x}| && (5.45)\\
&& ||\delta\mathbf{x}||/||\mathbf{x} + \delta\mathbf{x}|| &\leq (||A^{-1}||\ ||A||)\ ||\delta A||/||A|| && (5.46)\\
&& ||\delta\mathbf{x}||/||\mathbf{x} + \delta\mathbf{x}||/||\delta A||/||A|| &\leq C(A) = ||A^{-1}||\ ||A|| && (5.47)
\end{aligned}$$

Again,the condition number gives an upper bound on % change in solution to % error A. In simple terms, condition number of a matrix tells us how serious is the error in solution of $A\mathbf{x} = \mathbf{b}$ due to the truncation or round off errors in a computer. These inequalities mean that round off error comes from two sources

- Inherent or natural sensitivity of the problem,which is measured by $C(A)$
- Actual errors $\delta\mathbf{b}$ *or* δA.

It has been shown that the maximum pivoting strategy is adequate to keep (δA) in control so that the whole burden of round off errors is carried by the condition number $C(A)$. If condition number is high (>1000), the system is ill conditioned and is more sensitive to round off errors. If condition number is low (<100) system is well conditioned and you should check your algorithm for possible source of errors.

5.3.3. *Computations of condition number.* Let λ_n denote the largest magnitude eigenvalue of matrix A and λ_1 denote the smallest magnitude eigen value of A. Then,

$$||A||_2^2 = \rho(A^T A) = \lambda_n \tag{5.48}$$

Also,

$$||A^{-1}||_2^2 = \rho[(A^{-1})^T A^{-1}] = \rho\left[(AA^T)^{-1}\right] \tag{5.49}$$

This follows from identity

$$\begin{aligned} (A^{-1}A)^T &= I \\ A^T(A^{-1})^T &= I \\ (A^T)^{-1} &= (A^{-1})^T \qquad (5.50) \end{aligned}$$

Now, if λ is eigenvalue of A^TA and $\mathbf{v}$ is the corresponding eigenvector, then

$$(A^TA)\mathbf{v} = \lambda\mathbf{v} \qquad (5.51)$$

$$AA^T(A\mathbf{v}) = \lambda(A\mathbf{v}) \qquad (5.52)$$

λ is also eigenvalue of AA^T. Thus, we can write

$$||A^{-1}||_2^2 = \rho\left[(AA^T)^{-1}\right] = \rho\left[(A^TA)^{-1}\right] \qquad (5.53)$$

Also, since AA^T is a symmetric positive definite matrix, we can diagonalize it as

$$A^TA = \Psi\Lambda\Psi^T \qquad (5.54)$$

$$\Rightarrow (A^TA)^{-1} = [\Psi\Lambda\Psi^T]^{-1} = (\Psi^T)^{-1}\left[\Lambda^{-1}\right]\Psi^{-1} = \Psi\Lambda^{-1}\Psi^T$$

as Ψ is a unitary matrix. Thus, if λ is eigen value of A^TA then $1/\lambda$ is eigen value of $(A^TA)^{-1}$. If $\lambda_1$smallest eigenvalue of A^TA then $1/\lambda_1$ is largest magnitude eigenvalue of A^TA

$$\Rightarrow \rho[(A^TA)^{-1}] = 1/\ \lambda_1 \qquad (5.55)$$

Thus, the condition number of matrix A can be computed using 2-norm as

$$C(A) = ||A||\ ||A^{-1}|| = \ (\lambda_n/\lambda_1)^{1/2} \qquad (5.56)$$

where λ_n and λ_1 are largest and smallest magnitude eigenvalues of A^TA.

Example 43. *Consider matrix*

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad (5.57)$$

This ordinary looking matrix is near singular with eigen values (computed using MATLAB)

$$\lambda_1 = 16.117\ \ ; \lambda_2 = -1.1168\ \ ;\ \ \lambda_3 = -1.0307 \times 10^{-15} \qquad (5.58)$$

has the condition number of $C(A) = 3.8131e + 016$. *If we attempt to compute inverse of this matrix using MATLAB (which has arguably one of the best linear equations solvers) we get following result*

$$A^{-1} = 10^{16} \times \begin{bmatrix} -0.4504 & 0.9007 & -0.4504 \\ 0.9007 & -1.8014 & 0.9007 \\ -0.4504 & 0.9007 & -0.4504 \end{bmatrix} \tag{5.59}$$

with a warning: 'Matrix is close to singular or badly scaled. Results may be inaccurate.' *The difficulties in computing inverse of this matrix are apparent if we further compute product* $A \times A^{-1}$, *which yields*

$$A \times A^{-1} = \begin{bmatrix} 0 & -4 & 0 \\ 0 & 8 & 0 \\ 4 & 0 & 0 \end{bmatrix} \tag{5.60}$$

On the other hand, consider matrix

$$B = 10^{-17} \times \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \\ 1 & 1 & 3 \end{bmatrix} \tag{5.61}$$

with eigen values

$$\lambda_1 = 4.5616 \times 10^{-17} \ ; \lambda_2 = \ -1 \times 10^{-17}; \ \ \lambda_3 = 4.3845 \times 10^{-18} \tag{5.62}$$

Looking at small magnitude eigenvalues (near singularity), we may anticipate trouble in computations. However, the condition number of this matrix is $C(A) = 16.741$. *If we proceed to compute of* B^{-1} *and product* $B \times B^{-1}$ *using MATLAB, we get*

$$B = 10^{17} \times \begin{bmatrix} 0.5 & 1.5 & -2.5 \\ 0.5 & -0.5 & 0.5 \\ -0.5 & -0.5 & 1/5 \end{bmatrix} \tag{5.63}$$

and $B * B^{-1} = I$, *i.e. identity matrix. Thus, it is important to realize that each system of linear equations has a inherent character, which can be quantified using the condition number of the associated matrix. The best of the linear equation solvers cannot overcome the computational difficulties posed inherent ill conditioning of a matrix. As a consequence, when such ill conditioned matrices are encountered, the results obtained using any computer or any solver are unreliable.*

6. Solving Nonlinear Algebraic Equations

Consider set of n nonlinear equations

$$f_i(\mathbf{x}) = 0\,; \quad i = 1, ..., n \tag{6.1}$$

$$\text{or } F(\mathbf{x}) = \overline{0} \tag{6.2}$$

which have to be solved simultaneously. Among the various approaches available for solving these equations, the method of successive substitution and Newton Raphson method are based on successive solutions of subproblems involving linear algebraic equations of type (1.5).

6.1. Successive Substitution [5]. In many situations, equation (6.2) can be rearranged as

$$A\mathbf{x} = G(\mathbf{x}) \tag{6.3}$$

$$G(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) & g_2(\mathbf{x}) & .... & g_n(\mathbf{x}) \end{bmatrix}^T \tag{6.4}$$

such that the solution of equation (6.3) is also solution of equation (6.2). The nonlinear Equation (6.3) can be used to formulate iteration sequence of the form

$$A\mathbf{x}^{(k+1)} = G\left[\mathbf{x}^{(k)}\right] \tag{6.5}$$

Given a guess $\mathbf{x}^{(k)}$,the R.H.S. is a fixed vector, say $\mathbf{b}^{(k)} = G\left[\mathbf{x}^{(k)}\right]$, and computation of the next guess $\mathbf{x}^{(k+1)}$ essentially involves solving the linear algebraic equation

$$A\mathbf{x}^{(k+1)} = \mathbf{b}^{(k)}$$

at each iteration. Thus, the set of nonlinear algebraic equations is solved by formulating a sequence of linear sub-problems. Computationally efficient method of solving such sequence of linear problems would be to use LU decomposition of matrix A.

A special case of interest is when matrix $A = I$ in equation (6.5).In this case, if the set of equations given by (6.3) can be rearranged as

$$x_i = g_i(\mathbf{x}) \;\; ; \;\; (i = 1, ...n) \tag{6.6}$$

then method of successive substitution can be arranged as

- **Jacobi-Iterations**

$$x_i^{(k+1)} = g_i[\mathbf{x}^{(k)}] \;\; ; \;\; (\, i = 1, ..., n)$$

- **Gauss Seidel Iterations**

$$x_i^{(k+1)} = g_i[x_1^{(k+1)}, ...x_{i-1}^{(k+1)}, x_i^{(k)}, ...x_n^{(k)}] \tag{6.7}$$

$$i = 1, ..., n$$

- **Relaxation Method**

$$x_i^{(k+1)} = x_i^{(k)} + \omega[g_i(\mathbf{x}^k) - x_i^{(k)}]$$

A popular method of this type is **Wegstein iterations.** Given initial guess vector $\mathbf{x}^{(0)}$

$$x_i^{(1)} = g_i(\mathbf{x}^{(0)}) \ \ ; \ \ (\, i = 1, ..., n)$$

$$s_i^{(k)} = \frac{\left[g_i(\mathbf{x}^{(k)}) - g_i(\mathbf{x}^{(k-1)})\right]}{\left[x_i^{(k)} - x_i^{(k-1)}\right]}$$

$$\omega_i^{(k)} = \frac{s_i^{(k)}}{\left[1 - s_i^{(k)}\right]} \tag{6.8}$$

$$x_i^{(k+1)} = g_i(\mathbf{x}^{(k)}) + \omega_i^{(k)}[x_i^{(k)} - g_i(\mathbf{x}^{(k)})]$$

$$i = 1, ..., n$$

The iterations can be terminated when

$$\left\Vert \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\Vert < \varepsilon \tag{6.9}$$

EXAMPLE 44. *Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out.*

$$\frac{1}{Pe}\frac{d^2C}{dz^2} - \frac{dC}{dz} - DaC^2 = 0 \qquad (0 \leq z \leq 1) \tag{6.10}$$

$$\frac{dC}{dz} = Pe(C-1) \qquad at \quad z = 0; \tag{6.11}$$

$$\frac{dC}{dz} = 0 \qquad at \quad z = 1; \tag{6.12}$$

Using method of finite difference, we get following set of $n+1$ *nonlinear algebraic equations*

$$\frac{1}{Pe}\frac{C_{i+1} - 2C_i + C_{i-1}}{(\Delta z)^2} - \frac{C_{i+1} - C_{i-1}}{2\,(\Delta z)} = DaC_i^2 \tag{6.13}$$

$$or \qquad \alpha C_{i+1} - \beta C_i + \alpha C_{i-1} = DaC_i^2 \tag{6.14}$$

$$i = 1, 2, ...n - 1$$

where

$$\alpha = \left(\frac{1}{(\Delta z)^2 Pe} + \frac{1}{2(\Delta z)} \right) \quad ; \quad \beta = \left(\frac{2}{Pe(\Delta z)^2} \right) \quad ; \tag{6.15}$$

$$\frac{C_1 - C_0}{\Delta z} = Pe(C_0 - 1) \tag{6.16}$$

$$\frac{C_n - C_{n-1}}{\Delta z} = 0 \tag{6.17}$$

the above set of nonlinear algebraic equations can be arranged as

$$\begin{bmatrix} -(1+\Delta z Pe) & 1 & 0. & ..... & .... & 0 \\ \alpha & -\beta & \alpha & .... & ..... & ..... \\ .... & ..... & ..... & ..... & ..... & 0. \\ ..... & ..... & ..... & ..... & -\beta & \alpha \\ 0 & ..... & ..... & ... & -1 & 1 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ . \\ . \\ C_n \end{bmatrix} = \begin{bmatrix} -Pe(\Delta z) \\ DaC_1^2 \\ ..... \\ DaC_{n-1}^2 \\ 0 \end{bmatrix} \tag{6.18}$$

If we define

$$\mathbf{x} = \begin{bmatrix} C_0 & C_1 & ... & C_n \end{bmatrix}^T \tag{6.19}$$

then, equation (6.18) is of the form $A\mathbf{x} = G(\mathbf{x})$.

6.2. Newton Raphson Method. For a general set of simultaneous equations $F(\mathbf{x}) = \overline{\mathbf{0}}$, it may not be always possible to transform to form $A\mathbf{x} = G(\mathbf{x})$ by simple rearrangement of equations. Even when it is possible, the iterations may not converge. When the function vector $F(\mathbf{x})$ is once differentiable, Newton-Raphson method provides a way to transform an arbitrary set of equations $F(\mathbf{x}) = \overline{\mathbf{0}}$ to form $A\mathbf{x} = G(\mathbf{x})$ using Taylor series expansion.

The main idea behind the Newton-Raphson method is solving the set of nonlinear algebraic equations (6.2) by formulating a sequence of linear subproblems of type

$$A^{(k)}\Delta\mathbf{x}^{(k)} = \mathbf{b}^{(k)} \tag{6.20}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \triangle\mathbf{x}^{(k)} \quad ; \quad k = 0, 1, 2, ... \tag{6.21}$$

in such a way that sequence $\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$ converges to solution of equation (6.2). Suppose $\mathbf{x}^*$ is a solution such that $F(\mathbf{x}^*) = \overline{0}$. If each function

$f_i(\mathbf{x})$ is continuously differentiable, then, in the neighborhood of $\mathbf{x}^*$ we can approximate its behavior by Taylor series, as

$$\mathbf{F}(\mathbf{x}^*) = \mathbf{F}\left[\mathbf{x}^{(k)} + \left(\mathbf{x}^* - \mathbf{x}^{(k)}\right)\right] \tag{6.22}$$

$$= F(\mathbf{x}^{(k)}) + \left[\frac{\partial F}{\partial x}\right]_{\mathbf{x}=\mathbf{x}^{(k)}} \left[\Delta\mathbf{x}^{(k)}\right] + ... \tag{6.23}$$

Since $\mathbf{x}^{(k)}$ is assumed to be close to $\mathbf{x}^*$, second and higher order terms can be neglected. Defining

$$J^{(k)} = \left[\frac{\partial F}{\partial x}\right]_{\mathbf{x}=\mathbf{x}^{(k)}} \tag{6.24}$$

$$F^{(k)} = F(\mathbf{x}^{(k)}) \tag{6.25}$$

we can solve for

$$F(x^*) \cong F^{(k)} + J^{(k)} \triangle \mathbf{x}^{(k)} = \overline{0} \tag{6.26}$$

Now $\triangle\mathbf{x}^{(k)}$ can be interpreted as the error committed in approximating $\mathbf{x}^*$ by $\mathbf{x}^{(k)}$. We can obtain an improved approximations $\mathbf{x}^{(k+1)}$ of $\mathbf{x}^*$ as

$$J^{(k)} \triangle \mathbf{x}^{(k)} = -F^{(k)} \tag{6.27}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \triangle\mathbf{x}^{(k)} \tag{6.28}$$

Alternatively, iterations can be formulated by solving

$$\left[J^{(k)T} J^{(k)}\right] \triangle \mathbf{x}^{(k)} = -J^{(k)T} F^{(k)} \tag{6.29}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \triangle\mathbf{x}^{(k)} \tag{6.30}$$

where $\left[J^{(k)T} J^{(k)}\right]$ is symmetric and positive definite matrix. Iterations can be terminated when the following convergence criteria is satisfied

$$||F(\mathbf{x}^{(k+1)})|| < \varepsilon$$

Often Newton Raphson method finds a large step $\triangle\mathbf{x}^{(k)}$ such that approximation of the function vector by linear term in Taylor series is not valid in interval $\left[\mathbf{x}^{(k)}, \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}\right]$. In order to alleviate this problem,we find a corrected $\mathbf{x}^{(k+1)}$ as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \triangle \mathbf{x}^{(k)} \tag{6.31}$$

where $0 < \lambda^{(k)} \leq 1$ is chosen such that

$$||F(\mathbf{x}^{(k+1)})|| < ||F(\mathbf{x}^{(k)})|| \tag{6.32}$$

In practical problems this algorithm converges faster than the one that uses $\lambda^{(k)} = 1$.

Note that the Newton-Raphson iteration equation can be expressed in form (6.3) as follows

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \lambda[(\partial F/\partial \mathbf{x})^{(k)}]^{-1}F(\mathbf{x}^{(k)}) = G(\mathbf{x}^{(k)}) \tag{6.33}$$

i.e.

$$G(\mathbf{x}) = \mathbf{x} - \lambda[\partial F/\partial \mathbf{x}]^{-1}F(\mathbf{x})$$

It is easy to see that at the solution $\mathbf{x} = \mathbf{x}^*$ of $F(\mathbf{x})$, i.e. when $F(\mathbf{x}^*) = \overline{0}$, the iteration equation 6.33 has a fixed point $\mathbf{x}^* = G(\mathbf{x}^*)$.

Modified Newton Raphson Algorithm:

```
INITIALIZE: $\mathbf{x}^{(0)}, \varepsilon_1, \varepsilon_2,\alpha, k, k_{\max,}\delta_2$
$\delta_1 = \left\|F^{(0)}\right\|$
WHILE $[(\delta_1 > \varepsilon_1)\ AND\ (\delta_2 > \varepsilon_2)\ AND\ (k < k_{max})]$
    Solve
    $J^{(k)} \triangle \mathbf{x}^{(k)} = -F^{(k)}$  (OR $\left[J^{(k)T}J^{(k)}\right] \triangle \mathbf{x}^{(k)} = -J^{(k)T}F^{(k)}$)
    $\lambda^{(0)} = 1,\ j = 0$
    $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(0)} \triangle \mathbf{x}^{(k)}$
    $\delta_1 = \left\|F^{(k+1)}\right\|$
    WHILE$\left[\delta_1 > \left\|F^{(k)})\right\|\right]$
        $\lambda^{(j)} = \alpha\lambda^{(j-1)}$
        $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(j)} \triangle \mathbf{x}^{(k)}$
        $\delta_1 = \left\|F^{(k+1)}\right\|$
    END WHILE
    $\delta_{2=}||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||\,/\,||\mathbf{x}^{(k+1)}||$
    $k = k + 1$
END WHILE
```

6.2.1. *Quasi-Newton Method with Broyden Update.* A major difficulty with Newton Raphson method is that it requires calculation of Jacobian at each iteration. The quasi-Newton methods try to overcome this difficulty by generating approximate successive Jacobians using function vectors evaluated at previous iterations. While moving from iteration k to $(k+1)$, if $||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||$ is not too large, then it can be argued that $J^{(k+1)}$ is "**close"** to $J^{(k)}$. Under such situation ,we can use the following rank-one update of the Jacobian

$$J^{(k+1)} = J^{(k)} + \mathbf{y}^{(k)}[\mathbf{z}^{(k)}]^T \tag{6.34}$$

Here, $\mathbf{y}^{(k)}$ and $\mathbf{z}^{(k)}$ are two vectors that depend on $\mathbf{x}^{(k)}$, $\mathbf{x}^{(k+1)}$, $F^{(k)}$ and $F^{(k+1)}$. To arrive at the update formula, consider Jacobian $J^{(k)}$ that produces step $\triangle\mathbf{x}^{(k)}$ as

$$J^{(k)} \triangle \mathbf{x}^{(k)} = -F^{(k)} \tag{6.35}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \triangle\mathbf{x}^{(k)} \tag{6.36}$$

Step $\mathbf{x}^{(k+1)}$ predicts a function change

$$\Delta F^{(k)} = F^{(k+1)} - F^{(k)} \tag{6.37}$$

We impose the following two conditions to obtain estimate of $J^{(k+1)}$.

(1) In the direction perpendicular to $\triangle\mathbf{x}^{(k)}$, our knowledge about F is maintained by new Jacobian estimate $J^{(k+1)}$. This means for a vector, say $\mathbf{r}$, if $[\Delta\mathbf{x}^{(k)}]^T\ \mathbf{r} = 0$, then

$$J^{(k)}\ \mathbf{r} = J^{(k+1)}\ \mathbf{r} \tag{6.38}$$

In other words, both $J^{(k)}$and $J^{(k+1)}$ will predict some change in direction perpendicular to $\Delta\mathbf{x}^{(k)}$.

(2) $J^{(k+1)}$ predicts for $\triangle\mathbf{x}^{(k)}$, the same $\Delta F^{(k)}$ in linear expansion, i.e.,

$$F^{(k+1)} = F^{(k)} - J^{(k+1)} \triangle \mathbf{x}^{(k)} \tag{6.39}$$

or

$$J^{(k+1)} \triangle \mathbf{x}^{(k)} = \Delta F(k) \tag{6.40}$$

Now, for vector $\mathbf{r}$ perpendicular to$\triangle\mathbf{x}^{(k)}$, we have

$$J^{(k+1)}\ \mathbf{r} = J^{(k)}\mathbf{r} + y^{(k)}[\mathbf{z}^{(k)}]^T\ \mathbf{r} \tag{6.41}$$

As

$$J^{(k+1)}\ \mathbf{r} = J^{(k)}\ \mathbf{r} \tag{6.42}$$

We have

$$y^{(k)}[\mathbf{z}^{(k)}]^T\ \mathbf{r} = 0 \tag{6.43}$$

Since $\triangle\mathbf{x}^{(k)}$ is perpendicular to $\mathbf{r}$, we can choose $\mathbf{z}^{(k)} = \triangle\mathbf{x}^{(k)}$. Substituting this choice of $\mathbf{z}^{(k)}$ in equation (6.34) and post multiplying equation (6.34) by $\triangle\mathbf{x}^{(k)}$, we get

$$J^{(k+1)} \triangle \mathbf{x}^{(k)} = J^{(k)} \triangle \mathbf{x}^{(k)} + y^{(k)}[\triangle\mathbf{x}^{(k)}]^T \triangle \mathbf{x}^{(k)} \tag{6.44}$$

Using equation (3), we have

$$\Delta F^{(k)} = J^{(k)} \triangle \mathbf{x}^{(k)} + y^{(k)}[\triangle\mathbf{x}^{(k)}]^T \triangle \mathbf{x}^{(k)} \tag{6.45}$$

which yields

$$y^{(k)} = \frac{\left[\Delta F^{(k)} - J^{(k)} \triangle \mathbf{x}^{(k)}\right]}{[[\triangle\mathbf{x}^{(k)}]^T \triangle \mathbf{x}^{(k)}]} \tag{6.46}$$

Thus, the Broyden's update formula for the Jacobian is

$$J^{(k+1)} = J^{(k)} + \frac{[\Delta F^{(k)} - J^{(k)} \bigtriangleup \mathbf{x}^{(k)}][\bigtriangleup \mathbf{x}^{(k)}]^T}{[[\bigtriangleup \mathbf{x}^{(k)}]^T \bigtriangleup \mathbf{x}^{(k)}]} \tag{6.47}$$

This can be further simplified as

$$\Delta F^{(k)} - J^{(k)} \bigtriangleup \mathbf{x}^{(k)} = F^{(k+1)} - (F^{(k)} + J^{(k)} \bigtriangleup \mathbf{x}^{(k)}) = F^{(k+1)} \tag{6.48}$$

Thus, Jackobian can be updated as

$$J^{(k+1)} = J^{(k)} + \frac{1}{[[\bigtriangleup \mathbf{x}^{(k)}]^T \bigtriangleup \mathbf{x}^{(k)}]} \left[F^{(k+1)}[\bigtriangleup \mathbf{x}^{(k)}]^T \right] \tag{6.49}$$

Broyden's update derived by an alternate approach yields following formula [**8**]

$$J^{(k+1)} = J^{(k)} - \frac{1}{[[\mathbf{p}^{(k)}]^T J^{(k)} \Delta F^{(k)}]} \left[J^{(k)} \Delta F^{(k)} - \mathbf{p}^{(k)} \right] [\mathbf{p}^{(k)}]^T J^{(k)} \tag{6.50}$$

6.3. Convergence of Iteration Schemes.

Either by successive substitution approach or Newton -Raphson method, we generate an iteration sequence

$$\mathbf{x}^{(k+1)} = G(\mathbf{x}^{(k)}) \tag{6.51}$$

which has a fixed point

$$\mathbf{x}^* = G(\mathbf{x}^*) \tag{6.52}$$

at solution of $F(\mathbf{x}^*) = \overline{0}$.Once we formulate an iteration scheme of the form (6.51), we start from some initial guess solution, say $\mathbf{x}^{(0)}$, and generates a sequence of vectors $\left\{ \mathbf{x}^{(k)} \in R^n : k = 1, 2, 3, ... \right\}$. The iterations are terminated when $\left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\| \leq \varepsilon$ where ε is a small number, provided $\left\{ \mathbf{x}^{(k)} \right\}$ forms a convergent sequence. The main concern while dealing with such equations is whether this sequence of vector converges to some limit, say $\mathbf{x}^*$. In mathematical terms, we want to know conditions under which the iterative sequences starting from arbitrary initial guesses will converge to a **fixed point** $\mathbf{x}^*$of equation (6.5) such that

$$\mathbf{x}^* = G\left[\mathbf{x}^*\right] \tag{6.53}$$

When the equation (6.5) is linear, i.e. it can be expressed in the form

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} \tag{6.54}$$

where $\mathbf{B}$ is $n \times n$ matrix, then necessary and sufficient conditions for convergence can be derived using eigenvalue analysis as shown earlier. When $F\left[\mathbf{x}\right]$ is a nonlinear, contraction mapping theorem is used to derive sufficient conditions for convergence. We discuss both these cases in the following subsections.

6.3.1. *Contraction Mapping Principle.* Contraction mapping theorem develops sufficient conditions for convergence of general nonlinear iterative equation (6.5). Unlike the eigenvalue analysis used for analysis of its linear counterpart, this theorem is more useful for developing understanding of convergence process.

Consider general nonlinear iteration equation (6.5) which defines a mapping from a Banach space X into itself.

DEFINITION 28. ***(Contraction Mapping):*** *An operator $G : X \to X$ given by equation (6.5), mapping a Banach space X into itself, is called a contraction mapping of closed ball*

$$U(\mathbf{x}^{(0)}, r) = \left\{ \mathbf{x} \in X : \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\| \leq r \right\}$$

if there exists a real number θ $(0 \leq \theta < 1)$ such that

$$\left\| G\left(\mathbf{x}^{(1)}\right) - G\left(\mathbf{x}^{(2)}\right) \right\| \leq \theta \left\| \mathbf{x}^{(1)} - \mathbf{x}^{(2)} \right\|$$

for all $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in U(\mathbf{x}^{(0)}, r)$. The quantity θ is called contraction constant of G on $U(\mathbf{x}^{(0)}, r)$.

When the map $F(.)$ is differentiable, an exact characterization of the contraction property can be developed.

LEMMA 3. *Let the operator $G(.)$ on a Banach space X be differentiable in $U(\mathbf{x}^{(0)}, r)$. Operator $G(.)$ is a contraction of $U(\mathbf{x}^{(0)}, r)$ if and only if*

$$\left\| \frac{\partial G}{\partial \mathbf{x}} \right\| \leq \theta < 1 \quad \textit{for every } \mathbf{x} \in U(\mathbf{x}^{(0)}, r)$$

where $\|.\|$ is any induced operator norm.

The contraction mapping theorem is stated next. Here, $\mathbf{x}^{(0)}$ refers to the initial guess vector in the iteration process given by equation (6.5).

THEOREM 10. [**14, 9**] *If $G(.)$ maps $U(\mathbf{x}^{(0)}, r)$ into itself and $G(.)$ is a contraction mapping on the set with contraction constant θ, for*

$$\begin{aligned} r &\geq r_0 \\ r_0 &= \frac{1}{1-\theta} \left\| G\left[\mathbf{x}^{(0)}\right] - \mathbf{x}^{(0)} \right\| \end{aligned}$$

then:

1. *$G(.)$ has a fixed point $\mathbf{x}^*$ in $U(\mathbf{x}^{(0)}, r_0)$ such that $\mathbf{x}^* = G(\mathbf{x}^*)$*
2. *$\mathbf{x}^*$ is unique in $U(\mathbf{x}^{(0)}, r)$*
3. *The sequence $\mathbf{x}^{(k)}$ generated by equation $\mathbf{x}^{(k+1)} = G\left[\mathbf{x}^{(k)}\right]$ converges to $\mathbf{x}^*$ with*

$$\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \leq \theta^k \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|$$

(4) *Furthermore, the sequence* $\overline{\mathbf{x}}^{(k)}$ *generated by equation*

$\overline{\mathbf{x}}^{(k+1)} = G\left[\overline{\mathbf{x}}^{(k)}\right]$ *starting from any initial guess* $\overline{\mathbf{x}}^{(0)} \in U(\mathbf{x}^{(0)}, r_0)$

converges to $\mathbf{x}^*$ *with*

$$\left\|\overline{\mathbf{x}}^{(k)} - \mathbf{x}^*\right\| \leq \theta^k \left\|\overline{\mathbf{x}}^{(0)} - \mathbf{x}^*\right\|$$

The proof of this theorem can be found in Rall [**14**] and Linz [**9**].

Example 45. [**9**] *Consider simultaneous nonlinear equations*

$$z + \frac{1}{4}y^2 = \frac{1}{16} \tag{6.55}$$

$$\frac{1}{3}\sin(z) + y = \frac{1}{2} \tag{6.56}$$

We can form an iteration sequence

$$z^{(k+1)} = \frac{1}{16} - \frac{1}{4}\left(y^{(k)}\right)^2 \tag{6.57}$$

$$y^{(k+1)} = \frac{1}{2} - \frac{1}{3}\sin(z^{(k)}) \tag{6.58}$$

Using $\infty-$*norm In the unit ball* $U(\mathbf{x}^{(0)} = \overline{0}, 1)$ *in the neighborhood of origin, we have*

$$\left\|G\left(\mathbf{x}^{(i)}\right) - G\left(\mathbf{x}^{(j)}\right)\right\|_\infty$$

$$= \max\left(\frac{1}{4}\left|\left(y^{(i)}\right)^2 - \left(y^{(j)}\right)^2\right|, \frac{1}{3}\left|\sin(x^{(i)}) - \sin(x^{(j)})\right|\right) \tag{6.59}$$

$$\leq \max\left(\frac{1}{4}\left|y^{(i)} - y^{(j)}\right|, \frac{1}{3}\left|x^{(i)} - x^{(j)}\right|\right) \tag{6.60}$$

$$\leq \frac{1}{2}\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_\infty \tag{6.61}$$

Thus, G(.) is a contraction map with $\theta = 1/2$ *and the system of equation has a unique solution in the unit ball* $U(\mathbf{x}^{(0)} = \overline{0}, 1)$ *i.e.* $-1 \leq x \leq 1$ *and* $-1 \leq y \leq 1$. *The iteration sequence converges to the solution.*

Example 46. [**9**] *Consider system*

$$x - 2y^2 = -1 \tag{6.62}$$

$$3x^2 - y = 2 \tag{6.63}$$

which has a solution (1,1). The iterative method

$$x^{(k+1)} = 2\left(y^{(k)}\right)^2 - 1 \tag{6.64}$$

$$y^{(k+1)} = 3\left(x^{(k)}\right)^2 - 2 \tag{6.65}$$

is not a contraction mapping near (1,1) and the iterations do not converge even if we start from a value close to the solution. On the other hand, the rearrangement

$$x(k+1) = \sqrt{(y^{(k)}+2)/3} \tag{6.66}$$

$$y^{(k+1)} = \sqrt{(x^{(k)}+1)/2} \tag{6.67}$$

is a contraction mapping and solution converges if the starting guess is close to the solution.

6.3.2. *Convergence Criteria for Iteration Schemes.* Defining error

$$\mathbf{e}^{(k+1)} = \mathbf{x}^{(k+1)} - \mathbf{x}^* = G(\mathbf{x}^{(k)}) - G(\mathbf{x}^*) \tag{6.68}$$

and using Taylor series expansion, we can write

$$G(\mathbf{x}^*) = G[\mathbf{x}^{(k)} - (\mathbf{x}^{(k)} - \mathbf{x}^*)] \tag{6.69}$$

$$\simeq G(\mathbf{x}^{(k)}) - \left[\frac{\partial G}{\partial \mathbf{x}}\right]_{x=x^{(k)}} (\mathbf{x}^{(k)} - \mathbf{x}^*) \tag{6.70}$$

Substituting in (6.68)

$$\mathbf{e}^{(k+1)} = \left[\frac{\partial G}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \mathbf{e}^{(k)} \tag{6.71}$$

where

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*$$

and using definition of induced matrix norm, we can write

$$\frac{||\mathbf{e}^{(k+1)}||}{||\mathbf{e}^{(k)}||} < \left\| \left[\frac{\partial G}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \right\| \tag{6.72}$$

It is easy to see that the successive errors will reduce in magnitude if the following condition is satisfied at each iteration i.e.

$$\left\| \left[\frac{\partial G}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \right\| < 1 \text{ for } k = 1, 2, ... \tag{6.73}$$

Applying **contraction mapping theorem**, a sufficient condition for convergence of iterations in the neighborhood $\mathbf{x}^*$can be stated as

$$\left\| \left[\frac{\partial G}{\partial \mathbf{x}}\right] \right\|_1 \leq \theta_1 < 1 \tag{6.74}$$

or

$$\left\| \left[\frac{\partial G}{\partial \mathbf{x}}\right] \right\|_\infty \leq \theta_\infty < 1$$

Note that this is only a sufficient conditions. If the condition is not satisfied, then the iteration scheme may or may not converge. Also, note that introduction of step length parameter $\lambda^{(k)}$ in Newton-Raphson step as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)}\Delta\mathbf{x}^{(k)} \tag{6.75}$$

such that $\left\|F^{(k+1)}\right\| < \left\|F^{(k)})\right\|$ ensures that $G(\mathbf{x})$ is a contraction map and ensures convergence.

6.4. Condition Number of Nonlinear Set of Equations. Concept of condition number can be easily extended to analyze numerical conditioning of set on nonlinear algebraic equations. Consider nonlinear algebraic equations of the form

$$F(\mathbf{x},\mathbf{u}) = \overline{0} \;\; ; \;\; \mathbf{x} \in R^n \, , \;\; \mathbf{u} \in R^m \tag{6.76}$$

where F is $n \times 1$ function vector and $\mathbf{u}$ is a set of known parameters or independent variables on which the solution depends. The condition number measures the worst possible effect on the solution $\mathbf{x}$ caused by small perturbation in $\mathbf{u}$. Let $\delta\mathbf{x}$ represent the perturbation in the solution caused by perturbation $\delta\mathbf{u}$,i.e.

$$F(\mathbf{x}+\delta\mathbf{x}, \mathbf{u}+\delta\mathbf{u}) = \overline{0} \tag{6.77}$$

Then the condition number of the system of equations is defined as

$$C(\mathbf{x}) = \begin{array}{c} \sup \\ \delta\mathbf{u} \end{array} \frac{||\delta\mathbf{x}||/||\mathbf{x}||}{||\delta\mathbf{u}||/||\mathbf{u}||} \tag{6.78}$$

$$\Rightarrow \frac{||\delta\mathbf{x}||/||\mathbf{x}||}{||\delta\mathbf{u}||/||\mathbf{u}||} \leq C(\mathbf{x}) \tag{6.79}$$

If the solution does not depend continuously on $\mathbf{u}$, then the $C(\mathbf{x})$ becomes infinity and such systems are called as (numerically) unstable systems. Systems with large condition numbers are more susceptible to computational errors.

Example 47. [**1**] *Consider equation*

$$x - e^u = 0 \tag{6.80}$$

Then,

$$\delta x/x = \frac{e^{u+\delta u} - e^u}{e^y} = e^{\delta u} - 1 \tag{6.81}$$

$$C(x) = \begin{array}{c} \sup \\ \delta u \end{array} \left| u\frac{e^{\delta u} - 1}{\delta u} \right| \tag{6.82}$$

For small δu, we have $e^{\delta u} = 1 + \delta u$ and

$$C(x) = |u|$$

7. Solutions of ODE-BVP and PDEs by Orthogonal Collocation

In this section, we demonstrate how combination of Weierstrass theorem and Newton-Raphson method can be used to solve ODE boundary value problem and PDEs. Application of Weierstrass theorem facilitates conversion of ODE-BVP and certain class of PDEs to a set of nonlinear algebraic equations, which can be solved using Newton-Raphson method. Effectively, the ODE-BVP / PDE is solved by forming a sequence of linear algebraic sub-problems.

Consider ODE-BVP described by

$$\Psi[d^2y/dz^2, dy/dz, y, z] = 0 \quad ; \quad z \in \ (0,1) \tag{7.1}$$

$$f_1[dy/dz, y, z] = 0 \ \ at \ z = 0 \tag{7.2}$$

$$f_2[dy/dz, y, z] = 0 \ \ at \ z = 1 \tag{7.3}$$

The true solution to problem is a function, say $y^*(z) \in \mathbf{C^{(2)}[0,1]}$, which belongs to the set of twice differentiable continuous functions**.** According to the Weierstrass theorem, any continuous function over an interval can be approximated with arbitrary accuracy using a polynomial function of appropriate degree. Thus, we assume an (approximate) n'th order polynomial solution to ODE-BVP of the form

$$y(z) = \theta_0 \mathbf{p}_0(z) + ... + \theta_{n+1}\mathbf{p}_{n+1}(z) \tag{7.4}$$

$\{\mathbf{p}_i(z); i = 0, 1, 2, ...n + 1\}$ should be linearly independent vectors in $C^{(2)}[0, 1]$. A straightforward choice of such linearly independent vectors is

$$\mathbf{p}_0(z) = 1, \mathbf{p}_1(z) = z, ...\mathbf{p}_n(z) = z^n \tag{7.5}$$

Alternatively, orthogonal polynomials can be used such as

- **Shifted Legandre polynomials,** which are orthonormal with respect to

$$\langle \mathbf{p}_i(z), \mathbf{p}_j(z) \rangle = \int_0^1 \mathbf{p}_i(z)\mathbf{p}_j(z)dz \tag{7.6}$$

- **Jacobi polynomials,** which are orthonormal with respect to

$$\langle \mathbf{p}_i(z), \mathbf{p}_j(z) \rangle = \int_0^1 W(z)\mathbf{p}_i(z)\mathbf{p}_j(z)dz \tag{7.7}$$

$$W(z) = z^{\alpha}(1-z)^{\beta} \tag{7.8}$$

  where α and β are constants.

The next step is to convert the ODE-BVP to a set of nonlinear algebraic equations. We first select n internal collocation (grid) points at $z = z_i$, $(i = 1, 2, ...n)$ in the interval $[0, 1]$. These collocation points need not be equispaced. It has been shown that, if these collocation points are chosen at roots of n'th order orthogonal polynomial, then the error $|y^*(z) - y(z)|$ is evenly distributed in the entire domain of z. For example, one possibility is to choose the orthogonal collocation points at the roots of shifted Legandre polynomials.

Order (n)	Roots
1	0.5
2	0.21132, 0.78868
3	0.1127, 0.5, 0.8873
4	0.9305,0.6703, 0.3297, 0.0695
5	0.9543, 0.7662, 0.5034, 0.2286, 0.0475
6	0.9698, 0.8221, 0.6262, 0.3792, 0.1681, 0.0346
7	0.9740, 0.8667, 0.7151, 0.4853, 0.3076, 0.1246, 0.0267

In fact, the name *orthogonal collocation* can be attributed to the choice the collocation points at roots of orthogonal polynomials. After selecting the location of collocation points, the approximate solution (4.7) is used to convert the ODE-BVP together with the BCs into a set of nonlinear algebraic equations by setting **residuals** at n collocation (grid) points and the two boundary points equal to zero.

Approach 1

Let us denote

$$y_i(\boldsymbol{\theta}) = y(z_i) = \theta_0\mathbf{p}_0(z_i) + ... + \theta_{n+1}\mathbf{p}_{n+1}(z_i) \tag{7.9}$$

$$y_i'(\boldsymbol{\theta}) = [dy/dz]_{z=z_i} = \theta_0\mathbf{p}_0'(z_i) + ...\theta_{n+1}\mathbf{p}_{n+1}'(z_i) \tag{7.10}$$

$$y_i''(\boldsymbol{\theta}) = \left[d^2y/dz^2\right]_{z=z_i} = \theta_0\mathbf{p}_0''(z_i) + ...\theta_{n+1}\mathbf{p}_{n+1}''(z_i) \tag{7.11}$$

where

$$\boldsymbol{\theta} = [\theta_0..........\theta_{n+1}]^T \tag{7.12}$$

Substitute for y, $y^{'}$ and y'' in equation (3.1) and enforcing residual to be equal to zero at the grid points, we get

$$\Psi[y_i''(\boldsymbol{\theta}),\ y_i'(\boldsymbol{\theta}),\ y_i(\boldsymbol{\theta}),\ z_i] = 0 \tag{7.13}$$

$$i = 1, ...n$$

Similarly, enforcing residuals at boundary points equal to zero yields

$$f_1[y_0'(\boldsymbol{\theta}), y_0(\boldsymbol{\theta}), 0] \quad = 0 \tag{7.14}$$

$$f_2[[y_1'(\boldsymbol{\theta}), y_1(\boldsymbol{\theta}), 1] = 0 \tag{7.15}$$

Thus, we have $n+2$ nonlinear equations in $n+2$ unknowns, which can be solved simultaneously using Newton-Raphson method for estimating $\boldsymbol{\theta}$.

Approach 2

The above approach is not convenient from computational viewpoint. Note that we have to select an initial guess for vector $\boldsymbol{\theta}$ to start the Newton-Raphson method. Now, unlike linear algebraic equations, a set of nonlinear equations can have multiple solutions and the solution we reach by applying Newton Raphson method is governed by the choice of the initial guess vector $\boldsymbol{\theta}^{(0)}$. Instead of working with $\boldsymbol{\theta}$ as unknowns, this approach works with y_i as unknowns. It is much easy to generate initial guess for y_i using knowledge of underlying physics of the problem under consideration. In order to see how this is done, consider $n+2$ equations

$$y_0 = \theta_0\mathbf{p}_0(0) + ... + \theta_n\mathbf{p}_{n+1}(0) \tag{7.16}$$

$$y_i = \theta_0\mathbf{p}_0(z_i) + ... + \theta_n\mathbf{p}_{n+1}(z_i) \tag{7.17}$$

$$i = 1, ...n$$

$$y_{n+1} = \theta_0\mathbf{p}_0(1) + ... + \theta_n\mathbf{p}_{n+1}(1) \tag{7.18}$$

which can be rearranged as

$$\begin{bmatrix} \mathbf{p}_0(0) & \mathbf{p}_1(0) & .... & \mathbf{p}_{n+1}(0) \\ \mathbf{p}_0(z_1) & \mathbf{p}_1(z_1) & .... & \mathbf{p}_{n+1}(z_1) \\ .... & .... & .... & .... \\ \mathbf{p}_0(1) & \mathbf{p}_1(1) & .... & \mathbf{p}_{n+1}(1) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_{n+1} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ ... \\ y_{n+1} \end{bmatrix} \tag{7.19}$$

$$\text{or} \qquad M\boldsymbol{\theta} \;=\; \mathbf{y} \tag{7.20}$$

where the matrix M is computed at the internal collocation points and the boundary points. Using equation (7.20), we can write

$$\boldsymbol{\theta} = M^{-1}\mathbf{y} = R\mathbf{y} \tag{7.21}$$

Using equation (7.21), we can express vector of first derivatives as

$$\begin{bmatrix} y_0' \\ y_1' \\ ... \\ y_{n+1}' \end{bmatrix} = \begin{bmatrix} \mathbf{p}_0'(0) & \mathbf{p}_1'(0) & .... & \mathbf{p}_{n+1}'(0) \\ \mathbf{p}_0'(z_1) & \mathbf{p}_1'(z_1) & .... & \mathbf{p}_{n+1}'(z_1) \\ .... & .... & .... & .... \\ \mathbf{p}_0'(1) & \mathbf{p}_1'(1) & .... & \mathbf{p}_{n+1}'(1) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_{n+1} \end{bmatrix} \tag{7.22}$$

$$= N\boldsymbol{\theta} = [NR]\,\mathbf{y} = S\,\mathbf{y} \tag{7.23}$$

If we express the matrix S as

$$S = \begin{bmatrix} \left[\mathbf{s}^{(0)}\right]^T \\ \left[\mathbf{s}^{(1)}\right]^T \\ .... \\ \left[\mathbf{s}^{(n+1)}\right]^T \end{bmatrix} \tag{7.24}$$

where $\mathbf{s}^{(i)}$ represents $(i+1)$'th row vector of matrix S, then

$$y_i' = \left[\mathbf{s}^{(i)}\right]^T \mathbf{y} \tag{7.25}$$

Similarly,

$$\begin{bmatrix} y_0'' \\ y_1'' \\ ... \\ y_n'' \end{bmatrix} = \begin{bmatrix} \mathbf{p}_0''(0) & \mathbf{p}_1''(0) & .... & \mathbf{p}_{n+1}''(0) \\ \mathbf{p}_0''(z_1) & \mathbf{p}_1''(z_1) & .... & \mathbf{p}_{n+1}''(z_1) \\ .... & .... & .... & .... \\ \mathbf{p}_0''(1) & \mathbf{p}_1''(1) & .... & \mathbf{p}_{n+1}''(1) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_{n+1} \end{bmatrix} \tag{7.26}$$

$$= Q\boldsymbol{\theta} = [QR]\,\mathbf{y} = T\,\mathbf{y} \tag{7.27}$$

and

$$y_i'' = \left[\mathbf{t}^{(i)}\right]^T \mathbf{y} \tag{7.28}$$

where $\left[\mathbf{t}^{(i)}\right]$ represents $(i+1)$'th row vector of matrix T. Using these transformations, the equation (7.13-7.15) can be written as

$$\Psi\left[\left[\mathbf{s}^{(i)}\right]^T\mathbf{y},\ \left[\mathbf{t}^{(i)}\right]^T\mathbf{y},\ y_i,\ z_i\right] = 0 \tag{7.29}$$

$$i = 1, ...n$$

$$f_1\left[\left[\mathbf{s}^{(0)}\right]^T\mathbf{y}, y_0, 0\right] = 0 \tag{7.30}$$

$$f_2\left[\left[\mathbf{s}^{(n+1)}\right]^T\mathbf{y}, y_{n+1}, 1\right] = 0 \tag{7.31}$$

which can to be solved simultaneously for unknown vector $\mathbf{y}$ using Newton-Raphson method. In particular, if we choose

$$\mathbf{p}_0(z) = 1, \mathbf{p}_1(z) = z, ...\mathbf{p}_{n+1}(z) = z^{n+1} \tag{7.32}$$

then, matrices S and T can be computed using

$$M = \begin{bmatrix} \mathbf{1} & 0 & .... & 0 \\ \mathbf{1} & z_1 & .... & (z_1)^{n+1} \\ .... & .... & .... & .... \\ 1 & 1 & .... & 1 \end{bmatrix} \tag{7.33}$$

$$N = \begin{bmatrix} \mathbf{0} & 1 & .... & 0 \\ \mathbf{0} & 1 & .... & (n+1)(z_1)^n \\ .... & .... & .... & .... \\ 0 & 1 & .... & (n+1) \end{bmatrix} \tag{7.34}$$

$$Q = \begin{bmatrix} 0 & \mathbf{0} & \mathbf{2} & 6z_0 & .... & 0 \\ 0 & \mathbf{0} & \mathbf{2} & 6z_1 & .... & n(n+1)(z_1)^{n-1} \\ .... & .... & .... & .... & .... & .... \\ 0 & 0 & 2 & 6z_n & .... & n(n+1) \end{bmatrix} \tag{7.35}$$

where z_i are collocation points.

EXAMPLE 48. [**6**] *Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out*

$$\frac{1}{Pe}\frac{d^2C}{dz^2} - \frac{dC}{dz} - DaC^2 = 0 \qquad (0 \leq z \leq 1) \tag{7.36}$$

$$\frac{dC}{dz} = Pe(C-1) \qquad at \quad z = 0; \tag{7.37}$$

$$\frac{dC}{dz} = 0 \qquad at \quad z = 1; \tag{7.38}$$

Using method of orthogonal collocation with $n = 3$ and defining

$$\mathbf{C} = \begin{bmatrix} C_0 & C_1 & ... & C_4 \end{bmatrix}^T \tag{7.39}$$

we get following set of five nonlinear algebraic equations

$$\frac{1}{Pe}\left[\left[\mathbf{t}^{(i)}\right]^T \mathbf{C}\right] - \left[\left(\mathbf{s}^{(i)}\right)^T \mathbf{C}\right] - DaC_i^2 = 0 \tag{7.40}$$

$$i = 1, 2, 3$$

$$\left[\left[\mathbf{t}^{(0)}\right]^T \mathbf{C}\right] - Pe(C_0 - 1) = 0 \tag{7.41}$$

$$\left[\left[\mathbf{t}^{(4)}\right]^T \mathbf{C}\right] = 0 \tag{7.42}$$

where the matrices

$$S = \begin{bmatrix} -13 & 14.79 & -2.67 & 1.88 & -1 \\ -5.32 & 3.87 & 2.07 & -1.29 & 0.68 \\ 1.5 & -3.23 & 0 & 3.23 & -1.5 \\ -0.68 & 1.29 & -2.07 & -3.87 & 5.32 \\ 1 & -1.88 & 2.67 & -14.79 & 13 \end{bmatrix} \tag{7.43}$$

$$T = \begin{bmatrix} 84 & -122.06 & 58.67 & -44.60 & 24 \\ 53.24 & -73.33 & 26.27 & -13.33 & 6.67 \\ -6 & 16.67 & -21.33 & 16.67 & -6 \\ 6.76 & -13.33 & 26.67 & -73.33 & 53.24 \\ 24 & -44.60 & 58.67 & -122.06 & 84 \end{bmatrix} \tag{7.44}$$

Example 49. [**6**] *Consider the 2-dimensional Laplace equation*

$$\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = f(x,y) \tag{7.45}$$

$$0 < x < 1 \ ; \ 0 < y < 1$$

where $u(x,y)$ represents the dimensionless temperature distribution in a furnace and x, y are space coordinates. The boundary conditions are as follows:

$$x = 0 : u = u^*; \ \ x = 1 : u = u^* \tag{7.46}$$

$$y = 0 : u = u^*; \ \ y = 1 : k(\partial u/\partial y) = h(u_\infty - u(x,1)) \tag{7.47}$$

Using n_x internal grid lines parallel to y axis and n_y grid lines parallel to y-asix, we get $n_x \times n_y$ internal collocation points. Corresponding to the chosen collocation points, we can compute matrices (S_x, T_x) and (S_y, T_y) using equations (7.23) and (7.27). Using these matrices, the PDE can be transformed as

$$\left[\mathbf{t}_x^{(i)}\right]^T U^{(i)} + \left[\mathbf{t}_y^{(j)}\right]^T U^{(j)} = f(x_i, y_j) \tag{7.48}$$

$$i = 1, 2, ...n_x \ ; \ \ j = 1, 2, ...n_y$$

$$U^{(i)} = \left[\begin{array}{cccc} u_{0,i} & u_{1,i} & ... & u_{n_{x+1},i} \end{array} \right] \tag{7.49}$$

$$U^{(j)} = \left[\begin{array}{cccc} u_{j,o} & u_{j,1} & ... & u_{j,n_{y+1}} \end{array} \right] \tag{7.50}$$

At the boundaries, we have

$$u_{0,j} = u^* \ ; \ (j = 0, 1, ...n_{x+1}) \tag{7.51}$$

$$u_{1,j} = u^* \ ; \ (j = 0, 1, ...n_{x+1}) \tag{7.52}$$

$$u_{i,0} = u^* \ ; \ (i = 0, 1, ...n_{y+1}) \tag{7.53}$$

$$k\left[\mathbf{s}_y^{(n_y+1)}\right]^T U^{(i)} = h(u_\infty - u(x_i, 1)) \tag{7.54}$$

$$(i = 1, ...n_y)$$

The above procedure yields $(n_x + 1) \times (n_y + 1)$ *nonlinear equations in* $(n_x + 1) \times (n_y + 1)$ *unknowns, which can be solved simultaneously using Newton-Raphson method.*

Remark 3. *Are the two methods presented above, i.e. finite difference and collocation methods, doing something fundamentally different? Suppose we choose* $n'th$ *order polynomial (4.7), we are essentially approximating the true solution vector* $y^*(z) \in \mathbf{C^{(2)}[0,1]}$ *by another vector (i.e. the polynomial function) in* $(n+2)$ *dimensional subspace of* $\mathbf{C^{(2)}[0,1]}$. *If we choose n internal grid points by finite difference approach, then we are essentially finding a vector* $\mathbf{y}$ *in* R^{n+2} *that approximates* $y^*(z)$. *In fact, if we compare the Approach 2 presented above and the finite difference method, the similarities are more apparent as the underlying* $(n+2)$ *dimensional subspace used in approximations become identical. Let us compare the following two cases (a) finite difference method with 3 internal grid points (b) collocation with 3 internal grid points on the basis of expressions used for approximating the first and second order derivatives computed at one of the grid points. For the sake of comparison, we have taken equi-spaced grid points for collocation method instead of taking them at the roots of 3'rd order orthogonal polynomial. Thus, for both collocation and finite difference method, the grid (or collocation) points are at* $\{z_0 = 0, z_1 = 1/4, z_2 = 1/2, z_3 = 3/4, z_4 = 1\}$ *and we want to estimate the approximate solution vector* $\mathbf{y} = \begin{bmatrix} y_0 & y_1 & y_2 & y_3 & y_4 \end{bmatrix}$ *in both the cases. Let us compare expressions for derivative at* $z = z_2$ *used in both the approaches.*

Finite Difference

$$(dy/dz)_2 = \frac{(y_3 - y_1)}{2(\Delta z)} = 2y_3 - 2y_1 \quad ; \quad \Delta z = 1/4 \tag{7.55}$$

$$(d^2y/dz^2)_2 = \frac{(y_3 - 2y_2 + y_1)}{(\Delta z)^2} = 16y_3 - 32y_2 + 16y_1 \tag{7.56}$$

Collocation

$$
\begin{aligned}
&(7.57) \qquad & (dy/dz)_2 &= 0.33y_0 - 2.67y_1 + 2.67y_3 - 0.33y_4 \\
&(7.58) \qquad & (d^2y/dz^2)_2 &= -1.33y_0 + 21.33y_1 - 40y_2 + 21.33y_3 - 1.33y_4
\end{aligned}
$$

It becomes clear from the above expressions that the essential difference between the two approaches is the way the derivatives at any grid (or collocation) point is approximated. The finite difference method takes only immediate neighboring points for approximating the derivatives while the collocation method finds derivatives as weighted sum of all the collocation (grid) points. As a consequence, the approximate solutions generated by these approaches will be different.

8. Summary

In these lecture notes, we have developed methods for efficiently solving large dimensional linear algebraic equations. To begin with, we introduce induced matrix norms and use them to understand matrix ill conditioning and susceptibility of matrices to round of errors. The direct methods for dealing with sparse matrices are discussed next. Iterative solution schemes and their convergence characteristics are discussed in the subsequent section. We then present techniques for solving nonlinear algebraic equations, which are based on successive solutions of linear algebraic sub-problem. In the last section, we have discussed the orthogonal collocation technique, which converts ODE-BVPs or certain class of PDEs into a set of nonlinear algebraic equations, which can be solved using Newton-Raphson method.

9. Appendix

9.1. Proof of Theorem 2 [4]: For Jacobi method,

$$
\begin{aligned}
&(9.1) \qquad & S^{-1}T &= -D^{-1}\left[L+U\right] \\
&(9.2) \qquad & &= \begin{bmatrix} 0 & -\dfrac{a_{12}}{a_{11}} & ..... & -\dfrac{a_{1n}}{a_{11}} \\ -\dfrac{a_{12}}{a_{22}} & 0 & ..... & .... \\ ..... & ..... & .... & -\dfrac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\dfrac{a_{12}}{a_{nn}} & ..... & ..... & 0 \end{bmatrix}
\end{aligned}
$$

As matrix A is diagonally dominant, we have

$$\sum_{j=1(j\neq i)}^{n} |a_{ij}| \quad < \quad |a_{ii}| \quad \text{for } \; i=1,2.,...n \tag{9.3}$$

$$\Rightarrow \quad \sum_{j=1(j\neq i)}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1 \quad \text{for } \; i=1,2.,...n \tag{9.4}$$

$$\left\|S^{-1}T\right\|_{\infty} \quad = \quad \max_{i} \left[\sum_{j=1(j\neq i)}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|\right] < 1 \tag{9.5}$$

Thus, Jacobi iteration converges if A is diagonally dominant.

For Gauss Seidel iterations, the iteration equation for i'th component of the vector is given as

$$x_i^{(k+1)} = \left(\frac{1}{a_{ii}}\right)\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right] \tag{9.6}$$

Let $\mathbf{x}^*$ denote the true solution of $A\mathbf{x} = \mathbf{b}$. Then, we have

$$x_i^* = \left(\frac{1}{a_{ii}}\right)\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^* - \sum_{j=i+1}^{n} a_{ij}x_j^*\right] \tag{9.7}$$

Subtracting (9.7) from (9.6), we have

$$x_i^{(k+1)} - x_i^* = \left(\frac{1}{a_{ii}}\right)\left[\sum_{j=1}^{i-1} a_{ij}\left(x_j^* - x_j^{(k+1)}\right) + \sum_{j=i+1}^{n} a_{ij}\left(x_j^* - x_j^{(k)}\right)\right] \tag{9.8}$$

or

$$\left|x_i^{(k+1)} - x_i^*\right| \leq \left[\sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \left|\left(x_j^* - x_j^{(k+1)}\right)\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \left|\left(x_j^* - x_j^{(k)}\right)\right|\right] \tag{9.9}$$

Since

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_{\infty} = \max_{j} \left|\left(x_j^* - x_j^{(k)}\right)\right|$$

we can write

$$\left|x_i^{(k+1)} - x_i^*\right| \leq p_i \left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_{\infty} + q_i \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_{\infty} \tag{9.10}$$

where

$$p_i = \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \quad ; \quad q_i = \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \tag{9.11}$$

Let s be value of index i for which

$$\left|x_s^{(k+1)} - x_s^*\right| = \max_{j} \left|\left(x_j^* - x_j^{(k+1)}\right)\right| \tag{9.12}$$

Then, assuming $i = s$ in inequality (9.10), we get

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_\infty \leq p_i \left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_\infty + q_i \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \tag{9.13}$$

or

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_\infty \leq \frac{q_s}{1 - p_s} \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \tag{9.14}$$

Let

$$\mu = \begin{matrix} \max \\ j \end{matrix} \ \frac{q_j}{1 - p_j} \tag{9.15}$$

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_\infty \leq \frac{q_s}{1 - p_s} \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \tag{9.16}$$

then it follows that

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|_\infty \leq \mu \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \tag{9.17}$$

Now, as matrix A is diagonally dominant, we have

$$0 < p_i < 1 \text{ and } 0 < q_i < 1 \tag{9.18}$$

$$0 < p_i + q_i = \sum_{j=1 (j \neq i)}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \tag{9.19}$$

Let

$$\beta = \begin{matrix} \max \\ i \end{matrix} \left[\sum_{j=1 (j \neq i)}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \right] \tag{9.20}$$

Then, we have

$$p_i + q_i \leq \beta < 1 \tag{9.21}$$

It follows that

$$q_i \leq \beta - p_i \tag{9.22}$$

and

$$\mu = \frac{q_i}{1 - p_i} \leq \frac{\beta - p_i}{1 - p_i} \leq \frac{\beta - p_i \beta}{1 - p_i} = \beta < 1 \tag{9.23}$$

Thus, it follows from inequality (9.17) that

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \leq \mu^k \left\|\mathbf{x}^* - \mathbf{x}^{(0)}\right\|_\infty \tag{9.24}$$

i.e. the iteration scheme is a contraction map and

$$\begin{matrix} \lim \\ k \to \infty \end{matrix} \ \mathbf{x}^{(k)} = \mathbf{x}^* \tag{9.25}$$

9.2. Proof of Theorem 3. For Gauss-Seidel method, when matrix A is symmetric, we have

$$S^{-1}T = (L+D)^{-1}(-U) = -(L+D)^{-1}(L^T) \tag{9.26}$$

Now, let $\mathbf{e}$ represent unit eigenvector of matrix $S^{-1}T$ corresponding to eigenvalue λ, i.e.

$$-(L+D)^{-1}(L^T)\mathbf{e} = \lambda\mathbf{e} \tag{9.27}$$

$$\text{or} \qquad L^T\mathbf{e} = -\lambda(L+D)\mathbf{e} \tag{9.28}$$

Taking inner product of both sides with $\mathbf{e}$,we have

$$\left\langle L^T\mathbf{e},\mathbf{e}\right\rangle = -\lambda\left\langle (L+D)\mathbf{e},\mathbf{e}\right\rangle \tag{9.29}$$

$$\lambda = -\frac{\left\langle L^T\mathbf{e},\mathbf{e}\right\rangle}{\langle D\mathbf{e},\mathbf{e}\rangle + \langle L\mathbf{e},\mathbf{e}\rangle} = -\frac{\langle \mathbf{e},L\mathbf{e}\rangle}{\langle D\mathbf{e},\mathbf{e}\rangle + \langle L\mathbf{e},\mathbf{e}\rangle} \tag{9.30}$$

Defining

$$\alpha = \langle L\mathbf{e},\mathbf{e}\rangle = \langle \mathbf{e},L\mathbf{e}\rangle \tag{9.31}$$

$$\sigma = \langle D\mathbf{e},\mathbf{e}\rangle = \sum_{i=1}^{n} a_{ii}\left(e_i\right)^2 > 0 \tag{9.32}$$

we have

$$\lambda = -\frac{\alpha}{\alpha+\sigma} \Rightarrow |\lambda| = \left|\frac{\alpha}{\alpha+\sigma}\right| \tag{9.33}$$

Note that $\sigma > 0$ follows from the fact that trace of matrix A, is positive as eigenvalues of A are positive. Using positive definiteness of matrix A, we have

$$\langle A\mathbf{e},\mathbf{e}\rangle = \langle L\mathbf{e},\mathbf{e}\rangle + \langle D\mathbf{e},\mathbf{e}\rangle + \left\langle L^T\mathbf{e},\mathbf{e}\right\rangle \tag{9.34}$$

$$= \sigma + 2\alpha > 0 \tag{9.35}$$

This implies

$$-\alpha < (\sigma+\alpha) \tag{9.36}$$

Since $\sigma > 0$, we can say that

$$\alpha < (\sigma+\alpha) \tag{9.37}$$

i.e.

$$|\alpha| < (\sigma+\alpha) \tag{9.38}$$

This implies

$$|\lambda| = \left|\frac{\alpha}{\alpha+\sigma}\right| < 1 \tag{9.39}$$

10. Exercise

(1) A polynomial

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

passes through point (3, 2), (4, 3), (5, 4) and (6, 6) in an x-y coordinate system. Setup the system of equations and solve it for coefficients a_0 to a_3 by Gaussian elimination. The matrix in this example is (4 X 4) *Vandermonde matrix.* Larger matrices of this type tend to become ill-conditioned.

(2) Solve using Gauss Jordan method.

$$u + v + w = -2$$

$$3u + 3v - w = 6$$

$$u - v + w = -1$$

to obtain A^{-1}. What coefficient of v in the third equation, in place of present -1 , would make it impossible to proceed and force the elimination to break down?

(3) Decide whether vector **b** belongs to column space spanned by $\mathbf{x}^{(1),}\mathbf{x}^{(2)}, ...$

(a) $\mathbf{x}^{(1)} = (1, 1, 0);\ \mathbf{x}^{(2)} = (2, 2, 1); \mathbf{x}^{(3)} = (0, 2, 0); \mathbf{b} = (3, 4, 5)$

(b) $\mathbf{x}^{(1)} = (1, 2, 0);\ \mathbf{x}^{(2)} = (2, 5, 0); \mathbf{x}^{(3)} = (0, 0, 2); \mathbf{x}^{(4)} = (0, 0, 0);$ any **b**

(4) Find dimension and construct a basis for the four fundamental subspaces associated with each of the matrices.

$$A_1 = \begin{bmatrix} 0 & 1 & 4 & 0 \\ 0 & 2 & 8 & 0 \end{bmatrix} \quad ; \quad U_2 = \begin{bmatrix} 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad ; \quad A_3 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} \quad ; \quad U_1 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}$$

(5) Find a non -zero vector $\mathbf{x}^*$ orthogonal to all rows of

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 6 & 4 \end{bmatrix}$$

(In other words, find Null space of matrix A.) If such a vector exits, can you claim that the matrix is singular? Using above A matrix find one possible solution $\mathbf{x}$ for $A\mathbf{x} = \mathbf{b}$ when $b = \begin{bmatrix} 4 & 9 & 13 \end{bmatrix}^T$. Show that if vector x is a solution of the system $A\mathbf{x} = \mathbf{b}$, then $(\mathbf{x}{+}\alpha\mathbf{x}^*)$ is also a solution for any scalar α, i.e.

$$A(\mathbf{x} + \alpha\mathbf{x}^*) = b$$

Also, find dimensions of row space and column space of A.

(6) If product of two matrices yields null matrix, i.e. $AB = [0]$, show that column space of B is contained in null space of A and the row space of A is in the left null space of B.

(7) Why there is no matrix whose row space and null space both contain the vector

$$\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$$

(8) Find a 1×3 matrix whose null space consists of all vectors in R^3 such that $x_1 + 2x_2 + 4x_3 = 0$. Find a 3×3 matrix with same null space.

(9) If V is a subspace spanned by

$$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} ; \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} ; \begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix}$$

find matrix A that has V as its row space and matrix B that has V as its null space.

(10) Find basis for each of subspaces and rank of matrix A

(a)

$$A = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 4 & 6 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b)

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 3 & 2 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 & 2 & 1 \\ 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(11) Consider application of finite difference method to solving ODE-BVP with non-equidistant grid point i.e.

$$\Delta z_i = z_{i+1} - z_i \ ; i = 0, 1, 2, ...$$

Derive expressions for the first and second derivatives

$$y_i^{(2)} = \frac{2}{\Delta z_i + \Delta z_{i-1}} \left[\frac{y_{i+1-}y_i}{\Delta z_i} - \frac{y_{i-}y_i - 1}{\Delta z_{i-1}} \right] - \frac{1}{3} y_i^{(3)} \left(\Delta z_i - \Delta z_{i-1} \right) + ...$$

(12) In several fluid mechanics problems, one needs to solve for $\nabla^4\Psi = 0$ where Ψ is the stream function. Solving these equations numerically requires computation of higher derivatives. Derive following expressions for

$$\frac{d^3y}{dz^3} = \frac{y_{i+2} - 2y_{i+1} + y_i - 2y_{i-1} + y_{i-2}}{2\left(\Delta z\right)^3} + O\left(\Delta z\right)^2$$

$$\frac{d^4y}{dz^4} = \frac{y_{i+2} - 4y_{i+1} + 6y_i - 4y_{i-1} + y_{i-2}}{2\left(\Delta z\right)^4} + O\left(\Delta z\right)^2$$

(13) Consider the solution of the linear system

$$\begin{aligned} x + 0.9y + 0.1z &= 1 \\ 0.4x + y + 0.4z &= 0 \\ 0.8x + 0.1y + z &= 0 \end{aligned}$$

by iteration scheme

$$\begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \\ z^{(k+1)} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0.9 & 0.1 \\ 0.4 & 0 & 0.4 \\ 0.8 & 0.1 & 0 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ y^{(k)} \\ z^{(k)} \end{bmatrix}$$

Show that this iteration scheme converges to the unique solution for arbitrary starting guess.

(14) The true solution $Ax = b$ is slightly different from the elimination solution to $LUx_0 = b$; $A - LU$ misses zero matrix because of round off errors. One possibility is to do everything in double precision, but a better and faster way is iterative refinement: Compute only one vector $r = b - Ax_0$ in double precision, solve $LUy = r$,and add correction y to x_0 to generate an improved solution $x_1 = x_0 + y$.

Problem:Multiply $x_1 = x_0 + y$, by LU, write the result as splitting $Sx_1 = Tx_0 + b$, and explain why T is extremely small. This single step bring almost close to **x**.

(15) Consider system

$$A = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \quad ; \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

where A is Hilbert matrix with $a_{ij} = 1/(i+j-1)$, which is severely ill- conditioned. Solve using

(a) Gauss-Jordan elimination

(b) exact computations

(c) rounding off each number to 3 figures.

Perform 4 iterations each by

(a) Jacobi method

(b) Gauss- Seidel method

(c) Successive over-relaxation method with $\omega = 1.5$

Use initial guess $\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$ and compare in each case how close to the $\mathbf{x}^{(4)}$ is to the exact solution. (Use 2-norm for comparison).

Analyze the convergence properties of the above three iterative processes using eigenvalues of the matrix $(S^{-1}T)$ in each case. Which iteration will converge to the true solution?

(16) The Jacobi iteration for a general 2 by 2 matrix has

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad ; \quad D = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$$

If A is symmetric positive definite, find the eigenvalues of $J = S^{-1}T = D^{-1}(D - A)$ and show

that Jacobi iterations converge.

(17) It is desired to solve $Ax = b$ using Jacobi and Gauss-Seidel iteration scheme where

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 5 & 3 \\ 2 & 4 & 7 \end{bmatrix} \quad ; \quad A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix} \quad ; \quad A = \begin{bmatrix} -7 & 1 & -2 & 3 \\ 1 & 8 & 1 & 3 \\ -2 & 1 & -5 & 1 \\ 1 & 0 & -1 & -3 \end{bmatrix}$$

Will the Jacobi and Gauss-Seidel the iterations converge? Justify your answer. (Hint: Check for diagonal dominance before proceeding to compute eigenvalues).

(18) Given matrix

$$J = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

find powers J^2, J^3 by direct multiplications. For which matrix A is this a Jacobi matrix $J = I - D^{-1}A$? Find eigenvalues of J.

(19) The tridiagonal $n \times n$ matrix A that appears when finite difference method is used to solve second order PDE / ODE-BVP and the corresponding Jacobi matrix are as follows

$$A = \begin{bmatrix} 2 & -1 & 0 & ... & 0 \\ -1 & 2 & -1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & -1 & 2 & -1 \\ 0 & ... & 0 & 1 & 2 \end{bmatrix} \quad ; \quad J = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 1 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & 1 & 0 & 1 \\ 0 & ... & 0 & 1 & 0 \end{bmatrix}$$

Show that the vector

$$x = \begin{bmatrix} sin(\pi h) & sin(2\pi h) & ... & sin(nh) \end{bmatrix}^T$$

satisfies $J\mathbf{x} = \lambda \mathbf{x}$ with eigenvalue $\lambda = cos(\pi h)$. Here, $h = 1/(n+1)$ and hence $sin\left[(n+1)\pi h\right] = 0$.

Note: The other eigenvectors replace π by 2π, 3π,...., nπ. The other eigenvalues

are $cos(2\pi h)$, $cos(3\pi h)$,...... , $cos(n\pi h)$ all smaller than $cos(\pi h) < 1$.

(20) Consider the following system

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon \end{bmatrix}$$

Obtain A^{-1}, $\det(A)$ and also solve for $[x_1 \; x_2 \;]^T$. Obtain numerical values for $\varepsilon = 0.01$, 0.001 and 0.0001. See how sensitive is the solution to change in ε.

(21) If A is orthogonal matrix, show that $||A|| = 1$ and also $c(A) = 1$. Orthogonal matrices and their multiples(αA) are the only perfectly conditioned matrices.

(22) For the positive definite matrix, $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$, compute $||A^{-1}|| = 1/\lambda_1$; and $c(A) = \lambda_2/\lambda_{1.}$Find the right side b and a perturbation vector δb such that the error is worst possible, i.e. find vector x such that

$$||\ \delta x||/||x|| = c||\ \delta_b||/||b||$$

(23) Find a vector x orthogonal to row space, and a vector y orthogonal to column space, of

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 6 & 4 \end{bmatrix}$$

(24) Show that vector $\mathbf{x}-\mathbf{y}$ is orthogonal to vector $\mathbf{x}+\mathbf{y}$ if and only if $||\mathbf{x}|| = ||\mathbf{y}||$.

(25) For a positive definite matrix A, the Cholensky decomposition is $A = L\ D\ L^T = RR^T$ where $R = LD^{1/2}$. Show that the condition number of R is square root of condition number of A. It follows that Gaussian elimination needs no row exchanges for a positive definite matrix; the condition number does not deteriorate, since $c(A) = c(R^T)c(R)$.

(26) Show that for a positive definite symmetric matrix, the condition number can be obtained as

$$c(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$$

(27) If A is an orthonormal(unitary) matrix (i.e. $A^T A = I$), show that $||A|| = 1$ and also $c(A) = 1$. Orthogonal matrices and their multiples $(\alpha A)$are he only perfectly conditioned matrices.

(28) Show that $\lambda_{\max}$ (i.e. maximum magnitude eigen value of a matrix) or even $\max|\lambda_i|$,is not a satisfactory norm of a matrix, by finding a 2x2 counter example to

$$\lambda_{\max}(A+B) \leq \ \lambda_{\max}(A) + \ \lambda_{\max}(B)$$

and to

$$\lambda_{\max}(AB) \leq \ \lambda_{\max}(A)\ \lambda_{\max}(B)$$

(29) Prove the following inequalities/ identities

$$\|A+B\| \leq \|A\| + \|B\|$$

$$\|AB\| \leq \|A\|\ \|B\|$$

$$C(AB) \leq C(A)C(B)$$

$$\|A\|_2 = \left\|A^T\right\|_2$$

(30) If A is an orthonormal (unitary) matrix, show that $\|A\|_2 = 1$ and also condition number $C(A) = 1$. Note that orthogonal matrices and their multiples are the only perfectly conditioned matrices.

(31) Show that for a positive definite symmetric matrix, the condition number can be obtained as $C(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

(32) Show that $\lambda_{\max}$, or even $max\ |\lambda_i|$, is not a satisfactory norm of a matrix, by finding a 2×2 counter examples to following inequalities

$$\lambda_{\max}(A+B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$$

$$\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B)$$

(33) For the positive definite matrix

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

compute the condition number $C(A)$ and find the right hand side $\mathbf{b}$ of equation $A\mathbf{x} = \mathbf{b}$ and perturbation $\delta\mathbf{b}$ such that the error is worst possible, i.e.

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = C(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

(34) Consider the following two sets of nonlinear equations

(a) Set 1:

$$\begin{aligned} x_1^2 + x_2^2 - 4 &= 0 \\ x_1^2 - x_2^2 - 1.5 &= 0 \\ \mathbf{x}^{(0)} &= \begin{bmatrix} 1 & 1 \end{bmatrix}^T \end{aligned}$$

(b) Set 2 (Banana Function):

$$\begin{aligned} 10(-x_1^2 + x_2) &= 0 \\ x_1 - 1 &= 0 \\ \mathbf{x}^{(0)} &= \begin{bmatrix} -1 & 1 \end{bmatrix}^T \end{aligned}$$

(c) Set 3:

$$\begin{aligned} 2x &= \sin\left[(x+y)/2\right] \\ 2y &= \cos\left[(x-y)/2\right] \\ \mathbf{x}^{(0)} &= \begin{bmatrix} 10 & -10 \end{bmatrix}^T \end{aligned}$$

(d) Set 4:

$$\begin{aligned} \sin(x) + y^2 + \ln(z) &= 7 \\ 3x + 2^y - z^3 &= -1 \\ x + y + z &= 5 \\ \mathbf{x}^{(0)} &= \begin{bmatrix} 0 & 2 & 2 \end{bmatrix}^T \end{aligned}$$

(a) Perform 3 iterations using (a) successive substitution (b) Newton Raphson (c) Newton Raphson with Broyden's update and compare progress of $\mathbf{x}^{(k)}$ towards the true solution in each case.

(b) Check whether the sufficient condition for convergence is satisfied at each iteration by computing infinite norms of the Jacobians.

(35) Consider following set of nonlinear algebraic equations Perform 3 iterations using

(a) Set 2 (Banana Function):

$$\begin{aligned} 10(-x_1^2 + x_2) &= 0 \\ x_1 - 1 &= 0 \end{aligned}$$

$$\mathbf{x}^{(0)} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$$

(36) The following coupled differential equations characterize a system

$$\begin{aligned} \frac{d^2u}{dz^2} + 2ue^v &= f_1(u,v,z) \quad ; \quad u(0) = 0; \;\; u(1) = 5 \\ \frac{d^2v}{dz^2} + 5uv &= f_2(u,v,z) \quad ; \quad v(0) = 1; \;\; v(1) = 2 \end{aligned}$$

Obtain sets of nonlinear algebraic equations using (a) finite difference method with 2 internal grid points and (b) orthogonal collocation with 2 collocation points. Also, arrange these equations as $A\mathbf{x} = G(\mathbf{x})$ so that method of successive substitution can be used for formulating iterative scheme.

CHAPTER 4

ODE-IVPs and Related Numerical Schemes

1. Motivation

In these lecture notes, we undertake the study of solution techniques for multivariable and coupled ODE-IVPs. The numerical techniques for solving ODE-IVPs form basis for a number of numerical schemes and are used for solving variety of problems such as

- Dynamic simulation of lumped parameter systems
- Solution of ODE-BVP
- Solving Parabolic / Hyperbolic PDEs
- Solving simultaneous nonlinear algebraic equation

and so on. In order to provide motivation for studying the numerical methods for solving ODE-IVPs, we first formulate numerical schemes for the above problems in the following subsections.

1.1. Dynamic behavior of lumped parameter systems.

Example 50. *Three isothermal CSTRs in series*

Consider three isothermal CSTRs in series in which a first order liquid phase reaction of the form

$$A \longrightarrow B \tag{1.1}$$

is carried out. It is assumed that volume and liquid density remains constant in each tank and

$$V_1 \frac{dC_{A1}}{dt} = F(C_{A0} - C_{A1}) - kV_1 C_{A1} \tag{1.2}$$

$$V_2 \frac{dC_{A2}}{dt} = F(C_{A1} - C_{A2}) - kV_2 C_{A2} \tag{1.3}$$

$$V_3 \frac{dC_{A3}}{dt} = F(C_{A2} - C_{A3}) - kV_3 C_{A3} \tag{1.4}$$

Defining $\tau = F/V$, we can re arrange the above set of equations as

$$\begin{bmatrix} \frac{dC_{A1}}{dt} \\ \frac{dC_{A2}}{dt} \\ \frac{dC_{A3}}{dt} \end{bmatrix} = \begin{bmatrix} -(k+1/\tau_1) & 0 & 0 \\ 1/\tau & -(k+1/\tau_2) & 0 \\ 0 & 1/\tau & -(k+1/\tau_3) \end{bmatrix} \begin{bmatrix} C_{A1} \\ C_{A2} \\ C_{A3} \end{bmatrix} \tag{1.5}$$

$$+ \begin{bmatrix} 1/\tau_1 \\ 0 \\ 0 \end{bmatrix} C_{A0} \tag{1.6}$$

$$\begin{gathered} \mathbf{x} = [C_{A1}, C_{A2}, C_{A3}]^T \\ \frac{d\mathbf{x}}{dt} = A\mathbf{x} + BC_{A0} \end{gathered} \tag{1.7}$$

where matrices A and B are defined in the above equation. Now, suppose initially $C_{A0} = \bar{C}_{A0}$, till $t = 0$, and, for $t \geq 0$, C_{A0} was changed to $C_{A0} = 0$. Then we are required to solve

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}; \qquad \mathbf{x} = \mathbf{x}(0) \quad at \quad t = 0 \tag{1.8}$$

and generate trajectories $\mathbf{x}(t)$ (i.e. $C_{A1}(t), C_{A2}(t)$ and $C_{A3}(t)$) over interval $[0, t_f]$. This is a typical problem of dynamic simulation of lumped parameter system.

Example 51. *Continuous Fermenter*

Consider a continuously operated fermenter described by the following set of ODEs

$$\frac{dX}{dt} = F_1(X, S, P,\ D,\ S_f) = -DX + \mu X \tag{1.9}$$

$$\frac{dS}{dt} = F_2(X, S, P,\ D,\ S_f) = D(S_f - S) - \frac{1}{Y_{X/S}}\mu X \tag{1.10}$$

$$\frac{dP}{dt} = F_3(X, S, P,\ D,\ S_f) = -DP + (\alpha\mu + \beta)X \tag{1.11}$$

where X represents effluent cell-mass or biomass concentration, S represents substrate concentration and P denotes product concentration. It is assumed that product concentration (S) and the cell-mass concentration (X) are measured process outputs while dilution rate (D) and the feed substrate concentration (S_f) are process inputs which can be manipulated. Model parameter μ represents the specific growth rate, $Y_{X/S}$ represents the cell-mass yield, α and

β are the yield parameters for the product. The specific growth rate model is allowed to exhibit both substrate and product inhibition:

$$\mu = \frac{\mu_m(1 - \dfrac{P}{P_m})S}{K_m + S + \dfrac{S^2}{K_i}} \tag{1.12}$$

where μ_m represents maximum specific growth rate, P_m represents product saturation constant, K_m substrate saturation constant and the K_i represents substrate inhibition constant. Defining state and input vectors as

$$\mathbf{x} = \begin{bmatrix} X & S & P \end{bmatrix}^T \quad ; \ \mathbf{u} = \begin{bmatrix} D & S_f \end{bmatrix}^T \tag{1.13}$$

the above equation can be represented as

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}, \mathbf{u}) \tag{1.14}$$

A typical problem dynamic simulation problem is to find trajectories of product, biomass and substrate concentrations over an interval $[0, t_f]$, given their initial values and dilution rate $D(t)$ and feed substrate concentration S_f as a function of time over $[0, t_f]$.

In abstract terms, the dynamic simulation problem can be states as follows. Given time trajectories of independent variables $\{\mathbf{u}(t) : 0 \leq t \leq t_f\}$,and initial state, $\mathbf{x}(0)$, of the system, obtain state trajectories $\{\mathbf{x}(t) : 0 \leq t \leq t_f\}$ by integrating

$$\frac{d\mathbf{x}}{dt} = F\left[\mathbf{x}(t), \mathbf{u}(t)\right] \quad ; \ \mathbf{x} = \ \mathbf{x}(0) \ \text{ at } t = 0 \tag{1.15}$$

where $\mathbf{x} \in R^n$ represents dependent or state variables and $\mathbf{u} \in R^m$denote independent inputs. As the independent variable trajectories are known *a-priori* while solving ODE-IVP, the problem can be looked at as n-ODE's in n variables with variable coefficients. Thus, the above problem can be re-stated as

$$\frac{d\mathbf{x}}{dt} = F_u(\mathbf{x}, t) \ ; \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{1.16}$$

In other words, a forced dynamic systems can be looked upon as unforced systems with variable parameters.

1.2. Solving nonlinear algebraic equations using method of homotopy. Consider the problem of solving simultaneous nonlinear algebraic equation of the form

$$F(\mathbf{x}) = \overline{0} \tag{1.17}$$

where $\mathbf{x} \in R^m$ and $F(\mathbf{x})$ is a $m \times 1$ function vector. Introducing a parameter λ such that$(0 \leq \lambda \leq 1)$, we define

$$F(\mathbf{x}(\lambda)) = (1-\lambda)F(\mathbf{x}^{(0)}) \tag{1.18}$$

where $\mathbf{x}^{(0)}$ represents some arbitrary initial guess. Obviously, at $\lambda = 1$ we have $F(\mathbf{x}) = \overline{0}$. Differentiating w.r.t. λ and rearranging, we get

$$\frac{d\mathbf{x}}{d\lambda} = -[\frac{\partial F}{\partial \mathbf{x}}]^{-1}F(\mathbf{x}^{(0)}) \tag{1.19}$$

$$\mathbf{x}(0) = \mathbf{x}^{(0)} \tag{1.20}$$

Integrating above ODE-IVPs from $\lambda = 0$ to $\lambda = 1$ produces the solution of nonlinear the equations at $\lambda = 1$.

1.3. Solutions of Parabolic / Hyperbolic PDE's.

1.3.1. *Finite Difference Method.* Use of finite difference method to solve parabolic or hyperbolic equations with finite spatial boundaries results in set of coupled linear / nonlinear ODE-IVPs.

-

Example 52. *Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out.*

$$\frac{\partial C}{\partial t} = \frac{1}{Pe}\frac{\partial^2 C}{\partial z^2} - \frac{\partial C}{\partial z} - DaC^2 \qquad (0 \leq z \leq 1) \tag{1.21}$$

$$t = 0 \ : \ c(z,0) = 0 \tag{1.22}$$

$$\frac{\partial C(0,t)}{\partial z} = Pe\,(C(0,t) - 1) \qquad at \quad z = 0; \tag{1.23}$$

$$\frac{\partial C(1,t)}{\partial z} = 0 \qquad at \quad z = 1; \tag{1.24}$$

Using finite difference method along the spatial coordinate z with $m-1$ internal grid points, we have

$$\begin{aligned} \frac{dC_i(t)}{dt} &= \frac{1}{Pe}\left(\frac{C_{i+1}(t) - 2C_i(t) + C_{i-1}(t)}{(\Delta z)^2}\right) \\ &\quad - \left(\frac{C_{i+1}(t) - C_{i-1}(t)}{2\,(\Delta z)}\right) - Da\,[C_i(t)]^2 \\ i &= 1, 2, ...m-1 \end{aligned} \tag{1.25}$$

$$\frac{C_1(t) - C_0(t)}{\Delta z} = Pe\,(C_0(t) - 1) \tag{1.26}$$

$$\frac{C_{m+1}(t) - C_m(t)}{\Delta z} = 0 \tag{1.27}$$

The above set of ODEs, together with initial conditions,

$$C_1(0) = C_2(0) = ... = C_{m+1}(0) = 0 \tag{1.28}$$

defines an ODE-IVP of type (1.16).

Example 53. *Consider the 2-dimensional unsteady state heat transfer problem*

$$\frac{\partial T}{\partial t} = \alpha[\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}] \tag{1.29}$$

$$t = 0 : T = F(x, y) \tag{1.30}$$

$$x = 0 : T(0, y, t) = T^*; \;\; x = 1 : T(1, y, t) = T^* \tag{1.31}$$

$$y = 0 \quad : T(x, 0, t) = T^* \, ; \; y = 1 \;\; : \; k\frac{dT(x, 1, t)}{dy} = h(T_\infty - T(x, 1, t)) \tag{1.32}$$

Example 54. *where $T(x, y, t)$ is the temperature at locations (x, y) at time t and α is the thermal diffusivity. By finite difference approach, we construct a 2-dimensional grid with $n_x + 1$ equispaced grid lines parallel to the y-axis and $n_y + 1$ grid lines parallel to the x-axis. The temperature T at the (i, j)'th grid point is given by*

$$T_{ij}(t) = T(x_i, y_i, t) \tag{1.33}$$

Now, we force the residual to zero at each internal grid point to generate a set of coupled ODE-IVP's as

$$\frac{dT_{ij}}{dt} = \frac{\alpha}{(\Delta x)^2}[T_{i+1,j} - 2T_{i,j} + T_{i-1,j}] + \frac{\alpha}{(\Delta y)^2}[T_{i,j+1} - 2T_{i,j} + T_{i,j-1}] \tag{1.34}$$

$$i = 1, 2, ...n_x - 1 \qquad \text{and} \qquad j = 1, 2, ...n_y - 1$$

The values of T_{ij} at boundary points corresponding to $x = 0, x = 1,$ and $y = 0$ can be easily written using boundary conditions.

$$k\frac{dT_{i,n_y}(t)}{dy} = h(T_\infty - T_{i,n_y}(t)) \tag{1.35}$$

$$i = 1, 2, ...n_x - 1$$

1.3.2. *Orthogonal Collocation Method.*

Example 55. *Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) given in the above section. Using method of orthogonal collocation with m internal collocation points, we get*

$$\frac{dC_i(t)}{dt} = \frac{1}{Pe}\left[\left[\mathbf{t}^{(i)}\right]^T \mathbf{C}\right] - \left[\left(\mathbf{s}^{(i)}\right)^T \mathbf{C}\right] - DaC_i^2$$

$$i = 1, 2, 3, ...m$$

$$\begin{aligned} \left[\left[\mathbf{t}^{(0)}\right]^T \mathbf{C}\right] &= Pe(C_0(t) - 1) \\ \left[\left[\mathbf{t}^{(m+1)}\right]^T \mathbf{C}\right] &= 0 \end{aligned}$$

where the matrices $\left[\mathbf{t}^{(i)}\right]$ and $\left(\mathbf{s}^{(i)}\right)$ represent row vectors of matrices T and S, as defined in the Section 5 (example 6) of lecture notes on linear algebraic equations and related numerical scheme.. Here, $C_i(t)$ represents concentration at the i'th collocation point. The above set of ODEs, together with initial conditions,

$$C_1(0) = C_2(0) = ... = C_{m+1}(0) = 0$$

defines an ODE-IVP of type (1.16).

1.4. Solution of ODE-BVP: Shooting Method. By this approach, we reduce the 2nd or higher order ODE-BVP to a set of first order ODE's.

Example 56. *the ODE-BVP describing tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out is given as*

$$\frac{1}{Pe}\frac{d^2C}{dz^2} - \frac{dC}{dz} - DaC^2 = 0 \qquad (0 \leq z \leq 1) \tag{1.36}$$

$$\begin{aligned} \frac{dC}{dz} &= Pe(C - 1) \qquad at \quad z = 0; \\ \frac{dC}{dz} &= 0 \qquad at \quad z = 1; \end{aligned} \tag{1.37}$$

where C is the dimensionless concentration, z is axial position, Pe is the Peclet number for mass transfer and Da is the Damkohler number. Now, defining new state variables

$$x_1 = C \qquad and \qquad x_2 = \frac{dC}{dz} \tag{1.38}$$

we can transform the above ODE's as

$$\frac{dx_1}{dz} = x_2 \tag{1.39}$$

$$\frac{dx_2}{dz} = (Pe)x_2 + (Da.Pe)x_1^2 \tag{1.40}$$

$$x_2 = Pe(x_1 - 1) \qquad at \qquad z = 0 \tag{1.41}$$

$$x_2 = 0 \qquad at \qquad z = 1 \tag{1.42}$$

Shooting method attempts to solve this problem by converting it into and ODE-IVP problem as follows

- Step 1: Assume the 'missing' initial condition $x_1(0)$ at $z = 0$.
- Step 2: Integrate (shoot) the ODEs from $z = 0$ to $z = 1$ as if it is an ODE-IVP using any standard numerical integration method for solving ODE-IVP
- Step 3: Check whether all the specified boundary conditions are satisfied at $z = 1$.

  If not, use method such as Newton - Raphson or successive substitution to correct the guess value at $z = 0$ and go to step 1.

  If the BC at $z = 1$ is satisfied, terminate the iterations.

For TRAM problem, this method can be applied as follows:

- Assume $x_1(0) = s \Rightarrow x_2(0) = Pe(s - 1)$
- Integrate the two ODE's simultaneously using any standard integration method from z = 0 to z = 1.
- The check to be used comes from the given B.C.

$$f(s) = x_2(1) = 0 \tag{1.43}$$

- The value of s can be changed from iteration to iteration by the secant method.

$$s^{(k+1)} = s^{(k)} - f[s^{(k)}]\left[\frac{s^{(k)} - s^{(k-1)}}{f[s^{(k)}] - f[s^{(k-1)}]}\right] \tag{1.44}$$

Alternatively, we can use Newton-Raphson method for generating

$$s^{(k+1)} = s^{(k)} - \left[\frac{f[s^{(k)}]}{[df/ds]_{s=s^{(k)}}}\right] \tag{1.45}$$

The derivative $[df/ds]_{s=s^{(k)}}$ can be computed by simultaneously integrating sensitivity equations. Given a set of the first order nonlinear equations

$$\frac{d\mathbf{x}}{dz} = F(\mathbf{x}) \;\; ; \;\; \mathbf{x}(0) = \mathbf{x}_0 \; ; \; \mathbf{x} \in \mathbf{R}^n \tag{1.46}$$

and $F(\mathbf{x})$ is a $n \times 1$ vector, the associated sensitivity equations are defined as

$$\frac{d\Phi(z)}{dz} = \left[\frac{\partial F}{\partial \mathbf{x}}\right] \Phi(z) \ ; \quad \Phi(0) = \mathbf{I} \tag{1.47}$$

where

$$\Phi(z) = \left[\frac{\partial \mathbf{x}(z)}{\partial \mathbf{x}_0}\right] \tag{1.48}$$

represents the $n \times n$ sensitivity of solution vector $\mathbf{x}(z)$ with respect to the initial conditions and $\mathbf{I}$ denotes identity matrix. In the present case, the sensitivity equations are

$$\frac{d\Phi}{dz} = \begin{bmatrix} 0 & 1 \\ 2D_aP_ex_1 & P_e \end{bmatrix} \Phi(z) \tag{1.49}$$

$$\Phi(z) = \begin{bmatrix} \dfrac{\partial x_1(z)}{\partial x_1(0)} & \dfrac{\partial x_1(z)}{\partial x_2(0)} \\ \dfrac{\partial x_2(z)}{\partial x_1(0)} & \dfrac{\partial x_2(z)}{\partial x_2(0)} \end{bmatrix} \tag{1.50}$$

These equations have to be integrated from $z = 0$ to $z = 1$ to evaluate

$$[df/ds]_{s=s^{(k)}} = \Phi_{21}(1) = \frac{\partial x_2(1)}{\partial x_1(0)} \tag{1.51}$$

Example 57. ***Converting a PDE to an ODE-BVP*** [**6**]***:*** *Consider 2-D steady state heat transfer problem*

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \tag{1.52}$$

$$x = 0 : T = T^*; \ x = 1 : T = T^* \tag{1.53}$$

$$y = 0 : T = T^*; \ y = 1 : k\frac{dT}{dx} = h(T_\infty - T) \tag{1.54}$$

We construct $n_y + 1$ grid lines parallel to the x-axis. The temperature T along the j^{th} grid line is denoted as

$$T_j(x) = T(x, y_j) \tag{1.55}$$

Now, we equate residuals to zero at each internal grid line as

$$\begin{aligned} \frac{d^2T_j}{dx^2} &= -\frac{1}{(\Delta y)^2}[T_{j+1}(x) - 2T_j(x) + T_{j-1}(x)] \\ j &= 2, ...n_y - 1 \end{aligned} \tag{1.56}$$

The boundary conditions at $y = 0$ can be used to eliminate variables on the corresponding edge. At the boundary $y = 1$, using the B.C.s we get

$$k\frac{dT_{n_y}}{dx} = h(T_\infty - T_{n_y}) \tag{1.57}$$

$$j = 1, 2, ...n_y - 1$$

The above set of ODE's can be integrated from $x = 0$ to $x = 1$ with initial condition

$$T_j(0) = T^*, (j = 1, 2, ...n_y.) \tag{1.58}$$

and boundary conditions

$$T_j(1) = T^*, (j = 1, 2, ...n_y.) \tag{1.59}$$

The resulting ODE-BVP can be solved using shooting method.

2. Analytical Solutions of Multivariable Linear ODE-IVP

Consider the problem of solving simultaneous linear ODE-IVP

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}; \qquad \mathbf{x} = \mathbf{x}(0) \quad at \quad t = 0 \tag{2.1}$$

$$\mathbf{x} \in R^m, \quad A \; is \; a \; (m \times m) \; matrix$$

To begin with, we develop solution for the scalar case and generalize it to the multivariable case.

2.1. Scalar Case. Consider the scalar equation

$$\frac{dx}{dt} = ax; \qquad x = x(0) \quad at \quad t = 0 \tag{2.2}$$

Let the guess solution to this IVP be

$$x(t) = e^{\lambda t}v \quad ; \quad v \in R \tag{2.3}$$

Now,

$$x = x(0) \; at \; t = 0 \Rightarrow v = x(0) \tag{2.4}$$

$$\text{or} \quad x(t) = e^{\lambda t}x(0) \tag{2.5}$$

This solution also satisfies the ODE, i.e.

$$\frac{dx}{dt} = \lambda\left[e^{\lambda t}x(0)\right] = \lambda x(t) = ax(t) \tag{2.6}$$

$$\Rightarrow \lambda = a \;\text{ and }\; x(t) = e^{at}x(0) \tag{2.7}$$

Asymptotic behavior of solution can be predicted using the value of parameter a as follows

- Unstable behavior: $a > 0 \Rightarrow x(t) = e^{at}x(0) \rightarrow \infty \; as \; t \rightarrow \infty$
- Stable behavior:$a < 0 \Rightarrow x(t) = e^{at}x(0) \rightarrow 0 \; as \; t \rightarrow \infty$

2.2. Vector case. Now consider system of equations given by equation (2.1). Taking clues from the scalar case, let us investigate a candidate solution of the form

$$\mathbf{x}(t) = e^{\lambda t}\mathbf{v}; \quad \mathbf{v} \in R^m \tag{2.8}$$

where $\mathbf{v}$ is a constant vector. The above candidate solution must satisfy the ODE, i.e.,

$$\begin{aligned} &\frac{d}{dt}(e^{\lambda t}\mathbf{v}) = A(e^{\lambda t}\mathbf{v}) \\ &\Rightarrow \lambda \mathbf{v} e^{\lambda t} = A\mathbf{v}e^{\lambda t} \end{aligned} \tag{2.9}$$

Cancelling $e^{\lambda t}$ from both the sides, as it is a non-zero scalar, we get an equation that vector $\mathbf{v}$ must satisfy,

$$\lambda \mathbf{v} = A\mathbf{v} \tag{2.10}$$

This fundamental equation has two unknowns λ and $\mathbf{v}$ and the resulting problem is the well known eigenvalue problem in linear algebra. The number λ is called the eigenvalue of the matrix A and $\mathbf{v}$ is called the eigenvector. Now, $\lambda \mathbf{v} = A\mathbf{v}$ is a non-linear equation as λ multiplies $\mathbf{v}$. if we discover λ then the equation for $\mathbf{v}$ would be linear. This fundamental equation can be rewritten as

$$(A - \lambda I)\mathbf{v} = 0 \tag{2.11}$$

This implies that vector $\mathbf{v}$ should be $\perp$ to the row space of $(A - \lambda I)$. This is possible only when rows of $(A - \lambda I)$ are linearly dependent. In other words, λ should be selected in such a way that rows of $(A - \lambda I)$ become linearly dependent, i.e., $(A - \lambda I)$ is singular. This implies that λ is an eigenvalue of A if and only if

$$det(A - \lambda I) = 0 \tag{2.12}$$

This is the characteristic equation of A and it has m possible solutions $\lambda_1, ..., \lambda_m$. Thus, corresponding to each eigenvalue λ_i, there is a vector $\mathbf{v}^{(i)}$ that satisfies $(A - \lambda_i I)\mathbf{v}^{(i)} = 0$. This implies that each vector $e^{\lambda_i t}\mathbf{v}^{(i)}$ is a candidate solution to equation (2.1). Now, suppose we construct a vector as lineal combination of these fundamental solutions, i.e.

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v}^{(1)} + c_2 e^{\lambda_2 t}\mathbf{v}^{(2)} + ... + c_m e^{\lambda_m t}\mathbf{v}^{(m)} \tag{2.13}$$

Then, it can be shown that $\mathbf{x}(t)$ also satisfies equation (2.1). Thus, a general solution to the linear ODE-IVP can be constructed as a linear combination of the fundamental solutions $e^{\lambda_i t}\mathbf{v}^{(i)}$.

The next task is to see to it that the above equation reduces to the initial conditions at $t = 0$. Defining vectors C and matrix Ψ as

$$C = \begin{bmatrix} c_1 & c_2 & ... & c_m \end{bmatrix}^T \quad ; \quad \Psi = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & ..... & \mathbf{v}^{(m)} \end{bmatrix} \tag{2.14}$$

we can write

$$\mathbf{x}(0) = \Psi C \tag{2.15}$$

If the eigenvectors are linearly independent,

$$C = \Psi^{-1}\mathbf{x}(0) \tag{2.16}$$

Thus the solution can be written as

$$\begin{aligned} \mathbf{x}(t) &= [e^{\lambda_1 t}\mathbf{v}^{(1)} \quad e^{\lambda_2 t}\mathbf{v}^{(2)}.......e^{\lambda_m t}\mathbf{v}^{(m)}]\Psi^{-1}\mathbf{x}(0) \\ \Rightarrow \mathbf{x}(t) &= [\mathbf{v}^{(1)}\ \mathbf{v}^{(2)}.....\mathbf{v}^{(m)}] \begin{bmatrix} e^{\lambda_1 t} & 0 & ... & 0 \\ 0 & e^{\lambda_2 t} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & e^{\lambda_m t} \end{bmatrix} \Psi^{-1}\mathbf{x}(0) \end{aligned} \tag{2.17}$$

Now let us define the matrix $e\mathbf{x}p(At)$ as follows

$$e^{At} = I + At + \frac{1}{2!}(At)^2 + ... \tag{2.18}$$

Using the fact that matrix A can be diagonalized as

$$A = \Psi\Lambda\Psi^{-1} \tag{2.19}$$

where matrix Λ is

$$\Lambda = diag \begin{bmatrix} \lambda_1 & \lambda_2 & .... & \lambda_m \end{bmatrix}$$

we can write

$$\begin{aligned} e^{At} &= \Psi\Psi^{-1} + \Psi\Lambda\Psi^{-1}t + \tfrac{1}{2!}\Psi\Lambda^2\Psi^{-1}t^2 + ... \\ &= \Psi\Psi^{-1} + \Psi\Lambda\Psi^{-1}t + \tfrac{1}{2!}\Psi\Lambda^2\Psi^{-1}t^2 + ... \\ &= \Psi e^{\Lambda t}\Psi^{-1} \end{aligned} \tag{2.20}$$

Here, the matrix $e^{\Lambda t}$ is limit of infinite sum

$$e^{\Lambda t} = I + t\Lambda + \frac{1}{2!}t^2\Lambda^2 + ...$$
$$= \begin{bmatrix} e^{\lambda_1 t} & 0 & ... & 0 \\ 0 & e^{\lambda_2 t} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & e^{\lambda_m t} \end{bmatrix} \tag{2.21}$$

Thus, equation (2.17) reduces to

$$\mathbf{x}(t) = \Psi e^{\Lambda t}\Psi^{-1}\mathbf{x}(0) \tag{2.22}$$

With this definition, the solution to the ODE-IVP can be written as

$$\mathbf{x}(t) = \Psi e^{\Lambda t}\Psi^{-1}\mathbf{x}(0) = e^{At}\mathbf{x}(0) \tag{2.23}$$

2.3. Asymptotic behavior of solutions. In the case of linear multivariable ODE-IVP problems, it is possible to analyze asymptotic behavior of the solution by observing eigenvalues of matrix A.

$$\begin{gathered} \mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v}^{(1)} + c_2 e^{\lambda_2 t}\mathbf{v}^{(2)} + ... + c_m e^{\lambda_m t}\mathbf{v}^{(m)} \\ C = \Psi^{-1}\mathbf{x}(0) \end{gathered} \tag{2.24}$$

Let $\lambda_j = \alpha_j + i\beta_j$ represent j'th eigenvalue of matrix A. Then, we can write

$$e^{\lambda_j t} = e^{\alpha_j t}.e^{i\beta_j t} = e^{\alpha_j t}[\cos\beta_j t + i\sin\beta_j t] \tag{2.25}$$

As

$$\left|[\cos\beta_j t + i\sin\beta_j t]\right| \leq 1 \text{ for all } t \text{ and all } j \tag{2.26}$$

the asymptotic behavior of the solution $\mathbf{x}(t)$ as $t \rightarrow \infty$ is governed by the terms $e^{\alpha_j t}$. We have following possibilities here

- If $\alpha_j < 0$ then $e^{\alpha_j t} \rightarrow 0$ as $t \rightarrow \infty$
- If $\alpha_j > 0$ then $e^{\alpha_j t} \rightarrow \infty$ as $t \rightarrow \infty$
- If $\alpha_j = 0$ then $e^{\alpha_j t} \rightarrow 1$ as $t \rightarrow \infty$

Thus, we can deduce following three cases

- Case A: $||\mathbf{x}(t)|| \rightarrow 0$ as $t \rightarrow \infty$ if and only if $Re(\lambda_i) < 0$ for $i = 1, 2, ...m$ (Asymptotically stable solution)
- Case B: $||\mathbf{x}(t)|| \leq M < \infty$ as $t \rightarrow \infty$ if and only if $Re(\lambda_i) \leq 0$ for $i = 1, 2, ...m$ (Stable solution)
- Case C: $||\mathbf{x}(t)|| \rightarrow \infty \; at \; t \rightarrow \infty$ if for any λ_i,Re(λ_i) > 0 for i = 1,2,.......n (Unstable solution)

Note that asymptotic dynamics of linear ODE-IVP is governed by only eigenvalues of matrix A and is independent of the initial state $\mathbf{x}(t)$. Thus, based on the sign of real part of eignvalues of matrix A, the ODE-IVP is classified as asymptotically stable, stable or unstable.

REMARK 4. *The above approach can be extended to obtain local or perturbation solutions of nonlinear ODE-IVP systems*

$$\frac{d\mathbf{x}}{dt} = F\left[\mathbf{x}(t), u(t)\right] \quad ; \ \mathbf{x} = \ \mathbf{x}(0) \ \ at\ t = 0 \tag{2.27}$$

in the neighborhood of a steady state point $\overline{\mathbf{x}}$ *such that*

$$F(\overline{\mathbf{x}}) = 0 \tag{2.28}$$

Using Taylor expansion in the neighborhood of $\overline{\mathbf{x}}$ *and neglecting terms higher that first order, equation (2.27) can be approximated as*

$$\begin{aligned} \frac{d(\mathbf{x}-\overline{\mathbf{x}})}{dt} &= \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\overline{\mathbf{x}}} (\mathbf{x}-\overline{\mathbf{x}}) \\ \frac{d\delta\mathbf{x}}{dt} &= A\ \delta\mathbf{x} \ ; \ \ \delta\mathbf{x}(0) = \mathbf{x}(0) - \overline{\mathbf{x}} \\ A &= \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\overline{\mathbf{x}}} \end{aligned} \tag{2.29}$$

Note that the resulting equation is a linear multivariable system of type (2.1) and the perturbation solution can be computed as

$$\begin{aligned} \delta\mathbf{x}(t) &= e^{At}\delta\mathbf{x}(0) \\ \mathbf{x}(t) &= \overline{\mathbf{x}} + \delta\mathbf{x}(t) \end{aligned}$$

EXAMPLE 58. ***Stirred Tank Reactor***

The system under consideration is a Continuous Stirred Tank Reactor (CSTR) in which a non-isothermal, irreversible first order reaction

$$A \longrightarrow B$$

is taking place. The dynamic model for a non-isothermal CSTR is given as follows :

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{A0} - C_A) - k_0 \exp(-\frac{E}{RT})C_A \tag{2.30}$$

$$\frac{dT}{dt} = \frac{F}{V}(T_0 - T) + \frac{(-\Delta H_r)\,k_0}{\rho C_p}\exp(-\frac{E}{RT})C_A - \frac{Q}{V\rho C_p} \tag{2.31}$$

$$Q = \frac{aF_c^{b+1}}{F_c + \left(\dfrac{aF_c^b}{2\rho_c C_{pc}}\right)}(T - T_{cin}) \tag{2.32}$$

Table 1. Parameters and Steady State Operating Conditions of CSTR

Parameter (↓) Operating Point (→)		Stable	Unstable
Reaction rate constant (k_0)	min^{-1}	10^{10}	10^{10}
Inlet concentration of A (C_{A0})	$kmol/m^3$	2.0	2.0
Steady state flow rate of A (F)	m^3/min	1.0	1.0
Density of the reagent A (ρ)	g/m^3	10^6	10^6
Specific heat capacity of A(C_p)	cal/g^0C	1.0	1.0
Heat of reaction (ΔH$_r$)	$cal/kmol$	$-130 * 10^6$	$-130 * 10^6$
Density of the coolant (ρ_c)	g/m^3	10^6	10^6
Specific heat capacity of coolant (C_{pc})	cal/g^0C	1.0	1.0
Volume of the CSTR (V)	m^3	1.0	1.0
Coolant flow rate (F_c)	m^3/min	15	15
Inlet temperature of the coolant (T_{cin})	^0K	365	365
Inlet temperature of A (T_0)	^0K	323	343
Reactor temperature (T)	K	393.954	349.88
Reactor concentration of A (C_A)	$kmol/m^3$	0.265	1.372
a		$1.678X10^6$	$0.516X10^6$
Reaction Rate Parameter (E/R)	$(^0K)^{-1}$	8330	8330
b		0.5	0.5

This system exhibits entirely different dynamic characteristics for different set of parameter values (Marlin, 1995). The nominal parameter values and nominal steady state operating conditions of the CSTR for the *stable and unstable operating* points are given in Table 1.

- Perturbation model at stable operating point

$$\frac{d}{dt}\begin{bmatrix}\delta C_A \\ \delta T\end{bmatrix} = \begin{bmatrix}-7.559 & -0.09315 \\ 852.7 & 5.767\end{bmatrix}\begin{bmatrix}\delta C_A \\ \delta T\end{bmatrix} \tag{2.33}$$

Eigenvalues of $\left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\overline{\mathbf{x}}}$ are

$$\lambda_1 = -0.8960 + 5.9184i \ \ ; \ \ \lambda_2 = -0.8960 - 5.9184i \tag{2.34}$$

and all the trajectories for the unforced system (i.e. when all the inputs are held constant at their nominal values) starting in the neighborhood of the steady state operating point converge to the steady state.

- Perturbation model at unstable operating point

$$\frac{d}{dt}\begin{bmatrix}\delta C_A\\ \delta T\end{bmatrix}=\begin{bmatrix}-1.889 & -0.06053\\ 115.6 & 2.583\end{bmatrix}\begin{bmatrix}\delta C_A\\ \delta T\end{bmatrix} \tag{2.35}$$

Eigenvalues of $\left[\dfrac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\overline{\mathbf{x}}}$ are

$$\lambda_1 = 0.3468 + 1.4131i\ ;\quad \lambda_2 = 0.3468 - 1.4131i \tag{2.36}$$

and all the trajectories for the unforced system starting in any small neighborhood of the steady state operating point diverge from the steady state.

3. Numerical Methods for the solution of ODE-IVP

There are two basic approaches to integrating ODE-IVP numerically.

- Methods based on Taylor series expansion
  - Runge - Kutta methods
- Methods based on polynomial approximation
  - Predictor corrector methods
  - Orthogonal collocation based methods

In this section, we describe these methods in detail.

3.1. Why develop methods only for the set of first order ODE's?

In the above illustrations, the system of equations under consideration is the set of simultaneous first order ODEs represented as

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x},t)\ ;\quad \mathbf{x}(0)=\mathbf{x}_0\ ;\mathbf{x}\in R^n \tag{3.1}$$

In practice, not all models appear as first order ODEs. In general, one can get an m'th order ODE of the type:

$$\frac{d^m y}{dt^m} = f[y,\frac{dy}{dt},\frac{d^2y}{dt^2},.....,\frac{d^{m-1}y}{dt^{m-1}},t] \tag{3.2}$$

$$\text{Given } y(0),......\frac{d^{m-1}y}{dt^{m-1}}(0) \tag{3.3}$$

Now, do we develop separate methods for each order? It turns out that such a exercise in unnecessary as a m'th order ODE can be converted to m first order

ODEs. Thus, we can define auxiliary variables

$$
\begin{aligned}
x_1(t) &= y(t) \\
x_2(t) &= \frac{dy}{dt} \\
&....... \\
&....... \\
x_m(t) &= \frac{d^{m-1}y}{dt^{m-1}}
\end{aligned}
\tag{3.4}
$$

Using these variables, the original nth order ODE can be converted to n first order ODE's as,

$$
\begin{aligned}
\frac{dx_1}{dt} &= x_2 \\
\frac{dx_2}{dt} &= x_3 \\
&....... \\
\frac{dx_{m-1}}{dt} &= x_m \\
\frac{dx_m}{dt} &= f[x_1, x_2, x_3, ..., x_m, t]
\end{aligned}
\tag{3.5}
$$

Defining function vector

$$
F(\mathbf{x}) = \begin{bmatrix} x_2 \\ ..... \\ x_m \\ f[x_1, x_2, x_3, ..., x_m, t] \end{bmatrix}
\tag{3.6}
$$

we can write the above set of

$$
\frac{d\mathbf{x}}{dt} = F(\mathbf{x}, t)
\tag{3.7}
$$

$$
\mathbf{x}(0) = \begin{bmatrix} y(0) & \frac{dy}{dt}(0) .... \frac{d^{m-1y}}{dt^{m-1}}(0) \end{bmatrix}^T
\tag{3.8}
$$

Thus, it is sufficient to study only the solution methods for solving n first order ODE's. Any set of higher order ODEs can be reduced to a set of first order ODEs. Also, as forced systems (non-homogeneous systems) can be looked upon as unforced systems (homogenous systems) with variable parameters, it is sufficient to study the solution methods for homogenous set of equations of the type (3.1).

3.2. Basic concepts. Consideration is the set of equations defined by equation (3.1). Let $\{\mathbf{x}^*(t) : 0 \leq t \leq t_f\}$ denote the true / actual solution of the above ODE-IVP. In general, for a nonlinear ODE, it is seldom possible to obtain the true solution analytically. The aim of numerical methods is to find an approximate solution numerically. Let $t_1, t_2, ..., t_n$ be a sequence of numbers such that

$$0 < t_1 < t_2 < .... < t_n < .... < t_f \tag{3.9}$$

Instead of attempting to approximate the function $\mathbf{x}^*(t)$, which is defined for all values of t such that $0 \leq t \leq t_f$, we attempt to approximate the sequence of vectors $\{\mathbf{x}^*(t_n) : n = 1, ...f\}$. Thus, in order to integrate over a large interval $0 \leq t \leq t_f$,we solve a sequence of ODE-IVPs subproblems

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= F(\mathbf{x}, t) \ ; \quad \mathbf{x}(t_n) = \mathbf{x}(n) \ ; \\ t_n &\leq t \leq t_{n+1} \ ; \ (n = 1, 2, ...f) \end{aligned} \tag{3.10}$$

each defined over a smaller interval $[t_n, t_{n+1}]$. This generates a sequence of approximate solution vectors $\{\mathbf{x}(t_n) : n = 1, ...f\}$. The difference $h_n = t_n - t_{n-1}$ is referred to as the integration step size or the integration interval. Two possibilities can be considered regarding the choice of the sequence $\{t_n\}$

- Fixed integration interval: The numbers t_n are equispaced, i.e., $t_n = nh$ for some $h > 0$
- Variable size integration intervals

For the sake of convenience, we introduce the notation

$$F(n) \equiv F[\mathbf{x}(t_n), t_n] \tag{3.11}$$

$$\mathbf{x}(n) \equiv \mathbf{x}(t_n) \tag{3.12}$$

$$\left(\frac{\partial F}{\partial \mathbf{x}}\right)_n = \left(\frac{\partial F}{\partial \mathbf{x}}\right)_{(\mathbf{x}(t_n), t_n)} \tag{3.13}$$

and use it throughout in the rest of the text.

3.2.1. *Two Basic Approaches : Implicit and Explicit.* There are two basic approaches to numerical integrations. To understand these approaches, consider the integration of the equation (3.1) over the interval $[t_n, t_{n+1}]$ using Euler's method. Let us also assume that the numbers t_n are equi-spaced and h is the integration stepsize.

- **Explicit Euler method:**If the integration interval is *small*,

$$\begin{aligned} &\frac{d\mathbf{x}}{dt} \cong \frac{\mathbf{x}(n+1) - \mathbf{x}(n)}{h} = F[\mathbf{x}(n), t_n] \\ &\mathbf{x}(n+1) = \mathbf{x}(n) + hF(n), \qquad (n = 0, 1, ..., n-1) \end{aligned} \tag{3.14}$$

The new value $\mathbf{x}(n+1)$ is a function of only the past value of $\mathbf{x}$ i.e., $\mathbf{x}(n)$. This is a non-iterative scheme.

- **Implicit Euler method:**

$$\begin{aligned} &\frac{d\mathbf{x}}{dt} \cong \frac{\mathbf{x}(n+1)-\mathbf{x}(n)}{h} = F[\mathbf{x}(n+1), t_{n+1}] \\ &\mathbf{x}(n+1) = \mathbf{x}(n) + hF(n+1), \qquad (n = 0, 1, ..., n-1) \end{aligned} \tag{3.15}$$

Each of the above equation has to be solved by iterative method. For example if we use successive substitution method for solving the resulting nonlinear equation(s), the algorithm can be stated as follows:

```
Initialize: $\mathbf{x}(0), t_f, h, \in, N = t_f/h$
FOR n = 1 TO n = N
    $\mathbf{x}^{(0)}(n+1) = \mathbf{x}(n) + hF[\mathbf{x}(n), t_n]$


    WHILE ( $\delta > \in$)
        $\mathbf{x}^{(k+1)}(n+1) = \mathbf{x}(n) + hF[\mathbf{x}^{(k)}(n+1), t_{n+1}]$
        $\delta = \dfrac{||\mathbf{x}^{(k+1)}(n+1) - \mathbf{x}^{(k)}(n+1)||}{||\mathbf{x}^{(k)}(n+1)||}$
    END WHILE
    $\mathbf{x}(n+1) = \mathbf{x}^{(k)}(n+1)$
END FOR
```

3.2.2. *Variable stepsize implementation with accuracy monitoring.* One practical difficulty involved in the integration with fixed stepsize is the choice of stepsize such that the approximation errors are kept small. Alternatively, a variable stepsize algorithm is implemented with error monitoring as follows.

Given: $t_n, \mathbf{x}(n) = \mathbf{x}(t_n), \varepsilon$

- Step 1: Choose stepsize h_1 and let $t_{n+1}^{(1)} = t_n + h_1$
- Step 2: Compute $\mathbf{x}^{(1)}(n+1)$ using an integration method (say explicit Euler).
- Step 3: Define $h_2 = h_1/2$; $\quad t_{n+1}^{(2)} = t_n + h_2$
  $t_{n+2}^{(2)} = t_n + 2h_2 \quad (= t_{n+1}^{(1)})$
  Compute $\mathbf{x}^{(2)}$ *and* $\mathbf{x}_{n+2}^{(2)}$ by the same integration method.
- Step 4: IF $(\left\|\mathbf{x}^{(1)}(n+1) - \mathbf{x}^{(2)}(n+2)\right\| < \varepsilon)$,
  (Accept $\mathbf{x}^{(1)}(n+1)$ as the new value)
  Set $\mathbf{x}(n+1) = \mathbf{x}^{(1)}(n+1)$, and $n = n + 1$ and proceed to Step 1.
  ELSE
  set $h_1 = h_2$ and proceed to the step 2.
  END IF

3.2.3. *Stiffness of ODEs.* The problem of integrating multi-variable ODE-IVP with some variables changing very fast in time while others changing slowly, is difficult to solve. This is because, the stepsize has to be selected according to the fastest changing variable / mode. For example, consider the equation

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} -100 & 0 \\ 2 & -1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{3.16}$$

$$A = \begin{bmatrix} -100 & 0 \\ 2 & -1 \end{bmatrix} \quad ; \quad y(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \tag{3.17}$$

It can be shown that the solution for the above system of equations is

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 2e^{-100t} \\ \frac{103}{99}e^{-t} - \frac{4}{99}e^{-100t} \end{bmatrix} \tag{3.18}$$

It can be observed that the terms with e^{-100t} lead to a sharp decrease in $y_1(t)$ and to a small maximum in $y_2(t)$ at $t = 0.0137$. The term $y_2(t)$ is dominated by e^{-t} which decreases slowly. Thus,

$$y_1(t) < 0.01y_1(0) \qquad for \qquad t > 0.03 \tag{3.19}$$

$$y_2(t) < 0.01y_1(t) \qquad for \qquad t > 4.65 \tag{3.20}$$

Now, stepsize should be selected such that the faster dynamics can be captured. The stiffness of a given ODE-IVP is determined by finding the stiffness ratio defined as

$$S.R. = \frac{|Re\lambda_i(A)|_{\max}}{|Re\lambda_i(A)|_{\min}} \tag{3.21}$$

where matrix A is defined above. Systems with '*large*' stiffness ratio are called as stiff.

Remark 5. *This analysis can be extended to a general nonlinear systems only locally. Using Taylor's theorem, we can write*

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}) = F\left[\mathbf{x}(n) + \mathbf{x}(t) - \mathbf{x}(n)\right] \tag{3.22}$$

$$\cong F(\mathbf{x}_n) + \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}(n)} \left[\mathbf{x}(t) - \mathbf{x}(n)\right] \tag{3.23}$$

Local stiffness ratio can be calculated using eigenvalues of the Jacobian and the ODE-IVP is locally stiff if the local S.R. is high, i.e., the system has at least one eigenvalue which does not contribute significantly over most of the domain of interest. In general, eigenvalues of the Jacobian are time dependent and S.R.

is a function of time. Thus, for stiff systems it is better to use variable step size methods or special algorithms for stiff systems.

3.3. Taylor's series based methods. Consider a simple scalar case

$$\frac{dx}{dt} = f(x,t) \ ; \quad x \in R \tag{3.24}$$

Suppose we know the exact solution $x^*(t)$ at time t_n, i.e. $x^*(n)$, then we can compute $x^*(n+1)$ using Taylor series as follows:

$$x^*(n+1) = x^* + h\frac{dx^*(t_n)}{dt} + \frac{1}{2!}h^2\frac{d^2x^*(t_n)}{dt^2} + ... \tag{3.25}$$

The various derivatives in the above series can be calculated using the differential equation, as follows:

$$\frac{dx^*(t_n)}{dt} = f\left[t_n, x^*(n)\right] \tag{3.26}$$

$$\frac{d^2x^*(t_n)}{dt^2} = \left[\frac{\partial f}{\partial x}\right]_{(x^*(n),t_n)} f\left[x^*(n), t_n\right] + \frac{\partial f\left[x^*(n), t_n\right]}{\partial t} \tag{3.27}$$

and so on. Let us now suppose that instead of actual solution $x^*(n)$, we have available an approximation to $x^*(n)$, denoted as $x(n)$. With this information, we can construct $x(n+1)$, as

$$x(n+1) = x(n) + hf(n) + \frac{h^2}{2}\left[\left(\frac{\partial f}{\partial x}\right)_n f(n) + \left(\frac{\partial f}{\partial t}\right)_n\right] + ... \tag{3.28}$$

We can now make a further approximation by truncating the infinite series. If the Taylor series is truncated after the term involving h^k, then the Taylor's series method is said to be of order k.

- **Order 1(Euler explicit formula)**

$$x(n+1) = x(n) + hf(n) \tag{3.29}$$

- **Order 2**

$$x(n+1) = x(n) + hf(n) + \frac{h^2}{2}\left[\left(\frac{\partial f}{\partial x}\right)_n f(n) + \left(\frac{\partial f}{\partial t}\right)_n\right] \tag{3.30}$$

  Taylor's series methods are useful starting points for understanding more sophisticated methods, but are not of much computational use. First order method is too inaccurate and the higher order methods require calculation of a lot of partial derivatives.

3.4. Runge-Kutta (R-K) Methods.

3.4.1. *Univariate Case.* Runge-Kutta methods duplicate the accuracy of the Taylor series methods, but do not require the calculation of higher partial derivatives. For example, consider the second order method that uses the formula:

$$x(n+1) = x(n) + ak_1 + bk_2 \tag{3.31}$$

$$k_1 = hf(t_n, x(n)) = hf(n) \tag{3.32}$$

$$k_2 = hf(t_n + \alpha h, x(n) + \beta k_1) \tag{3.33}$$

The real numbers a, b, α, β, are chosen such that the RHS of (3.31) approximates the RHS of Taylor series method of order 2 (ref. 3.30). To achieve this, consider Taylor series expansion of the function k_2, about $(t_n, x(n))$.

$$\frac{k_2}{h} = f(t_n + \alpha h, x(n) + \beta hf(n)) \tag{3.34}$$

$$= f(t_n, x(n)) + \alpha h \left(\frac{\partial f}{\partial t}\right)_n + \beta h \left(\frac{\partial f}{\partial x}\right)_n f(n) + O(h^3) \tag{3.35}$$

Substituting this in equation (3.31), we have

$$\begin{aligned} x(n+1) &= x(n) + ahf(n) \\ &\quad +bh\left[f(t_n, x(n)) + \alpha h \left(\frac{\partial f}{\partial t}\right)_n + \beta h \left(\frac{\partial f}{\partial x}\right)_n f(n)\right] + O(h^3) \end{aligned} \tag{3.36}$$

$$= x(n) + (a+b)hf(n) + \alpha bh^2 \left(\frac{\partial f}{\partial t}\right)_n + \beta bh^2 \left(\frac{\partial f}{\partial x}\right)_n f(n) + O(h^3) \tag{3.37}$$

Comparing 3.30 and 3.37, we get

$$\begin{gathered} a + b = 1 \\ \alpha b = \beta b = \tfrac{1}{2} \end{gathered} \tag{3.38}$$

There are 4 unknowns and 3 equations and so we can assign one arbitrarily giving:

$$\begin{gathered} a = 1 - b; \\ \alpha = \tfrac{1}{2b}; \quad \beta = \tfrac{1}{2b}; \quad b \neq 0 \end{gathered} \tag{3.39}$$

Thus, the general 2nd order algorithm can be stated as

$$x(n+1) = x(n) + h\left[(1-b)f(n) + bf\left(t_n + \frac{h}{2b}, x(n) + \frac{h}{2b}f(n)\right)\right] \tag{3.40}$$

- **Heun's modified algorithm:**Set b = 1/2.

$$x(n+1) = x(n) + h\left[(\frac{1}{2}f(n) + \frac{1}{2}f\left(t_n + h, x(n) + hf(n)\right)\right] \tag{3.41}$$

- **Modified Euler-Cauchy Method:** Set b = 1.

$$x(n+1) = x(n) + hf\left[t_n + \frac{h}{2}, x(n) + \frac{h}{2}f(n)\right] \tag{3.42}$$

It must be emphasized that 3.40. and 3.30 do not give identical results. However, if we start from the same $x(n)$, then $x(n+1)$ given by 3.30 and 3.40 would differ only by $O(h^3)$.

3.4.2. *Multivariate Case.* Even though the above derivation has been worked for one dependent variable case, the method can be easily extended to multi-variable case. For example, the most commonly used fourth order R-K method for one variable can be stated as

$$x(n+1) = x(n) + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + +k_4\right) \tag{3.43}$$

$$\begin{aligned}
k_1 &= f(t_n, x(n)) = f(n) && (3.44)\\
k_2 &= f\left(t_n + \frac{h}{2}, x(n) + \frac{h}{2}k_1\right) && (3.45)\\
k_3 &= f\left(t_n + \frac{h}{2}, x(n) + \frac{h}{2}k_2\right) && (3.46)\\
k_4 &= f\left(t_n + h, x(n) + hk_3\right) && (3.47)
\end{aligned}$$

Now, suppose we want to use this method for solving m simultaneous ODE-IVPs

$$\begin{aligned}
\frac{d\mathbf{x}}{dt} &= \mathbf{F}(\mathbf{x}, t) && (3.48)\\
\mathbf{x}(0) &= \mathbf{x}_0 && (3.49)
\end{aligned}$$

where $\mathbf{x} \in R^m$ and $F(\mathbf{x}, t)$ is a $m \times 1$ function vector. Then, the above algorithm can be modified as follows

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \frac{h}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + +\mathbf{k}_4\right) \tag{3.50}$$

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{F}\left(t_n, \mathbf{x}(n)\right) = \mathbf{F}(n) && (3.51)\\
\mathbf{k}_2 &= \mathbf{F}\left(t_n + \frac{h}{2}, \mathbf{x}(n) + \frac{h}{2}\mathbf{k}_1\right) && (3.52)\\
\mathbf{k}_3 &= \mathbf{F}\left(t_n + \frac{h}{2}, \mathbf{x}(n) + \frac{h}{2}\mathbf{k}_2\right) && (3.53)\\
\mathbf{k}_4 &= \mathbf{F}\left(t_n + h, \mathbf{x}(n) + h\mathbf{k}_3\right) && (3.54)
\end{aligned}$$

Note that $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3$ and $\mathbf{k}_4$ are $n \times 1$ function vectors.

Note that Runge-Kutta methods can be implemented using variable step size with accuracy monitoring. Thus, these methods (with variable step size implementation) are preferred when $\mathbf{x}(t)$ is expected to change rapidly in some regions and slowly in others.

3.5. Multistep Methods.

3.5.1. *Univariate Case.* The multi-step methods are based on the Weierstrass theorem, which states that any continuous function over a finite interval can be uniformly approximated to any desired degree of accuracy by a polynomial of appropriate degree. Thus, we approximate the solution of a given differential equation by a polynomial in the independent variable t. In order to understand how this is achieved, consider the scalar differential equation

$$\frac{dx}{dt} = f(x,t); \qquad x(t_n) = x(n); \quad x \in R \tag{3.55}$$

with uniformly spaced integration (time) intervals. At time $t = t_n$, we have state and derivative information in the 'past' i.e.

$$\{x(n), x(n-1), x(n-2).........x(0)\}$$

and

$$\{f(n), f(n-1), f(n-2).........f(0)\}$$

which can be used to construct the polynomial approximation. We approximate $x(t)$ in the neighborhood of $t = t_n$ by constructing a local polynomial approximation of type

$$x^{(n)}(t) = a_{0,n} + a_{1,n}t + a_{2,n}t^2 + ... + a_{m,n}t^m \tag{3.56}$$

and use it to estimate or extrapolate $x(n+1)$.(Note that subscript 'n' used for coefficients indicate that the coefficients are corresponding to polynomial approximation at time $t = t_n$). Here, the coefficients of the polynomial are estimated using the state and derivative information from the past and possibly $f(n+1)$. In order to see how this can be achieved, consider a simple case where we want construct a second order approximation

$$x^{(n)}(t) = a_{0,n} + a_{1,n}t + a_{2,n}t^2 \tag{3.57}$$

at instant $t = t_n$. This implies the derivative $f(x,t)$ at time t can be computed as

$$f(x,t) = \frac{dx^{(n)}(t)}{dt} = a_{1,n} + 2a_{2,n}\,t \tag{3.58}$$

For the sake of computational convenience, we choose a shifted time scale as follows:

$$t_n = 0; \quad t_{n+1} = h; \quad t_{n-1} = -h, \tag{3.59}$$

Now, there are several ways we could go about estimating the unknown parameters of the polynomial.

- **Explicit algorithm:** Let us use only the current and the past information of state and derivatives, which will lead to an explicit algorithm.

$$\begin{aligned} f(n-1) &= a_{1,n} - 2a_{2,n}\,h \\ f(n) &= a_{1,n} \end{aligned} \tag{3.60}$$

$$x(n) = a_{0,n} \tag{3.61}$$

Solving above equations simultaneously, we get coefficients

$$a_{0,n} = x(n) \;\; ; \; a_{1,n} = f(n) \;\; ; \;\; a_{2,n} = \frac{f(n) - f(n-1)}{2h} \tag{3.62}$$

which can be used to extrapolate the value of $x(t)$ at $t = t_{n+1}$ as

$$\begin{aligned} x(n+1) &= a_{0,n} + a_{1,n}h + a_{2,n}h^2 \\ &= x(n) + f(n)\,h + \left[\frac{f(n) - f(n-1)}{2h}\right] h^2 \\ &= x(n) + h\left[\frac{3}{2}f(n) - \frac{1}{2}f(n-1)\right] \end{aligned} \tag{3.63}$$

- **Implicit algorithm:** Alternatively, we can choose to estimate $x(n+1)$ based on derivative at t_{n+1},i.e.

$$f(n+1) = a_{1,n} + 2a_{2,n}\,h \tag{3.64}$$

$$f(n) = a_{1,n} \tag{3.65}$$

$$x(n) = a_{0,n} \tag{3.66}$$

These equations yield following set of coefficients

$$a_{0,n} = x(n) \;\; ; \; a_{1,n} = f(n) \;\; ; \;\; a_{2,n} = \frac{f(n+1) - f(n)}{2h} \tag{3.67}$$

and $x(n+1)$ can be estimated as

$$x(n+1) = a_{0,n} + a_{1,n}h + a_{2,n}h^2 \tag{3.68}$$

$$= x(n) + f(n)\,h + \left[\frac{f(n+1) - f(n)}{2h}\right] h^2 \tag{3.69}$$

The above expression can be rearranged as

$$x(n+1) = x(n) + \frac{h}{2}\left[f(n) + f(n+1)\right] \tag{3.70}$$

which is popularly known as trapezoidal rule or Crank-Nicholson algorithm.

Thus, a more general expression for computational form of $x(n+1)$ can be stated as

$$\begin{aligned} x(n+1) &= \alpha_0 x(n) + \alpha_1 x(n-1) + ... + \alpha_p x(n-p) \\ & \quad +h\left[\beta_{-1} f(n+1) + \beta_0 f(n) + \beta_1 f(n-1) + ... + \beta_p f(n-p)\right] \end{aligned} \tag{3.71}$$

or

$$x(n+1) = \sum_{i=0}^{p} \alpha_i x(n-i) + h \sum_{i=-1}^{p} \beta_i f(n-i) \tag{3.72}$$

where p is an integer and α_i, β_i are real numbers to be selected. Note that if $\beta_{-1} \neq 0$, we get an implicit formula for the unknown quantity $x(n+1)$, else we get an explicit formula. An algorithm of the type 3.71 is called $(p+1)$ step algorithm because $x(n+1)$ is given in terms of the values of x at previous $(p+1)$ steps $[x(n), ..., x(n-p)]$. Formula 3.71 can also be thought of arising from the discrete approximation of the expression

$$x(t) = x_0^* + \int_0^t f[x(\tau), \tau] d\tau \tag{3.73}$$

Order of the algorithm is the degree of the highest-degree polynomial for which 3.71 gives an exact expression of $x(n+1)$. To see how this definition of order is used, consider the development of the m'th order algorithm in scalar case. i.e., $x \in R$. (similar arguments can be used in vector case). Suppose polynomial solution of initial value problem is given by

$$x^{(n)}(t) = a_{0,n} + a_{1,n} t + a_{2,n} t^2 + ... + a_{m,n} t^m = \sum_{j=0}^{m} a_{j,n} (t)^j \tag{3.74}$$

$$f(x,t) = \frac{dx^{(n)}}{dt} = a_{1,n} + 2a_{2,n} t + ... + m\, a_{m,n} t^{m-1} = \sum_{j=1}^{m} j a_{j,n} (t)^{j-1} \tag{3.75}$$

For the sake of convenience, we choose a shifted time scale as follows:

$$t_n = 0; t_{n+1} = h; t_{n-1} = -h, ...t_{n-i} = -ih \tag{3.76}$$

Thus we have,

$$x(n+1) = a_{0,n} + a_{1,n} h + a_{2,n} h^2 + ... + a_{m,n} h^m \tag{3.77}$$

$$x(n-i) = a_{0,n} + a_{1,n}(-ih) + a_{2,n}(-ih)^2 + ... + a_{m,n}(-ih)^m \tag{3.78}$$

$$i = 0, 1, ..., p \tag{3.79}$$

$$f(n-i) = a_{1,n} + 2a_2, n(-ih) + ... + m\, a_{m,n}(-ih)^{m-1} \tag{3.80}$$

$$i = -1, 0, ...p \tag{3.81}$$

Substitution of equations (3.77),(3.78) and (3.80) into (3.71) gives what is known as the exactness constraints for the algorithm as

$$\begin{aligned}
\sum_{j=0}^{m} a_{j,n}(h)^j &= \sum_{i=0}^{p} \alpha_i \left[\sum_{j=0}^{m} a_{j,n}(-ih)^j\right] + h\sum_{i=-1}^{p} \beta_i \left[\sum_{j=1}^{m} j a_{j,n}(-ih)^{j-1}\right] \\
&= \left(\sum_{i=0}^{p} \alpha_i\right) a_{0,n} + \left(\sum_{i=0}^{p}(-i)\alpha_i + \sum_{i=-1}^{p}(-i)^0 \beta_i\right) a_{1,n}h + ... \\
&\quad ... + \left(\sum_{i=0}^{p}(-i)^m \alpha_i + m\sum_{i=-1}^{p}(-i)^{m-1}\beta_i\right) a_{m,n}h^m \qquad (3.82)
\end{aligned}$$

Because we would like (3.82) to hold independently of any stepsize, we obtain the following equations (constraints) by equating like powers of h.

$$\sum_{i=0}^{p} \alpha_i = 1 \; ; \qquad (j=0) \qquad (3.83)$$

$$\sum_{i=0}^{p}(-i)^j \alpha_i + j\sum_{i=-1}^{p}(-i)^{j-1}\beta_i = 1 \;\; ; \quad (j = 1, 2, ..., m) \qquad (3.84)$$

$$\text{Note : } (i)^j = 1 \text{ when } i = j = 0$$

Thus, equations (3.88-3.90) gives $m+1$ constraints and the number of variables are $2p + 3$, namely $\alpha_0, ...\alpha_p, \beta_{-1}, \beta_0, ...\beta_p$. Any choice of these constants makes the corresponding algorithm 3.71 exact for m'th order polynomial. The set of equations (3.83) provide $(m+1)$ constraints on $(2p+3)$ variables. Thus, in order for the algorithm (3.71) to be exact in case of the mth degree polynomial we must have

$$(m+1) \leq 2p+3 \qquad (3.85)$$

If equality holds, i.e. when

$$m = 2(p+1) \qquad (3.86)$$

then we can solve for $\{\alpha_i\}$ and $\{\beta_i\}$ exactly.

Now, let us re-derive the 2'nd order implicit algorithm again using the above approach. Constraints for this case can be generated by equating coefficients of

$$\begin{aligned}
a_{0,n} + a_{1,n}h + a_{2,n}h^2 &= \sum_{i=0}^{p} \alpha_i[a_{0,n} + a_{1,n}(-ih) + a_{2,n}(-ih)^2] \\
&\quad + h\sum_{i=-1}^{p} \beta_i[a_{1,n} + 2a_{2,n}(-ih)] \qquad (3.87)
\end{aligned}$$

The resulting constraints are

$$\sum_{i=0}^{p}\alpha_i = 1 \tag{3.88}$$

$$\sum_{i=0}^{p}(-i\alpha_i)+\sum_{i=-1}^{p}\beta_i = 1 \tag{3.89}$$

$$\sum_{i=0}^{p}i^2\alpha_i+\sum_{i=-1}^{p}(-2i\beta_i) = 1 \tag{3.90}$$

Clearly for (3.88-3.90) to hold, we must have $2p+3 \geq 3$. The second order algorithm with the smallest number of constants α_i, β_i is obtained by setting $2p+3=3$, i.e., $p=0$. In this case,

$$\begin{aligned}&\alpha_0 = 1\\ &\beta_{-1}+\beta_0 = 1\\ &2\beta_{-1} = 1\end{aligned} \tag{3.91}$$

which gives

$$\alpha_0 = 1; \qquad \beta_{-1} = 1/2; \qquad \beta_0 = 1/2 \tag{3.92}$$

and the second order algorithm becomes

$$x(n+1) = x(n) + \frac{h}{2}\left[f(n)+f(n+1)\right] \tag{3.93}$$

3.5.2. *Examples of Multi-step methods.* A number of multi-step algorithms can be obtained by making suitable choices of the parameters $\{\alpha_i\}$ and $\{\beta_i\}$. Some of the popular algorithms are discussed in this sub-section.

Adams-Bashworth (explicit method): Choose

$$\alpha_1 = \alpha_2 = ... = \alpha_p = 0 \tag{3.94}$$

$$\beta_{-1} = 0 \tag{3.95}$$

$$p = m-1 \tag{3.96}$$

These are additional $(p+1)$ equations.

$$\begin{aligned}\text{Total number of constraints} &= (m+1)+(p+1). = 2m+1\\ \text{Total number of variables} &= (2p+3) = 2m+1\end{aligned}$$

Out of these, $(p+1=m)$ variables are selected to be zero and $(m+1)$ constants namely, $\alpha_0, \beta_0, ...\beta_p$ are to be detected. Using constraints for $j=0$,

$$\sum_{i=0}^{p}\alpha_i = 1; \Rightarrow \alpha_0 = 1 \tag{3.97}$$

Using the other constraints,

$$\begin{bmatrix} 1 & 1 & ... & 1 \\ 0 & (-1) & ... & (-p) \\ ... & ... & ... & ... \\ 0 & (-1)^{m-1} & ... & (-p)^{m-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_p \end{bmatrix} = \begin{bmatrix} 1/j \\ 1/j \\ ... \\ 1/j \end{bmatrix} \tag{3.98}$$

Solving for the $\beta' s$, we can write the algorithm as

$$x(n+1) = x(n) + h\left[\beta_0 f(n) + \beta_1 f(n-1) + ... + \beta_p f(n-p)\right] \tag{3.99}$$

Adam-Moulton Implicit Algorithms: Choose

$$p = m - 2 \tag{3.100}$$

$$\alpha_1 = \alpha_2 = ... = \alpha_p = 0 \tag{3.101}$$

For j = 0, we have

$$\sum_{i=0}^{p} \alpha_i = 1; \Rightarrow \alpha_0 = 1 \tag{3.102}$$

Remaining m variables $\beta_{-1}, ..., \beta_{m-2}$ can be determined by solving

$$\begin{bmatrix} 1 & 1 & ... & 1 \\ 0 & (-1) & ... & (-p) \\ ... & ... & ... & ... \\ 0 & (-1)^{m-1} & ... & (-p)^{m-1} \end{bmatrix} \begin{bmatrix} \beta_{-1} \\ \beta_0 \\ ... \\ \beta_p \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ ... \\ 1/m \end{bmatrix} \tag{3.103}$$

The algorithm can be written as

$$x(n+1) = x(n) + h\begin{bmatrix} \beta_0 f(n) + \beta_1 f(n-1) + \\ ... + \beta_p f(n-p)] + \beta_{-1} f(n+1) \end{bmatrix} \tag{3.104}$$

$$= y_n + h\beta_{-1} f\left[x(n+1), t_{n+1}\right] \tag{3.105}$$

where $y(n)$ represents sum of all terms which are known from the past data. The above implicit equation has to be solved iteratively to obtain $x(n+1)$.

3.5.3. *Predictor-Corrector Algorithms.* We saw that a m step Adams-Bashworth algorithm is exact for polynomials of order m, while a m-step Adams-Moulton algorithm is exact for the polynomials of order $(m+1)$. However, the Adams-Moulton algorithm is implicit, i.e.,

$$x(n+1) = y(n) + h\beta_{-1} f\left[x(n+1), t_{n+1}\right] \tag{3.106}$$

where the quantity $y(n)$ depends on $x(n), ..., x(n-p)$ and is known. The above implicit equation can be solved iteratively as

$$x^{(k+1)}(n+1) = y(n) + h\beta_{-1} f\left[x^{(k)}(n+1), t_{n+1}\right] \tag{3.107}$$

where iterations are terminated when

$$|x^{(k+1)}(n+1) - x^{(k)}(n+1)| < \in \tag{3.108}$$

If we choose the initial guess $x^{(0)}(n+1)$ reasonably close to the solution, the convergence of the iterations is accelerated. To achieve this, we choose $x^{(0)}(n+1)$ as the value generated by an explicit $m-$step algorithm and then apply the iterative formula. This is known as the predictor-corrector method. For example, a two-step predictor-corrector algorithm can be given as

$$x^{(0)}(n+1) = x(n) + h\left[\frac{3}{2}f(n) - \frac{1}{2}f(n-1)\right] \qquad \text{(Predictor)} \tag{3.109}$$

(3.110)

$$x^{(k+1)}(n+1) = x(n) + h\left[\frac{1}{2}f(x^{(k)}(n+1), t_{n+1}) + \frac{1}{2}f(n)\right] \qquad \text{(Corrector)}$$

If the stepsize is selected properly, relatively few applications of the correction formula are enough to determine $x(n+1)$, with a high degree of accuracy.

Gear's Predictor-Corrector Algorithms: A popular algorithm used for numerical integration is Gear's predictor corrector. The equations for this algorithm are as follows:

- Gear's m-th order predictor algorithm is an explicit algorithm, with

$$p = m-1 \tag{3.111}$$

$$\beta_{-1} = \beta_1 = ... = \beta_p = 0; \quad \beta_0 \neq 0 \tag{3.112}$$

$$x(n+1) = \alpha_0 x(n) + \alpha_1 x(n-1) + ... + \alpha_p x(n-p) + h\beta_0 f(n) \tag{3.113}$$

- Gear's m-th order corrector

$$p = m-1 \tag{3.114}$$

$$\beta_0 = \beta_1 = ... = \beta_p = 0; \quad \beta_{-1} \neq 0 \tag{3.115}$$

$$x(n+1) = \alpha_0 x(n) + \alpha_1 x(n-1) + ... + \alpha_p x(n-p) + h\beta_{-1} f(n+1) \tag{3.116}$$

Coefficients of the above algorithm can be computed by setting up appropriate constraint equations as shown above.

3.5.4. *Multivariate Case.* Even though the above derivations have been worked for one dependent variable case, these methods can be easily extended to multivariable case

$$\frac{dx}{dt} = F(\mathbf{x}, t) \ ; \quad \mathbf{x} \in R^n \tag{3.117}$$

where $F(\mathbf{x}, t)$ is a $n \times 1$ function vector. In the multivariable extension, the scalar function $f(x, t)$ is replaced by the function vector $F(\mathbf{x}, t)$, i.e.

(3.118)

$$\begin{aligned} \mathbf{x}(n+1) &= \alpha_0 \mathbf{x}(n) + \alpha_1 \mathbf{x}(n-1) + ... + \alpha_p \mathbf{x}(n-p) \\ &\quad + h\left[\beta_{-1} F(n+1) + \beta_0 F(n) + \beta_1 F(n-1) + ... + \beta_p F(n-p)\right] \end{aligned}$$

where

$$\begin{aligned} F(n-i) &\equiv F\left[\mathbf{x}(t_n - ih), (t_n - ih)\right] \\ i &= -1, 0, 1, ...p \end{aligned} \tag{3.119}$$

and the scalar coefficients $\left\{\alpha_0 ....\alpha_p, \beta_{-1}, \beta_0, \beta_1, ...\beta_p\right\}$ are identical with the coefficients derived for the scalar case as described in the above section.

The main advantages and limitations of multi-step methods can be summarized as follows

- **Advantages:**

  There are no extraneous 'inter-interval' calculations as in the case of Runge-Kutta methods.

  Can be used for stiff equations if integration interval is chosen carefully.

- **Limitations:**

  Time instances should be uniformly spaced and selection of integration interval is a critical issue.

4. Stability Analysis and Selection of Integration Interval

Selection of integration interval is a crucial parameter while solving ODE-IVPs numerically. In order to see how choice of integration interval can affect solution behavior consider a scalar linear equation

$$\frac{dx}{dt} = ax; \qquad x(0) = x_0 \tag{4.1}$$

Analytical (true) solution of the above equation is given as

$$x^*(t) = e^{at} x(0) \tag{4.2}$$

Defining $x^*(t_n) = x^*(n)$, we can write true solution as a difference equation

$$x^*(n) = e^{anh} x(0) \Rightarrow x^*(n+1) = e^{ah} x^*(n) \tag{4.3}$$

Now consider the approximate solution of the above ODE-IVP by explicit Euler methods

$$\begin{aligned} x(n+1) &= x(n) + hf(n) \\ &= (1+ah)x(n) \\ \Rightarrow x(n) &= (1+ah)^n x(0) \end{aligned} \tag{4.4}$$

Defining approximation error introduced due to numerical integration,

$$e(n) = x^*(n) - x(n) \tag{4.5}$$

we can write

$$e(n+1) = (1+ah)e(n) + \left[e^{ah} - (1+ah)\right] x^*(n) \tag{4.6}$$

Thus, the combined equation becomes

$$\begin{bmatrix} e(n+1) \\ x^*(n+1) \end{bmatrix} = \begin{bmatrix} (1+ah) & \left[e^{ah} - (1+ah)\right] \\ 0 & e^{ah} \end{bmatrix} \begin{bmatrix} e(n) \\ x^*(n) \end{bmatrix} \tag{4.7}$$

Now, let us consider the situation where $a < 0$, i.e. the ODE is asymptotically stable and $x^*(n) \rightarrow 0$ and $n \rightarrow \infty$ as $\left|e^{ah}\right| < 1$. Thus, we can expect that the approximate solution $\{x(n)\}$ should exhibit similar behavior qualitatively and $e(n) \rightarrow 0$ as $n \rightarrow \infty$. This requires that the difference equation given by (4.7) should be asymptotically stable, i.e., all eigen values of matrix

$$\begin{bmatrix} (1+ah) & \left[e^{ah} - (1+ah)\right] \\ 0 & e^{ah} \end{bmatrix}$$

should have magnitude strictly less than one. Thus, the approximation error $e(n) \rightarrow 0$ as $n \rightarrow \infty$ provided the following condition holds

$$|1+ah| < 1 \Rightarrow -2 < ah < 0 \tag{4.8}$$

This inequality gives constraint on the choice of integration interval h, which will ensure that approximation error will vanish asymptotically.

Following similar line of arguments, we can derive conditions for choosing integration interval for different methods. For example,

- **Implicit Euler**

$$\begin{bmatrix} e(n+1) \\ x^*(n+1) \end{bmatrix} = \begin{bmatrix} \dfrac{1}{(1-ah)} & \left[e^{ah} - \dfrac{1}{(1-ah)}\right] \\ 0 & e^{ah} \end{bmatrix} \begin{bmatrix} e(n) \\ x^*(n) \end{bmatrix} \tag{4.9}$$

$$\left|\frac{1}{1-ah}\right| < 1 \Rightarrow ah < 0 \tag{4.10}$$

- **Trapeziodal Rule (Simpson's method)**

$$\begin{bmatrix} e(n+1) \\ x^*(n+1) \end{bmatrix} = \begin{bmatrix} 1+ah+\dfrac{(ah)^2}{2} & \left[e^{ah} - \dfrac{1+(ah/2)}{1-(ah/2)}\right] \\ 0 & e^{ah} \end{bmatrix} \begin{bmatrix} e(n) \\ x^*(n) \end{bmatrix} \tag{4.11}$$

$$\left|\frac{1+(ah/2)}{1-(ah/2)}\right| < 1 \Rightarrow ah < 0 \tag{4.12}$$

- **2'nd Order Runge Kutta Method**

$$\begin{bmatrix} e(n+1) \\ x^*(n+1) \end{bmatrix} \tag{4.13}$$

$$= \begin{bmatrix} \left(1+ah+\dfrac{(ah)^2}{2}\right) & \left[e^{ah} - \left(1+ah+\dfrac{(ah)^2}{2}\right)\right] \\ 0 & e^{ah} \end{bmatrix} \begin{bmatrix} e(n) \\ x^*(n) \end{bmatrix} \tag{4.14}$$

$$\left|1+ah+\frac{(ah)^2}{2}\right| < 1 \Rightarrow -2 < ah + \frac{(ah)^2}{2} < 0 \tag{4.15}$$

Thus, choice of integration interval depends on the parameters of the equation to be solved and the method used for solving ODE IVP. These simple example also demonstrates that the approximation error analysis gives considerable insight into relative merits of different methods. For example, in the case of implicit Euler or Simpson's rule, the approximation error asymptotically reduces to zero for any choice of $h > 0$. (Of course, larger the value of h, less accurate is the numerical solution.) Same is not true for explicit Euler method. This clearly shows that implicit Euler method and Simpson's rule are superior to explicit Euler method.

It may be noted that we start solving ODE-IVP from a point $x(0) = x^*(0)$ i.e. $e(0) = 0$.

The above analysis can be easily extended to a coupled system of linear ODE-IVP of the form

$$\begin{gathered} \frac{d\mathbf{x}}{dt} = A\mathbf{x} \\ \mathbf{x} = \mathbf{x}(0) \text{ at } t = 0 \\ \mathbf{x} \in R^n \qquad A \equiv (n \times n) \text{ matrix} \end{gathered} \tag{4.16}$$

Following similar arguments as in the scalar case, it can be shown that condition for choosing integration interval are as follows

- **Explicit Euler**

$$\mathbf{x}(n+1) = (I + hA)\mathbf{x}(n) \tag{4.17}$$

$$\rho\left[I + hA\right] < 1 \tag{4.18}$$

where $\rho(.)$ represents spectral radius of the matrix $[I + hA]$. When matrix A is diagonalizable, i.e. $A = \Psi\Lambda\Psi^{-1}$, we can write

$$I + hA = \Psi\left[I + h\Lambda\right]\Psi^{-1} \tag{4.19}$$

and eigen values of matrix $I + hA$ are $\{(1 + h\lambda_i) : i = 1, 2, ..., n\}$ where $\{\lambda_i : i = 1, 2, ..., n\}$ represent eigenvalues of matrix A. Thus, the stability requirement reduces to

$$|1 + h\lambda_i| < 1 \text{ for } i = 1, 2, ..., n \tag{4.20}$$

- **Implicit Euler**

$$\mathbf{x}(n+1) = (I - hA)^{-1}\mathbf{x}(n) \tag{4.21}$$

$$\rho\left[(I - hA)^{-1}\right] < 1 \tag{4.22}$$

- **Trapeziodal Rule:**

$$\mathbf{x}(n+1) = \left(I - \frac{h}{2}A\right)^{-1}\left(I + \frac{h}{2}A\right)\mathbf{x}(n) \tag{4.23}$$

$$\rho\left[\left(I - \frac{h}{2}A\right)^{-1}\left(I + \frac{h}{2}A\right)\right] < 1 \tag{4.24}$$

Similar error analysis (or stability analysis) can be performed for other integration methods. For example, when the 3-step algorithm is used for obtaining the numerical solution of

$$\begin{aligned}
x(n+1) &= \beta_{-1}hf(n+1) + \alpha_0 x(n) + \alpha_1 x(n-1) + \alpha_2 x(n-2) \\
&= a\beta_{-1}hx(n+1) + \alpha_0 x(n) + \alpha_1 x(n-1) + \alpha_2 x(n-2) \\
&= \frac{1}{1 - a\beta_{-1}h} + \left[\alpha_0 x(n) + \alpha_1 x(n-1) + \alpha_2 x(n-2)\right] \\
&= \eta_0 x(n) + \eta_1 x(n-1) + \eta_2 x(n-2) \qquad (4.25)
\end{aligned}$$

The above difference equation can be rearranged in the following form.

$$\begin{bmatrix} x(n-1) \\ x(n) \\ x(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \eta_2 & \eta_1 & \eta_0 \end{bmatrix} \begin{bmatrix} x(n-2) \\ x(n-1) \\ x(n) \end{bmatrix} \tag{4.26}$$

Defining

$$\mathbf{z}(n) = \begin{matrix} x(n-2) \\ x(n-1) \\ x(n) \end{matrix} \ ; \quad \mathbf{z}(n+1) = \begin{bmatrix} x(n-1) \\ x(n) \\ x(n+1) \end{bmatrix} \ ; \quad B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \eta_2 & \eta_1 & \eta_0 \end{bmatrix} \tag{4.27}$$

we have

$$\mathbf{z}(n+1) = B\mathbf{z}(n) \tag{4.28}$$

$$\begin{aligned} x(n+1) &= \mathbf{z}_3(n+1) \\ &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \mathbf{z}(n) = C\mathbf{z}(n) \end{aligned} \tag{4.29}$$

Similarly, the true solution can be expressed as

$$\mathbf{z}^*(n+1) = B^*\mathbf{z}^*(n) \tag{4.30}$$

$$x(n+1) = C\mathbf{z}^*(n) \tag{4.31}$$

where

$$B^* = \begin{bmatrix} e^{ah} & 0 & 0 \\ 0 & e^{ah} & 0 \\ 0 & 0 & e^{ah} \end{bmatrix} \tag{4.32}$$

The evolution of the approximation error is given as

$$\mathbf{e}(n+1) = B\mathbf{e}(n) + [B^* - B]\,\mathbf{z}^*(n) \tag{4.33}$$

$$\mathbf{e}(n) = \mathbf{z}^*(n) - \mathbf{z}(n) \tag{4.34}$$

If the stability criterion that can be used to choose integration interval h can be derived as

$$\rho(B) < 1 \tag{4.35}$$

Note that characteristic equation for matrix B is given as

$$\lambda^3 - \eta_0\lambda^2 - \eta_1\lambda - \eta_2 = 0 \tag{4.36}$$

Thus, eigenvales of matrix B can be directly computed using the coefficients η_0, η_1 and $\eta_{2,}$ which are functions of integration interval h.

Equations such as (4.18), (4.22) and (4.36) can be used to generate *stability envelopes* for each method in the complex plane (eigenvalues of a matrix can be complex). Stability envelopes for most of the methods are available in literature. The following general conclusions can be reached by studying these plots [**6**].

- Even though the first and second order Adams-Moulton methods (implicit Euler and Crank-Nicholson) are A-stable, the higher order techniques have restricted regions of stability. These regions are larger than the Adams-Bashworth family of the same order.
- All forms of the R-K algorithms with order ≤ 4 have identical stability envelopes.
- Explicit R-K techniques have better stability characteristics than explicit Euler.
- For predictor-corrector schemes, accuracy of scheme improves with order. However, stability region shrinks with order.

Remark 6. *The conclusions reached from studying linear systems can be extended to general nonlinear systems locally using Taylor expansion.*

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}) \tag{4.37}$$

can be approximated as

$$\begin{aligned}
\frac{d\mathbf{x}}{dt} &\cong F(\mathbf{x}(n)) + \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}(n)} (\mathbf{x} - \mathbf{x}(n)) && (4.38)\\
&\cong \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}(n)} \mathbf{x} + \left[F\left[\mathbf{x}(n)\right] - \left[\frac{\partial F}{\partial \mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}(n)} \mathbf{x}(n)\right] && (4.39)\\
&\cong (A)_n \mathbf{x} + (\mathbf{d})_n && (4.40)
\end{aligned}$$

Applying some numerical technique to solve this problem will lead to

$$\mathbf{x}(n+1) = (B)_n \mathbf{x}(n+1) + (\mathbf{c})_n \tag{4.41}$$

and stability will depend on the choice of h such that $\rho[(B)_n] < 1$ for all n. Note that, it is difficult to perform global analysis for general nonlinear systems.

5. Summary

In these lecture notes, we undertake the study of solutions of multivariable and coupled ODE-IVPs. To begin with, we show that variety of problems, such as solving ODE-BVP, hyperbolic / parabolic PDEs or set of nonlinear algebraic equations, can be reduced to solving ODE-IVP. A special class of problems, i.e. solutions of coupled linear ODE-IVPs can be solved analytically. Thus, before we start the development of the numerical methods, we will develop analytical solutions for unforced (homogeneous) linear ODE-IVP problem and investigate their asymptotic behavior using eigenvalue analysis. We later discuss development of numerical algorithms based on

- Taylor series approximations (Runge-Kutta methods)

- Polynomial approximation based algorithms (Predictor-corrector type methods).

In the end, we provide a brief introduction to the stability analysis of the numerical algorithms for solving ODE-IVPs.

6. Exercise

(1) Express the following set of equations in the standard form

$$dx/dt = Ax; \;\; x(0) = x^{(0)}$$

and solve the resulting initial value problem analytically

(a) Set 1

$$d^2y/dt^2 + 4dy/dt + 3y = 0; \;\; y(0) = 1; dy/dt = 0 \text{ at } t = 0$$

(b) Set 2

$$d^3y/dt^3 + 6d^2y/dt^2 + 11dy/dt + 6y = 0$$

$$y(0) = 1; \;\; dy/dt = d^2y/dt^2 = 0 \text{ at } t = 0;$$

(c) Set 3

$$dy/dt + 3y + z = 0; y(0) = 1$$

$$d^2z/dt^2 + 3dz/dt + 2z = 0$$

$$z(0) = 1; dz/dt = 0$$

Compare the coefficients of the characteristic equation, i.e. $det(\lambda I - A) = 0$, and those of the ODE(s) for the first two sets. Also, comment upon the asymptotic behavior of the solution in each case based on eigenvalues of matrix A.

(2) Consider the dynamic model of three isothermal CSTRs in series (Example 1). The model parameters are Residences time values: $\tau_1 = 1$ min $\tau_1 = 2$ min $\tau_3 = 3$ min and the reaction rate constant $k = 0.5$ (min^{-1})

(a) Assuming that the CSTR is at a steady state initially (i.e., $dc/dt = 0$) with $C_{A0} = 1.8$, find the corresponding steady state concentration by solving the resulting linear algebraic equations.

(b) Suppose, we decide to shutdown the reactor system and reduce $C_{A0} = 0$ at $t = 0$. Integrate the resulting set of homogeneous ODEs analytically to obtain the concentration profile $C(t)$, starting with the steady state obtain above.

(c) Use explicit Euler method to integrate the above set of equations from t =0 to t = 2 with integration interval of 0.1, 0.25, 0.5, 1.0 and compare the approximate solutions with the corresponding analytical solution in each case.

(d) Repeat (c) for the case h=0.25 using implicit Euler method.

(3) Consider the PDE given below

$$\partial C/\partial t = \partial^2 C/\partial z^2$$

$$C(0,t) = C(1,t) = 0 \text{ for all } 0 \leq t \leq \infty$$

$$C(z,0) = 1 \text{ for } 0 \leq z \leq 1$$

(a) Use the finite difference technique on the dimensionless diffusion equation obtain a set of ODE-IVPs assuming N internal grid points. Particularly for the case $N = 3$, obtain the analytical solution to the resulting ODE-IVP.

(b) Repeat the above exercise using orthogonal collocation to discretize in space with two internal collocation points.

(4) Consider Van der Pol equation given below

$$d^2y/dt^2 - (1-y^2)dy/dt + 3y = 0$$

$$y(0) = 2; dy/dt = 0 \; at \; t = 0$$

(a) Express the above ODE-IVP in standard form

$$d\mathbf{x}/dt = F(\mathbf{x});\; \mathbf{x} = \mathbf{x}(0) \text{ at } t = 0$$

(b) Linearize the resulting equation in the neighborhood of $x = \begin{bmatrix} 0 & 0 \end{bmatrix}$ and obtain the perturbation solution analytically. Comment upon the asymptotic behavior of the solution.

(5) Consider the quadruple tank setup shown in Figure 1.

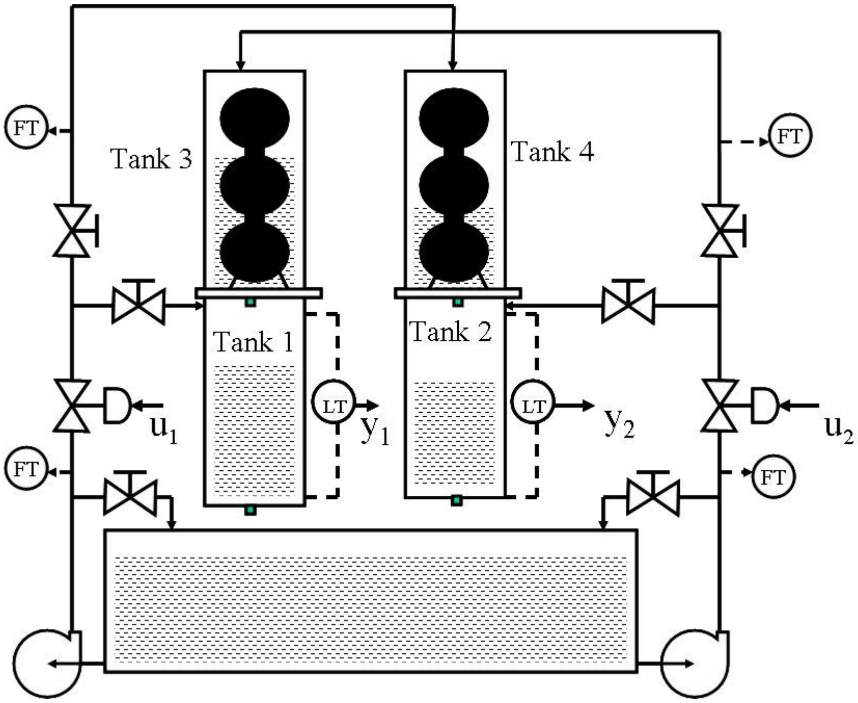


FIGURE 1. Quadruple tank setup: Schematic diagram

CHAPTER 5

Optimization and Related Numerical Schemes

1. Introduction

These lecture notes deal with multivariable unconstrained optimization techniques and their application to computing numerical solutions of various types of problems. One of the major application of unconstrained optimization is model parameter estimation (function approximation or multivariate regression). Thus, we begin by providing a detailed description of the model parameter estimation problem. We then derive necessary and sufficient conditions for optimality for a general multivariable unconstrained optimization problem. If the model has a *nice* structure, such as it is linear in parameters or can be transformed to a linear in parameter form, then the associated parameter estimation problem can be solved analytically. The parameter estimation of linear in parameter models (multivariate linear regression) problem is treated next. Geometric and statistical properties of the linear least square problem are discussed in detail to provide further insights into this formulation. Numerical methods for estimating parameters of the nonlinear-in-parameter models are presented in subsequent section. The applications of optimization formulations for solving problems such as solving linear/ nonlinear equations and solution of PDEs using finite element method are discussed in the last section.

2. Principles of Optimization

2.1. Necessary Conditions for Optimality. Given a real valued scalar function $F(\mathbf{z}) : R^n \rightarrow R$ defined for any $\mathbf{z} \in R^n$.

Definition 29. ***(Global Minimum):*** *If there exists a point* $\mathbf{z}^* \in R^n$ *such that* $F(\mathbf{z}^*) < F(\mathbf{z})$ *for any* $\mathbf{z} \in R^{\mathbf{N}}$, *then* $\mathbf{z}^*$ *is called as global minimum of* $F(\mathbf{z})$.

Definition 30. ε*-neighborhood of a point* $\mathbf{a}$ *be defined as the set* $N_e(\mathbf{a}) = \{\mathbf{z} : \|\mathbf{z} - \mathbf{a}\| \leq \varepsilon\}$

Definition 31. ***(Local Minimum)*** *: If there exists an* $\varepsilon-$*neighborhood* $N_C(\mathbf{z}^*)$ *round* $\mathbf{z}^*$ *such that* $F(\mathbf{z}^*) < F(\mathbf{z})$ *for each* $\mathbf{z} \in N_e(\mathbf{z})$, *then* $\mathbf{z}^*$ *is called local minimum.*

Theorem 11. *If $F(\mathbf{z})$ is continuous and differentiable and has an extreme (or stationary) point (i.e. maximum or minimum) point at $\mathbf{z} = \mathbf{z}^*$, then*

$$\nabla F(\mathbf{z}^*) = \left[\frac{\partial F}{\partial z_1} \ \ \frac{\partial F}{\partial z_2} .... \frac{\partial F}{\partial z_{\mathbf{N}}}\right]^T_{\mathbf{z}=\mathbf{z}^*} = \overline{0} \tag{2.1}$$

.

Proof: Suppose $\mathbf{z} = \mathbf{z}^*$ is a minimum point and one of the partial derivatives, say the k^{th} one, does not vanish at $\mathbf{z} = \mathbf{z}^*$, then by Taylor's theorem

$$F(\mathbf{z}^* + \Delta\mathbf{z}) = F(\mathbf{z}^*) + \sum_{i=1}^{\mathbf{N}} \frac{\partial F}{\partial z_i}(\mathbf{z}^*)\Delta z_i + R_2(\mathbf{z}^*, \Delta\mathbf{z}) \tag{2.2}$$

$$i.e. \ \ F(\mathbf{z}^* + \Delta\mathbf{z}) - F(\mathbf{z}^*) = \Delta z_k \frac{\partial F}{\partial z_i}(\mathbf{z}^*) + R_2(\mathbf{z}^*, \Delta\mathbf{z}) \tag{2.3}$$

Since $R_2(\mathbf{z}^*, \Delta\mathbf{z})$ is of order $(\Delta z_i)^2$, the terms of order Δz_i will dominate over the higher order terms for sufficiently small $\Delta\mathbf{z}$. Thus, sign of $F(\mathbf{z}^*+\Delta\mathbf{z})-F(\mathbf{z}^*)$ is decided by sign of

$$\Delta z_k \frac{\partial F}{\partial z_k}(\mathbf{z}^*)$$

Suppose,

$$\frac{\partial F}{\partial z_k}(\mathbf{z}^*) > 0 \tag{2.4}$$

then, choosing $\Delta z_k < 0$ implies

$$F(\mathbf{z}^* + \Delta\mathbf{z}) - F(\mathbf{z}^*) < 0 \Rightarrow F(\mathbf{z}^* + \Delta\mathbf{z}) < F(\mathbf{z}^*) \tag{2.5}$$

and $F(\mathbf{z})$ can be further reduced by reducing Δz_k. This contradicts the assumption that $\mathbf{z} = \mathbf{z}^*$ is a minimum point. Similarly, if

$$\frac{\partial F}{\partial z_k}(\mathbf{z}^*) < 0 \tag{2.6}$$

then, choosing $\Delta z_k > 0$ implies

$$F(\mathbf{z}^* + \Delta\mathbf{z}) - F(\mathbf{z}^*) < 0 \Rightarrow F(\mathbf{z}^* + \Delta\mathbf{z}) < F(\mathbf{z}^*) \tag{2.7}$$

and $F(\mathbf{z})$ can be further reduced by increasing Δz_k. This contradicts the assumption that $\mathbf{z} = \mathbf{z}^*$ is a minimum point. Thus, $\mathbf{z} = \mathbf{z}^*$ will be a minimum of $F(\mathbf{z})$ only if

$$\frac{\partial F}{\partial z_k}(\mathbf{z}^*) = 0 \quad For \quad k = 1, 2, ...\mathbf{N} \tag{2.8}$$

Similar arguments can be made if $\mathbf{z} = \mathbf{z}^*$ is a maximum of $F(\mathbf{z})$.

2.2. Sufficient Conditions for Optimality. Before we prove sufficient condition for optimality, we revise some relevant definitions from linear algebra.

DEFINITION 32. ***(Positive Definite Matrix)*** *A $n \times n$ matrix A is called positive definite if for every $\mathbf{z} \in R^n$*

$$\mathbf{z}^T A\mathbf{z} > 0 \tag{2.9}$$

whenever $\mathbf{z} \neq \overline{0}$.

DEFINITION 33. ***(Positive Semi-definite Matrix)*** *A $n \times n$ matrix A is called positive semi-definite if for every $\mathbf{z} \in R^n$ we have*

$$\mathbf{z}^T A\mathbf{z} \geq 0 \tag{2.10}$$

DEFINITION 34. .***(Negative Definite Matrix)*** *A $n \times n$ matrix A is called negative definite if for every $\mathbf{z} \in R^n$*

$$\mathbf{z}^T A\mathbf{z} < 0 \tag{2.11}$$

whenever $\mathbf{z} \neq \overline{0}$.

DEFINITION 35. ***(Negative Semi-definite Matrix)*** *A $n \times n$ matrix A is called negative semi-definite if for every $\mathbf{z} \in R^n$ we have*

$$\mathbf{z}^T A\mathbf{z} \leq 0 \tag{2.12}$$

The sufficient condition for optimality, which can be used to establish whether a stationary point is a maximum or a minimum, is given by the following theorem.

THEOREM 12. *A sufficient condition for a stationary point $\mathbf{z} = \mathbf{z}^*$ to be an extreme point is that matrix $\left[\dfrac{\partial^2 F}{\partial z_i \partial z_j}\right]$ (Hessian of F)evaluated at $\mathbf{z} = \mathbf{z}^*$ is*

(1) positive definite when $\mathbf{z} = \mathbf{z}^*$ is minimum
(2) negative definite when $\mathbf{z} = \mathbf{z}^*$ is maximum

Proof: Using Taylor series expansion, we have

$$F(\mathbf{z}^* + \Delta\mathbf{z}) = F(\mathbf{z}^*) + \sum_{i=1}^{\mathbf{N}} \frac{\partial F}{\partial z_i}(\mathbf{z}^*)\Delta\mathbf{z} + \frac{1}{2!}\sum_{i=1}^{\mathbf{N}}\sum_{j=1}^{\mathbf{N}} \frac{\partial^2 F(\mathbf{z}^* + \lambda\Delta\mathbf{z})}{\partial z_i \partial z_j}\Delta z_i \Delta z_j$$

$$(0 < \lambda < 1) \tag{2.13}$$

.Since $\mathbf{z} = \mathbf{z}^*$ is a stationary point we have

$$\nabla F(\mathbf{z}^*) = \overline{0} \tag{2.14}$$

Thus, above equation reduces to

$$F(\mathbf{z}^{*}+\Delta\mathbf{z})-F(\mathbf{z}^{*}) \;=\; \frac{1}{2!}\sum_{i=1}^{\mathbf{N}}\sum_{j=1}^{\mathbf{N}}\frac{\partial^{2}F(\mathbf{z}^{*}+\lambda\Delta\mathbf{z})}{\partial z_{i}\partial z_{j}}\Delta z_{i}\Delta z_{j} \qquad (0 \;<\; \lambda<1) \tag{2.15}$$

This implies that sign of $F(a+\Delta z)-F(a)$at extreme point $\mathbf{z}^{*}$ is same as sign of R.H.S. Since the 2'nd partial derivative $\left[\dfrac{\partial^{2}F}{\partial z_{i}\partial z_{j}}\right]$ is continuous in the neighborhood of $\mathbf{z}=\mathbf{z}^{*}$, its value at $\mathbf{z}=\mathbf{z}^{*}+\lambda\Delta\mathbf{z}$ will have same sign as its value at $\mathbf{z}=\mathbf{z}^{*}$ for all sufficiently small $\Delta\mathbf{z}$. If the quantity

$$\sum_{i=1}^{\mathbf{N}}\sum_{j=1}^{\mathbf{N}}\frac{\partial^{2}F(\mathbf{z}^{*}+\lambda\Delta z)}{\partial z_{i}\partial z_{j}}\Delta z_{i}\Delta z_{j}\simeq(\Delta\mathbf{z})^{T}[\nabla^{2}F(\mathbf{z}^{*})]\Delta\mathbf{z}\geq 0 \tag{2.16}$$

for all $\Delta\mathbf{z}$, then $\mathbf{z}=\mathbf{z}^{*}$ will be a local minimum. In other words, if Hessian matrix $[\nabla^{2}F(\mathbf{z}^{*})]$ **positive semi-definite,** then $\mathbf{z}=\mathbf{z}^{*}$ will be a local minimum. If the quantity

$$\sum_{i=1}^{\mathbf{N}}\sum_{j=1}^{\mathbf{N}}\frac{\partial^{2}F(\mathbf{z}^{*}+\lambda\Delta z)}{\partial z_{i}\partial z_{j}}\Delta z_{i}\Delta z_{j}\simeq(\Delta\mathbf{z})^{T}[\nabla^{2}F(\mathbf{z}^{*})]\Delta\mathbf{z}\leq 0 \tag{2.17}$$

for all $\Delta\mathbf{z}$, then $\mathbf{z}=\mathbf{z}^{*}$ will be a local maximum. In other words, if Hessian matrix $[\nabla^{2}F(\mathbf{z}^{*})]$ **negative semi-definite,** then $\mathbf{z}=\mathbf{z}^{*}$ will be a local maximum.

REMARK 7. *It should be noted that the need to define a* positive definite *or* negative definite *matrix naturally arises from the geometric considerations while qualifying a stationary point in multi-dimensional optimization. When the Hessian is positive definite, a better insight into the local geometry can be obtained by plotting function*

$$q(\Delta\mathbf{z})=(\Delta\mathbf{z})^{T}[\nabla^{2}F(\mathbf{z}^{*})]\Delta\mathbf{z}=1 \tag{2.18}$$

To begin with, consider a special case when $z\in R^{2}$ and $[\nabla^{2}F(\mathbf{z}^{*})]=I$. In this case

$$(\Delta\mathbf{z})^{T}[\nabla^{2}F(\mathbf{z}^{*})]\Delta\mathbf{z}=(\Delta z_{1})^{2}+(\Delta z_{2})^{2}=1 \tag{2.19}$$

which represents a unit circle in the neighborhood of $\mathbf{z}=\mathbf{z}^{*}$. In R^{3}, $(\Delta\mathbf{z})^{T}\Delta\mathbf{z}=1$ represents a sphere in the neighborhood of $\mathbf{z}=\mathbf{z}^{*}$. Now, suppose we consider

$$[\nabla^{2}F(\mathbf{z}^{*})]=diag\begin{bmatrix}4 & 1 & 1/9\end{bmatrix}$$

then

$$(\Delta \mathbf{z})^T[\nabla^2 F(\mathbf{z}^*)]\Delta \mathbf{z} = 4\,(\Delta z_1)^2 + (\Delta z_2)^2 + (1/9)\,(\Delta z_3)^2 = 1 \tag{2.20}$$

As the coefficients of quadratic terms are unequal and positive, we get an ellipsoid instead of a sphere in the neighborhood of $\mathbf{z} = \mathbf{z}^*$.

Now, suppose we consider a dense and positive definite $[\nabla^2 F(\mathbf{z}^*)]$ *such as*

$$[\nabla^2 F(\mathbf{z}^*)] = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \tag{2.21}$$

then, we have

$$(\Delta \mathbf{z})^T[\nabla^2 F(\mathbf{z}^*)]\Delta \mathbf{z} = 5\,(\Delta z_1)^2 + 8\,(\Delta z_1 \Delta z_2) + 4\,(\Delta z_2)^2 = 1 \tag{2.22}$$

This is still an ellipse in the neighborhood of $\mathbf{z} = \mathbf{z}^*$, *however, its axis are not aligned parallel to the coordinate axis. Matrix* $[\nabla^2 F(\mathbf{z}^*)]$ *can be diagonalized as*

$$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T \tag{2.23}$$

Defining a rotated coordinates $\Delta \mathbf{y}$ *as*

$$\Delta \mathbf{y} = \begin{bmatrix} \Delta y_1 \\ \Delta y_2 \end{bmatrix} = \Psi^T \Delta \mathbf{z} \tag{2.24}$$

$$= \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T \begin{bmatrix} \Delta z_1 \\ \Delta z_2 \end{bmatrix} \tag{2.25}$$

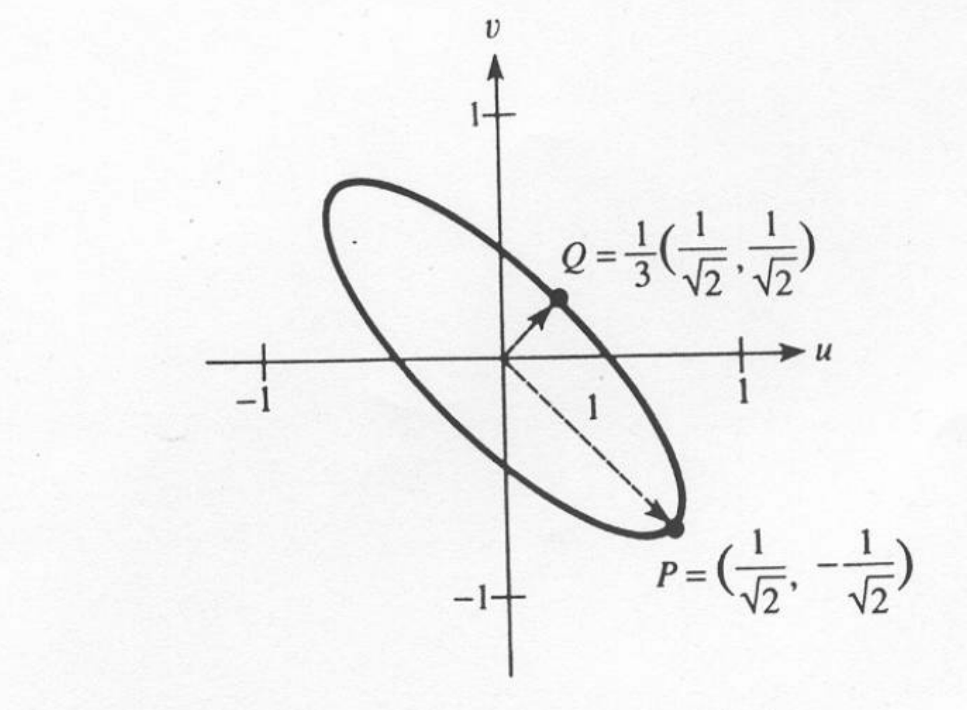
we have

$$\begin{aligned} (\Delta \mathbf{z})^T[\nabla^2 F(\mathbf{z}^*)]\Delta \mathbf{z} &= \Delta \mathbf{y}^T \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \Delta \mathbf{y} \\ &= (\Delta y_1)^2 + 9\,(\Delta y_2)^2 = 1 \qquad (2.26) \end{aligned}$$

Figure 1 shows ellipsoid when $\mathbf{z}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. *Note that the coordinate transformation* $\Delta \mathbf{y} = \Psi^T \Delta \mathbf{z}$ *has rotated the axis of the space to match the axes of the ellipsoid. Moreover, the major axis is aligned along the eigenvector corresponding to the largest magnitude eigenvalue and the minor axis is aligned along the smallest magnitude eigenvalue.*

In more general settings, when $z \in R^n$, *let* $0 < \lambda_1 \leq \lambda_2 \leq .... \leq \lambda_n$ *represent eigenvalues of the Hessian matrix. Using the fact that Hessian is positive definite, we can write*

$$[\nabla^2 F(\mathbf{z}^*)] = \Psi \Lambda \Psi^T \tag{2.27}$$


v
1
$Q = \frac{1}{3}(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$
u
−1
1
1
$P = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$
−1

The ellipse $5u^2 + 8uv + 5v^2 = 1$ and its principal axes.

FIGURE 1

where

$$\Psi = \left[\begin{array}{cccc} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & .... & \mathbf{v}^{(n)} \end{array} \right]$$

represents unitary matrix with eigenvectors of the Hessian as its columns and Λ *is a diagonal matrix with eigenvalues on the diagonal. Using this transformation, we have*

$$q(\Delta\mathbf{z}) = (\Psi^T\Delta\mathbf{z})^T\Lambda\left(\Psi^T\Delta\mathbf{z}\right) = (\Delta\mathbf{y})^T\Lambda\,(\Delta\mathbf{y}) \tag{2.28}$$

$$= \lambda_1\left(\Delta y_1\right)^2 + \lambda_2\left(\Delta y_2\right)^2 + ...\lambda_n\left(\Delta y_n\right)^2 = 1 \tag{2.29}$$

From the above expression, it is easy to see that $q(\Delta\mathbf{y})$ *is an ellipsoid in* n *dimensions with its major axis aligned along the eigenvector* $\mathbf{v}^{(n)}$ *and the minor axis along the eigenvector* $\mathbf{v}^{(1)}$. *Thus, function* $F(\mathbf{z})$ *looks like an ellipsoid in the neighborhood of* $\mathbf{z} = \mathbf{z}^*$ *when* $\mathbf{z}^*$ *is a minimum.*

REMARK 8. *Whether a matrix is positive (semi)definite, negative (semi) definite or indefinite can be established using eigen values of the matrix. If eigenvalues of a matrix are all real positive (i.e.* $\lambda_i \geq 0$ *for all* i*) then, the matrix is positive semi-definite. If eigenvalues of a matrix are all real negative* *(i.e. $\lambda_i \leq 0$ for all i) then, the matrix is negative semi-definite. When eigen values have mixed signs, the matrix is indefinite.*

3. Model Parameter Estimation

3.1. Mathematical Models in Engineering.

Mathematical modeling is an elegant tool for describing various processes and phenomena occurring in a processing plant or a manufacturing system. Mathematical models play important role in design and scaling of a new process or understanding static/dynamic behavior of an existing plant. Typically, such models are **gray box** models **and** are developed by judiciously combining first principles (i.e. energy, momentum and material balances) with semi-empirical correlations developed from experimental data. As a consequence, such models involve a number of parameters, such as heat and mass transfer coefficients, parameters, reaction kinetics, correlation coefficients etc., which have to be estimated from the experimental data. In general, these models can be expressed in abstract form as

$$y = f(\mathbf{x}, \boldsymbol{\theta})$$

where $\mathbf{x} \in \mathbf{R}^m$ represents vector of independent variables (e.g.. temperature, pressure, concentration, current, voltage etc.) and let $y \in R$ denotes dependent variable, $f(.)$ represents proposed functional relationship that relates y with $\mathbf{x}$ and $\boldsymbol{\theta} \in R^l$ represent vector of model parameters.

Example 59. ***Correlations***

(1) *Specific heat capacity at constant pressure ($C_{p,}$), as a function of temperature*

$$C_P = a + bT + cT^2 \tag{3.1}$$

$$y \equiv C_p \;\; ; \;\; \mathbf{x} \equiv T \;\; ; \boldsymbol{\theta} \equiv \begin{bmatrix} a & b & c \end{bmatrix}^T$$

(2) *Dimensionless analysis is mass transfer / heat transfer*

$$Sh = \alpha_0 \, Re^{\alpha_1} \; Sc^{\alpha_2} \tag{3.2}$$

$$y = Sh \quad ; \quad \mathbf{x} = [Re \;\; Sc]^T \quad ; \quad \boldsymbol{\theta} \equiv \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \end{bmatrix}^T$$

$$Nu = \alpha_0 Re^{\alpha_1} Pr^{\alpha_2} \left(\mu_a/\mu_p\right)^{\alpha_3} \tag{3.3}$$

$$y = Nu \quad ; \quad \mathbf{x} = [Re \;\; \mathrm{Pr}]^T \quad ; \quad \boldsymbol{\theta} \equiv \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}^T$$

(3) *Friction factor as a function of Reynold's number for turbulent flow*

$$1/\sqrt{f} = \alpha \log(\mathrm{Re}\sqrt{f}) - \beta \tag{3.4}$$

$$y = f \;\; ; \quad \mathbf{x} = \mathrm{Re} \;\; ; \; \boldsymbol{\theta} \equiv \begin{bmatrix} \alpha & \beta \end{bmatrix}^T$$

(4) *Equation(s) of state: e.g. Redlish Kwong equation*

$$P = \frac{RT}{V-b} - \frac{a}{T^{1/2}(V+b)V} \tag{3.5}$$

$$y = P \;\; ; \quad \mathbf{x} = \begin{bmatrix} T & V \end{bmatrix}^{\mathbf{T}} \; ; \; \boldsymbol{\theta} \equiv \begin{bmatrix} a & b \end{bmatrix}^T$$

or Van der Waals equation

$$(P + \frac{a}{V^2})(V-b) = RT \tag{3.6}$$

$$y = P \;\; ; \quad \mathbf{x} = \begin{bmatrix} T & V \end{bmatrix}^{\mathbf{T}} \; ; \; \boldsymbol{\theta} \equiv \begin{bmatrix} a & b \end{bmatrix}^T$$

(5) *Antonie equation for estimating vapor pressure of a pure component*

$$\log(P_v) = A - \frac{B}{T+C} \tag{3.7a}$$

EXAMPLE 60. *Reaction rate models:*

$$-r_A = -\left(\frac{dC_A}{dt}\right) = k_o \exp(-E/RT) \;\; (C_A)^n \tag{3.8}$$

$$y \equiv -r_A \, ; \quad \mathbf{x} \equiv [C_A \; T] \;^T ; \quad \boldsymbol{\theta} \equiv \begin{bmatrix} n & E & k_o \end{bmatrix}^T$$

EXAMPLE 61. *Step response of a second order process (dynamic modeling for process control)*

$$\delta T(t) = K\left[1 + \alpha_1 \exp(-t/\tau_1) + \alpha_2 \exp(-t/\tau_2)\right] \Delta F_c \tag{3.9}$$

$$y = \delta T \;\; ; \;\; x \equiv t \;\; ; \;\; \boldsymbol{\theta} = \begin{bmatrix} K & \alpha_1 & \alpha_2 & \tau_1 & \tau_2 \end{bmatrix}$$

Here δT represents deviation temperature obtained in response to step change in cooling water flow rate of magnitude ΔF_c. This is a nonlinear in parameter model.

3.2. Classification of Models. Based on the manner in which the parameters appear in model equations, we can categorize the model as follows:

- **Linear in parameter models**: The most common type of approximation considered is from the class of functions

$$\widehat{y} \ = \ \theta_1 f_1(\mathbf{x}) + \theta_2 f_2(\mathbf{x}) + ... + \theta_m f_m(\mathbf{x}) \tag{3.10}$$

  As the parameters $\theta_1, ...\theta_m$ appear linearly in the model, the model is called as linear in parameter model. Note that $f_i(\mathbf{x})$ can be nonlinear functions of $\mathbf{x}$. More commonly used linear forms are

  - Simple polynomials

$$\widehat{y} = \theta_1 + \theta_2 x + \theta_3 x^2 + ... + \theta_m x^{m-1}$$

  - Legendre polynomials

$$\widehat{y} = \theta_1 L_0(x) + \theta_2 L_1(x) + ... + \theta_m L_{m-1}(x) \tag{3.11}$$

  - Chebysheve polynomials

$$\widehat{y} = \theta_1 T_0(x) + \theta_2 T_1(x) + ... + \theta_m T_{m-1}(x) \tag{3.12}$$

  - Fourier series

$$\widehat{y} = \theta_1 \sin(\omega x) + \theta_2 \sin(2\omega x) + ... + \theta_m \sin(m\omega x) \tag{3.13}$$

  - Exponential form with $\alpha_1....\alpha_m$ specified

$$\widehat{y} = \theta_1 e^{\alpha_1 x} + \theta_2 e^{\alpha_2 x} + ... + \theta_m e^{\alpha_m x} \tag{3.14}$$

- **Nonlinear in parameter models:** In many problems the parameters appear nonlinearly
  in the model, i.e.

$$\widehat{y} \ = f \ \ (\mathbf{x} \ ; \ \theta_1, .., \theta_m) \quad ; \quad (i = 1, ...\mathbf{N}) \tag{3.15}$$

  where f is a nonlinear function of parameters $\theta_1 ..., \theta_m$.

EXAMPLE 62. • ***Linear and Nonlinear in Parameter Models***

- *C_p as a function of temperature (equation 3.1) is a linear in parameter model.*
- *Reaction rate model (equation 3.8), model for friction factor (equation 3.4), Antonie equation (equation 3.7a), second order dynamics (equation 3.9), heat and mass transfer correlations (equations 3.3 and 3.2) are examples of nonlinear in parameter models. However, some of these*

models can be transformed to linear in parameter models. For example, the transformed reaction rate model

$$\log(-r_A) = \log(k_o) + n\log C_A - \frac{E}{R}\left(\frac{1}{T}\right)$$

(1)

Example 63. • *is a linear in parameter model.*

3.3. Formulation of Parameter Estimation Problem. Estimation of model parameter from experimental data is not an easy task as the data obtained from experiments is always influenced by uncertain variation of uncontrollable and unknown variables, which occur while conducting experiments and collecting data. In modeling parlance, the data is corrupted with *unmeasured disturbances* and *measurement errors.* For example,

- If we measure flow, pressure, temperature etc., through electronic transmitters, there are errors or **noise** in the measurements due to local electrical disturbances.
- While conducting experiments involving heating with a steam coil, unmeasured fluctuations in steam header pressure may introduce variations in the rate of heating

In any experimental evaluation, we can list many such unknown factors which influence the data. Apart from these influences, the proposed mathematical models are often approximate descriptions of underlying phenomenon and additional errors are introduced due to limitations imposed by modeling assumptions. Thus, when we develop a model from experimental data, we can identify three possible sources of error

- **Measurement errors :** Errors in measurements of various recorded variables
- **Unmeasured disturbances:** Unrecorded influences
- **Modeling Errors :** Errors arising due to fact that the model equation(s) represents only an approximate description of the reality.

When we develop mathematical models using data corrupted with measurement errors and unmeasured disturbances, it becomes necessary to characterize the unknown component in the data for estimating the model parameters accurately. Let $\mathbf{x}_t \in R^n$ denote a vector of *true values* of independent variables (e.g.. temperature, pressure, concentration, current, voltage etc.) and let $y_t \in R$ denote the true value of dependent variable. This can be expressed as

$$y_t = F_T(\mathbf{x}_t, \mathbf{\Theta})$$

where $F_T(.)$ represents true functional relationship that relates y_t with $\mathbf{x}_t$ and $\mathbf{\Theta}$ represent the parameter vector. When we collect data from an experiment, we get a set of measurements $\mathbf{x}^{(k)}$ and y_k such that

$$\begin{aligned}\mathbf{x}^{(k)} &= \mathbf{x}_t^{(k)}+\boldsymbol{\varepsilon}^{(k)} \\ y_k &= y_{t,k}+v_k \\ k &= 1,2,......N\end{aligned}$$

where $\boldsymbol{\varepsilon}^{(k)}$ and $v^{(k)}$ represent errors in measurements of independent and dependent variables, respectively, and N represents the size of data sample. Given these measurements, the model relating these measured quantities can be stated as

$$y_{t,k} = F\left(\mathbf{x}_t^{(k)},\boldsymbol{\theta}\right)+e_k$$

where $f(.)$ represents the proposed approximate functional relationship, $\boldsymbol{\theta}$ represent the parameter vector and e_k represents the *equation error* for k'th sample point. Thus, the most general problem of estimating model parameters from experimental data can be stated as follows:

Estimate of $\boldsymbol{\theta}$ such that

$$\begin{matrix}\min \\ \boldsymbol{\theta}\end{matrix} \; f\left(e_1,e_2,.....e_N,\boldsymbol{\varepsilon}^{(1)},....\boldsymbol{\varepsilon}^{(N)},v_1,...v_N\right)$$

subject to

$$\begin{aligned}e_i &= y_i - F\left[\mathbf{x}_t^{(i)},\boldsymbol{\theta}\right] \\ \boldsymbol{\varepsilon}^{(i)} &= \mathbf{x}^{(i)} - \mathbf{x}_t^{(i)} \\ v_i &= y_i - y_{t,i} \\ &\text{for } \; i=1,2,....N\end{aligned}$$

where, $f\,(.)$ represents some scalar objective function.

Given a data set, formulation and solution of the above general modeling problem is not an easy task. In these lecture notes, we restrict ourselves to a special class of models, which assume that

(1) Measurement errors in all independent variables are negligible i.e. $\mathbf{x} = \mathbf{x}_T$
(2) Effect of all unknown disturbances and modeling errors can be adequately captured using equation error model, i.e.

$$y = F\left(\mathbf{x},\boldsymbol{\theta}\right)+e$$

The term e on R.H.S. can be interpreted either as modeling error or as measurement error while recording y.

These assumption considerably simplifies problem formulation. Under these assumptions, the model parameter estimation problem can be stated as estimation of $\boldsymbol{\theta}$ such that

$$\underset{\boldsymbol{\theta}}{\min}\ f\left(e_1, e_2, ...e_N\right)$$

$$e_i \;=\; y_i - \widehat{y}\left[\mathbf{x}^{(i)}, \boldsymbol{\theta}\right] \quad ; \quad i = 1, 2, ...N$$

3.4. Interpolation and Approximation. Problem definition : Let a set $\{\mathbf{x}^{(i)} : i = 1, ...N\}$ of variable x corresponds to a set $\{y_i : i = 1, ...\mathbf{N}\}$ of variable y and let

$$\widehat{y} = f\left(\mathbf{x}, \theta_1, ...\theta_m\right) \tag{3.16}$$

represent the proposed model where f is a continuous function linear in parameters. Substituting x_i we get

$$\widehat{y_i} = \ f\left(\mathbf{x}^{(i)}, \theta_1, ...\theta_m\right) \qquad (\ i = 1, 2......\mathbf{N}) \tag{3.17}$$

a set of $\mathbf{N}$ linear equations in m unknowns.

Case $(\mathbf{N} = m)$: If $\mathbf{N}$ equations are independent then a unique solution exists and we can obtain the function $\widehat{y} = f\left(\mathbf{x}, \theta_1, ...\theta_m\right)$ passing through each of these $\mathbf{N}$ points i.e. we have done an interpolation.

Case $(\mathbf{N} > m)$: When we have an overdetermined set of equations , in general there is neither a solution nor a function $f\left(\mathbf{x}, \theta_1, ...\theta_m\right)$ passing through all points. Define the errors e_i as

$$e_i \;=\; y_i - \widehat{y_i} \qquad for \quad i = 1, 2, ...\mathbf{N} \tag{3.18}$$

$$\;=\; y_i - f\left(\mathbf{x}^{(i)}, \theta_1, ...\theta_m\right) \tag{3.19}$$

This is a system of $\mathbf{N}$ linear equations in $(m + \mathbf{N})$ unknowns, namely m parameters θ_j and the $\mathbf{N}$ error terms e_i. Such a system has infinite solutions. Among these, we choose the one that minimizes

$$\|\ e\ \|_2 = \sum_{i=1}^{\mathbf{N}} e_i^2\ w_i \tag{3.20}$$

Mathematically, we have two different problems for a given set of $\mathbf{N}$ pairs (x_i, y_i) with m parameters.

- **Interpolation** $(m = \mathbf{N})$: Interpolation is used in the case of ‘precisely’ known values (x_i, y_i) and when it is required to predict y at points x other than support points x_i. e.g. when using logarithmic tables.

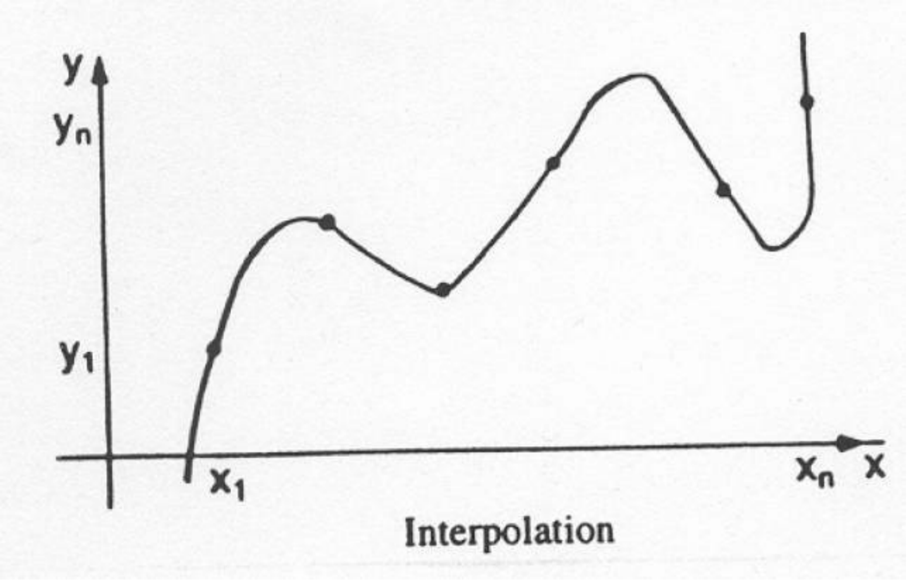


FIGURE 2. Interpolation

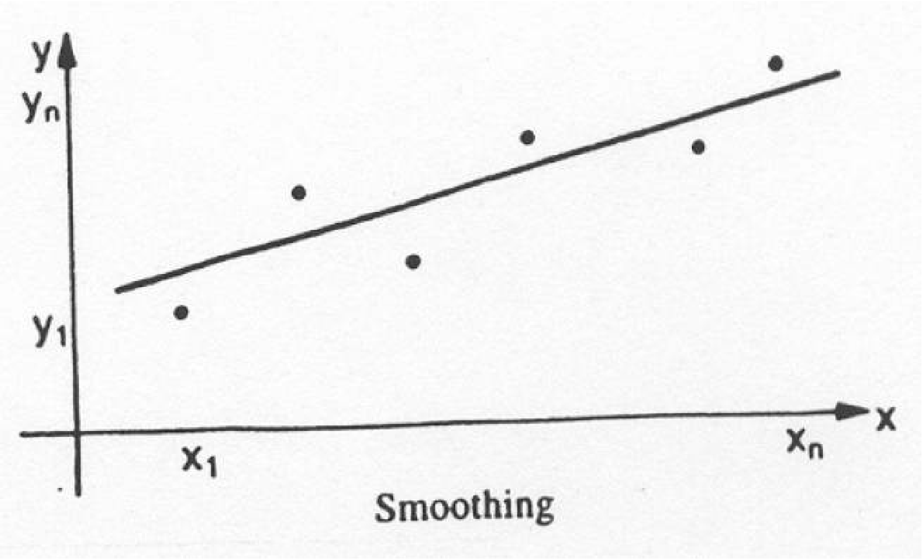


FIGURE 3. Approximation

- **Approximation** ($\mathbf{N} > m$) : In case of experimental values, imprecise values are obtained due to measurement noise and other uncontrolled perturbations. The aim of approximation (which is also called smoothing) is to try to eliminate the effects of noise as much as possible from experimental data.

Example 64. *Consider energy consumption of a boat as a function of distance (see Figures 2 and 3). It is clear that total energy consumed is directly proportional to distance but the atmospheric factors namely wind, water current etc. create perturbations. Figures 2 and 3 also shows results of interpolation and smoothing, respectively. Clearly, in this case the interpolation is unable to bring out the trend in energy consumption while smoothing produces acceptable results.*

In the case of imprecise data, the result of interpolation is absurd when the original problem is sensitive to errors. Consequently, these results cannot be used for any future predictions. On the other hand, smoothing allows us to bring out a tendency to reduce the probable energy consumption y for any arbitrary distance x.

3.5. Quality of Approximation. When we approximate a real continuous function$\{y(\mathbf{x}) : \mathbf{x} \in [\mathbf{a}, \mathbf{b}]\}$ or a set of numerical data $\{y_i \ : \ i = 1, 2......N \ \}$ by an analytic function $\widehat{y}(\mathbf{x})$, it is desirable to choose $\widehat{y}(\mathbf{x})$ such that error between the approximate function and real function / measurements is small in some sense. The 'distance' between the real function/ measurements and its model predictions can be measured by the **norm** function. For the set of numerical data $\left(\mathbf{x}^{(i)}, y_i\right)$where $y_i = y\left(\mathbf{x}^{(i)}\right)$, let us define vectors Y of measurements

$$\mathbf{y} = \left[\begin{array}{cccc} y_1 & y_2 & .... & y_N \end{array} \right]_{N\times 1}$$

and $\widehat{\mathbf{y}}$ of model predictions as

$$\widehat{\mathbf{y}} = \left[\begin{array}{cccc} \widehat{y}_1 & \widehat{y}_2 & .... & \widehat{y}_N \end{array} \right]_{N\times 1}$$

Then, we choose parameters of (3.10) such that

$$f\left(e_1, e_2, ...e_N\right) = \left\|\mathbf{y} - \widehat{\mathbf{y}}\right\|_p$$

is minimized. Commonly used norms are

- Laplace - norm (1 - norm)

$$|| \ \mathbf{y} - \widehat{\mathbf{y}} \ ||_1 = \sum_{i=1}^{N} | \ y_i - \widehat{y}_i \ | \tag{3.21}$$

- Euclidean - norm (2 - norm)

$$|| \ \mathbf{y} - \widehat{\mathbf{y}} \ ||_2 = \sum_{i=1}^{N} (\ y_i - \widehat{y}_i)^2 \tag{3.22}$$

- Laplace - Chebyshev norm ($\propto$ - norm)

$$\| \mathbf{y} - \widehat{\mathbf{y}} \|_{\propto} = \max_{i=1....\mathbf{N}} | y_i - \widehat{y}_i | \tag{3.23}$$

By definition, the best approximation $\widehat{y}$ of y in the least squares will imply that $\| y - \widehat{y} \|_2$ is minimum. The 2 - norm defined above is often inadequate in practice as some values can be more precise than the others and approximation $\widehat{y}$ should be more influenced by precise values than the others. For the above reasons, a weighted least square norm is introduced as

$$\| \mathbf{y} - \widehat{\mathbf{y}} \|_{2,w} = \sum_{i=1}^{\mathbf{N}} (y_i - \widehat{y}_i)^2 w_i \tag{3.24}$$

where $w_i \geq 0$ (for $i = 1, 2$...........$\mathbf{N}$)are weights associated with experimental values.

In the case of approximation of a continuous function defined on an interval $[a, b]$ we define the following norms.

$$\| y(z) - \widehat{y}(z) \|_1 = \int_a^b | y(z) - \widehat{y}(z) | \, dz \tag{3.25}$$

$$\| y(z) - \widehat{y}(z) \|_2 = \int_a^b [y(z) - \widehat{y}(z)]^2 \, dz \tag{3.26}$$

$$\| y(z) - \widehat{y}(z) \|_{2,w} = \int_a^b [y(z) - \widehat{y}(z)]^2 \, w(z) \, dz \tag{3.27}$$

$$\| y(z) - \widehat{y}(z) \|_{\propto} = \max_{z \in [a,b]} | y(z) - \widehat{y}(z) | \tag{3.28}$$

The procedures for the determination of the best approximation are both for continuous and discrete cases.

4. Multivariate Linear Regression

4.1. Least Square Formulation for Linear In Parameter Models.

Suppose the following model is proposed for a phenomenon

$$\widehat{y} = \sum_{j=1}^{m} \theta_j \, f_j \, (\mathbf{x}) \tag{4.1}$$

where $x \in R^r$ and we have N experimental data sets $\{ \ (x^{(i)}, y_i) : \ i = 1, ...N \ \}$. Defining the i^{th} approximation error as

$$e_i = y_i - \widehat{y}_i = \quad y_i - \sum_{j=1}^{m} \theta_j \ f_j(\mathbf{x}^{(i)}) \tag{4.2}$$

it is desired to choose a solution that minimizes the scalar quantity

$$\begin{array}{c} \min \\ \theta_1....\theta_m \end{array} \left[f = \sum_{i=1}^{\mathbf{N}} e_i^2 w_i \right] \tag{4.3}$$

where $w_i \geq 0$ are the weights associated with the individual measurements. These weights can be chosen to reflect reliability of each experimental data. A relatively large weight w_i can be selected for the experimental data set $(x^{(i)}, y_i)$ that is more reliable and vice-versa.

The above optimization problem can be expressed in compact vector-matrix notation as follows

$$\widehat{y}_1 = \theta_1 f_1(\mathbf{x}^{(1)}) + \theta_2 f_2\left(\mathbf{x}^{(1)}\right) + ... + \theta_m f_m(\mathbf{x}^{(1)}) \tag{4.4}$$

$$\widehat{y}_2 = \theta_1 f_1(\mathbf{x}^{(2)}) + \theta_2 f_2\left(\mathbf{x}^{(2)}\right) + ... + \theta_m f_m(\mathbf{x}^{(2)}) \tag{4.5}$$

$$..... = ...$$

$$\widehat{y}_{\mathbf{N}} = \theta_1 f_1(\mathbf{x}^{(N)}) + \theta_2 f_2\left(\mathbf{x}^{(N)}\right) + ... + \theta_m f_m(\mathbf{x}^{(N)}) \tag{4.6}$$

Defining

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & .... & \theta_m \end{bmatrix}^T \in R^m \tag{4.7}$$

$$\varphi^{(i)} = \begin{bmatrix} f_1(\mathbf{x}^{(i)}) & f_2(\mathbf{x}^{(i)}) & .... & f_m(\mathbf{x}^{(i)}) \end{bmatrix}^T \in R^m \tag{4.8}$$

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 & \widehat{y}_2 & .... & \widehat{y}_{\mathbf{N}} \end{bmatrix} \in R^{\mathbf{N}} \tag{4.9}$$

and

$$\Phi = \begin{bmatrix} f_1(\mathbf{x}^{(1)}) & .... & f_m(\mathbf{x}^{(1)}) \\ ...... & .... & ...... \\ f_1(\mathbf{x}^{(N)}) & .... & f_m(\mathbf{x}^{(N)}) \end{bmatrix}_{\mathbf{N} \ x \ m} = \begin{bmatrix} \left(\varphi^{(1)}\right)^T \\ ...... \\ \left(\varphi^{(N)}\right)^T \end{bmatrix} \tag{4.10}$$

we get an over - determined set of equations

$$\widehat{\mathbf{y}} = \Phi \boldsymbol{\theta} \tag{4.11}$$

Let the vector of measurements be defined as

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & .... & y_{\mathbf{N}} \end{bmatrix}^T \in R^{\mathbf{N}} \tag{4.12}$$

Now, defining approximation error vector $\mathbf{e} \in R^{\mathbf{N}}$ as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \Phi\boldsymbol{\theta} \tag{4.13}$$

(4.14)

it is desired to choose θ such that the quantity $\Phi = \mathbf{e}^T\mathbf{W}\mathbf{e}$ is minimized, i.e.

$$\widehat{\boldsymbol{\theta}} = \begin{matrix} \min \\ \boldsymbol{\theta} \end{matrix} \ \mathbf{e}^T\mathbf{W}\mathbf{e} \tag{4.15}$$

where

$$\mathbf{W} = diag \left[\begin{matrix} w_1 & w_2 & .... & w_{\mathbf{N}} \end{matrix} \right] \tag{4.16}$$

is a diagonal weighting matrix.

4.2. Solution of Linear Least Square Problem. Consider the minimization problem

$$\widehat{\boldsymbol{\theta}} = \begin{matrix} \min \\ \boldsymbol{\theta} \end{matrix} \ \mathbf{e}^T\mathbf{W}\mathbf{e} \tag{4.17}$$

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - A\boldsymbol{\theta}$$

Using the necessary condition for optimality, we have

$$\frac{\partial\left[\mathbf{e}^T\mathbf{W}\mathbf{e}\right]}{\partial\boldsymbol{\theta}} = \overline{0}$$

Rules of differentiation of a scalar function $f = \mathbf{u}^T B\mathbf{v}$ with respect to vectors $\mathbf{u}$ and $\mathbf{v}$ can be stated as follows

$$\frac{\partial}{\partial\mathbf{u}}\left(\mathbf{u}^T B\mathbf{v}\right) = B\mathbf{v}\ ; \qquad \frac{\partial}{\partial\mathbf{v}}[\mathbf{u}^T B\mathbf{v}] = B^T\mathbf{u} \tag{4.18}$$

$$\frac{\partial}{\partial\mathbf{u}}\left[\mathbf{u}^T B\mathbf{u}\right] = 2\ B\mathbf{u} \qquad \text{when } B \text{ is symmetric} \tag{4.19}$$

Now, applying the above rules

$$\mathbf{e}^T\mathbf{W}\mathbf{e} = [\mathbf{y} - \Phi\boldsymbol{\theta}]^T\ \mathbf{W}\ [\mathbf{y} - \Phi\boldsymbol{\theta}] \tag{4.20}$$

$$= \mathbf{y}^T\mathbf{W}\mathbf{y} - (\Phi\boldsymbol{\theta})^T\mathbf{W}\mathbf{y} - \mathbf{y}^T\mathbf{W}\Phi\boldsymbol{\theta} + \boldsymbol{\theta}^T(\Phi^T\mathbf{W}\Phi)\boldsymbol{\theta}$$

$$\frac{\partial\left[\mathbf{e}^T\mathbf{W}\mathbf{e}\right]}{\partial\boldsymbol{\theta}} = -\Phi^T\mathbf{W}\mathbf{y} - \Phi^T\mathbf{W}\mathbf{y} + 2(\Phi^T\mathbf{W}\Phi)\widehat{\boldsymbol{\theta}} = \overline{0} \tag{4.21}$$

$$\Rightarrow\ (\Phi^T\mathbf{W}\Phi)\ \widehat{\boldsymbol{\theta}}_{LS} = \Phi^T\mathbf{W}\mathbf{y} \tag{4.22}$$

In the above derivation, we have used the fact that matrix $\Phi^T\mathbf{W}\Phi$ is symmetric. If matrix $(\Phi^T\mathbf{W}\Phi)$ is invertible, the least square estimate of parameters $\widehat{\boldsymbol{\theta}}$ can computed as

$$\widehat{\boldsymbol{\theta}}_{LS} = \left[\Phi^T\mathbf{W}\Phi\right]^{-1}\left(\Phi^T\mathbf{W}\right)\mathbf{y} \tag{4.23}$$

Using sufficient condition for optimality, Hessian matrix should be positive definite or positive semi-definite for the stationary point to be a minimum. Now,

$$\left[\frac{\partial^2\left[\mathbf{e}^T\mathbf{W}\mathbf{e}\right]}{\partial\boldsymbol{\theta}^2}\right] = 2(\Phi^T\mathbf{W}\Phi) \tag{4.24}$$

It can be easily shown that

$$\mathbf{v}^T\left(\Phi^T\mathbf{W}\Phi\right)\mathbf{v} \geq 0 \quad \text{for any } \mathbf{v} \in R^m \tag{4.25}$$

and the sufficiency condition is satisfied and the stationary point is a minimum. As Φ is a convex function, it can be shown that the solution $\hat{\boldsymbol{\theta}}$ is the global minimum of Φ.

Thus, linear least square estimation problem is finally reduced to solving equation of the form $\mathbf{Ax} = \mathbf{b}$ where

$$\mathbf{A} = \Phi^T\mathbf{W}\Phi \quad \text{and} \quad \mathbf{b} = \Phi^T\mathbf{W}\mathbf{y} \tag{4.26}$$

Note that $\Phi^T\mathbf{W}\Phi$ is symmetric and positive semi-definite and Cholensky decomposition method can be used to solve the resulting linear system.

5. Projections: Geometric Interpretation of Linear Regression

5.1. Distance of a Point from a Line.

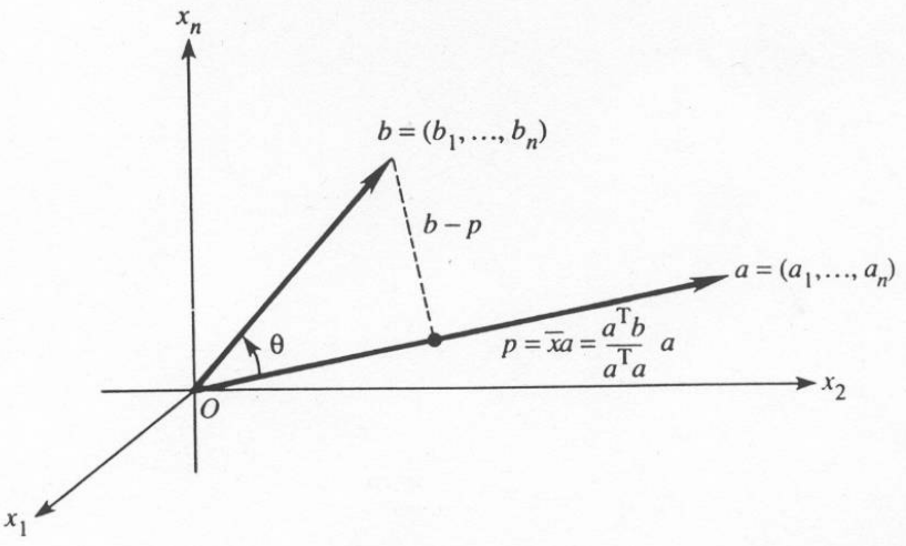
Suppose we are given a point $\mathbf{y} \in R^3$ in space and we want to find its distance from the line in the direction of vector $\mathbf{a} \in R^3$. In other words, we are looking for a point $\mathbf{p}$ along the line that is closest to $\mathbf{y}$ (see Figure 4),i.e $\mathbf{p} = \theta\mathbf{a}$ such that

$$\Phi = \|\mathbf{p}-\mathbf{y}\|_2 = \|\theta\mathbf{a}-\mathbf{y}\|_2 \tag{5.1}$$

is minimum. This problem can be solved by minimizing Φ with respect to θ,which is equivalent to

$$\min_{\theta} \Phi^2 = \min_{\theta} \langle\theta\mathbf{a}-\mathbf{y},\theta\mathbf{a}-\mathbf{y}\rangle \tag{5.2}$$

$$= \min_{\theta} \left[\theta^2\langle\mathbf{a},\mathbf{a}\rangle - 2\theta\langle\mathbf{a},\mathbf{y}\rangle + \langle\mathbf{y},\mathbf{y}\rangle\right] \tag{5.3}$$



The projection of b onto a, with $\cos\theta = \frac{Op}{Ob} = \frac{a^\mathrm{T} b}{\|a\|\,\|b\|}$.

FIGURE 4

Using necessary condition for optimality,

$$\frac{\partial \Phi^2}{\partial \theta} = \theta \langle \mathbf{a}, \mathbf{a} \rangle - \langle \mathbf{a}, \mathbf{y} \rangle = 0 \tag{5.4}$$

$$\Rightarrow \theta = \frac{\langle \mathbf{a}, \mathbf{y} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle} \tag{5.5}$$

$$\mathbf{p} = \theta\, \mathbf{a} = \frac{\langle \mathbf{a}, \mathbf{y} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle} \mathbf{a} \tag{5.6}$$

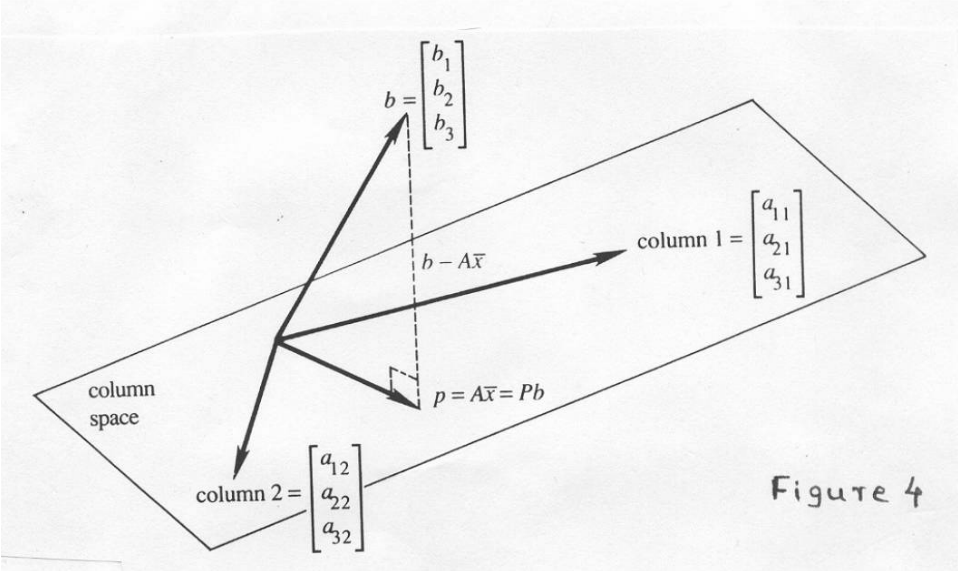
Now, equation (5.4) can be rearranged as

$$\langle \mathbf{a}, \theta \mathbf{a} \rangle - \langle \mathbf{a}, \mathbf{y} \rangle = \langle \mathbf{a}, \theta \mathbf{a} - \mathbf{y} \rangle = \langle \mathbf{a}, \mathbf{p} - \mathbf{y} \rangle = 0 \tag{5.7}$$

From school geometry we know that if $\mathbf{p}$ is such a point, then the vector $(\mathbf{y} - \mathbf{p})$ is perpendicular to direction $\mathbf{a}$. We have derived this geometric result using optimization. Equation (5.6) can be rearranged as

$$\mathbf{p} = \left\langle \frac{\mathbf{a}}{\sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}}, \mathbf{y} \right\rangle \frac{\mathbf{a}}{\sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}} = \langle \widehat{\mathbf{a}}, \mathbf{y} \rangle \, \widehat{\mathbf{a}} \tag{5.8}$$

where $\widehat{\mathbf{a}} = \dfrac{\mathbf{a}}{\sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}}$ is unit vector along direction of $\mathbf{a}$.and point $\mathbf{p}$ is the projection of vector $\mathbf{y}$ along direction $\widehat{\mathbf{a}}$. Note that the above derivation holds in any general n dimensional space $\mathbf{a}, \mathbf{y} \in R^n$ or even any infinite dimensional vector space.



$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$
$b - A\bar{x}$
column 1 = $\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}$
$p = A\bar{x} = Pb$
column space
column 2 = $\begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}$
Figure 4

Projection onto the column space of a 3 by 2 matrix.

Figure 5

The equation can be rearranged as

$$\mathbf{p} = \ \mathbf{a}\left(\frac{\mathbf{a}^T\mathbf{y}}{\mathbf{a}^T\mathbf{a}}\right) = \left[\frac{1}{\mathbf{a}^T\mathbf{a}}\right]\left[\ \mathbf{a}\mathbf{a}^T\right]\ \mathbf{y} = \mathbf{P}_r.\mathbf{y} \tag{5.9}$$

where $\mathbf{P}_r = \frac{\mathbf{1}}{\mathbf{a^Ta}}\mathbf{a}\mathbf{a}^T$ is a $n \times n$ matrix is called as **projection matrix,** which projects vector $\mathbf{y}$ into its column space.

5.2. Distance of a point from Subspace. The situation is exactly same when we are given a point $\mathbf{y} \in R^3$ and plane S in R^3 passing through origin,we want to find distance of $\mathbf{y}$ from S,i.e. a point $\mathbf{p} \in S$ such that $\|\mathbf{p} - \mathbf{y}\|_2$ is minimum (see Figure 5). Again, from school geometry, we know that such point can be obtained by drawing a perpendicular from $\mathbf{y}$ to S ; $\mathbf{p}$ is the point where this perpendicular meets S (see Figure 5). We would like to formally derive this result using optimization.

More generally, given a point $\mathbf{y} \in R^m$ and subspace S of R^m, the problem is to find a point $\mathbf{p}$ in subspace S such that it is closest to vector $\mathbf{y}$. Let $S = span\left\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, ..., \mathbf{a}^{(m)}\right\}$ and as $\mathbf{p} \in S$ we have

$$\mathbf{p} = \theta_1\mathbf{a}^{(1)} + \theta_2\mathbf{a}^{(2)} + ... + \theta_m\mathbf{a}^{(m)} = \sum_{i=1}^{m} \theta_i\mathbf{a}^{(i)} \tag{5.10}$$

We want to find a point find $\mathbf{p}$ such that

$$\Phi = \|\mathbf{p}-\mathbf{y}\|_2 = \left\| \left(\sum_{i=1}^{m} \theta_i \mathbf{a}^{(i)} \right) - \mathbf{y} \right\|_2 \tag{5.11}$$

is minimum. This problem is equivalent to

$$\begin{matrix}\min \\ \boldsymbol{\theta}\end{matrix}\ \Phi^2 = \begin{matrix}\min \\ \boldsymbol{\theta}\end{matrix}\ \left\langle \left(\sum_{i=1}^{m} \theta_i \mathbf{a}^{(i)} - \mathbf{y} \right), \left(\sum_{i=1}^{m} \theta_i \mathbf{a}^{(i)} - \mathbf{y} \right) \right\rangle \tag{5.12}$$

Using necessary condition for optimality,

$$\begin{aligned} \frac{\partial \Phi^2}{\partial \theta_j} &= \left\langle \mathbf{a}^{(j)}, \left(\sum_{i=1}^{m} \theta_i \mathbf{a}^{(i)} - \mathbf{y} \right) \right\rangle = \left\langle \left(\mathbf{a}^{(i)}\right), \left(\mathbf{p}-\mathbf{y}\right) \right\rangle = 0 \\ j &= 1, 2, ...m \end{aligned} \tag{5.13}$$

Equation (5.13) has a straight forward geometric interpretation. Vector $\mathbf{p}-\mathbf{y}$ is orthogonal to each vector $\mathbf{a}^{(i)}$, which forms the basis of S. The point $\mathbf{p}$ is a projection of $\mathbf{y}$ onto subspace S. This is exactly what we learn in the school geometry (see Figure 5).

Now, let us calculate optimal parameters θ_i using equation (5.13). Equation (5.13) can be rearranged as

$$\left\langle \mathbf{a}^{(j)}, \sum_{i=1}^{m} \theta_i \mathbf{a}^{(i)} \right\rangle = \sum_{i=1}^{m} \theta_i \left\langle \mathbf{a}^{(j)}, \mathbf{a}^{(i)} \right\rangle = \left\langle \mathbf{a}^{(j)}, \mathbf{y} \right\rangle \tag{5.14}$$

$$j = 1, 2, ...m \tag{5.15}$$

Collecting the above set of equations and using vector-matrix notation, we have

$$\begin{bmatrix} \langle \mathbf{a}^{(1)}, \mathbf{a}^{(1)} \rangle & \langle \mathbf{a}^{(1)}, \mathbf{a}^{(2)} \rangle & .... & \langle \mathbf{a}^{(1)}, \mathbf{a}^{(m)} \rangle \\ \langle \mathbf{a}^{(2)}, \mathbf{a}^{(1)} \rangle & \langle \mathbf{a}^{(2)}, \mathbf{a}^{(2)} \rangle & .... & \langle \mathbf{a}^{(2)}, \mathbf{a}^{(m)} \rangle \\ ..... & ..... & ..... & ..... \\ \langle \mathbf{a}^{(m)}, \mathbf{a}^{(1)} \rangle & \langle \mathbf{a}^{(m)}, \mathbf{a}^{(2)} \rangle & ..... & \langle \mathbf{a}^{(m)}, \mathbf{a}^{(m)} \rangle \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ .... \\ \theta_m \end{bmatrix} = \begin{bmatrix} \langle \mathbf{a}^{(1)}, \mathbf{y} \rangle \\ \langle \mathbf{a}^{(2)}, \mathbf{y} \rangle \\ .... \\ \langle \mathbf{a}^{(m)}, \mathbf{y} \rangle \end{bmatrix} \tag{5.16}$$

This is nothing but the classic **normal equation** derived in the above subsection.

Let us now interpret the least square parameter estimation problem stated in the last section using the above geometric arguments. The least square problem was posed as choosing the parameter vector $\widehat{\boldsymbol{\theta}}$ such that

$$\| \mathbf{e} \|_{W,2} = \| \mathbf{y} - \boldsymbol{\Phi\theta} \|_{W,2} = \sqrt{[\mathbf{y} - \boldsymbol{\Phi\theta}]^T \mathbf{W} [\mathbf{y} - \boldsymbol{\Phi\theta}]} \tag{5.17}$$

is minimized. The subscript $(_{W,2})$ indicates that $\|.\|_{W,2}$ is a 2-norm defined using matrix $\mathbf{W}$. This is exactly the geometrical problem of finding distance of vector Y from a subspace S . The sub-space involved is nothing but the column space of $\mathbf{\Phi}$. Let $\mathbf{\Phi}$ be represented as

$$\Phi = \begin{bmatrix} \mathbf{a}^{(1)} & \mathbf{a}^{(2)} & .... & \mathbf{a}^{(m)} \end{bmatrix} \tag{5.18}$$

where $\mathbf{a}^{(i)} \in R^N$ are columns of matrix Φ. Let us define the inner product for any $\mathbf{u}, \mathbf{v} \in R^N$ as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{W} \mathbf{v} \tag{5.19}$$

Let the vector $\mathbf{p}$ such that

$$\mathbf{p} = \widehat{\theta}_1 \mathbf{a}^{(1)} + \widehat{\theta}_2 \mathbf{a}^{(2)} + ... + \widehat{\theta}_m \mathbf{a}^{(m)} = \Phi \widehat{\boldsymbol{\theta}} \tag{5.20}$$

Then, using geometric arguments, we have

$$\begin{aligned}
\left\langle \mathbf{a}^{(1)}, \mathbf{y} - \mathbf{\Phi}\widehat{\boldsymbol{\theta}} \right\rangle &= \left[\mathbf{a}^{(1)}\right]^T \mathbf{W}\mathbf{y} - \left[\mathbf{a}^{(1)}\right]^T \mathbf{W} \left[\mathbf{\Phi}\widehat{\boldsymbol{\theta}}\right] = 0 && (5.21)\\
\left\langle \mathbf{a}^{(2)}, \mathbf{y} - \mathbf{\Phi}\widehat{\boldsymbol{\theta}} \right\rangle &= \left[\mathbf{a}^{(2)}\right]^T \mathbf{W}\mathbf{y} - \left[\mathbf{a}^{(2)}\right]^T \mathbf{W} \left[\mathbf{\Phi}\widehat{\boldsymbol{\theta}}\right] = 0 && (5.22)\\
.... &= .... = 0 \\
\left\langle \mathbf{a}^{(m)}, \mathbf{y} - \mathbf{\Phi}\widehat{\boldsymbol{\theta}} \right\rangle &= \left[\mathbf{a}^{(m)}\right]^T \mathbf{W}\mathbf{y} - \left[\mathbf{a}^{(m)}\right]^T \mathbf{W} \left[\mathbf{\Phi}\widehat{\boldsymbol{\theta}}\right] = 0 && (5.23)
\end{aligned}$$

Rearranging above equations, we have

$$\begin{bmatrix} \left[\mathbf{a}^{(1)}\right]^T \\ \left[\mathbf{a}^{(2)}\right]^T \\ .... \\ \left[\mathbf{a}^{(m)}\right]^T \end{bmatrix} \mathbf{W\Phi} \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ .... \\ \widehat{\theta}_m \end{bmatrix} = \begin{bmatrix} \left[\mathbf{a}^{(1)}\right]^T \mathbf{Wy} \\ \left[\mathbf{a}^{(2)}\right]^T \mathbf{Wy} \\ .... \\ \left[\mathbf{a}^{(m)}\right]^T \mathbf{Wy} \end{bmatrix} = \begin{bmatrix} \left[\mathbf{a}^{(1)}\right]^T \\ \left[\mathbf{a}^{(2)}\right]^T \\ .... \\ \left[\mathbf{a}^{(m)}\right]^T \end{bmatrix} \mathbf{Wy} \tag{5.24}$$

$$\left[\mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi}\right] \widehat{\boldsymbol{\theta}} = \left[\mathbf{\Phi}^T \mathbf{W}\right] \mathbf{y} \tag{5.25}$$

It can be easily shown that

$$\left[\mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi}\right] = \begin{bmatrix} \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(2)} \right\rangle & .... & \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(m)} \right\rangle \\ \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(2)} \right\rangle & .... & \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(m)} \right\rangle \\ ..... & ..... & ..... & ..... \\ \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(2)} \right\rangle & ..... & \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(m)} \right\rangle \end{bmatrix} \tag{5.26}$$

$$\left[\mathbf{\Phi}^T \mathbf{W}\right] \mathbf{y} = \begin{bmatrix} \left\langle \mathbf{a}^{(1)}, \mathbf{y} \right\rangle \\ \left\langle \mathbf{a}^{(2)}, \mathbf{y} \right\rangle \\ .... \\ \left\langle \mathbf{a}^{(m)}, \mathbf{y} \right\rangle \end{bmatrix} \tag{5.27}$$

5.3. Additional Geometric Insights.

- Let us consider the special case where $\mathbf{W} = \mathbf{I}$,i.e. identity matrix. If columns of $\mathbf{\Phi}$ are linearly independent then $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible and, the point $\mathbf{p}$, which is projection of $\mathbf{y}$ onto column space of $\mathbf{\Phi}$ (i.e. $R(\mathbf{\Phi})$) is given as

$$\mathbf{p} = \mathbf{\Phi}\widehat{\boldsymbol{\theta}} = \mathbf{\Phi}\left[\mathbf{\Phi}^T\mathbf{\Phi}\right]^{-1}\left[\mathbf{\Phi}^T\right]\mathbf{y} = [P_r]\,\mathbf{y} \tag{5.28}$$

$$P_r = \mathbf{\Phi}\left[\mathbf{\Phi}^T\mathbf{\Phi}\right]^{-1}\left[\mathbf{\Phi}^T\right] \tag{5.29}$$

Here matrix P_r is the projection matrix, which projects matrix $\mathbf{y}$ onto $R(\mathbf{\Phi})$, i.e. the column space of $\mathbf{\Phi}$. Note that $[P_r]\,\mathbf{y}$ is the component of $\mathbf{y}$ in $R(\mathbf{\Phi})$

$$\mathbf{y} - (P_r)\mathbf{y} = [I - P_r]\,\mathbf{y} \tag{5.30}$$

is component of $\mathbf{y}$ $\perp$ to $R(\mathbf{\Phi})$. Thus we have a matrix formula of splitting a vector into two orthogonal components.

- Projection matrix has two fundamental properties.
  (1) $[P_r]^2 = P_r$
  (2) $[P_r]^T = P_r$

Conversely, any symmetric matrix with $\mathbf{\Phi}^2 = \mathbf{\Phi}$ represents a projection matrix.

- Suppose then $\mathbf{y} \in R(\mathbf{\Phi})$, then $\mathbf{y}$ can be expressed as linear combination of columns of $\mathbf{\Phi}$ i.e.,the projection of $\mathbf{y}$ is still $\mathbf{y}$ itself.

$$\mathbf{p} = \mathbf{\Phi}\widehat{\boldsymbol{\theta}} = \mathbf{y} \tag{5.31}$$

This implies

$$\mathbf{p} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)\widehat{\boldsymbol{\theta}} = \mathbf{\Phi}\widehat{\boldsymbol{\theta}} = \mathbf{y} \tag{5.32}$$

The closest point of $\mathbf{p}$ to $\mathbf{y}$ is $\mathbf{y}$ itself

- At the other extreme, suppose $\mathbf{y} \perp R(\mathbf{\Phi})$. Then

$$p = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \quad \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\bar{0} = \bar{0} \tag{5.33}$$

- When $\mathbf{\Phi}$ is square and invertible, every vector projects onto itself, i.e.

$$\mathbf{p} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \quad (\mathbf{\Phi}\mathbf{\Phi}^{-1})(\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}^T\mathbf{y} = \mathbf{y} \tag{5.34}$$

Suppose $\mathbf{\Phi}$ is only a column vector. Then

$$\theta = \frac{a^T\mathbf{y}}{a^Ta} \tag{5.35}$$

Remark 9. *Matrix* $\left[\mathbf{\Phi}^T\mathbf{\Phi}\right]^{-1}\left[\mathbf{\Phi}^T\right]$ *is called as pseudo-inverse of matrix* $\mathbf{\Phi}$.

5.4. Projection Theorem in a general Hilbert Space. Equations we have derived in the above sub-sections are special cases of a very general result called **projection theorem,** which holds in any Hilbert space. Although we state this result here without giving a formal proof, the discussion in the above subsections provided sufficient basis for understanding the theorem.

Theorem 13. ***Classical Projection Theorem :*** *Let X be a Hilbert space and S be a finite dimensional subspace of X. Corresponding to any vector $\mathbf{y} \in X$, there is unique vector $\mathbf{p} \in S$ such that $\|\mathbf{y}-\mathbf{p}\|_2 \leq \|\mathbf{y}-\mathbf{s}\|_2$ for any vector $\mathbf{s} \in S$. Furthermore, a necessary and sufficient condition for $\mathbf{p} \in S$ be the unique minimizing vector is that vector $(\mathbf{y}-\mathbf{p})$ is orthogonal to S.*

Thus, given any finite dimensional sub-space S spanned by linearly independent vectors $\left\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, ..., \mathbf{a}^{(m)}\right\}$ and an arbitrary vector $\mathbf{y} \in X$ we seek a vector $\mathbf{p} \in S$

$$\mathbf{p} = \theta_1 \mathbf{a}^{(1)} + \theta_2 \mathbf{a}^{(2)} + ... + \theta_m \mathbf{a}^{(m)}$$

such that

$$\left\|\mathbf{y} - \left(\theta_1 \mathbf{a}^{(1)} + \theta_2 \mathbf{a}^{(2)} + ... + \theta_m \mathbf{a}^{(m)}\right)\right\|_2 \tag{5.36}$$

is minimized with respect to scalars $\widehat{\theta}_{1,}....\widehat{\theta}_m$. Now, according to the projection theorem, the unique minimizing vector $\mathbf{p}$ is the orthogonal projection of $\mathbf{y}$ on S. This translates to the following set of equations

$$\left\langle \mathbf{y}-\mathbf{p}, \mathbf{a}^{(i)} \right\rangle = \left\langle \mathbf{y} - \left(\theta_1 \mathbf{a}^{(1)} + \theta_2 \mathbf{a}^{(2)} + ... + \theta_m \mathbf{a}^{(m)}\right), \mathbf{a}^{(i)} \right\rangle = 0 \tag{5.37}$$

for $i = 1, 2, ...m$. This set of m equations can be written as $\mathbf{G}\boldsymbol{\theta} = \mathbf{b}$. i.e.
(5.38)

$$\begin{bmatrix} \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(2)} \right\rangle & .... & \left\langle \mathbf{a}^{(1)}, \mathbf{a}^{(m)} \right\rangle \\ \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(2)} \right\rangle & .... & \left\langle \mathbf{a}^{(2)}, \mathbf{a}^{(m)} \right\rangle \\ ..... & ..... & ..... & ..... \\ \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(1)} \right\rangle & \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(2)} \right\rangle & ..... & \left\langle \mathbf{a}^{(m)}, \mathbf{a}^{(m)} \right\rangle \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ .... \\ \theta_m \end{bmatrix} = \begin{bmatrix} \left\langle \mathbf{a}^{(1)}, \mathbf{y} \right\rangle \\ \left\langle \mathbf{a}^{(2)}, \mathbf{y} \right\rangle \\ .... \\ \left\langle \mathbf{a}^{(m)}, \mathbf{y} \right\rangle \end{bmatrix}$$

This is the general form of **normal equation** resulting from the minimization problem. The $m \times m$ matrix $\mathbf{G}$ on L.H.S. is called as Gram matrix. If vectors $\left\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, ..., \mathbf{a}^{(m)}\right\}$ are linearly independent, then Gram matrix is nonsingular. Moreover, if the set $\left\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, ..., \mathbf{a}^{(m)}\right\}$ is chosen to be an orthonormal set, say $\left\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, ..., \mathbf{e}^{(m)}\right\}$, then Gram matrix reduces to identity matrix i.e. $\mathbf{G} = I$ and we have

$$\mathbf{p} = \theta_1 \mathbf{e}^{(1)} + \theta_2 \mathbf{e}^{(2)} + ... + \theta_m \mathbf{e}^{(m)} \tag{5.39}$$

where

$$\theta_i = \left\langle \mathbf{e}^{(i)}, \mathbf{y} \right\rangle$$

as $\left\langle \mathbf{e}^{(i)}, \mathbf{e}^{(j)} \right\rangle = 0$ when $i \neq j$. It is important to note that, if we choose orthonormal set $\left\{ \mathbf{e}^{(1)}, \mathbf{e}^{(2)}, ..., \mathbf{e}^{(m)} \right\}$ and we want to include an additional orthonormal vector, say $\mathbf{e}^{(m+1)}$, to this set, then we can compute θ_{m+1} as

$$\theta_{m+1} = \left\langle \mathbf{e}^{(m+1)}, \mathbf{y} \right\rangle$$

without requiring to recompute $\theta_1, ...\theta_m$.

Remark 10. *Given any Hilbert space X and a orthonormal basis for the Hilbert space $\left\{ \mathbf{e}^{(1)}, \mathbf{e}^{(2)}, .., \mathbf{e}^{(m)}, ... \right\}$ we can express any vector $\mathbf{u} \in X$ as*

$$\mathbf{u} = \alpha_1 \mathbf{e}^{(1)} + \alpha_2 \mathbf{e}^{(2)} + ... + \alpha_m \mathbf{e}^{(m)} + ... \tag{5.40}$$

$$\alpha_i = \left\langle \mathbf{e}^{(i)}, \mathbf{u} \right\rangle \tag{5.41}$$

The series

$$\mathbf{u} = \left\langle \mathbf{e}^{(\mathbf{1})}, \mathbf{u} \right\rangle \mathbf{e}^{(1)} + \left\langle \mathbf{e}^{(2)}, \mathbf{u} \right\rangle \mathbf{e}^{(2)} + ... + \left\langle \mathbf{e}^{(i)}, \mathbf{u} \right\rangle \mathbf{e}^{(i)} + ... \tag{5.42}$$

$$= \sum_{i=1}^{\infty} \left\langle \mathbf{e}^{(\mathbf{i})}, \mathbf{u} \right\rangle \mathbf{e}^{(i)} \tag{5.43}$$

*which converges to element $\mathbf{u} \in X$ is called as **generalized Fourier series expansion** of element $\mathbf{u}$ and coefficients $\alpha_i = \left\langle \mathbf{e}^{(i)}, \mathbf{u} \right\rangle$ are the corresponding **Fourier coefficients**. The well known Fourier expansion or a continuous function over interval $[-\pi, \pi]$ using $\sin(it)$ and $\cos(it)$ is a special case of this more general result.*

6. Statistical Interpretations of Linear Regression

6.1. Review of Fundamentals of Statistics. Consider a vector of random variables (RVs), say $\boldsymbol{\varphi} \in R^m$, where each element $\boldsymbol{\varphi}_i$ of $\boldsymbol{\varphi}$ is a random variable. Let $\overline{\boldsymbol{\varphi}}$ represents *population mean* (average of all possible outcomes of $\boldsymbol{\varphi}$**),** i.e.

$$\overline{\boldsymbol{\varphi}} = \mathbf{E}\left(\boldsymbol{\varphi}\right) = \begin{bmatrix} \mathbf{E}\left(\boldsymbol{\varphi}_1\right) & ... & \mathbf{E}\left(\boldsymbol{\varphi}_m\right) \end{bmatrix}^T \tag{6.1}$$

where operator $\mathbf{E}\left(.\right)$ represents expected value of RVs $\boldsymbol{\varphi}$ defined as

$$\mathbf{E}\left(\boldsymbol{\varphi}_i\right) = \int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} \boldsymbol{\varphi}_i p(\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_m) d\boldsymbol{\varphi}_1 ... d\boldsymbol{\varphi}_m \tag{6.2}$$

where $p(\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_m) = p(\boldsymbol{\varphi})$ represent multivariate probability density function of $\boldsymbol{\varphi}$.Also, let Σ represents population covariance, which is defined as

$$\begin{aligned}(6.3)\boldsymbol{\Sigma} &= \mathbf{E}\left[(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})^T\right] \\ &= \begin{bmatrix} \mathbf{E}\left[(\boldsymbol{\varphi}_1-\overline{\boldsymbol{\varphi}}_1)(\boldsymbol{\varphi}_1-\overline{\boldsymbol{\varphi}}_1)^T\right] & .... & \mathbf{E}\left[(\boldsymbol{\varphi}_1-\overline{\boldsymbol{\varphi}_1})(\boldsymbol{\varphi}_m-\overline{\boldsymbol{\varphi}}_m)^T\right] \\ .... & .... & .... \\ \mathbf{E}\left[(\boldsymbol{\varphi}_m-\overline{\boldsymbol{\varphi}}_m)(\boldsymbol{\varphi}_1-\overline{\boldsymbol{\varphi}}_1)^T\right] & .... & \mathbf{E}\left[(\boldsymbol{\varphi}_m-\overline{\boldsymbol{\varphi}}_m)(\boldsymbol{\varphi}_m-\overline{\boldsymbol{\varphi}}_m)^T\right] \end{bmatrix}\end{aligned}$$

In fact, operator $\mathbf{E}(.)$ can be used for computing ensemble (population) expectation of any arbitrary function of $\boldsymbol{\varphi}$, say $g(\boldsymbol{\varphi})$,i.e.,

$$(6.4) \qquad \mathbf{E}[g(\boldsymbol{\varphi})] = \int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} g(\boldsymbol{\varphi})p(\boldsymbol{\varphi})d\boldsymbol{\varphi}_1...d\boldsymbol{\varphi}_m$$

Mean and covariance are specific forms function $g(\boldsymbol{\varphi})$.

Given a vector of random variables (RVs) $\boldsymbol{\varphi}$, in general, we may not have exact knowledge of $p(\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_m), \overline{\boldsymbol{\varphi}}$ and $\boldsymbol{\Sigma}$. However, we can generate estimates of $\overline{\boldsymbol{\varphi}}$ and $\boldsymbol{\Sigma}$ if we can conduct experiments and collect data for $\boldsymbol{\varphi}$ in each experiments. Let us assume that we have carried out *experiments* and collected N data vectors $\left\{\boldsymbol{\varphi}^{(1)}, ..., \boldsymbol{\varphi}^{(N)}\right\}$, which can be can be arranged into a matrix as follows

$$(6.5) \qquad \Phi = \begin{bmatrix} \left(\boldsymbol{\varphi}^{(1)}\right)^T \\ .... \\ \left(\boldsymbol{\varphi}^{(N)}\right)^T \end{bmatrix}$$

$$(6.6) \qquad \left(\boldsymbol{\varphi}^{(i)}\right)^T = \begin{bmatrix} \phi_{i,1} & \phi_{i,2} & ..... & \phi_{i,m} \end{bmatrix}$$

which contains various values random variable $\boldsymbol{\varphi}_i$ takes during the experiments. The data matrix Φ can also be expressed as

$$(6.7) \qquad \Phi = \begin{bmatrix} \boldsymbol{\eta}^{(1)} & \boldsymbol{\eta}^{(2)} & ..... & \boldsymbol{\eta}^{(m)} \end{bmatrix}$$

$$(6.8) \qquad \boldsymbol{\eta}^{(j)} = \begin{bmatrix} \phi_{1,j} \\ .... \\ \phi_{N,j} \end{bmatrix}$$

where $\boldsymbol{\eta}^{(j)}$ represents j^{th} column vector of matrix Φ consists of vector.

The arithmetic average or *sample mean* of $\boldsymbol{\varphi}_i$ can be computed as

$$(6.9) \qquad \widehat{\overline{\boldsymbol{\varphi}}}_i = \frac{1}{N}\sum_{j=1}^{N}\phi_{j,i} \quad ; \quad i = 1, 2, ...m$$

Here, $\overline{\boldsymbol{\varphi}}_i$ represents *true mean* or (entire) *population mean* and $\widehat{\overline{\boldsymbol{\varphi}}}_i$ denotes estimate of population mean generated using the given data set. The sample mean can also be viewed as coefficient of projection of vector $\eta^{(i)}$ on unit vector $\mathbf{1}$ defined as

$$\mathbf{1} = \begin{bmatrix} 1 \\ .... \\ 1 \end{bmatrix} \tag{6.10}$$

$$\widehat{\overline{\boldsymbol{\varphi}}}_j = \frac{\langle \boldsymbol{\eta}^{(j)}, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} \quad ; \quad j = 1, 2, ...m \tag{6.11}$$

Alternatively, sample mean of random variable vector $\boldsymbol{\varphi}$ can be estimated as follows

$$\widehat{\overline{\boldsymbol{\varphi}}} = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{\varphi}^{(j)} \tag{6.12}$$

It may be noted that the vector $\mathbf{1}$ makes makes equal angles with all coordinate axis in R^N and the vector $\widehat{\overline{\boldsymbol{\varphi}}}_j \mathbf{1}$ represents the projection of $\eta^{(j)}$ along vector $\mathbf{1}$.

A measure of spread of data elements $\left\{\phi_{j,i} : j = 1, 2...N\right\}$ around the estimated sample mean $\widehat{\overline{\boldsymbol{\varphi}}}_i$ is given by *sample variance* defined as

$$s_i^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left[\phi_{j,i} - \overline{\boldsymbol{\varphi}}_i\right]^2 \tag{6.13}$$

The sample variance can also be estimated as follows

$$\mathbf{e}^{(i)} = \boldsymbol{\eta}^{(i)} - \widehat{\overline{\boldsymbol{\varphi}}}_i \mathbf{1} \tag{6.14}$$

$$s_i^2 = \frac{1}{N-1} \left[\mathbf{e}^{(i)}\right]^T \mathbf{e}^{(i)} \tag{6.15}$$

Note that the vector $\left(\boldsymbol{\eta}^{(i)} - \widehat{\overline{\boldsymbol{\varphi}}}_i \mathbf{1}\right)$ is orthogonal to vector $\left(\widehat{\overline{\boldsymbol{\varphi}}}_i \mathbf{1}\right)$, which is best approximation of $\boldsymbol{\eta}^{(i)}$ along vector $\mathbf{1}$. Square root of sample variance is called *sample standard deviation.*

Now consider data obtained for two different random variables, say $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varphi}_k$. A measure of linear association between these two variables can be estimated as

$$s_{i,k} = \frac{1}{N-1} \sum_{j=1}^{N} \left[\phi_{j,i} - \widehat{\overline{\boldsymbol{\varphi}}}_i\right] \left[\phi_{j,k} - \widehat{\overline{\boldsymbol{\varphi}}}_k\right] = \frac{1}{N-1} \left[\mathbf{e}^{(i)}\right]^T \mathbf{e}^{(k)} \tag{6.16}$$

$$i = 1, 2, ...m \quad ; \quad i = 1, 2, ...m$$

Here $s_{i,k}$ are $(i,k)^{th}$ elements of *sample covariance* matrix $\mathbf{S}_{\boldsymbol{\varphi}}$ of the random variable vector $\boldsymbol{\varphi}$. Alternatively, this matrix can also be estimated as

$$\mathbf{S}_{\boldsymbol{\varphi}} = Cov(\boldsymbol{\varphi}) = \frac{1}{N-1}\sum_{j=1}^{N}\left[\boldsymbol{\varphi}^{(j)} - \widehat{\overline{\boldsymbol{\varphi}}}\right]\left[\boldsymbol{\varphi}^{(j)} - \widehat{\overline{\boldsymbol{\varphi}}}\right]^T \tag{6.17}$$

$$= \frac{1}{N-1}\left[\Phi - \mathbf{1}\left(\widehat{\overline{\boldsymbol{\varphi}}}\right)^T\right]^T\left[\Phi - \mathbf{1}\left(\widehat{\overline{\boldsymbol{\varphi}}}\right)^T\right] \tag{6.18}$$

It may be noted that sample covariance matrix $\mathbf{S}_{\boldsymbol{\varphi}}$ is an estimated of population covariance matrix $\boldsymbol{\Sigma}$.

It may be noted that Finally*, sample correlation coefficients* (normalized or standardized covariances) are defined as

$$\begin{aligned} r_{i,k} &= \frac{\sum_{j=1}^{N}\left[\phi_{j,i} - \widehat{\overline{\boldsymbol{\varphi}}}_i\right]\left[\phi_{j,k} - \widehat{\overline{\boldsymbol{\varphi}}}_k\right]}{\sqrt{\sum_{j=1}^{N}\left[\phi_{j,i} - \widehat{\overline{\boldsymbol{\varphi}}}_i\right]^2}\sqrt{\sum_{j=1}^{N}\left[\phi_{j,k} - \widehat{\overline{\boldsymbol{\varphi}}}_k\right]^2}} \\ &= \frac{s_{i,k}}{\sqrt{s_{i,i}}\sqrt{s_{k,k}}} = \frac{s_{i,k}}{\sqrt{s_i}\sqrt{s_k}} \\ &= \frac{\left[\mathbf{e}^{(i)}\right]^T\mathbf{e}^{(k)}}{\sqrt{\left[\mathbf{e}^{(i)}\right]^T\mathbf{e}^{(i)}}\sqrt{\left[\mathbf{e}^{(k)}\right]^T\mathbf{e}^{(k)}}} = \cos(\theta_{\iota,k}) \\ i &= 1,2,...m \quad ; \quad i = 1,2,...m \end{aligned} \tag{6.19}$$

where s_i and $s_k$represents sample standard deviations of $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varphi}_k$, respectively, and $r_{i,k}$ represents $(i,k)^{th}$ element of *sample correlation matrix* $\mathbf{R}_{\boldsymbol{\varphi}}$. Here $\theta_{\iota,k}$ is angle between vectors $\mathbf{e}^{(i)}$ and $\mathbf{e}^{(k)}$. From the above equations it follows that

$$r_{i,k} = \cos(\theta_{\iota,k}) \Rightarrow -1 \leq r_{i,k} \leq 1 \tag{6.20}$$

If two deviation vectors $\mathbf{e}^{(i)}$ and $\mathbf{e}^{(k)}$ have nearly same orientation, the sample correlation coefficient will be close to 1. If two deviation vectors $\mathbf{e}^{(i)}$ and $\mathbf{e}^{(k)}$ are nearly perpendicular, the sample correlation coefficient will be close to 0. if two deviation vectors $\mathbf{e}^{(i)}$ and $\mathbf{e}^{(k)}$ have nearly opposite orientation, the sample correlation coefficient will be close to -1. Thus, $r_{i,k}$ is a measure of linear association between two random variables.

$$\begin{array}{rc} r_{i,k} = 0 \Rightarrow & \text{Indicates Lack of } \textbf{linear} \text{ association between } \boldsymbol{\varphi}_i \text{ and } \boldsymbol{\varphi}_k \\ r_{i,k} < 0 \Rightarrow & \begin{array}{c}\text{tendency for one value in pair to be larger than} \\ \text{its average when the other is smaller than its average}\end{array} \\ r_{i,k} > 0 \Rightarrow & \begin{array}{c}\text{tendency for one value in pair to be large when the} \\ \text{other is large and for both values to be small together}\end{array} \end{array}$$

It may be noted that two random variables having **nonlinear** association may have $r_{i,k} = 0$ i.e. lack of **linear** association. Thus, $r_{i,k} = 0$ implies only lack of **linear** association and **not** lack of association between two random variables.

The sample correlation matrix $\mathbf{R}_{\boldsymbol{\varphi}}$ can also be estimated as

$$\mathbf{R}_{\boldsymbol{\varphi}} = \mathbf{D}^{-1/2}\mathbf{S}_{\boldsymbol{\varphi}}\mathbf{D}^{-1/2} \tag{6.21}$$

where

$$\mathbf{D}^{1/2} = diag\left[\begin{array}{cccc} \sqrt{s_1} & \sqrt{s_2} & ... & \sqrt{s_m} \end{array}\right] \tag{6.22}$$

Suppose we define a variable transformation where $\phi_{i,j}$ are replaced by *normalized* or *standardized* variables defined as

$$\upsilon_{i,j} = \frac{\left(\phi_{i,j} - \widehat{\overline{\boldsymbol{\varphi}}}_j\right)}{\sqrt{s_j}} \tag{6.23}$$

Then the vector $\boldsymbol{\upsilon}$ of these scaled random variables will have zero mean and its sample covariance will be identical to its sample correlation matrix.

6.2. Commonly used probability distributions.

6.2.1. *Multivariable Gaussian / Normal Distribution.* Consider a vector or random variables $\boldsymbol{\varphi}$ with *multivariate normal* (Gaussian) distribution, which is generalization of univariate normal density to dimensions higher that one. Let $\overline{\boldsymbol{\varphi}}$ and Σ denote population mean and population covariance matrix, respectively. Then, the probability density function for this distribution has form

$$f(\boldsymbol{\varphi}) = C \exp\left[\|(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})\|_{2,\Sigma^{-1}}\right] \tag{6.24}$$

where C is a constant and $\|(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})\|_{2,\Sigma^{-1}}$ is a normalized distance measure of vector $\boldsymbol{\varphi}$ from its mean defined as

$$\|(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})\|_{2,\Sigma^{-1}} = (\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})^T \Sigma^{-1} (\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}}) \tag{6.25}$$

Note that Σ is a positive definite and symmetric matrix. From the requirement that

$$\int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} f(\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_m) d\boldsymbol{\varphi}_1 ... d\boldsymbol{\varphi}_m = 1 \tag{6.26}$$

it follows that

$$C = \frac{1}{(2\pi)^{m/2}\left[\det(\Sigma)\right]^{1/2}} \tag{6.27}$$

Thus, multivariate Gaussian distribution has form

$$f(\boldsymbol{\varphi}) = \frac{1}{(2\pi)^{m/2}\left[\det(\Sigma)\right]^{1/2}} \exp\left[(\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})^T \Sigma^{-1} (\boldsymbol{\varphi}-\overline{\boldsymbol{\varphi}})\right] \tag{6.28}$$

In the univariate case, this reduces to familiar form

$$f(\varphi) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[\frac{(\varphi - \overline{\varphi})}{\sigma^2}\right] \tag{6.29}$$

6.2.2. *Chi-square (χ^2) distribution.* This distribution is used in connection with

- Testing goodness of fit of experimental observations to hypothesized probability distributions
- Obtaining confidence limits for the variance and the standard deviations
- Testing independence of variables.

Let $(\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_m)$ represent a set of m independent normally distributed random variables with parameters $(\mu_1, \sigma_1^2), ..., (\mu_m, \sigma_m^2)$. If we calculate the squares of the standard normal variables

$$\mathbf{u}_i^2 = \left(\frac{\boldsymbol{\varphi}_i - \overline{\boldsymbol{\varphi}}_i}{\sigma_i}\right)^2 \tag{6.30}$$

and sum the $\mathbf{u}_i^2 s$, then we have a new random variable χ^2 as follows

$$\chi^2 = \sum_{i=1}^{m} \mathbf{u}_i^2 \tag{6.31}$$

Here, m is called *degrees of freedom* for χ^2. The distribution of χ^2 depends only on m because $\mathbf{u}_i$ are standardized. The probability density function for χ^2 can be shown to be

$$\begin{aligned} p\left(\chi^2\right) &= \frac{1}{(2)^{m/2}\Gamma(m/2)} \left(\chi^2\right)^{\frac{m}{2}-1} \exp\left(-\frac{\chi^2}{2}\right) \\ 0 &< \chi^2 < \infty \end{aligned} \tag{6.32}$$

For $m > 30$, $\sqrt{2\chi^2}$ is approximately distributed as a normal variable with $\mu = \sqrt{2m-1}$ and $\sigma^2 = 1$.

6.2.3. *Student t distribution.* Given a random variable φ, the random variable t represents the ratio of two independent ransom variables, $\mathbf{u}$, and $\sqrt{\frac{\chi^2}{m}}$

$$t = \frac{\mathbf{u}}{\sqrt{\frac{\chi^2}{m}}} = \frac{\mathbf{u}}{s_\varphi / \sigma_\varphi} = \frac{\widehat{\overline{\varphi}} - \overline{\varphi}}{s_{\overline{\varphi}}} \tag{6.33}$$

where $\widehat{\overline{\varphi}}$ is the sample mean and $s_{\overline{\varphi}}$ is the sample standard deviation. The probability density function for t is

$$p(t) = \frac{1}{\sqrt{\pi m}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \left[1 + \frac{t^2}{m}\right]^{-\frac{m+1}{2}} \tag{6.34}$$

where m is the degrees of freedom associated with s_{φ}^2.

6.2.4. *Fisher distribution.* If two samples are taken, one consisting of n_1 independent measurements of a normal random variable φ_1, which has mean μ_1 and variance σ_1^2, and other sample consisting of n_2 independent measurements of a normal random variable φ_2, which has mean μ_2 and variance σ_2^2, then the random variable F is defined as

$$F(n_1-1,n_2-1)=\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{6.35}$$

with degrees of freedom (n_1-1) and (n_2-1). If $\sigma_1^2=\sigma_2^2=\sigma^2$, then F can be related to χ^2 as

$$F(n_1-1,n_2-1)=\frac{s_1^2}{s_2^2}=\frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)} \tag{6.36}$$

The probability density function of F is given by

(6.37)

$$p(F)=\frac{\Gamma\left(\frac{n_1+n_2-2}{2}\right)}{\Gamma\left(\frac{n_1-1}{2}\right)\Gamma\left(\frac{n_2-1}{2}\right)}\left[n_1-1\right]^{\frac{n_1-1}{2}}\left[n_2-1\right]^{\frac{n_2-1}{2}}\frac{F^{\left(\frac{n_1+1}{2}\right)}}{\left[(n_2-1)+(n_1-1)F\right]^{\left(\frac{n_1+n_2-2}{2}\right)}}$$

Ensemble mean and variances of F are

$$\mathbf{E}\left(F\right) = \frac{n_2-1}{n_2-3} \text{ for } n_2>2 \tag{6.38}$$

$$Var(F) = \frac{2(n_2-1)^2(n_1+n_2-4)}{(n_1-1)(n_2-3)^2(n_2-5)} \tag{6.39}$$

6.3. Statistical Interpretation of Linear Least Squares. In order to interpret the linear least square approximation from statistical viewpoint, we make following *assumptions :*

- The weighting matrix is an identity matrix, i.e. $W = I$. In other words, all the measurements are assumed to be equally likely.
- Only equation / modelling errors are present. Alternatively, measurement errors are present only in dependent variable and there are no measurement errors in independent variables.

We further hypothesize that

- Error vector $\mathbf{e}=\ \mathbf{y}-\widehat{\mathbf{y}}$ has a Gaussian / normal distribution $\mathbf{N}(0,\sigma^2I)$.i.e.,

$$mean(\mathbf{e}) = E(\mathbf{e}) = \overline{0} \quad \text{and} \quad cov(\mathbf{e}) = \sigma^2 I \tag{6.40}$$

Note that the above covariance expression implies that each element e_i of vector $\mathbf{e}$, are independent and normally distributed variables such that

$$y_i = \left[\boldsymbol{\varphi}^{(i)}\right]^T \boldsymbol{\theta} + e_i \qquad \text{for} \quad i = 1, 2...N \tag{6.41}$$

$$cov(e_i, e_j) = 0 \text{ when } i \neq j \;\; ; \;\; i = 1, 2...N, \; j = 1, 2...N \tag{6.42}$$

$$var(e_i) = \sigma^2 \quad \text{for} \quad i = 1, 2...N \tag{6.43}$$

- The parameter vector $\boldsymbol{\theta}$ and the error vector $\mathbf{e}$ are not correlated.

Thus, statistical model for experimental data given by equation (6.41) can be expressed as

$$\mathbf{y} = \Phi\boldsymbol{\theta}_T + \mathbf{e} \tag{6.44}$$

$$\mathbf{\Phi} = \begin{bmatrix} \left[\boldsymbol{\varphi}^{(1)}\right]^T \\ .... \\ \left[\boldsymbol{\varphi}^{(N)}\right]^T \end{bmatrix} \tag{6.45}$$

where $\boldsymbol{\theta}_T$ represent the **true** parameter values and vectors $\mathbf{y}$ and $\mathbf{e}$ have been defined by equations (4.12) and (4.13), respectively. Taking expectation (mean) on both sides of the above equation., we have

$$E(\mathbf{y}) = E(\mathbf{\Phi}\boldsymbol{\theta}_T + \mathbf{e}) = \mathbf{\Phi}\boldsymbol{\theta}_T \tag{6.46}$$

From least square analysis, we have least square solution , given as

$$\widehat{\boldsymbol{\theta}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{6.47}$$

$$E(\widehat{\boldsymbol{\theta}}) = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T E(\mathbf{y}) = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T E(\mathbf{\Phi}\boldsymbol{\theta}_T + \mathbf{e}) \tag{6.48}$$

$$= (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Phi}\boldsymbol{\theta}_T \;\; = \boldsymbol{\theta}_T \tag{6.49}$$

The above result guarantees that, if we collect sufficiently large number of samples, the least square estimate will approach the true values of the model parameters. In statistical terms, $\widehat{\boldsymbol{\theta}}$ is an **unbiased estimate** of $\boldsymbol{\theta}_T$. To calculate covariance of $\widehat{\boldsymbol{\theta}}$, we proceed as follows

$$\mathbf{V} = cov\;(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}) = E\;[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T)^T] \tag{6.50}$$

Now

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T(\mathbf{\Phi}\boldsymbol{\theta}_T + \mathbf{e}) - \boldsymbol{\theta}_T \tag{6.51}$$

$$= \boldsymbol{\theta}_T + \; (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e} - \boldsymbol{\theta}_T \tag{6.52}$$

$$= (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e} \tag{6.53}$$

This implies

$$\mathbf{V} = cov(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}) = E\;[\;(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\;\mathbf{\Phi}^T\;(\mathbf{e}\mathbf{e}^T)\;\mathbf{\Phi}\;(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\;] \tag{6.54}$$

$$= (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\;\mathbf{\Phi}^T\;E(\mathbf{e}\mathbf{e}^T)\;\mathbf{\Phi}\;(\mathbf{\Phi}^T\mathbf{\Phi})^{-1} \tag{6.55}$$

Using the hypothesis

$$E(\mathbf{e}\mathbf{e}^T) \;=\; \sigma^2 I \tag{6.56}$$

we have

$$\mathbf{V} = cov(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\theta}}) \;=\; \sigma^2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \tag{6.57}$$

Thus, covariance matrix of regression coefficients $\widehat{\boldsymbol{\theta}}$ is proportional to $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$. If $\mathbf{N}$ is large, there is negligible bias in the estimation of σ^2. The knowledge of $cov(\widehat{\boldsymbol{\theta}})$ is important because it constitutes an estimation on precision with which each θ_i are determined. Note that diagonal elements of $cov(\widehat{\boldsymbol{\theta}})$ are variances of individual parameters θ_i, i.e.

$$var\left[\theta_i\right] = cov(\theta_i,\theta_i) = \mathbf{V}_{ii} = \left[\sigma^2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\right]_{ii} \tag{6.58}$$

Now, since $\mathbf{e}$ is zero mean and normally distributed vector, it can be shown that $\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_T\right)$ is also a zero mean and normally distributed vector with the covariance matrix given by equation (6.57). Thus, given matrix $\boldsymbol{\Phi}$ and σ^2,we can determine **confidence interval** of each parameter using the parameter variances..

In general, σ^2 may not be known *apriori* for a given set of data. However, an estimate of σ^2 (denoted as $\widehat{\sigma}^2$) can be obtained from the estimated error vector as

$$\widehat{\sigma}^2 \;=\; \frac{1}{\mathbf{N}-\mathbf{m}}\sum_{i=1}^{\mathbf{N}}\widehat{e}_i^2 \;=\; \frac{\widehat{\mathbf{e}}^T\widehat{\mathbf{e}}}{\mathbf{N}-\mathbf{m}} \tag{6.59}$$

$$\widehat{\mathbf{e}} \;=\; \mathbf{y}-\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} \tag{6.60}$$

where $\widehat{\boldsymbol{\theta}}$ is the least square estimate. Estimate of $\mathbf{V}$ (denoted as $\widehat{\mathbf{V}}$) can be computed as

$$\widehat{\mathbf{V}} = \left(\frac{\widehat{\mathbf{e}}^T\widehat{\mathbf{e}}}{\mathbf{N}-\mathbf{m}}\right)(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \tag{6.61}$$

To appreciate the validity of uniformity of a model after smoothing, normally we calculate the correlation coefficient R.

$$R^2 \;= \frac{\mathbf{y}^T\,\widehat{\mathbf{y}}}{\mathbf{y}^T\,\mathbf{y}} \;=\; \frac{\mathbf{y}^T\,(\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}})}{\mathbf{y}^T\,\mathbf{y}} \tag{6.62}$$

This quantity indicates the ratio of response $\mathbf{y}$ known from the model. For example, if $R^2 = 0.95$ then we can say that 95% of variation of y_i is known from $\widehat{y}_i$. Thus, this correlation coefficient can be used to compare different models developed using same data. Further it can be shown that the quantity

$$f = \frac{N-m}{m+1}\left(\frac{R^2}{1-R^2}\right) \tag{6.63}$$

has Fisher distribution with $(m+1)$ and $(N-m)$ degrees of freedom. When regression is significant, we can test the hypothesis H_0 by rejecting H_0 with risk of $\alpha\%$. The test statistic is with risk of $\alpha\%$ can be computed as

$$\varepsilon = \mathbf{F}_\alpha(m+1, N-m) \tag{6.64}$$

where $\mathbf{F}$ denotes Fisher distribution. If $f > \varepsilon$,then we conclude that f is probably not a the Fisher distribution and the model is not suitable. If $f < \varepsilon$,then smoothing is significant and the model is satisfactory.

In many situations, the model parameters have physical meaning. In such cases, it is important to determine confidence interval for parameter θ_i. Defining variable

$$t_i = \frac{\widehat{\theta}_i - \theta_i}{\sqrt{\widehat{V}_{ii}}} \tag{6.65}$$

where θ_i is the true value of parameter and $\widehat{\theta}_i$ is the estimated value of parameter, it can be shown that t_i is a t (Student) distribution with $N-m$ degrees of freedom. Thus, confidence interval with the risk of $\alpha\%$ for θ_i is

$$\widehat{\theta}_i - \sqrt{\widehat{V}_{ii}}\, t(\alpha/2, N-m) < \theta_i < \widehat{\theta}_i + \sqrt{\widehat{V}_{ii}}\, t(\alpha/2, N-m) \tag{6.66}$$

In many situations, when we are fitting a functional form to explain variation in experimental data, we do not know *apriori* the exact function form (e.g. polynomial order) that fits data best. In such a case, we need to assess whether a particular term $f_j(\mathbf{x})$ contributes significantly to $\widehat{y}$ or otherwise. In order to measure the importance of the contribution of θ_i in

$$\widehat{y} = \sum_{j=1}^{m} \theta_j \; f_j(\mathbf{x}) \tag{6.67}$$

one can test hypothesis H_0 namely: $\theta_i = 0$ (i.e. $f_i(\mathbf{x})$ does not influence $\widehat{y}$). Thus, we have

$$\tau_i = \frac{\widehat{\theta}_i - 0}{\widehat{\sigma}\sqrt{\widehat{V}_{ii}}} \tag{6.68}$$

If

$$-t(\alpha/2, N-m) < \tau_i < t(\alpha/2, N-m) \tag{6.69}$$

then $f_i(\mathbf{x})$ has significant influence on $\widehat{y}$. If not, then we can remove term $f_i(\mathbf{x})$ from $\widehat{y}$.

7. Simple Polynomial Models and Hilbert Matrices

7.1. Approximation of Continuous Function by a Polynomial. Consider problem of approximating a continuous function, say $f(z)$, over interval $[0,1]$ by a simple polynomial model of the form

$$\begin{aligned} (7.1)\quad \widehat{y}(z) &= \theta_1\mathbf{p}^{(1)}(z) + \theta_2\mathbf{p}^{(2)}(z) + \theta_3\mathbf{p}^{(3)}(z) + ... + \theta_m\mathbf{p}^{(m)}(z) \\ &= \theta_1 + \theta_2 z + \theta_3 z^2 + ... + \theta_m z^{m-1} \end{aligned}$$

Let the inner product on $C_2[0,1]$ is defined as

$$\langle h(z), g(z)\rangle = \int\limits_0^1 h(z)g(z)dz \tag{7.2}$$

We want to find a polynomial of the form (7.1) which approximates $f(z)$ in the least square sense. Geometrically, we want to project $f(z)$ in the m dimensional subspace of $C_2[0,1]$ spanned by vectors

$$\mathbf{p}^{(1)}(z) = 1;\ \ \mathbf{p}^{(2)}(z) = z\ \ ;\ \ \mathbf{p}^{(3)}(z) = z^2, ..., \mathbf{p}^{(m)}(z) = z^{m-1} \tag{7.3}$$

Using projection theorem, we get the normal equation

$$\begin{bmatrix} \langle 1,1\rangle & \langle 1,z\rangle & .... & \langle 1,z^{m-1}\rangle \\ \langle z,1\rangle & \langle z,z\rangle & .... & \langle z,z^{m-1}\rangle \\ ..... & ..... & ..... & ..... \\ \langle z^{m-1},1\rangle & \langle z^{m-1},z\rangle & ..... & \langle z^{m-1},z^{m-1}\rangle \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ .... \\ \theta_m \end{bmatrix} = \begin{bmatrix} \langle 1,f(z)\rangle \\ \langle z,f(z)\rangle \\ .... \\ \langle z^{m-1},f(z)\rangle \end{bmatrix} \tag{7.4}$$

Element h_{ij} of the matrix on L.H.S. can be computed as

$$h_{ij} = \left\langle x^{(i)}(z), x^{(j)}(z)\right\rangle = \int\limits_0^1 z^{j+i-2}\ dz\ = \frac{1}{i+j-1} \tag{7.5}$$

and this reduces the above equation to

$$H\begin{bmatrix} \theta_1 \\ \theta_2 \\ .... \\ \theta_m \end{bmatrix} = \begin{bmatrix} \langle 1,f(z)\rangle \\ \langle z,f(z)\rangle \\ .... \\ \langle z^{m-1}, f(z)\rangle \end{bmatrix} \tag{7.6}$$

where

$$H = \begin{bmatrix} 1 & 1/2 & 1/3 & ... & ... & 1/m \\ 1/2 & 1/3 & 1/4 & ... & ... & 1/(m+1) \\ ... & ... & ... & ... & ... & ... \\ 1/m & ... & ... & ... & ... & 1/(2m-1) \end{bmatrix} \tag{7.7}$$

The matrix H is known as Hilbert matrix and this matrix is highly ill-conditioned for $m > 3$. The following table shows condition numbers for a few values of m.

(7.8)

m	3	4	5	6	7	8
$c(H)$	524	1.55e4	4.67e5	1.5e7	4.75e8	1.53e10

Thus, for polynomial models of small order, say $m = 3$ we obtain good situation, but beyond this order, what ever be the method of solution, we get approximations of less and less accuracy. This implies that approximating a continuous function by polynomial of type (7.1) with the choice of basis vectors as (7.3) is extremely ill-conditioned problem from the viewpoint of numerical computations. Also, note that if we want to increase the degree of polynomial to say $(m+1)$from m, then we have to recompute $\theta_1, ..., \theta_m$ along with θ_{m+1}.

On the other hand, consider the model

$$\widehat{y}(z) = \alpha_1 \mathbf{p}_1(z) + \alpha_2 \mathbf{p}_2(z) + \alpha_3 \mathbf{p}_3(z) + ... + \alpha_m \mathbf{p}_m(z) \tag{7.9}$$

where $\mathbf{p}_i(z)$ represents the i'th order orthonormal basis function on $C_2[0,1]$.This corresponds to choice of basis functions as

$$x^{(1)}(z) = \mathbf{p}_1(z); \;\; x^{(2)}(z) = \mathbf{p}_2(z) \;\; ;........, x^{(m)}(z) = \mathbf{p}_m(z) \tag{7.10}$$

and since

$$\langle \mathbf{p}_i(z), \mathbf{p}_j(z) \rangle = \left\{ \begin{array}{ccc} 1 & if & i = j \\ 0 & if & i \neq j \end{array} \right\} \tag{7.11}$$

the normal equation reduces to

$$\begin{bmatrix} 1 & 0 & .... & 0 \\ 0 & 1 & .... & 0 \\ ..... & ..... & ..... & ..... \\ 0 & 0 & ..... & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ .... \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \langle \mathbf{p}_1(z), f(z) \rangle \\ \langle \mathbf{p}_2(z), f(z) \rangle \\ .... \\ \langle \mathbf{p}_m(z), \; f(z) \rangle \end{bmatrix} \tag{7.12}$$

or simply

$$\alpha_i = \langle \mathbf{p}_i(z), f(z) \rangle \quad ; \quad i = 1, 2, ...m \tag{7.13}$$

Obviously, the approximation problem is extremely well conditioned in this case. In fact, if we want to increase the degree of polynomial to say $(m+1)$from m, then we do not have to recompute $\alpha_1, ..., \alpha_m$ as in the case basis (7.3) where vectors are linearly independent but not orthogonal. We simply have to compute the α_{m+1} as

$$\alpha_{m+1} = \langle \mathbf{p}_{m+1}(z), f(z) \rangle \tag{7.14}$$

The above illustration of approximation of a function by orthogonal polynomials is a special case of what is known as generalized Fourier series expansion.

7.2. Approximation of Numerical Data by a Polynomial. Suppose we only know numerical $\{y_1, y_2, ...y_N\}$ at points $\{z_1, z_2, ...z_N\} \in [0,1]$ and we want to develop a simple polynomial model of the form given by equation (7.1). Substituting the data into the polynomial model leads to an overdertermined

$$y_i = \theta_1 + \theta_2 z_i + \theta_3 z_i^2 + ... + \theta_m z_i^{m-1} + e_i \tag{7.15}$$

$$i = 1, 2, ...N \tag{7.16}$$

The least square estimates of the model parameters (for $W = I$) can be obtained by solving normal equation

$$(\Phi^T \Phi)\hat{\boldsymbol{\theta}} = \Phi^T \mathbf{y} \tag{7.17}$$

where

$$\Phi = \begin{bmatrix} 1 & z_1 & z_1^2 & ... & z_1^{m-1} \\ ... & ... & ... & ... & .... \\ 1 & z_{\mathbf{N}} & z_{\mathbf{N}}^2 & ... & z_{\mathbf{N}}^{m-1} \end{bmatrix} \tag{7.18}$$

$$\Phi^T \Phi = \begin{bmatrix} \mathbf{N} & \sum z_i & \sum z_i^2 & .... & \sum z_i^{m-1} \\ \sum z_i & \sum z_i^2 & ..... & .... & \sum z_i^m \\ ..... & ..... & ..... & .... & .... \\ ..... & ..... & ..... & .... & .... \end{bmatrix} \tag{7.19}$$

i.e.,

$$(\Phi^T \Phi)_{ik} = \sum_{i=1}^{\mathbf{N}} z_i^{j+k-2} \tag{7.20}$$

Let us assume that z_i is uniformly distributed in interval $[0,1]$. For large N, approximating $dz \simeq 1/\mathbf{N}$, we can write

$$[\Phi]_{jk} = \sum_{i=1}^{\mathbf{N}} z_i^{j+k-2} \cong \mathbf{N} \int_0^1 z^{j+k-2}\, dz = \frac{\mathbf{N}}{j+k-1} \tag{7.21}$$

$$(\ j, k = 1, 2, ...m \ ) \tag{7.22}$$

Thus, we can approximate $(\Phi^T\Phi)$ matrix by the Hilbert matrix

$$(7.23)\qquad (\Phi^T\Phi) = \mathbf{N}(H) \;=\; \mathbf{N}\begin{bmatrix} 1 & 1/2 & 1/3 & ... & ... & 1/m \\ 1/2 & 1/3 & 1/4 & ... & ... & 1/(m+1) \\ ... & ... & ... & ... & ... & ... \\ 1/m & ... & ... & ... & ... & 1/(2m-1) \end{bmatrix}$$

which is highly ill- conditioned for large m. Thus, whether we have a continuous function or numerical data over interval $[0,1]$, the numerical difficulties persists as the Hilbert matrix appears in both the cases.

Similar to the previous case, modelling in terms of orthogonal polynomials can considerably improve numerical accuracy. The orthogonal set under consideration is now different. Let $p_m(z_i)$ denote a orthogonal polynomial of order m. The inner product is now defined as

$$(7.24)\qquad \langle \mathbf{p}_j(z)\ ,\mathbf{p}_k(z)\rangle = \sum_{i=1}^{\mathbf{N}} w_i\ \mathbf{p}_j(z_i)\ \mathbf{p}_k(z_i)$$

By definition, a set of polynomials $\{\mathbf{p}_j(z_i)\}$are orthogonal over a set of points $\{z_i\}$ with weights w_i, if

$$\sum_{i=1}^{\mathbf{N}} w_i\ \mathbf{p}_j(z_i)\ \mathbf{p}_k(z_i) = 0 \quad ; \qquad j,k = 1...............m \ \text{ and } (\ j \neq\ k)$$

Let a linear polynomial model be defined as

$$(7.25)\qquad y_i \;=\; \sum_{j=1}^{\mathbf{m}} \alpha_j\ \mathbf{p}_j(z_i)\ +\ e_i \;\;;\;\; i = 1,..................\mathbf{N}$$

Then the normal equation becomes

$$(7.26)$$

$$\begin{bmatrix} \sum\limits_{i=1}^{\mathbf{N}} w_i\ \mathbf{p}_1^2(z_i) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum\limits_{i=1}^{\mathbf{N}} w_i\ \mathbf{p}_m^2(z_i) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{\mathbf{N}} w_i\ y_i\ p_1(z_i) \\ \vdots \\ \sum\limits_{i=1}^{\mathbf{N}} w_i\ y_i\ p_m(z_i) \end{bmatrix}$$

Thus, each term $\alpha_j\ \mathbf{p}_j(z_i)$ is (statistically) independent of any other term and contains information that other terms do not have and the resulting parameter estimation problem is highly well conditioned. A set of orthogonal polynomials can be obtained by different approaches. One simple algorithm is

INITIALIZE

$$p_{-1}(z_i) = 0$$
$$p_0(z_i) = 1 \qquad (for\ \ i =\ 1,............\mathbf{N})$$

```
          $\Psi_0 = 0$
FOR   $(k = 0, 1, ............m-1)$
     FOR   $(i = 0, 1, ............N)$
          $\Psi_k = \dfrac{\sum_{i=1}^{\mathbf{N}} w_i \; [\; p_k(z_i) \;]^2}{\sum_{i=1}^{\mathbf{N}} w_i \; [\; p_{k-1}(z_i) \;]^2}$
          $\Gamma_{k+1} = \dfrac{\sum_{i=1}^{\mathbf{N}} w_i \; z_i[\; p_k(z_i) \;]^2}{\sum_{i=1}^{\mathbf{N}} w_i \; [\; p_k(z_i) \;]^2}$
          $p_{k+1}(z_i) = [z_i - \Gamma_{k+1}] \; p_k(z_i) - [\Psi_k] \; \rho_{k-1}(z_i)$
     END FOR
END FOR
```

8. Nonlinear in Parameter Models

In many problems the parameters appear nonlinearly in the model

$$\widehat{y}_i = f\left(\mathbf{x}^{(i)} \; ; \; \theta_1........, \theta_m\right) \quad ; \quad (i = 1, ...\mathbf{N}) \tag{8.1}$$

or in the vector notation

$$\widehat{\mathbf{y}} = F\left[\mathbf{X}, \boldsymbol{\theta}\right] \tag{8.2}$$

where

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 & \widehat{y}_2 & .... & \widehat{y}_N \end{bmatrix}^T \tag{8.3}$$

$$F = \begin{bmatrix} f\left(\mathbf{x}^{(1)}, \boldsymbol{\theta}\right) & f\left(\mathbf{x}^{(2)}, \boldsymbol{\theta}\right) & .... & f\left(\mathbf{x}^{(N)}, \boldsymbol{\theta}\right) \end{bmatrix}^T \tag{8.4}$$

and $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\right\}$represents data set. The problem is to determine vector$\widehat{\boldsymbol{\theta}}$ such that

$$\Psi = \mathbf{e}^T\mathbf{W}\mathbf{e} \tag{8.5}$$

$$\mathbf{e} = \widehat{\mathbf{y}} - F\;(\mathbf{X}, \boldsymbol{\theta}) \tag{8.6}$$

is minimized. Note that. in general, the above problem cannot be solved analytically and we have to resort to iterative procedures. There are three solution approaches:

- Approximate solution using weighted least square when the model is analytically linearizable
- Successive linear least square approach (Gauss-Newton method)
- Use of direct optimization (nonlinear programming)

The first two approaches use the linear least square formulation as basis while the nonlinear programming approaches is a separate class of algorithms.

8.1. Weighted Least Square For Analytically Linearizable Models. Many nonlinear-in-parameter forms can be converted into linear-in-parameter forms by means of some linearizing transformations. In such cases, it appears that we can apply linear least square to obtain parameter estimates. However, this may not minimize $\mathbf{e}^T\mathbf{We}$ as explained below

Consider linearizable form of model

$$\widehat{y} = \theta_0 \ [\ f_1(\mathbf{x})]^{\theta_1} \ [\ f_2(\mathbf{x})]^{\theta_2} \qquad \cdots\cdots \ [f_m(\mathbf{x})]^{\theta_m} \tag{8.7}$$

which can be transformed as

$$\ln \ \widehat{y} = \ln \theta_1 \ + \ \theta_2 \ln [\ f_1(\mathbf{x})] \ + \ \theta_3 \ln [\ f_2(\mathbf{x})] \ \ \cdots \ \ \theta_m \ln [\ f_m(\mathbf{x})] \tag{8.8}$$

Now, parameter set that minimizes

$$\widetilde{\Psi} = \sum_{i=1}^{\mathbf{N}} (\ \ln y_i - \ln \widehat{y}_i)^2 \ = \sum_{i=1}^{\mathbf{N}} (\widetilde{e}_i)^2 \tag{8.9}$$

may not minimize

$$\Psi = \sum_{i=1}^{\mathbf{N}} (\ y_i - \widehat{y}_i)^2 \ = \sum_{i=1}^{\mathbf{N}} (e_i)^2 \tag{8.10}$$

A rigorous approach to avoid this problem is to use nonlinear programming of Gauss-Newton approach. However, it is often possible to use weighted coefficients w_i to account for the linearizing transformation.

Let e_i denote the error associated with nonlinear model and let $\widetilde{e}_i$ denote the error associated with transformed model .We want to approximate e_i^2 in terms of $\widetilde{e}_i^2$, i.e.

$$e_i^2 \simeq \widetilde{e}_i^2 \ w_i \tag{8.11}$$

Note that $\widetilde{e}_i$ is a complex function of e_i. Let us denote this function

$$\widetilde{e}_i \ \ = \ \ g(e_i) \qquad (i = 1, \cdots \mathbf{N}) \tag{8.12}$$

$$\Rightarrow \ \ d\widetilde{e}_i = \frac{\partial g}{\partial e_i} de_i \tag{8.13}$$

One possibility is to use Taylor series expansion of function $g(.)$ at point $(e_i = 0, \widetilde{e}_i = 0)$

$$\widetilde{e}_i \ \ \simeq \ \ g(0) + \left[\frac{\partial g}{\partial e_i}\right]_{(e_i=0)} (e_i - 0) \tag{8.14}$$

$$\simeq \ \ \left[\frac{\partial g}{\partial e_i}\right]_{(e_i=0)} e_i \tag{8.15}$$

Then, we can use

$$w_i = \left[\left(\left[\frac{\partial \bar{e}_i}{\partial e_i} \right]_{e_i=0} \right)^2_{\bar{e}_i=0} \right]^{-1} \tag{8.16}$$

$$(i = 1, ...\mathbf{N}) \tag{8.17}$$

For example, in the above case

$$\widetilde{e}_i = \ln y_i - \ln \widetilde{y} = \ln(\widetilde{y}_i + e_i) - \ln \widetilde{y}_i \tag{8.18}$$

$$\frac{\partial \bar{e}_i}{\partial e_i} = \frac{1}{\widetilde{y}_i + e_i} = \frac{1}{y_i} \tag{8.19}$$

$$\Rightarrow \quad w_i = (y_i)^2 \tag{8.20}$$

and optimization problem can be formulated as minimization of

$$\widetilde{\Psi} = \sum_{i=1}^{\mathbf{N}} (y_i)^2 (\ln y_i - \ln \widehat{y}_i)^2 = \sum_{i=1}^{\mathbf{N}} (y_i \widetilde{e}_i)^2 \tag{8.21}$$

The resulting weighted least square problem can be solved analytically.

8.2. Method of Successive Linearization (Gauss-Newton method). This approach is iterative. Start with an initial guess vector $\boldsymbol{\theta}^{(0)}$. By some process, generate improved guess $\boldsymbol{\theta}^{(k)}$ from $\boldsymbol{\theta}^{(k-1)}$. At k^{th} iteration let $\boldsymbol{\theta}^{(k-1)}$ be the guess solution. By expanding the model as Taylor series in the neighborhood of $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k-1)}$ and neglecting higher order terms we have

$$\widetilde{\mathbf{y}}^{(k)} \simeq F\left(\mathbf{X}, \boldsymbol{\theta}^{(k-1)}\right) + \left[\frac{\partial F}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k-1)}} \left(\triangle \boldsymbol{\theta}^{(k)}\right) \tag{8.22}$$

where

$$\mathbf{J}^{(k-1)} = \left[\frac{\partial F}{\partial \boldsymbol{\theta}}\right] \tag{8.23}$$

is a $(\mathbf{N} \times m)$ matrix with elements

$$\left[\frac{\partial F}{\partial \boldsymbol{\theta}}\right]_{ij} = \left[\frac{\partial F\left(\mathbf{x}^{(i)}, \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}_i}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k-1)}} \qquad i = 1, \ldots \mathbf{N} \text{ and } j = 1, \ldots m \tag{8.24}$$

Let us denote

$$\mathbf{J}^{(k-1)} = \left[\frac{\partial F}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k-1)}} \tag{8.25}$$

and

$$F^{(k-1)} = F\left(\mathbf{X}, \boldsymbol{\theta}^{(k-1)}\right) \tag{8.26}$$

Then approximate error vector at k^{th} iteration can be defined as

$$\widetilde{\mathbf{e}}^{(k)} = \mathbf{y} - \widetilde{\mathbf{y}}^{(k)} = \left[\mathbf{y} - F^{(k-1)}\right] - \mathbf{J}^{(k-1)} \triangle\boldsymbol{\theta}^{(k)} \tag{8.27}$$

and k^{th} linear sub-problem is defined as

$$\min_{\triangle\boldsymbol{\theta}^{(j)}} \left[\widetilde{\mathbf{e}}^{(k)}\right]^T \mathbf{W}\, \widetilde{\mathbf{e}}^{(k)} \tag{8.28}$$

The least square solution to above sub problem can be obtained by solving the normal equation

$$\left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W}\ \ \mathbf{J}^{(k-1)}\triangle\boldsymbol{\theta}^{(k)} = \left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W}\left[\mathbf{y} - F^{(k-1)}\right] \tag{8.29}$$

$$\triangle\boldsymbol{\theta}^{(k)} = \left[\left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W}\ \ \mathbf{J}^{(k-1)}\right]^{-1} \left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W}\left[\mathbf{y} - F^{(k-1)}\right] \tag{8.30}$$

and an improved guess can be obtained as

$$\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)} + \triangle\boldsymbol{\theta}^{(k)} \tag{8.31}$$

Termination criterion : Defining $\mathbf{e}^{(k)} = \mathbf{y} - F^{(k)}$ and

$$\Phi^{(k)} = \left[\mathbf{e}^{(k)}\right]^T \mathbf{W}\, \mathbf{e}^{(k)} \tag{8.32}$$

terminate iterations when $\Phi^{(k)}$changes only by a small amount, i.e.

$$\frac{\mid \Phi^{(k)} - \Phi^{(k-1)} \mid}{\mid \Phi^{(k)} \mid} < \varepsilon \tag{8.33}$$

Gauss Newton Algorithm:

```
INITIALIZE: $\boldsymbol{\theta}^{(0)}, \varepsilon_1, \varepsilon_2, \alpha, k, \delta_1, \delta_2, k_{\max}$
$\mathbf{e}^{(0)} = y - F\ [\ \mathbf{X}, \boldsymbol{\theta}^{(0)}\ ]$
$\delta_1 = \mathbf{e}^{(0)T} W \mathbf{e}^{(0)}$
WHILE $[(\delta_1 > \varepsilon_1)\ AND\ (\delta_2 > \varepsilon_2)\ AND\ (k < k_{max})]$
    $\mathbf{e}^{(k)} = y - F\ [\ \mathbf{X}, \boldsymbol{\theta}^{(k)}\ ]$
    $\left[J^{(k)T} W J^{(k)}\right] \triangle\, \boldsymbol{\theta}^{(k)} = J^{(k)T} W \mathbf{e}^{(k)})$
    $\lambda^{(0)} = 1,\ j = 0$
    $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda^{(0)} \triangle\, \boldsymbol{\theta}^{(k)}$
    $\delta_0 = \mathbf{e}^{(k)} W \mathbf{e}^{(k)}$
    $\delta_1 = \mathbf{e}^{(k+1)} W \mathbf{e}^{(k+1)}$
    WHILE$[\delta_1 > \delta_0]$
        $\lambda^{(j)} = \alpha \lambda^{(j-1)}$
        $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda^{(j)} \triangle\, \boldsymbol{\theta}^{(k)}$
        $\delta_1 = \mathbf{e}^{(k+1)} W \mathbf{e}^{(k+1)}$
    END WHILE
    $\delta_{2=} || \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} || \,/\, || \boldsymbol{\theta}^{(k+1)} ||$
    $k = k + 1$
```

END WHILE

9. Unconstrained Nonlinear Programming

Any iterative method for unconstrained optimization consists of two steps.

- **Step 1**: given a guess solution vector $\mathbf{x}^{(k)}$,find a search direction $\mathbf{s}^{(k)}$ to move in the direction where f(x) decreases (or increases).
- **Step 2:** After the direction $\mathrm{S}^{(k)}$ is determined, estimate the step length λ to calculate

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda \mathbf{s}^{(k)} \tag{9.1}$$

i.e. new guess. Typically, λ is selected by minimizing $f(\mathbf{x}^{(k)}+\lambda\mathbf{s}^{(k)})$ with respect to, i.e. by carrying out one dimensional optimization with respect to λ.

The iterations are terminated when either one of the following termination criteria are satisfied:

$$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{|f(x^{(k+1)})|} < \epsilon_1 \tag{9.2}$$

$$\frac{||(x^{(k+1)}) - (x^{(k)})||}{||(x^{(k+1)})||} < \epsilon_2 \tag{9.3}$$

The iterative nonlinear unconstrained optimization algorithms can be categorized as follows

- **Analytical Methods :** Solve $\nabla\Phi(\mathbf{z}) = \overline{0}$ analytically
- **Direct Methods:** need only computation of $\Phi(\mathbf{z})$ at each step
- **Gradient based methods:** Require computation of $\nabla\Phi(\mathbf{z})$ and $\Phi(\mathbf{z})$at each step
- **Hessian based methods:** require computation of $\nabla^2\Phi(\mathbf{z})$, $\nabla\Phi(\mathbf{z})$and $\Phi(\mathbf{z})$ at each step

In this section, we a brief introduction to some of the direct, gradient and Hessian based approaches.

9.1. Simplex Method (Direct Method). The geometric figure formed by $(n+1)$ points in n dimensional are called as simplex. Simplex in two dimensional space is a triangle and three dimensional space is tetrahedron. Basic idea in simplex method is to evaluate $f(\mathbf{x})$ at $(n+1)$ vertices of a simplex and use this information to move towards the optimum. Let $\mathbf{x}^{(m)}$ be such that

$$\mathbf{x}^{(m)} = \begin{array}{c} \max \\ i \in (1, ...n+1) \end{array} f(\mathbf{x}^{(i)}) \tag{9.4}$$

and $\mathbf{x}^{(c)}$ represent centroid of the remaining points defined as

$$\mathbf{x}^{(c)} = \frac{1}{n} \sum_{i=1, i \neq m}^{n+1} \mathbf{x}^{(i)} \tag{9.5}$$

Then, we find a new point $\mathbf{x}^{(new)}$ new such that by moving in direction of $\mathbf{x}^{(c)} - \mathbf{x}^{(m)}$ as follows

$$\mathbf{x}^{(new)} = \mathbf{x}^{(m)} + \lambda \left(\mathbf{x}^{(c)} - \mathbf{x}^{(m)} \right) \tag{9.6}$$

The choice of $\lambda = 1 \Rightarrow \mathbf{x}^{(new)} = \mathbf{x}^{(c)}$ and $\lambda = 0 \Rightarrow \mathbf{x}^{(new)} = \mathbf{x}^{(m)}$. In order to maintain the regularity of simplex, the reflection should be symmetric and $\lambda = 2$ is selected for achieving symmetry. The property of simplex used by algorithm is that a new simplex- can be generated on any face of the old one by projecting any chosen vertex a suitable distance through the centroid of the remaining vertices of old simplex. The new Simplex is then formed by replacing old vertices with newly generated projected point.

Initialization step involves generation of a regular simplex given a starting point $\mathbf{x}^{(0)}$ and a scale factor α. Remaining n vertices of a regular simplex (with equidistant neighboring points) are given by

$$\mathbf{x}_j^{(i)} = \left\{ \begin{array}{c} x_j^{(0)} + \delta_1 \; if \; j \neq i \\ x_j^{(0)} + \delta_2 \; if \; j = i \end{array} \right\} \tag{9.7}$$

$$\delta_1 = \left[\frac{(N+1)^{1/2} + N - 1}{N\sqrt{2}} \right] \alpha \tag{9.8}$$

$$\delta_2 = \left[\frac{(N+1)^{1/2} - 1}{N\sqrt{2}} \right] \alpha \tag{9.9}$$

Scale factor α is selected by user. Choice $\alpha = 1$ generates a simplex with unit length. If cycling occurs, i.e., a given vertex remains unchanged for more than M iterations , then the simplex size is reduced by some factor. A new simplex is setup with the currently lowest points as base points. This rule requires specification of some reduction factor. The search is terminated when

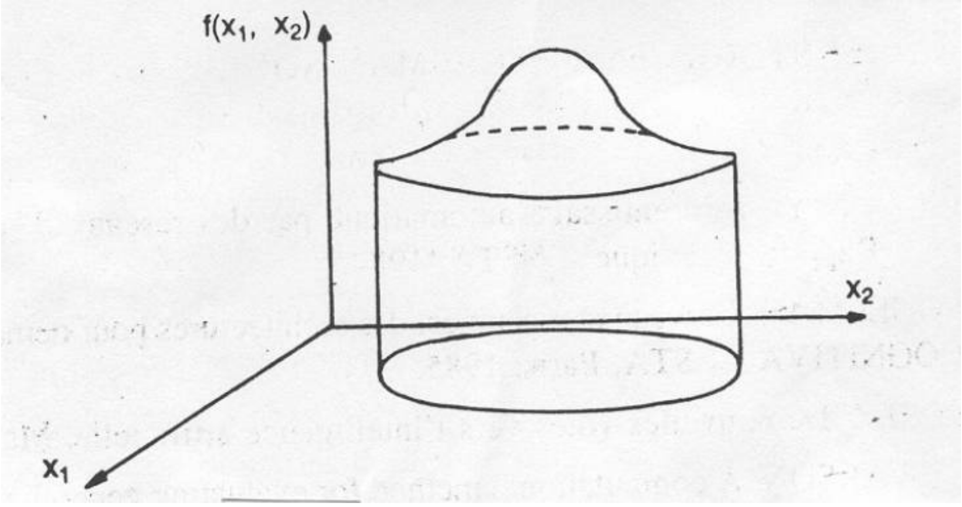


FIGURE 6. $f(\mathbf{x})$ curves for n = 2

the simplex gets small enough or else if

$$\sum_{i=1}^{n+1}\frac{[f(x_i)-f(x_c)]^2}{n+1}\leq\epsilon \tag{9.10}$$

There are other modifications of this method where by λ is selected smaller or larger than 2, i.e. contraction or expansion of simplex. Main advantages of simplex method are (a) simple calculations and uncomplicated logic (b) few adjustable parameters(α,β,λ). It may be noted that the variables should be scaled to avoid difficulties due to different ranges of variables. The main disadvantage of this method is that the movement towards the optimum can be very slow.

9.2. Gradient / Steepest Descent / Cauchy's method.

DEFINITION 36. *Set of vectors* $\mathbf{z}\in R^{\mathbf{N}}$*such that* $\Phi(\mathbf{z})=C$ *where* C *is a constant, is called the level surface of* $\Phi(\mathbf{z})$ *for value* C.

By tracing out one by one level surfaces we obtain contour plot (see Figure 6 and Figure 7). Suppose $\mathbf{z}=\overline{\mathbf{z}}$ is a point lying on one of the level surfaces. If $\Phi(\mathbf{z})$ is continuous and differentiable then, using Taylor series expansion in a neighborhood of $\overline{\mathbf{z}}$ we can write

$$\Phi(\mathbf{z})=\Phi(\overline{\mathbf{z}})+\left[\nabla\Phi(\overline{\mathbf{z}})\right]^T(\mathbf{z}-\overline{\mathbf{z}})+\frac{1}{2}(\mathbf{z}-\overline{\mathbf{z}})^T\left[\nabla^2\Phi(\overline{\mathbf{z}})\right](\mathbf{z}-\overline{\mathbf{z}})+.... \tag{9.11}$$

If we neglect the second and higher order terms, we obtained

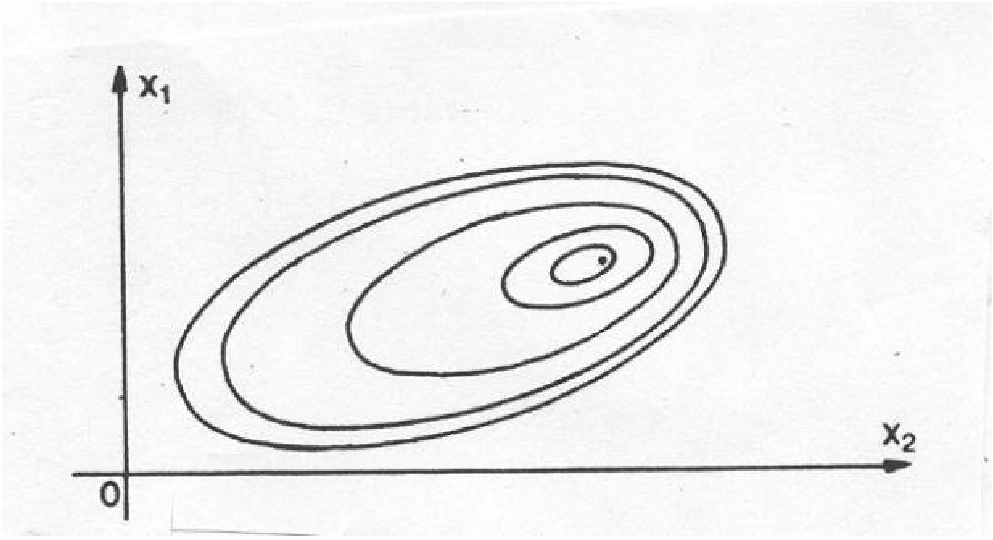


Figure 7. Level surfaces for n = 2

$$\Phi(\mathbf{z}) = \Phi(\overline{\mathbf{z}}+\Delta\mathbf{z}) \simeq \Phi(\overline{\mathbf{z}}) + \left[\nabla\Phi(\overline{\mathbf{z}})\right]^T \Delta\mathbf{z} = C \tag{9.12}$$

$$\Delta\mathbf{z} = (\mathbf{z}-\overline{\mathbf{z}}) \tag{9.13}$$

This is equation of the plane tangent to surface $\Phi(\mathbf{z})$ at point $\overline{\mathbf{z}}$. The equation of level surface through $\overline{\mathbf{z}}$ is

$$C = \Phi(\mathbf{z}) = \Phi(\overline{\mathbf{z}}) \tag{9.14}$$

Combining above two equations are obtain the equation of tangent surface at $\mathbf{z} =\overline{\mathbf{z}}$ as

$$(\mathbf{z}-\overline{\mathbf{z}})^T\nabla\Phi(\overline{\mathbf{z}}) = 0 \tag{9.15}$$

Thus, gradient at $\mathbf{z} = \overline{\mathbf{z}}$ is perpendicular to the level surface passing through $\Phi(\overline{\mathbf{z}})$ (See Figure 8).

We will now show that it points in the direction in which the function increases most rapidly, and in fact, $\nabla\Phi(\overline{\mathbf{z}})$ is the direction of maximum slope. If

$$\left[\nabla\Phi(\overline{\mathbf{z}})\right]^T \Delta\mathbf{z} < 0$$

then

$$\Phi(\overline{\mathbf{z}}+\Delta\mathbf{z}) < \Phi(\overline{\mathbf{z}}) \tag{9.16}$$

and $\Delta\mathbf{z}$ is called as descent direction. Suppose we fix our selves to unit sphere in the neighborhood of $\mathbf{z} =\overline{\mathbf{z}}$ i.e. set of all $\mathbf{z}$ such that $\|\Delta\mathbf{z}\| \leq 1$ and want to

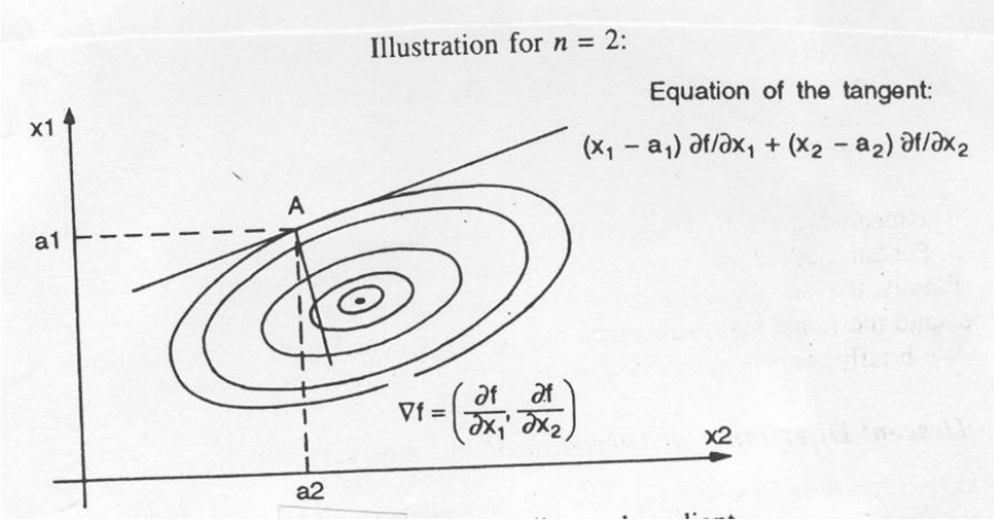


Illustration for n = 2:
Equation of the tangent:
$(x_1 - a_1)\,\partial f/\partial x_1 + (x_2 - a_2)\,\partial f/\partial x_2$
x1
a1
A
$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)$
x2
a2

FIGURE 8

find direction Δz such that $\Delta\Phi(z)^T\Delta z$ algebraically minimum. Using Cauchy-Schwartz inequality together with $\|\Delta\mathbf{z}\| \leq 1$, we have

$$\left|[\nabla\Phi(\overline{\mathbf{z}})]^T\,\Delta\mathbf{z}\right| \leq \|\nabla\Phi(\overline{\mathbf{z}})\|\,\|\Delta\mathbf{z}\| \leq \|\nabla\Phi(\overline{\mathbf{z}})\| \tag{9.17}$$

This implies

$$-\|\nabla\Phi(\overline{\mathbf{z}})\| \leq [\nabla\Phi(\overline{\mathbf{z}})]^T\,\Delta\mathbf{z} \leq \|\nabla\Phi(\overline{\mathbf{z}})\| \tag{9.18}$$

and minimum value $[\nabla\Phi(\overline{\mathbf{z}})]^T\,\Delta\mathbf{z}$ can attain when $\Delta\mathbf{z}$ is restricted within the unit ball equals $-\|\nabla\Phi(\overline{\mathbf{z}})\|$. In fact, the equality will hold if and only if $\Delta\mathbf{z}$ is colinear with $\nabla\Phi(\overline{\mathbf{z}})$, i.e.

$$\Delta\widehat{\mathbf{z}} = -\frac{\nabla\Phi(\overline{\mathbf{z}})}{\|\nabla\Phi(\overline{\mathbf{z}})\|} \tag{9.19}$$

Thus, unit direction $\Delta\widehat{\mathbf{z}}$ given by the above equation is the direction of steepest or maximum descent in which $\Phi(\overline{\mathbf{z}}+\Delta\mathbf{z}) - \Phi(\overline{\mathbf{z}})$ reduces at the maximum rate.

Algorithm

Given a point $\mathbf{z}^{(z)}$,the steepest descent method performs a line search in the direction of $-\dfrac{\nabla\Phi(\overline{\mathbf{z}})}{\|\nabla\Phi(\overline{\mathbf{z}})\|}$, i.e. the direction of steepest descent.

```
INITIALIZE: $\mathbf{z}^{(0)}, \varepsilon, k_{\max}, \lambda^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon) \ \ AND \ \ (k < k_{\max})]$
    $\mathbf{s}^{(k)} = \dfrac{\nabla\Phi(\mathbf{z}^{(k)})}{\|\nabla\Phi(\mathbf{z}^{(k)})\|}$
```

```
        $\lambda_k^* = \begin{matrix} \min \\ \lambda \end{matrix} \; \Phi\left(\mathbf{z}^{(k)} - \lambda \mathbf{s}^{(k)}\right)$
        $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - \lambda_k^* \mathbf{s}^{(k)}$
        $\delta = \left\| \nabla\Phi(z^{k+1}) \right\|_2$
    END WHILE
```

A numerical approach to the above one dimensional minimization problem is given in the Appendix. Alternate criteria which can be used for termination of iterations are as follows

$$\begin{aligned} \frac{\left|\Phi(\mathbf{z}^{(k+1)}) - \Phi(\mathbf{z}^{(k)})\right|}{\left|\Phi(z^{(k+1)})\right|} &\leq \varepsilon \\ \underset{i}{Maz} \left| \frac{\partial \Phi(\mathbf{z}^{(k+1)})}{\partial z_i} \right| &\leq \varepsilon \\ \left\| \mathbf{z}^{(k+1)} - \mathbf{z}^{(k)} \right\| &\leq \varepsilon \end{aligned}$$

The Method of steepest descent may appear to be the best unconstrained minimization method. However due to the fact that steepest descent is a local property, this method is not effective in many problems. If the objective function are distorted, then the method can be hopelessly slow.

9.3. Method of Conjugate Gradients. Convergence characteristics of steepest descent can be greatly improved by modifying it into a conjugate gradient method. Method of conjugate gradients display positive characteristics of Cauchy's and second order (i.e. Newton's) method with only first order information. This procedure sets up each new search direction as a linear combination of all previous search directions and newly determined gradient . The set of directions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}$........ are called as **A-conjugate** if they satisfy the condition.

$$[\mathbf{p}^{(k)}]^T \mathbf{A} \mathbf{p}^{(k-1)} = 0 \quad \text{for all } k \tag{9.20}$$

where $\mathbf{A}$ is a symmetric positive definite matrix. In conjugate directions method the successive directions are selected such that successive directions are conjugate with respect to the local Hessian .

Theorem 14. *If a quadratic function*

$$\Phi(\mathbf{z}) = 0.5\mathbf{z}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{z} + \mathbf{c} \tag{9.21}$$

is minimized sequentially, once along each direction of a set of n linearly independent A-conjugate directions,the global minimum of $\Phi(\mathbf{z})$ will be reached at or before n^{th} step regardless of the starting point (initial guess) $\mathbf{z}^{(0)}$.

Proof. Let $\mathbf{z} =\overline{\mathbf{z}}$ minimize $\Phi(\mathbf{z})$, then applying necessary condition for optimality, we have

$$\nabla\Phi(\overline{\mathbf{z}}) = \mathbf{b} + \mathbf{A}\overline{\mathbf{z}} = \overline{0} \tag{9.22}$$

Now , given a point $\mathbf{z}^{(0)}$ and n linearly independent directions $\mathbf{s}^{(0)}$,.....,$\mathbf{s}^{(n-1)}$, constants β_i can be found such that

$$\overline{\mathbf{z}} = \mathbf{z}^{(0)} + \sum_{i=0}^{n-1}\beta_i\mathbf{s}^{(i)} \tag{9.23}$$

If $\{\mathbf{s}^{(i)}\}$are A-conjugate and none of them is zero, then

$$\nabla\Phi(\overline{\mathbf{z}}) = \mathbf{b} + \mathbf{A}\mathbf{z}^{(0)} + \mathbf{A}\sum_{i=1}^{n}\beta_i\mathbf{s}^{(i)} = 0 \tag{9.24}$$

$$[\mathbf{s}^{(j)}]^T\nabla\Phi(\overline{\mathbf{z}}) = [\mathbf{s}^{(j)}]^T\left[\mathbf{b} + \mathbf{A}\mathbf{x}^{(1)}\right] + \beta_j\left[\mathbf{s}^{(j)}\right]^T\mathbf{A}\mathbf{s}^{(j)} = 0 \tag{9.25}$$

$$\beta_j = -\frac{[\mathbf{b} + \mathbf{A}\mathbf{x}^{(0)}]^T\mathbf{s}^{(j)}}{[\mathbf{s}^{(j)}]^T\mathbf{A}\mathbf{s}^{(j)}} \tag{9.26}$$

Now, consider an iterative procedure for minimization that starts at $\mathbf{z}^{(0)}$ and successively minimizes $\Phi(\mathbf{z})$ in directions $\mathbf{s}^{(0)}$ ....$\mathbf{s}^{(n-1)}$ where the directions are A-conjugate .The successive points are determined by relation

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \lambda_{\min}^{(k)}\mathbf{s}^{(k)} \tag{9.27}$$

Where$\lambda_{\min}^{(k)}$ is found by minimizing $f(\mathbf{z}^{(k)} + \lambda\mathbf{s}^{(k)})$ with respect to λ. At the optimum λ, we have

$$\frac{\partial\Phi}{\partial\lambda} = \left[\frac{\partial\Phi}{\partial\mathbf{z}}\right]^T_{\mathbf{z}=\mathbf{z}^{(k+1)}}\frac{\partial\mathbf{z}^{(k+1)}}{\partial\lambda} = 0 \tag{9.28}$$

$$\Rightarrow [\nabla\Phi(\mathbf{z}^{(k+1)})]^T[\mathbf{s}^{(k)}] = 0 \tag{9.29}$$

$$\nabla\Phi(\mathbf{z}^{(k+1)}) = \mathbf{b} + \mathbf{A}\{\mathbf{z}^{(k)} + \lambda\mathbf{s}^{(k)}\} \tag{9.30}$$

$$\Rightarrow \lambda_{\min}^{k} = -\frac{\left[\mathbf{b} + \mathbf{A}\mathbf{z}^{(k)}\right]^T\mathbf{s}^{(k)}}{[\mathbf{s}^{(k)}]^T\mathbf{A}\mathbf{s}^{(k)}} \tag{9.31}$$

Now

$$\mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} + \lambda_{\min}^{i}\mathbf{s}^{(i)} \qquad \text{for } i = 1...n$$

$$\Rightarrow \mathbf{z}^{(k)} = \mathbf{z}^{(0)} + \sum_{j=1}^{k-1}\lambda^{(j)}\mathbf{s}^{(j)} \tag{9.32}$$

$$[\mathbf{z}^{(k)}]^T\mathbf{A}\mathbf{s}^{(k)} = [\mathbf{z}^{(0)}]^T\mathbf{A}\mathbf{s}^{(k)} + \sum_{j=1}^{i-1}\lambda_{\min}^{k}\,\mathbf{s}^{(j)T}\mathbf{A}\mathbf{s}^{(k)} \tag{9.33}$$

$$= [\mathbf{z}^{(0)}]^T\mathbf{A}\mathbf{s}^{(k)} \tag{9.34}$$

$$\Rightarrow \lambda_{\min}^{k} = -\frac{\left[\mathbf{b}+\mathbf{A}\mathbf{z}^{(0)}\right]^T\mathbf{s}^{(k)}}{[\mathbf{s}^{(k)}]^T\mathbf{A}\mathbf{s}^{(k)}} \tag{9.35}$$

which is identical with β_j. Thus $\overline{\mathbf{z}}$ can be expressed as

$$\overline{\mathbf{z}} = \mathbf{z}^{(0)} + \sum_{i=1}^{n-1}\beta_j\mathbf{s}^{(j)} = \mathbf{z}^{(0)} + \sum_{i=1}^{n-1}\lambda_{\min}^{(i)}\mathbf{s}^{(i)} \tag{9.36}$$

This implies that $\overline{\mathbf{z}}$ can be reached in n steps or less. Since above equation holds good for any $\mathbf{z}^{(0)}$, the process converges independent of any choice of starting point $\mathbf{z}^{(0)}$ and any arbitrary set of conjugate directions. □

Now for development of algorithm we assume that the objective function to be quadratic.

$$\Phi(\mathbf{z}) = 0.5\mathbf{z}^{(T)}\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{z} + \mathbf{c} \tag{9.37}$$

Defining

$$\mathbf{g}(\mathbf{z}) = \nabla\Phi(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b} \tag{9.38}$$

we have

$$\mathbf{g}^{(k)} = \mathbf{g}(\mathbf{z}^{(k)}) = \mathbf{A}\mathbf{z}^{(k)} + \mathbf{b} \tag{9.39}$$

$$\mathbf{g}^{(k-1)} = \mathbf{g}(\mathbf{z}^{(k-1)}) = \mathbf{A}\mathbf{z}^{(k-1)} + \mathbf{b} \tag{9.40}$$

$$\Rightarrow \triangle\mathbf{g}(\mathbf{z}) = \mathbf{g}^{(k)} - \mathbf{g}^{(k-1)} = \triangle\mathbf{g}^{(k)} = A\triangle\mathbf{z}^{(k)}$$

Now, let us form an iterative scheme

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \lambda^{(k)}\mathbf{s}^{(k)} \tag{9.41}$$

Where we would like $\mathbf{s}^{(0)}$.......$\mathbf{s}^{(k)}$ to be A-conjugate. To generate such directions, consider

$$\mathbf{s}^{(k)} = -\bigtriangledown\Phi(\mathbf{z}^{(k)}) + \sum_{i=0}^{k-1}\alpha_i\mathbf{s}^{(i)} \tag{9.42}$$

$$= -\mathbf{g}^{(k)} + \sum_{i=0}^{k-1}\alpha_i\mathbf{s}^{(i)} \tag{9.43}$$

which is linear combination of all previous directions with $\mathbf{s}^{(0)} = -\mathbf{g}^{(0)}$. Note that at each iteration, a line search is performed to determine optimum step length λ^k such that

$$\frac{\delta\Phi(\mathbf{z}^{(k+1)})}{\delta\lambda} = \bigtriangledown\Phi[\mathbf{z}^{(k+1)}]^T\mathbf{s}^{(k)} \tag{9.44}$$

$$= \left[\mathbf{g}^{(k+1)}\right]^T\mathbf{s}^{(k)} = 0 \tag{9.45}$$

Now, we want to choose α_i such that $(i = 1, ...k-1)$ such that $\mathbf{s}^{(k)}$ is a conjugate to all previous directions. Consider first direction

$$\mathbf{s}^{(1)} = -\mathbf{g}^{(1)} + \alpha_0\mathbf{s}^{(0)} = -\mathbf{g}^{(1)} \pm \alpha_0\mathbf{g}^{(0)} \tag{9.46}$$

We force it to be A-conjugate with respect to $\mathbf{s}^{(0)}$, *i.e.*,

$$[\mathbf{s}^{(1)}]^T\mathbf{A}\mathbf{s}^{(0)} = 0 \tag{9.47}$$

$$[\mathbf{g}^{(1)} + \alpha_0\mathbf{g}^{(0)}]^T\mathbf{A}\mathbf{s}^{(0)} = 0 \tag{9.48}$$

At iteration $k = 1$, we have

$$\Delta x^{(1)} = \lambda_0\mathbf{s}^{(0)}$$

$$\mathbf{s}^{(0)} = \frac{\Delta x^{(1)}}{\lambda_0} \tag{9.49}$$

$$[\mathbf{g}^{(1)} + \alpha_0\mathbf{g}^{(0)}]^T\mathbf{A}\frac{\Delta\mathbf{z}^{(1)}}{\lambda_0} = 0 \tag{9.50}$$

But, we have

$$\mathbf{A}\Delta\mathbf{z}^{(1)} = \Delta I^{(1)} \tag{9.51}$$

$$[\mathbf{g}^{(1)} + \alpha_0\mathbf{g}^{(0)}]^T\Delta I^{(1)} = 0 \tag{9.52}$$

$$\alpha_0 = \frac{-(\Delta I^{(1)})^T\mathbf{g}^{(1)}}{(\Delta I^{(1)})^T\mathbf{g}^{(0)}} \tag{9.53}$$

$$\Rightarrow \ \alpha_0 = \frac{-(\Delta I^{(1)} - \mathbf{g}^{(0)})^T\mathbf{g}^{(1)}}{(\Delta I^{(1)} - \mathbf{g}^{(0)})^T\mathbf{g}^{(0)}} \tag{9.54}$$

$$\Rightarrow [\mathbf{g}^{(1)}]^T\mathbf{g}^{(1)} + \alpha_0(\mathbf{g}^{(0)})^T\mathbf{g}^{(1)} - \mathbf{g}^{(1)}\mathbf{g}^{(0)} - \ \alpha_0(\mathbf{g}^{(0)})^T\mathbf{g}^{(0)} = 0 \tag{9.55}$$

Now, we have

$$[\mathbf{g}^{(k+1)}]^T\mathbf{s}^{(k)} = 0 \tag{9.56}$$

$$[\mathbf{g}^{(1)}]^T\mathbf{g}^{(0)} = 0$$

$$\alpha_0 = \frac{||\mathbf{g}^{(1)}||^2}{||\mathbf{g}^{(0)}||^2} = \frac{||\nabla f(x^1)||^2}{||\nabla f(x^0)||^2} \tag{9.57}$$

Continuing the process, we form next direction as

$$\mathbf{s}^{(2)} = -\mathbf{g}^{(2)} + \alpha_0 \mathbf{s}^{(0)} + \alpha_1 \mathbf{s}^{(1)} \tag{9.58}$$

we want to choose α_0 and $\alpha 1$ such that

$$[\mathbf{s}^{(2)}]^T \mathbf{A}\mathbf{s}^{(0)} = 0 \; and \; [\mathbf{s}^{(2)}]^T \mathbf{A}\mathbf{s}^{(1)} = 0$$

$$[-\mathbf{g}^{(2)} + \alpha_0 \mathbf{s}^{(0)} + \alpha_1 \mathbf{s}^{(1)}]^T \mathbf{A}\mathbf{s}^{(0)} = 0 \tag{9.59}$$

$$-\mathbf{g}^{(2)} A\mathbf{s}^{(0)} + \alpha_1 [\mathbf{s}^{(2)}]^T \mathbf{A}\mathbf{s}^{(0)} = 0 \tag{9.60}$$

$$\alpha_1 = \frac{(\mathbf{g}^{(2)})^T \mathbf{A}\mathbf{s}^{(0)}}{[\mathbf{s}^{(0)}]^T \mathbf{A}\mathbf{s}^{(0)}} = \frac{(\mathbf{g}^{(2)})^T \mathbf{A}\Delta\mathbf{z}^{(1)}}{[\mathbf{s}^{(1)}]^T \mathbf{A}\mathbf{s}^{(1)}} \tag{9.61}$$

$$\frac{(\mathbf{g}^{(2)})^T [\mathbf{g}^{(1)} - \mathbf{g}^{(0)}]}{\mathbf{g}^{(0)} \mathbf{A}\mathbf{g}^{(0)}} = \frac{(\mathbf{g}^{(2)})^T \Delta I^{(1)}}{[\mathbf{s}^{(1)}]^T \mathbf{A}\mathbf{s}^{(1)}} \tag{9.62}$$

Continuing this process leads to the following general iteration scheme

$$\mathbf{s}^{(k)} = -\mathbf{g}^{(k)} + \left[\frac{||\mathbf{g}^{(k)}||^2}{||\mathbf{g}^{(k-1)}||^2}\right] \mathbf{s}^{(k+1)} \tag{9.63}$$

These are called Fletcher and Reev's iterations**.** For quadratic $\Phi(\mathbf{z})$, the minimum reached in maximum n steps. For non-quadratic $\Phi(\mathbf{z})$, additional line searches and steps may be required.

Algorithm

```
INITIALIZE: $\mathbf{z}^{(0)}, \varepsilon, k_{\max}, \lambda^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon) \;\; AND \;\; (k < k_{\max})]$
    $\alpha^{(k)} = \frac{\left\|\nabla\Phi\left(\mathbf{z}^{(k)}\right)\right\|^2}{\left\|\nabla\Phi\left(\mathbf{z}^{(k-1)}\right)\right\|^2}$
    $\mathbf{s}^{(k)} = \frac{\nabla\Phi(\mathbf{z}^{(k)})}{\left\|\nabla\Phi(\mathbf{z}^{(k)})\right\|}$
    $\mathbf{p}^{(k)} = \mathbf{s}^{(k)} - \alpha^{(k)}\mathbf{p}^{(k-1)}$
    $\lambda_k^* = \begin{matrix}\min \\ \lambda\end{matrix} \; \Phi\left(\mathbf{z}^{(k)} - \lambda\mathbf{p}^{(k)}\right)$
    $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - \lambda_k^* \, \mathbf{p}^{(k)}$
    $\delta = \left\|\nabla\Phi(\mathbf{z}^{(k+1)})\right\|_2$
END WHILE
```

Polak and Ribiere have developed conjugate gradient method by considering more general $\Phi(\mathbf{x})$ than quadratic. They have suggested use of α_k as

$$\alpha_k = \frac{[\Delta\mathbf{g}^{(k)}]^T\mathbf{g}^{(k)}}{||\mathbf{g}^{(k-1)}||_2^2} \tag{9.64}$$

9.4. Newton's Method. The necessary condition for optimization of a scalar function $\Phi(\mathbf{z})$ is

$$\nabla\Phi(\overline{\mathbf{z}}) = \overline{0} \tag{9.65}$$

if $\mathbf{z} = \overline{\mathbf{z}}$ is the optimum. Note that equation (9.65) defines a system of m equations in m unknowns. If $\nabla\Phi(\mathbf{z})$ is continuously differentiable in the neighborhood of $\mathbf{z} = \overline{\mathbf{z}}$, then, using Taylor series expansion, we can express the optimality condition (9.65) as

$$\nabla\Phi(\overline{\mathbf{z}}) = \nabla\Phi\left[\mathbf{z}^{(k)} + (\overline{\mathbf{z}} - \mathbf{z}^{k})\right] \simeq \nabla\Phi[\mathbf{z}^{(k)}] + \left[\nabla^2\Phi(\mathbf{z}^{(k)})\right]\Delta\mathbf{z}^{(k)} = \overline{0} \tag{9.66}$$

Defining Hessian matrix $H^{(k)}$ as

$$H^{(k)} = \left[\nabla^2\Phi(\mathbf{z}^{(k)})\right]$$

an iteration scheme an iteration scheme can be developed by solving equation (9.66)

$$\begin{aligned}\mathbf{z}^{(k+1)} &= \mathbf{z}^k + \lambda\Delta\mathbf{z}^k; \\ \Delta\mathbf{z}^{(k)} &= -\left[H^{(k)}\right]^{-1}\nabla\Phi[\mathbf{z}^{(k)}]\end{aligned}$$

In order that $\nabla\mathbf{z}^{(k)}$ is a descent direction it should satisfy the condition

$$\left[\nabla\Phi[\mathbf{z}^{(k)}]\right]^T\Delta\mathbf{z}^{(k)} < 0 \tag{9.67}$$

or

$$\left[\nabla\Phi[\mathbf{z}^{(k)}]\right]^T\left[H^{(k)}\right]^{-1}\nabla\Phi[\mathbf{z}^{(k)}] > 0 \tag{9.68}$$

i.e. in order that $\Delta\mathbf{z}^{(k)}$ is a descent direction, Hessian $H^{(k)}$ should be a positive definite matrix. This method has good convergence but demands large amount of computations i.e. solving a system of linear equations and evaluation of Hessian at each step

Algorithm

```
INITIALIZE: $\mathbf{z}^{(0)}, \varepsilon, k_{\max}, \lambda^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon)$ $AND$ $(k < k_{\max})]$
    $Solve\ \ H^{(k)}\mathbf{s}^{(k)} = -\nabla\Phi^{(k)}$
```

```
        $\lambda_k^* = \underset{\lambda}{\min}\ \Phi\left(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)}\right)$
        $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - \lambda_k^* \mathbf{s}^{(k)}$
        $\delta = \left\|\nabla\Phi\left[\mathbf{z}^{(k+1)}\right]\right\|_2$
    END WHILE
```

9.4.1. *Quasi- Newton Method.* Major disadvantage of Newtons method or its variants is the need to compute the Hessian of each iteration . The quasi-Newton methods over come this difficulty by constructing an approximate Hessian from the gradient information available at the successive iteration. Quasi-Newton methods thus mimic positive characteristics of Newton's method using only gradient information. All methods of this class generate search direction as

$$\mathbf{s}^{(k)} = -D^{(k)}\nabla f^{(k)} \tag{9.69}$$

which is similar to the Newton's update, i.e.

$$\mathbf{s}^{(k)} = -[H^{(k)}]^{-1}\nabla f^{(k)}$$

Here, $D^{(k)}$ is a $n \times n$ matrix (called metric), which changes with iterations (variable metric methods). A variable metric method is quasi-Newton method if it is designed so that the iterates satisfy the following quadratic function property

$$\Delta\mathbf{z} = A^{-1}\Delta\mathbf{g} \tag{9.70}$$

Let us assume a recursion for the estimate to inverse of Hessian

$$D^{(k+1)} = D^{(k)} + D_c^{(k)} \tag{9.71}$$

Basic idea is to form $D_c^{(k)}$ such that $D^{(0)}, D^{(1)}..........D^{(k+1)} \rightarrow [H(\overline{\mathbf{z}})]^{-1} = [\nabla^2 f(\overline{\mathbf{z}})]^{-1}$. We know that for a quadratic $f(\mathbf{z})$ of the form

$$f(\mathbf{z}) = 0.5\mathbf{z}^T A\mathbf{z} + \mathbf{bz} + \mathbf{c} \tag{9.72}$$

we can show

$$\Delta\mathbf{z} = A^{-1}\Delta\mathbf{g} \tag{9.73}$$

Let us assume that our approximation for A^{-1} is of the form

$$A^{-1} = \beta D^{(k)} \tag{9.74}$$

where β is a scalar. We would like $D^{(k)}$ to satisfy

$$\begin{aligned} \Delta\mathbf{z}^{(k)} &= \mathbf{z}^{(k+1)} - \mathbf{z}^{(k)} = \beta D^{(k+1)}\Delta\mathbf{g}^{(k)} \\ &= \beta D^{(k+1)}[\mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}] \end{aligned} \tag{9.75}$$

$$= D^{(k+1)}\Delta\mathbf{g}^{(k)} = \frac{\Delta\mathbf{z}^{(k)}}{\beta} \tag{9.76}$$

Now

$$D^{(k+1)}\Delta\mathbf{g}^{(k)} = D^{(k)}\Delta\mathbf{g}^{(k)} + D_c^{(k)}\Delta\mathbf{g}^{(k)} \tag{9.77}$$

$$\frac{\Delta\mathbf{z}^{(k)}}{\beta} = [D^{(k)} + D_c^{(k)}]\Delta\mathbf{g}^{(k)} \tag{9.78}$$

$$A_c^{(k)}\Delta\mathbf{g}^{(k)} = \frac{\Delta\mathbf{z}^{(k)}}{\beta} - D^{(k)}\Delta\mathbf{g}^{(k)} \tag{9.79}$$

One can verify by direct substitution that

$$A_c^{(k)} = \frac{1}{\beta}\left[\frac{\Delta\mathbf{z}^{(k)}y^T}{y^T\Delta I^{(k)}} - \frac{A^{(k)}\Delta I^{(k)}\boldsymbol{\eta}^T}{\boldsymbol{\eta}^T\Delta I^{(k)}}\right] \tag{9.80}$$

is a solution. As $\mathbf{y}$ and $\boldsymbol{\eta}$ are arbitrary vectors, this is really a family of solutions .If we let

$$\mathbf{y} = \Delta\mathbf{z}^{(k)} \text{ and } \boldsymbol{\eta} = A^{(k)}\Delta I^{(k)} \tag{9.81}$$

we get the Davidon-Fletcher-Powel update

$$A^{(k)} = A^{(k-1)} + \frac{\Delta\mathbf{z}^{(k-1)}[\Delta\mathbf{z}^{(k-1)}]^T}{[\Delta\mathbf{z}^{(k-1)}]^T\Delta I^{(k-1)}} - \frac{A^{(k-1)}\Delta I^{(k-1)}[\Delta I^{(k-1)}A^{(k-1)}]^T}{[\Delta I^{(k-1)}]^T A^{(k-1)}\Delta I^{(k-1)}} \tag{9.82}$$

Thus, matrix $A^{(k)}$ is iteratively computed

$$A^{(k+1)} = A^{(k)} + M^{(k)} - N^{(k)} \tag{9.83}$$

$$\mathbf{q}^{(k)} = \nabla\Phi^{k+1} - \nabla\Phi^{(k)} \tag{9.84}$$

$$M^{(k)} = \left(\frac{\lambda_k^*}{\left[\Delta\mathbf{z}^{(k)}\right]^T\mathbf{q}^{(k)}}\right)\left[\left[\Delta\mathbf{z}^{(k)}\right]\left[\Delta\mathbf{z}^{(k)}\right]^T\right] \tag{9.85}$$

$$N^k = \left(\frac{1}{\left[\mathbf{q}^{(k)}\right]^T L^{(k)}\mathbf{q}^{(k)}}\right)\left[L^{(k)}\mathbf{q}^{(k)}\right]\left[L^{(k)}\mathbf{q}^{(k)}\right]^T \tag{9.86}$$

starting from some initial guess, usually a positive definite matrix. Typical choice is

$$L^{(0)} = \mathbf{I} \tag{9.87}$$

Properties

- It can be shown that if $A^{(0)}$ is symmetric and positive definite definite, then $A^{(1)}, A^{(2)}$...............$A^{(k)}$ will all be symmetric and positive definite.(Typically $A^0 = I$).

- If f(x)is quadratic, it can be shown that $D^{(k)}$ converges to A^{-1} at optimum, i.e., $D^{(n+1)} = A^{-1}$
- The directions generated are A-conjugate

There are several variations of this method depending upon choice of $\beta, y, \ and\ z.$
Broyder-Fletcher-Shanno update(BSF).

$$A^{(k)} = \left[I - \frac{\Delta x^{(k-1)}[\Delta I^{(k-1)}]^T}{[\Delta x^{(k-1)}]^T \Delta I^{(k-1)}}\right] A^{(k-1)} \left[I - \frac{\Delta x^{(k-1)}[\Delta I^{(k-1)}]^T}{[\Delta x^{(k-1)}]^T \Delta I^{(k-1)}}\right] + \frac{\Delta x^{(k-1)}[\Delta x^{(k-1)}]^T}{[\Delta x^{(k-1)}]^T \Delta x^{(k)}} \tag{9.88}$$

9.5. Leverberg-Marquardt Method. It is known from the experience that the steepest descent method produces large reduction in objective function when $\mathbf{z}^{(0)}$ is far away from $\overline{\mathbf{z}}$,The optimal solution. However, steepest descent method becomes notoriously slow near optimum, on the other hand Newtons method generates ideal search directions near the solution. The Leverberg- Marquardt approach combines as follows.

$$\begin{aligned}\left[H^{(k)} + \lambda_k I\right] \Delta\mathbf{z}^{(k)} &= -\nabla\Phi[\mathbf{z}^{(k)}] \\ \mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} + \Delta\mathbf{z}^{(k)}\end{aligned}$$

Here λ_k is used to set the search direction and step length. To begin the search, a large value of $\lambda_o(\cong 10^4)$is selected so that

$$\left[H^{(0)} + \lambda^{(0)} I\right] \cong [\lambda_o I]$$

Thus, for sufficiently large $\lambda^{(0)}, \Delta\mathbf{z}^{(0)}$is in the negative of the gradient direction i.e. $-\nabla\Phi^{(k)}$.As $\lambda_k \to 0, \Delta\mathbf{z}^{(k)}$ goes from steepest descent to Newtons method.

- **Advantages:**Simplicity, excellent convergence near $\overline{z}$
- **Disadvantages:** Need to compute Hessian matrix, $H^{(k)}$and set of linear equations has to be solved at each iteration

Algorithm

```
INITIALIZE: $\mathbf{z}^{(0)}, \varepsilon, k_{\max}, \lambda^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon)\ \ AND\ \ (k < k_{\max})]$
    STEP 1 : Compute $H^{(k)}$ and $\nabla\Phi[\mathbf{z}^{(k)}]$
    STEP 2 : $Solve\ \ \left[H^{(k)} + \lambda_k I\right] \mathbf{s}^{(k)} = -\nabla\Phi[\mathbf{z}^{(k)}]$
    IF $\left(\Phi[\mathbf{z}^{(k+1)}] < \Phi[\mathbf{z}^{(k)}]\right)$
        $\lambda^{(k+1)} = \frac{1}{2}\lambda^{(k)}$
```

```
            $\delta = \left\|\nabla\Phi[\mathbf{z}^{(k+1)}]\right\|$
            $k = k + 1$
        ELSE
            $\lambda^{(k)} = 2\lambda^{(k)}$
        GO TO STEP 2
    END WHILE
```

9.6. Line Search: One Dimensional Optimization. In any multi-dimensional optimization problem, we have to solve a one dimensional minimization problem

$$\underset{\lambda}{\min}\ \Phi(\lambda) = \underset{\lambda}{\min}\ \Phi\left(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)}\right) \tag{9.89}$$

where $\mathbf{s}^{(k)}$ is the search direction. There are several numerical approaches available for solving this one dimensional optimization problem. In this sub-section, we discuss the cubic interpolation method for performing the line search.

The first step in the line search is to find bounds on the optimal step size λ^*. These are established by finding two points, say α and β, such that the slope $d\Phi/d\lambda$

$$d\Phi/d\lambda = \left[\frac{\partial\Phi(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)})}{\partial(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)})}\right]^T \frac{\partial(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)})}{\partial\lambda} \tag{9.90}$$

$$= -\left(\nabla\Phi(\mathbf{z}^{(k)} - \lambda\mathbf{s}^{(k)})\right)^T \mathbf{s}^{(k)} \tag{9.91}$$

has opposite signs at these points. We know that at $\lambda = 0$,

$$d\Phi(0)/d\lambda = -\left(\nabla\Phi(\mathbf{z}^{(k)})\right)^T \mathbf{s}^{(k)} < 0 \tag{9.92}$$

as $s^{(k)}$ is assumed to be a descent direction. Thus, we take α corresponding to $\lambda = 0$ and try to find a point $\lambda = \beta$ such that $d\Phi/d\lambda > 0$. The point β can be taken as the first value out of $\lambda = h,\ 2h, 4h, 8h, ...$ for which $d\Phi/d\lambda > 0$,where h is some pre-assigned initial step size. As $d\Phi/d\lambda$ changes sign in the interval $[0, \beta]$, the optimum λ^* is bounded in the interval $[0, \beta]$.

The next step is to approximate $\Phi(\lambda)$ over interval $[0, \beta]$ by a cubic interpolating polynomial for the form

$$\Phi(\lambda) = a + b\lambda + c\lambda^2 + d\lambda^3 \tag{9.93}$$

The parameters a and b be computed as

$$\begin{aligned} \Phi(0) &= a = \Phi(\mathbf{z}^{(k)}) \\ d\Phi(0)/d\lambda &= b = -\left(\nabla\Phi(\mathbf{z}^{(k)})\right)^T \mathbf{s}^{(k)} \end{aligned}$$

The parameters c and d can be computed by solving

$$\begin{aligned} \Phi(\beta) &= a + b\beta + c\beta^2 + d\beta^3 \\ d\Phi(\beta)/d\lambda &= b + 2c\beta + 3d\beta^2 \end{aligned}$$

i.e. by solving

$$\begin{bmatrix} \beta^2 & \beta^3 \\ 2\beta & 3\beta^2 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} \Phi\left(\mathbf{z}^{(k)} - \beta\mathbf{s}^{(k)}\right) - a - \beta b \\ -\left(\mathbf{s}^{(k)}\right)^T \nabla\Phi\left(\mathbf{z}^{(k)} - \beta\mathbf{s}^{(k)}\right) - b \end{bmatrix}$$

The application of necessary condition for optimality yields

$$d\Phi/d\lambda = b + 2c\lambda + 3d\lambda^2 = 0 \tag{9.94}$$

i.e.

$$\lambda^* = \frac{-c \pm \sqrt{(c^2 - 3bd)}}{3d} \tag{9.95}$$

One of the two values correspond to the minimum. The sufficiency condition for minimum requires

$$d^2\Phi/d\lambda^2 = 2c + 6d\lambda^* > 0 \tag{9.96}$$

The fact that $d\Phi/d\lambda$ has opposite sign at $\lambda = 0$ and $\lambda = \beta$ ensures that the equation 9.94 does not have imaginary roots.

Algorithm

```
INITIALIZE: $\mathbf{z}^{(k)}, \mathbf{s}^{(k)}, h$
Step 1: Find $\beta$
    $\beta = h$
    WHILE $[d\Phi(\beta)/d\lambda < 0]$
        $\beta = 2\beta$
    END WHILE
Step 2: Solve for $a, b, c$ and $d$ using $\mathbf{z}^{(k)}, \mathbf{s}^{(k)}$ and $\beta$
Step 3: Find $\lambda^*$ using sufficient condition for optimality
```

10. Numerical Methods Based on Optimization Formulation

10.1. Simultaneous Solutions of Linear Algebraic Equations. Consider system of linear algebraic equations

$$A\mathbf{x} = \mathbf{b} \;\; ; \;\; \mathbf{x}, \mathbf{b} \in \mathbf{R}^n \tag{10.1}$$

where A is a non-singular matrix. Defining objective function

$$\Phi(\mathbf{x}) = \frac{1}{2}(A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) \tag{10.2}$$

the necessary condition for optimality requires that

$$A^T(A\mathbf{x}-\mathbf{b})=\overline{0} \tag{10.3}$$

Since A is assumed to be nonsingular, the stationarity condition is satisfied only at the solution of $A\mathbf{x}=\mathbf{b}$. The stationary point is also a minimum as A^TA is a positive definite matrix. Thus, we can compute the solution of $A\mathbf{x}=\mathbf{b}$ by minimizing

$$\Phi(\mathbf{x})=\frac{1}{2}\mathbf{x}^T\left(A^TA\right)\mathbf{x}-\left(A^T\mathbf{b}\right)^T\mathbf{x} \tag{10.4}$$

If conjugate gradient method is used for solving the optimization problem, it can be theoretically shown that the minimum can be reached in n steps. In practice, however, we require more than n steps to achieve $\Phi(\mathbf{x})<\varepsilon$ due to the rounding off errors in computation of conjugate directions. Nevertheless, when n is large, this approach can generate a reasonably accurate solution with considerably less computations.

10.2. Simultaneous Solutions of Nonlinear Algebraic Equations. Consider problem of solving n nonlinear algebraic equations in n variables, which have to be solved simultaneously. These can be expressed in the following abstract form

$$\begin{aligned}
f_1(x_1,x_2,x_{3,}......x_n) &= 0 && (10.5)\\
f_2(x_1,x_2,x_{3,}......x_n) &= 0 && (10.6)\\
.... &= 0 &&\\
f_n(x_1,x_2,x_{3,}......x_n) &= 0 && (10.7)
\end{aligned}$$

Or

$$\begin{aligned}
F(\mathbf{x}) &= \overline{0} \quad ; \quad \mathbf{x}\in R^n && (10.8)\\
\mathbf{x} &= \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}^T
\end{aligned}$$

where $\overline{0}$ represents $n\times 1$ zero vector. Here $F(\mathbf{x})\in R^n$ represents n dimensional function vector defined as

$$F(\mathbf{x})=\begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & ... & f_n(\mathbf{x}) \end{bmatrix}^T \tag{10.9}$$

Defining a scalar function

$$\Phi(\mathbf{x})=[F(\mathbf{x})]^T F(\mathbf{x})=[f_1(\mathbf{x})]^2+[f_2(\mathbf{x})]^2+....+[f_n(\mathbf{x})]^2 \tag{10.10}$$

finding solution to above set of equations can be formulated as minimization of

$$\underset{\mathbf{x}}{\min}\ \Phi\left(\mathbf{x}\right) \tag{10.11}$$

The necessary condition for unconstrained optimality is

$$\frac{\partial\Phi\left(\mathbf{x}\right)}{\partial\mathbf{x}} = \left[\frac{\partial F(\mathbf{x})}{\partial\mathbf{x}}\right]^T F(\mathbf{x}) = \overline{0} \tag{10.12}$$

If we ignore the degenerate case where matrix $\left[\dfrac{\partial F(\mathbf{x})}{\partial\mathbf{x}}\right]$ is singular and vector $F(\mathbf{x})$ belongs to null space of this matrix, the necessary condition for optimality is satisfied when

$$F(\mathbf{x}) = \overline{0} \tag{10.13}$$

which also corresponds to the global minimum of $\Phi\left(\mathbf{x}\right)$. The Leverberg-Marquardt method is known to work well for solving optimization formulation of this type.

10.3. Finite Element Method for Solving ODE-BVP and PDEs [15, 16]. The finite element method is a powerful tool for solving PDEs particularly when the system under consideration has complex geometry. This method is based on the optimization formulation. In this section, we provide a very brief introduction to the method of finite element.

10.3.1. *Raleigh-Ritz method.* Consider linear system of equations

$$A\mathbf{x} = \mathbf{b} \tag{10.14}$$

where A is a $n \times n$ positive definite and symmetric matrix and it is desired to solve for vector $\mathbf{x}$. We can pose this as a minimization problem by defining objective function

$$\Phi(\mathbf{x}) = (1/2)\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} \tag{10.15}$$

$$= (1/2)\left\langle \mathbf{x},\, A\mathbf{x}\right\rangle - \left\langle \mathbf{x}, \mathbf{b}\right\rangle \tag{10.16}$$

If $\Phi(\mathbf{x})$ minimum at $\mathbf{x} = \mathbf{x}^*$, the necessary condition for optimality requires

$$\partial\Phi/\partial\mathbf{x} = A\mathbf{x}^* - \mathbf{b} = \overline{0} \tag{10.17}$$

which is precisely the equation we want to solve. Since the Hessian matrix

$$\partial^2\Phi/\partial\mathbf{x}^2 = A$$

is positive definite, the solution of $\mathbf{x} = \mathbf{x}^*$ of $A\mathbf{x} = \mathbf{b}$ is the global minimum of objective function $\Phi(\mathbf{x})$.

In the above demonstration, we were working in space R^n. Now, let us see if a similar formulation can be worked out in another space, namely $C^{(2)}[0,1]$, i.e. the set of twice differentiable continuous functions on $[0,1]$. Consider ODE-BVP

$$Lu = -d^2u/dz^2 = f(z); \tag{10.18}$$

$$B1 : u(0) = 0 \tag{10.19}$$

$$B2 : u(1) = 0 \tag{10.20}$$

Similar to the linear *operator* (matrix) A,which operates on vector $\mathbf{x} \in R^n$ to produce another vector $\mathbf{b} \in R^n$, the linear operator $L = [-d^2/dz^2\,]$operates on vector $u(z) \in C^{(2)}[0,1]$ to produce $f(z) \in C^{(2)}[0,1]$. Note that the matrix A in our example is symmetric and positive definite, i.e.

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x} \rangle &> 0 \text{ for all } \mathbf{x} \neq \overline{0} \\ \text{and } A^T &= A \end{aligned}$$

In order to see how the concept of symmetric matrix can be generalized to operators on infinite dimensional spaces, let us first define adjoint of a matrix.

Definition 37. ***(Adjoint of Matrix):*** *A matrix A^* is said to be adjoint of matrix A if it satisfies $\langle \mathbf{x}, A\mathbf{y} \rangle = \langle A^*\mathbf{x}, \mathbf{y} \rangle$. Further, the matrix A is called self adjoint if $A^* = A$.*

When matrix A has all real elements, we have

$$\mathbf{x}^T(A\mathbf{y}) = (A^T\mathbf{x})^T\mathbf{y}$$

and it is easy to see that $A^* = A^T$, i.e.

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \left\langle A^T\mathbf{x}, \mathbf{y} \right\rangle \tag{10.21}$$

The matrix A is called *self-adjoint* if $A^T = A$. Does operator L defined by equations (10.18-10.20) have some similar properties of *symmetry* and *positiveness*? Analogous to the concept of adjoint of a matrix, we first introduce the concept of adjoint of an operator L on any inner product space.

Definition 38. ***(Adjoint of Operator)*** *An operator L^* is said to be adjoint of operator L if it satisfies*

$$\langle v, Lu \rangle = \langle L^*v, u \rangle$$

Further, the operator L is said to be self-adjoint, if $L^* = L$, $B1^* = B1$ and $B2^* = B2$.

To begin with, let us check whether the operator L defined by equations (10.18-10.20) is self-adjoint.

$$\begin{aligned}(10.22)\quad \langle v, Lu\rangle &= \int_0^1 v(z)(-d^2u/dz^2)dz \\ &= \left[-v(z)\frac{du}{dz}\right]_0^1 + \int_0^1 \frac{dv}{dz}\frac{du}{dz}dz \\ &= \left[-v(z)\frac{du}{dz}\right]_0^1 + \left[\frac{dv}{dz}u(z)\right]_0^1 + \int_0^1 \left(-\frac{d^2v}{dz^2}\right)u(z)dz\end{aligned}$$

Using the boundary conditions $u(0) = u(1) = 0$, we have

$$(10.23)\qquad \left[\frac{dv}{dz}u(z)\right]_0^1 = \frac{dv}{dz}u(1) - \frac{dv}{dz}u(0) = 0$$

If we set

$$(10.24)\qquad B1^* \ : \ v(0) = 0$$

$$(10.25)\qquad B2^* \ : \ v(1) = 0$$

then

$$(10.26)\qquad \left[\frac{du}{dz}v(z)\right]_0^1 = 0$$

and we have

$$(10.27)\qquad \langle v, Lu\rangle = \int_0^1 \left(-\frac{d^2v}{dz^2}\right)u(z)dz = \langle L^*v, u\rangle$$

In fact, it is easy to see that the operator L is self adjoint as $L^* = L$, $B1^* = B1$ and $B2^* = B2$. In addition to the self-adjointness of L, we have

$$(10.28)\qquad \langle u, Lu\rangle = \left[-u(z)\frac{du}{dz}\right]_0^1 + \int_0^1 \left(\frac{du}{dz}\right)^2 dz$$

$$(10.29)\qquad = \int_0^1 \left(\frac{du}{dz}\right)^2 dz > 0 \text{ for all } u(z)$$

when $u(z)$ is a non-zero vector in $C^{(2)}[0,1]$. In other words, solving the ODE-BVP is analogous to solving $A\mathbf{x} = \mathbf{b}$ by optimization formulation where A is symmetric and positive definite matrix, i.e.

$$A \leftrightarrow \left[-d^2/dz^2\right] \ ; \ \ \mathbf{x} \leftrightarrow u(z); \ \mathbf{b} \leftrightarrow f(z)$$

Let $u(z) = u^{*}(z)$ represent the true solution of the ODE-BVP. Now, taking motivation from the optimization formulation for solving $A\mathbf{x} = \mathbf{b}$, we can formulate a minimization problem to compute the solution

$$\Phi\left[u(z)\right] = (1/2)\left\langle u(z), -d^2u/dz^2\right\rangle - \langle u(z), f(z)\rangle \tag{10.30}$$

$$= 1/2\int_0^1 u(z)(-d^2u/dz^2)dz - \int_0^1 u(z)f(z)dz$$

$$u^{*}(z) = \begin{matrix} Min \\ u(z) \end{matrix} \; \Phi[u(z)]$$

$$= \begin{matrix} Min \\ u(z) \end{matrix} \; (1/2)\left\langle u(z), Lu\right\rangle - \langle u(z), f(z)\rangle \tag{10.31}$$

$$u(z) \in C^{(2)}[0,1] \tag{10.32}$$

$$\text{subject to } \; u(0) = u(1) = 0$$

Thus, solving the $ODE - BVP$ has been converted solving a minimization problem. Integrating the first term in equation (**??**) by parts, we have

$$\int_0^1 u(z)(-d^2u/dz^2)dz = \int_0^1 (du/dz)^2 dz - \left[u(du/dz)\right]_0^1 \tag{10.33}$$

Now, using boundary conditions, we have

$$\left[u(du/dz)\right]_0^1 = \left[u(0)(du/dz)_{z=0} - u(1)(du/dz)_{z=1}\right] = 0 \tag{10.34}$$

This reduces $\Phi(u)$ to

$$\Phi(u) = \left[1/2\int_0^1 (du/dz)^2 dz\right] - \left[\int_0^1 uf(z)dz\right] \tag{10.35}$$

The above equation is similar to an *energy function,* where the first term is analogous to kinetic energy and the second term is analogous to potential energy.

As

$$\int_0^1 (du/dz)^2 dz$$

is *positive and symmetric*, we are guaranteed to find the minimum. The main difficulty in performing the search is that, unlike the previous case where we were working in R^n, the search space is infinite dimensional as $u(z) \in C^{(2)}[0,1]$. One remedy to alleviate this difficulty is to reduce the infinite dimensional search problem to a finite dimensional search space by constructing an approximate solution using n trial functions. Let $v^{(1)}(z), ..., v^{(n)}(z)$ represent trial function. The approximate solution is constructed as

$$\widehat{u}(z) = c_0 v^{(0)}(z) + ... + c_n v^{(n)}(z) \tag{10.36}$$

where $v^{(i)}(z)$ represents *trial functions.* Using this approximation, we convert the infinite dimensional optimization problem to a finite dimensional optimization problem as follows

$$\underset{\mathbf{C}}{Min} \widehat{\Phi}(\mathbf{c}) = \left[1/2 \int_0^1 (d\widehat{u}/dz)^2 dz \right] - \left[\int_0^1 \widehat{u} f(z) dz \right] \tag{10.37}$$

$$= 1/2 \int_0^1 \left[c_0 \left(dv^{(0)}(z)/dz \right) + ... + c_n \left(dv^{(n)}(z)/dz \right) \right]^2 dz \tag{10.38}$$

$$- \int_0^1 f(z)[c_0 v^{(0)}(z) + ... + c_n v^{(n)}(z)] dz \tag{10.39}$$

The trial functions $v^{(i)}(z)$ are chosen in advance and coefficients $c_1, ...c_m$ are treated as unknown. Then, the above optimization problem can be recast as

$$\underset{\mathbf{c}}{Min} \quad \widehat{\Phi}(\mathbf{c}) = \underset{\mathbf{c}}{Min} \quad \left[1/2 \; \mathbf{c}^T A \mathbf{c} - \mathbf{c}^T b \right] \tag{10.40}$$

$$\mathbf{c} = \begin{bmatrix} c_0 & c_2 & ... & c_n \end{bmatrix}^T \tag{10.41}$$

$$A = \begin{bmatrix} \left\langle \dfrac{dv^{(0)}}{dz}, \dfrac{dv^{(0)}}{dz} \right\rangle & .... & \left\langle \dfrac{dv^{(0)}}{dz}, \dfrac{dv^{(n)}}{dz} \right\rangle \\ .... & .... & .... \\ \left\langle \dfrac{dv^{(n)}}{dz}, \dfrac{dv^{(0)}}{dz} \right\rangle & .... & \left\langle \dfrac{dv^{(n)}}{dz}, \dfrac{dv^{(n)}}{dz} \right\rangle \end{bmatrix} \tag{10.42}$$

$$\mathbf{b} = \begin{bmatrix} \langle v^{(1)}(z), f(z) \rangle \\ .... \\ \langle v^{(n)}(z), f(z) \rangle \end{bmatrix} \tag{10.43}$$

It is easy to see that matrix A is positive definite and symmetric and the global minimum of the above optimization problem can be found by using necessary condition for optimality as follows

$$\partial\widehat{\Phi}/\partial\mathbf{c} = A\mathbf{c}^* - \mathbf{b} = \overline{0} \tag{10.44}$$

or

$$\mathbf{c}^* = A^{-1}\mathbf{b} \tag{10.45}$$

Note the similarity of the above equation with the normal equation arising from the projection theorem. Thus, steps in the Raleigh-Ritz method can be summarized as follows

1. Choose an approximate solution.
2. Compute matrix A and vector $\mathbf{b}$
3. Solve for $A\mathbf{c} = \mathbf{b}$

Similar to finite difference method, we begin by choosing $(n-1)$ equidistant internal node (grid) points as follows

$$z_i = i\Delta z \quad (i = 0, 1, 2, ...n)$$

and defining $(n-1)$ finite elements

$$z_{i-1} \leq z \leq z_i \quad \text{for} \;\; i = 1, 2, ...n-1 \tag{10.46}$$

Then, by invoking Weierstarss theorem, we formulate the approximate solution using piecewise constant polynomials on each finite element.. The simplest possible choice is a line

$$\widehat{u}_i(z) \;\; = \;\; a_i + b_i z \tag{10.47}$$

$$z_{i-1} \;\; \leq \;\; z \leq z_i \quad \text{for} \;\; i = 0, 2, ...n-1 \tag{10.48}$$

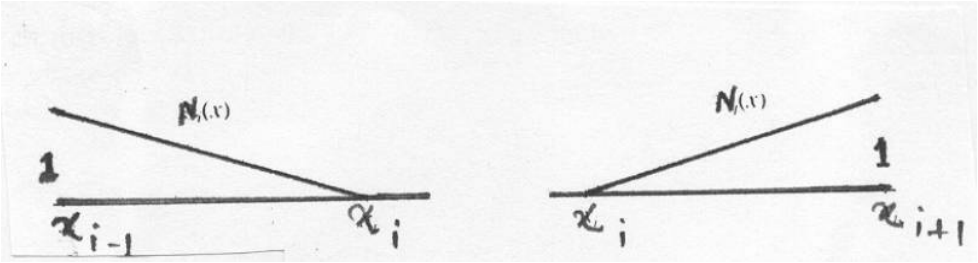


FIGURE 9

At he boundary points of the element, we have

$$\widehat{u}_i(z_{i-1}) = \widehat{u}_{i-1} = a_i + b_i z_{i-1} \tag{10.49}$$

$$\widehat{u}_i(z_i) = \widehat{u}_i = a_i + b_i z_i \tag{10.50}$$

This implies

$$a_i = \frac{\widehat{u}_{i-1} z_i - \widehat{u}_i z_{i-1}}{\Delta z} \quad ; \quad b_i = \frac{\widehat{u}_i - \widehat{u}_{i-1}}{\Delta z} \tag{10.51}$$

Thus, the polynomial on the i'th segment can be written as

$$\begin{aligned} \widehat{u}_i(z) &= \frac{\widehat{u}_{i-1} z_i - \widehat{u}_i z_{i-1}}{\Delta z} + \left(\frac{\widehat{u}_i - \widehat{u}_{i-1}}{\Delta z}\right) z \\ z_{i-1} &\leq z \leq z_i \quad \text{for } i = 1, 2, ...n-1 \end{aligned} \tag{10.52}$$

and now we can work in terms of unknown values $\{\widehat{u}_{0,}\widehat{u}_1, ...\widehat{u}_n\}$ instead of parameters a_i and $b_{i.}$ A more elegant and useful form of equation (10.52) can be found by defining shape functions (see Figure 9)

$$M_i(z) = \frac{z - z_i}{\Delta z} \quad ; \quad N_i(z) = \frac{z - z_{i-1}}{\Delta z} \tag{10.53}$$

The graphs of these shape functions are straight lines and they have fundamental properties

$$M_i(z) = \left\{ \begin{array}{ccc} 1 & ; & z = z_{i-1} \\ 0 & ; & z = z_i \end{array} \right\} \tag{10.54}$$

$$N_i(z) = \left\{ \begin{array}{ccc} 0 & ; & z = z_{i-1} \\ 1 & ; & z = z_i \end{array} \right\} \tag{10.55}$$

This allows us to express $\widehat{u}_i(z)$ as

$$\begin{aligned} \widehat{u}_i(z) &= \widehat{u}_{i-1} M_i(z) + \widehat{u}_i N_i(z) \\ i &= 1, 2, ...n-1 \end{aligned}$$

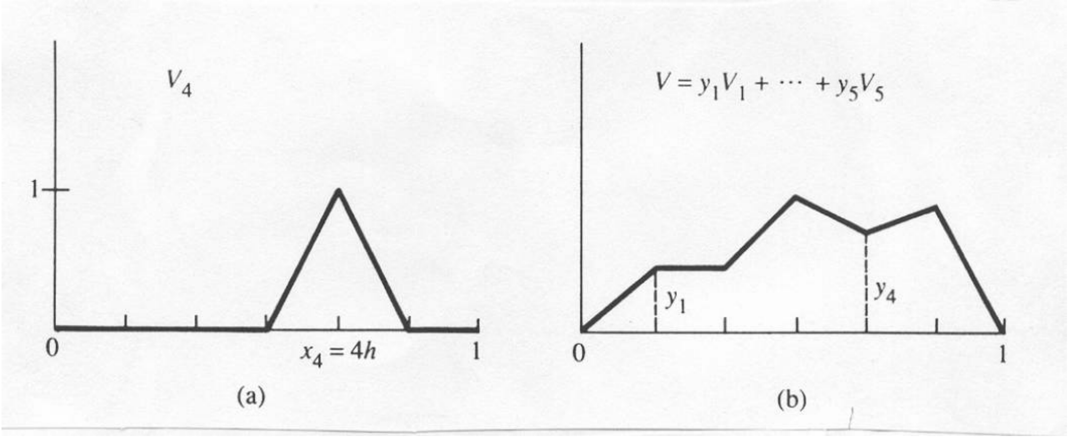


FIGURE 10

Note that the coefficient $\widehat{u}_i$ appears in polynomials $\widehat{u}_i(z)$ and $\widehat{u}_{i+1}(z)$, i.e.

$$\begin{aligned} \widehat{u}_i(z) &= \widehat{u}_{i-1}M_i(z) + \widehat{u}_iN_i(z) \\ \widehat{u}_{i+1}(z) &= \widehat{u}_iM_{i+1}(z) + \widehat{u}_{i+1}N_{i+1}(z) \end{aligned}$$

Thus, we can define a continuous *trial function* by combining $N_i(z)$ and $M_{i+1}(z)$ as follows

$$\begin{aligned} (10.56) \quad v^{(i)}(z) &= \left\{ \begin{array}{lc} N_i(z) = \dfrac{z - z_{i-1}}{\Delta z} & ; \quad z_{i-1} \leq z \leq z_i \\ -M_{i+1}(z) = 1 - \dfrac{z - z_i}{\Delta z} & ; \quad z_i \leq z \leq z_{i+1} \\ 0 & \text{Elsewhere} \end{array} \right\} \\ i &= 1, 2, ...n - 1 \end{aligned}$$

The simplest and most widely used are these piecewise linear functions (or hat function) as shown in Figure 10. This is a continuous linear function of z, but, it is not differentiable at $z_{i-1,}z_{i,}$and z_{i+1}. Also, note that at $z = z_i$, we have

$$\begin{aligned} (10.57) \qquad v^{(i)}(z_i) &= \left\{ \begin{array}{ll} 1 & \text{if } \ i = j \\ 0 & \text{otherwise} \end{array} \right\} \\ i &= 1, ...n - 1 \end{aligned}$$

Thus, plot of this function looks like a symmetric triangle. The two functions at the boundary points are defined as ramps

$$(10.58) \qquad v^{(0)}(z) = \left\{ \begin{array}{lc} -M_1(z) = 1 - \dfrac{z}{\Delta z} & ; \quad 0 \leq z \leq z_1 \\ 0 & \text{Elsewhere} \end{array} \right\}$$

$$v^{(n)}(z) = \left\{ \begin{array}{ll} N_n(z) = \dfrac{z - z_{n-1}}{\Delta z} & ; \quad z_{n-1} \leq z \leq z_n \\ 0 & \text{Elsewhere} \end{array} \right\} \tag{10.59}$$

Introduction of these trial functions allows us to express the approximate solution as

$$\widehat{u}(z) = \widehat{u}_0 v^{(0)}(z) + ... + \widehat{u}_n v^{(n)}(z) \tag{10.60}$$

and now we can work with $\widehat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_0 & \widehat{u}_2 & ... & \widehat{u}_n \end{bmatrix}^T$ as unknowns. The optimum parameters $\widehat{\mathbf{u}}$ can be computed by solving equation

$$A\widehat{\mathbf{u}} - \mathbf{b} = \overline{0} \tag{10.61}$$

where

$$(A)_{ij} = \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(j)}}{dz} \right\rangle \tag{10.62}$$

and

$$\frac{dv^{(i)}}{dz} = \left\{ \begin{array}{ll} 1/\Delta z & \text{on interval left of } z_i \\ -1/\Delta z & \text{on interval right of } z_i \end{array} \right\}$$

If intervals do not overlap, then

$$\left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(j)}}{dz} \right\rangle = 0 \tag{10.63}$$

The intervals overlap when

$$i = j \ : \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i)}}{dz} \right\rangle = \int_{z_{i-1}}^{z_i} (1/\Delta z)^2 dz + \int_{z_{i-1}}^{z_i} (-1/\Delta z)^2 dz = 2/\Delta z \tag{10.64}$$

or

$$i = j+1 : \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i-1)}}{dz} \right\rangle = \int_{z_{i-1}}^{z_i} (1/\Delta z).(-1/\Delta z) dz = -1/\Delta z \tag{10.65}$$

$$i = j-1 : \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i+1)}}{dz} \right\rangle = \int_{z_{i-1}}^{z_i} (1/\Delta z).(-1/\Delta z) dz = -1/\Delta z \tag{10.66}$$

Thus, the matrix A is a tridiagonal matrix

$$A = 1/\Delta z \begin{bmatrix} 2 & -1 & .... & .... & 0 \\ -1 & 2 & -1 & .... & ... \\ .... & .... & .... & .... & ... \\ 0 & .... & .... & -1 & 2 \end{bmatrix} \tag{10.67}$$

which is similar to the matrix obtained using finite difference method. The components of vector **b** on **the** R.H.S. is computed as

$$b_i = \left\langle v^{(i)},\ f(z) \right\rangle \tag{10.68}$$

$$= \int\limits_{z_{i-1}}^{z_i} f(z)\frac{z - z_{i-1}}{\Delta z} dz + \int\limits_{z_{i-1}}^{z_i} f(z) \left(1 - \frac{z - z_i}{\Delta z} \right) dz \tag{10.69}$$

which is a weighted average of $f(z)$ over the interval $z_{i-1} \leq z \leq z_{i+1}$. Note that the R.H.S. is different from finite difference method.

The Raleigh-Ritz method can be easily applied to problems in higher dimensions when the operators are self-adjoin. Consider Laplace / Poisson's equation

$$Lu = -\partial^2 u/\partial x^2 - \partial^2 u/\partial y^2 = f(x, y) \tag{10.70}$$

in open set S and $u(x, y) = 0$ on the boundary. Let the inner product on the space $C^{(2)}[0, 1] \times C^{(2)}[0, 1]$ be defined as

$$\langle f(x, y), g(x, y) \rangle = \int\limits_0^1 \int\limits_0^1 f(x, y)\ g(x, y)\ dxdy \tag{10.71}$$

We formulate and optimization problem

$$\Phi(u) = 1/2 \left\langle u(x, y), -\partial^2 u/\partial x^2 - \partial^2 u/\partial y^2 \right\rangle - \langle u(x, y), f(x, y) \rangle \tag{10.72}$$

Integrating by parts, we can show that

$$\Phi(u) = \int \int [1/2(\partial u/\partial x)^2 + 1/2(\partial u/\partial y)^2 - fu] dxdy \tag{10.73}$$

$$= (1/2) \langle \partial u/\partial x, \partial u/\partial x \rangle + 1/2 \langle \partial u/\partial y, \partial u/\partial y \rangle - \langle f(x, y), u(x, y) \rangle \tag{10.74}$$

We begin by choosing $(n - 1) \times (n - 1)$ equidistant (with $\Delta x = \Delta y = h$) internal node (grid) points at (x_i, y_j) where

$$\begin{aligned} x_i &= ih \quad (i = 1, 2, ...n - 1) \\ y_j &= ih \quad (j = 1, 2, ...n - 1) \end{aligned}$$

In two dimension, the simplest element divides region into triangles on which simple polynomials are fitted. For example, $u(x,y)$ can be approximated as

$$\widehat{u}(x,y) = a + bx + cy \tag{10.75}$$

where vertices a, b, c can be expressed in terms of values of $\widehat{u}(x,y)$ at the triangle vertices. For example, consider triangle defined by (x_i, y_j), (x_{i+1}, y_j) and (x_i, y_{j+1}). The value of the approximate solution at the corner points is denoted by

$$\widehat{u}_{i,j} = \widehat{u}(x_i, y_j) \ ; \widehat{u}_{i+1,j} = \widehat{u}(x_{i+1}, y_j) \ ; \ \widehat{u}_{i,j+1} = \widehat{u}(x_i, y_{j+1})$$

Then, $\widehat{u}(x,y)$ can be written in terms of shape functions as follows

$$\begin{aligned} \widehat{u}(x,y) &= \widehat{u}_{i,j} + \frac{\widehat{u}_{i+1,j} - \widehat{u}_{i,j}}{h}(x - x_{i,j}) + \frac{\widehat{u}_{i,j+1} - \widehat{u}_{i,j}}{h}(y - y_{i,j}) \qquad (10.76) \\ &= \widehat{u}_{i,j}\left[1 - \frac{(x - x_{i,j})}{h} - \frac{(y - y_{i,j})}{h}\right] \\ &\quad + \widehat{u}_{i+1,j}\left[\frac{(x - x_{i,j})}{h}\right] + \widehat{u}_{i,j+1}\left[\frac{(y - y_{i,j})}{h}\right] \qquad (10.77) \end{aligned}$$

Now, coefficient $\widehat{u}_{i,j}$ appears in the shape functions of four triangular element around (x_i, y_j). Collecting these shape functions, we can define a two dimensional trial function as follows

(10.78)

$$v^{(i,j)}(z) = \left\{ \begin{array}{lc} 1 - \dfrac{(x - x_{i,j})}{h} - \dfrac{(y - y_{i,j})}{h} & ; \ x_i \leq x \leq x_{i+1} \ ; \ y_j \leq y \leq y_{j+1} \\ 1 + \dfrac{(x - x_{i,j})}{h} - \dfrac{(y - y_{i,j})}{h} & ; \ x_{i-1} \leq x \leq x_i \ ; \ y_j \leq y \leq y_{j+1} \\ 1 - \dfrac{(x - x_{i,j})}{h} + \dfrac{(y - y_{i,j})}{h} & ; \ x_i \leq x \leq x_{i+1} \ ; \ y_{j-1} \leq y \leq y_j \\ 1 + \dfrac{(x - x_{i,j})}{h} + \dfrac{(y - y_{i,j})}{h} & ; \ x_{i-1} \leq x \leq x_i \ ; \ y_{j-1} \leq y \leq y_j \\ 0 & \text{Elsewhere} \end{array} \right\}$$

The shape of this trial function is like a pyramid (see Figure 11). We can define trial functions at the boundary points in a similar manner. Thus, expressing the approximate solution using trial functions and using the fact that $\widehat{u}(x,y) = 0$ at the boundary points, we get

$$\widehat{u}(x,y) = \widehat{u}_{1,1}v^{(1,1)}(x,y) + ... + \widehat{u}_{n-1,n-1}v^{(n-1,n-1)}(x,y)$$

where $v^{(i,j)}(z)$ represents the (i,j)'th trial function. For the sake of convenience, let us re-number these trial functions and coefficients using a new index $l =$

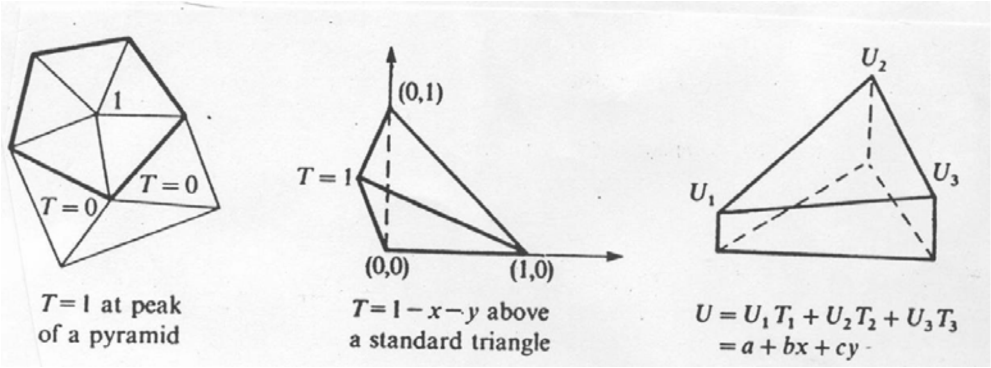


FIGURE 11

$0, 1, \ldots\ldots, N$ such that

$$\begin{aligned} l &= i + (n-1)j \\ i &= 1, \ldots n-1 \text{ and } j = 0, 1, \ldots n-1 \\ N &= (n-1) \times (n-1) \end{aligned}$$

The approximate solution can now be expressed as

$$\widehat{u}(x,y) = \widehat{u}_0 v^0(x,y) + \ldots. + \widehat{u}_N v^N(x,y) \tag{10.79}$$

The minimization problem an be reformulated as

$$\underset{\widehat{\mathbf{u}}}{Min}\ \Phi(\widehat{u}) = \underset{\widehat{\mathbf{u}}}{Min}\ \left[\frac{1}{2}\left\langle \frac{\partial \widehat{u}}{\partial x}, \frac{\partial \widehat{u}}{\partial x} \right\rangle + \frac{1}{2}\left\langle \frac{\partial \widehat{u}}{\partial y}, \frac{\partial \widehat{u}}{\partial y} \right\rangle - \langle f(x,y), \widehat{u}(x,y) \rangle \right] \tag{10.80}$$

where

$$\widehat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_0 & \widehat{u}_2 & \ldots & \widehat{u}_N \end{bmatrix}^T \tag{10.81}$$

Thus, the above objective function can be reformulated as

$$\underset{\widehat{\mathbf{u}}}{Min}\ \Phi(\widehat{\mathbf{u}}) = \underset{\widehat{\mathbf{u}}}{Min}\ \left(1/2\widehat{\mathbf{u}}^T A \widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T \mathbf{b}\right)$$

where

$$(A)_{ij} = (1/2)\left\langle \partial v^{(i)}/\partial x, \partial v^{(j)}/\partial x \right\rangle + (1/2)\left\langle \partial v^{(i)}/\partial y, \partial v^{(j)}/\partial y \right\rangle \tag{10.82}$$

$$b_i = \left\langle f(x,y), v^{(j)}(x,y) \right\rangle \tag{10.83}$$

Again, the matrix A is symmetric and positive definite matrix and this guarantees that stationary point of $\Phi(\mathbf{u})$ is the minimum. At the minimum, we have

$$\partial\Phi/\partial\widehat{\mathbf{u}} = A\widehat{\mathbf{u}} - \mathbf{b} = 0 \tag{10.84}$$

The matrix A will also be a sparse matrix. The main limitation of Relay-Ritz method is that it works only when the operator L is *symmetric* or self adjoint.

10.3.2. *Gelarkin's method.* The Gelarkin's method can be applied for any problem where differential operator is not self adjoint or symmetric. Instead of minimizing $\Phi(\widehat{\mathbf{u}})$, we solve for

$$\begin{aligned} \left\langle v^{(i)}(z), L\widehat{u}(z)\right\rangle &= \left\langle v^{(i)}(z), f(z)\right\rangle \\ i &= 0, 2, ...n \end{aligned} \tag{10.85}$$

where $\widehat{u}(z)$ is chosen as finite dimensional approximation to $u(z)$

$$\widehat{u}(z) = \widehat{u}_0 v^{(0)}(z) + ... + \widehat{u}_n v^{(n)}(z) \tag{10.86}$$

Rearranging above equations as

$$\left\langle v^{(i)}(z), (L\widehat{u}(z) - f(z))\right\rangle = 0\,; \quad (i = 0, 1, 2, ...n) \tag{10.87}$$

we can observe that parameters $u_0, ...c_n$ are computed such that the error or residual vector

$$e(z) = (L\widehat{u}(z) - f(z)) \tag{10.88}$$

is perpendicular to the subspace $(n+1)$ dimensional subspace spanned by set S defined as

$$S = \left\{v^{(i)}(z)\ :\ i = 0, 1, 2, ...n\right\} \tag{10.89}$$

This results in a linear algebraic equation of the form

$$A\widehat{\mathbf{u}} = \mathbf{b} \tag{10.90}$$

where

$$A = \begin{bmatrix} \left\langle v^{(0)}, L(v^{(0)})\right\rangle & .... & \left\langle v^{(1)}, L(v^{(n)})\right\rangle \\ .... & .... & .... \\ \left\langle v^{(n)}, L(v^{(0)})\right\rangle & .... & \left\langle v^{(n)}, L(v^{(n)})\right\rangle \end{bmatrix} \tag{10.91}$$

$$\mathbf{b} = \begin{bmatrix} \left\langle v^{(0)}(z), f(z)\right\rangle \\ .... \\ \left\langle v^{(n)}(z), f(z)\right\rangle \end{bmatrix}$$

Solving for $\widehat{\mathbf{u}}$ gives approximate solution given by equation (10.86).When the operator is L self adjoint, the Gelarkin's method reduces to the Raleigh-Ritz method.

EXAMPLE 65. *Consider ODE-BVP*

$$Lu = \partial^2 u/\partial z^2 - \partial u/\partial z = f(z) \tag{10.92}$$

$$in\ (\ 0 < x < 1) \tag{10.93}$$

$$subject\ to\ u(0) = 0;\ u(1) = 0 \tag{10.94}$$

It can be shown that

$$L^*(= \partial^2/\partial z^2 + \partial/\partial z) \neq (\partial^2/\partial z^2 - \partial/\partial z) = L$$

Thus, Raleigh-Ritz method cannot be applied to generate approximate solution to this problem, however, Gelarkin's method can be applied.

EXAMPLE 66. [**6**]*Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out.*

$$LC = \frac{1}{Pe}\frac{d^2C}{dz^2} - \frac{dC}{dz} = DaC^2 \qquad (0 \leq z \leq 1) \tag{10.95}$$

$$\frac{dC}{dz} = Pe(C-1) \qquad at \quad z = 0; \tag{10.96}$$

$$\frac{dC}{dz} = 0 \qquad at \quad z = 1; \tag{10.97}$$

The approximate solution is chosen as

$$\widehat{C}(z) = \widehat{C}_0 v^{()}(z) + ... + \widehat{C}_n v^{(n)}(z) = \sum_{i=0}^{n} \widehat{C}_i v^{(i)}(z) \tag{10.98}$$

and elements of matrix A are computed as

$$A_{ij} = \left\langle v^{(i)}, L(v^{(j)}) \right\rangle \tag{10.99}$$

$$b_i = \left\langle v^{(i)}, Da\left(\sum_{i=0}^{n} \widehat{C}_i v^{(i)}(z)\right)^2 \right\rangle \tag{10.100}$$

and using the boundary conditions. This yields nonlinear algebraic equations, which have to be solved simultaneously to compute the unknown coefficients $\widehat{C}_0, ...\widehat{C}_n$.

11. Summary

In these lecture notes, we have presented various numerical schemes based on multivariate unconstrained optimization formulation. One of the major application of unconstrained optimization is function approximation or multivariate regression. Thus, we begin by providing a detailed description of the model parameter estimation problem. We then derive necessary and sufficient conditions for optimality for a general multivariable unconstrained optimization problem. If the model has some nice structure, such as it is linear in parameters, then the parameter estimation problem can be solved analytically. Thus, the linear model parameter estimation (linear regression) problem is treated elaborately. Geometric and statistical properties of the linear least square problem are discussed in detail to provide further insights into this formulation. Numerical methods for estimating parameters of the nonlinear-in-parameter models are presented next. Other applications of optimization formulations, such as solving nonlinear algebraic equations and finite element method for solving PDEs and ODE-BVPs, have been discussed at the end.

12. Exercise

(1) Square the matrix $P = aa^T/a^Ta$, which projects onto a line and show that $p^2 = p$. Is projection matrix invertible?

(2) Compute a matrix that projects every point in the plane onto line $x + 2y = 0$.

(3) Solve $Ax = b$ by least square and find $p = Ax$ if

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

verify that $b - p$ is perpendicular to the columns of A.

(4) Given the measurements $b_1, .., b_n$ at distant points $t_1, ..., t_n$. Show that the straight line

$$b = C + Dt \; + e$$

which minimizes $\sum e^2$ comes from the least squares:

$$\begin{bmatrix} n & \sum t_i \\ \sum t_i & (\sum t_i)^2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \sum b \\ \sum b_i t_i \end{bmatrix}$$

(5) The following system has no solution

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 9 \end{bmatrix}$$

sketch and solve a straight line fit that leads to minimization of the quadratic

$$(C-D-4)^2+(C-5)^2+(C+D-9)^2$$

(6) Find the best straight line fit to the following measurements, and sketch your solution

$$\begin{aligned} y &= 2\ at\ t=-1\ \ ;\ \ y=0\ at\ t=0 \\ y &= -3\ at\ t=-1\ \ ;\ y=-5\,at\ t=2 \end{aligned}$$

(7) Suppose that instead of a straight line, we fit in the previous exercise by a parabola:

$$y=C+Dt+Et^2$$

In the inconsistent set of systems $A\mathbf{x}=\mathbf{b}$ that comes from the four measurements, what are the coefficient matrix A, the unknown vector x and the data vector $\mathbf{b}$? Compute the least square solution.

(8) (a) If P is the projection matrix onto k-dimensional subspace S of the whole space R^n, then what is the column space of P and its rank?

(b) If $P=A(A^TA)^{-1}A^T\mathbf{b}$ is the projection onto the column space of A, what is the projection of vector $\mathbf{a}$ onto row space of A?

(9) It is desired to fit the experimental data in the table below with a second -degree polynomial of the form

$$y=a_0+a_1x+a_2x^2$$

using least squares. It is known from external information that the fitted equation should have a maximum value at $x=0.6$ and it should pass through the point $x=1, y=0$. Determine the fitted equation that satisfies these requirements.

x :	0.2	0.6	0.8	1.0
y :	0.7	2.0	1.0	0.5

(10) It is desired to fit the heat capacity data for methylecyclohexane (C_7H_{14}) to a polynomial function in temperature

model 1: $c_p=a+bT$

model 2: $c_p=a+bT+cT^2$

where c_p is the heat capacity, T is the absolute temperature.

- Determine the model parameters in each case using least square estimation method using the following data:
- Formulate the problem as minimization of $||e||_2 = ||Ac - Y||_2$ and find the projection matrix P_r that projects vector Y onto the column space of A.
- Find the model prediction vector $\overline{Y} = Ac$ and the error vector $e = Y - \overline{Y}$ using the projection matrix.
- Assuming that the errors are normally distributed, estimate the covariance matrix of the estimated parameters and correlation coefficient in each case and compare the two models.

$c_p(KJ/kgK)$	$T(K)$	$c_p(KJ/kgK)$	$T(K)$
1.426	150	1.627	230
1.469	170	1.661	240
1.516	190	1.696	250
1.567	210	1.732	260

(11) Very common model for a dimensionless first-order chemical reaction is

$$dC/dt = -kC$$

with $C(0) = 1$. The integrated form of this model is $C = exp(-kt)$, which is non linear in the parameter k. Solve this problem by first transforming the model to linear form to first estimate of k to be used in calculating the nonlinear value. How much different are the linear and nonlinear values of k

t	0.2	0.5	1	1.5	2
C	0.75	0.55	0.21	0.13	0.04

Reconsider the above problem. Since we do not have a measurement of the initial concentration at $t = 0$, it is possible that there is a systematic error (bias) in the concentrations, so that C(0) is not equal to unity. To test for this possibility, extend the previous model to form $C = A\,exp(-kt)$, where both A and k are to be determined. A simple test for bias involves re-computing to see if the two-parameter model causes the value of "Error" to be significantly reduced. First transform the original model to a linear form and estimate the model parameters. Using these estimates of the parameters, solve the nonlinear model. Is there evidence of a significant bias?

(12) Use all of the data given in the following table to fit the following two-dimensional models for diffusion coefficient D as a function of temperature (T) and weight fraction (X).

T(^0C)	20	20	20	25	25	25	30	30	30
X	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
D $\times 10^5 cm^2/s$	0.823	0.639	0.43	0.973	0.751	0.506	1.032	0.824	0.561

Model 1 : $D = c_1 + c_2T + c_3X$

Model 2 : $D = c_1 + c_2T + c_3X + c_2T^2 + c_5TX + c_6X^2$

In each case, calculate D at $T = 22, X = 0.36$ and compare the two results.

(13) Weighed least square method is to be used in the following linearizable model

$$y_1 = \exp(-\theta_1 x_{1,i} - \theta_2/x_{2,i}); \qquad i = 1, 2, , ...n$$

Estimate the weights that should be used to formulate the weighted least square problem.

(14) Consider the problem of minimizing the following functions

- Function 1

$$f(x) = 4[(x_1)^2 + (x_2)^2] - 2x_1x_2 - 6(x_1 + x_2)$$

Initial guess $x^{(0)} = [3 \;\; 2]^T$

- Rosenbrok banana function

$$f(x) = [(x_1)^2 - (x_2)^2 + (1 - x_1)^2]$$

Initial guess $x^{(0)} = [-0.5 \;\; 0.5 \; 2]^T$

Determine the minimum by the following approaches (a) Analytical approach (b) Steepest descent method (with 1-D minimization for step length) (c) Conjugate gradient method (with 1-D minimization for step length) (d) Newton's optimization method

(15) Apply Newton's optimization method and perform 3 iterations each (without 1-D minimization and t$^{(k)}$ = 1 for all k) to minimize

- $f(x) = x_1 - x_2 + 2(x_1)^2 + 2x_1x_2 + (x_2)^2$
  $x^{(0)} = [0 \; 0]^T$
- f(x)= -1/$[(x_1)^2 + (x_2)^2 + 2]$
  $x^{(0)} = [4 \; 0]^T$

Comment upon the convergence of iterations in case (ii). What measure will ensure convergence?

(16) Solve the following equations using Steepest descent method and Newton's optimization method

$$2x_1 + x_2 = 4; x_1 + 2x_2 + x_3 = 8;\ x_2 + 3x_3 = 11$$

by taking the starting point as $x = [0\ 0\ 0]^T$.

(17) In optimization algorithms, instead of performing 1-D minimization to find step length $\lambda^{(k)}$ at each iteration, an alternate approach is to choose $\lambda^{(k)}$ such that

$$f[x^{(k+1)}] = f[x^{(k)} + \lambda^{(k)} s^{(k)}] < f[x^{(k)}]$$

Develop an algorithm to choose $\lambda^{(k)}$ at each iteration step by the later approach.

Bibliography

[1] Atkinson, K. E.; An Introduction to Numerical Analysis, John Wiley, 2001.

[2] Axler, S., Linear Algebra Done Right, Springer, San Francisco, 2000.

[3] Bazara, M.S., Sherali, H. D., Shetty, C. M., Nonlinear Programming, John Wiley, 1979.

[4] Demidovich, B. P. and I. A. Maron; Computational Mathematics. Mir Publishers, Moskow, 1976.

[5] Gourdin, A. and M Boumhrat; Applied Numerical Methods. Prentice Hall India, New Delhi.

[6] Gupta, S. K.; Numerical Methods for Engineers. Wiley Eastern, New Delhi, 1995.

[7] Kreyzig, E.; Introduction to Functional Analysis with Applications,John Wiley, New York, 1978.

[8] Linfield, G. and J. Penny; Numerical Methods Using MATLAB, Prentice Hall, 1999.

[9] Linz, P.; Theoretical Numerical Analysis, Dover, New York, 1979.

[10] Luenberger, D. G.; Optimization by Vector Space Approach , John Wiley, New York, 1969.

[11] Luenberger, D. G.; Optimization by Vector Space Approach , Joshn Wiley, New York, 1969.

[12] Moursund, D. G., Duris, C. S., Elementary Theory and Application of Numerical Analysis, Dover, NY, 1988.

[13] Rao, S. S., Optimization: Theory and Applications, Wiley Eastern, New Delhi, 1978.

[14] Rall, L. B.; Computational Solutions of Nonlinear Operator Equations. John Wiley, New York, 1969.

[15] Strang, G.; Linear Algebra and Its Applications. Harcourt Brace Jevanovich College Publisher, New York, 1988.

[16] Strang, G.; Introduction to Applied Mathematics. Wellesley Cambridge, Massachusetts, 1986.

[17] Vidyasagar, M.; Nonlinear Systems Analysis. Prentice Hall, 1978.